Understanding Freeway Crashes through Data-Driven Solutions

John Eugene Ash

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Yinhai Wang (Chair)

Xuegang (Jeff) Ban

Ed McCormack

Program Authorized to Offer Degree:

Civil & Environmental Engineering

University of Washington

Abstract

Understanding Freeway Crashes through Data-Driven Solutions

John Eugene Ash

Chair of the Supervisory Committee:
Yinhai Wang, Professor
Civil & Environmental Engineering

Traffic safety has been and continues to be one of the most active research areas within transportation engineering as government agencies consistently name safety their top priority. While fundamental problems in the field (e.g., crash frequency modeling) often remain the same, advances in statistical methodologies, data availability, and computing continue to enable new solutions to these problems, as well as options for framing these problems in a new and different manner. Notably, real-time crash prediction modeling (RTCPM) has been an area gaining attention over recent years. RTCPM studies the relationship between crash risk and changes in traffic conditions (measured by different sensors) over short-duration time periods; it thus assumes the occurrence of a crash is related to the traffic conditions occurring in some time period before the crash takes place. While several studies have indicated correlation between traffic conditions and crashes, there is still much work to be done especially when it comes to critical evaluation of appropriate study design and application of traffic sensing data to derive appropriate and

representative features describing traffic conditions. This dissertation examines this question, along with others related to crash frequency modeling as part of a broader effort to investigate and gain a better understanding of the nature of the relationship between traffic operations and crashes, as well as better understanding of variation in crash frequency estimates.

A key component of the RTCPM effort in this work is application of probe vehicle trajectory data derived from GPS trace points provided by mobile location services, consumer GPS devices, and commercial vehicle transponders. Such data have not been used in this application before (to the author's knowledge) and provide finer spatial/temporal measurement resolution than obtainable through conventional traffic sensing infrastructure (e.g., loop detectors). Use of this trajectory data also provides novelty in that it (1) only describes a sample of the traffic stream, so thus, there are questions as to if it can be used to make population-level inference and (2) the dataset is substantially larger than that used in previous studies, necessitating an efficient data processing method. The RTCPM component of this study takes a comprehensive look at study design, feature extraction, modeling techniques, and interpretation of results.

A final component of this dissertation focuses on how to better understand and account for variation in crash frequency modeling efforts. The bulk of existing studies produce point estimates for crash frequency, which only tell part of the story. At their core, crash frequency models produce estimates for a hierarchy of parameters, each of which can exhibit substantial variation. As such, this study derives confidence and prediction intervals for several types of mixed-Poisson models commonly used for crash frequency estimation in order to better capture and show the variation associated with crash estimates as one varies different factors. This study begins with the formulation of a mixed-Poisson model and discussion of several key mixture distributions used in

crash frequency modeling efforts. Then, the intervals are derived based on the variance of the safety (also known as the Poisson parameter), and a case study is presented for a real crash dataset to show how the method can be applied, as well to demonstrate the variation in estimates between and within models.

TABLE OF CONTENTS

Chapter 1. Background 1

1.1 Problem Statement 5

1.2 Research Objectives 7

1.3 Overview of Document 8

Chapter 2. State of the art 10

2.1 Real-Time Crash Detection 10

2.2 Exposure and Consideration of Traffic Operations in Crash Frequency Modeling 18

2.3 Crash Frequency Modeling 23

Chapter 3. Data Resources 28

3.1 Operational Data 28

3.1.1 Overview of Vehicle Trajectory Data 28

3.1.2 Trajectory Data Processing (Map Matching and Conflation) 34

3.1.3 Some Notes on Trajectory Data 46

3.2 Crash Data and Roadway Geometrics 49

3.2.1 Crash Report Data from Law Enforcement (Disaggregate) 49

3.2.2 Aggregate Crash Data for Crash Frequency Analysis 50

3.2.3 Ancillary Data 52

Chapter 4. Real-Time Crash Prediction Study Design 53

4.1 Problem Statement and Motivation 53

4.2 Case-Control Study Design 55

4.3 Modeling Methodology 56

4.4 Data Preparation and Feature Design 62

4.4.1 Data Overview 62

4.4.2 Data Preprocessing 63

4.4.3 Considerations for Study Design 63

4.4.4 Feature Design 69

4.4.5 Modeling Process and Data Summary 72

Chapter 5. Real-Time Crash Prediction Modeling Results 89

5.1 Presentation of Results For Initial Models 89

5.2 Discussion and Interpretation of Results 97

5.3 A Deeper Dive into Real-Time Crash Prediction Modeling 99

5.4 Conclusion and Extensions 113

Chapter 6. Development of Confidence and Prediction intervals for mixed-Poisson Regression models 114

6.1 Problem Statement and Motivation 114

6.2 Derivation of Confidence and Prediction Models 115

6.2.1 Mixed-Poisson Models and Formulation 115

6.2.2 Parametrizations of the Mixture Distributions 117

6.2.3 Derivation of Confidence Intervals for Poisson mean (True mean Crash Frequency) (μ) 119

6.2.4 Derivation of Prediction Intervals for Poisson Parameter (m)............................... 121

6.2.5 Derivation of Prediction Intervals for Predicted Crash Count (y) 124

6.3 Case Study 125

6.3.1 Model Development.. 126

6.3.2 Confidence and Prediction Intervals 127

6.4 Summary and Conclusions 135

Chapter 7. Conclusions and Future work.. 138

References.. 141

Appendix.. 148

Summary Data Tables for RTCPM Models in Chapter 4 148

Phase I Model Results Not Shown in Section 5.2 166

Derivation of CI-PIs.. 178

List of Figures

Figure 2-1 Types of Detection used in Real-Time Crash Prediction Models (based on Hossain et al. (2019)) 12

Figure 3-1 Bounding Box for INRIX Data Request (Seattle-Area Freeways), Image via Mapbox/© OpenStreetMap contributors 30

Figure 3-2 Distribution of Trips by Provider Type 32

Figure 3-3 Distribution of Trips by Vehicle Weight Class 33

Figure 3-4 Distribution of Trip Waypoint Sampling Frequency 33

Figure 3-5 INRIX Waypoints Visualized in WGS 84 Coordinate System (credit: © OpenStreetMap contributors) 35

Figure 3-6 OSM Relation 1071195 (I-405, both directions) (credit: © OpenStreetMap contributors) 38

Figure 3-7 OSM Way 1071195 (along I-405 SB) (credit: © OpenStreetMap contributors)38

Figure 3-8 OSM Node 2477186164 (along Way 1071195, I-405 SB) (credit: © OpenStreetMap contributors) 39

Figure 3-9 WSDOT Public Highway Network as Visualized in Q-GIS 40

Figure 3-10 Link along I-405 SB as Viewed in OSM (credit: © OpenStreetMap contributors) 41

Figure 3-11 Schematic of Relation between WSDOT and OSM Roadway Networks..... 43

Figure 3-12 Schematic of Trajectory Data Preparation Process (Seattle map via © OpenStreetMap contributors) 45

Figure 4-1 Array Representation of Case-Control Study Design Data 58

Figure 5-1 KDE Plots for 1.5 mi Up-stream (5-10 minutes), Control Definition #1 101

Figure 5-2 KDE Plots for 0.5 mi Down-stream (5-10 minutes), Control Definition #1. 102

Figure 5-3 KDE Plots for Upstream-Downstream Difference, Control Definition #1... 103

Figure 5-4 KDE Plots for 1.5 mi Up-stream (5-10 minutes), Control Definition #2 104

Figure 5-5 KDE Plots for 0.5 mi Down-stream (5-10 minutes), Control Definition #2. 105

Figure 5-6 KDE Plots for Upstream-Downstream Difference, Control Definition #2... 106

Figure 6-1 95% CIs and PI for Negative Binomial Model 130

Figure 6-2 95% CIs and PI for Poisson-Inverse-Gaussian Model 131

Figure 6-3 95% CIs and PI for Sichel Model 131

Figure 6-4 95% CIs and PI for Poisson-Lognormal Model 132

Figure 6-5 95% CIs and PI for Poisson-Weibull Model 132

List of Tables

Table 3-1 Data Elements in Trips Files 31

Table 3-2 Data Elements in Waypoints Files 31

Table 3-3 Initial Data Counts for Data Request (May 2017, Seattle Area)...................... 32

Table 3-4 Data Elements of PostGIS Tabular Representation of WSDOT 24k Road Map41

Table 3-5 Milepost Range for Study by Route Number and Direction 45

Table 3-6 Final Counts of Data after Trajectory Processing 45

Table 3-7 Distribution of May 2017 Crashes by Highway for Project Study Area.......... 49

Table 3-8 Summary of Data Elements in WSDOT-Provided Police Crash Report Data. 50

Table 3-9 Summary Statistics for Animal-Vehicle Collison Dataset 51

Table 4-1 Criteria for Different Datasets used in RTCPM Efforts.................................. 68

Table 4-2 Key Variables used in RTCPM Modeling.. 71

Table 4-3 Dataset for Control Def. #1 Upstream=0.5 mi, downstream=0.5mi (UP-stream portion only) 75

Table 4-4 Dataset for Control Def. #1 Upstream=0.5 mi, downstream=0.5mi (DOWN-stream portion only).. 76

Table 4-5 Dataset for Control Def. #2 Upstream=0.5 mi, downstream=0.5mi (UP-stream portion only) 77

Table 4-6 Dataset for Control Def. #2 Upstream=0.5 mi, downstream=0.5mi (DOWN-stream portion only).. 78

Table 4-7 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=0.5mi (UP-stream portion only) 79

Table 4-8 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=0.5mi (DOWN-stream portion only).. 80

Table 4-9 Dataset for Control Def. #2 Upstream=1.0 mi, downstream=0.5mi (UP-stream portion only) 81

Table 4-10 Dataset for Control Def. #2 Upstream=1.0 mi, downstream=0.5mi (DOWN-stream portion only).. 82

Table 4-11 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=0.5mi (UP-stream portion only).. 83

Table 4-12 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=0.5mi (DOWN-stream portion only)............ 84
Table 4-13 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=0.5mi (UP-stream portion only)............ 85
Table 4-14 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=0.5mi (UP-stream portion only)............ 86
Table 5-1 Model Results for Dataset 0.5 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #1 91
Table 5-2 Model Results for Dataset 0.5 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #1 91
Table 5-3 Model Results for Dataset 0.5 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #2 92
Table 5-4 Model Results for Dataset 0.5 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #2 92
Table 5-5 Model Results for Dataset 1.0 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #1 93
Table 5-6 Model Results for Dataset 1.0 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #1 94
Table 5-7 Model Results for Dataset 1.0 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #2 94
Table 5-8 Model Results for Dataset 1.0 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #2 95
Table 5-9 Model Results for Dataset 1.5 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #1 96
Table 5-10 Model Results for Dataset 1.5 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #1............ 96
Table 5-11 Model Results for Dataset 1.5 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #2 97
Table 5-12 Model Results for Dataset 1.5 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #2............ 97

Table 5-13 Summary of Significant Predictors (0.05 level) across the 12 Models in Section 5.1 98
Table 5-20 Primary Model for 1.5mi Upstream, 0.5mi Downstream Data (Control Def. #1) 110
Table 5-21 Secondary Model for 1.5mi Upstream, 0.5mi Downstream Data (Control Def. #1) 111
Table 5-22 Primary Model for 1.5mi Upstream, 0.5mi Downstream Data (Control Def. #2) 112
Table 5-23 Secondary Model for 1.5mi Upstream, 0.5mi Downstream Data (Control Def. #2) 112
Table 6-1 Variance of m for Mixture Distributions 122
Table 6-2 95% Prediction Intervals for m 123
Table 6-3 95% Predictions Intervals for y 125
Table 6-4 Model Results Estimated with ML Approach 127
Table 6-5 Model Results Estimated with Bayesian Approach 127
Table 6-6 Default Values of Variables used in Models for Interval Construction 128
Table 6-7 Summary of Values for Mixed-Poisson CIs and PIs at AADT=120,000 133
Table 6-8 MSE Values Calculated between μ and 95% CI and PI Lower and Upper Bounds for the Animal-Vehicle Collision Dataset 135
Table 0-1 Dataset for Control Def. #1 Upstream=0.5 mi, downstream=0.0mi 148
Table 0-2 Dataset for Control Def. #2 Upstream=0.5 mi, downstream=0.0mi 149
Table 0-3 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=0.0mi 150
Table 0-4 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=0.0mi 151
Table 0-5 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=0.0mi 152
Table 0-6 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=0.0mi 153
Table 0-7 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=1.0 mi (UP-stream portion only) 154
Table 0-8 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=1.0 mi (DOWN-stream portion only) 155
Table 0-9 Dataset for Control Def. #2 Upstream=1.0 mi, downstream=1.0 mi (UP-stream portion only) 156

Table 0-10 Dataset for Control Def. #2 Upstream=1.0mi, downstream=1.0 mi (DOWN-stream portion only)... 157

Table 0-11 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=1.0 mi (UP-stream portion only)... 158

Table 0-12 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=1.0 mi (DOWN-stream portion only)... 159

Table 0-13 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=1.0 mi (UP-stream portion only)... 160

Table 0-14 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=1.0 mi (DOWN-stream portion only)... 161

Table 0-15 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=1.5 mi (UP-stream portion only)... 162

Table 0-16 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=1.5 mi (DOWN-stream portion only)... 163

Table 0-17 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=1.5 mi (UP-stream portion only)... 164

Table 0-18 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=1.5 mi (DOWN-stream portion only)... 165

Table 0-19 Model Results for Dataset 0.5 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #1 166

Table 0-20 Model Results for Dataset 0.5 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #1 166

Table 0-21 Model Results for Dataset 0.5 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #2 167

Table 0-22 Model Results for Dataset 0.5 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #2 167

Table 0-23 Model Results for Dataset 1.0 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #1 168

Table 0-24 Model Results for Dataset 1.0 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #1 168

Table 0-25 Model Results for Dataset 1.0 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #2 169

Table 0-26 Model Results for Dataset 1.0 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #2 169

Table 0-27 Model Results for Dataset 1.5 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #1 170

Table 0-28 Model Results for Dataset 1.5 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #1 170

Table 0-29 Model Results for Dataset 1.5 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #2 171

Table 0-30 Model Results for Dataset 1.5 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #2 171

Table 0-31 Model Results for Dataset 1.0 mi Upstream, 1.0 mi downstream, ALL Crashes, Control Definition #1 172

Table 0-32 Model Results for Dataset 1.0 mi Upstream, 1.0 mi downstream, MV Crashes, Control Definition #1 172

Table 0-33 Model Results for Dataset 1.0 mi Upstream, 1.0 mi downstream, ALL Crashes, Control Definition #2 173

Table 0-34 Model Results for Dataset 1.0 mi Upstream, 1.0 mi downstream, MV Crashes, Control Definition #2 173

Table 0-35 Model Results for Dataset 1.5 mi Upstream, 1.0 mi downstream, ALL Crashes, Control Definition #1 174

Table 0-36 Model Results for Dataset 1.5 mi Upstream, 1.0 mi downstream, MV Crashes, Control Definition #1 174

Table 0-37 Model Results for Dataset 1.5 mi Upstream, 1.0 mi downstream, ALL Crashes, Control Definition #2 175

Table 0-38 Model Results for Dataset 1.5 mi Upstream, 1.0 mi downstream, MV Crashes, Control Definition #2 175

Table 0-39 Model Results for Dataset 1.5 mi Upstream, 1.5 mi downstream, ALL Crashes, Control Definition #1 176

Table 0-40 Model Results for Dataset 1.5 mi Upstream, 1.5 mi downstream, MV Crashes, Control Definition #1 176

Table 0-41 Model Results for Dataset 1.5 mi Upstream, 1.5 mi downstream, ALL Crashes, Control Definition #2 177

Table 0-42 Model Results for Dataset 1.5 mi Upstream, 1.5 mi downstream, MV Crashes, Control Definition #2 177

ACKNOWLEDGEMENTS

There are so many people who have helped me in innumerable ways throughout this journey to whom I owe a huge debt of gratitude. First and foremost, I thank Dr. Yinhai Wang who has served as my committee chair and advisor for the past several years. His mentorship and support have helped me grow as a student, researcher, and person in general, and I would not be where I am today without his encouragement and choice to take me on as a student. Second, I would like to thank my committee, Dr. Xuegang (Jeff) Ban, Dr. Ed McCormack, Dr. Anne Vernez-Moudon, and Dr. Simon (Shaolei) Du. I am very thankful for the insight and guidance they provided both on the dissertation and throughout my time at the University of Washington. I also must thank Ted Trepanier at INRIX for supplying me with the data used to complete much of this study.

Beyond those mentioned above, I am very grateful for the support provided to me by many others during my graduate studies. To Dr. David Noyce (at the University of Wisconsin) thank you for instilling a passion for transportation in me, for your continued support since I was an undergrad, and for being the role model you are. To Dr. Yajie Zou, thank you for helping me get my first paper published and allowing me to work with and learn from you along the way. To Dr. Zhixia (Richard) Li, thanks for all of your support during my MS in Wisconsin and helping me learn valuable research skills. To Dr. John Bohte, thanks for all of your support and guidance through this journey as well. Finally, to Dr. C. Mathews thank you for your guidance through this overall situation.

Within the Smart Transportation Applications and Research Laboratory (STAR Lab), I thank all I have worked with on anything from projects to coursework to papers over the past several years. I especially learned a lot and had a lot of fun working with Wenbo Zhu, Ruimin Ke,

Zhiyong Cui, Mayuree Binjolkar, Cole Kopca, and Ziyuan Pu, and I appreciate their friendship as well. Further, I cannot express how important the mentorship, guidance, and friendship provided by Kristian Henrickson has been for me. He, more than almost anyone else, has helped me through many things in my research journey and for that I truly thank him. An additional thanks to my friend, Allen Hernandez, who helped me in many cases with a host of computing issues.

Finally, and from the bottom of my heart, I would like to thank my partner Cynthia Chiou (and our dog Eddy), my parents, John and Catherine Ash, and my brother, Eric Ash, for their continued support and understanding over the past several years. I definitely could not have done it without you!

DISCLAIMER

Traffic and Driver Services Information Provided by INRIX, ©2021. All rights reserved by INRIX, Inc.

Some of the traffic data used in various analyses within this dissertation was provided by INRIX. All such cases are indicated within the text. Further, this dissertation is an independent publication and has not been authorized, sponsored, or otherwise approved by INRIX. All views on and interpretations of the data are solely those of the author and do not reflect the positions of INRIX. The name and wordmark "INRIX" are registered trademarks of INRIX.

Chapter 1. BACKGROUND

Traffic safety has been and continues to be one of the most active research areas within transportation engineering. This is not surprising as governments and the transportation agencies operating under them typically place a high priority on traffic safety. For example, the United States Department of Transportation (USDOT) lists safety as their top strategic goal in their most current strategic plan. Specifically, they are advocating for a "systemic safety approach" to decreasing fatalities and injuries associated with traffic crashes through numerous strategies including data-driven safety analyses, involving stakeholders in the safety-improvement process etc. (U.S. Department of Transportation 2018). Transportation agencies at various levels of government have almost universally adopted safety goals in alignment with those of USDOT. At the highest level, the Federal Highway Administration (FHWA) is embracing "a vision of zero deaths" on the country's highways (FHWA 2019). Additionally, many states and cities have adopted similar goals. For example, the Washington State Department of Transportation (WSDOT) refers to their program as "Target Zero," while the Seattle Department of Transportation named their program "Vision Zero" (SDOT 2019; WSDOT 2019).

While agencies are advocating for traffic safety programs to accomplish the zero transportation-associated-deaths goal, the fact of the matter is that achieving such a goal seems quite challenging when considering the current state of traffic safety. Based on data collected across the country, the National Highway Traffic Safety Administration (NHTSA) reported a total of 34,247 fatal traffic crashes resulting in 37,133 deaths in the year of 2017. In the same year, approximately 6.45 million crashes were reported to police, resulting in approximately 2.75 million injuries (NHTSA 2019). While these numbers (and the corresponding rates per 100 million vehicle miles traveled (VMT) and per 100,000 population) represent a slight decrease over the 2016 values, there is substantial room for safety improvement. In addition to the tremendous impact on human life, traffic crashes also affect mobility in terms of congestion and travel time reliability.

In order to help improve the state of traffic safety, researchers across the world have been working for decades on a variety of efforts. One major research direction is and has consistently been the use of statistical models to better understand crashes. Typically, such models are developed based upon crash data, often collected by law enforcement officers at the crash site and recorded in a crash report. Crash reports typically contain data on the location of the crash, the time of day, the type of crash (e.g., rear-end, sideswipe, etc.), the type(s) and number(s) of the vehicle(s) involved, the state of the occupants (i.e., number in each vehicle, seatbelt usage, age, intoxication status, etc.), injury types sustained and associated severity (if any), and many other factors. Other data that can also be used in the statistical modeling of crashes includes roadway geometrics and attributes (e.g., number of lanes, speed limit, presence/magnitude of vertical curvature, presence/magnitude of horizontal curvature, etc.), weather data, lighting data, and operational data such as data on speed/volume/occupancy from traffic sensors (Mannering and Bhat 2014).

The aforementioned models applying crash report data have typically sought to gain insight on factors associated with crash frequency and crash severity. Crash frequency models typically formulate a regression problem to model the dependent variable (crash frequency over some time period, often years, on a continuous scale for a given road segment) as a function of several covariates such as roadway geometric characteristics and traffic volumes (corresponding to the given segment) (Mannering and Bhat 2014). These models have applications in before/after analyses (e.g., to examine effectiveness of a countermeasure) and safety estimation for facilities (Bonneson and Ivan 2013). Crash severity models, on the other hand, formulate a classification problem where the dependent variable (discrete level of crash severity for a given crash) is modeled as a function of several covariates such as vehicle type, weather condition, occupant attributes (e.g., age) (Savolainen et al. 2011). Such models have numerous applications including but not limited to evaluating effectiveness of countermeasures. Numerous advances in the modeling of crash frequency and crash severity have been made been over time with advances in statistical methodologies, data availability, and computing.

In addition to the previously-described modeling efforts in traffic safety, another important research area is real-time crash prediction (Hossain et al. 2019; Roshandel et al. 2015). Real-time crash prediction models seek to predict the occurrence of a crash (alternately, to examine crash risk) as a function of several covariates and are typically framed as a binary classification problem (i.e., each data point used in model development either has or has not resulted in a crash). While the crash frequency and severity models typically apply a series of static factors in their formulation, real-time crash prediction often incorporate more dynamic, and microscopic features such as speed and occupancy as measured by loop detectors. Hossain et al. (2019) note that real-time crash prediction relies on an underlying belief that changes in traffic flow parameters over some spatial and temporal window can be used to predict the occurrence of a crash. This belief also highlights another important view of the real-time crash prediction problem, that being that rather than simply just predict the occurrence of a crash, one can get a better understanding of traffic patterns and sets of traffic flow parameters that are correlated with crashes. Insight gained from framing the problem as a means to better understand traffic patterns correlated with crashes has implications in control (e.g., variable speed limits) (Pande 2005).

Real-time crash, or even simply incident, prediction models are an important component of incident management programs for traffic operators due to the impact that traffic incidents have on traffic operations. As part of a project for the Strategic Highway Research Program 2, List et al. (2014) cited traffic incidents as one of the seven main sources of unreliable travel times. Another key source of decreased travel time reliability is changes in demand, a factor directly impacted by incidents when drivers choose to take alternate routes to bypass incidents. In terms of delay, incidents are estimated to contribute to approximately 60% of the total-vehicle hours of delay (Ozbay and Kachroo 1999). Knoop et al. showed that freeway capacity could be reduced by approximately 50% across all lanes due to an incident, even those in the opposite direction of travel (a result they attributed to rubbernecking). Ultimately, the nature of the relationship between traffic crashes and traffic operations is complex and appears to be bi-directional (i.e., crashes have an impact on operations and operations have an impact on crashes). As was true for the general trend of safety modeling, advances on the methodological front, as well as increased data availability, and rapid growth in computing technologies present new means to address, model, and fundamentally understand these complicated issues.

With regard to advances in data availability and computing, the increasing ubiquity of mobile phones and other mobile global positioning system (GPS) devices, traffic data are becoming available at a rate greater than ever before. Hererra et al. (2010) were among the first to make use of vehicle trajectory data derived from mobile phones in the "Mobile Century Field Experiment." They gathered data from a sample of 100 vehicles traveling on the I-880 freeway near Union City, California for an 8-hour period of study. With the data, they visualized trajectories and time-space plots, as well as compared speed estimation from the mobile GPS data with speed estimates from loop data. This study was instrumental in showing how GPS data obtained from mobile devices could be used to develop vehicle trajectories and use them in studies of traffic flow and operations. Since then, numerous studies have applied mobile GPS trajectory data, typically with a focus on traffic flow and operations (Herrera et al. 2010; Herring et al. 2010; Hofleitner et al. 2012b; a), but in some limited cases they have been applied to safety analysis as well (Stipancic et al. 2017, 2018a; b).

1.1 PROBLEM STATEMENT

This dissertation seeks to examine and better understand freeway crashes and their implications on traffic operations through a series of data-driven analyses. Further, this dissertation seeks to develop a means to better quantify uncertainty with estimates from safety models, notably mixed-Poisson models used in crash frequency applications.

A central component of several of these analyses will be the use and application of mobile GPS data as a means to construct vehicle trajectories and extract detailed information (i.e., at high level of spatial and temporal resolution). Specifically, several GB of GPS trace point data for the Seattle-area freeway network have been obtained via a partnership with INRIX, Inc. for use in this study. Such data is often referred to as probe data as the vehicles collecting it are samples, or probes, within the traffic stream. The source of the original GPS data points can vary depending on the data provider, however, mobile location services (e.g., within cellular phones), consumer GPS devices, and commercial vehicle transponders (e.g., in fleet vehicles, taxis, etc.) are common sources (Henrickson et al. 2019). While these data provide tremendous opportunity to study the spatial and temporal dynamics of traffic in new ways, they are not without their limitations. Notably, these data have issues surrounding missingness (depending on factors such as sampling rate, penetration rate, etc.), sampling bias/self-selection, and overall quality/accuracy of data that must be accounted for and well understood in order to address biases their use may result in

(Henrickson 2018; Henrickson et al. 2019). How effective use of data from a sample of the vehicle population may be in real-time crash prediction is a challenging question that remains to be answered.

As aforementioned, real-time crash prediction assumes a relationship between the probability of crash occurrence and the traffic conditions observed during some time period before the crash. While many studies have sought to investigate this issue and determine potentially influential factors impacting crash occurrence, there are still many issues that need to be addressed. First of all, to the best of the author's knowledge, only two previous studies have applied vehicle trajectory data for real-time crash prediction (Hourdos 2005; Hourdos et al. 2006). In these studies, the data was derived from video, not mobile GPS devices and the sample sizes of crashes/number of sites examined was relatively small. Hence, there is a gap to fill by applying this novel data source to a larger-scale freeway network and seeing (1) what features (especially new ones that cannot be obtained via loop data) can be derived from it and (2) how they may possibly be related to crash occurrence. The vast majority of other studies on real-time crash prediction have derived features describing traffic operations from loop detector data. This data is inherently constrained in spatial (where the loops are located and how they are spaced) and temporal resolution (what the level of aggregation is), hence, use of trajectory data will allow increased resolution along both of these dimensions allowing potential to custom define spatial/temporal windows of analysis with a high level of detail. Another key challenge related to real-time crash prediction that deserves critical attention is the issue of investigating how the study design itself, notably the use of/choice in case-control analyses, may influence results (Roshandel et al. 2015). There are numerous parameters to define/tune in the study design for real-time crash prediction models including how to define normal and pre-crash traffic conditions, how to select the case-control ratio, level of temporal aggregation of data, etc. Further, the issue of omitted variable bias, notably with respect to human-factors-related features also demands attention (Roshandel et al. 2015). With better understanding of these open issues, it is hoped that a stronger link can be established between traffic operations and safety.

Omitted variable bias is also a key issue in crash frequency modeling, and it is sometimes mentioned as a consequence of unobserved heterogeneity (Mannering and Bhat 2014). Upon performing a quick scan of the literature on crash frequency modeling, one will notice that models typically incorporate a variety of relatively static (i.e., they are not changing on a small time scale, such as on the order of seconds) features such as lane width, number of lanes, speed limit, etc. Examples of crash frequency models with parameters describing traffic operations, with exception of annual average daily traffic (AADT) or some similar measure of exposure, are quite limited. One way to better understand the potential for variation in crash frequency estimates is to expand beyond the notion of only looking at a point estimate as most typical models provide. Instead, one can investigate confidence and prediction intervals for different parameters in the hierarchy of mixed-Poisson models (a family that includes the negative binomial model, a common choice in crash frequency analyses). In this study, the author explores the key results derived in Ash et al. (2019), which explore a variety of commonly used mixed-Poisson model formulations, as well as application of the newly derived intervals to real crash frequency data.

1.2 RESEARCH OBJECTIVES

The primary objective of this dissertation is to investigate the nature of the relationship between traffic operations and crashes. While this relationship has been investigated in previous work, there are still many gaps to fill on both the real-time crash prediction and crash frequency modeling fronts. With regard to the other direction of the relationship, i.e., the impact of crashes on traffic operations, the problem is less well-studied. In addition to studying the aforementioned relationship to better understand freeway crashes, another objective of this dissertation is to investigate the value of vehicle trajectory data (collected from various GPS-enabled devices) in traffic safety analyses.

The specific proposed objectives for this dissertation are as follows:

- Develop a framework for real-time crash prediction based on detailed vehicle trajectory data and conventional and data-driven modeling strategies (a key component of this step will be feature design and extraction based on the large-scale GPS trajectory data);
- Critically analyze study design used in previous real-time cash prediction efforts and see how parameters of the design may impact results;
- Investigate the relationship between traffic operations and crash occurrence by developing a modeling framework that properly accounts for traffic dynamics;
- Demonstrate that data derived from a sample of probe vehicles can be used to produce results consistent with previous RTCPM studies that rely on substantially larger samples from loop data; and
- Study the issue of variance in crash frequency estimates and how to better describe it via derivation of confidence and prediction intervals for a variety of model types, and different hierarchical parameters in each model.

1.3 OVERVIEW OF DOCUMENT

This proposal document is arranged as follows. Chapter 2 presents the state-of-the art (i.e., background information and literature review) on the core topic areas of this dissertation to provide context, help frame the problems, and highlight some of the gaps in the research. Chapter 3 introduces the data resources available for this project and provides some summary information on each. A key part of this section is a brief summary of the vehicle trajectory data provided for use in this project. Chapter 4 presents initial discussion on the real-time crash prediction and issues related to the study design. Next, Chapter 5 presents results and interpretation for the real-time crash prediction models. Chapter 6 moves to focus on the crash frequency issue and how one can better understand the variance associated with crash prediction estimates through the use of derived confidence and prediction intervals. Finally, Chapter 7 draws conclusions from the study and suggests a few topics for future related work. An appendix also follows Chapter 7. In the appendix are summary data tables and model results for a selection of the real-time crash prediction models (as there are too many to include the main body) and derivations for the intervals in Chapter 6.

Chapter 2. STATE OF THE ART

The following sections provide a brief review of the literature and key background information on the core topic areas of this dissertation.

2.1 REAL-TIME CRASH DETECTION

As aforementioned, the primary focus of real-time crash detection is to detect or, more correctly stated, predict the occurrence of a crash based on a set of covariates, and most typically including several variables describing traffic flow parameters. As the outcome of any evaluation point is discrete (i.e., either the occurrence or non-occurrence of a crash), the problem is inherently framed as a classification problem. One of the first works to address this problem used a Bayesian non-parametric model that combined crash data and loop data to demonstrated how speed variation has a direct correlation with crash likelihood (Oh et al. 2001; Stylianou et al. 2019). Recently, several review articles have been published describing the previous work in real-time crash prediction, its limitations, and future directions. The following will provide a summary of those publications, followed by an overview of several important individual studies on real-time crash prediction.

Hossain et al. (2019) reviewed a series of existing studies on real-time crash prediction, provided a detailed analysis of study design, and put forth several requirements for future modeling efforts based on best-practices of the preceding work. They note how central to all such modeling efforts is a hypothesis or assumption that "the probability of a crash occurring on a specific road section within a very short time window can be predicted using the instantaneous traffic dynamics" (Hossain et al. 2019). Initial studies in the area focused simply on the prediction task, most often for freeways. More recent studies have expanded the scope and investigated issues including application of models in other areas (i.e., transferability), model development for types of road segments beyond the typical basic freeway segment (such as weaving areas), and incorporating severity into the prediction (Hossain et al. 2019).

A key contribution of Hossain et al. (2019) is their systematic definition of the typical components in the real-time crash prediction modeling process based on their review of more than 70 sources. While not explicitly mentioned, the first step is essentially choosing the spatial and temporal boundaries for the study area. Then, the analyst must define the variables to be used in their model(s) and collect the data from their study site(s). Next, the criteria to classify "pre-crash and normal traffic conditions" are established and data points mapped to one of these two categories. After the data are prepared, the analyst must select a model type (typically a binary classification model), train the model, and validate/measure its performance (Hossain et al. 2019).

The variable definition step is itself contingent upon what types of traffic flow data are available to the analyst, which itself depends on the available sensing infrastructure. Figure 2-1 shows the distribution of sensor types used in the 77 studies reviewed by Hossain et al. (2019). It can be seen that the vast majority of the studies used traffic flow data from loop detectors, but other studies also applied automatic vehicle identification (AVI) systems, Bluetooth sensors, microwave vehicle detection systems (MVDS), probe vehicles, remote traffic microwave sensors (RTMS), or some combination of the aforementioned, sometimes in conjunction with video or radar detection. Besides types of detectors, detector spacing and relative location of detectors with respect to the crash location are also important considerations in the study design. In their review of 77 articles, Hossain et al. (2019) found that more than one fourth of the studies did not discuss detector spacing, and for those that did, that the average spacing for the most commonly-used detector type, i.e., loop detectors, was 0.8 kilometers. Further, there was variation in which detectors were considered to pull data from. In addition to the nearest detector to the crash location, studies considered, detectors upstream of the crash, downstream of the crash, and in some cases both upstream and downstream (Hossain et al. 2019).

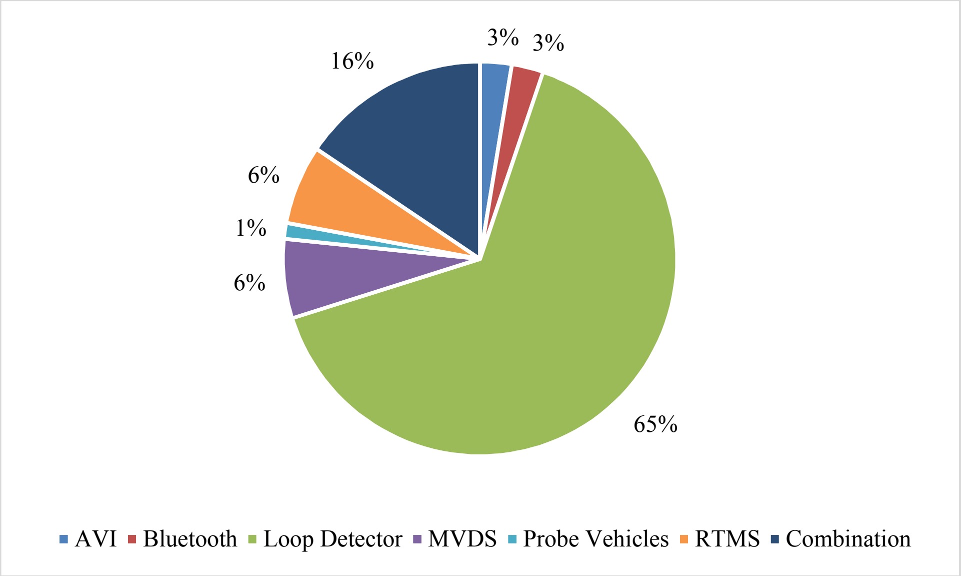


Figure 2-1 Types of Detection used in Real-Time Crash Prediction Models (based on Hossain et al. (2019))

Traffic conditions were classified as “pre-crash” or “normal” in many ways in different studies. One of the most common approaches was examining a 5-minute time period occurring between 5 and 10 minutes before the crash itself as the temporal data point describing pre-crash conditions (Hossain et al. 2019). Normal time periods were found to be defined based on a variety of criteria such as data measured no less than 30 minutes before the crash at the same detectors measuring the pre-crash condition, data measured 5 hours after the crash, as well as data chosen randomly in periods when no crash had occurred (Hossain et al. 2019).

For the variable selection step, as noted previously, traffic flow variables were the most commonly used type of variables in the models. Depending on the type of sensor(s) used and configuration of the sensor, different variables were considered. As the majority of studies applied loop detector data, they considered variables including flow, occupancy, and speed; other studies applying different sensor types included variables such as density and queue lengths (Hossain et al. 2019). Further, transformations and comparisons of variables (e.g., coefficient of variation of speed, the difference in a metric between upstream and downstream locations with respect to the crash, etc.) were also considered. Non-traffic flow related variables considered in some studies included roadway geometrics (e.g., number of lanes, grade, shoulder width, etc.) and weather effects (Hossain et al. 2019). Finally, the modeling efforts reviewed by Hossain et al. (2019) were classified into two main categories: classical statistical and artificial intelligence (AI)/data-mining based. For statistical models, logit and probit models were most common; neural networks, directed graphical models, classification trees, and support vector machines (SVM), were among the choices of alternate models used.

In addition to the review article by Hossain et al., (2019), a variety of other recent review articles have provided a great overview of advances in the field/state-of-the-practice. Stylianou et al. (2019) examined 48 studies on real-time crash prediction published between 2001-2018; all of the studies they consider apply data aggregated in intervals of between 20 seconds and 6 minutes. For each study, they summarized the key purpose, method(s) used, variables considered, the data source (including type of detection), as well as the main conclusions. They also note the main modeling approaches used in said studies, many of which overlap with those considered in Hossain et al. (2019), are either statistical or data-mining/AI-based. For the statistical approaches, they found that logistic regression of the regular or matched case control variety, as well as Bayesian models were the most commonly used. They also pointed out that one of the main tradeoffs between the statistical methods compared to the data-mining-based methods is that the former provides greater interpretability in terms of the impact of the independent variables often at the expense of limiting assumptions such as linear relations between dependent and independent variables, while the latter trades interpretability for often-improved accuracy (Stylianou et al. 2019). Types of detection found in the review by Stylianou et al. (2019) include loop detectors, MVDS, AVI, RTMS, radar, vehicle detection stations, video detectors, traffic message channel (TMC) data.

While the preceding two review articles give a good overview of data types, variables used, types of models used, and key conclusions of previous real-time crash prediction studies, they are not especially critical of issues and limitations of said studies. Hossain et al. (2019) does, however, mention that sample size of crashes is a key limitation of existing studies. In their meta-analysis of real-time crash prediction work, Roshandel et al. (2015) review 13 studies published between 2001 and 2012. Like the other review articles, they found a majority of the studies (11 out of 13) to have applied loop detector data, and noted that the studies they consider that did not apply loop data used vehicle trajectory data obtained from video. They also reported that the majority of the studies they considered applied a case-control study design (12 out of 13). In terms of limitations of the existing studies, Roshandel et al. (2015) highlight a number of issues that deserve attention in future work. For one they bring attention the fact that existing studies almost universally ignore human-factors related variables in the models, hence leading to omitted variable bias. Additionally, the real-time crash prediction models typically only consider first-order terms and fail to consider interactions. Another issue they point out is how, at the time of their review, many papers were developed from a single dataset, leading to further bias issues. As a final limitation of existing work, Roshandel et al. (2015) highlight issues associated with the use of case-control study designs, including the ratio of cases to controls itself. They discuss how the traffic crash prediction scenario is not completely analogous to the typical, medical usage of case-control studies where access to the full control set is usually impossible. They also call into question issues regarding the impact of spatial and temporal resolution of data and data aggregation, in terms of suppression of signal in the data. Further, the lack of principled manner in which to choose a proper data resolution is discussed (Roshandel et al. 2015).

Hourdos et al. (2006) used a logit model to estimate crash risk for a freeway segment with loop detectors, as well as video camera surveillance on I-94 in Minnesota. The segment was 450 feet long, and the upstream and downstream segments under camera surveillance were both 300 feet in length. Video surveillance made it possible to (1) extract vehicle trajectories from the study segment and (2) observe 110 crashes. The logit model considered variables such as average speed, coefficient of variation in speed, kinetic energy (density multiplied by mean speed), mean velocity gradient (ratio of acceleration noise to mean velocity) which was developed based upon trajectory data, traffic pressure (variance in speed multiplied by density), an indicator for wet/dry pavement, a categorical variable describing position of the sun, an indicator for reduced/clear visibility. Upon testing their model, they determined it correctly identified 58 percent of crashes and had a false alarm rate of 6.8 percent.

Hossain and Muromachi (2012) developed a Bayesian belief network (BBN) to predict real-time crash risk on expressways in Japan; the two segments were 11.9 km and 13.5 km in length, and had a total of 250 loop detectors between the two locations (leading to a detector spacing of approximately 250 m). A total of 722 crashes and a corresponding control set of 26,899 periods with normal traffic were used in the model development, and it was determined that variables including downstream congestion index, difference in downstream and upstream occupancy, and difference in downstream and upstream speed had an impact on crash risk. The model ultimately correctly labeled two thirds of crashes in the testing phase with less than 20 percent false positives. Ahmed and Abdel-Aty (2012) used toll tag reader, also known as AVI, data in real-time crash prediction modeling efforts over 78 miles of expressway in Orlando. They built the model on 670 crashes matched to 2,680 controls (i.e., normal traffic periods). They showed that the coefficient of variation in speed for the segment in which the crash occurred on was positively correlated with crash risk and that their models could correctly classify between approximately 68-69 percent of crashes correctly.

Yu and Abdel-Aty (2013a) used Bayesian logistic regression (with considerations for addressing unobserved heterogeneity, namely random parameters and random effects) and SVMs to predict real-time crash risk for a 15-mile segment of I-770 in Colorado. Variables considered in the logistic regression models include average speed at the nearest downstream detector to the crash, as well the standard deviation of occupancy and the standard deviation of volume also both calculated based on the nearest downstream detectors to the crash. Variables considered in the SVM model include average speed at the nearest downstream detector to the crash, average speed at a downstream RTMS station, as well as the average values of occupancy and volume at the nearest downstream RTMS to the crash. Ultimately, they concluded that the SVM with RBF kernel had the best goodness-of-fit amongst all models and that inclusion of random parameters or random effects within the logistic regression models had little impact on results. Xu et al. (2015) used a sequential logit model to predict crash risk for severity levels including: fatal/incapacitating injury, non-incapacitating/possible injury, and property damage only. They studied a 29-mile segment of I-880 in San Francisco with 119 loop detectors total over both directions of travel, 5 weather stations, and 794 crashes occurring during 2008. They used data from loop detectors (both upstream and downstream of the crash) including upstream occupancy, standard deviation in speed both upstream and downstream, downstream average absolute difference in adjacent lane occupancy, average absolute difference in vehicle volumes between upstream and downstream locations, average absolute difference in vehicle volumes between upstream and downstream locations, a peak hour indicator, and a rain/fog indicator, among other variables in their model. Their main conclusions were that non-injury crashes had a higher probability of occurring in congested periods with high standard deviation in speed and there are many lane-change maneuvers; severe injury crashes were most probable in uncongested time periods with high speeds and speed variations among lanes.

Sun and Sun (2015) applied a dynamic Bayesian network (DBN) to six freeway segments in Shanghai, China ranging in length from 0.8-2.2 kilometers in length, over which 551 crashes occurred (April-December 2010); the crashes were matched to 2,755 control time periods. They used 5-minute loop data and considered traffic states defined based on speeds upstream and downstream of the crash site. They demonstrated that their DBN model with 9 traffic state combinations based on speed had the best performance compared to a static Bayesian network, a DBN with 4 states, and a DBN considering volume, speed, and occupancy variables, with an overall accuracy of 0.763 (i.e., proportion of correct predictions of crashes and non-crashes). Shi and Abdel-Aty (2015) developed real-time crash prediction models based on data from 275 MVDS detectors over 75-miles of freeway with segments along State Route (SR) 408, SR 417, and SR 528 in Florida. Their study considered 243 rear-end crashes matched to 962 non-crash cases. They developed three types of Bayesian logit models: random effects, fixed effects, and random parameters, and each model considered the following variables: logarithm of volume at second nearest upstream detector station to crash, average speed at second nearest upstream detector station, and the congestion index (defined in Equation 2-1) measured at the nearest downstream detector station to the crash. Goodness-of-fit was similar between all models and key conclusions were that increases in upstream volume and downstream CI, as well decreases in downstream speed were correlated with increased probability of rear-end crash occurrence.

$$Congestion\ Index\ (CI) = Max\left(\frac{free\ flow\ speed - actual\ speed}{free\ flow\ speed}, 0\right) \tag{2-1}$$

2.2 Exposure and Consideration of Traffic Operations in Crash Frequency Modeling

If one considers real-time crash prediction as micro-scale crash modeling, then, at the other end of the spectrum (i.e., macro-scale modeling) would be cash frequency modeling. Crash frequency models seek to estimate the number of crashes that will occur on a given road segment or at an intersection based on a variety of factors. Over the past few decades much of the focus on crash frequency modeling has involved applying different model formulations to enable better fit to the data, and account for issues such as over-dispersion, endogeneity, and unobserved heterogeneity among many others (Lord and Mannering 2010; Mannering and Bhat 2014). Models applied have evolved tremendously from the initial usage of the Poisson model to the negative binomial model to advanced formulations such as finite mixture models, Markov switching models, and machine-learning-/AI-based approaches such as neural networks (Lord and Mannering 2010). Despite the continual advances, there is much work to be done addressing a series of core issues including further consideration of the unobserved heterogeneity issue, as well as how to best handle and account for variation of variables over time periods used in crash modeling.

Mannering and Bhat (2014) describe the unobserved heterogeneity issue in the context of crash frequency modeling as follows: "[when] unobserved factors … are correlated with observed factors, biased parameters will be estimated and incorrect inferences could be drawn." The unobserved variables generally refer to variables whose values are not or cannot be measured for any variety of reasons. As an example, consider a model for crash frequency whose only variables is AADT. For such a model, AADT is likely correlated with many other variables such as functional classification of the roadway, number of lanes, urban/rural location, etc. that may impact crash frequency as well. As such, the resultant model will (1) posit that the impact of AADT on crash frequency is constant for the entire population and (2) force AADT to be a proxy for the omitted variables, who will also almost certainly exhibit variation from the population perspective (Mannering and Bhat 2014).

In regards to the issue of variation in variables within time periods, Stylianou et al. (2019) note that only considering the mean value of variables over some time period fails to address the importance of considering how variation of said variable in some time period may impact crash frequency. Lord and Mannering (2010) echo this sentiment and note that data availability may be partially to blame for this issue. They further describe how use of aggregated data (depending on the level, of course) can lead to unobserved heterogeneity.

Crash frequency models typically apply a series of static, or at least, relatively static predictor variables such as AADT, number of lanes, lane width, shoulder width, etc. Often missing from these equations is variables describing traffic operations, which at least based on the aforementioned review of real-time crash prediction, appear to have some correlation with crash occurrence and in turn crash frequency. By omitting such variables from crash frequency models, the issue of unobserved heterogeneity arises. That said, how to best account for time-variation of these variables when considering modeling frameworks with “macro” time scales, typically on the order of a year or greater can be challenging. The following provides an overview of applications including traffic-flow/traffic-operations-related variables in frequency models.

While detailed traffic flow information is often not included in the predictor set for crash frequency models, one typical operational variable included is AADT, often considered to provide a measure of exposure. Exposure measures are included in most if not all crash frequency models; however, the specific metric used varies between studies. To further complicate things, the definition of exposure is not well agreed upon in the literature. Hauer (1982) states that "exposure is used to estimate risk" and that "a unit of exposure corresponds to a trial. The result of which is the occurrence or non-occurrence of [some type of] an accident." Based on his definition, AADT would not qualify as a measure of exposure. Elvik et al. (2009), however, note that AADT is one of several so-called "summary measures of exposure" as it does not give a direct indication of trials in which a crash could occur. They further define elementary exposure measures as one of the following four possibilities: encounters (where vehicles traveling in different directions pass), arrivals at conflict points for conflicting movements, lane changes, and braking/stopping.

In their spatial analysis of urban crashes Bao et al. (2017) considered exposure metrics including AADT, truck AADT, trip productions by traffic analysis zone (TAZ), and trip attractions by TAZ for their geographically weighted regression models. Stipancic et al. (2018b) used the number of GPS trips as an exposure metric in their study of crash frequency applying GPS data. Saunier and Sayed (2008) mentioned that typical exposure metrics include or are based upon numbers of peoples as well as quantities describing travel (i.e., in time or distance, such as road user-hours or road user-kilometers). They also discuss how the idea of exposure was raised a means to make comparing conditions between locations fair, and such comparisons were often based on a collision rate defined as the ratio of crashes to exposure. It is important to note however, that exposure and crashes may not actually be linearly related (Hauer 1995). Indeed, Qin et al. (2004) showed a non-linear relationship existed between crash count and AADT for two-lane highways, when modeling crashes by the following crash types: single-vehicle, multi-vehicle same direction of travel, multi-vehicle opposite direction of travel, and multi-vehicle intersecting paths.

Besides consideration of exposure metrics being derived from traffic flow/operational data, other studies have sought to include predictors derived from traffic flow data in their crash frequency modeling efforts. While these studies are limited in number, perhaps due to consideration of the within period variation issue aforementioned, there are still several of note. Recently, a series of studies by Stipancic et al. applied trajectory data derived from mobile phone GPS traces in order to examine relations between traffic operations and crashes (Stipancic et al. 2017, 2018a; b). Stipancic et al. (2017) investigated GPS data collected over 21 days between April and May 2014 in Quebec City, Quebec. Their dataset had 21,939 trips completed by over 4,000 total drivers, and data points in each trip were recorded with an average frequency of one to two seconds. Data were recorded for both road segments and intersections, and for each unit, functional classifications included were as follows: motorway, primary, secondary, tertiary, and residential. It is further important to note that data were collected through a mobile phone app which drivers had to choose to install. The crash data used in their study was collected between 2000 and 2010, and it contained 9,248 crashes. For their initial study, they investigated correlation between crash frequency and the following surrogate safety measures (SSMs) defined based on the trajectory data for links and intersections: congestion index, average speed, and coefficient of variation of speed. Correlations found were not particularly strong, but it was observed that CI and crash frequency had positive correlation (ranging from 0.02 to 0.21, depending on functional classification of the analysis road unit). Additionally, CVS was found to have positive correlation with crash frequency (ranging from 0.10 to 0.38, depending on functional classification of the analysis road unit) (Stipancic et al. 2017).

Stipancic et al. (2018a) examined the correlation between hard breaking events (HBEs) and hard acceleration events (HAEs), two types of SSMs, derived from GPS data collected from more than 4,000 drivers and 21,939 trips taking place in Quebec City for 21 day period in 2014. Collision data was gathered between 2006-2010 and contained 9,238 total crashes. They observed that both HBEs and HAEs had positive correlation with crash frequency for links and intersections, and between the two types of analysis units, correlations were stronger at intersections. In the preceding two articles, univariate correlations were investigated between crash frequency and different SSMs. In Stipancic et al. (2018b), multivariate regression models were developed to examine the impact of SSMs on crash frequency. The GPS data used was collected over less than one month in 2014 in Quebec City and contained more than 4,000 drivers and close to 22,000 trips. Crash data were extracted between 2000-2010, a period in which 14,278 crashes occurred on the network. In their study, they developed latent Gaussian spatial negative binomial models for links and intersections based on features derived from the GPS trajectory data; such models were used as they can account for spatial correlations. Both models included the following predictors: an intercept, logarithm of number of trips on each spatial unit (taken as an exposure measure), number of HBEs per trip, CI, coefficient of variation of speed, average speed, logarithm of length (link-based model only), and a categorical variable describing functional classification. For the link model, considering the variables derived from the trajectories only, coefficient of variation of speed was found to have positive correlation with crash frequency (and the variable was significant at the 95% confidence level). For the intersection-based model, logarithm of trips, HBEs per trip, and CI were positively correlated with crash frequency, while average speed showed negative correlation (all variables mentioned here were significant at the 95% confidence level). A crucial issue with the three preceding studies is that the time periods over which the crash data and trajectory data were collected were not the same, nor even overlapping at all.

2.3 Crash Frequency Modeling

As aforementioned, crash frequency models are used by analysts to predict crash counts over time based on factors such as roadway geometry (e.g., lane width, shoulder width, etc.), traffic volumes, etc. (Mannering and Bhat 2014). As both statistical methodologies and computing power have advanced over time, many different types of models can be considered candidates for crash frequency analysis (Lord and Mannering 2010). For early crash frequency modeling applications, simple count data models such as Poisson regression were commonly applied (Gustavsson 1969; Gustavsson and Svensson 1976; Jovanis and Chang 1986). While such models are simple, intuitive, and relatively easy to apply/estimate, one of their biggest drawbacks is the inability to handle overdispersion in crash data, a phenomenon which is relatively common (Lord and Mannering 2010). Overdispersion is defined to occur when the variance of the crash counts is observed to be greater than the mean (Lord and Mannering 2010). When one assumes that crash data result from Poisson trials (i.e., Bernoulli trials with a non-constant crash probability for each trial), as outlined by Lord et al. (2005), overdispersion can indeed occur. Overdispersed crash datasets often have numerous datapoints (i.e., roadway entities such as segments) with zero observed crashes and/or a high number of crashes and thus are often not able to be accommodated by assuming a simple underlying Poisson distribution (Lord et al. 2005). Just as overdispersion is possible, underdispersion exists in crash datasets and occurs when the variance is less than the mean in observed crash counts. Study of underdispersed crash data is not a focus of this dissertation, however, it has been investigated in other studies. For a sample of such studies, interested readers are directed to Lord et al. (2010a) and Giuffrè et al. (2011) who applied the Conway-Maxwell Poisson model to study underdispersed crash data.

As overdispersed crash data is quite common, many models/methodologies have been developed to properly account for said phenomenon. To date, it seems the most popular/most commonly applied is the negative binomial (NB) model, a model that many researchers and analysts have used in applications with overdispersed crash data (Connors et al. 2013; El-Basyouny and Sayed 2006; Hauer et al. 1988; Lord and Mannering 2010; Maycock and Hall 1984; Park et al. 2012; Srinivasan et al. 2010; Ye et al. 2013). In the NB model, a crucial underlying assumption requires that the mean crash frequency (i.e., the Poisson parameter) for any roadway entity (i.e., site) i, λ_i, follows a gamma distribution (Hauer 1992). As such, the marginal mean and variance for a given crash count, y_i, can be formulated in a manner in where the variance can be greater than the mean (Hauer 1992; Lawless 1987; Lord and Mannering 2010). Another name for the NB model is the Poisson-Gamma model due to the fact that the crash count for a given site i, y_i, when conditioned on the Poisson parameter λ_i (again, following the gamma distribution), itself follows a Poisson distribution. With this in mind, it is important to note that one does not have to assume that λ_i must always be gamma distributed (Hauer 1997; Lord et al. 2005). Indeed, many studies have investigated the use of several other distributions for the λ parameter, and these in turn lead to different types of mixed-Poisson regression models (i.e., any models where the crash count when conditioned on the Poisson parameter, whose distribution is referred to as the mixing or mixture distribution, follows a Poisson distribution) (Cameron and Trivedi 2013; Lawless 1987). One example of an additional choice for the mixture distribution is the generalized inverse Gaussian (GIG) distribution, and use of this mixture distribution leads to the Sichel (SI) model which can be used to model overdispersed count data (Rigby et al. 2008). Zou et al. (2015)

demonstrated an application of a Sichel model in the analysis of a crash dataset from Texas with a high degree of overdispersion and then compared those results to the results from a traditional NB model estimated on the same dataset. They observed that the SI model led to lower values for both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (i.e., better statistical goodness-of-fit) than those for the NB model. An additional candidate mixture distribution is the inverse Gaussian (IG) distribution, which when used in a mixed-Poisson model leads to the Poisson-Inverse-Gaussian (PIG) model (Dean et al. 1989). Zha et al. (2016) observed that when applied to the Texas crash dataset mentioned previously, as well as a crash dataset collected in Washington State, the PIG regression models yielded a better fit (as measured by AIC and BIC) than the standard NB models.

Still, there are additional candidate mixture distributions that include, but are not limited to, the Weibull and lognormal distributions, which yield the Poisson-Weibull (PW) and Poisson-Lognormal (PLN) models, when applied respectively. Both the PW and PLN models, like their previously-mentioned counterparts, are able to be used in modeling applications with overdispersed data. Cheng et al. (2013) investigated an application of the PW model, while many others have investigated applications of the PLN model, including Lord and Miranda-Moreno (2008), Aguero-Valverde and Jovanis (2008), Lan and Persaud (2012), and Zhao et al. (2018), among others.

In additional to being able to accommodate overdispersion, another shared characteristic between each of the aforementioned mixed-Poisson regression models is that by default, they yield only a sole point estimate for expected crash frequency on a given roadway entity (i.e., segment, intersection, etc.). While these estimates still have their uses in different tasks such as prediction and before/after studies, it is not difficult to imagine cases where having a confidence interval for a particular crash estimate is better (Casella and Berger 2001). Previous studies have noted how confidence intervals have important applications in safety decision-making since they can as they can help convey the uncertainty associated with a particular point estimate (Lord 2008; Lord et al. 2010b). In the traffic safety domain, Wood (2005) was among the first to present formulae for prediction intervals (PIs) for the predicted response (i.e., crash frequency for a new site, y_i) and the gamma mean (m_i), in addition to confidence intervals (CIs) for the true mean crash frequency (additionally named the mean response or Poisson mean, μ_i), for the commonly-used NB (Poisson-gamma) regression model. Here, one must be clear on the difference between the Poisson parameter and the Poisson mean. When estimating a typical Poisson regression, these two values are indeed the same. That said, in the context of a mixed-Poisson model, the error term in the formulation of the Poisson parameter negates the former equality of these two terms. This issue is described further in this dissertation as well as in Ash et al. (2019). Wood (2005) concluded that the use of PIs and CIs may be quite useful for crash prediction applications at different sites with similar features to those found at the sites considered in the initial model development. Lord (2008) presented a method to estimate the predicted confidence intervals for NB regression models when multiplied by crash modification factors (CMFs). Geedipally and Lord (2008) examined PIs for the predicted response (y) and the gamma mean (m), as well as CIs for the mean response/Poisson mean (μ), that were developed from NB models with dispersion parameters that were both fixed and allowed to vary, in addition to those for univariate and bivariate NB models (Geedipally and Lord 2010). Lord et al. (2010b) calculated PIs for crash counts (y) estimated from a multivariate NB model in comparison to those estimated based upon a "baseline model" (i.e., flow-only crash prediction model) with crash-modification factors (CMFs) applied to it. Connors et al. (2013) estimated and plotted the CIs and PIs for predicted values of μ_i and m for a range of flow and segment length values in NB and PLN model applications. In their study, however, it did not appear that any explicit formulae for the CIs and PIs for the PLN model were provided.

Chapter 3. DATA RESOURCES

The main data components for this study can be classified into one of the following three categories: (1) operational data, safety/crash data, and (3) roadway inventory and descriptive data. The operational data primarily consists of vehicle trajectory data collected from probe vehicles. The safety data consists of data obtained from crash reports about the occurrence of specific crashes, including such features as time of type of crash, crash severity, time of crash occurrence, etc. Both disaggregate and aggregate forms of this data are used in different applications which will be detailed later. The final category of data used in this study consists of roadway inventory data (e.g., roadway geometrics) as well as other descriptive data, such as information on the weather. The following sections provide a detailed overview of the data used in this study.

3.1 OPERATIONAL DATA

3.1.1 *Overview of Vehicle Trajectory Data*

In the simplest sense, a vehicle trajectory is defined as data describing the motion of an object in a time-space surface (Daganzo 1997). In transportation applications, vehicle trajectories are often mapped to a two-dimensional plane, space versus time, in which case their location is given with respect to a linear reference. That said, motion of vehicles can be described in higher-dimensional terms as well; hence the use of a time-space surface. Often, vehicle locations are collected over time as pairs of longitude (abscissa) and latitude (ordinate) given in decimal degrees. Depending on the data resolution, trajectory data may be preferrable to data collected from loop detectors as it describes the movement of individual vehicles and can be aggregated to provide summary statistics over time, which is the primary form in which loop data is provided (i.e., loop data is most often aggregated over some time window).

Vehicle trajectory data provided for this project were supplied by INRIX, Inc., a traffic data company operating out of Kirkland, WA. At its core, the data are a series of GPS points in the form of (longitude, latitude), with supporting attributes, provided within a bounding box requested by the user. Sources of the data include cellphone providers, fleet vehicles, dedicated in-vehicle GPS navigation systems, etc. For this project, data were requested for the following geographical bounding box in the form of (xMin, yMin, xMax, yMax): (-122.34116, 47.44634, -121.970028, 47.840953), as shown in Figure 3-1. For this project's data request, INRIX provided data within these geographic bounds for the entire month of May 2017. The data provided come in two separate types of csv files, the "trips" files and the "waypoints" files; here each file represented one week (seven days) of data. A trips file contains summary information on any trip that traveled within (i.e., had at least one GPS point recorded in) the bounding box during the specified time period. Each row in the file represents one such trip and has attributes including, but not limited to the following, as shown in Table 3-1 Data Elements in Trips File. The waypoints file contains the individual GPS points that make up a given trip; the attributes of the waypoints file are show in Table 3-2. Both Trip ID and Device ID can be linked between the two files. In summary, the trips file provides broad overview/aggregate data on a given trip traveling into/through the bounding box, while the waypoints file describes the individual GPS trace points that make up a trip.

A final point of note about the waypoints data used in this study is that it is a probe vehicle dataset representing just a sample of the broader traffic population. That is to say, unlike loop detectors, all vehicles in the traffic stream at a given time-space location are not accounted for, rather only a sample of them report their location data over time via on-board GPS devices. The penetration rate (i.e., sample percentage) for probe data is often in the range of 1.5 to 5.5 percent of the total traffic population. Of course this factor brings with it its own set of challenges and considerations, and these will be addressed later.

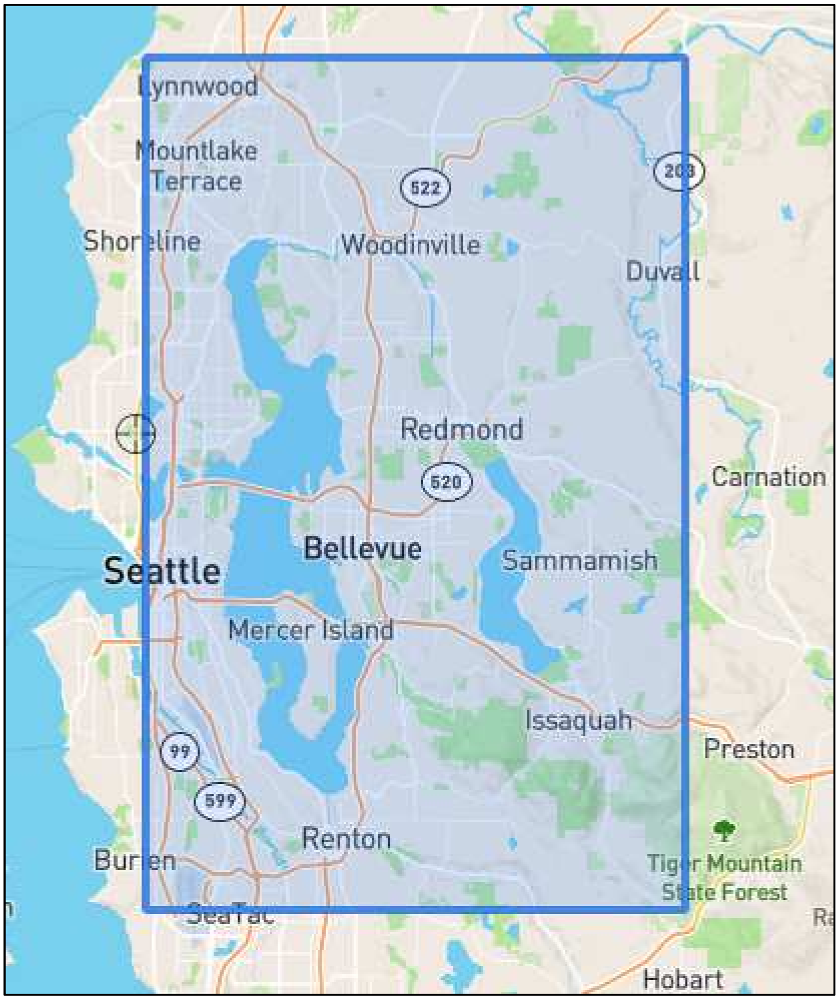


Figure 3-1 Bounding Box for INRIX Data Request (Seattle-Area Freeways), Image via Mapbox/© OpenStreetMap contributors

Table 3-1 Data Elements in Trips Files

Variable Name	Description [units]
Trip ID	Unique identifier for the given trip
Device ID	Unique ID for the specific device recording the GPS data
Provider ID	Unique ID for the provider of the GPS data for the given trip
Mode	Mode of travel
Start date	Start date and time of trip [down to 1/100 sec]
End date	End date time of trip [down to 1/100 sec]
Start location	Location of trip start (lat, lon)
End location	Location of trip end (lat, lon)
Provider type	GPS provider type
Vehicle weight class	Weight group classification
Trip mean speed	Mean speed [kph]
Trip max speed	Max speed [kph]
Trip distance	Distance traveled [m]

Table 3-2 Data Elements in Waypoints Files

Variable Name	Description [units]
Trip ID	Unique identifier for the given trip
Waypoint sequence	Integer describing temporal order of GPS points
Capture date	Date and time
Location	Location of GPS point (lat, lon)
Device ID	Unique ID for the specific device recording the GPS data
Raw speed	Speed at moment of data capture [kph]

An initial examination of the INRIX data was conducted to summarize some key aspects of the data (note the following summary measures were computed before map matching and filtering only to the Seattle area freeway network. i.e., metrics shown are for all trips occurring in May 2017 that travel into, through, and/or out of the bounding box shown in Figure 3-1). The freeway network of interest in this area includes I-5 and I-405 North and South running between Lynwood and Tukwila, I-90 East and West from its western terminus to an area near Issaquah, and SR-520 East and West from its western terminus to its eastern terminus. In total, the dataset for May 2017 had 1,902,603 individual trips, comprised of 92,545,591 waypoints (i.e., individual GPS points); see Table 3-3. The majority of the data came from consumer vehicles, and the distribution of trips by provider type is shown in Figure 3-2. In terms of vehicle type, most of the trips were from light duty vehicles (weight less than 14,000 pounds); the distribution of trips by vehicle weight class is shown in Figure 3-3. While the distribution of waypoint frequency (i.e., GPS point sampling rate) varies tremendously, this section begins by examining a summary of trips. In total, there are 451,000 trips (each composed of numerous GPS waypoints) with waypoint sampling rates less than or equal to 30 seconds. A bar chart showing the distribution of trips by sampling rate is shown in Figure 3-4.

Table 3-3 Initial Data Counts for Data Request (May 2017, Seattle Area)

Field	Count
Number of Trips	1,902,603
Number of Waypoints (GPS points)	92,451,591

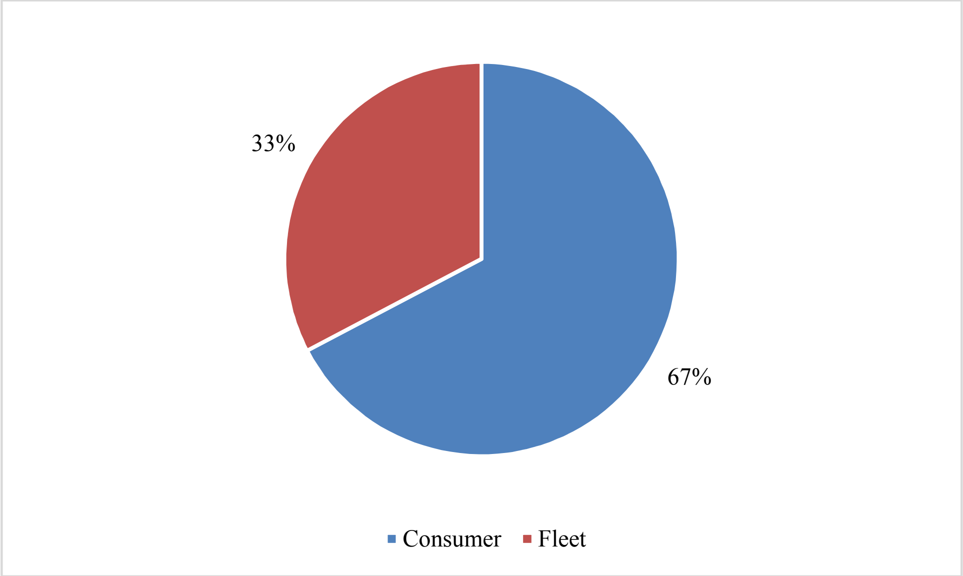


Figure 3-2 Distribution of Trips by Provider Type

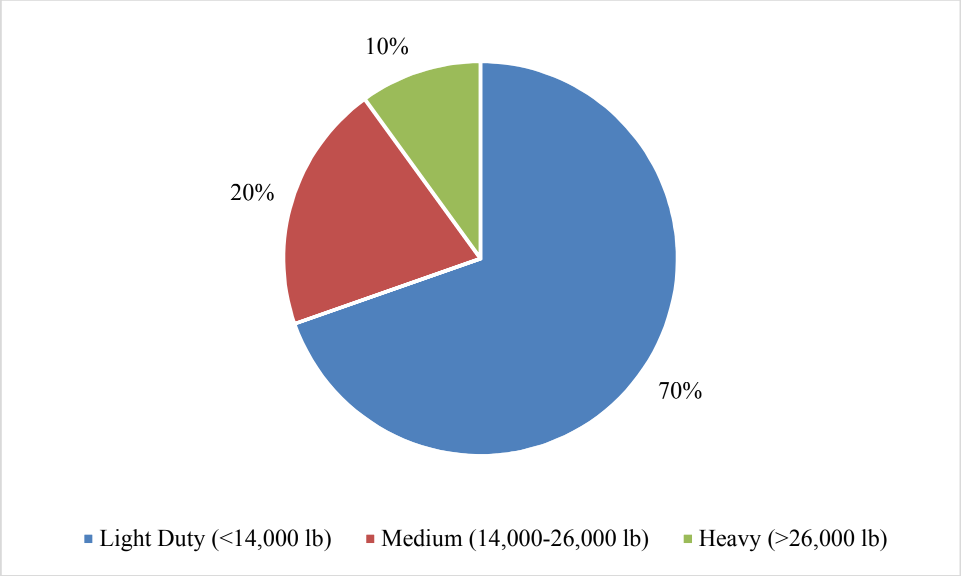


Figure 3-3 Distribution of Trips by Vehicle Weight Class

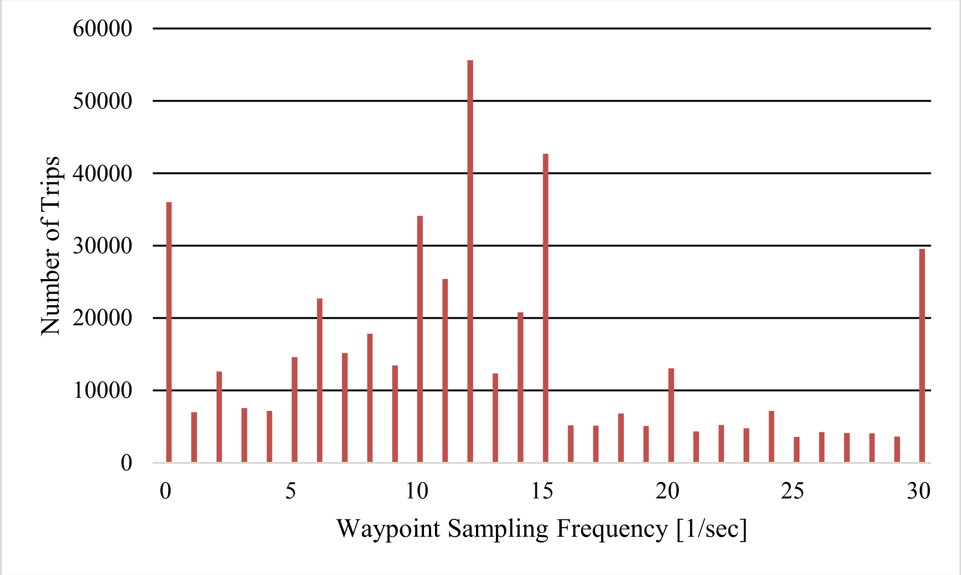


Figure 3-4 Distribution of Trip Waypoint Sampling Frequency

The data provided by INRIX solely consisted of raw GPS points; that is to say, none of the data were map-matched or linked to/associated with any given part of a roadway network. In order to analyze the data in the context of the Seattle freeway network, the data points had to be mapped to the roadway network, and later filtered down to only the subject roadways of this study. At a high level, this process was achieved as follows. First, all data from the trips and waypoints files were loaded into respective "trips" and "waypoints" tables in a PostgreSQL database with the

PostGIS extension to enable use of spatial data types and analyses. These databases were created solely for this project. Then, the GPS data from the waypoints file were map matched to the Washington roadway network using a base map from OpenStreetMap (OSM). The map matching was accomplished through use of the Open Street Routing Machine (OSRM), a data routing engine provided and developed by OpenStreetMap (OSM). OSRM applies a map-matching algorithm developed by Newson and Krum (2009) that uses a hidden Markov model (HMM) to link GPS points to the most probable roadway link on which they were recorded. For each measured GPS point with associated time stamp, the algorithm examines possible road segments in the OSM network the point may fall on and it also considers transitions to determine the most probable path. Finally, the GPS points were associated with a linear referencing system (i.e., one-dimensional space), so mileposts could be calculated. The aforementioned process is described in detail in the following.

3.1.2 *Trajectory Data Processing (Map Matching and Conflation)*

As previously described, the trajectory data processing process has three main steps described in the following.

3.1.2.1 Map Matching

The first step of the trajectory process is known as map matching, and here, it involves matching the GPS (lon, lat) points provided by INRIX to the OSM roadway network. First, the INRIX trip and waypoint data are loaded into relational database tables in the PostGRESQL database management system (referred to henceforth as PostGRES). These tables can then be linked via a foreign key defined on the Trip ID attribute and further queried to get things like summary statistics, as well as spatial statistics, the latter of which can be calculated in a more efficient way by establishing a spatial index on the longitude and latitude of the GPS data points. Once the data are loaded into their respective tables, the longitude and latitude data in the Waypoints table are converted from their decimal (i.e., type double) representation in two different columns to a PostGIS-specific data type, known as a Geometry. The data are then converted to a representation on the World Geodetic System 84 (WGS 84) datum surface (spatial reference identifier (SRID) 4326), allowing them to be easily worked with and visualized in OSM, which also uses the WGS 84 coordinate system/reference. As an example of a series of waypoints records (i.e., GPS points from probe vehicles) being visualized in OSM with respect to the WGS 84 datum, see Figure 3-5.

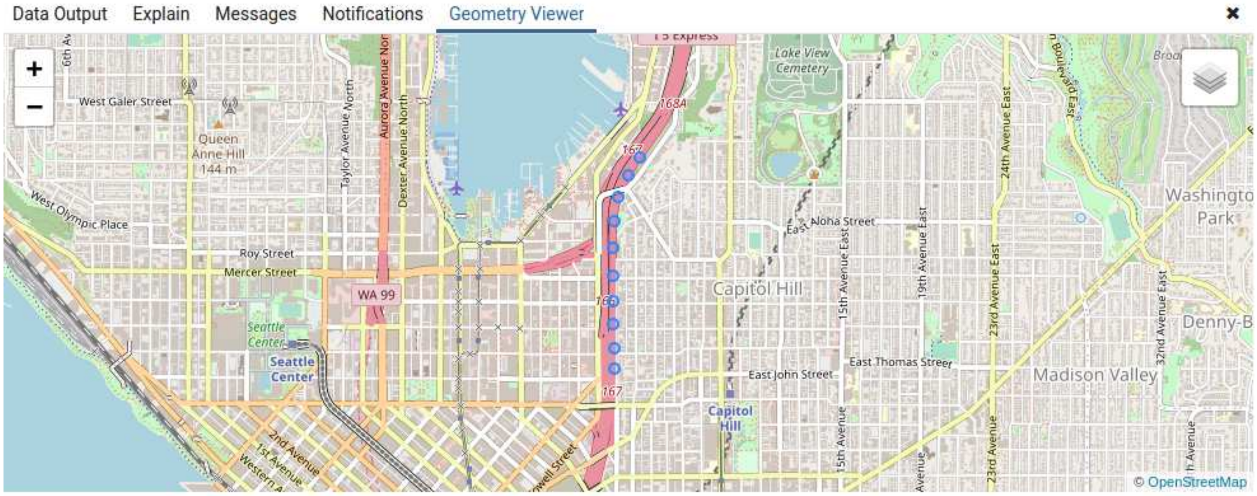


Figure 3-5 INRIX Waypoints Visualized in WGS 84 Coordinate System (credit: © OpenStreetMap contributors)

The next step of the map matching process involves using the Open Street Routing Machine to match the waypoints to the OSM roadway network. The OSM network is comprised of three primary elements including nodes, ways, and relations. These elements follow a hierarchy in that a series of nodes (i.e., individual GPS points) makes up a way, and a series of ways (i.e., lines) makes up a relation. OSRM works by ingesting a series of GPS coordinates and outputting a JSON file showing their relation to the OSM roadway network. This file shows elements including the degree of confidence of the match a list of one-to-one matches of the input GPS points to points along the OSM network (assuming a match is possible; call these match points), the nodes in the OSM network that bound each match point, and a representation of the trip (i.e., list of input points) in terms of legs defined by the OSM nodes that bound a match between two input GPS points. When every input point is matched to the OSM network, the results will show n-1 legs for n input points, and each leg will be defined by at least two OSM nodes that bound the path between the two matched input points. To clarify this step, raw GPS points (i.e., the INRIX waypoints) are input into OSRM which matches them (as possible) to the OSM network and provides output associating the input points with a nodal representation along the OSM network. A Python script was written to automate this process and run it in parallel by pulling data from the Waypoints table in PostGRES for each Trip ID, passing the waypoints and corresponding timestamps to OSRM, and storing the output, namely the coordinates of the match point for every input point, as well as the OSM nodes that bound each match point. These data were stored in a separate table in PostGRES. At his point, each input waypoint is now matched to a GPS (lon, lat) pair corresponding to the OSM network and the IDs of the OSM nodes that bound it are also associated with each point.

3.1.2.2 Map Conflation

The second key step of the trajectory data processing is map conflation, which means finding a means to link the WSDOT roadway network with the OSM network so that one can determine the names of the roadways (and later the mileposts) the points fall on. Consider the following two map representations, each of which can be represented in a PostGIS table or series of tables. In the previous step, the representation of the OSM map in terms of nodes (points), ways (sets of points, i.e., links), and relations (sets of the aforementioned) was described. For the state of Washington, one can generate a tabular representation of the OSM nodes, ways, and relations that comprise the state's roadway network. The nodes table is comprised of individual points (nodes) designated by a unique ID as well as their longitude and latitude in decimal degrees. Similarly, the ways table is comprised of a set of ways (i.e., links made up of one or more nodes) and each entry gives a unique identifier for the way (Way ID) as well as an array of nodes that comprise the way. At the highest level of the hierarchy is the relations table. In this table, each relation (designated by a unique ID) is listed along with an array of ways that comprise it. As an example, I-5 North in Washington is designated as a relation in OSM and it is composed of several ways, each of which are composed of several nodes. In some cases, the relations are bidirectional, meaning there is not a separate relation for the increasing and decreasing milepost directions along the roadway. Such bidirectional relations are often used for shorter length roadways, such as I-405 and SR-520 for this project. Ultimately, however, even a relation is bidirectional, each direction of travel is comprised of several ways, so one is still easily able to distinguish travel direction. As an example of the hierarchy, consider the relation describing I-405, an example way that is a member of said relation, and one node that is a member of said way in Figure 3-6, Figure 3-7, and Figure 3-8, respectively. An extremely important point of note here is that each direction of travel for a given roadway is represented as one unit with no lateral (i.e., lane) separation. For example, I-405 has a four-lane cross section in several locations, but the OSM relation/way/nodal representation does not distinguish one lane of travel from another.

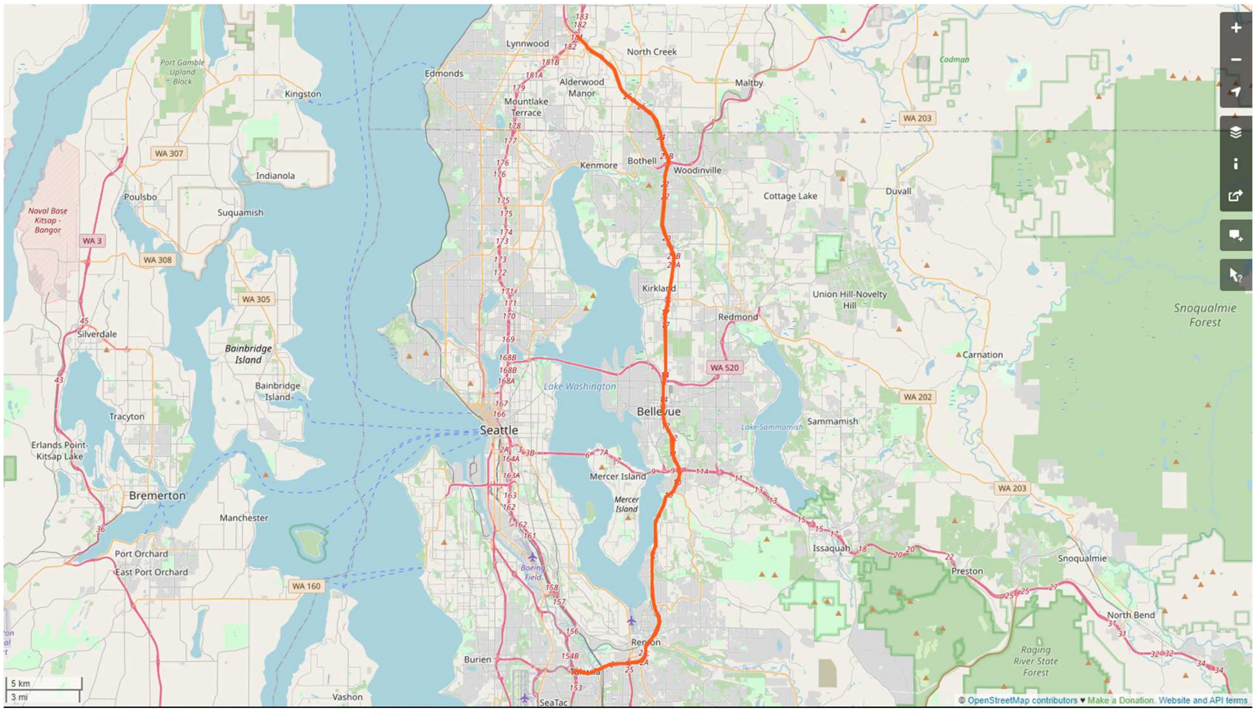


Figure 3-6 OSM Relation 1071195 (I-405, both directions) (credit: © OpenStreetMap contributors)

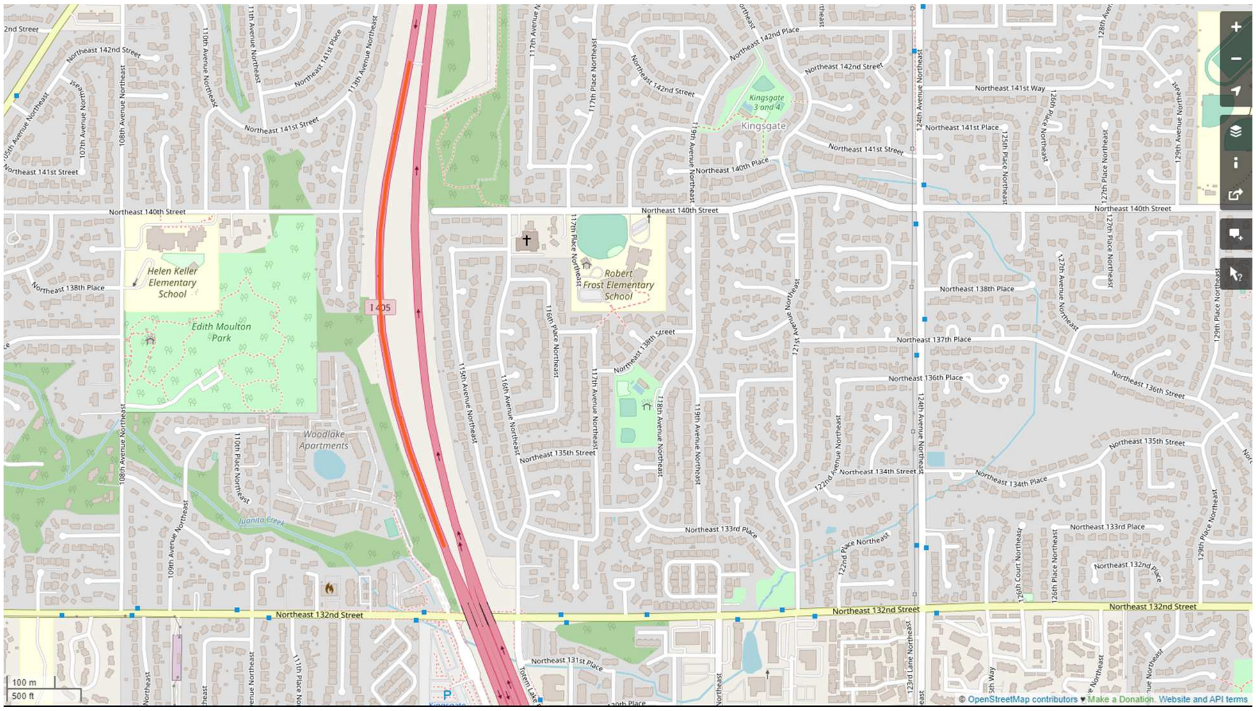


Figure 3-7 OSM Way 1071195 (along I-405 SB) (credit: © OpenStreetMap contributors)

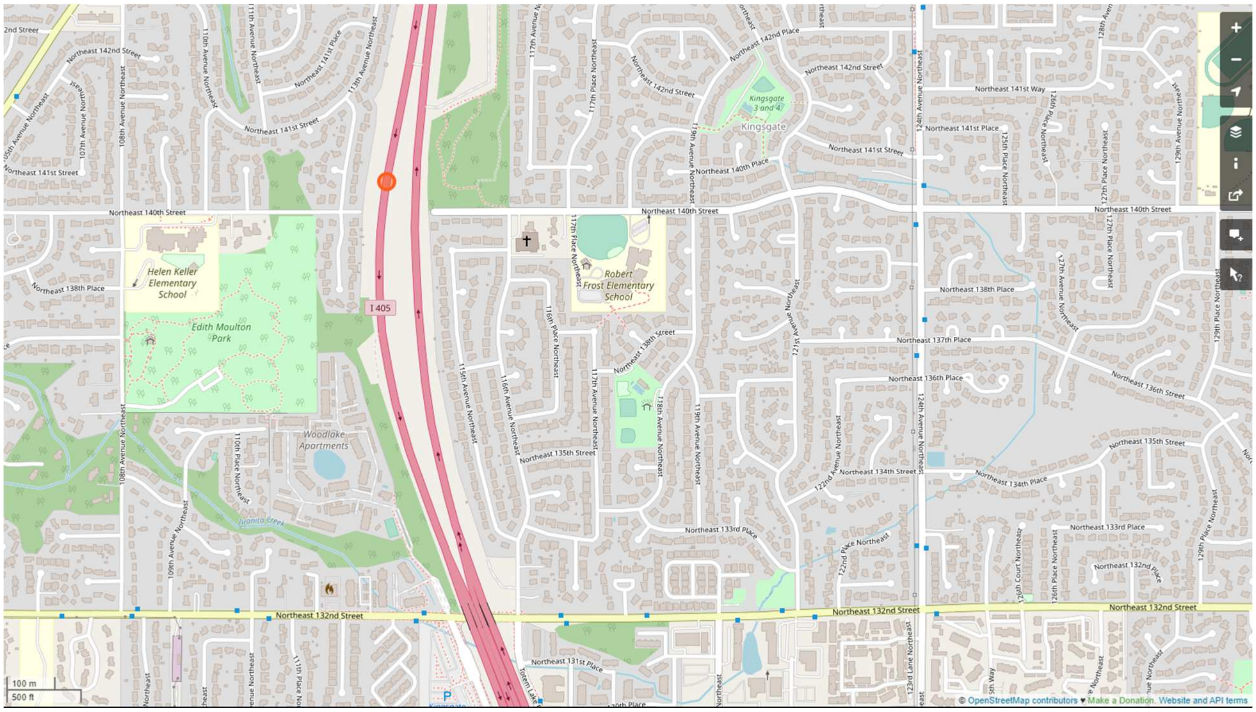


Figure 3-8 OSM Node 2477186164 (along Way 1071195, I-405 SB) (credit: © OpenStreetMap contributors)

The other part of the conflation task, from a cartographic perspective, is a representation of the WSDOT roadway network. Like most public transportation agencies, WSDOT maintains GIS files of the roadway network they, and other entities, manage. For this project, the 24k representation (i.e., a scale of 1:24,000) of the state roadway network is used; note that the final analysis is only concerned with roadways functionally classified as freeways in the Seattle area. A representation of the Washington public roadway network can be seen in Figure 3-9, note the area of the map most densely populated with roadways is not surprisingly Seattle.

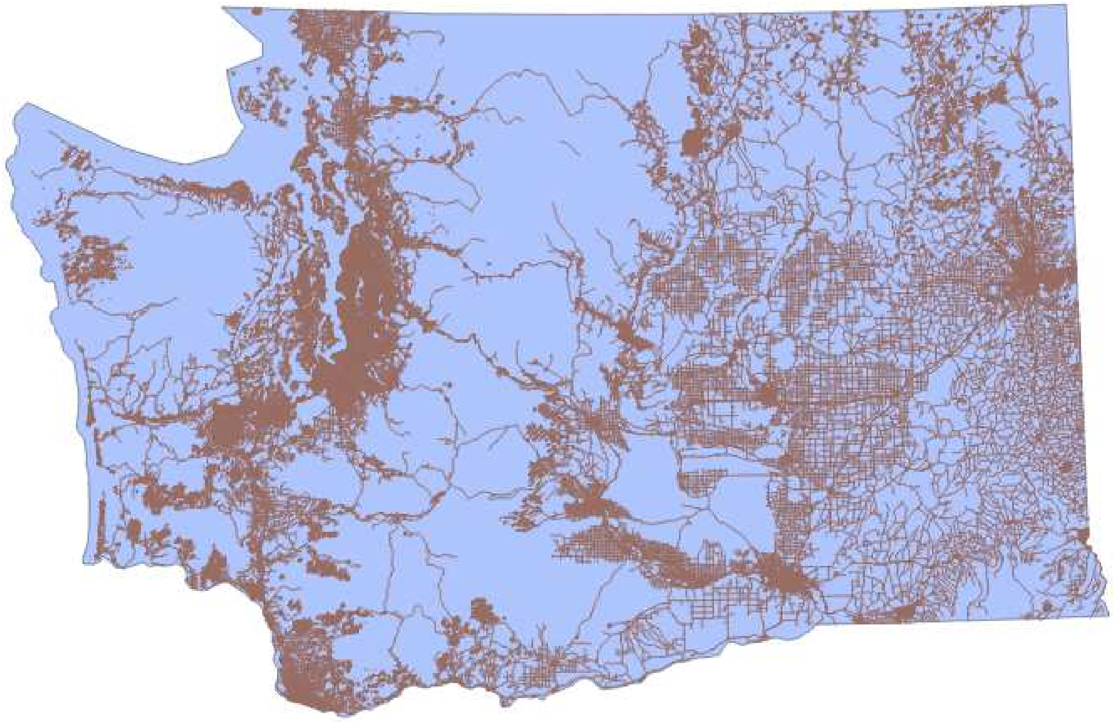


Figure 3-9 WSDOT Public Highway Network as Visualized in Q-GIS

While one may often think of geographic data like the aforementioned 24k map in terms of visual representation, i.e., as shapefiles, they can also be represented in a tabular form. Use of such representation was critical for this project, and a PostGIS table of the 24k map was developed. Each row (i.e., element) of the table represented a unidirectional link with the following attributes (as well as some additional attributes deemed unnecessary to show here) as shown in Table 3-4, and an image of a representative link as viewed in the OSM interface is shown in Figure 3-10 Link along I-405 SB as Viewed in OSM. Similar to the OSM representation of the roadway network, lanes are not distinguished for any link of the WSDOT network.

Table 3-4 Data Elements of PostGIS Tabular Representation of WSDOT 24k Road Map

Field Name	Data Type	Description
GID	Integer	Unique ID of link
Direction	Character	Direction of travel (inc or dec)
Barm	Begin Route Accumulated MP	Segment start MP
Earm	End Route Accumulated MP	Segment end MP
Region	Text	Region of state
Route ID	Integer	Route number
Route Type	Text	Type of roadway (e.g., I, SR, US, etc.)
Geom	Geometry	Array of GPS points comprising road segment (SRID=4326)

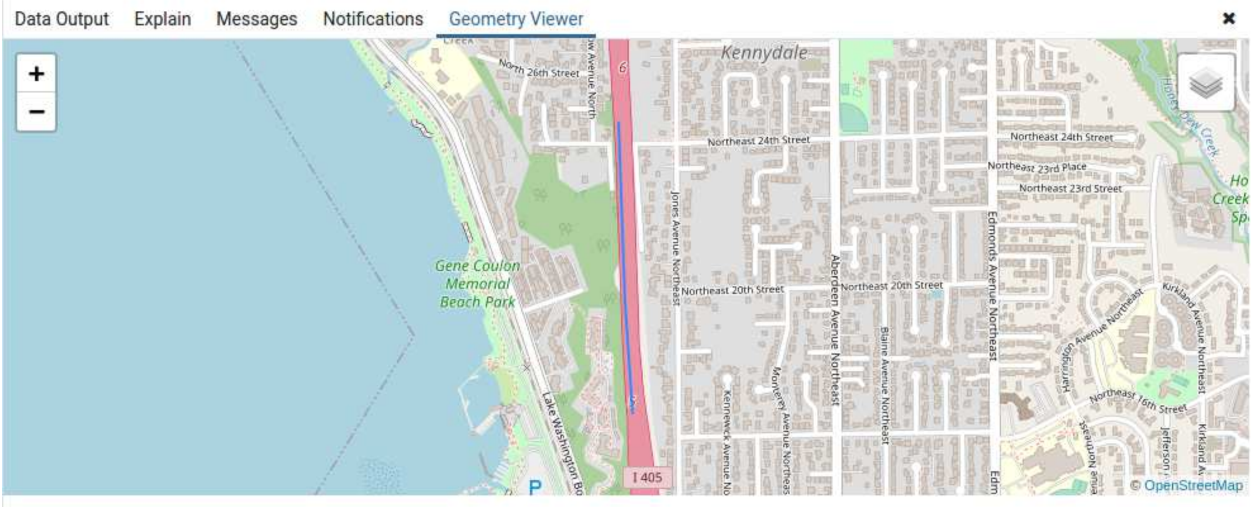


Figure 3-10 Link along I-405 SB as Viewed in OSM (credit: © OpenStreetMap contributors)

Once the map data for the OSM and WSDOT roadway networks are loaded into their respective PostGRES tables, the main task of conflation involves finding a means to link the two roadway networks. For example, one may need to determine how to match the segment from milepost 0.00-0.37 on I-405 SB to the corresponding relation/way/nodes in the OSM network. Similar to the previous task of map matching the vehicle trajectory GPS points, OSRM is also used as a key part of this task. For each of the following unidirectional roadways in the WSDOT map:

I-5 NB, I-5 SB, I-90 EB, I-90 WB, I-405 NB, I-405 SB, SR-520 EB, and SR-520 WB, the collection of segments comprising it (between a given milepost range defined in the next section) is fed into OSRM on a pointwise basis. That is to say, one can imagine the GPS points defining the WSDOT roadway links are nothing more than GPS points observed in a vehicle trajectory. Similar to the map matching process, for every unidirectional segment fed into OSRM on a pointwise basis, the output is a JSON file describing the including the confidence of the match results, a list of one-to-one matches of the input GPS points (here, defining unidirectional roadway links) to points along the OSM network (assuming a match is possible, again, call these match points), the nodes in the OSM network that bound each match point, and a representation of the roadway path (i.e., list of input points) in terms of legs defined by the OSM nodes that bound a match between two input GPS points. Again, when every input point is matched to the OSM network, the results will show n-1 legs for n input points, and each leg will be defined by at least two OSM nodes that bound the path between the two matched input points. Consider Figure 3-11 for a more detailed explanation of the process. In this figure, consider a segment of a WSDOT roadway from milepost 0.00 to 0.37, itself comprised of four GPS points. These points are matched to points in the OSM network, denoted by the green squares. These green points themselves are each bounded by two nodes of the OSM network denoted as gray or black circles. The gray circles represent intermediate points defining a way in OSM and the black points represent end points of ways; such black points are also referred to as junction nodes. The ultimate goal here is to determine what way in OSM a given point defining part of a freeway link's (with route number and direction) geometry corresponds to. Since there exists a tabular representation of the nodes and ways in OSM, one can determine what way the WSDOT points are on by (1) seeing what OSM nodes bound (gray or black points in the picture) them and (2) determining what OSM way these bounding nodes are on. At the end of this step, one is able to associate each point along a WSDOT freeway (itself identified according to a start and end milepost as well as route number and direction) with a way in OSM on which it lies. Once one knows what ways the points defining the roadway network lie on in OSM, he can calculate the mileposts of the junction nodes for each OSM way by interpolating based on the mileposts of the WSDOT points and the geometries of the respective segments (i.e., curvature and shape are defined via the geometry). For each route number and direction, one can begin at milepost 0.0 and calculate the mileposts of the junction nodes of each way in the OSM map based on the WSDOT points that fall along them (whose mileposts are known).

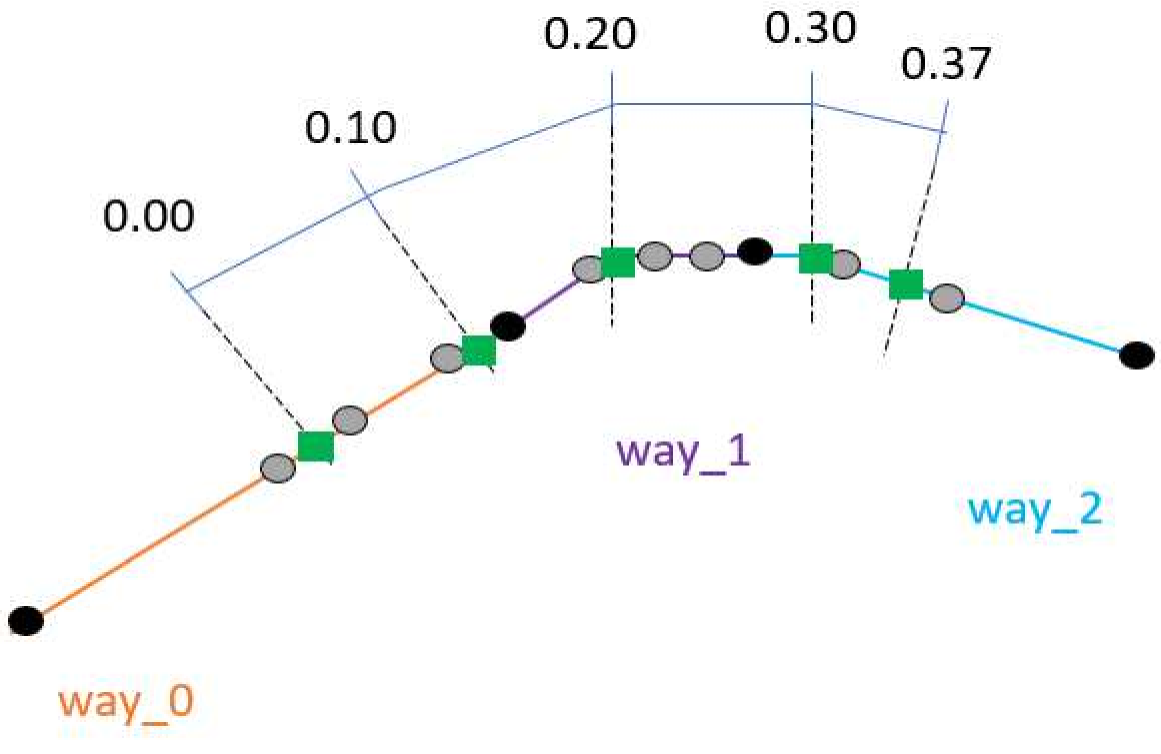


Figure 3-11 Schematic of Relation between WSDOT and OSM Roadway Networks

3.1.2.3 Milepost Calculation for Vehicle Trajectory Datapoints

The next step of the trajectory data processing involves determining the route number, direction of travel, and milepost corresponding to each input GPS point along a probe vehicle's trajectory. As a reminder, when a vehicle trajectory is provided from INRIX, it is solely a series of raw GPS points, none of which have any notion of being associated with a given roadway. For this step, an analog will be drawn to the output of the map matching step for the trajectories. Recall that for each matched trajectory, the output from OSRM provides (and have saved in a database table) the point in the OSM network that most closely matches the GPS input point, as well as the nodes that bound said point. As was the case for the map conflation step, one can use the bounding nodes to determine what way each matched point in the trajectory falls upon. In turn one is able to associate it with a route number direction of travel, and further calculate its milepost location based on the start and end milepost of the way and the geometry (i.e., shape/curvature of the way). Since the only roadways in the WSDOT network associated with the OSM network in the previous step were I-5, I-90, I-405, and SR-520, points matched to ways that do not comprise the relations defining these highways were filtered out at this step. Thus, the final output of this step and the map matching and conflation process was a PostGRES table of GPS points (for a given Trip ID) with timestamps, as well as their associated GPS point matches in OSM, and the route number, direction of travel, and milepost to which each was associated. This table only contained points along the aforementioned freeways, and the range of mileposts along which points can lie for each direction of travel is defined in Table 3-5. Further, it is important to note that in all cases, the increasing (I) direction of travel refers to NB or EB, while the decreasing (D) direction of travel refers to SB or WB.

Table 3-5 Milepost Range for Study by Route Number and Direction

Route Name/Number	Direction	Start MP	End MP
I-5	NB (I)	149.00	186.00
	SB (D)	186.00	149.00
I-90	EB (I)	0.00	30.00
	WB (D)	30.00	0.00
I-405	NB (I)	0.00	30.30
	SB (D)	30.30	0.00
SR-520	EB (I)	0.00	12.82
	WB (D)	12.82	0.00

To conclude the trajectory data processing section, a schematic summarizing the map matching, conflation, and MP calculation process is shown in Figure 3-12, and a summary of the number of trips and waypoints remaining following the data processing (i.e., those on the routes shown in Figure 3-12) is shown in Table 3-6.

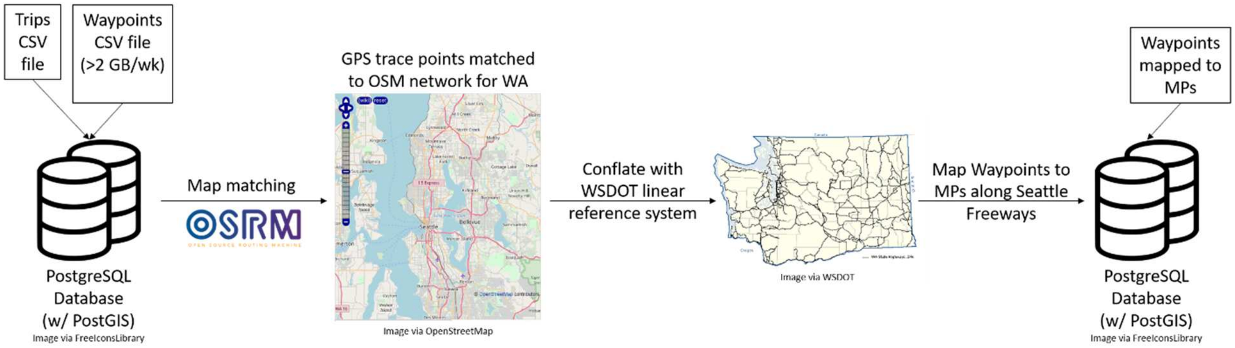

Trips CSV file
Waypoints CSV file (>2 GB/wk)
PostgreSQL Database (w/ PostGIS)
Image via FreeIconsLibrary
Map matching
OSRM
GPS trace points matched to OSM network for WA
Image via OpenStreetMap
Conflate with WSDOT linear reference system
Image via WSDOT
Map Waypoints to MPs along Seattle Freeways
Waypoints mapped to MPs
PostgreSQL Database (w/ PostGIS)
Image via FreeIconsLibrary

Figure 3-12 Schematic of Trajectory Data Preparation Process (Seattle map via © OpenStreetMap contributors)

Table 3-6 Final Counts of Data after Trajectory Processing

Field	Count
Number of Trips	730,909
Number of Waypoints (GPS points)	18,600,169

3.1.3 *Some Notes on Trajectory Data*

While working with trajectory data from probe vehicles has many benefits, there are also some limitations that must be discussed. First, it is important to talk a bit more about data resolution in both the spatial and temporal domains. In terms of spatial resolution, trajectories offer benefits over fixed, point-based detectors in allowing for data to be collected at any point in a road network, assuming a probe either (a) reported a location at said point or (b) passed through said point as determined by interpolation. That said, while loop detectors can provide data per lane, assuming detectors in each lane are wired to report data separately, the probe data used herein does not have the spatial resolution to determine what lane the vehicle is in. Further, the notion of lanes is lost in the map matching process when the GPS points are mapped to OSM ways, themselves which do not differentiate between ways. Additionally, as the location data is reported from GPS devices there is an inherent potential for errors in the location data due to issues such as multi-path and the urban canyon effect, as well as general limitations on GPS precision. With the exception of small portions of I-5 in downtown Seattle and portions of I-90 crossing Lake Washington, both of which pass through tunnels, the majority of the roadway miles in the study area are generally unobstructed from so-called urban canyons. With regard to inherent error in the GPS data, this issue is closely related with the map matching algorithm. In principle, one is trying to match a point (the GPS waypoint from a trajectory) to a line (an OSM way representing a roadway link), neither of which have any associated width. A perfect match is thus impossible due to floating point error, hence all GPS points matched to the OSM network are provided along with an error distance to the snapped, matched point on the way on which they would fall. The error distance is a Euclidean distance and can be further amplified by the loss of distinction of the travel lane the vehicle is in. As an example, consider a highway with a 5-lane cross section where each lane has a 12-foot width; it is easy to see how the true lateral location could affect the matching accuracy. While seemingly subtle, these errors can add up and impact latter calculations for measures such as speed. Finally, since the map-matching algorithm is itself prone to some errors. In cases where a match cannot be found, such a fact will be reported and the input data point will likely be rendered useless. In cases where a match was found, but it seems unlikely, the user will be alerted via the confidence of the match. In some cases, however, a match will be reported with a high degree of confidence yet examining the points more closely will show impossible travel behavior. This error most commonly manifested in monotonicity issues within a trajectory. Specifically, cases were identified in which a vehicle trajectory was matched to the network in such a way that one or more of its points in an increasing sequence of time would be mapped upstream of the preceding point(s), thus leading to non-sensical negative distance (and in turn speed calculations). Such cases, which may have arisen from location data subject to measurement noise being reported by a very slow-moving vehicle (i.e., one in a jam state), were rare. Nonetheless, they were filtered out of the final dataset when identified.

In terms of temporal resolution, the main benefit with respect to probe vehicle trajectories is that some data providers report location data at a very high resolution, about one second at the lowest, whereas loop detectors typically report data at intervals of no less than 20 seconds to as high was several minutes. On the other hand, some probe data providers report location data at a frequency of several minutes. Depending on the application, this data may or may not be appropriate to use as if one wants to examine the trajectory at a higher-sampling rate, accuracy would be lost as a result of interpolation. A final factor that many would see as a benefit from using probe vehicle trajectory data is that such data are provided in a disaggregate manner, allowing scaling to any spatial-temporal window for which data are available. This can be hugely beneficial for allowing customization of study parameters and the associated benefits will only grow as more and more vehicle location data is collected.

When one uses data collected from probe vehicles it is important to note that such data only represents a sample of the population, and while it may be representative of broader trends, it can also be subject to several biases as follows. First is the issue of data sparsity. Depending on the penetration rate of vehicles reporting location data (for the INRIX Trips data, 1.5-5.5% of the total vehicle population is estimated to be reporting data at any time), there can often be several locations and time periods across a network for which no data or data from a very small sample of probes are available. While traffic volume is often correlated with penetration rate, there are certainly cases in which it is not. If such correlations are present and can be shown over time however, a common assumption in periods of no data reported may be free-flow conditions (i.e., those associated with low volume). In cases where such correlation is not present and more generally in cases with small sample sizes of probe data, one must be aware of the inherent potential for bias in tasks like speed calculation depending on numerous factors including vehicle type, roadway functional classification, driver profile, roadway geometry. As a simple example, imagine a scenario on a low-volume rural freeway where only one tractor trailer is reporting location data. If such vehicle traverses sections of the roadway with large upgrades and reports location data in said areas, one may get an inaccurate picture that such roadway is congested if they based their decision on travel speed alone. While the preceding describes just a few of many potential biases associated with probe data, they are important to look out for. A much more detailed summary of such issues and the larger concept of probe vehicle sampling can be found in (Henrickson 2018).

3.2 CRASH DATA AND ROADWAY GEOMETRICS

For this study, two different forms of crash data were applied. The first form involves a disaggregate dataset comprised of crash reports for individual freeway crashes. The second primary crash dataset investigated was comprised of crash counts (specifically animal-vehicle collision counts) on roadway segments over time (i.e., aggregate data). Each dataset as well as relevant supporting data (e.g., roadway geometrics) is described in the following sub-sections.

3.2.1 *Crash Report Data from Law Enforcement (Disaggregate)*

Police crash report data were requested from and provided by WSDOT for this research. Data were requested for state routes in King, Pierce, and Snohomish counties in Washington between 1/1/2015 and 12/31/2017. Spatially, these are the crashes that occurred on the Seattle area freeway network shown in Figure 3-1 (I-5, I-90, I-450, and SR-520). In total, 46,791 crashes occurred on state routes in the 3 aforementioned counties over the 3-year time period of the data request. Ultimately, however, crash report data was only used from the May 2017 for the real-time crash prediction analysis. Such dataset consisted of a total of 607 crashes, distributed across the study area by highway as shown in Table 3-7.

Table 3-7 Distribution of May 2017 Crashes by Highway for Project Study Area

Highway	Number of Crashes
I-5	291
I-90	72
I-405	217
SR-520	27

Data elements shown on the crash report records are quite typical and include attributes such as roadway on which the crash occurred, milepost at which the crash occurred, date and time of crash, crash severity, vehicle types involved, weather, description of the crash events etc. A summary of the key data elements in the police crash reports is shown in Table 3-8.

Table 3-8 Summary of Data Elements in WSDOT-Provided Police Crash Report Data

Variable Name	Description
City	City where crash occurred
Primary trafficway	Roadway where crash occurred
Milepost	Milepost where crash occurred
Report number	Crash report ID
Date	Date on which crash occurred
Time	Time at which crash occurred/was reported
Most severe injury type	Crash injury severity
# Inj	Number of injuries
# Fat	Number of fatalities
# Veh	Number of vehicles involved
Vehicle 1 type	Type of first vehicle
Vehicle 2 type	Type of second vehicle
Junction relationship	Whether or not crash was at/related to intersection
Weather	Weather at time of crash
Roadway surface condition	Roadway surface condition at time of crash
Lighting condition	Lighting condition at time of crash
First collision type/object struck	Type of primary collision
Vehicle 1 action	Action of first vehicle prior to crash
Vehicle 2 action	Action of first second prior to crash
MV driver contributing circumstance 1 (vehicle 1)	Circumstance of crash for first motor vehicle
MV driver contributing circumstance 1 (vehicle 2)	Circumstance of crash for second motor vehicle
First impact location	Lane number or shoulder indicator, increasing/decreasing travel direction

3.2.2 *Aggregate Crash Data for Crash Frequency Analysis*

The dataset used in this study was based upon that used in Lao et al. (2011). Specifically, it was collected to model animal-vehicle collisions along ten highways in Washington State over a total of 752 road segments. Highways for which data were collected include US-2, SR-8, SR-20, I-90, US-97, US-101, US-395, SR-525, and SR-970, and in total. The dependent variable represents the number of animal (white-tailed deer, mule deer, or elk) carcasses removed from each road segment over a five-year period from 2002 to 2006; in total, there were 2,607 reported animal-vehicle collisions. A summary of the data, including the explanatory variables and relevant summary statistics can be seen in Table 3-9. For binary variables, the mean shows the proportion of "yes" values in the data. In general, the dataset is rather typical of that used in a crash frequency analysis as the predictors include things like AADT, properties of the roadway geometry, and other roadway-inventory-related predictors. More detailed information on the dataset can be found in Lao et al. (2011).

Table 3-9 Summary Statistics for Animal-Vehicle Collison Dataset

Variable	Description	Min	Max	Mean	SD
Carcass	Number of carcasses per segment	0	53	3.47	6.86
AADT	Annual average daily traffic	612	120173	7721.85	10820.29
Access	Restrictive access control (0=No, 1=Yes)			0.15	
Spd_limt	Speed limit (miles per hour)	25	70	58.60	6.81
Trkpcts	Truck percentage (%)	0	54.16	15.54	8.88
Nolanes	Number of lanes	2	7	2.48	0.95
Seg_lng	Segment length (miles)	0.5	1	0.69	0.14
TerRol	Terrain type rolling (0=No, 1=Yes)			0.76	
TerMou	Terrain type mountainous (0=No, 1=Yes)			0.13	
Lanewid	Lane width (feet)	10	17	11.78	0.55
Lshlw	Left shoulder width (feet)	0	20	6.02	2.77
Rshlw	Right shoulder width (feet)	0	26	8.64	6.24
White	White-tailed deer habitat (0=No, 1=Yes)			0.36	
Elk	Elk deer habitat (0=No, 1=Yes)			0.36	
Mule	Mule deer habitat (0=No, 1=Yes)			0.60	

3.2.3 *Ancillary Data*

While a variety of different data types can be used to supplement vehicle crash and roadway inventory data, the only such set considered here described the weather in Seattle in May of 2017. Weather data were retrieved from http://www.weatherunderground.com for the entire month of May, and the data collection location was the Seattle-Tacoma International Airport (SEA-TAC). Historical weather data were typically reported on an hourly basis and measurements were taken for variables including temperature, dew point, humidity, wind direction, wind speed, amount of precipitation, weather condition (e.g., cloudy, foggy, etc.), etc. Ultimately, only precipitation and weather condition data were used to supplement the crash report data used in the real-time crash prediction modeling section. On a final note, it is important to point out that there is certainly potential for bias/issues with only using one weather station, given (1) the breadth of the study area and (2) the use of hourly data. The furthest location away from the airport in the study area is approximately 30 miles, and weather patterns can obviously change in intervals much less than an hour. While weather data with higher spatial and temporal resolution have their advantages, the hourly data was used due to its easy availability and more broadly as the consideration of weather impact on crashes itself is not a primary focus of this study.

Chapter 4. REAL-TIME CRASH PREDICTION STUDY DESIGN

4.1 PROBLEM STATEMENT AND MOTIVATION

As discussed in the literature review section, real-time crash prediction modeling refers to crash modeling on a micro-scale. Specifically, factors shown to be correlated with crash occurrence are used to model the likelihood of a crash, typically in a binary classification problem where success event (1) is described as a crash and a failure event (2) would be a non-crash. Typically, traffic-flow-related variables are calculated based upon data from traffic detectors (most commonly loop detectors) and grouped according to whether they preceded a crash event or not. This aggregated data is then used to model the aforementioned outcomes in a variety of models, though binary logit models are quite common.

This dissertation does not deviate from the general concept of real-time crash prediction modeling, but rather tries to bring some novelty to the solution through several different ways. First of all, this study applies vehicle trajectory data collected from probe vehicles. Unlike other studies that have applied loop data or examined trajectories extracted from video data, not all vehicles are captured in the sample of probes. Hence, the goal is to determine if a small sample of the traffic can be used to draw inferences for the larger population when it comes to association between traffic conditions and crash occurrence. Additionally, the use of trajectory data allows calculation of disaggregate features (i.e., at the vehicle trip level) over a much broader spatial and temporal domain than that available via the fixed-location and fixed-time interval reporting from most traffic detectors (e.g., loops); that is to say the trajectory data is often sampled at higher resolution than loop data. These disaggregate features can also include metrics describing higher-order position derivatives than just velocity, namely those related to acceleration and sometimes jerk. It is of interest in this study to see if such new variables add any predictive power to the crash prediction models.

Further, while it has been noted that preceding studies rarely investigated datasets with more than 500 crashes, the initial dataset considered in this study contains 607 crashes distributed over 220.24 roadway miles (Hossain et al. 2019). Additionally, several factors in the study design are varied to see their impact on the dataset generation and in turn, the modeling results. This is a first step towards the issue of investigating the way cases and controls are chosen as described by Roshandel et al. (2015). Finally, this study focuses on providing a clear and proper interpretation of results. Notably, conditional logit models are not intended for use in prediction applications in the conventional sense (i.e., predicting the probability of success (here a crash) conditioned on the given set of covariate values) due to the fixed nature of the case control design. Despite this issue, some studies still present a confusion matrix and describe the predictive power their conditional logit models (Abdel-Aty et al. 2004).

In the following chapter, the steps of the RTCPM study with probe vehicle data are presented. First, a necessary background on case-control study design is provided as such design is the basis for nearly all datasets used in RTCPM applications. Next, the modeling framework used to estimate crash risk by comparing variables describing traffic conditions associated with crashes (i.e., pre-crash conditions) and non-crashes (i.e., non-crash normal conditions) is presented. Once the statistical framework is laid out, the data preparation and feature design for this study are discussed. Then, models are developed on a variety of datasets to examine the impact of several factors on RTCPM. These results are then discussed, interpreted, and compared to those from preceding studies. Finally, the chapter concludes with a discussion of some limitations of the work and next steps.

4.2 CASE-CONTROL STUDY DESIGN

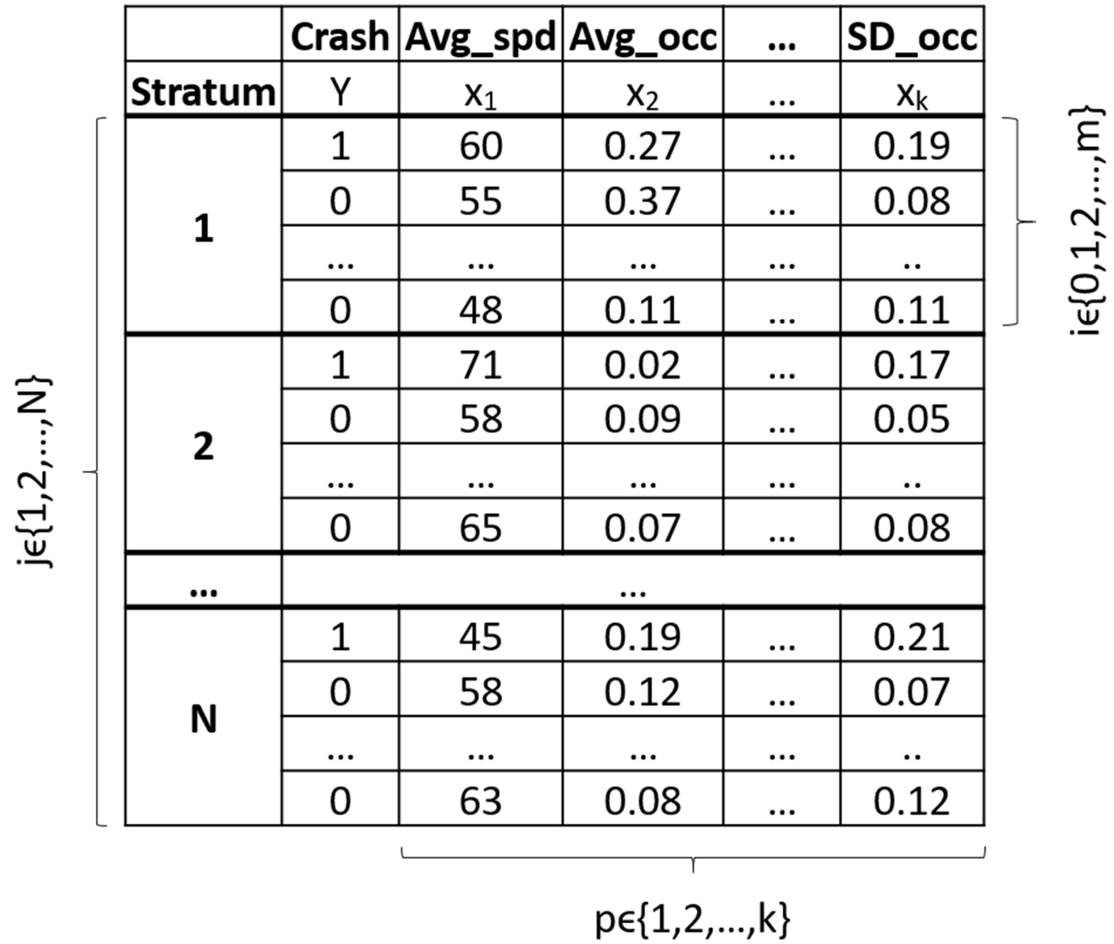
As previously mentioned, datasets used in RTCPM applications are typically developed under a case-control study design. Such study designs are extremely common in medical literature, where the goal may be to examine some propensity for a disease or condition based on several risk factors (Brinton et al. 1992; Schlesselman 1982; Selby et al. 1992; Sokejima and Kagamimori 1998). The analog in RTCPM thus being investigating the correlation between traffic-related features and crashes. Commonly, the basic idea behind a case-control study design is to collect data in a way such that one is attempting to control for some number of potential confounding variables in the data selection process as opposed to controlling for the confounders via effects in the model itself. With this control application, the logic is that one can isolate and study the relations between the desired variables only and the outcome.

The aforementioned effort towards control is typically conducted via a procedure known as matching in the study design. Here, the experimenter selects cases (entities who are associated with a "success" outcome) and controls (entities who are associated with a "failure" outcome) such that all entities in a given group are selected due to having similar values for one or more pre-defined confounding/matching variables. Two types of matching are common in case control designs, those being individual and frequency matching. Individual matching takes a case and matches one or more controls to it based on a shared/similar value for one or more matching variables; each set of one case and one or more controls is known as a stratum. Frequency matching refers to the ensuring an equal distribution in values/ranges of values of the matching variable between the cases and controls across each stratum (Kleinbaum et al. 2007). In either case, the idea the main idea is that intra-stratum (i.e., within stratum) variance in the matching variables should be small, while inter-strata variance in the matching variables can be larger. This dissertation will focus on the application of an individual matched case-control study design. It is important to note that the numbers of controls in each stratum do not have to be equal across strata, however, each stratum is required to have one case (Breslow and Day 1980; Gould 2000). That said, number of controls (just like sample size in general) will affect the statistical power. On a final note, one does not have to necessarily attempt to account for potential confounders in the study design; in such cases, the design is referred to as an unmatched analysis. In such case, there is only one stratum.

In terms of matching variables, medical studies often match on factors such as age, gender, socioeconomic status, race, etc. (Breslow and Day 1980; Stevens 2020). For RTCPM applications, matching variables used to define non-crash/normal conditions primarily include time of day, day of week, functional classification of roadway, and location (Hossain et al. 2019). In general, the matching factors are chosen such that they represent confounders whose impact may not be able to be directly measured (e.g., what driver population typically comprises the traffic stream at a certain time of day at a certain milepost on a freeway?) (Graaf et al. 2011).

4.3 MODELING METHODOLOGY

In alignment with the preceding section, the data to be used in this study was collected under a case-control study design. Assume the dataset is comprised of N cases (i.e., crashes), each of which is matched to at most m controls (note, one must acknowledge the situation as at most since not all cases will be matched to the same number of controls, nor do they need to be) (Gould 2000; Stevens 2020); note also that the matching procedure will be defined in a forthcoming section. Such data can be visualized in a two-dimensional array such as that shown in Figure 4-1. In this array, one can see there is one stratum per crash, each of which has at most m+1 data points (one crash and m controls), and a total of N strata. Each row in a given stratum represents either a case or control, based on the value of the dependent variable, and each other column represents the values recorded for that data point over up to k different predictors or features. In RTCPM applications, these features are typically descriptors of traffic flow for conditions preceding a crash (i.e., pre-crash conditions, cases) or conditions in which no crash took place within a specified spatial-temporal window. Since most preceding studies used loop detector data, that features shown in Figure 4-1 are attributes commonly available from loop detectors. It is clear from the data layout that for RTCPM, one desires to model a binary outcome (crash/no-crash) as a function of several covariates. Thus, a logical choice of model would be one permitting binary classification, e.g., a logit model. That said, the stratified nature of the design makes it such that the conditional logit model is more commonly used, as it can account for the effects of stratification in the study design (i.e., the matched study design).



	Crash	Avg_spd	Avg_occ	…	SD_occ
Stratum	Y	x_1	x_2	…	x_k
1	1	60	0.27	…	0.19
	0	55	0.37	…	0.08
	…	…	…	…	..
	0	48	0.11	…	0.11
2	1	71	0.02	…	0.17
	0	58	0.09	…	0.05
	…	…	…	…	..
	0	65	0.07	…	0.08
…	…				
N	1	45	0.19	…	0.21
	0	58	0.12	…	0.07
	…	…	…	…	..
	0	63	0.08	…	0.12

jϵ{1,2,…,N}
iϵ{0,1,2,…,m}
pϵ{1,2,…,k}

Figure 4-1 Array Representation of Case-Control Study Design Data

In the following, a detailed explanation of the conditional logit model and its interpretation is provided. First, recall the typical logistic regression model, a generalized linear model (GLM) with a log link function, shown in the following. Equation 4-1 can be derived via maximum likelihood estimation (MLE) (Agresti 2007).

$$logit[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \boldsymbol{X\beta} \tag{4-1}$$

Where,

$\pi(x)$ = probability of success at given value of x;

β_0 = intercept term;

$\boldsymbol{X}$ = design matrix (n by p);

$\boldsymbol{\beta}$ = vector of regression coefficients (p by 1).

A key interpretation of the logit model is that the odds of "success" are proportional to the exponentiation of the linear predictor. This still holds true for conditional logit models, however, in the conditional logit models there are other key differences. Consider an example where the probability (unconditional) of a "success" (i.e., crash) is defined by the logit model as shown in Equation 4-1. This can be described via the following:

$$\frac{\pi(y_i)}{1-\pi(y_i)} = \exp(\beta_0 + \boldsymbol{x_i\beta}) \tag{4-2a}$$

Stated in another way, the aforementioned can be written as follows.

$$\pi(y_i) = \frac{\exp(\beta_0+\boldsymbol{x_i\beta})}{1+\exp(\beta_0+\boldsymbol{x_i\beta})} \tag{4-2b}$$

The issue with Equation 4-2b is that there is no consideration of the stratified sampling, that is to say, some sort of conditioning must be considered in each group (i.e., stratum) (Gould 2000). As a simple example to understand the conditioning, consider a stratum defined with one case (i.e., crash) and one control (i.e., non-crash). Based on the data design, what is actually estimated is as follows (Gould 2000):

$$P(1\ crash\ and\ 1\ noncrash|1\ crash) = p(y_1 = 1\ and\ y_2 = 0|y_i = 1) \tag{4-3a}$$

Equation 4-3a can be expanded as follows, via Bayes' theorem.

$$\frac{P(y_1=1)*P(y_2=0)}{P(y_1=1)*P(y_2=0)+P(y_1=0)*P(y_2=0)} \tag{4-3b}$$

Then, finally, by substitution of (4-2b) into (4-3b), the result is the following.

$$P(y_1 = 1, y_2 = 0|y_1 = 1) = \frac{\exp(\boldsymbol{x_1\beta})}{\exp(\boldsymbol{x_1\beta})+\exp(\boldsymbol{x_2\beta})} \tag{4-4}$$

It is very important to note that in Equation 4-4 no intercept term is present. This is due to the fact that it would always cancel, and so it is not estimated/presented in output from statistical programs. Another consideration to think about here is that the nature of design fixes the ratio of success to failures, the exact data on which an intercept would depend.

For the full presentation of the conditional logit model, return to the case for a sample of N crashes. Each of the N crashes (and its corresponding traffic conditions) will be matched to a total of up to m data points representing non-crash scenarios. The matching is done based on criteria such as location, time period, etc. Now define a stratum as a set of one crash data point and its corresponding matched non-crash data points. Thus, there is N strata and each individual stratum is comprised of m+1 data points. Each data point consists of a set of independent variables describing traffic flow and other variables (e.g., weather conditions) and a dependent variable (a binary indicator noting if the condition was associated with a crash or not) (Abdel-Aty et al. 2004; Collett 1991).

Now, consider the case where the i^th^ data point in the j^th^ stratum corresponds to a crash. Equation 4-5 parametrizes this probability as a logistic regression model, specifically as the conditional logit model (again, note there is no intercept) (Abdel-Aty et al. 2004; Collett 1991).

$$logit\left[\pi_j\left(x_{ij}\right)\right] = \alpha_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} \tag{4-5a}$$

Where,

logit = link function, log(odds ratio);

π_j = probability of a crash in the jth stratum (j=1, 2, …, N);

α_j = stratum-specific intercept;

x_{ij} = ith data point in the jth stratum (i=0,1, 2, …, m);

β_k = regression coefficient corresponding to kth feature (e.g., average speed, k=1, 2, …, p); and

x_{kij} = value of the kth feature for the ith data point in the jth stratum.

Additionally, the likelihood function that when maximized will give the model coefficients in Equation 4-5a is shown in Equation 4-5b Abdel-Aty et al. 2004).

$$L(\beta) = \prod_{j=1}^{N}\left[1 + \sum_{i=1}^{m} \exp\left\{\sum_{k=1}^{p} \beta_k\left((x_{kij} - x_{0ij})\right)\right\}\right]^{-1} \tag{4-5b}$$

Finally, with regard to the conditional logit model, the odds ratio is defined in the same way as for the standard logit model. That is to say, any of the exponentiated regression coefficients (for continuous variables, at least) can be interpreted as a multiple on the odds of success corresponding to a one unit increase in the continuous predictor (while holding all other predictors constant) (Agresti 2007).

In addition to the conditional logit model, itself a large part of the analysis, herein, kernel density estimation (KDE) is introduced in the following, as it was used in exploratory analyses to compare distributions of features between crash and non-crash events. At its core, KDE is a non-parametric density estimation technique. For a random variable X, from which N randomly sampled observations are assumed to be independent and identically distributed, the kernel density estimator is defined in the following equation (Hastie et al. 2009). In this study, bandwidth is selected automatically through procedures in (Bowman and Azzalini 2019).

$$\widehat{f_h}(x) = \frac{1}{N}\sum_{i=1}^{N} K_h(x - x_i) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{4-6}$$

Where,

$\widehat{f_h}(x)$ = kernel density estimate at x;

K = kernel function; and

h = bandwidth parameter.

For this study, a Gaussian kernel (Equation 4-7) is applied in all KDE applications.

$$K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \qquad (4\text{-}7)$$

Where,

$z = x\text{-}x_i$;

And all other variables are as previously described.

A final important point to discuss in the methodology for the RTCPM portion of the dissertation is that when using conditional logistic regression, prediction and use of goodness of fit tests based on prediction (e.g., Hosmer-Lemeshow) cannot be used since no intercept is estimated. Despite this important point, some studies have still presented predictions (i.e., confusion matrices) for crash occurrence under a conditional logit model (Abdel-Aty et al. 2004).

4.4 Data Preparation and Feature Design

With the background on matched case-control logistic regression, the next step is to consider other important parts of the study design. Specifically, this section will focus on issues such as feature design, data processing, criteria for matching cases and controls, and some important issues inherent to the probe vehicle trajectory data to be aware of when applying it.

4.4.1 *Data Overview*

The main data sources to be used for the matched case-control study were the (1) probe vehicle trajectory data provided by INRIX and (2) the disaggregate crash data provided by WSDOT. Both datasets were examined under the time frame of May 1-31 2017 (the entire month).

4.4.2 *Data Preprocessing*

The first task in the RTCPM modeling work-flow was data pre-processing. Specifically, the crash data from WSDOT was reduced to only include crashes taking place on the specified highways and within the specified milepost ranges as described in Chapter 3. Following this, the total crash count was 607. For each crash, key data elements were extracted from the respective crash reports as follows:

- Crash location: route number, direction, milepost;
- Time of crash (converted to Unix epoch, seconds since Jan. 1 00:00, 1970, for convenience);
- Crash report number (used later to define strata); and
- Whether the crash was a single-vehicle or multi-vehicle crash.

The matrix consisting of the individual crashes and their respective features was hence defined as the list of "cases." The next key step was to match each case with one or more controls, a topic to be discussed in the following sub-section. Before moving on though, it is important to note that here, the time of the crash provided on the crash report was correct. While some may doubt the accuracy of this time, a buffer time period before the crash occurred was used to collect pre-crash-related traffic variables (Hossain et al. 2019).

4.4.3 *Considerations for Study Design*

The following section outlines the key attributes of the study design used for the real-time crash prediction models. The first choice to make was how to define the spatial and temporal windows for data collection for cases and controls. In this study, it was decided to collect data over 5-minute intervals for the following two periods for cases: (1) 5-10 minutes before the time of crash and (2) 10-15 minutes before the time of the crash. The choice of aggregating data at the 5-minute level was made based on the use of said interval in many other studies that used said interval and aggregated data across all lanes (Abdel-Aty et al. 2005, 2012; Hossain et al. 2019; Liu and Chen 2017; Yu et al. 2013) and the desire to select a "longer" interval in an effort to accumulate more probe samples and more data from them in that interval. Again, it is important to point out here that for the trajectory data used in this study, there is no option but to aggregate data across all travel lanes as data is not available on a per lane basis due to the spatial resolution of the map matching process. Further, five minutes is a commonly-studied interval for the reporting of aggregated loop detector data. With regard to (2), many previous studies have investigated data in 5-minute intervals anywhere between 5 to 30 minutes (and sometimes other ranges) before a crash, and many have found that variables computed for the time period of 5-10 minutes before the crash were significant (at a given confidence level) in their regression models (Abdel-Aty et al. 2004; Abdel-Aty and Abdalla 2004; Hossain et al. 2019; Pande and Abdel-Aty 2006).

With discussion of the time intervals over which data was collected, the next important decision was to define the spatial window. A benefit of this study compared to the vast majority of previous studies that applied data from point-based detectors was that, pending availability (i.e., the presence of probes at a given time), trajectory data could be queried in any spatial window with regard to the crash location. Previous studies applying loop data collected pre-crash data over a range of spatial locations including as many as five upstream loop stations (Abdel-Aty et al. 2004; Abdel-Aty and Abdalla 2004; Abdel-Aty and Pande 2005) and three downstream loop stations (Ahmed and Abdel-Aty 2012; Shew et al. 2013). For this study, a range of upstream and downstream distances over which to collect data were chosen. For the traffic data collected upstream of the crash, distances of data collection spanning 0.5 miles, 1.0 miles, and 1.5 miles were considered. Downstream of the crash location, a distance of 0.5 miles was considered if data were collected at all (in some cases, no downstream data was considered). Different combinations of upstream and downstream data collection were used in the modeling efforts, a topic explained at the end of this section.

A key part of any matched case-control study is defining the factors on which to match the cases and controls (i.e., how to define the controls). Hossain et al. (2019) define a variety of spatial and temporal criteria for matching and control definition. For this study two different definitions for controls were considered as follows:

1. Control Definition 1: A control is defined as data collected in the same intervals (i.e., 5-minute intervals from 5-10 and 10-15 minutes before the crash time) as the pre-crash data, at the same location (same highway, direction, and milepost range), at the same time, and on the same day of the week. For this dataset, this meant each pre-crash case could be matched to at most four other controls depending on the day of week on which it fell and data availability.
2. Control Definition 2: A control is defined as data collected in the same intervals (i.e., 5-minute intervals from 5-10 and 10-15 minutes before the crash time) as the pre-crash data, at the same location (same highway, direction, and milepost range), and at seven random times occurring between 5-1-17 00:00:00 and 5-31-17 23:59:59. For this dataset, this meant each pre-crash case could be matched to at most seven other controls depending on data availability.

In both definitions of controls, another matching factor was that no crash took place within a range of five miles upstream over a window of at least two hours before the time of the crash. A final factor to consider in these matching definitions is what confounders are potentially accounted for. By collecting data at the same location, the design is attempting to adjust for a variety of roadway inventory variables that would take on the same values for the case and controls in one stratum included but not limited to: number of lanes at the crash location, shoulder widths, ramp density, AADT, etc. One can imagine how many of their variables may have an impact on crash occurrence and as such including them in a regression model would result in a very high number of terms and likely a poor fit. For the first matching definition, by collecting data at the same time as the crash occurred for the controls, the effects of driver population and trends in volume (e.g., peaking) are intended to be implicitly accounted for. This is not the case for the second definition of controls, where only the spatial factors are considered in the matching criteria. A final and intuitive note on this point is that even if one wanted to model the effects of some of the aforementioned variables (e.g., roadway geometry), if said variables were coded as binary or categorical variables, they would cancel out in the conditional logit model if both the case and all matched controls in a given stratum had the same value for said variables. The effects of these non-continuous variables can, however, be investigated via interaction terms if one desires.

One final factor that was selected for examination in the study design was whether to separate out single-vehicle crashes from the data or not as was done in (Yu and Abdel-Aty 2013b). It was hypothesized that multi-vehicle crashes could be more probably described by traffic conditions than single-vehicle crashes, where choices of one driver may be the main factor. As such, this dissertation considered datasets with all crash types (i.e., single- and multi-vehicle) and multi-vehicle only crashes. Per discussion in Roshandel et al. (2015), this dissertation examined many potential study designs to surmise their impact on the RTCPM process. The conditions that define the datasets considered in this study are shown in Table 4-1.

In total, 36 datasets were considered in the modeling efforts; each row in Table 4-1 shows conditions to represent four total datasets (as the two right-most columns each take one of two values at a given time). The variety of datasets examined were intended to cover a range of conditions and ensure several datasets of adequate sample size were available as some stratum had to be removed prior to modeling based on missing data issues.

Table 4-1 Criteria for Different Datasets used in RTCPM Efforts

Dataset Indices	Time Periods for Data Collection	Distance for Upstream Data Collection (mi)	Distance for Downstream Data Collection (mi)	Control Definition Used	Crash Types Considered
1-4	5-10 AND 10-15 min. before crash	0.5	0.0	Control Def. 1 OR Control Def. 2	ALL OR Multi-Veh Only
5-8	5-10, 10-15 min. before crash	1.0	0.0	Control Def. 1 OR Control Def. 2	ALL OR Multi-Veh Only
9-12	5-10, 10-15 min. before crash	1.5	0.0	Control Def. 1 OR Control Def. 2	ALL OR Multi-Veh Only
13-16	5-10, 10-15 min. before crash	0.5	0.5	Control Def. 1 OR Control Def. 2	ALL OR Multi-Veh Only
17-20	5-10, 10-15 min. before crash	1.0	0.5	Control Def. 1 OR Control Def. 2	ALL OR Multi-Veh Only
21-24	5-10, 10-15 min. before crash	1.5	0.5	Control Def. 1 OR Control Def. 2	ALL OR Multi-Veh Only
25-28	5-10, 10-15 min. before crash	1.0	1.0	Control Def. 1 OR Control Def. 2	ALL OR Multi-Veh Only
29-32	5-10, 10-15 min. before crash	1.5	1.0	Control Def. 1 OR Control Def. 2	ALL OR Multi-Veh Only
33-36	5-10, 10-15 min. before crash	1.5	1.5	Control Def. 1 OR Control Def. 2	ALL OR Multi-Veh Only

4.4.4 *Feature Design*

Once the controls were defined, datasets designed based on the characteristics in Table 4-1 were developed. Recall, based on the case control design, for N crashes, each matched with at most m controls (assume m here for convenience), thus the dataset has dimension (N*m) by p dataset where N*m is the row count (with each case having y=1 and each control having y=0) and p being the number of features to be calculated. Also, recall that features are aggregated over time and spatial windows with time periods of five minutes and spatial windows defined both up- and downstream of the crash at varying distances (to be clear upstream and downstream data were collected and aggregated separately).

The majority of features considered in the RTCPMs, i.e., the conditional logit models, were continuous variables based on the vehicle position, and higher order position derivatives, along the WSDOT linear referencing system as ascertained from the map-matched trajectories. Essentially, this step involved querying a large PostGRES database table containing the data for 730,909 trips made up of a total of 18,600,169 waypoints (i.e., individual GPS points) for a given spatial-temporal window, computing values on a per trajectory/trip basis, and finally aggregating said values. In order to ensure some level of data quality, portions of trajectories that were not monotonically increasing in space-time were filtered out of the dataset. The process then involved calculating variables primarily defined based on vehicle speed, vehicle acceleration, and jerk. In all cases, these values were computed on a pointwise basis per trajectory prior to aggregating over time and space. Additionally, in no cases were points added to the trajectory based on interpolating over time or space. The effect of this choice would have little to no effect on high-sample-frequency data (i.e., vehicles reporting location data every second), but would certainly have more of an effect as the sampling frequency decreases. That said, vehicles with low-sampling-frequency data would often not even be considered in the calculations due to the relatively small spatial-temporal windows considered.

When calculating speed, as another quality control metric, data points with speeds greater than 100 miles per hour were removed from the dataset (similar to (Abdel-Aty and Abdalla 2004)), the hope being that this would also have an impact on the data for the higher order position derivatives. When considering acceleration, two additional variables were defined, count of hard braking events (HBE) and count of hard acceleration events (HAE). The threshold used to define an HBE was any deceleration event with value less than -10 ft/s^2 and the threshold for an HAE was any acceleration event with a value exceeding 10 ft/s^2 (Fazeen et al. 2012; Stipancic et al. 2018a). Additionally, a count of severe jerk events was defined based on any events that exceeded the threshold value of -32.4 ft/s^3 (Wolshon et al. 2015).

Now, for an example, assume one wants to examine a dataset that considers all crashes (i.e., single- and multi-vehicle), defines controls under the first definition (same location, same time, same day of week), and aggregates data over a length of 1.5 miles upstream of the crash and 0.5 miles downstream of the crash. For every row in the case-control design matrix (again, of dimension (N*m) by p), one can query data for the given spatial/temporal window, calculate speed, acceleration, and jerk vectors for each vehicle trajectory in said window, and finally aggregate data at the five-minute level to calculate all desired variables both upstream and downstream of the crash location separately. Variables considered in the study, which were calculated for both cases and controls 5-10 and 10-15 minutes before the crash at varying up- and downstream locations are shown in Table 4-2.

Table 4-2 Key Variables used in RTCPM Modeling

Variable Name	Description [units]
Crash Report Number	Unique ID of crash, used to define strata
Crash/no crash (Y)	Dependent variable, whether conditions under which data were collected correspond to a crash (1) or not (0)
Probe Vehicle Sample Size	Count of probe vehicles returned in query of waypoint data
Average Speed	Average speed across all vehicles [mph]
Minimum Speed	Min. speed across all vehicles [mph]
Maximum Speed	Max. speed across all vehicles [mph]
Standard Deviation of Speed	SD of speed across all vehicles [mph]
HBE Count	Count of HBEs across all vehicles
HAE Count	Count of HAEs across all vehicles
Severe Jerk Event Count	Count of severe jerk events across all vehicles
Standard Deviation of Acceleration	SD of acceleration across all vehicles [ft/s^2]
Standard Deviation of Jerk	SD of jerk across all vehicles [ft/s^3]
Rain	Binary indicator (1=rain at given time, 0=no rain)
Fog	Binary indicator (1=fog at given time, 0=no fog)

Besides the variables shown in the preceding table, additional variables were calculated based on those in the table. For example, one such variable was the coefficient of variation of speed defined as follows.

$$CVS_{speed} = \frac{\sigma_{speed}}{\mu_{speed}} \tag{4-8}$$

Other variables calculated included differences between upstream and downstream locations in terms of variables such as average speed and sample size.

To conclude this section, it is important to raise a few further points about the data to be used in the forthcoming modeling sections. First of all, the majority of categorical variables (e.g., those defining roadway geometry) were excluded from the analysis since inherently, their impact cannot be estimated under the conditional logit model structure (as cases and controls typically have same values). Second, it is important to recall that different vehicles have different sampling frequencies and no interpolation as done to add points to any trajectories. Vehicles with very low sampling frequencies would likely be excluded from consideration in the dataset as the definition of the spatial/temporal window would filter them out. As one example, consider a vehicle with a sampling frequency of 5-minutes traveling at 60 mph, and assume the area of investigation is one mile long. In such case, this vehicle would report at most one sample point in the data and thus it would not bias any calculations on speed/acceleration/jerk. A final point of note is that the probe data being applied, while high in volume compared to that used in a lot of previous RTCPM studies, is relatively sparse in terms of availability (depending of course on time and location). As a result, there were several time intervals for which no probes were detected and thus no metrics could be calculated. Any cases or controls for which this occurred were filtered out of the final dataset, under the additional condition that each case was matched to at least one control.

4.4.5 *Modeling Process and Data Summary*

As described previously, a total of 36 different matched case-control datasets were generated for this study, the goal being to examine the impact design factors may have on models. In order to accomplish this goal and the larger goal of developing meaningful real-time crash prediction models, two phases of modeling were conducted. The first phase involved developing 36 different conditional logit models, one for each of the aforementioned datasets. The same procedure was used to develop all of the models. The second phase involved a less procedural and deeper look at the model development process. In this phase, hypotheses on influential variables were examined, and a preliminary analysis of important factors in RTCP was conducted by comparing non-parametric estimates of the random variables calculated separately for cases only and controls only. A permutation test was used, in addition to visual inspection, to assess the differences in the densities between crash/pre-crash and non-crash variables.

As previously mentioned, a relatively static procedure was used to develop the models for the first phase of the RTCPM efforts. The idea here was to have a relatively repeatable process and associated model selection criteria to use to develop each of the 36 models, such that they could be compared after development. In order to do this, a forward stepwise regression procedure considering first-order terms (i.e., no interactions) only was applied (Hastie et al. 2009). The steps of the stepwise procedure were as follows. The use of such a modeling procedure is not to necessarily advocate for its usage as stepwise procedures can sometimes lead to non-intuitive results, but rather to rely on the use of a consistent modeling method.

1. Begin with no terms in the conditional logit model. Determine what single term will lead to the best fit (here, assessed based on the value of the Akaike Information Criterion (AIC)), and add that term (if p-val<0.05).
2. From the remaining predictors, first determine which predictors are substantially correlated (define based on a threshold value of the Pearson correlation coefficient, ρ, exceeding 0.4) with the current predictor in the model and remove them from consideration so as to avoid multi-collinearity. From the remaining set of predictors, determine which one predictor, if any, will improve model fit (again based on AIC). If such a predictor is found, add it (if p-val<0.05). If not, stop here.
3. Repeat Step 2 until no further predictors (that meet the criteria for preventing multi-collinearity as previously defined) can be added (with p-val<0.05) to improve the model fit (i.e., decrease the AIC). At this point, stop and take the final model.

The procedure used for the second phase of the modeling will be described in the modeling results section in advance of the second phase results.

In the following tables, summary data for the predictors used in a selection of the 36 datasets are provided, and general trends are discussed after. As there are a lot of tables, the following only shows summary statistics for datasets with an upstream data collection range of 0.5, 1.0, and 1.5 miles, and a downstream range of 0.5 miles; these data are shown in Table 4-3 through Table 4-14. All other tables can be seen in the appendix. In cases where the variable involved of interest is binary, the "Mean" column is used to show the proportion of ones in the data. Each individual table shows:

1. Data collected based on either Control Definition #1 or #2;
2. Summary statistics for upstream (denoted with a green highlighted cell) or downstream (denoted with an orange highlighted cell) data; and
3. Data from all crash types and data from multi-vehicle crashes only.

Table 4-3 Dataset for Control Def. #1 Upstream=0.5 mi, downstream=0.5mi (UP-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	Time Period	Min	Max	Mean	SD	Min	Max	Mean	SD
Sample Size	5-10 min. before crash	1	15	4.35	2.54	1	15	4.40	2.56
Avg. Speed		8.51	72.30	41.28	13.77	8.51	72.30	40.83	13.61
Min. Speed		0.00	69.77	6.82	13.50	0.00	69.77	6.49	13.15
Max. Speed		27.79	99.98	76.55	10.96	27.79	99.98	76.46	10.91
HBE Count		0	244	4.45	16.00	0	244	4.32	15.56
HAE Count		0	266	3.85	15.77	0	266	3.70	15.35
Severe Jerk Count		0	120	1.08	5.28	0	120	1.02	5.01
SD Speed		2.08	29.75	16.53	5.17	2.08	29.75	16.67	5.06
SD Acceleration		0.04	46.19	2.82	3.39	0.04	46.19	2.81	3.35
SD Jerk		0.00	49.58	2.44	4.43	0.00	49.58	2.40	4.37
CVS		0.03	1.65	0.47	0.24	0.03	1.65	0.48	0.24
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	18	4.42	2.65	1	18	4.48	2.67
Avg. Speed		8.65	74.01	41.67	13.51	8.65	74.01	41.22	13.39
Min. Speed		0.00	69.08	7.23	14.29	0.00	69.08	6.87	13.89
Max. Speed		19.82	99.98	76.34	10.85	19.82	99.98	76.30	10.88
HBE Count		0	240	4.86	16.58	0	240	4.82	16.37
HAE Count		0	260	4.28	16.25	0	260	4.25	16.07
Severe Jerk Count		0	120	1.12	5.38	0	120	1.10	5.36
SD Speed		1.50	31.94	16.42	5.36	1.50	31.94	16.55	5.26
SD Acceleration		0.03	32.84	2.77	3.24	0.03	32.84	2.78	3.27
SD Jerk		0.00	50.70	2.50	4.46	0.00	50.70	2.52	4.49
CVS		0.02	1.44	0.46	0.24	0.02	1.44	0.47	0.24

Table 4-4 Dataset for Control Def. #1 Upstream=0.5 mi, downstream=0.5mi (DOWN-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	17	4.38	2.66	1	17	4.41	2.66
Avg. Speed		8.64	72.30	41.12	13.79	8.64	72.30	40.68	13.68
Min. Speed		0.00	69.77	7.26	14.38	0.00	69.77	6.97	14.03
Max. Speed		28.20	99.98	76.60	11.03	28.20	99.98	76.48	11.02
HBE Count		0	244	4.75	16.63	0	244	4.56	15.95
HAE Count		0	266	4.13	16.32	0	266	3.93	15.69
Severe Jerk Count		0	120	1.15	5.48	0	120	1.09	5.20
SD Speed		0.98	32.68	16.45	5.35	0.98	32.68	16.55	5.24
SD Acceleration		0.02	41.74	2.79	3.29	0.02	41.74	2.80	3.30
SD Jerk		0.00	77.03	2.46	4.77	0.00	77.03	2.46	4.78
CVS		0.02	1.42	0.47	0.24	0.02	1.42	0.48	0.24
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	16	4.40	2.62	1	16	4.46	2.62
Avg. Speed		8.65	73.85	41.50	13.46	8.65	73.85	41.09	13.36
Min. Speed		0.00	64.07	6.96	13.75	0.00	64.07	6.63	13.34
Max. Speed		28.06	99.97	76.37	10.93	28.06	99.97	76.34	10.94
HBE Count		0	272	4.99	17.58	0	272	4.93	17.41
HAE Count		0	165	4.22	15.97	0	165	4.19	15.80
Severe Jerk Count		0	57	1.11	4.88	0	57	1.10	4.83
SD Speed		1.73	29.08	16.45	5.27	1.73	29.08	16.55	5.19
SD Acceleration		0.02	28.30	2.79	3.20	0.02	28.30	2.81	3.24
SD Jerk		0.00	47.90	2.48	4.47	0.00	47.90	2.49	4.51
CVS		0.03	1.20	0.47	0.24	0.03	1.20	0.47	0.24

Table 4-5 Dataset for Control Def. #2 Upstream=0.5 mi, downstream=0.5mi (UP-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	17	3.89	2.52	1	17	3.93	2.54
Avg. Speed		8.51	74.85	46.36	14.18	8.51	74.85	46.07	14.25
Min. Speed		0.00	66.55	10.82	17.99	0.00	66.55	10.57	17.86
Max. Speed		22.39	99.99	77.12	10.83	22.39	99.99	77.06	10.82
HBE Count		0	315	5.02	19.92	0	315	5.11	20.33
HAE Count		0	314	4.41	19.79	0	314	4.52	20.22
Severe Jerk Count		0	120	1.29	7.05	0	120	1.33	7.24
SD Speed		0.36	30.51	14.89	6.14	0.36	30.51	14.95	6.08
SD Acceleration		0.01	45.42	2.96	3.91	0.01	45.42	2.97	3.89
SD Jerk		0.00	75.36	2.63	5.45	0.00	75.36	2.64	5.47
CVS		0.01	1.65	0.39	0.25	0.01	1.65	0.40	0.25
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	17	3.81	2.47	1	17	3.84	2.49
Avg. Speed		8.65	77.31	46.70	14.32	8.65	77.31	46.33	14.38
Min. Speed		0.00	66.70	11.61	18.42	0.00	66.70	11.33	18.27
Max. Speed		19.20	99.98	77.38	10.93	19.20	99.98	77.26	10.92
HBE Count		0	240	4.36	15.71	0	240	4.36	15.86
HAE Count		0	260	3.81	15.70	0	260	3.81	15.90
Severe Jerk Count		0	120	1.05	5.70	0	120	1.07	5.84
SD Speed		0.68	30.70	14.81	6.34	0.68	30.70	14.90	6.31
SD Acceleration		0.02	33.30	2.73	3.49	0.02	33.30	2.75	3.54
SD Jerk		0.00	41.88	2.42	4.64	0.00	41.88	2.43	4.69
CVS		0.01	1.59	0.39	0.25	0.01	1.59	0.39	0.25

Table 4-6 Dataset for Control Def. #2 Upstream=0.5 mi, downstream=0.5mi (DOWN-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	21	3.79	2.49	1	21	3.82	2.50
Avg. Speed		6.98	74.85	46.59	14.32	6.98	74.85	46.27	14.40
Min. Speed		0.00	66.14	12.12	18.85	0.00	66.14	11.87	18.65
Max. Speed		15.56	99.98	77.19	10.71	15.56	99.98	77.06	10.72
HBE Count		0	316	5.14	19.91	0	316	5.19	20.11
HAE Count		0	315	4.57	19.89	0	315	4.64	20.17
Severe Jerk Count		0	120	1.30	6.80	0	120	1.33	6.96
SD Speed		0.59	30.27	14.65	6.26	0.75	30.27	14.71	6.19
SD Acceleration		0.01	44.05	2.76	3.46	0.01	44.05	2.75	3.45
SD Jerk		0.00	78.05	2.44	4.90	0.00	78.05	2.44	4.91
CVS		0.01	1.45	0.39	0.25	0.01	1.45	0.39	0.25
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	16	3.73	2.46	1	16	3.75	2.46
Avg. Speed		8.65	77.31	46.81	14.13	8.65	77.31	46.47	14.21
Min. Speed		0.00	66.70	11.81	18.37	0.00	66.70	11.59	18.24
Max. Speed		25.65	99.85	76.80	10.69	25.65	99.85	76.72	10.74
HBE Count		0	248	4.06	14.51	0	248	4.07	14.65
HAE Count		0	253	3.54	14.27	0	253	3.55	14.44
Severe Jerk Count		0	155	0.96	5.37	0	155	0.97	5.49
SD Speed		1.33	39.15	14.76	6.30	1.33	39.15	14.82	6.27
SD Acceleration		0.01	28.30	2.67	3.35	0.01	28.30	2.66	3.32
SD Jerk		0.00	47.41	2.34	4.62	0.00	47.41	2.33	4.62
CVS		0.02	1.59	0.38	0.25	0.02	1.59	0.39	0.25

Table 4-7 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=0.5mi (UP-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				**Multi-Veh Only**			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	22	6.37	3.65	1	22	6.44	3.66
Avg. Speed		8.51	72.30	41.52	13.28	8.51	72.30	41.05	13.14
Min. Speed		0.00	69.77	5.18	12.10	0.00	69.77	5.03	11.91
Max. Speed		33.66	99.98	78.74	10.89	33.66	99.98	78.67	10.91
HBE Count		0	244	5.08	15.91	0	244	4.99	15.56
HAE Count		0	266	4.27	15.62	0	266	4.16	15.29
Severe Jerk Count		0	120	1.18	5.12	0	120	1.14	4.89
SD Speed		0.93	29.86	16.88	5.02	0.93	29.86	16.99	4.94
SD Acceleration		0.02	38.53	3.12	3.50	0.02	38.53	3.11	3.47
SD Jerk		0.00	66.99	2.85	4.81	0.00	66.99	2.83	4.77
CVS		0.02	1.84	0.47	0.24	0.02	1.84	0.48	0.23
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	22	6.45	3.65	1.00	22.00	6.52	3.68
Avg. Speed		8.81	70.60	41.82	13.16	8.81	70.60	41.32	13.06
Min. Speed		0.00	65.27	5.17	12.30	0.00	65.27	5.02	12.13
Max. Speed		14.12	99.98	78.54	10.69	14.12	99.98	78.41	10.68
HBE Count		0	240	5.85	17.71	0	240	5.72	17.19
HAE Count		0	260	5.01	17.27	0	260	4.87	16.76
Severe Jerk Count		0	120	1.35	5.67	0	120	1.30	5.41
SD Speed		1.47	28.93	16.73	5.04	1.47	28.93	16.84	4.97
SD Acceleration		0.02	32.14	3.09	3.21	0.02	32.14	3.10	3.24
SD Jerk		0.00	42.20	2.86	4.42	0.00	42.20	2.86	4.44
CVS		0.02	1.44	0.47	0.23	0.02	1.44	0.47	0.23

Table 4-8 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=0.5mi (DOWN-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	17	4.20	2.64	1	17	4.24	2.64
Avg. Speed		8.64	72.30	41.60	13.78	8.64	72.30	41.13	13.68
Min. Speed		0.00	69.77	7.97	15.05	0.00	69.77	7.62	14.63
Max. Speed		17.53	99.98	76.12	11.22	17.53	99.98	76.01	11.23
HBE Count		0	244	4.40	15.84	0	244	4.24	15.22
HAE Count		0	266	3.79	15.53	0	266	3.61	14.96
Severe Jerk Count		0	120	1.06	5.21	0	120	1.00	4.96
SD Speed		0.28	32.68	16.28	5.49	0.28	32.68	16.40	5.37
SD Acceleration		0.00	41.74	2.76	3.33	0.00	41.74	2.77	3.36
SD Jerk		0.00	77.03	2.41	4.75	0.00	77.03	2.42	4.77
CVS		0.00	1.42	0.46	0.24	0.00	1.42	0.47	0.24
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	16	4.23	2.59	1	16	4.29	2.60
Avg. Speed		8.65	73.85	41.96	13.43	8.65	73.85	41.52	13.33
Min. Speed		0.00	64.07	7.58	14.35	0.00	64.07	7.22	13.92
Max. Speed		28.06	99.97	76.07	11.14	28.06	99.97	76.00	11.15
HBE Count		0	272	4.71	17.07	0	272	4.68	16.95
HAE Count		0	165	3.98	15.62	0	165	3.96	15.50
Severe Jerk Count		0	57	1.04	4.68	0	57	1.04	4.64
SD Speed		0.28	30.32	16.31	5.42	0.28	30.32	16.41	5.34
SD Acceleration		0.00	34.30	2.77	3.32	0.00	34.30	2.79	3.37
SD Jerk		0.00	47.90	2.43	4.54	0.00	47.90	2.45	4.58
CVS		0.00	1.20	0.46	0.24	0.00	1.20	0.46	0.24

Table 4-9 Dataset for Control Def. #2 Upstream=1.0 mi, downstream=0.5mi (UP-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	26	5.41	3.53	1	26	5.47	3.57
Avg. Speed		8.51	74.85	46.86	13.67	8.51	74.85	46.60	13.75
Min. Speed		0.00	65.94	8.66	16.68	0.00	65.94	8.48	16.56
Max. Speed		26.19	99.99	79.06	10.60	26.19	99.99	79.01	10.62
HBE Count		0	315	5.27	19.14	0	315	5.37	19.51
HAE Count		0	314	4.43	18.97	0	314	4.53	19.38
Severe Jerk Count		0	120	1.30	6.69	0	120	1.34	6.86
SD Speed		0.36	36.62	15.34	5.96	0.36	36.62	15.39	5.92
SD Acceleration		0.00	42.84	3.21	3.83	0.00	42.84	3.23	3.85
SD Jerk		0.00	69.06	2.83	5.24	0.00	69.06	2.84	5.28
CVS		0.01	1.84	0.39	0.24	0.01	1.84	0.40	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	25	5.46	3.51	1	25	5.50	3.52
Avg. Speed		9.10	77.27	47.28	13.82	9.10	77.27	46.96	13.88
Min. Speed		0.00	66.70	9.30	17.11	0.00	66.70	9.09	16.92
Max. Speed		19.20	99.98	79.30	10.30	19.20	99.98	79.17	10.33
HBE Count		0	240	5.00	15.96	0	240	5.03	16.15
HAE Count		0	260	4.18	15.83	0	260	4.21	16.06
Severe Jerk Count		0	120	1.20	5.75	0	120	1.22	5.88
SD Speed		1.23	30.63	15.15	6.02	1.23	30.63	15.22	5.98
SD Acceleration		0.01	41.68	3.06	3.77	0.02	41.68	3.07	3.78
SD Jerk		0.00	59.48	2.71	4.88	0.00	59.48	2.71	4.91
CVS		0.02	1.60	0.39	0.24	0.02	1.60	0.39	0.24

Table 4-10 Dataset for Control Def. #2 Upstream=1.0 mi, downstream=0.5mi (DOWN-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	21	3.56	2.43	1	21	3.57	2.44
Avg. Speed		6.98	74.85	47.45	14.13	6.98	74.85	47.19	14.19
Min. Speed		0.00	67.43	13.24	19.86	0.00	67.43	12.96	19.66
Max. Speed		15.56	99.98	76.83	10.62	15.56	99.98	76.73	10.63
HBE Count		0	316	4.71	18.71	0	316	4.76	18.90
HAE Count		0	315	4.15	18.71	0	315	4.21	18.97
Severe Jerk Count		0	120	1.18	6.37	0	120	1.21	6.51
SD Speed		0.59	36.62	14.41	6.46	0.75	36.62	14.47	6.40
SD Acceleration		0.01	44.05	2.77	3.52	0.01	44.05	2.77	3.51
SD Jerk		0.00	78.05	2.37	4.79	0.00	78.05	2.36	4.80
CVS		0.01	1.45	0.37	0.25	0.01	1.45	0.38	0.25
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	16	3.51	2.40	1	16	3.53	2.40
Avg. Speed		8.65	77.31	47.63	14.00	8.65	77.31	47.36	14.07
Min. Speed		0.00	66.70	13.02	19.38	0.00	66.70	12.84	19.28
Max. Speed		25.65	99.85	76.49	10.72	25.65	99.85	76.41	10.77
HBE Count		0	248	3.74	13.67	0	248	3.76	13.80
HAE Count		0	253	3.24	13.49	0	253	3.27	13.65
Severe Jerk Count		0	155	0.88	5.00	0	155	0.90	5.12
SD Speed		0.53	39.15	14.48	6.44	0.53	39.15	14.52	6.41
SD Acceleration		0.00	41.67	2.69	3.60	0.00	41.67	2.68	3.59
SD Jerk		0.00	47.41	2.28	4.68	0.00	47.41	2.27	4.68
CVS		0.01	1.59	0.37	0.24	0.01	1.59	0.37	0.25

Table 4-11 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=0.5mi (UP-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				**Multi-Veh Only**			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	27	7.96	4.44	1	27	8.07	4.45
Avg. Speed		8.51	72.30	41.85	13.19	8.51	72.30	41.33	13.05
Min. Speed		0.00	69.77	4.67	11.74	0.00	69.77	4.42	11.37
Max. Speed		33.66	99.98	79.91	10.51	33.66	99.98	79.88	10.52
HBE Count		0	244	5.77	16.76	0	244	5.72	16.52
HAE Count		0	266	4.81	16.32	0	266	4.73	16.08
Severe Jerk Count		0	120	1.32	5.28	0	120	1.28	5.08
SD Speed		2.14	29.49	16.96	4.93	2.14	29.49	17.11	4.82
SD Acceleration		0.02	35.64	3.19	3.30	0.02	35.64	3.20	3.30
SD Jerk		0.00	61.20	2.99	4.63	0.00	61.20	2.98	4.61
CVS		0.03	1.83	0.47	0.23	0.03	1.83	0.48	0.23
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	32	8.11	4.55	1	32	8.22	4.56
Avg. Speed		9.10	69.56	42.09	13.05	9.10	69.56	41.59	12.95
Min. Speed		0.00	62.91	4.44	11.45	0.00	62.91	4.24	11.14
Max. Speed		42.06	99.98	79.93	10.45	42.06	99.98	79.84	10.44
HBE Count		0	240	6.79	19.26	0	240	6.69	18.84
HAE Count		0	260	5.79	18.79	0	260	5.68	18.37
Severe Jerk Count		0	120	1.56	6.02	0	120	1.52	5.81
SD Speed		1.50	28.94	16.85	4.96	1.50	28.94	16.97	4.87
SD Acceleration		0.03	26.29	3.22	3.10	0.03	26.29	3.23	3.13
SD Jerk		0.00	36.63	3.07	4.44	0.00	36.63	3.07	4.47
CVS		0.02	1.39	0.47	0.23	0.02	1.39	0.47	0.23

Table 4-12 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=0.5mi (DOWN-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	17	4.13	2.63	1	17	4.17	2.64
Avg. Speed		8.64	72.30	41.90	13.80	8.64	72.30	41.40	13.69
Min. Speed		0.00	69.77	8.38	15.38	0.00	69.77	8.01	14.96
Max. Speed		17.53	99.98	75.94	11.22	17.53	99.98	75.82	11.23
HBE Count		0	244	4.28	15.60	0	244	4.12	15.01
HAE Count		0	266	3.68	15.29	0	266	3.52	14.74
Severe Jerk Count		0	120	1.03	5.13	0	120	0.98	4.88
SD Speed		0.28	32.68	16.18	5.55	0.28	32.68	16.32	5.43
SD Acceleration		0.00	41.74	2.73	3.34	0.00	41.74	2.74	3.37
SD Jerk		0.00	77.03	2.38	4.73	0.00	77.03	2.38	4.76
CVS		0.00	1.42	0.46	0.25	0.00	1.42	0.47	0.24
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	16	4.17	2.58	1	16	4.23	2.59
Avg. Speed		8.65	73.85	42.26	13.44	8.65	73.85	41.81	13.36
Min. Speed		0.00	64.07	7.91	14.70	0.00	64.07	7.52	14.25
Max. Speed		28.06	99.97	76.03	11.12	28.06	99.97	75.95	11.12
HBE Count		0	272	4.64	16.86	0	272	4.60	16.76
HAE Count		0	165	3.92	15.44	0	165	3.90	15.34
Severe Jerk Count		0	57	1.01	4.61	0	57	1.01	4.58
SD Speed		0.28	30.32	16.22	5.48	0.28	30.32	16.32	5.39
SD Acceleration		0.00	34.30	2.74	3.29	0.00	34.30	2.75	3.33
SD Jerk		0.00	47.90	2.40	4.50	0.00	47.90	2.42	4.54
CVS		0.00	1.20	0.45	0.24	0.00	1.20	0.46	0.24

Table 4-13 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=0.5mi (UP-stream portion only)

Control Definition #2 (same location, random time)									
		All				**Multi-Veh Only**			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	26	6.70	4.27	1	26	6.78	4.31
Avg. Speed		8.51	74.73	47.16	13.50	8.51	74.73	46.92	13.58
Min. Speed		0.00	67.02	7.86	16.09	0.00	67.02	7.70	16.01
Max. Speed		25.26	99.99	80.04	10.21	25.26	99.99	80.00	10.21
HBE Count		0	315	5.65	19.35	0	315	5.77	19.72
HAE Count		0	314	4.70	19.07	0	314	4.81	19.47
Severe Jerk Count		0	120	1.38	6.76	0	120	1.43	6.93
SD Speed		0.36	36.62	15.46	5.86	0.36	36.62	15.51	5.82
SD Acceleration		0.00	36.56	3.28	3.68	0.00	36.56	3.30	3.70
SD Jerk		0.00	60.96	2.92	5.11	0.00	60.96	2.95	5.15
CVS		0.01	1.83	0.39	0.24	0.01	1.83	0.39	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	27	6.83	4.27	1	27	6.88	4.28
Avg. Speed		9.10	73.56	47.54	13.61	9.10	73.56	47.27	13.68
Min. Speed		0.00	66.78	8.48	16.64	0.00	66.78	8.30	16.47
Max. Speed		45.63	99.98	80.37	10.04	45.63	99.98	80.26	10.05
HBE Count		0	240	5.74	17.78	0	240	5.77	17.96
HAE Count		0	260	4.81	17.54	0	260	4.83	17.75
Severe Jerk Count		0	120	1.40	6.42	0	120	1.42	6.55
SD Speed		1.23	29.51	15.23	5.94	1.23	29.51	15.29	5.90
SD Acceleration		0.02	41.68	3.16	3.67	0.02	41.68	3.16	3.69
SD Jerk		0.00	58.17	2.88	4.84	0.00	58.17	2.88	4.87
CVS		0.02	1.60	0.38	0.24	0.02	1.60	0.39	0.24

Table 4-14 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=0.5mi (UP-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1.00	21.00	3.51	2.42	1.00	21.00	3.52	2.42
Avg. Speed		6.98	74.85	47.70	14.13	6.98	74.85	47.45	14.19
Min. Speed		0.00	67.43	13.76	20.23	0.00	67.43	13.49	20.03
Max. Speed		15.56	99.98	76.66	10.63	15.56	99.98	76.55	10.63
HBE Count		0.00	316.00	4.53	18.34	0.00	316.00	4.58	18.51
HAE Count		0.00	315.00	3.99	18.33	0.00	315.00	4.05	18.57
Severe Jerk Count		0.00	120.00	1.14	6.24	0.00	120.00	1.16	6.38
SD Speed		0.59	36.62	14.27	6.53	0.75	36.62	14.34	6.47
SD Acceleration		0.01	44.05	2.72	3.49	0.01	44.05	2.72	3.48
SD Jerk		0.00	78.05	2.30	4.72	0.00	78.05	2.30	4.72
CVS		0.01	1.45	0.37	0.25	0.01	1.45	0.37	0.25
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1.00	16.00	3.45	2.38	1.00	16.00	3.47	2.38
Avg. Speed		8.65	77.69	47.94	14.00	8.65	77.69	47.68	14.07
Min. Speed		0.00	70.48	13.68	19.87	0.00	70.48	13.52	19.79
Max. Speed		25.65	99.91	76.38	10.74	25.65	99.91	76.29	10.79
HBE Count		0.00	248.00	3.64	13.41	0.00	248.00	3.65	13.53
HAE Count		0.00	253.00	3.15	13.23	0.00	253.00	3.17	13.38
Severe Jerk Count		0.00	155.00	0.85	4.90	0.00	155.00	0.87	5.01
SD Speed		0.53	39.15	14.30	6.50	0.53	39.15	14.34	6.48
SD Acceleration		0.00	41.67	2.65	3.58	0.00	41.67	2.64	3.57
SD Jerk		0.00	47.41	2.24	4.64	0.00	47.41	2.23	4.64
CVS		0.01	1.59	0.36	0.25	0.01	1.59	0.37	0.25

Here, a brief overview of the aggregate data collected from the vehicle trajectories used in the different datasets is provided. In terms of the sample size of probe vehicles, for the tables in Section 4.4.5, one will notice that in most cases the minimum number of probe vehicles observed in a given five-minute period describing a pre-crash or normal traffic condition is one (recall all intervals with zero vehicles were filtered out). This low number certainly presents chances for bias in the data, depending on many characteristics of the sampled vehicle such as vehicle type, driver behavior, etc. In a free-flow state on a freeway, traffic conditions are less variable than during periods with worse levels of service; as such, sampling one vehicle in said intervals may give an accurate representation of the true conditions. On the other hand, if the lone sampled vehicle demonstrates behavior substantially different than the average behavior of the traffic stream, such measurements would be biased. The maximum number of vehicles sampled in a five-minute interval in Table 4-3 through Table 4-14 was 27, and in general, as the spatial extent of the study area increased (i.e., data were collected over a greater distance upstream and/or downstream), the sample size of probe vehicles increased.

In terms of speed values in Table 4-3 through Table 4-14, three speed metrics were considered, those being average, minimum, and maximum. Each of these (and their supporting metrics) were calculated using a pool of all sampled vehicle speeds in a given spatial-temporal window. Minimum speeds were as low as 0.00 mph (most likely during jam states), while maximum speeds approached 100.00 mph. That said, all speed values greater than 100.00 were filtered out of the dataset similar to (Abdel-Aty and Abdalla 2004). Average speeds ranged from less than 10.00 mph to values approaching 80.00 mph in cases where the sample of probes was small (i.e., 1).

For values of the counts of hard-braking events, hard-acceleration events, and severe jerk events, values of each ranged from 0 in the case of sampling 1 vehicle, to values in the low hundreds for periods with higher sample sizes. Average values of standard deviations of acceleration and jerk, however, were typically lower than the averages of standard deviation of speed. That said, it is important to interpret the results surrounding higher-order derivatives of speed with caution and consider the potential impacts of lower sample size and lower sampling resolutions. Finally, in terms of the weather-related variables, rain was present for about one percent of the data points for most datasets, while fog was present for no more than two percent of data points.

Chapter 5. REAL-TIME CRASH PREDICTION MODELING RESULTS

The following section presents the modeling results and supporting discussion for the models based upon the study design in the previous chapter.

5.1 PRESENTATION OF RESULTS FOR INITIAL MODELS

As described in Chapter 4, a matched case-control study design was used to generate datasets to study factors associated with crash occurrence, via conditional logit models. The first phase of the modeling involved using a forward stepwise regression procedure (focusing on main effects only) in an effort to compare how different factors including choice/definition of controls, type of crash (all vs. multi-vehicle only), upstream range of data collection (0.5, 1.0, 1.5 miles), downstream range of data collection (0.0, 0.5, 1.0, 1.5 miles) may affect model results and interpretation based on the probe vehicle dataset used herein. For each model, the following are presented:

- Coefficient (coeff) estimates with standard error (SE), Z-statistic (Z stat) value, p-value, estimated odds ratio (i.e., exp(coeff)), and a 95% confidence interval for the odds ratio;
- Test statistics, supporting degrees of freedom (DF), and associated p-values;
- AIC value to measure goodness of fit (the AIC presented was the lowest value found in the stepwise procedure;
- The number of samples (cases [data collected in pre-crash time periods] + controls [data collected under normal conditions]), denoted as n; and

- The number of cases, denoted n_{crash}.

As was the case with the previous section, the presentation of results will focus on a subset of the 36 models, specifically all of those with a downstream data collection distance of 0.5 mi. These models are relatively representative of the entire group of 36 in terms of the main effects, coefficient values, etc. Data on the remaining 24 models is presented in the appendix, and a discussion of trends across all 36 models is provided later in this chapter.

The models with a downstream data collection distance of 0.5 miles are presented in Table 5-1 through Table 5-12. Following their presentation, the results are interpreted.

The models shown in Table 5-1 through Table 5-4 are based upon datasets with a 0.5 mile upstream and 0.5 mile downstream distance. All four of the models had downstream average speed (computed 5-10 minutes before the crash) as a significant variable at the 0.05 level. The downstream speed had a negative coefficient in all cases and an odds ratio ranging from 0.954-0.969, meaning that for every mile per hour increase in average speed (all else held constant), the odds of a crash multiply by between 95-96 percent (i.e., they decrease by 4 percent). Upstream sample size of probe vehicles (calculated in the interval of 5-10 minutes before the crash) was found to be positively correlated with log odds of a crash in three of the models, with an odds ratio implying for every additional probe vehicle sampled (all else held constant), the odds of crash multiply by between 2.1 to 2.9. Finally, two models found downstream maximum speed (calculated 10-15 minutes before the crash) to have a positive coefficient as well.

Table 5-1 Model Results for Dataset 0.5 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #1

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	0.5 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
dn_05_510_avg_speed	-0.032	0.968	0.007	-4.821	1.43E-06	0.956	0.981
up_05_510_sample_size	0.078	1.081	0.028	2.736	0.006	1.022	1.143
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	35.76		2		2.00E-08		
Wald	33.35		2		6.00E-08		
Score (logrank)	34.87		2		3.00E-08		
AIC	998.6994						
n	1507						
n_{crash}	395						

Table 5-2 Model Results for Dataset 0.5 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #1

Multi-Veh Crashes Only, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	0.5 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
dn_05_510_avg_speed	-0.031	0.969	0.007	-4.641	3.46E-06	0.956	0.982
up_05_510_sample_size	0.085	1.088	0.029	2.932	0.003	1.029	1.152
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	35.09		2		2.00E-08		
Wald	32.76		2		8.00E-08		
Score (logrank)	34.26		2		4.00E-08		
AIC	956.3969						
n	1446						
n_{crash}	376						

Table 5-3 Model Results for Dataset 0.5 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #2

All Crashes, Controls=Def. #2 (same location, random time)							
Distance Upstream	0.5 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
dn_05_510_avg_speed	-0.047	0.954	0.005	-10.001	<2e-16	0.945	0.963
dn_05_1015_max_speed	0.016	1.016	0.006	2.763	0.006	1.005	1.028
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	114.50		2		<2e-16		
Wald	100.50		2		<2e-16		
Score (logrank)	111.60		2		<2e-16		
AIC	984.0348						
n	1661						
n_{crash}	406						

Table 5-4 Model Results for Dataset 0.5 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #2

Multi-Veh Crashes Only, Controls=Def. #2 (same location, random time)							
Distance Upstream	0.5 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
dn_05_510_avg_speed	-0.046	0.955	0.005	-8.914	<2e-16	0.945	0.965
dn_05_1015_max_speed	0.016	1.016	0.006	2.682	0.007	1.004	1.028
up_05_510_sample_size	0.057	1.059	0.027	2.113	0.035	1.004	1.116
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	122.30		3		<2e-16		
Wald	105.20		3		<2e-16		
Score (logrank)	118.30		3		<2e-16		
AIC	917.1349						
n	1569						
n_{crash}	384						

The models shown in Table 5-5 through Table 5-8 are based upon datasets with a 1.0 mile upstream and 0.5 mile downstream distance. In two of the models, upstream average speed (in the 5-10 minute period) was found to be negatively correlated with the log odds of a crash; and in the other two models, downstream average speed for the same time interval was found to be negatively correlated as well. In all cases, the odds ratios were quite similar and imply that for every mile per hour increase in average speed (all else held constant), the odds of a crash multiply by approximately 96 percent. Similar to the aforementioned models, downstream maximum speed (as calculated in the 5-10 minute time period) was found to be a significant predictor with an odds ratio ranging of roughly 1.01, suggesting a small impact on the odds of a crash increasing with increasing maximum speed. Finally, one model found downstream sample size of probe vehicles (in the 10-15 minute period) to be a significant predictor at the 0.05 level, with an odds ratio of approximately 1.07.

Table 5-5 Model Results for Dataset 1.0 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #1

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.0 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_10_510_avg_speed	-0.037	0.964	0.007	-5.652	1.59E-08	0.951	0.976
dn_05_510_max_speed	0.014	1.014	0.006	2.464	0.014	1.003	1.025
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	35.86		2		2.00E-08		
Wald	33.66		2		5.00E-08		
Score (logrank)	35.25		2		2.00E-08		
AIC	1110.326						
n	1677						
n_{crash}	431						

Table 5-6 Model Results for Dataset 1.0 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #1

<table>
<tr><th colspan="8">Multi-Veh Crashes Only, Controls=Def. #1 (same time, day of week, and location)</th></tr>
<tr><td><b>Distance Upstream</b></td><td colspan="2">1.0 mi</td><td colspan="3"><b>Distance Downstream</b></td><td colspan="2">0.5 mi</td></tr>
<tr><td colspan="8"><b>MODEL RESULTS</b></td></tr>
<tr><td><b>Variable Name</b></td><td><b>Coefficient</b></td><td><b>exp(coeff)</b></td><td><b>SE(coeff)</b></td><td><b>Z stat</b></td><td><b>p-val</b></td><td><b>Lower 95%</b></td><td><b>Upper 95%</b></td></tr>
<tr><td>up_10_510_avg_speed</td><td>-0.037</td><td>0.964</td><td>0.007</td><td>-5.487</td><td>4.08E-08</td><td>0.951</td><td>0.977</td></tr>
<tr><td>dn_05_510_max_speed</td><td>0.013</td><td>1.013</td><td>0.006</td><td>2.312</td><td>0.021</td><td>1.002</td><td>1.024</td></tr>
<tr><td colspan="8"><b>MODEL FIT</b></td></tr>
<tr><td><b>Test Name</b></td><td colspan="2"><b>Test Statistic</b></td><td colspan="2"><b>DF</b></td><td colspan="3"><b>p-val</b></td></tr>
<tr><td>Likelihood Ratio</td><td colspan="2">33.46</td><td colspan="2">2</td><td colspan="3">5.00E-08</td></tr>
<tr><td>Wald</td><td colspan="2">31.42</td><td colspan="2">2</td><td colspan="3">2.00E-07</td></tr>
<tr><td>Score (logrank)</td><td colspan="2">32.87</td><td colspan="2">2</td><td colspan="3">7.00E-08</td></tr>
<tr><td>AIC</td><td colspan="7">1061.795</td></tr>
<tr><td>n</td><td colspan="7">1604</td></tr>
<tr><td>n_{crash}</td><td colspan="7">410</td></tr>
</table>

Table 5-7 Model Results for Dataset 1.0 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #2

<table>
<tr><th colspan="8">All Crashes, Controls=Def. #2 (same location, random time)</th></tr>
<tr><td><b>Distance Upstream</b></td><td colspan="2">1.0 mi</td><td colspan="3"><b>Distance Downstream</b></td><td colspan="2">0.5 mi</td></tr>
<tr><td colspan="8"><b>MODEL RESULTS</b></td></tr>
<tr><td><b>Variable Name</b></td><td><b>Coefficient</b></td><td><b>exp(coeff)</b></td><td><b>SE(coeff)</b></td><td><b>Z stat</b></td><td><b>p-val</b></td><td><b>Lower 95%</b></td><td><b>Upper 95%</b></td></tr>
<tr><td>dn_05_510_avg_speed</td><td>-0.045</td><td>0.956</td><td>0.005</td><td>-9.388</td><td><2e-16</td><td>0.948</td><td>0.965</td></tr>
<tr><td>dn_05_1015_sample_size</td><td>0.068</td><td>1.070</td><td>0.026</td><td>2.628</td><td>0.009</td><td>1.017</td><td>1.126</td></tr>
<tr><td>dn_05_510_max_speed</td><td>0.012</td><td>1.012</td><td>0.006</td><td>2.186</td><td>0.029</td><td>1.001</td><td>1.023</td></tr>
<tr><td colspan="8"><b>MODEL FIT</b></td></tr>
<tr><td><b>Test Name</b></td><td colspan="2"><b>Test Statistic</b></td><td colspan="2"><b>DF</b></td><td colspan="3"><b>p-val</b></td></tr>
<tr><td>Likelihood Ratio</td><td colspan="2">143.10</td><td colspan="2">3</td><td colspan="3"><2e-16</td></tr>
<tr><td>Wald</td><td colspan="2">125.70</td><td colspan="2">3</td><td colspan="3"><2e-16</td></tr>
<tr><td>Score (logrank)</td><td colspan="2">141.80</td><td colspan="2">3</td><td colspan="3"><2e-16</td></tr>
<tr><td>AIC</td><td colspan="7">1148.22</td></tr>
<tr><td>n</td><td colspan="7">1996</td></tr>
<tr><td>n_{crash}</td><td colspan="7">443</td></tr>
</table>

Table 5-8 Model Results for Dataset 1.0 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #2

Multi-Veh Crashes Only, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.0 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
dn_05_510_avg_speed	-0.047	0.954	0.005	-9.501	<2e-16	0.945	0.964
dn_05_1015_sample_size	0.081	1.085	0.026	3.072	0.002	1.030	1.143
dn_05_510_max_speed	0.013	1.013	0.006	2.251	0.024	1.002	1.024
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	152.90		3		<2e-16		
Wald	131.30		3		<2e-16		
Score (logrank)	150.40		3		<2e-16		
AIC	1069.422						
n	1889						
n_{crash}	419						

The models in Table 5-9 through Table 5-12 are based upon datasets with a 1.5 mile upstream and 0.5 mile downstream distance. All models found average speed (in the 5-10 minute period) to be negatively associated with crash risk. In some cases, the predictor was upstream average speed and for others, it was downstream average speed; regardless, the odds ratio ranged between 1.01 and 1.09. Three of the four models found upstream minimum speed (in the 10 to 15 minute period) to have a negative coefficient, suggesting that the odds of a crash decrease with minimum speed. Downstream sample size (in the 10-15 minute period) was found to be significant in three of the four models with an odds ratio of approximately 1.07-1.09, suggesting a roughly 8 percent increase in the odds of a crash with every unit increase in probe vehicle sample size (all else held constant). Finally, one model found maximum speed (in the 5 to 10 minute period) to have a positive coefficient similar to a previously mentioned model.

Table 5-9 Model Results for Dataset 1.5 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #1

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.5 mi		Distance Downstream			0.5 mi	
MODEL RESULTS							
Variable Name	Coefficient	exp(coeff)	SE(coeff)	Z stat	p-val	Lower 95%	Upper 95%
up_15_510_avg_speed	-0.037	0.964	0.007	-5.491	4.01E-08	0.951	0.977
dn_05_510_max_speed	0.013	1.013	0.005	2.353	0.019	1.002	1.024
up_15_1015_min_speed	-0.013	0.987	0.006	-2.096	0.036	0.974	0.999
MODEL FIT							
Test Name	Test Statistic		DF		p-val		
Likelihood Ratio	46.16		3		5.00E-10		
Wald	42.03		3		4.00E-09		
Score (logrank)	44.65		3		1.00E-09		
AIC	1139.461						
n	1734						
n_{crash}	443						

Table 5-10 Model Results for Dataset 1.5 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #1

Multi-Veh Crashes Only, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.5 mi		Distance Downstream			0.5 mi	
MODEL RESULTS							
Variable Name	Coefficient	exp(coeff)	SE(coeff)	Z stat	p-val	Lower 95%	Upper 95%
up_15_510_avg_speed	-0.035	0.966	0.007	-5.228	1.71E-07	0.954	0.979
dn_05_510_sample_size	0.064	1.066	0.026	2.398	0.017	1.012	1.122
MODEL FIT							
Test Name	Test Statistic		DF		p-val		
Likelihood Ratio	38.98		2		3.00E-09		
Wald	36.52		2		1.00E-08		
Score (logrank)	38.26		2		5.00E-09		
AIC	1089.348						
n	1654						
n_{crash}	420						

Table 5-11 Model Results for Dataset 1.5 mi Upstream, 0.5 mi downstream, All Crashes, Control Definition #2

All Crashes, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.5 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
dn_05_510_avg_speed	-0.038	0.963	0.005	-8.043	0.000	0.954	0.972
dn_05_1015_sample_size	0.071	1.073	0.025	2.783	0.005	1.021	1.129
up_15_1015_min_speed	-0.013	0.987	0.005	-2.579	0.010	0.977	0.997
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	153.10		3		<2e-16		
Wald	130.70		3		<2e-16		
Score (logrank)	148.60		3		<2e-16		
AIC	1183.36						
n	2083						
n_{crash}	451						

Table 5-12 Model Results for Dataset 1.5 mi Upstream, 0.5 mi downstream, MV Crashes, Control Definition #2

Multi-Veh Crashes Only, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.5 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
dn_05_510_avg_speed	-0.040	0.961	0.005	-8.229	<2e-16	0.952	0.970
dn_05_1015_sample_size	0.085	1.089	0.026	3.245	0.001	1.034	1.146
up_15_1015_min_speed	-0.012	0.988	0.005	-2.261	0.024	0.977	0.998
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	160.60		3		<2e-16		
Wald	135.40		3		<2e-16		
Score (logrank)	156.10		3		<2e-16		
AIC	1105.717						
n	1974						
n_{crash}	427						

5.2 Discussion and Interpretation of Results

In this section, general trends across the first phase of models are discussed. Table 5-13 shows a summary of all of the main effects found to be significant in the 12 aforementioned models. In addition to simply naming the variables, an average value of the coefficient is computed, along with standard deviation, a check of whether the sign of the coefficient is consistent across all models, and the odds ratio computed based on the average value of the coefficient.

Table 5-13 Summary of Significant Predictors (0.05 level) across the 12 Models in Section 5.1

Variable	# Models	Avg. Coeff.	SD. Coeff	Sign Consistent?	Exp(avg_coef)
up_05_510_sample_size	8	0.069	0.024	Y	1.071
up_05_510_avg_speed	4	-0.039	0.007	Y	0.962
dn_05_510_sample_size	1	0.064	N/A	Y	1.066
dn_05_510_avg_speed	8	-0.041	0.006	Y	0.960
dn_05_510_max_speed	5	0.013	0.001	Y	1.013
dn_05_1015_sample_size	4	0.076	0.008	Y	1.079
dn_05_1015_max_speed	1	0.016	N/A	Y	1.016
up_10_510_avg_speed	10	-0.044	0.010	Y	0.957
up_10_1015_max_speed	2	0.013	0.001	Y	1.013
dn_10_510_sample_size	4	0.054	0.001	Y	1.056
dn_10_1015_hbe_count	2	0.007	0.000	Y	1.007
up_15_510_avg_speed	14	-0.045	0.009	Y	0.956
up_15_1015_max_speed	5	0.011	0.001	Y	1.011

From the table, one will note that when possible to calculate, the standard deviations for the coefficients are rather small. This suggests that for the given probe vehicle dataset, the methods of defining controls, as well as considering all crashes versus multi-vehicle crashes only did not have a very large impact on the results. Clearly there is also overlap in the data between the sets, a fact that further reinforces the preceding point. Additionally, for all predictors in the final models, the signs of the coefficients are consistent across all models. Further, the signs on each predictor make intuitive sense and appear to be in line with previous studies. For example, in a meta-analysis of 11 RTCPM studies, Roshandel et al. (2015) noted that over 10 studies that investigated the impact of average speed on crash potential, the summary odds ratio was found to be 0.952, a value quite close to the values associated with average speed in Table 5-13 that range from 0.956-0.962. Additionally, Roshandel et al. (2015) noted that across six studies that investigated average volume in RTCPM applications, such variable was found to be positively associated with crash risk and had a summary odds ratio of 1.001. In this study, the summary odds ratio for the average volume-like variable (here, sampled probe vehicle count, not actual volume) was found to be similar in magnitude 1.056-1.079.

5.3 A Deeper Dive into Real-Time Crash Prediction Modeling

In the preceding section, the groundwork for a series of real-time crash prediction models was laid, and despite the procedural means of model fitting, results were seemingly reasonable and intuitive, at least when compared to past studies. In this section, additional models are built upon taking a closer look into the data, in terms of the distribution of variables between pre-crash and normal time periods, as well as the potential for inclusion of binary/categorical variables in the model through interaction terms. Recall that if a binary/categorical variable has a constant value within a stratum, then its effect cannot be estimated. Further, some additional variables such as upstream-downstream sample size difference and upstream-downstream speed difference were considered. Finally, since the results from last section appeared to be similar between datasets, for this section, only the datasets corresponding to an upstream data collection distance of 1.5 miles and a downstream data collection distance of 0.5 miles, for all crashes (i.e., single- and multi-vehicle) were applied.

First, kernel density estimation was used to examine how values of each of the predictors were distributed when comparing pre-crash versus normal conditions. In each case, a Gaussian kernel was used and bandwidth was automatically selected as discussed in (Bowman and Azzalini 2019). For all variables, the density plots were inspected and the results of the permutation test (i.e., H_0: the two densities are the same) described in Bowman and Azzalini (2019) were also considered. The logic behind this portion of the investigation was to uncover variables whose distributions are markedly different between pre-crash and normal traffic conditions, as it was believed such variables may be good predictors in the conditional logit models for RTCPM. For brevity, Figure 5-1 through Figure 5-6 shows the KDE plots for the following scenarios, and in each plot, the red line represents normal conditions and the green, dashed line represents pre-crash conditions:

- Control definition #1 (same location, same day of week, same time periods), UP-stream 1.5 miles, 5-10 minutes before crash;
- Control definition #1, DOWN-stream 0.5 miles, 5-10 minutes before crash;
- Control definition #1, upstream-downstream differences, 5-10 and 10-15 minutes before crash;
- Control definition #2 (same location, random time), UP-stream 1.5 miles, 5-10 minutes before crash;
- Control definition #2, DOWN-stream 0.5 miles, 5-10 minutes before crash; and
- Control definition #2 upstream-downstream differences, 5-10 and 10-15 minutes before crash.

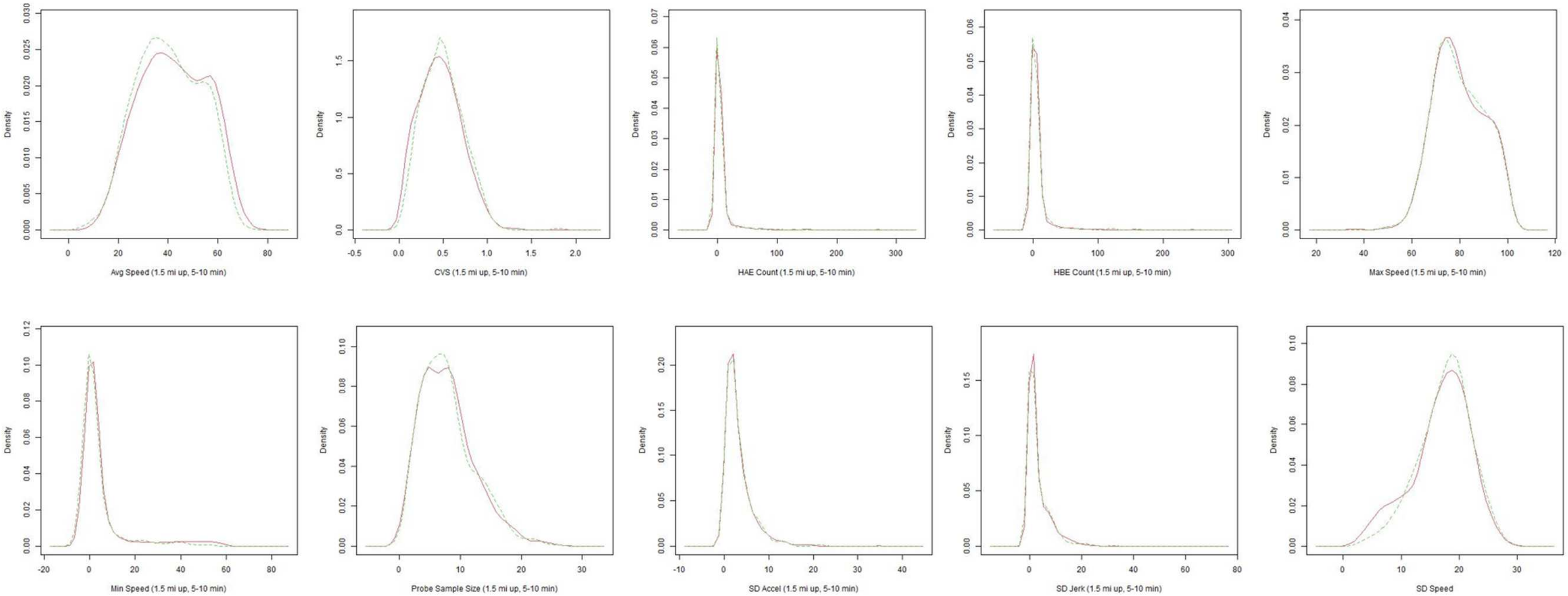


Figure 5-1 KDE Plots for 1.5 mi Up-stream (5-10 minutes), Control Definition #1

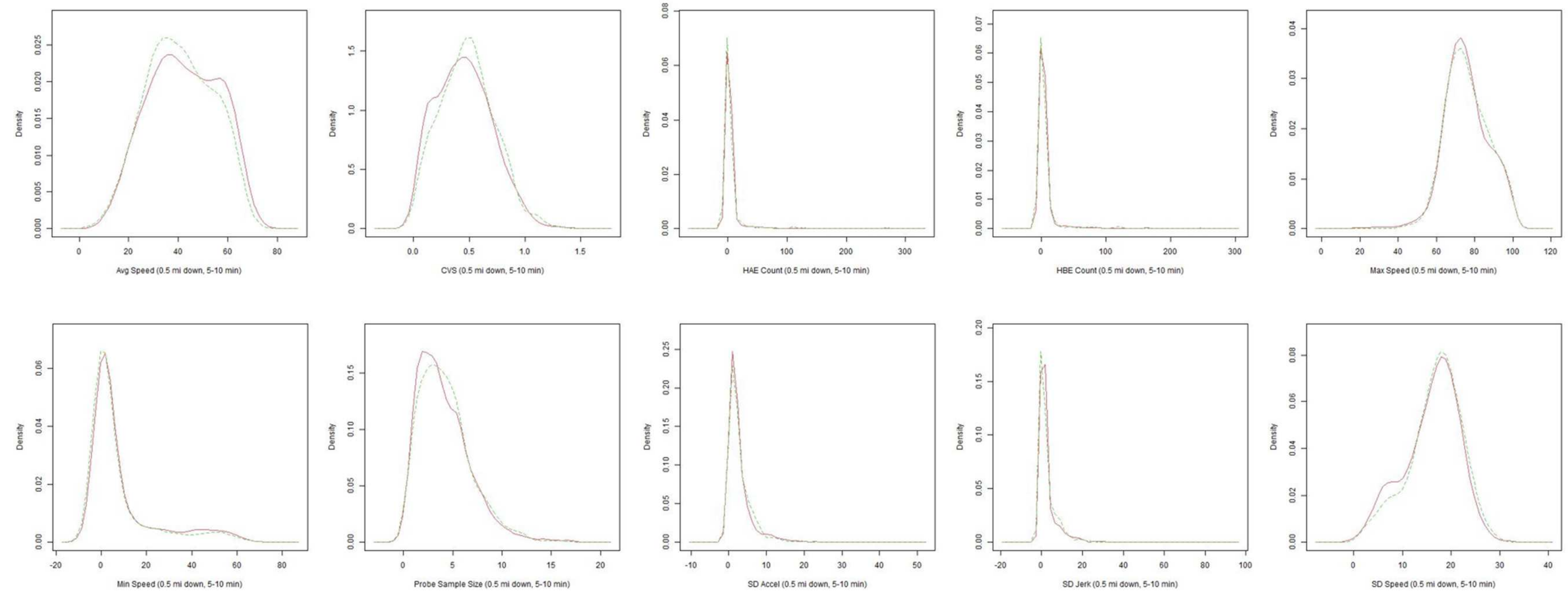


Figure 5-2 KDE Plots for 0.5 mi Down-stream (5-10 minutes), Control Definition #1

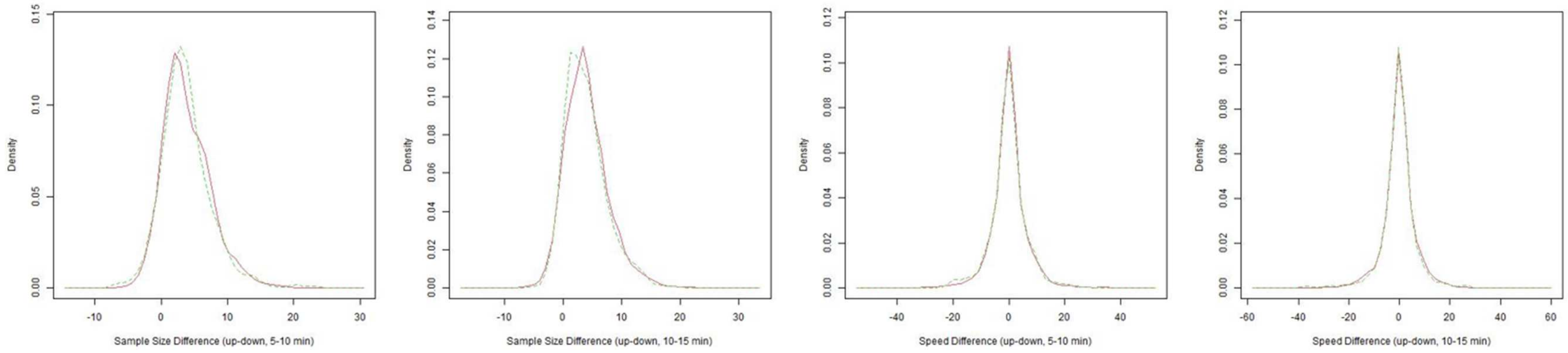


Figure 5-3 KDE Plots for Upstream-Downstream Difference, Control Definition #1

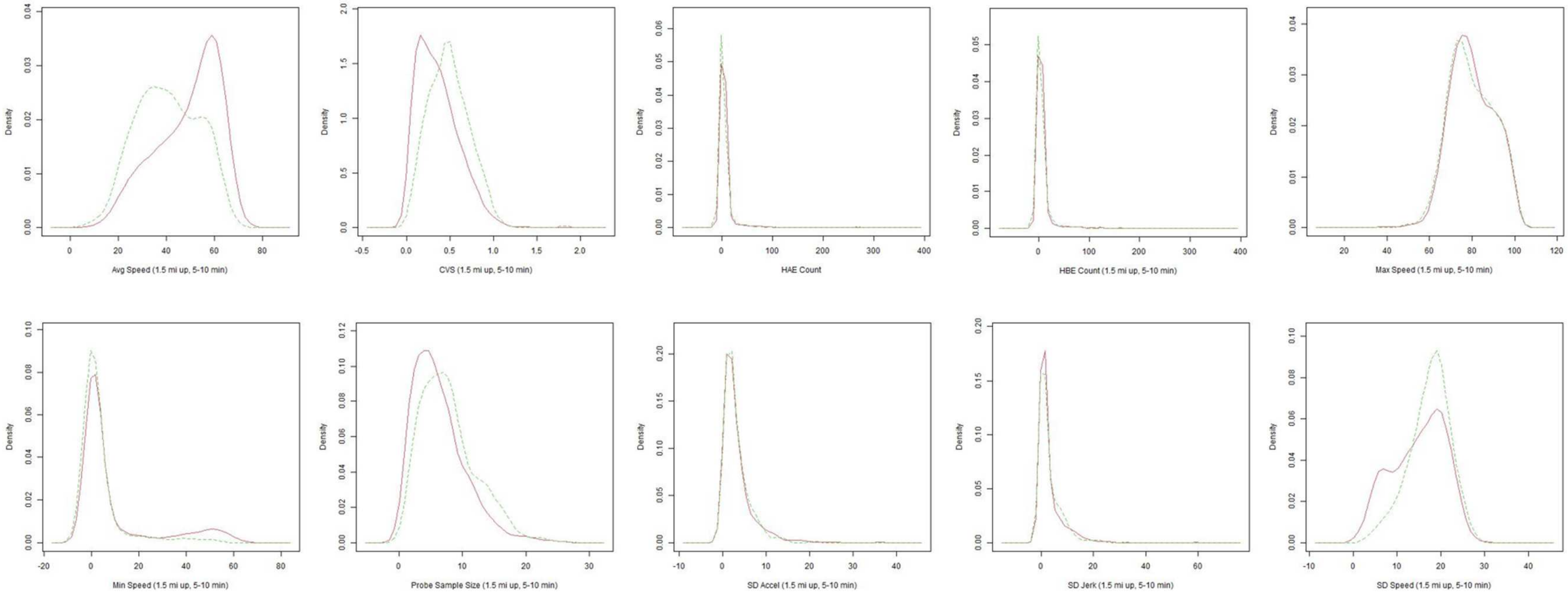


Figure 5-4 KDE Plots for 1.5 mi Up-stream (5-10 minutes), Control Definition #2

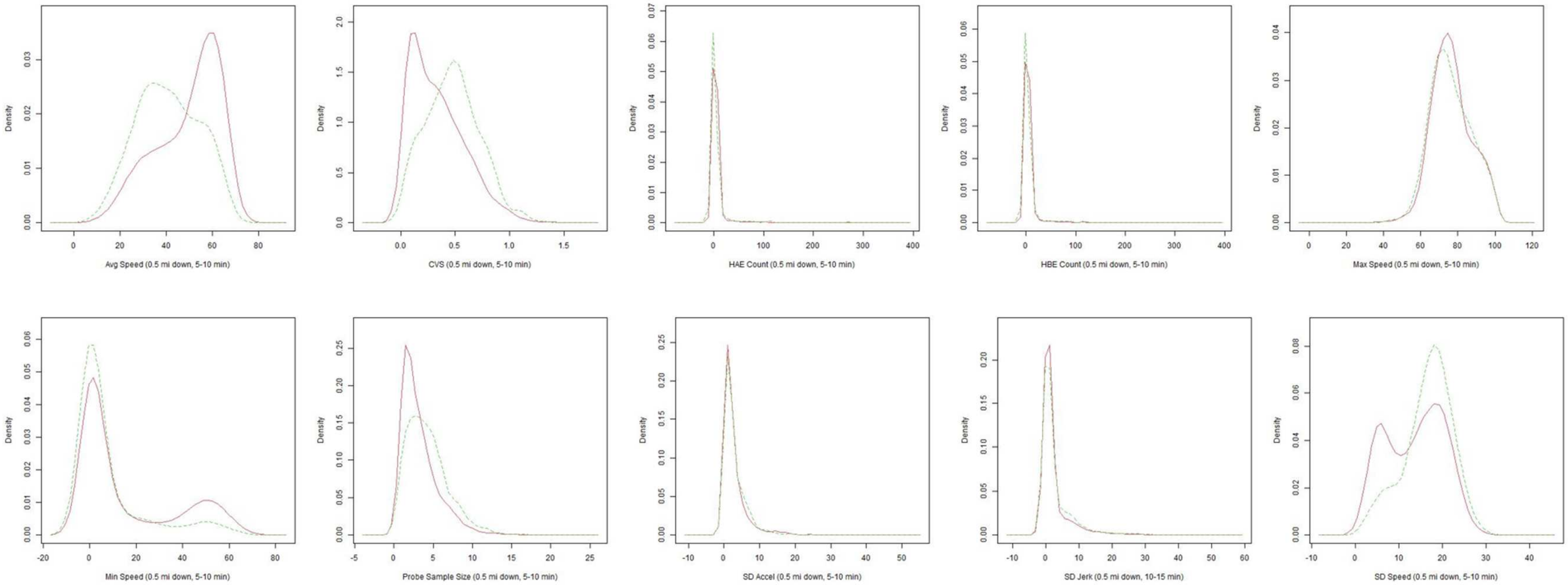


Figure 5-5 KDE Plots for 0.5 mi Down-stream (5-10 minutes), Control Definition #2

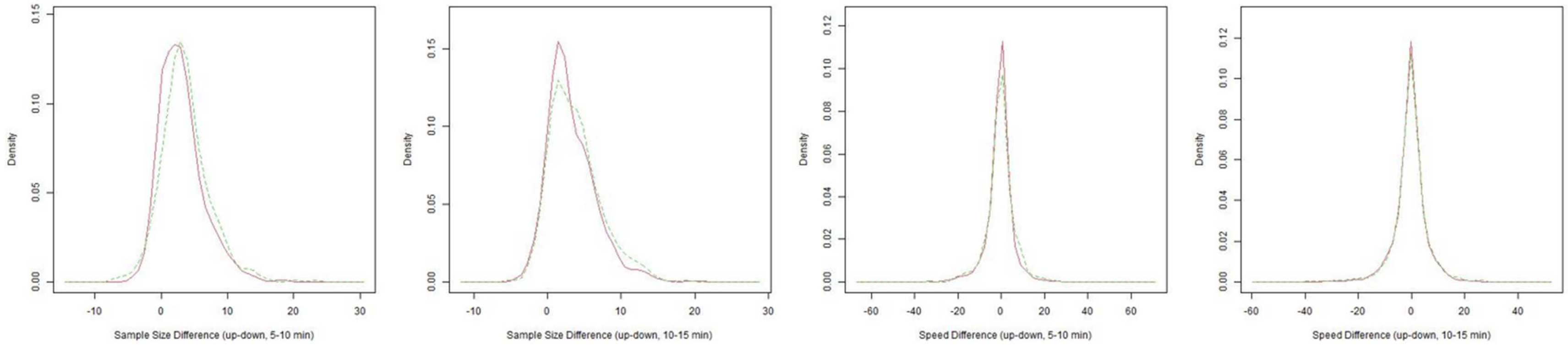


Figure 5-6 KDE Plots for Upstream-Downstream Difference, Control Definition #2

For variables defined based on data collected upstream, the density plots for average speed, coefficient of variation in speed (CVS), and probe sample size show substantial differences between pre-crash and normal conditions, with the differences appearing more prevalent for the datasets based on Control definition #2. This is not entirely surprising since the second definition of controls selects data random points in time at the same location as the crash, unlike the first control definition which selects data at the location of the crash on the same day of the week and at the same time as the crash throughout the period for which data are available; put simply, one would expect more variance in measurements when looking at the random time period data. Similar behavior is also observed for the downstream KDE plots, and minimal to no differences are observed for the plots describing differences between up- and down-stream data.

With this in mind, it seems that variables including CVS, probe sample size, and average speed appear to have significant differences in their distributions when considering pre-crash and normal time periods. Further, this re-affirms the findings in the previous section. With this in mind, these factors were investigated, along with potential interactions, in order to fit additional models that perhaps make more intuitive sense than those in the aforementioned section. For example, the first round of variable selection in the stepwise procedure typically had to choose between many significant and correlated predictors. As such, the variable selected might be correlated to another predictor that makes the results more interpretable, but was wiped out due to multi-collinearity issues. For this model building process, the goal was to conclude with interpretable models for crash risk, whose predictors are significant (at a prescribed level), and all variables in the model had between-variable correlations (Pearson's rho) of less than +/- 0.40 (in an effort to prevent multi-collinearity). In the following, the final models derived from this process are shown.

Similar to the preceding section, each of the models shown presents goodness of fit statistics as well as odds ratios and 95% confidence intervals for said odds ratios. The first two models were the output of the model fitting process for the matched case-control data defined according to control definition #1 (i.e., same location, same day of week, same time).

Table 5-14 shows a model with two significant predictors at the 0.05 level: upstream coefficient of variation in speed (ratio of standard deviation to mean) (5-10 minute period) and downstream sample size (5-10 minute period). The sign and magnitude of the CVS coefficient suggest that as speed variation increases, that is either the mean speed decreases and/or the standard deviation of speed increases, the odds of a crash increase, and multiply by a factor of 5.53. For the downstream sample size predictor, a result indicating increases in volume positively affect the odds of a crash is intuitive and in line with the preceding findings. The downstream sample size was chosen as opposed to the upstream sample size as the upstream value was substantially correlated with CVS, but less so with the downstream value.

Table 5-15 begins with the same base model as Table 5-14, but ultimately, not all predictors in the model were significant at the 0.05 level. However, the model was kept since most terms were significant at the 0.05 level and the interaction term was significant at the 0.10 level. Additionally, this model found upstream to downstream speed difference to be positively correlated with crash risk, with an odds ratio of approximately 1.84. Finally, while the main effect for the rain indicator was not significant at the 0.10 level, the interaction term between the rain and upstream to downstream speed difference variable was found to have a negative coefficient, perhaps indicating that when it is raining and upstream speed is greater than downstream speed, the odds of a crash decrease, suggesting drivers may be more cautious in the rain.

Table 5-14 shows a model with two significant predictors at the 0.05 level, those being upstream coefficient of variation in speed (ratio of standard deviation to mean) (5-10 minute period) and downstream sample size (5-10 minute period). With regard to CVS, the results indicate that as speed variation increases, that is either the mean speed decreases and/or the standard deviation of speed increases, the odds of a crash increase, and multiply by a factor of 8.35. For the downstream sample size predictor, a result indicating increases in volume positively affect the odds of a crash is intuitive and in line with the preceding findings. The downstream sample size was chosen as opposed to the upstream sample size as the upstream value was substantially correlated with CVS, but less so with the downstream value.

Table 5-14 shows a model with two significant predictors at the 0.05 level: upstream coefficient of variation in speed (ratio of standard deviation to mean) (5-10 minute period) and downstream sample size (5-10 minute period). The sign and magnitude of the CVS coefficient suggest that as speed variation increases, that is either the mean speed decreases and/or the standard deviation of speed increases, the odds of a crash increase, and multiply by a factor of 5.53. For the downstream sample size predictor, a result indicating increases in volume positively affect the odds of a crash is intuitive and in line with the preceding findings. The downstream sample size was chosen as opposed to the upstream sample size as the upstream value was substantially correlated with CVS, but less so with the downstream value.

Table 5-15 begins with the same base model as Table 5-14, but ultimately, not all predictors in the model were significant at the 0.05 level. However, the model was kept since most terms were significant at the 0.05 level and the interaction term was significant at the 0.10 level. Additionally, this model found upstream to downstream speed difference to be positively correlated with crash risk, with an odds ratio of approximately 1.84. Finally, while the main effect for the rain indicator was not significant at the 0.10 level, the interaction term between the rain and upstream to downstream speed difference variable was found to have a negative coefficient, perhaps indicating that when it is raining and upstream speed is greater than downstream speed, the odds of a crash decrease, suggesting drivers may be more cautious in the rain.

Table 5-14 Primary Model for 1.5mi Upstream, 0.5mi Downstream Data (Control Def. #1)

<table>
<tr><th colspan="8">All Crashes, Controls=Def. #1 (same time, day of week, and location)</th></tr>
<tr><th>Distance Upstream</th><td colspan="2">1.5 mi</td><th colspan="3">Distance Downstream</th><td colspan="2">0.5 mi</td></tr>
<tr><th colspan="8">MODEL RESULTS</th></tr>
<tr><th>Variable Name</th><th>Coefficient</th><th>exp(coeff)</th><th>SE(coeff)</th><th>Z stat</th><th>p-val</th><th>Lower 95%</th><th>Upper 95%</th></tr>
<tr><td>up_15_510_coeff_var_speed</td><td>1.710</td><td>5.531</td><td>0.357</td><td>4.792</td><td>1.65E-06</td><td>2.748</td><td>11.133</td></tr>
<tr><td>dn_05_510_sample_size</td><td>0.063</td><td>1.065</td><td>0.026</td><td>2.433</td><td>0.010</td><td>1.012</td><td>1.121</td></tr>
<tr><th colspan="8">MODEL FIT</th></tr>
<tr><th>Test Name</th><th colspan="2">Test Statistic</th><th colspan="2">DF</th><th colspan="3">p-val</th></tr>
<tr><td>Likelihood Ratio</td><td colspan="2">34.72</td><td colspan="2">2</td><td colspan="3">3.00E-08</td></tr>
<tr><td>Wald</td><td colspan="2">32.63</td><td colspan="2">2</td><td colspan="3">8.00E-08</td></tr>
<tr><td>Score (logrank)</td><td colspan="2">34.84</td><td colspan="2">2</td><td colspan="3">3.00E-08</td></tr>
<tr><td>AIC</td><td colspan="7">1148.901</td></tr>
<tr><td>n</td><td colspan="7">1734</td></tr>
<tr><td>n_{crash}</td><td colspan="7">443</td></tr>
</table>

Table 5-15 Secondary Model for 1.5mi Upstream, 0.5mi Downstream Data (Control Def. #1)

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.5 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_coeff_var_speed	1.792	5.999	0.368	4.870	1.12E-06	2.917	12.337
dn_05_510_sample_size	0.062	1.064	0.026	2.391	0.017	1.011	1.120
up_down_speed_diff_510	0.226	1.254	0.123	1.843	0.065	0.986	1.595
rain	-0.326	0.722	0.591	-0.551	0.582	0.227	2.300
up_down_speed_diff_510*rain	-0.215	0.807	0.121	-1.772	0.076	0.636	1.023
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	39.38		5		2.00E-07		
Wald	36.07		5		9.00E-07		
Score (logrank)	38.60		5		3.00E-07		
AIC	1150.242						
n	1734						
n_{crash}	443						

Table 5-16 shows a model with two significant predictors at the 0.05 level, those being upstream coefficient of variation in speed (5-10 minute period) and downstream sample size (5-10 minute period). With regard to CVS, the results indicate that as speed variation increases, the odds of a crash increase, and multiply by a factor of 8.35. For the downstream sample size predictor, a result indicating increases in volume positively affect the odds of a crash is intuitive and in line with the preceding findings.

Table 5-17 begins with the same base model as that in Table 5-16, but ultimately, not all predictors in the model were significant at the 0.05 level. Again, the model was evaluated since most terms were significant at the 0.05 level and the interaction term was significant at the 0.10 level. Like the preceding model with the interaction term, this model found upstream to downstream speed difference to be positively correlated with crash risk, with an odds ratio of approximately 1.22. Finally, while the main effect for the rain indicator was not significant at the

0.10 level, the interaction term between the rain and upstream to downstream speed difference variable was found to have a negative coefficient.

Table 5-16 Primary Model for 1.5mi Upstream, 0.5mi Downstream Data (Control Def. #2)

All Crashes, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.5 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_coeff_var_speed	2.122	8.346	0.260	8.150	3.64E-16	5.010	13.901
dn_05_510_sample_size	0.111	1.118	0.024	4.561	5.09E-06	1.065	1.172
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	130.2		2		<2e-16		
Wald	115.1		2		<2e-16		
Score (logrank)	130.5		2		<2e-16		
AIC	1204.329						
n	2083						
n_{crash}	451						

Table 5-17 Secondary Model for 1.5mi Upstream, 0.5mi Downstream Data (Control Def. #2)

All Crashes, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.5 mi		**Distance Downstream**			0.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_coeff_var_speed	2.199	9.019	0.265	8.292	<2e-16	5.363	15.168
dn_05_510_sample_size	0.103	1.108	0.025	4.171	3.03E-05	1.056	1.163
up_down_speed_diff_510	0.198	1.218	0.099	2.002	0.045	1.004	1.479
rain	-0.342	0.710	0.559	-0.612	0.541	0.237	2.125
up_down_speed_diff_510*rain	-0.178	0.837	0.097	-1.830	0.067	0.691	1.013
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	138.1		5		<2e-16		
Wald	120.6		5		<2e-16		
Score (logrank)	138.3		5		<2e-16		
AIC	1202.385						
n	2083						
n_{crash}	451						

5.4 CONCLUSION AND EXTENSIONS

Through the real-time crash prediction modeling portion of this study, several key conclusions must be addressed. First of all, it was demonstrated that probe vehicle trajectory data produced results similar to preceding studies in terms of significant variables (e.g., volume, average speed, coefficient of variation in speed, etc.), time periods in which variables are significant (e.g., 5-10 minutes before the time of the crash) as well as sign and magnitude of coefficients (Hossain et al. 2019; Roshandel et al. 2015). Another thing to note here, and in other studies as well including many described in the literature review, is that the model developed herein and most of those in other studies have a rather small number of variables in the final model. In some senses, this is not surprising as that in many cases, the effects of categorical variable cannot be estimated in case-control designs (as discussed previously). Further, when one is looking to model crash likelihood based on traffic flow related variables, many of these variables are correlated (e.g., consider the linear model for a speed-density curve) with each other, as well as exhibiting auto-correlation over time and space, hence many get removed from consideration in the final model in order to prevent multi-collinearity. While many of these models present seemingly reasonable results in terms of variables selected and their impacts, it is quite clear that said problem is suffering from omitted variable bias, especially with respect to consideration of human factors.

Chapter 6. DEVELOPMENT OF CONFIDENCE AND PREDICTION INTERVALS FOR MIXED-POISSON REGRESSION MODELS

6.1 Problem Statement and Motivation

For full transparency, the author notes that the work in this section is derived from/based upon the work presented in Ash et al. (2019). In the study, the primary goal was to show additional applications building on the key work of Wood (2005) and thus develop the CIs and PIs for several other mixed-Poisson regression models that many researchers have previously investigated and that are used in practice by safety analysts. For comparison purposes and ensuring clear alignment with Wood (2005), his notation/naming of the PI in reference to y and m (dependent variables) and the CI in reference to μ (a model parameter) will also be used in this dissertation. To be explicit, this work will begin by reviewing how to derive the PIs for y and m and the CI for μ, respectively, for the NB model based Wood (2005); then, it will present derivations for the CIs and PIs for the same three values (where m in these cases is generalized to be the Poisson parameter following a given mixture distribution, and such parameter is also called the safety) for the Poisson-Inverse-Gaussian, Sichel, Poisson-Weibull, and Poisson-Lognormal regression models. Following the methodology, the mixed-Poisson regression models of interest will be estimated in case study based on an animal-vehicle collision dataset a case study making use of an animal-vehicle collision dataset will be conducted. After the model development, the methodology will be demonstrated by estimating and plotting the associated PIs and CIs for y, m, and μ under each different model type. Then, a discussion and comparison of the PIs and CIs will be provided based on the case study and key results will be noted. Put simply, the key contribution of this work is developing a means to express/quantify uncertainty for estimates from safety-modeling efforts (notably those from mixed-Poisson regression models), as opposed to just having access to point estimates.

6.2 Derivation of Confidence and Prediction Models

To begin, the confidence and prediction intervals for each type of mixed-Poisson model evaluated in this work are derived. Before this part of the methodology, though, a quick discussion of requisite information on mixed-Poisson models is given. One may view a mixed-Poisson model as a sort of hierarchy with three levels. At the bottom/first level of the hierarchy there is the mean response (μ_i), also referred to as the Poisson mean. The mean response follows a normal distribution, $N(\mu_0, \sigma_0^2)$. At the next level up in the hierarchy is the Poisson parameter (m_i), also called the safety, which when conditioned on the Poisson mean can be shown to follow the mixture distribution that is being applied. Lastly, at the third/top level of the hierarchy is the predicted response (y_i), the crash frequency at site i. When conditioned on the Poisson parameter (m_i), the predicted response can be shown to follow a Poisson distribution.

6.2.1 *Mixed-Poisson Models and Formulation*

There are two main points to the definition of a mixed-Poisson model. First, the count being modeled (the number of crashes y_i) follows a Poisson distribution, conditional on the Poisson parameter λ_i (Cameron and Trivedi 2013). To stay in alignment with the terminology and notation of Wood (2005), this study will refer to the Poisson parameter λ_i as the "safety" and denote it as m_i as shown in the following equation:

$$f(y_i|m_i) = \frac{\exp(-m_i)*m_i^{y_i}}{y_i!}, y_i = 0,1,2 \ldots \qquad (6\text{-}1)$$

Second, the Poisson parameter (safety), m_i, has a multiplicative error term that follows the specified mixture distribution (e.g., gamma, inverse Gaussian, lognormal, etc.) and is represented in the conditional mean (the safety conditioned on the mean response μ_i) (Cameron and Trivedi 2013). Readers may notice here that without the error term, the formulation of the safety reduces to that of the mean response (μ_i), also known as the Poisson mean. The Poisson parameter (m_i) is defined in the following equation.

$$
\begin{aligned}
m_i &= \exp(\beta_0 + \textstyle\sum_{j=1}^{K} x'_{ij} * \beta_j + \varepsilon_i) \\
&= \exp(\beta_0 + \varepsilon_i) * \exp\left(\textstyle\sum_{j=1}^{K} x_{ij} * \beta_j\right) \\
&= \exp\left(\beta_0 + \textstyle\sum_{j=1}^{K} x_{ij} * \beta_j\right) * \exp(\varepsilon_i) \\
&= \mu_i * \nu_i \qquad (6\text{-}2)
\end{aligned}
$$

Where,

i = site index;

β_j = jth regression coefficient;

x_{ij} = jth covariate for site i; and

ε_i = error term such that $\exp(\varepsilon_i)$, itself referred to as ν_i, follows the chosen mixture distribution.

As previously noted, $Y|m_i \sim Poisson(m_i)$. The marginal distribution for Y (outlined in the equation as follows) is developed by integrating out the error term ν_i, and in this equation, $h(\nu_i)$ is the mixture distribution (Cameron and Trivedi 2013):

$$
\begin{aligned}
f(y_i|\mu_i) &= \int_0^\infty g(y_i|\mu_i, \nu_i) * h(\nu_i)\ d\nu_i \\
&= E_\nu[g(y_i|\mu_i, \nu_i)] \qquad (6\text{-}3)
\end{aligned}
$$

Then by noting the key equality $m_i = \mu_i \nu_i$, one can show that the Poisson parameter m_i clearly follows the mixture distribution (just like ν_i) (Cameron and Trivedi 2013).

6.2.2 *Parametrizations of the Mixture Distributions*

For this study, a total of five mixed-Poisson models were examined. These models, their corresponding mixture distributions for m_i and ν_i, and the parameterizations of each mixture distribution are outlined in the following sections. Each section is named in alignment with the model type described and the mixture distribution is shown in parentheses. Lastly, the subscript i is left out without any loss of generality.

6.2.2.1 Negative Binomial [NB] Model (Gamma)

When the choice of mixture distribution used in the mixed-Poisson model is the gamma distribution, the ultimate model is the NB model. Specifically, $\nu \sim Gamma(\delta, \phi)$; that said, if one wants to be able to properly identify the intercept in the regression model, one must have $E[\nu] = 1$. In order for this equality to result, one can set $\delta = \phi$, and thus a one-parameter gamma distribution is obtained. Finally, it follows that $Var[\nu] = 1/\phi$, that is, $Var[\nu] = \alpha$, and further that $m|\phi, \mu \sim Gamma\left(\phi, \frac{\phi}{\mu}\right)$ (Cameron and Trivedi 2013).

6.2.2.2 Poisson-Inverse-Gaussian [PIG] Model (Inverse Gaussian)

The PIG model is obtained when the inverse Gaussian (IG) distribution is used for the mixture distribution. Specifically, $\nu \sim IG(\mu_{IG}, \lambda)$, where the subscript "IG" is needed to distinguish μ_{IG} (the mean of the IG distribution) from the Poisson mean (μ). Just as was true for the NB model. the intercept identification condition necessitates that $E[\nu] = 1$. If $\mu_{IG} = 1$, then $E[\nu] = 1$, and also $Var[\nu] = 1/\lambda$ (Rigby and Stasinopolous 2009).

6.2.2.3 Sichel [SI] (Generalized Inverse Gaussian)

When the chosen mixture distribution is the generalized inverse Gaussian (GIG) distribution, the resultant model is the Sichel model. Here, $\nu \sim GIG(\mu_{GIG}, \sigma_{GIG}, \nu_{GIG})$. Again, $E[\nu] = 1$ (in order to meet the intercept identification condition), and the variance of ν is expressed as shown in the following (Rigby et al. 2008; Rigby and Stasinopolous 2009):

$$Var[\nu] = \frac{2\sigma_{GIG}(\nu_{GIG}+1)}{c} + \frac{1}{c^2} - 1 \tag{6-4}$$

Where,

$c = R_{\nu_{GIG}}\left(\frac{1}{\sigma_{GIG}}\right)$;

$R_\lambda(t) = K_{\lambda+1}(t)/K_\lambda(t)$; and

$K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp\left[-\frac{1}{2}t(x+x^{-1})\right] dx$ (where, $K_\lambda(t)$ is the modified Bessel function of the third kind).

6.2.2.4 Poisson-Lognormal [PLN] (Lognormal)

A selection of the mixture distribution as the lognormal distribution, leads to the Poisson-Lognormal model. In this case, $\nu \sim \log N(d, \sigma_{LN}^2)$, and the mean of the lognormal distribution is shown as follows (Connors et al. 2013):

$$E[\nu] = \exp\left(d + \frac{\sigma_{LN}^2}{2}\right) \tag{6-5}$$

In order to satisfy the intercept identification condition, $E[\nu] = 1$, one can require that $d = -\sigma_{LN}^2/2$. The variance of ν can then be calculated by substituting the preceding value for d into the following equation for $Var[\nu]$:

$$Var[\upsilon] = \left(e^{\sigma_{LN}^2} - 1\right)e^{2d+\sigma_{LN}^2}$$

$$= e^{\sigma_{LN}^2} - 1 \tag{6-6}$$

6.2.2.5 Poisson-Weibull [PW] (Weibull)

When ν, the error term, is chosen to follow the Weibull distribution, the resultant model is the Poisson-Weibull model; in this case, $\nu \sim Weibull(\mu_W, \sigma_W)$. The mean of the Weibull distribution is shown in the following (Cheng et al. 2013; Rigby and Stasinopolous 2009):

$$E[\nu] = \frac{1}{\mu_W^{1/\sigma_W}} \Gamma\left(\frac{1}{\sigma_W} + 1\right) \tag{6-7}$$

To establish the intercept identification condition of $E[\nu] = 1$, the following equality must be established:

$$\mu_W = \left(\Gamma\left(\frac{1}{\sigma_W} + 1\right)\right)^{\sigma_W} \tag{6-8}$$

Having $E[\nu] = 1$, the variance of ν can then be calculated by substituting the preceding value for μ_W into the equation for $Var[\nu]$ as follows.

$$Var[\nu] = \frac{1}{\mu_W^{2/\sigma_W}} \left[\Gamma\left(\frac{2}{\sigma_W} + 1\right) - \left(\Gamma\left(\frac{1}{\sigma_W} + 1\right)\right)^2\right]$$

$$= \frac{\Gamma\left(\frac{2}{\sigma_W}+1\right)}{\left(\Gamma\left(\frac{1}{\sigma_W}+1\right)\right)^2} - 1 \tag{6-9}$$

6.2.3 *Derivation of Confidence Intervals for Poisson mean (True mean Crash Frequency) (μ)*

To begin, a generalized linear model (GLM) for crash prediction will be examined, For the model, each site of interest is a road segment, and the model is of form shown as shown in Equations (6-10 a) and (6-10 b).

$$\eta = \log\left(\frac{\mu}{L*t}\right) = \log(\beta_0) + \sum_{i=1}^{n} x_{ij} * \beta_j \tag{6-10 a}$$

$$\eta = \log\left(\frac{\mu}{L*t}\right) = \beta'_0 + \sum_{i=1}^{n} x_{ij} * \beta_j \tag{6-10 b}$$

Where,

η = linear predictor;

β_j = jth regression coefficient;

x_{ij} = jth predictor for segment (site);

L = segment length (in miles); and

t = time period over which crash data was collected.

An offset term can be calculated by considering the product of segment length and the duration over which crash data was collected for each site, and this results in a reformulation for Equation 6-10 describing the regression as shown in the following:

$$\eta = \log(\mu) = \log(\beta_0) + \sum_{i=1}^{n} x_{ij} * \beta_j + \log(L * t) \tag{6-11}$$

Under Equation 6-11, if x_1 is the traffic volume (F), then:

$$\mu = \beta_0 F^{\beta_1}(L * t) * \exp\left(\sum_{i=2}^{n} x_{ij} * \beta_j\right)$$

For the GLM, the estimators of the regression coefficients, $\hat{\beta}_j$, are assumed to follow a multivariate normal distribution, $[\widehat{\beta'}_0, \dots, \hat{\beta}_n]' \sim N([\beta'_0, \dots, \beta_n]', \mathbf{\Sigma})$. Then, from Equation 6-11, one can see that $\mu = \exp(\eta)$. As noted in the preceding, the subscript i for the values of η and μ at every site i are omitted without any loss of generality. With this in mind, one can then use this information to derive an approximate (1-α)% confidence interval (CI) for the Poisson mean (also known as the true mean crash count), μ, as follows (where $Z_{1-\alpha/2}$ is the critical value for the $1-\alpha/2$ quantile of the standard normal distribution) (Wood 2005):

$$\exp(\hat{\eta} \pm Z_{1-\alpha/2}\sqrt{Var(\hat{\eta})})$$

$$= \exp(\hat{\eta}) * \exp(\pm Z_{1-\alpha/2}\sqrt{Var(\hat{\eta})})$$

$$= \hat{\mu} * \exp(\pm Z_{1-\alpha/2}\sqrt{Var(\hat{\eta})})$$

$$= \left[\frac{\hat{\mu}}{\exp\left(Z_{1-\alpha/2}\sqrt{Var(\hat{\eta})}\right)}, \hat{\mu} * \exp\left(Z_{1-\alpha/2}\sqrt{Var(\hat{\eta})}\right)\right] \tag{6-12}$$

Ultimately, no matter the choice of mixture distribution applied, the approximate $(1-\alpha)\%$ CI for the Poisson mean (i.e., the true mean crash count), μ, is as shown in Equation 6-12. Interested readers are referred to Wood (2005) for more detailed steps as to how to calculate linear predictor's variance.

6.2.4 *Derivation of Prediction Intervals for Poisson Parameter (m)*

Next, the derivation of an approximate $(1-\alpha)\%$ confidence interval for the Poisson parameter, also known as the safety, m is provided, again following the procedure originally outlined in Wood (2005). Prior to the derivation, a useful point for later calculations involves considering that though the distribution of the estimator for the Poisson mean ($\hat{\mu}$) is technically lognormal, it can be approximated as normal (Wood 2005). Hence, $\hat{\mu} \sim N(\mu_0 = \mu, \sigma_0^2 = \mu^2 Var(\hat{\eta}))$.

Based on Wood (2005), one can obtain the formulation for an approximate $(1-\alpha)\%$ PI for the Poisson parameter, also known as the safety, m that is presented in Equation 6-13.

$$\hat{\mu} \pm Z_{1-\alpha/2} * \sqrt{Var(m)} \tag{6-13}$$

From a mathematical/calculation perspective, it is indeed possible for the lower bound of the prediction interval for Equation 6-13 to be negative, but intuitively, this does not make sense as negative values of m are not sensible. Thus, one can reformulate the PI for m in Equation 6-13 as follows:

$$\left[\max\left\{0, \hat{\mu} - Z_{1-\frac{\alpha}{2}} * \sqrt{Var(m)}\right\}, \hat{\mu} + Z_{1-\frac{\alpha}{2}} * \sqrt{Var(m)}\right] \tag{6-14}$$

The variance of m is formulated as follows:

$$Var(m) = Var(\mu\nu)$$

$$= E(\mu^2\nu^2) - E(\mu\nu)^2$$

$$= E(\mu^2)E(\upsilon^2) - E(\mu)^2E(\upsilon)^2 \qquad [by\ independence\ of\ \mu\ and\ \upsilon]$$

$$= [Var(\mu) + E(\mu)^2] * [Var(\upsilon) + E(\upsilon)^2] - E(\mu)^2E(\upsilon)^2 \tag{6-15}$$

Table 6-1 shows the full set of derived expressions for the variance of m for each of the mixture distributions under analysis in this study.

Table 6-1 Variance of m for Mixture Distributions

Mixture Distribution	$\boldsymbol{Var(m)}$
Gamma	$\alpha * (\sigma_0^2 + \mu_0^2) + \sigma_0^2\ \left[Note: \alpha = \frac{1}{\varphi}\right]$
Inverse Gaussian	$\frac{1}{\lambda} * (\sigma_0^2 + \mu_0^2) + \sigma_0^2$
Generalized Inverse Gaussian	$[\sigma_0^2 + \mu_0^2] * \left(\frac{2\sigma_{GIG}(\nu_{GIG}+1)}{c} + \frac{1}{c^2}\right) - \mu_0^2$
Lognormal	$e^{\sigma_{LN}^2} * [\sigma_0^2 + \mu_0^2] - \mu_0^2$
Weibull	$[\sigma_0^2 + \mu_0^2] * \left[\frac{\Gamma\left(\frac{2}{\sigma_W}+1\right)}{\left(\Gamma\left(\frac{1}{\sigma_W}+1\right)\right)^2}\right] - \mu_0^2$

Now that the formulations for $Var(m)$ have been established, and if one recalls that the distribution of $\hat{\mu}$ can be approximated as normal (which provides the useful algebraic substitution $\sigma_0^2 = \mu^2 Var(\hat{\eta})$), the derived PIs for m for each type of mixed-Poisson model considered in this study can be formulated and they are shown in Table 6-2. For simplicity, only 95% PIs are shown Table 6-2.

Table 6-2 95% Prediction Intervals for m

Model	95% PI for m
Negative Binomial (NB)	$\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2[\hat{\alpha}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})]}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2[\hat{\alpha}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})]}\right]$
Poisson-Inverse-Gaussian (PIG)	$\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2\left[\frac{1}{\hat{\lambda}}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})\right]}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2\left[\frac{1}{\hat{\lambda}}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})\right]}\right]$
Sichel (SI)	$\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2\left\{[Var(\hat{\eta}) + 1] * \left(\frac{2\hat{\sigma}_{GIG}(\hat{\nu}_{GIG}+1)}{c} + \frac{1}{c^2}\right) - 1\right\}}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2\left\{[Var(\hat{\eta}) + 1] * \left(\frac{2\hat{\sigma}_{GIG}(\hat{\nu}_{GIG}+1)}{c} + \frac{1}{c^2}\right) - 1\right\}}\right]$
Poisson-Lognormal (PLN)	$\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2[e^{\hat{\sigma}_{LN}^2}(Var(\hat{\eta}) + 1) - 1]}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2[e^{\hat{\sigma}_{LN}^2}(Var(\hat{\eta}) + 1) - 1]}\right]$
Poisson-Weibull (PW)	$\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2\left([Var(\hat{\eta}) + 1] * \left[\frac{\Gamma\left(\frac{2}{\hat{\sigma}_W} + 1\right)}{\left(\Gamma\left(\frac{1}{\hat{\sigma}_W} + 1\right)\right)^2}\right] - 1\right)}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2\left([Var(\hat{\eta}) + 1] * \left[\frac{\Gamma\left(\frac{2}{\hat{\sigma}_W} + 1\right)}{\left(\Gamma\left(\frac{1}{\hat{\sigma}_W} + 1\right)\right)^2}\right] - 1\right)}\right]$

6.2.5 *Derivation of Prediction Intervals for Predicted Crash Count (y)*

The last type of prediction interval of interest in this study for a mixed-Poisson model is the interval for the predicted crash count (y) at a new site. Here, the PI is formulated based upon Chebyshev's inequality, and one must further assume the following: (1) The lower bound for y is zero (this is done in order to make a more conservative PI and since it follows the practice of Wood (2005)); (2) y must take on integer values only (Wood 2005). Thus, a $(1-\alpha)\%$ PI for y is shown in Equation 6-16, and in the formulation the floor of the upper bound is applied to enforce the constraint of ensuring an integer-value. Consider the following, where one wants to estimate a 95% PI for y; in such case, the expression under the first radical in the equation would come out to 19.

$$\left[0, \left\lfloor \hat{\mu} + \sqrt{\alpha^{-1}-1}\sqrt{Var(y)} \right\rfloor\right] \tag{6-16}$$

The variance of Y is evaluated as follows:

$$Var(Y) = E\{Var(Y|M)\} + Var\{E(Y|M)\}$$

$$= E(M) + Var(M)$$

$$= E(\mu v) + Var(M)$$

$$= E(\mu) * E(v) + Var(M)$$

$$= \mu_0 + Var(M) \tag{6-17}$$

Thus, a $(1-\alpha)\%$ PI for Y can be re-formulated as is shown in Equation 6-18.

$$\left[0, \left\lfloor \hat{\mu} + \sqrt{\alpha^{-1}-1}\sqrt{\hat{\mu} + Var(m)} \right\rfloor\right] \tag{6-18}$$

Then, by applying the expressions for $Var(m)$ as outlined in Equation 6-15, one can estimate the 95% PIs for Y for each of the five mixed-Poisson models, and these results are shown in Table 6-3.

Table 6-3 95% Predictions Intervals for y

Model	95% PI for y
Negative Binomial (NB)	$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2[\hat{\alpha}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})]} \right\rfloor\right]$
Poisson-Inverse-Gaussian (PIG)	$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2 \left[\frac{1}{\hat{\lambda}}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})\right]} \right\rfloor\right]$
Sichel (SI)	$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2 \left\{[Var(\hat{\eta}) + 1] * \left(\frac{2\hat{\sigma}_{GIG}(\hat{\nu}+1)}{c} + \frac{1}{c^2}\right) - 1\right\}} \right\rfloor\right]$
Poisson-Lognormal (PLN)	$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2[e^{\hat{\sigma}_{LN}^2}(Var(\hat{\eta}) + 1) - 1]} \right\rfloor\right]$
Poisson-Weibull (PW)	$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2 \left([Var(\hat{\eta}) + 1] * \left[\frac{\Gamma\left(\frac{2}{\hat{\sigma}}+1\right)}{\left(\Gamma\left(\frac{1}{\hat{\sigma}}+1\right)\right)^2}\right] - 1\right)} \right\rfloor\right]$

Interested readers can find the full derivations for each of the aforementioned confidence and prediction intervals in the appendix of this dissertation.

6.3 CASE STUDY

In the following section, a case study is presented where the mixed-Poisson models are estimated based upon a crash dataset from Washington State. After the models are estimated, the corresponding confidence and prediction intervals for the different components of the mixed-Poisson model hierarchy are estimated and displayed graphically. In this case study, the data applied is the animal-vehicle collision data that is further described in the preceding Section 3.2.2 of this dissertation.

6.3.1 *Model Development*

For the case study, five different mixed-Poisson models were estimated based upon the aforementioned animal-vehicle collision dataset. For each model, the variables included were those found to be significant at the $\alpha = 0.05$ level for the Negative Binomial model (in order to allow for comparison), and all models also had an offset term. The Negative Binomial, Poisson-Inverse-Gaussian, and Sichel models were estimated via a maximum likelihood (ML) estimation approach in the GAMLSS package within the R statistical software program (Rigby and Stasinopoulos 2005). Since the Poisson-Lognormal and Poisson-Weibull likelihood functions do not have a closed form, these models had were estimated via a Bayesian approach in the WinBUGS software package (Lunn et al. 2000) (the estimation used 20,000 iterations following 5000 iterations for the burn in process). For each model, the estimated model parameters, associated standard errors (SE) (and posterior values for the Bayesian models), p-values, and goodness-of-fit statistics such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are presented in Table 6-4 and Table 6-5.

Table 6-4 Model Results Estimated with ML Approach

	NB			PIG			SI		
	Estimate	SE	p-val	Estimate	SE	p-val	Estimate	SE	p-val
Intercept log(β_0)	-10.36	1.11	< 2.00E-16	-10.38	1.13	< 2.00E-16	-10.42	1.12	< 2.00E-16
log(AADT) β_1	0.55	0.08	1.45E-10	0.59	0.09	2.12E-11	0.56	0.09	2.00E-10
Access β_2	-1.12	0.30	1.98E-04	-1.02	0.30	5.92E-04	-1.11	0.30	2.45E-04
Spd_limt β_3	0.09	0.02	1.15E-07	0.08	0.02	5.49E-07	0.09	0.02	1.22E-07
Nolanes β_4	-0.40	0.11	1.65E-04	-0.39	0.12	8.22E-04	-0.41	0.11	1.71E-04
Lshlw β_5	0.12	0.03	4.07E-06	0.11	0.03	1.07E-04	0.12	0.03	7.06E-06
White β_6	1.39	0.13	< 2.00E-16	1.59	0.13	< 2.00E-16	1.41	0.13	< 2.00E-16
Elk β_7	0.37	0.13	3.54E-03	0.50	0.14	2.43E-04	0.38	0.13	3.47E-03
Distribution Parameter(s)	$\alpha = 1/\varphi = 1.85$			$\lambda = 0.106$			$\nu_{GIG} = 0.4716$, $\sigma_{GIG} = 271$		
Global Deviance	2986.46			3010.00			2986.27		
AIC	3004.46			3028.00			3006.27		
BIC	3046.06			3069.60			3052.49		

Table 6-5 Model Results Estimated with Bayesian Approach

	PLN					PW				
	Mean	SE	2.5%	50%	97.5%	Mean	SE	2.5%	50%	97.5%
Intercept log(β_0)	-10.28	0.60	-11.20	-10.38	-8.87	-9.08	1.37	-10.75	-9.42	-6.87
log(AADT) β_1	0.58	0.06	0.44	0.59	0.67	0.53	0.08	0.38	0.53	0.66
Access β_2	-1.05	0.32	-1.69	-1.04	-0.45	-0.79	0.31	-1.43	-0.79	-0.21
Spd_limt β_3	0.08	0.01	0.06	0.08	0.10	0.07	0.02	0.04	0.07	0.10
Nolanes β_4	-0.35	0.14	-0.57	-0.35	-0.10	-0.41	0.10	-0.61	-0.41	-0.22
Lshlw β_5	0.12	0.03	0.07	0.13	0.17	0.13	0.03	0.08	0.13	0.19
White β_6	1.69	0.15	1.40	1.69	2.00	1.51	0.13	1.25	1.51	1.74
Elk β_7	0.60	0.14	0.32	0.60	0.88	0.39	0.14	0.13	0.39	0.66
Distribution Parameter(s)	$\sigma_{LN} = 1.35$					$\sigma_W = 0.70$				

6.3.2 *Confidence and Prediction Intervals*

As a means to compare and contrast the prediction intervals for y and m, in addition to the confidence intervals for μ for each of the five mixed-Poisson models examined in this study, plots were made created so the intervals for each model could be shown graphically (Figure 6-1 through Figure 6-5). In every plot the results shown apply the model based on the coefficients shown in Table 6-6. In each case, AADT is varied (in increments of 50) between 0 and 120,000 vehicles per day (this was determined to be the approximate range in the animal-vehicle collision dataset). Finally, the segment length and time period were fixed and considered to be one mile and 5 years, respectively, in order to calculate the offset term. The remaining variables were set to constant values based upon the values that were the most common value of each variable, respectively, in the dataset. Thus, the intervals show the numbers of animal-vehicle collisions within the five-year period, for a one-mile road segment with AADT values that vary, and marginalizing across all other variables.

Table 6-6 Default Values of Variables used in Models for Interval Construction

Variable	Value
Access β_2	0
Spd_limt β_3	60
Nolanes β_4	2
Lshlw β_5	8
White β_6	0
Elk β_7	0

To begin, the 95% CI for the Poisson mean (μ) is examined as follows. From Table 6-4 and Table 6-5 one can see that no matter what model is considered, estimates for the model regression coefficients are indeed rather similar. As such, it is not surprising that the estimates for the Poisson means across models were rather similar. For these means, the maximum values were found to range between 17.97 for the Poisson-Weibull model to as much as 21.53 for the Sichel model, where AADT=120,000 (the maximum value considered). As seen in Figure 6-1 through Figure 6-5, both the lower and upper bounds of the 95% CI for the Poisson mean take on similar values, respectively, for each of the different models considered. For AADT=120,000, the smallest/tightest interval around the estimate of μ was obtained for the Poisson-Lognormal model ([11.94, 31.25], width=19.31), and on the other end, the largest interval was for the Sichel model

([11.62, 39.87], width=28.25). Since the true value of the Poisson mean is unknown for all sites, one cannot necessarily develop conclusions as to if the narrowest interval is "best" or not.

Figure 6-1 through Figure 6-5 show the plots of the 95% PIs for the safety, m. In each case, the lower bound values are technically not shown since in every model, no matter the value of AADT considered, the lower bounds were found to be zero. Technically, the models produced negative values for the calculated lower bounds, but as noted in Equation 6-14, any negative value for the safety (m) is not sensible and as such, the lowest reasonable value must be zero. In the case of the Poisson-Inverse-Gaussian model, the width of the interval for an AADT value of 120,000 was 149.10, and this was found to be the maximum value of m estimated from any of the 95% CIs. On the other hand, the lowest value for any upper bound for the 95% CIs for m at 120,000 AADT, when considering all models, was 72.89 and this was obtained from the Poisson-Weibull model. Again, like in the case for the Poisson mean, the true value of m is not known, and thus no conclusions on which interval is tightest, while still capturing the true parameter value can be developed.

No matter what model examined, the lower bound for the 95% prediction intervals for the predicted response at a new site (y) was always found zero, and thus, this interval does not appear in any of the plots. Further, one may observe how the upper bounds for the PIs for y are much greater (specifically, 1.92 to 2.06 times greater at 120,000 ADT) than the respective upper bounds for the 95% PIs for m. In addition to having the largest values, the curves for the PIs for y are also much less smooth than the curves for the CIs for μ and PIs for m. In these cases, the curve appears as a step-function due to the use of the floor function in the formulation of the upper bound of the PI as is shown in Equation 6-18. Again, this is due to the fact that one assumes that the number of crashes predicted for a new site should take on an integer value. In the case of upper bounds for the PIs for y, they ranged from a maximum of 307 for the Poisson-Inverse-Gaussian model to a minimum of 141 for the Poisson-Weibull model. For the Negative Binomial, Sichel, and Poisson-Weibull models the calculated upper bounds for the PIs for y appeared to be more similar in value than compared to those for the other models.

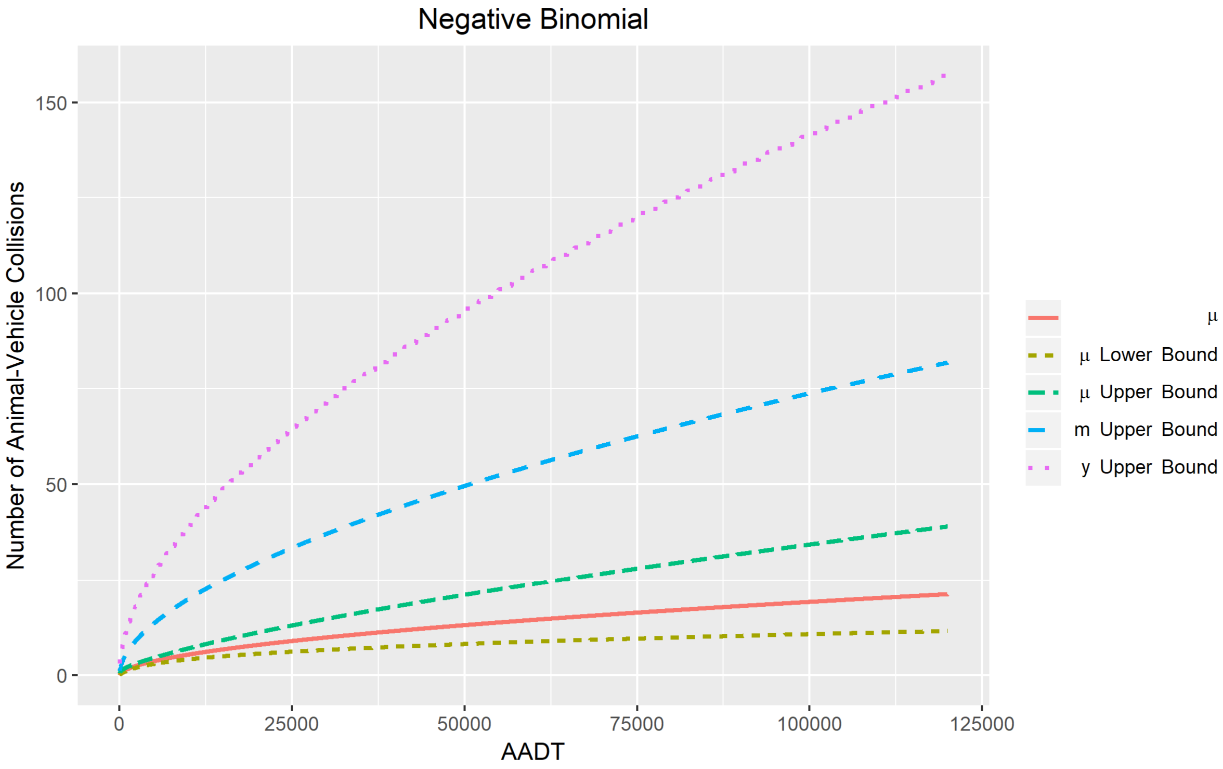


Figure 6-1 95% CIs and PI for Negative Binomial Model

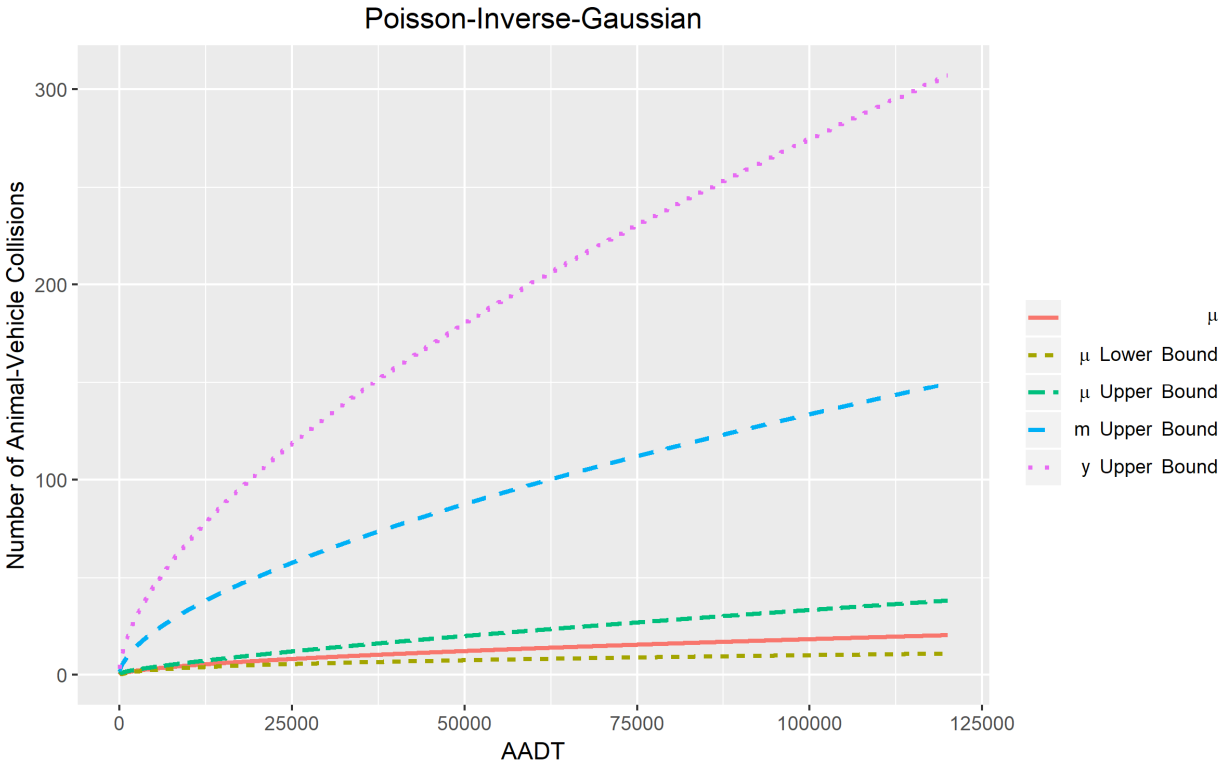


Figure 6-2 95% CIs and PI for Poisson-Inverse-Gaussian Model

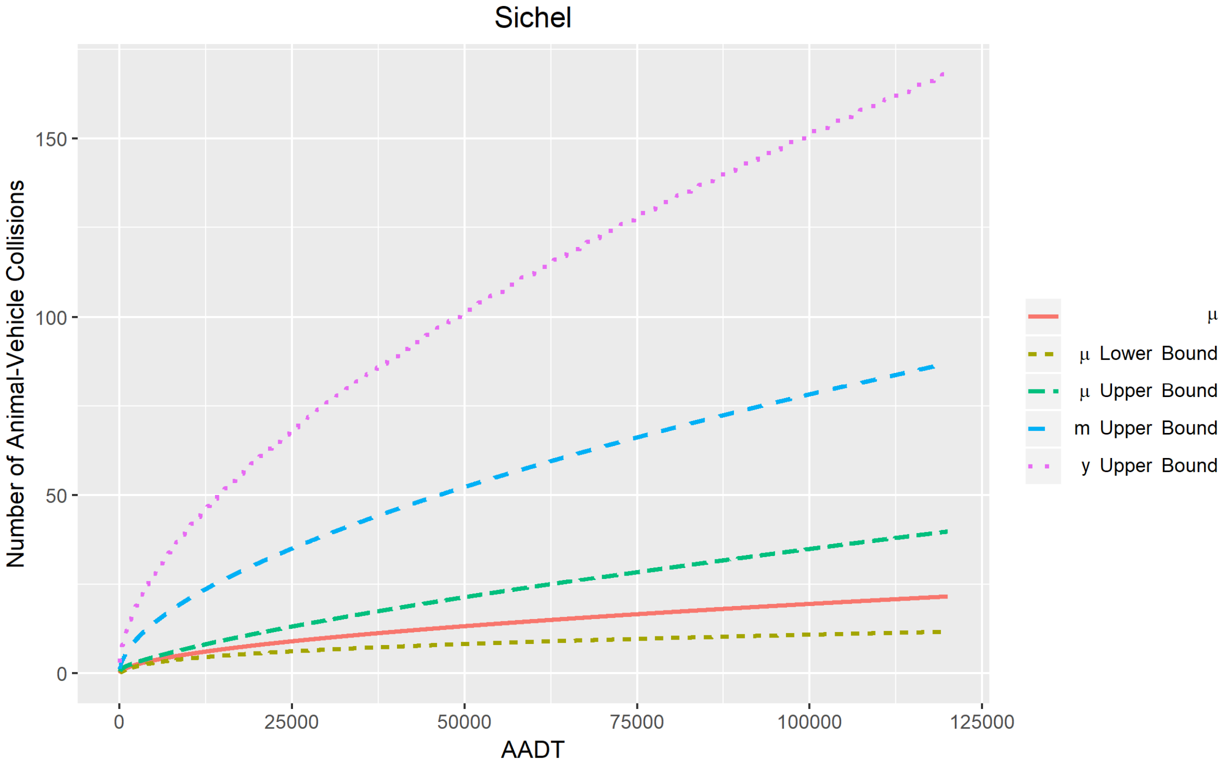


Figure 6-3 95% CIs and PI for Sichel Model

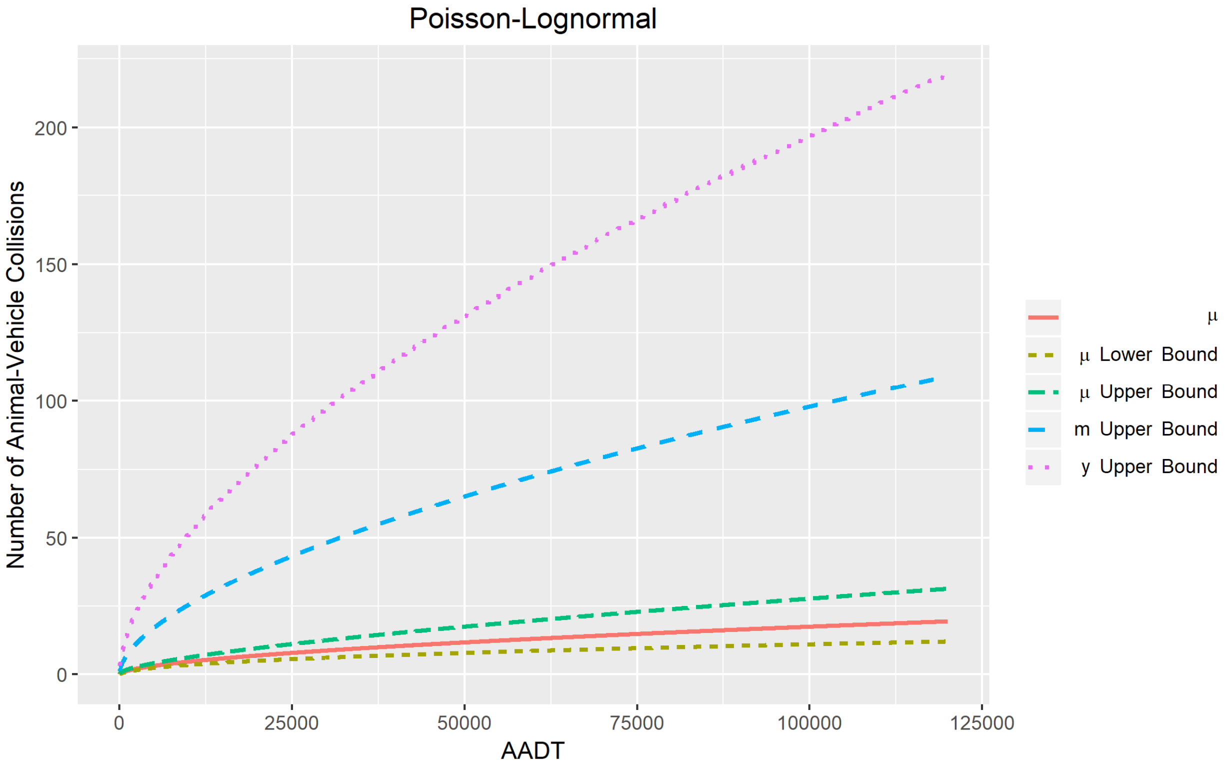


Figure 6-4 95% CIs and PI for Poisson-Lognormal Model

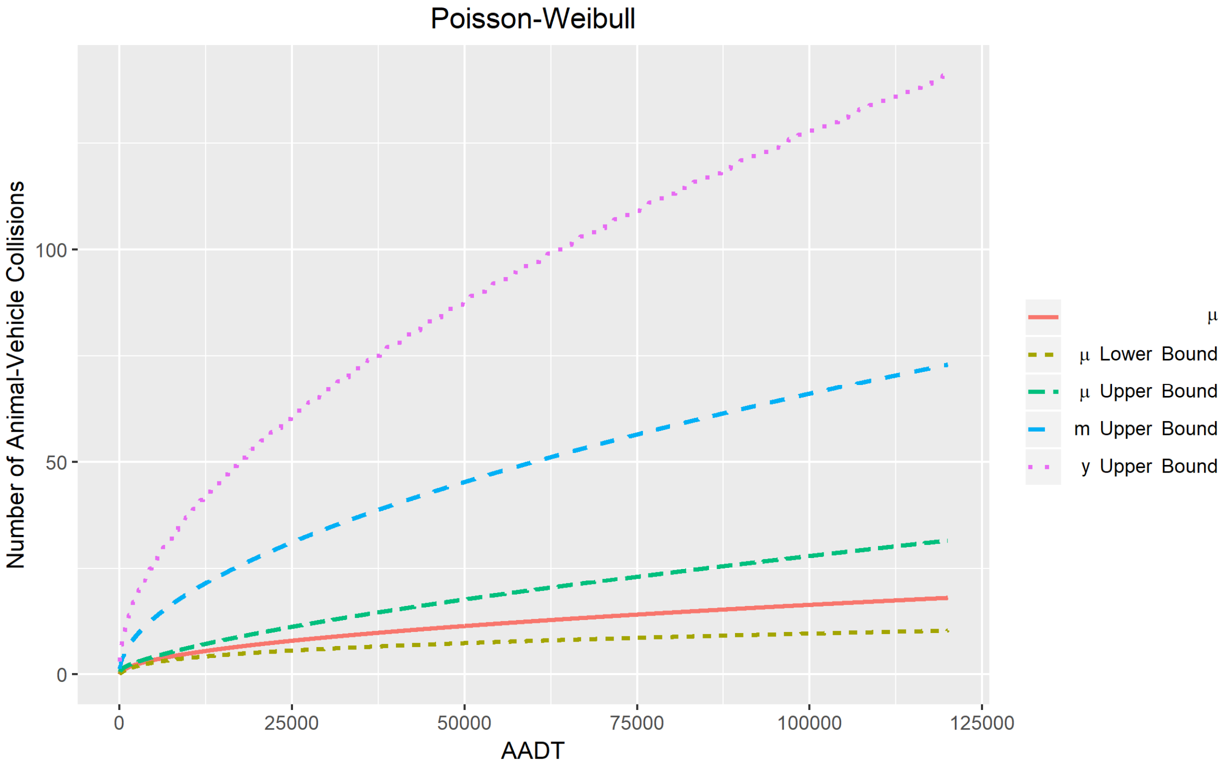


Figure 6-5 95% CIs and PI for Poisson-Weibull Model

Table 6-7 shows a summary of values for all of the CIs and PIs (shown in Figure 6-1 through Figure 6-5) developed based upon the five different types of mixed-Poisson models considered in this study, and in this case, the value of AADT was fixed at 120,000 (i.e., the maximum value on the plots).

Table 6-7 Summary of Values for Mixed-Poisson CIs and PIs at AADT=120,000

	NB	PIG	SI	PLN	PW
μ Lower Bound	11.55	10.82	11.62	11.94	10.24
μ Max	21.23	20.27	21.53	19.32	17.97
μ Upper Bound	39.02	38.00	39.87	31.25	31.53
m Lower Bound	0.00	0.00	0.00	0.00	0.00
m Upper Bound	81.92	149.10	87.06	109.07	72.89
y Lower Bound	0.00	0.00	0.00	0.00	0.00
y Upper Bound	157	307	168	219	141

For the last section of the case study, the estimates of the Poisson mean, μ, and the lower and upper bounds of the 95% CIs and PIs for each model across the range of values of covariates found in the dataset described in Table 6-4 and Table 6-5 were compared. Recall that in the previous section of the case study, the CIs and PIs considered were based on varying the AADT while holding all other variable values fixed (to those values shown in Table 6-6). For this section, the estimates were developed based on the full dataset as described in Table 6-4 and Table 6-5 (i.e., all covariates covered a range of values based on the data collected, and none were fixed to any particular value was done in the preceding section). Here, the mean squared error (MSE) was then calculated (for each form of mixed-Poisson model examined in this study) between the estimated values of μ, and the lower and upper bounds of the CI for μ and the PIs for the safety (m) and the predicted response (y) considering all data points, respectively. The model coefficients in Table 6-4 and Table 6-5were applied to calculate the corresponding values for the Poisson mean

(μ) for all data points in the animal-vehicle collision dataset in the context of each model type. After this, the 95% CI for the Poisson mean (μ) and the PIs for the safety (m) and the predicted response (y) were calculated, again for every data point under each model. Lastly, the MSE values were computed between the estimates of μ and the lower and upper bounds for all of the CIs and PIs, over all data points.

The calculated MSE results considering the estimated values of the Poisson mean (μ) and the lower and upper bounds of the CIs and PIs, by model, are presented in Table 6-8. In this table, one may note that with the animal-vehicle collision dataset considered, the Negative Binomial model provided the estimates for the Poisson mean (μ) that led to the smallest MSE values for each of the confidence and prediction intervals. Hence, it would seem as if the Negative Binomial model may be able to provide smaller variation for the CIs and PIs than that found corresponding to the other models considered in this study. The Poisson-Lognormal model produced the largest MSE values for the 95% CI for Poisson mean (μ), in the context of both the lower and upper bounds of said interval. If one examines the variation between the estimate of μ and the lower bounds for the PI for the safety (m) and the PI for the predicted response (y), over all models, the largest MSE values resulted from the Poisson-Lognormal model. That said, when considering the variation between the estimate of μ and the upper bounds for the PI for the safety (m) and the PI for the predicted response (y), again over all models, the largest MSE values were obtained for the Poisson-Inverse-Gaussian model. For each of the five models examined, the values of MSE for the lower bounds of the PIs for the safety (m) and the predicted response (y) were shown to be equal across all models. Again, this result was due to the simple fact that in all cases, the lower bound of the PIs for m and y was zero.

Table 6-8 MSE Values Calculated between μ and 95% CI and PI Lower and Upper Bounds for the Animal-Vehicle Collision Dataset

	NB	PIG	SI	PLN	PW
μ Lower Bound	2.24	3.28	2.41	4.09	2.37
μ Upper Bound	4.49	6.96	4.95	8.83	4.76
m Lower Bound	30.56	38.35	31.37	48.26	32.68
m Upper Bound	226.46	1439.31	264.18	1008.96	282.80
y Lower Bound	30.56	38.35	31.37	48.26	32.68
y Upper Bound	1165.45	7141.38	1351.75	5025.57	1443.71

6.4 Summary and Conclusions

Following the pioneering work of Wood (2005), this study developed confidence intervals for the Poisson mean (μ), safety or Poisson parameter/safety (m), and predicted response (i.e., number of crashes at a new site, y) for four new types of mixed-Poisson models. This extended beyond the initial work that considered just the Negative Binomial (Poisson-Gamma) and regular Poisson models that was outlined in by Wood (2005). Expressions for these intervals are provided herein and they are now able to be easily applied by safety analysts so they can estimate windows/ranges of uncertainty corresponding to their predictions, instead of a simple point estimate. In this study, the types of mixed-Poisson models examined were the Poisson-Inverse Gaussian, Sichel, Poisson-Lognormal, and Poisson-Weibull models. Each of these mixed-Poisson models is obtained when one allows for a multiplicative error term that follows a chosen mixture distribution to enter the functional form/expression of the Poisson parameter m. Following a brief background and motivation on mixed-Poisson models, the aforementioned confidence and prediction intervals were derived and shown.

After the expressions for the CIs and PIs were developed, a case study was conducted in which the intervals calculated for each of the five aforementioned types of mixed-Poisson models were examined. As this study applied real-world, actually observed crash data, one cannot know the true values of the Poisson mean (μ) and safety or Poisson parameter (m), and as such, they cannot necessarily decide on which intervals perform "best." That said, a series of conclusions can be still be developed from the case study based on the animal-vehicle collision data:

(1) No matter what model type was investigated, the regression coefficient estimates were found to be relatively similar and as such, the values of the Poisson mean (μ) were also similar over the range of AADT values;

(2) When considering all models developed in this study, the Sichel model produced the widest interval for the Poisson mean (μ). Further, the Poisson-Inverse-Gaussian model produced the widest intervals for the safety (m) and predicted response at a new site (y). This model also produced the second greatest value for the upper bound when considering all 95% PIs for the predicted response, y. Ultimately, however, it is important to remember that there is no way to say with full certainty that narrower intervals on the μ, n, and y are necessarily better as the true values of these parameters are not known;

(3) In the case of the Poisson mean (μ), the Poisson-Lognormal model led to the narrowest 95% CI. The Poisson-Weibull model led to the narrowest 95% PIs for m and for y across all model types examined in this study;

(4) All models estimated led to negative values being calculated for the lower bound on the safety (m) (this was of course prior to coercing them to be zero);

(5) For the maximum values of AADT examined, the upper bounds for the PIs for y ranged from 1.92 to 2.06 times the values of the upper bounds of m at the same AADT in a different model;

(6) MSE values were calculated to examine the variation in the estimated values of the Poisson mean (μ) when compared to the lower and upper bounds of the 95% CIs and PIs for each model, considering all datapoints in the animal-vehicle collision dataset. Here, the Negative Binomial model led to estimates of the Poisson mean (μ) that yielded the smallest MSE values across all confidence and prediction intervals. On the other hand, the Poisson-Lognormal model produced the biggest MSE values for the 95% CI for Poisson mean (μ), when considering both the lower and upper bounds for the interval;

(7) When examining the variation between estimates of μ and the lower bounds for the PI for the safety (m) and the PI for the predicted response (y), again over all models, the Poisson-Lognormal model yielded the largest MSE values; and

(8) Lastly, when studying the differences between the estimate of μ and the upper bounds for the PI for the safety (m) and the PI for the predicted response (y), over all models, the Poisson-Inverse-Gaussian model yielded the largest MSE values.

This study concludes naturally with some ideas for future work. As one example, a simulation study where simulation of values for the Poisson mean (μ), safety (m), and response (i.e., crash count, y) at a new site could be conducted in order to help determine which CIs and PIs "best" represent the true intervals. Additionally, as modeling tools and methodologies continue to advance, the estimated CIs and PIs should be extended and further developed for models such as multiparameter models (Geedipally et al. 2012; Lord and Geedipally 2018), random parameters models (Anastasopoulos and Mannering 2009; Rista et al. 2018; Shaon et al. 2018), and semi-parametric models (Heydari et al. 2016; Shirazi et al. 2016; Ye et al. 2018; Zou et al. 2018).

Chapter 7. CONCLUSIONS AND FUTURE WORK

This dissertation focused on crash modeling at two different levels, real-time crash prediction modeling and crash frequency modeling. The primary difference between the two beyond the obvious temporal differences is the type of features involved. RTCPMs generally apply traffic-flow-related features that are often quite dynamic. Crash frequency models, on the other hand, generally look at relatively static features such as descriptors of roadway inventory variables.

With regard to real-time crash prediction modeling, this study was among the first to the author's knowledge to apply large scale probe vehicle trajectory data. Such data allowed (1) greater spatial-temporal resolution and (2) calculation of variables unable to be derived from aggregate loop data. Additionally, an efficient data processing procedure was developed to ingest the raw trajectory data (lon, lat, and timestamp) and after a series of steps and manipulations (map-matching, conflation, etc.), map it to the Washington State Highway linear referencing system. Once the data was converted from its initial raw form, numerous datasets were generated from it in a robust examination of RTCPM study design. In this component, the dissertation investigated the impact of spatial-temporal data collection window size, control definition in a matched case-control study design, as well as potential differences between modeling for different crash types (e.g., all crashes versus multi-vehicle crashes only). Then, a series of real-time crash prediction models were built and observed findings from them were in alignment with many previous studies in terms of impactful predictor variables, as well as their coefficient signs and magnitudes. Notably, it was observed that variables including, but not limited to, coefficient of variation in speed, average speed, and probe vehicle sample size could be used in RTCPM applications. As one example for the use of such findings, that related to CVS provides some further evidence in favor of variable speed limits. Additionally, as this study applied probe vehicle data that only represents a small fraction of the full traffic stream, it was demonstrated that such an approach based on a large volume of data describing a relatively small number of vehicles could produce results in alignment with previous RTCPM studies. This is important as the transportation world continues to push towards a connected state and away from one of fixed sensing infrastructure.

Further, variation in the different hierarchical levels of various kinds of mixed-Poisson models used for crash frequency analysis was investigated. Notably, confidence and prediction intervals for the aforementioned models were developed based on expressions for the variance of the Poisson parameter (also known as the "safety"). Such expressions are based upon the parameters obtained from the regression analyses and can help applicants better understand potential variation in their crash frequency estimates as they vary model parameters.

While this dissertation presented some new approaches to conventional crash modeling problems, it is not to say it is without some drawbacks that can themselves serve as topics of future work. Notably, the sparsity issue posed by the probe vehicle data made it such that only a limited set of traffic-flow variables could be derived from the data. While higher-order position derivatives could be obtained, unlike from loops, there is still tremendous potential for capturing variables describing car-following as well as other driver-behavior and surrogate safety related variables as the penetration rate of vehicles reporting location data via GPS increases. The ultimate goal here would be to study how individual vehicle trajectories correlate with crash occurrence. Then, one may be able to derive new surrogate safety/conflict metrics and get a better understanding of how said conflicts relate to crashes. When such data are available, a substantial amount of work will need to be done on the data processing end as the volume of data associated with a relatively small sample of the traffic population used herein was quite large, compared to that used in previous RTCPM studies.

REFERENCES

Abdel-Aty, M. A., Hassan, H. M., Ahmed, M., and Al-Ghamdi, A. S. (2012). “Real-time prediction of visibility related crashes.” *Transportation Research Part C: Emerging Technologies*, 24, 288–298.

Abdel-Aty, M., and Abdalla, M. F. (2004). “Linking Roadway Geometrics and Real-Time Traffic Characteristics to Model Daytime Freeway Crashes: Generalized Estimating Equations for Correlated Data.” *Transportation Research Record*, SAGE Publications Inc, 1897(1), 106–115.

Abdel-Aty, M., and Pande, A. (2005). “Identifying crash propensity using specific traffic speed conditions.” *Journal of Safety Research*, 36(1), 97–108.

Abdel-Aty, M., Uddin, N., and Pande, A. (2005). “Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways.” *Transportation Research Record*, (1908), 51–58.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., and Hsia, L. (2004). “Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression.” *Transportation Research Record*, 1897(1), 88–95.

Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley series in probability and mathematical statistics, Wiley-Interscience, Hoboken, NJ.

Aguero-Valverde, J., and Jovanis, P. (2008). “Analysis of Road Crash Frequency with Spatial Models.” *Transportation Research Record: Journal of the Transportation Research Board*, 2061, 55–63.

Ahmed, M. M., and Abdel-Aty, M. A. (2012). “The Viability of Using Automatic Vehicle Identification Data for Real-Time Crash Prediction.” *IEEE Transactions on Intelligent Transportation Systems*, 13(2), 459–468.

Anastasopoulos, P. Ch., and Mannering, F. L. (2009). “A note on modeling vehicle accident frequencies with random-parameters count models.” *Accident Analysis & Prevention*, 41(1), 153–159.

Ash, J. E., Zou, Y., Lord, D., and Wang, Y. (2019). “Comparison of confidence and prediction intervals for different mixed-Poisson regression models.” *Journal of Transportation Safety & Security*, 1–23.

Bao, J., Liu, P., Yu, H., and Xu, C. (2017). “Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas.” *Accident Analysis & Prevention*, 106, 358–369.

Bonneson, J., and Ivan, J. (2013). *Theory, Explanation, and Prediction in Road Safety: Promising Directions*. Transportation Research Board, Washington, D.C., 50.

Bowman, A. W., and Azzalini, A. (2019). *The “sm” package for R*.

Breslow, N. E., and Day, N. E. (1980). *Statistical Methods in Cancer Research*. International Agency for Research on Cancer, Lyon.

Brinton, L. A., Berman, M. L., Mortel, R., Twiggs, L. B., Barrett, R. J., Wilbanks, G. D., Lannom, L., and Hoover, R. N. (1992). “Reproductive, menstrual, and medical risk factors for endometrial cancer: Results from a case-control study.” *American Journal of Obstetrics and Gynecology*, 167(5), 1317–1325.

Cameron, A. C., and Trivedi, P. K. (2013). *Regression Analysis of Count Data (Econometric Society Monographs)*. Cambridge University Press, Cambridge.

Casella, G., and Berger, R. L. (2001). *Statistical Inference*. Cengage Learning, Australia ; Pacific Grove, CA.

Cheng, L., Geedipally, S. R., and Lord, D. (2013). "The Poisson–Weibull generalized linear model for analyzing motor vehicle crash data." *Safety Science*, 54, 38–42.

Collett, D. (1991). *Modeling Binary Data*. Chapman and Hall, London, United Kingdom.

Connors, R. D., Maher, M., Wood, A., Mountain, L., and Ropkins, K. (2013). "Methodology for fitting and updating predictive accident models with trend." *Accident Analysis & Prevention*, 56, 82–94.

Daganzo, C. F. (1997). *Fundamentals of Transportation and Traffic Operations*. Elsevier, New York, N.Y.

Dean, C., Lawless, J. F., and Willmot, G. E. (1989). "A Mixed Poisson-Inverse-Gaussian Regression Model." *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 17(2), 171–181.

El-Basyouny, K., and Sayed, T. (2006). "Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models." *Transportation Research Record: Journal of the Transportation Research Board*, 1950, 9–16.

Elvik, R., Erke, A., and Christensen, P. (2009). "Elementary Units of Exposure." 1–13.

Fazeen, M., Gozick, B., Dantu, R., Bhukhiya, M., and González, M. C. (2012). "Safe Driving Using Mobile Phones." *IEEE Transactions on Intelligent Transportation Systems*, 13(3), 1462–1468.

FHWA. (2019). "Safety Culture and the Zero Deaths Vision - Safety | Federal Highway Administration." *FHWA Safety*, <https://safety.fhwa.dot.gov/zerodeaths/> (Oct. 25, 2019).

Geedipally, S. R., and Lord, D. (2008). "Effects of Varying Dispersion Parameter of Poisson–Gamma Models on Estimation of Confidence Intervals of Crash Prediction Models." *Transportation Research Record: Journal of the Transportation Research Board*, 2061(1), 46–54.

Geedipally, S. R., and Lord, D. (2010). "Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson–gamma models." *Accident Analysis & Prevention*, 42(4), 1273–1282.

Geedipally, S. R., Lord, D., and Dhavala, S. S. (2012). "The negative binomial-Lindley generalized linear model: Characteristics and application using crash data." *Accident Analysis & Prevention*, 45, 258–265.

Giuffrè, O., Granà, A., Roberta, M., and Corriere, F. (2011). "Handling Underdispersion in Calibrating Safety Performance Function at Urban, Four-Leg, Signalized Intersections." *Journal of Transportation Safety & Security*, 3(3), 174–188.

Gould, W. (2000). "STATA Technical Bulletin: sg124 Interpreting Logistic Regression in All its Forms." STATA Corporation.

Graaf, M. A. de, Jager, K. J., Zoccali, C., and Dekker, F. W. (2011). "Matching, an Appealing Method to Avoid Confounding?" *Nephron Clinical Practice*, Karger Publishers, 118(4), c315–c318.

Gustavsson, J. (1969). "On the use of regression models in the study of road accidents." *Accident Analysis & Prevention*, 1(4), 315–321.

Gustavsson, J., and Svensson, Å. (1976). "A Poisson Regression Model Applied to Classes of Road Accidents with Small Frequencies." *Scandinavian Journal of Statistics*, 3(2), 49–60.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics, Springer, New York, NY.

Hauer, E. (1982). "Traffic conflicts and exposure." *Accident Analysis & Prevention*, 14(5), 359–364.

Hauer, E. (1992). "Empirical bayes approach to the estimation of 'unsafety': The multivariate regression method." *Accident Analysis & Prevention*, 24(5), 457–477.

Hauer, E. (1995). "On Exposure and Accident Rate." *Traffic Engineering and Control*, 36(3), 134–138.

Hauer, E. (1997). *Observational Before/After Studies in Road Safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*.

Hauer, E., Ng, J. C. N., and Lovell, J. (1988). "ESTIMATION OF SAFETY AT SIGNALIZED INTERSECTIONS (WITH DISCUSSION AND CLOSURE)." *Transportation Research Record*, (1185).

Henrickson, K. C. (2018). "A Framework for Understanding and Addressing Bias and Sparsity in Mobile Location-Based Traffic Data." Ph.D., University of Washington, United States -- Washington.

Henrickson, K. C., Rodrigues, F., and Pereira, F. C. (2019). "Data Preparation." *Mobility Patterns, Big Data and Transport Analytics: Tools and Applications for Modeling*, Elsevier, Amsterdam, Netherlands, 73–106.

Herrera, J. C., Work, D. B., Herring, R., Ban, X. (Jeff), Jacobson, Q., and Bayen, A. M. (2010). "Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment." *Transportation Research Part C: Emerging Technologies*, 18(4), 568–583.

Herring, R., Hofleitner, A., Abbeel, P., and Bayen, A. (2010). "Estimating arterial traffic conditions using sparse probe data." *13th International IEEE Conference on Intelligent Transportation Systems*, 929–936.

Heydari, S., Fu, L., Lord, D., and Mallick, B. K. (2016). "Multilevel Dirichlet process mixture analysis of railway grade crossing crash data." *Analytic Methods in Accident Research*, 9, 27–43.

Hofleitner, A., Herring, R., Abbeel, P., and Bayen, A. (2012a). "Learning the Dynamics of Arterial Traffic From Probe Data Using a Dynamic Bayesian Network." *IEEE Transactions on Intelligent Transportation Systems*, 13(4), 1679–1693.

Hofleitner, A., Herring, R., and Bayen, A. (2012b). "Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning." *Transportation Research Part B: Methodological*, 46(9), 1097–1122.

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., and Sadeek, S. N. (2019). "Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements." *Accident Analysis & Prevention*, 124, 66–84.

Hossain, M., and Muromachi, Y. (2012). "A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways." *Accident Analysis & Prevention*, 45, 373–381.

Hourdos, J. N. (2005). "Crash prone traffic flow dynamics: Identification and real -time detection." Ph.D., University of Minnesota, United States -- Minnesota.

Hourdos, J. N., Garg, V., Michalopoulos, P. G., and Davis, G. A. (2006). “Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations.” *Transportation Research Record*, 1968, 9.

Jovanis, P. P., and Chang, H.-L. (1986). “MODELING THE RELATIONSHIP OF ACCIDENTS TO MILES TRAVELED.” *Transportation Research Record*, (1068).

Kleinbaum, D. G., Sullivan, K. M., and Barker, N. D. (Eds.). (2007). “Matching - Seems Easy, But not that Easy.” *A Pocket Guide to Epidemiology*, Springer, New York, NY, 257–275.

Lan, B., and Persaud, B. (2012). “Evaluation of Multivariate Poisson Log Normal Bayesian Methods for Before-After Road Safety Evaluations.” *Journal of Transportation Safety & Security*, 4(3), 193–210.

Lao, Y., Wu, Y.-J., Corey, J., and Wang, Y. (2011). “Modeling animal-vehicle collisions using diagonal inflated bivariate Poisson regression.” *Accident Analysis & Prevention*, 43(1), 220–227.

Lawless, J. F. (1987). “Negative binomial and mixed poisson regression.” *Canadian Journal of Statistics*, 15(3), 209–225.

List, G. F., Williams, B., Hranac, R., Barkley, T., Mai, E., Ciccarelli, A., Rodegerdts, L., Pincus, K., Nevers, B., Karr, A. F., Zhou, X., Wojtowicz, J., Schofer, J., and Khattak, A. (2014). *Establishing Monitoring Programs for Travel Time Reliability*. Transportation Research Board, Washington, D.C.

Liu, M., and Chen, Y. (2017). “Predicting Real-Time Crash Risk for Urban Expressways in China.” *Mathematical Problems in Engineering*, Research Article, Hindawi, <https://www.hindawi.com/journals/mpe/2017/6263726/> (Dec. 21, 2020).

Lord, D. (2008). “Methodology for estimating the variance and confidence intervals for the estimate of the product of baseline models and AMFs.” *Accident Analysis & Prevention*, 40(3), 1013–1017.

Lord, D., and Geedipally, S. R. (2018). “Chapter 14. Safety Prediction with Datasets Characterised with Excess Zero Responses and Long Tails.” *Safe Mobility: Challenges, Methodology and Solutions*, Transport and Sustainability, Emerald Publishing Limited, 297–323.

Lord, D., Geedipally, S. R., and Guikema, S. D. (2010a). “Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Underdispersion.” *Risk Analysis*, 30(8), 1268–1276.

Lord, D., Kuo, P.-F., and Geedipally, S. R. (2010b). “Comparison of Application of Product of Baseline Models and Accident-Modification Factors and Models with Covariates: Predicted Mean Values and Variance.” *Transportation Research Record: Journal of the Transportation Research Board*, (2147).

Lord, D., and Mannering, F. (2010). “The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives.” *Transportation Research Part A: Policy and Practice*, 44(5), 291–305.

Lord, D., and Miranda-Moreno, L. F. (2008). “Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective.” *Safety Science*, 46(5), 751–770.

Lord, D., Washington, S. P., and Ivan, J. N. (2005). “Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory.” *Accident Analysis & Prevention*, 37(1), 35–46.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). “WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility.” *Statistics and Computing*, 10(4), 325–337.

Mannering, F. L., and Bhat, C. R. (2014). “Analytic methods in accident research: Methodological frontier and future directions.” *Analytic Methods in Accident Research*, 1, 1–22.

Maycock, G., and Hall, R. D. (1984). “ACCIDENTS AT 4-ARM ROUNDABOUTS.”

Newson, P., and Krumm, J. (2009). “Hidden Markov map matching through noise and sparseness.” *In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 336–343.

NHTSA. (2019). “Quick Facts 2017.” NHTSA CrashStats, <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812747> (Oct. 25, 2019).

Oh, C., Oh, J. S., and Chang, M. (2001). “Real-time estimation of freeway accident likelihood.” *Proceedings of the 80th Annual Meeting of the Transportation Research Board*, Washington, D.C.

Ozbay, K., and Kachroo, P. (1999). *Incident Management in Intelligent Transportation Systems*. Artech House, Inc., Norwood, MA.

Pande, A. (2005). “Estimation of hybrid models for real-time crash risk assessment on freeways.” Ph.D., University of Central Florida, United States -- Florida.

Pande, A., and Abdel-Aty, M. (2006). “Assessment of freeway traffic parameters leading to lane-change related collisions.” *Accident Analysis & Prevention*, 38(5), 936–948.

Park, E. S., Carlson, P. J., Porter, R. J., and Andersen, C. K. (2012). “Safety effects of wider edge lines on rural, two-lane highways.” *Accident Analysis & Prevention*, Intelligent Speed Adaptation + Construction Projects, 48, 317–325.

Qin, X., Ivan, J. N., and Ravishanker, N. (2004). “Selecting exposure measures in crash rate prediction for two-lane highway segments.” *Accident Analysis & Prevention*, 36(2), 183–191.

Rigby, R. A., and Stasinopoulos, D. M. (2005). “Generalized additive models for location, scale and shape.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3), 507–554.

Rigby, R. A., Stasinopoulos, D. M., and Akantziliotou, C. (2008). “A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution.” *Computational Statistics & Data Analysis*, 53(2), 381–393.

Rigby, R., and Stasinopolous, D. (2009). *A flexible regression approach using GAMLSS in R*. London Metropolitan University, London.

Rista, E., Goswamy, A., Wang, B., Barrette, T., Hamzeie, R., Russo, B., Bou-Saab, G., and Savolainen, P. T. (2018). “Examining the safety impacts of narrow lane widths on urban/suburban arterials: Estimation of a panel data random parameters negative binomial model.” *Journal of Transportation Safety & Security*, 10(3), 213–228.

Roshandel, S., Zheng, Z., and Washington, S. (2015). “Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis.” *Accident Analysis & Prevention*, 79, 198–211.

Saunier, N., and Sayed, T. (2008). “Probabilistic framework for automated analysis of exposure to road collisions.” *Transportation Research Record: Journal of the Transportation Research Board*, (2083), 96–104.

Savolainen, P. T., Mannering, F. L., Lord, D., and Quddus, M. A. (2011). “The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives.” *Accident Analysis & Prevention*, 43(5), 1666–1676.

Schlesselman, J. J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press.

SDOT. (2019). “Vision Zero - Transportation | seattle.gov.” *City of Seattle*, <https://www.seattle.gov/visionzero> (Oct. 25, 2019).

Selby, J. V., Friedman, G. D., Quesenberry, C. P., and Weiss, N. S. (1992). “A Case–Control Study of Screening Sigmoidoscopy and Mortality from Colorectal Cancer.” *New England Journal of Medicine*, Massachusetts Medical Society, 326(10), 653–657.

Shaon, M. R. R., Qin, X., Shirazi, M., Lord, D., and Geedipally, S. R. (2018). “Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data.” *Analytic Methods in Accident Research*, 18, 33–44.

Shew, C., Pande, A., and Nuworsoo, C. (2013). “Transferability and robustness of real-time freeway crash risk assessment.” *Journal of Safety Research*, 46, 83–90.

Shi, Q., and Abdel-Aty, M. (2015). “Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways.” *Transportation Research Part C: Emerging Technologies*, Big Data in Transportation and Traffic Engineering, 58, 380–394.

Shirazi, M., Lord, D., Dhavala, S. S., and Geedipally, S. R. (2016). “A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data.” *Accident Analysis & Prevention*, 91, 10–18.

Sokejima, S., and Kagamimori, S. (1998). “Working hours as a risk factor for acute myocardial infarction in Japan: case-control study.” *BMJ*, British Medical Journal Publishing Group, 317(7161), 775–780.

Srinivasan, R., Baek, J., and Council, F. (2010). “Safety Evaluation of Transverse Rumble Strips on Approaches to Stop-Controlled Intersections in Rural Areas.” *Journal of Transportation Safety & Security*, 2(3), 261–278.

Stevens, J. R. (2020). “STAT 5500/6500 Conditional Logistic Regression for Matched Pairs.” *Dr. John R. Stevens*, <https://math.usu.edu/jrstevens/biostat/projects2013/rep_CondlLogReg.pdf> (Jul. 31, 2020).

Stipancic, J., Miranda-Moreno, L., and Saunier, N. (2017). “Impact of Congestion and Traffic Flow on Crash Frequency and Severity: Application of Smartphone-Collected GPS Travel Data.” *Transportation Research Record: Journal of the Transportation Research Board*, 2659(1), 43–54.

Stipancic, J., Miranda-Moreno, L., and Saunier, N. (2018a). “Vehicle manoeuvers as surrogate safety measures: Extracting data from the gps-enabled smartphones of regular drivers.” *Accident Analysis & Prevention*, 115, 160–169.

Stipancic, J., Miranda-Moreno, L., Saunier, N., and Labbe, A. (2018b). “Surrogate safety and network screening: Modelling crash frequency using GPS travel data and latent Gaussian Spatial Models.” *Accident Analysis & Prevention*, 120, 174–187.

Stylianou, K., Dimitriou, L., and Abdel-Aty, M. (2019). “Big Data and Road Safety: A Comprehensive Review.” *Mobility Patterns, Big Data and Transport Analytics: Tools and Applications for Modeling*, Elsevier, Amsterdam, Netherlands, 297–343.

Sun, J., and Sun, J. (2015). "A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data." *Transportation Research Part C: Emerging Technologies*, 54, 176–186.

U.S. Department of Transportation. (2018). *U.S. Department of Transportation Strategic Plan for FY 2018-2022*. U.S. Department of Transportation, Washington, D.C.

Wolshon, B., Parr, S. A., and Mousavi, S. M. (2015). *Identifying High-Risk Roadways for Infrastructure Investment Using Naturalistic Driving Data 10/01 2013 – 06/31/2015*. University of Arkansas, Fayetteville, AR.

Wood, G. R. (2005). "Confidence and prediction intervals for generalised linear accident models." *Accident Analysis & Prevention*, 37(2), 267–273.

WSDOT. (2019). "Target Zero: Strategic Highway Safety Plan | WSDOT." *Washington State Department of Transportation*, <https://www.wsdot.wa.gov/planning/SHSP.htm> (Oct. 25, 2019).

Ye, X., Pendyala, R. M., Shankar, V., and Konduri, K. C. (2013). "A simultaneous equations model of crash frequency by severity level for freeway sections." *Accident Analysis & Prevention*, 57, 140–149.

Ye, X., Wang, K., Zou, Y., and Lord, D. (2018). "A semi-nonparametric Poisson regression model for analyzing motor vehicle crash data." *PLOS ONE*, 13(5), e0197338.

Yu, R., and Abdel-Aty, M. (2013a). "Utilizing support vector machine in real-time crash risk evaluation." *Accident Analysis & Prevention*, 51, 252–259.

Yu, R., and Abdel-Aty, M. (2013b). "Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes." *Accident Analysis & Prevention*, 58, 97–105.

Yu, R., Abdel-Aty, M., and Ahmed, M. (2013). "Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors." *Accident Analysis & Prevention*, 50, 371–376.

Zha, L., Lord, D., and Zou, Y. (2016). "The Poisson inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data." *Journal of Transportation Safety & Security*, 8(1), 18–35.

Zhao, M., Liu, C., Li, W., and Sharma, A. (2018). "Multivariate Poisson-lognormal model for analysis of crashes on urban signalized intersections approach." *Journal of Transportation Safety & Security*, 10(3), 251–265.

Zou, Y., Ash, J. E., Park, B.-J., Lord, D., and Wu, L. (2018). "Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety." *Journal of Applied Statistics*, 45(9), 1652–1669.

Zou, Y., Wu, L., and Lord, D. (2015). "Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in Negative Binomial models." *Analytic Methods in Accident Research*, 5–6, 1–16.

APPENDIX

SUMMARY DATA TABLES FOR RTCPM MODELS IN CHAPTER 4

Table 0-1 Dataset for Control Def. #1 Upstream=0.5 mi, downstream=0.0mi

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	15	4.10	2.51	1	15	4.17	2.53
Avg. Speed		8.51	72.30	42.44	13.93	8.51	72.30	41.89	13.76
Min. Speed		0.00	69.77	8.35	15.25	0.00	69.77	7.80	14.70
Max. Speed		27.79	99.98	76.12	10.96	27.79	99.98	76.07	10.85
HBE Count		0	244	4.12	15.13	0	244	4.00	14.70
HAE Count		0	266	3.55	14.86	0	266	3.42	14.48
Severe Jerk Count		0	120	0.98	4.90	0	120	0.93	4.68
SD Speed		0.79	39.72	16.12	5.55	0.79	30.35	16.32	5.39
SD Acceleration		0.01	59.57	2.81	3.65	0.01	46.19	2.77	3.36
SD Jerk		0.00	59.11	2.41	4.61	0.00	49.58	2.35	4.37
CVS		0.01	1.65	0.45	0.25	0.01	1.65	0.46	0.25
Rain				0.02				0.02	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	18	4.13	2.60	1	18	4.22	2.62
Avg. Speed		8.65	79.68	42.81	13.98	8.65	74.01	42.25	13.85
Min. Speed		0.00	75.83	9.05	16.37	0.00	69.08	8.52	15.87
Max. Speed		19.82	99.98	75.85	10.89	19.82	99.98	75.83	10.90
HBE Count		0	240	4.41	15.46	0	240	4.41	15.35
HAE Count		0	260	3.88	15.14	0	260	3.88	15.06
Severe Jerk Count		0	120	1.00	5.00	0	120	1.00	5.00
SD Speed		1.26	31.94	15.83	5.74	1.26	31.94	16.01	5.62
SD Acceleration		0.02	32.84	2.70	3.27	0.02	32.84	2.69	3.21
SD Jerk		0.00	50.70	2.39	4.41	0.00	50.70	2.39	4.38
CVS		0.02	1.44	0.44	0.25	0.02	1.44	0.45	0.24

Table 0-2 Dataset for Control Def. #2 Upstream=0.5 mi, downstream=0.0mi

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	17	3.56	2.43	1	17	3.48	2.40
Avg. Speed		8.51	74.85	48.10	14.04	8.65	77.31	48.24	14.21
Min. Speed		0.00	69.03	13.30	19.98	0.00	68.62	14.00	20.32
Max. Speed		22.39	99.99	76.62	10.63	19.20	99.98	76.81	10.78
HBE Count		0	315	4.54	18.65	0	240	3.71	14.22
HAE Count		0	314	3.96	18.23	0	260	3.22	14.27
Severe Jerk Count		0	132	1.17	6.85	0	120	0.90	5.21
SD Speed		0.36	39.72	14.21	6.52	0.68	30.70	14.13	6.59
SD Acceleration		0.01	59.57	2.91	4.17	0.00	41.85	2.66	3.54
SD Jerk		0.00	75.36	2.53	5.56	0.00	45.89	2.25	4.60
CVS		0.01	1.65	0.36	0.25	0.01	1.59	0.36	0.25
Rain				0.10				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	17	3.45	2.38	1	17	3.48	2.40
Avg. Speed		8.65	77.31	48.56	14.11	8.65	77.31	48.24	14.21
Min. Speed		0.00	68.62	14.34	20.48	0.00	68.62	14.00	20.32
Max. Speed		19.20	99.98	76.89	10.78	19.20	99.98	76.81	10.78
HBE Count		0	240	3.70	14.13	0	240	3.71	14.22
HAE Count		0	260	3.22	14.11	0	260	3.22	14.27
Severe Jerk Count		0	120	0.88	5.07	0	120	0.90	5.21
SD Speed		0.68	34.61	14.03	6.63	0.68	30.70	14.13	6.59
SD Acceleration		0.00	41.85	2.63	3.48	0.00	41.85	2.66	3.54
SD Jerk		0.00	45.89	2.23	4.53	0.00	45.89	2.25	4.60
CVS		0.01	1.59	0.36	0.25	0.01	1.59	0.36	0.25

Table 0-3 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=0.0mi

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	22	5.85	3.59	1	22	5.97	3.61
Avg. Speed		8.51	72.30	43.18	13.52	8.51	72.30	42.53	13.37
Min. Speed		0.00	69.77	6.97	14.53	0.00	69.77	6.51	13.92
Max. Speed		33.66	99.98	78.13	10.84	33.66	99.98	78.12	10.81
HBE Count		0	244	4.56	14.74	0	244	4.51	14.46
HAE Count		0	266	3.82	14.40	0	266	3.77	14.16
Severe Jerk Count		0	120	1.04	4.64	0	120	1.01	4.47
SD Speed		0.85	31.14	16.34	5.49	0.93	31.14	16.55	5.34
SD Acceleration		0.00	49.95	3.07	3.72	0.01	38.53	3.05	3.54
SD Jerk		0.00	66.99	2.76	4.91	0.00	66.99	2.74	4.77
CVS		0.02	1.84	0.45	0.24	0.02	1.84	0.46	0.24
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	22	5.89	3.59	1	22	6.00	3.62
Avg. Speed		8.81	75.37	43.48	13.59	8.81	75.37	42.85	13.52
Min. Speed		0.00	67.53	7.30	15.12	0.00	67.53	6.97	14.77
Max. Speed		14.12	99.98	77.86	10.74	14.12	99.98	77.79	10.67
HBE Count		0	240	5.13	16.09	0	240	5.08	15.75
HAE Count		0	260	4.40	15.67	0	260	4.35	15.35
Severe Jerk Count		0	120	1.17	5.12	0	120	1.15	4.93
SD Speed		1.26	32.00	16.09	5.55	1.26	32.00	16.26	5.46
SD Acceleration		0.02	32.14	2.98	3.27	0.02	32.14	2.97	3.26
SD Jerk		0.00	42.20	2.70	4.39	0.00	42.20	2.70	4.39
CVS		0.02	1.44	0.44	0.24	0.02	1.44	0.45	0.24

Table 0-4 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=0.0mi

Control Definition #2 (same location, random time)									
		All				**Multi-Veh Only**			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	26	4.77	3.39	1	26	4.82	3.44
Avg. Speed		8.51	74.85	49.29	13.40	8.51	74.85	49.06	13.50
Min. Speed		0.00	70.25	12.29	19.97	0.00	70.25	12.02	19.81
Max. Speed		15.87	99.99	78.22	10.52	15.87	99.99	78.27	10.54
HBE Count		0	315	4.58	17.41	0	315	4.67	17.86
HAE Count		0	314	3.82	17.00	0	314	3.91	17.46
Severe Jerk Count		0	132	1.14	6.34	0	132	1.18	6.55
SD Speed		0.36	41.15	14.37	6.56	0.36	36.62	14.44	6.51
SD Acceleration		0.00	64.42	3.12	4.28	0.00	54.15	3.09	4.06
SD Jerk		0.00	69.06	2.65	5.35	0.00	69.06	2.64	5.34
CVS		0.01	1.84	0.35	0.24	0.01	1.84	0.35	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	25	4.71	3.37	1	25	4.74	3.39
Avg. Speed		9.10	77.27	49.67	13.45	9.10	77.27	49.39	13.56
Min. Speed		0.00	67.85	12.99	20.32	0.00	67.85	12.78	20.17
Max. Speed		19.20	99.99	78.37	10.40	19.20	99.99	78.32	10.42
HBE Count		0	240	4.20	14.78	0	240	4.26	15.01
HAE Count		0	260	3.48	14.39	0	260	3.53	14.64
Severe Jerk Count		0	134	1.04	5.75	0	134	1.08	5.93
SD Speed		0.48	30.63	14.08	6.51	0.48	30.63	14.16	6.48
SD Acceleration		0.00	41.85	2.91	3.77	0.00	41.85	2.94	3.83
SD Jerk		0.00	59.48	2.46	4.80	0.00	59.48	2.49	4.88
CVS		0.01	1.60	0.34	0.24	0.01	1.60	0.35	0.24

Table 0-5 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=0.0mi

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	27	7.22	4.39	1	27	7.42	4.40
Avg. Speed		8.51	72.30	43.82	13.45	8.51	72.30	43.04	13.28
Min. Speed		0.00	69.85	6.58	14.41	0.00	69.85	5.89	13.45
Max. Speed		33.66	99.98	79.29	10.47	33.66	99.98	79.33	10.43
HBE Count		0	244	5.31	15.92	0	244	5.35	15.84
HAE Count		0	266	4.46	15.50	0	266	4.48	15.45
Severe Jerk Count		0	120	1.21	4.91	0	120	1.20	4.80
SD Speed		0.85	30.93	16.34	5.49	1.27	29.49	16.62	5.28
SD Acceleration		0.01	49.95	3.12	3.51	0.01	36.34	3.12	3.36
SD Jerk		0.00	61.20	2.89	4.74	0.00	61.20	2.88	4.63
CVS		0.02	1.83	0.44	0.24	0.02	1.83	0.45	0.24
Rain				0.01				0.02	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	32	7.29	4.47	1	32	7.48	4.49
Avg. Speed		9.10	75.37	44.04	13.47	9.10	75.37	43.31	13.36
Min. Speed		0.00	67.53	6.79	14.73	0.00	67.53	6.15	13.87
Max. Speed		39.06	99.98	79.07	10.46	39.06	99.98	79.10	10.42
HBE Count		0	240	5.78	17.22	0	240	5.79	17.03
HAE Count		0	260	4.90	16.75	0	260	4.91	16.58
Severe Jerk Count		0	120	1.31	5.36	0	120	1.30	5.23
SD Speed		1.26	30.30	16.13	5.56	1.26	30.30	16.38	5.39
SD Acceleration		0.01	27.98	3.07	3.13	0.01	27.98	3.08	3.14
SD Jerk		0.00	36.63	2.86	4.36	0.00	36.63	2.87	4.37
CVS		0.02	1.39	0.43	0.24	0.02	1.39	0.45	0.24

Table 0-6 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=0.0mi

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	26	5.71	4.12	1	26	5.79	4.18
Avg. Speed		4.87	74.73	49.96	13.17	4.87	74.73	49.73	13.29
Min. Speed		0.00	69.03	12.02	19.90	0.00	69.03	11.85	19.85
Max. Speed		14.04	99.99	78.96	10.36	14.04	99.99	79.00	10.34
HBE Count		0	315	4.77	17.23	0	315	4.86	17.65
HAE Count		0	314	3.94	16.71	0	314	4.03	17.14
Severe Jerk Count		0	132	1.17	6.22	0	132	1.21	6.44
SD Speed		0.36	41.15	14.22	6.53	0.36	36.62	14.29	6.50
SD Acceleration		0.00	64.42	3.14	4.19	0.00	54.15	3.12	4.01
SD Jerk		0.00	60.96	2.70	5.25	0.00	60.96	2.70	5.27
CVS		0.01	1.83	0.34	0.24	0.01	1.83	0.35	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	27	5.71	4.12	1	27	5.75	4.15
Avg. Speed		9.10	73.73	50.21	13.18	9.10	73.73	49.96	13.31
Min. Speed		0.00	67.85	12.63	20.37	0.00	67.85	12.44	20.26
Max. Speed		43.07	99.99	79.14	10.19	43.07	99.99	79.09	10.22
HBE Count		0	240	4.60	15.84	0	240	4.65	16.07
HAE Count		0	260	3.82	15.39	0	260	3.86	15.62
Severe Jerk Count		0	134	1.15	6.06	0	134	1.18	6.23
SD Speed		0.17	33.67	14.01	6.50	0.17	29.51	14.09	6.47
SD Acceleration		0.00	41.68	2.97	3.61	0.00	41.68	2.98	3.66
SD Jerk		0.00	58.17	2.55	4.67	0.00	58.17	2.57	4.73
CVS		0.00	1.60	0.34	0.24	0.00	1.60	0.34	0.24

Table 0-7 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=1.0 mi (UP-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	22	6.08	3.60	1	22	6.18	3.62
Avg. Speed		8.51	72.30	42.43	13.30	8.51	72.30	41.92	13.19
Min. Speed		0.00	69.77	5.97	13.20	0.00	69.77	5.71	12.86
Max. Speed		33.66	99.98	78.40	10.89	33.66	99.98	78.40	10.85
HBE Count		0	244	4.76	15.10	0	244	4.73	14.87
HAE Count		0	266	3.98	14.80	0	266	3.93	14.59
Severe Jerk Count		0	120	1.09	4.81	0	120	1.07	4.62
SD Speed		0.85	31.14	16.65	5.20	0.93	31.14	16.80	5.09
SD Acceleration		0.02	38.53	3.10	3.54	0.02	38.53	3.10	3.49
SD Jerk		0.00	66.99	2.79	4.77	0.00	66.99	2.77	4.71
CVS		0.02	1.84	0.46	0.24	0.02	1.84	0.47	0.23
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	22	6.14	3.60	1	22	6.23	3.63
Avg. Speed		8.81	70.69	42.75	13.34	8.81	70.69	42.24	13.28
Min. Speed		0.00	65.27	6.05	13.51	0.00	65.27	5.88	13.36
Max. Speed		14.12	99.98	78.28	10.74	14.12	99.98	78.22	10.67
HBE Count		0	240	5.45	16.70	0	240	5.37	16.30
HAE Count		0	260	4.65	16.27	0	260	4.58	15.88
Severe Jerk Count		0	120	1.25	5.32	0	120	1.21	5.10
SD Speed		1.46	28.93	16.42	5.25	1.46	28.93	16.53	5.19
SD Acceleration		0.02	32.14	3.07	3.27	0.02	32.14	3.06	3.26
SD Jerk		0.00	42.20	2.80	4.42	0.00	42.20	2.80	4.41
CVS		0.02	1.44	0.45	0.23	0.02	1.44	0.46	0.23

Table 0-8 Dataset for Control Def. #1 Upstream=1.0 mi, downstream=1.0 mi (DOWN-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	25	6.18	3.62	1	25	6.25	3.61
Avg. Speed		8.79	73.95	42.29	13.43	8.79	73.95	41.81	13.31
Min. Speed		0.00	69.77	6.27	13.68	0.00	69.77	5.91	13.19
Max. Speed		28.20	99.98	78.28	10.79	28.20	99.98	78.20	10.69
HBE Count		0	244	5.19	16.47	0	244	5.11	16.11
HAE Count		0	266	4.40	16.04	0	266	4.32	15.71
Severe Jerk Count		0	120	1.25	5.36	0	120	1.21	5.19
SD Speed		0.28	29.39	16.47	5.22	0.28	29.39	16.61	5.10
SD Acceleration		0.00	41.74	3.05	3.34	0.00	41.74	3.06	3.34
SD Jerk		0.00	77.03	2.74	4.71	0.00	77.03	2.75	4.71
CVS		0.00	2.44	0.46	0.26	0.00	2.44	0.47	0.26
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	23	6.20	3.64	1	23	6.28	3.64
Avg. Speed		8.93	71.80	42.79	13.21	8.93	71.80	42.33	13.09
Min. Speed		0.00	64.75	6.28	13.74	0.00	64.75	5.89	13.14
Max. Speed		36.86	99.97	78.46	10.66	36.86	99.97	78.41	10.64
HBE Count		0	272	5.48	17.39	0	272	5.51	17.37
HAE Count		0	168	4.56	16.09	0	168	4.61	16.12
Severe Jerk Count		0	57	1.19	4.68	0	57	1.21	4.68
SD Speed		0.28	30.32	16.41	5.34	0.28	30.32	16.55	5.25
SD Acceleration		0.00	28.11	3.05	3.25	0.00	28.11	3.07	3.27
SD Jerk		0.00	43.69	2.77	4.55	0.00	43.69	2.81	4.59
CVS		0.00	1.34	0.45	0.23	0.00	1.34	0.46	0.23

Table 0-9 Dataset for Control Def. #2 Upstream=1.0 mi, downstream=1.0 mi (UP-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	26	5.06	3.44	1	25	5.07	3.41
Avg. Speed		8.51	74.85	48.18	13.45	9.10	77.27	48.65	13.54
Min. Speed		0.00	70.25	10.16	18.13	0.00	66.70	10.84	18.51
Max. Speed		26.19	99.99	78.66	10.55	19.20	99.98	78.96	10.31
HBE Count		0	315	4.96	18.26	0	240	4.57	15.37
HAE Count		0	314	4.16	17.88	0	260	3.79	14.97
Severe Jerk Count		0	132	1.23	6.65	0	134	1.12	5.92
SD Speed		0.36	36.62	14.88	6.24	0.79	30.63	14.63	6.24
SD Acceleration		0.00	42.84	3.16	3.89	0.01	41.85	3.02	3.84
SD Jerk		0.00	69.06	2.75	5.24	0.00	59.48	2.62	4.93
CVS		0.01	1.84	0.37	0.24	0.01	1.60	0.36	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	26	5.12	3.49	1	25	5.09	3.43
Avg. Speed		8.51	74.85	47.98	13.54	9.10	77.27	48.37	13.63
Min. Speed		0.00	70.25	10.02	18.08	0.00	66.70	10.65	18.40
Max. Speed		26.19	99.99	78.68	10.57	19.20	99.98	78.86	10.32
HBE Count		0	315	5.04	18.65	0	240	4.59	15.51
HAE Count		0	314	4.24	18.30	0	260	3.81	15.14
Severe Jerk Count		0	132	1.27	6.85	0	134	1.15	6.08
SD Speed		0.36	36.62	14.93	6.21	0.79	30.63	14.72	6.22
SD Acceleration		0.00	42.84	3.16	3.89	0.01	41.85	3.03	3.89
SD Jerk		0.00	69.06	2.74	5.28	0.00	59.48	2.63	4.98
CVS		0.01	1.84	0.37	0.24	0.01	1.60	0.37	0.24

Table 0-10 Dataset for Control Def. #2 Upstream=1.0mi, downstream=1.0 mi (DOWN-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	24	5.00	3.44	1	24	5.01	3.45
Avg. Speed		8.79	74.85	48.33	13.43	8.79	74.85	48.12	13.53
Min. Speed		0.00	67.43	10.76	18.60	0.00	67.43	10.70	18.61
Max. Speed		43.81	99.99	78.64	10.49	43.81	99.99	78.58	10.49
HBE Count		0	316	4.94	17.70	0	316	5.00	17.95
HAE Count		0	315	4.18	17.60	0	315	4.26	17.91
Severe Jerk Count		0	120	1.18	5.90	0	120	1.21	6.06
SD Speed		0.40	33.09	14.75	6.26	0.40	33.09	14.80	6.24
SD Acceleration		0.01	42.80	3.05	3.72	0.01	42.80	3.03	3.72
SD Jerk		0.00	59.04	2.60	4.87	0.00	59.04	2.59	4.90
CVS		0.01	1.32	0.37	0.24	0.01	1.32	0.37	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	21	4.97	3.33	1	21	4.99	3.32
Avg. Speed		8.93	75.21	48.78	13.49	8.93	75.21	48.58	13.56
Min. Speed		0.00	66.70	11.45	18.91	0.00	66.70	11.26	18.76
Max. Speed		39.02	99.98	78.74	10.41	39.02	99.98	78.68	10.39
HBE Count		0	248	4.13	13.78	0	248	4.11	13.77
HAE Count		0	253	3.44	13.47	0	253	3.42	13.47
Severe Jerk Count		0	155	0.97	4.96	0	155	0.97	5.05
SD Speed		0.29	39.15	14.51	6.36	0.29	39.15	14.56	6.33
SD Acceleration		0.00	41.67	3.01	3.80	0.00	41.67	3.01	3.82
SD Jerk		0.00	47.41	2.62	4.88	0.00	47.41	2.61	4.90
CVS		0.00	1.48	0.36	0.24	0.00	1.48	0.36	0.24

Table 0-11 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=1.0 mi (UP-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	Time Period	Min	Max	Mean	SD	Min	Max	Mean	SD
Sample Size	5-10 min. before crash	1	27	7.58	4.39	1	27	7.73	4.41
Avg. Speed		8.51	72.30	42.85	13.22	8.51	72.30	42.31	13.10
Min. Speed		0.00	69.77	5.32	12.62	0.00	69.77	4.98	12.14
Max. Speed		33.66	99.98	79.66	10.52	33.66	99.98	79.69	10.48
HBE Count		0	244	5.58	16.27	0	244	5.56	16.05
HAE Count		0	266	4.66	15.82	0	266	4.62	15.64
Severe Jerk Count		0	120	1.27	5.06	0	120	1.25	4.91
SD Speed		0.85	29.49	16.73	5.12	2.04	29.49	16.91	4.99
SD Acceleration		0.02	35.64	3.18	3.37	0.02	35.64	3.18	3.34
SD Jerk		0.00	61.20	2.96	4.65	0.00	61.20	2.95	4.61
CVS		0.02	1.83	0.46	0.23	0.03	1.83	0.47	0.23
Rain				0.01				0.02	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	32	7.70	4.46	1	32	7.83	4.49
Avg. Speed		9.10	70.69	43.08	13.22	9.10	70.69	42.58	13.15
Min. Speed		0.00	62.91	5.26	12.61	0.00	62.91	5.02	12.30
Max. Speed		42.06	99.98	79.66	10.40	42.06	99.98	79.63	10.35
HBE Count		0	240	6.25	18.05	0	240	6.21	17.75
HAE Count		0	260	5.30	17.58	0	260	5.25	17.29
Severe Jerk Count		0	120	1.42	5.62	0	120	1.40	5.45
SD Speed		1.50	28.94	16.56	5.18	1.50	28.94	16.69	5.09
SD Acceleration		0.02	26.29	3.20	3.15	0.02	26.29	3.20	3.15
SD Jerk		0.00	36.63	3.00	4.43	0.00	36.63	3.00	4.43
CVS		0.02	1.39	0.45	0.23	0.02	1.39	0.46	0.23

Table 0-12 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=1.0 mi (DOWN-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	25	6.07	3.60	1	25	6.14	3.60
Avg. Speed		8.79	73.95	42.69	13.47	8.79	73.95	42.18	13.36
Min. Speed		0.00	69.77	6.58	13.99	0.00	69.77	6.20	13.48
Max. Speed		28.20	99.98	78.17	10.83	28.20	99.98	78.08	10.75
HBE Count		0	244	5.11	16.34	0	244	4.99	15.91
HAE Count		0	266	4.33	15.91	0	266	4.22	15.54
Severe Jerk Count		0	120	1.23	5.30	0	120	1.19	5.15
SD Speed		0.28	29.39	16.35	5.29	0.28	29.39	16.50	5.17
SD Acceleration		0.00	41.74	3.04	3.37	0.00	41.74	3.04	3.37
SD Jerk		0.00	77.03	2.73	4.73	0.00	77.03	2.72	4.72
CVS		0.00	1.43	0.45	0.24	0.00	1.43	0.46	0.24
Rain				0.01				0.02	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	26	6.09	3.66	1	26	6.17	3.66
Avg. Speed		8.93	71.80	43.15	13.23	8.93	71.80	42.68	13.12
Min. Speed		0.00	64.75	6.63	14.15	0.00	64.75	6.16	13.48
Max. Speed		36.86	99.97	78.34	10.62	36.86	99.97	78.27	10.60
HBE Count		0	272	5.36	17.12	0	272	5.39	17.13
HAE Count		0	168	4.46	15.85	0	168	4.52	15.91
Severe Jerk Count		0	57	1.16	4.60	0	57	1.17	4.60
SD Speed		0.28	30.32	16.31	5.43	0.28	30.32	16.46	5.32
SD Acceleration		0.00	28.11	3.01	3.23	0.00	28.11	3.04	3.25
SD Jerk		0.00	43.69	2.73	4.52	0.00	43.69	2.77	4.57
CVS		0.00	1.34	0.44	0.23	0.00	1.34	0.45	0.23

Table 0-13 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=1.0 mi (UP-stream portion only)

Control Definition #2 (same location, random time)									
		All				**Multi-Veh Only**			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	26	6.20	4.19	1	26	6.27	4.23
Avg. Speed		8.51	74.73	48.64	13.27	8.51	74.73	48.44	13.35
Min. Speed		0.00	68.89	9.55	17.77	0.00	68.89	9.40	17.71
Max. Speed		25.26	99.99	79.59	10.24	25.26	99.99	79.63	10.24
HBE Count		0	315	5.28	18.32	0	315	5.39	18.72
HAE Count		0	314	4.39	17.82	0	314	4.49	18.23
Severe Jerk Count		0	132	1.30	6.64	0	132	1.34	6.83
SD Speed		0.36	36.62	14.89	6.18	0.36	36.62	14.95	6.14
SD Acceleration		0.00	38.02	3.21	3.76	0.00	38.02	3.22	3.78
SD Jerk		0.00	60.96	2.83	5.12	0.00	60.96	2.85	5.17
CVS		0.01	1.83	0.37	0.24	0.01	1.83	0.37	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	27	6.28	4.17	1	27	6.32	4.18
Avg. Speed		9.10	73.56	48.96	13.33	9.10	73.56	48.72	13.41
Min. Speed		0.00	66.78	9.99	18.08	0.00	66.78	9.79	17.95
Max. Speed		45.63	99.98	79.95	10.03	45.63	99.98	79.86	10.04
HBE Count		0	240	5.17	16.82	0	240	5.19	16.97
HAE Count		0	260	4.30	16.35	0	260	4.31	16.51
Severe Jerk Count		0	134	1.28	6.38	0	134	1.30	6.53
SD Speed		1.02	29.51	14.70	6.17	1.02	29.51	14.77	6.14
SD Acceleration		0.01	41.68	3.10	3.67	0.01	41.68	3.11	3.71
SD Jerk		0.00	58.17	2.74	4.79	0.00	58.17	2.75	4.83
CVS		0.01	1.60	0.36	0.24	0.01	1.60	0.36	0.24

Table 0-14 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=1.0 mi (DOWN-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	24	4.86	3.40	1	24	4.87	3.42
Avg. Speed		8.79	74.85	48.77	13.41	8.79	74.85	48.57	13.51
Min. Speed		0.00	67.43	11.55	19.26	0.00	67.43	11.47	19.24
Max. Speed		43.81	99.99	78.45	10.53	43.81	99.99	78.40	10.52
HBE Count		0	316	4.68	17.17	0	316	4.73	17.40
HAE Count		0	315	3.96	17.07	0	315	4.02	17.36
Severe Jerk Count		0	120	1.11	5.72	0	120	1.14	5.87
SD Speed		0.40	33.09	14.53	6.38	0.40	33.09	14.58	6.36
SD Acceleration		0.01	42.80	2.99	3.72	0.01	42.80	2.98	3.73
SD Jerk		0.00	59.04	2.52	4.82	0.00	59.04	2.51	4.85
CVS		0.01	1.32	0.36	0.24	0.01	1.32	0.36	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	26	4.85	3.32	1	26	4.87	3.32
Avg. Speed		8.93	77.69	49.21	13.44	8.93	77.69	49.01	13.53
Min. Speed		0.00	68.55	12.17	19.51	0.00	68.55	11.96	19.36
Max. Speed		31.79	99.98	78.51	10.47	31.79	99.98	78.45	10.45
HBE Count		0	248	4.00	13.57	0	248	3.98	13.55
HAE Count		0	253	3.34	13.28	0	253	3.32	13.28
Severe Jerk Count		0	155	0.94	4.85	0	155	0.94	4.93
SD Speed		0.29	39.15	14.30	6.44	0.29	39.15	14.35	6.41
SD Acceleration		0.00	41.67	2.97	3.80	0.00	41.67	2.97	3.82
SD Jerk		0.00	47.41	2.57	4.86	0.00	47.41	2.56	4.88
CVS		0.00	1.48	0.35	0.24	0.00	1.48	0.35	0.24

Table 0-15 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=1.5 mi (UP-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	27	7.45	4.38	1	27	7.60	4.39
Avg. Speed		8.51	72.30	43.15	13.30	8.51	72.30	42.55	13.17
Min. Speed		0.00	69.77	5.66	13.04	0.00	69.77	5.27	12.49
Max. Speed		33.66	99.98	79.52	10.46	33.66	99.98	79.54	10.43
HBE Count		0	244	5.50	16.19	0	244	5.49	16.03
HAE Count		0	266	4.60	15.74	0	266	4.59	15.60
Severe Jerk Count		0	120	1.25	4.99	0	120	1.23	4.86
SD Speed		0.85	30.93	16.63	5.25	2.04	29.49	16.83	5.11
SD Acceleration		0.02	49.95	3.16	3.47	0.02	35.64	3.14	3.30
SD Jerk		0.00	61.20	2.94	4.70	0.00	61.20	2.92	4.56
CVS		0.02	1.83	0.45	0.24	0.03	1.83	0.46	0.23
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	32	7.54	4.46	1	32	7.68	4.49
Avg. Speed		9.10	73.98	43.43	13.36	9.10	73.98	42.86	13.28
Min. Speed		0.00	64.66	5.86	13.46	0.00	64.66	5.51	12.99
Max. Speed		42.06	99.98	79.38	10.38	42.06	99.98	79.36	10.34
HBE Count		0	240	6.03	17.70	0	240	6.01	17.44
HAE Count		0	260	5.12	17.24	0	260	5.10	16.98
Severe Jerk Count		0	120	1.37	5.51	0	120	1.35	5.35
SD Speed		1.26	28.94	16.39	5.34	1.26	28.94	16.55	5.23
SD Acceleration		0.01	26.29	3.14	3.12	0.01	26.29	3.14	3.12
SD Jerk		0.00	36.63	2.94	4.37	0.00	36.63	2.94	4.38
CVS		0.02	1.39	0.44	0.24	0.02	1.39	0.45	0.23

Table 0-16 Dataset for Control Def. #1 Upstream=1.5 mi, downstream=1.5 mi (DOWN-stream portion only)

Control Definition #1 (same location, day of week, and time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	30	7.59	4.48	1	30	7.70	4.48
Avg. Speed		8.99	72.30	42.98	13.35	8.99	72.30	42.39	13.21
Min. Speed		0.00	62.87	5.86	13.30	0.00	62.24	5.46	12.70
Max. Speed		28.20	99.98	79.30	10.54	28.20	99.98	79.20	10.46
HBE Count		0	244	5.87	17.13	0	244	5.81	16.83
HAE Count		0	266	5.01	16.75	0	266	4.95	16.51
Severe Jerk Count		0	120	1.40	5.69	0	120	1.39	5.59
SD Speed		0.28	31.65	16.46	5.28	0.28	31.65	16.64	5.15
SD Acceleration		0.00	41.74	3.13	3.32	0.00	41.74	3.13	3.26
SD Jerk		0.00	77.03	2.92	4.81	0.00	77.03	2.92	4.78
CVS		0.00	1.51	0.45	0.23	0.00	1.51	0.46	0.23
Rain				0.01				0.01	
Fog				0.01				0.01	
Sample Size	10-15 min. before crash	1	30	7.66	4.54	1	30	7.79	4.55
Avg. Speed		9.22	79.68	43.32	13.30	9.22	71.80	42.78	13.14
Min. Speed		0.00	75.83	6.19	14.04	0.00	64.75	5.67	13.28
Max. Speed		39.36	99.97	79.35	10.36	41.12	99.97	79.33	10.35
HBE Count		0	272	6.10	18.98	0	272	6.19	19.14
HAE Count		0	224	5.07	17.68	0	224	5.19	17.88
Severe Jerk Count		0	156	1.39	6.24	0	156	1.43	6.35
SD Speed		1.46	30.27	16.34	5.43	1.46	30.27	16.52	5.31
SD Acceleration		0.01	34.23	3.08	3.23	0.01	34.23	3.11	3.23
SD Jerk		0.00	41.20	2.85	4.54	0.00	41.20	2.90	4.56
CVS		0.02	1.45	0.44	0.24	0.02	1.45	0.45	0.23

Table 0-17 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=1.5 mi (UP-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	26	6.04	4.17	1	26	6.11	4.22
Avg. Speed		4.87	74.73	49.03	13.31	4.87	74.73	48.83	13.40
Min. Speed		0.00	68.89	10.53	18.69	0.00	68.89	10.40	18.67
Max. Speed		14.04	99.99	79.31	10.33	14.04	99.99	79.34	10.32
HBE Count		0	315	5.07	17.82	0	315	5.14	18.15
HAE Count		0	314	4.21	17.34	0	314	4.27	17.68
Severe Jerk Count		0	132	1.24	6.43	0	132	1.27	6.61
SD Speed		0.36	36.62	14.66	6.32	0.36	36.62	14.71	6.29
SD Acceleration		0.00	49.95	3.18	3.86	0.00	38.02	3.16	3.78
SD Jerk		0.00	60.96	2.79	5.15	0.00	60.96	2.78	5.12
CVS		0.01	1.83	0.36	0.24	0.01	1.83	0.36	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	27	6.08	4.16	1	27	6.12	4.17
Avg. Speed		9.10	73.56	49.37	13.31	9.10	73.56	49.13	13.42
Min. Speed		0.00	66.78	10.90	18.90	0.00	66.78	10.72	18.79
Max. Speed		43.07	99.98	79.65	10.12	43.07	99.98	79.56	10.13
HBE Count		0	240	4.94	16.39	0	240	4.94	16.51
HAE Count		0	260	4.11	15.93	0	260	4.10	16.05
Severe Jerk Count		0	134	1.22	6.20	0	134	1.23	6.33
SD Speed		0.48	33.67	14.47	6.29	0.48	29.51	14.53	6.26
SD Acceleration		0.00	41.68	3.05	3.64	0.00	41.68	3.05	3.67
SD Jerk		0.00	58.17	2.67	4.72	0.00	58.17	2.67	4.75
CVS		0.01	1.60	0.35	0.24	0.01	1.60	0.36	0.24

Table 0-18 Dataset for Control Def. #2 Upstream=1.5 mi, downstream=1.5 mi (DOWN-stream portion only)

Control Definition #2 (same location, random time)									
		All				Multi-Veh Only			
Variable Name	**Time Period**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
Sample Size	5-10 min. before crash	1	30	5.99	4.18	1	30	6.01	4.19
Avg. Speed		8.99	74.85	49.14	13.23	8.99	74.85	48.92	13.35
Min. Speed		0.00	66.02	10.84	18.79	0.00	66.02	10.79	18.79
Max. Speed		43.07	99.99	79.37	10.37	43.07	99.99	79.31	10.36
HBE Count		0	316	5.08	17.18	0	316	5.12	17.39
HAE Count		0	315	4.26	17.05	0	315	4.32	17.31
Severe Jerk Count		0	120	1.19	5.65	0	120	1.21	5.78
SD Speed		0.40	33.09	14.51	6.30	0.40	33.09	14.57	6.28
SD Acceleration		0.01	41.64	3.12	3.74	0.01	41.64	3.10	3.71
SD Jerk		0.00	55.21	2.71	4.86	0.00	55.21	2.69	4.86
CVS		0.01	1.33	0.35	0.24	0.01	1.33	0.36	0.24
Rain				0.01				0.01	
Fog				0.02				0.02	
Sample Size	10-15 min. before crash	1	30	6.00	4.12	1	30	6.02	4.13
Avg. Speed		9.22	77.69	49.45	13.38	9.22	77.69	49.25	13.47
Min. Speed		0.00	68.55	11.16	18.99	0.00	68.55	10.90	18.80
Max. Speed		31.79	99.99	79.40	10.27	31.79	99.99	79.39	10.25
HBE Count		0	249	4.52	14.78	0	249	4.50	14.82
HAE Count		0	253	3.75	14.54	0	253	3.73	14.59
Severe Jerk Count		0	155	1.08	5.74	0	155	1.08	5.85
SD Speed		0.58	33.85	14.30	6.33	0.58	33.85	14.36	6.28
SD Acceleration		0.00	41.67	3.07	3.81	0.00	41.67	3.07	3.83
SD Jerk		0.00	59.56	2.71	4.92	0.00	59.56	2.70	4.95
CVS		0.01	1.48	0.35	0.24	0.01	1.48	0.35	0.24

Phase I Model Results Not Shown in Section 5.2

Table 0-19 Model Results for Dataset 0.5 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #1

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	0.5 mi		**Distance Downstream**			0.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_05_510_avg_speed	-0.034	0.966	0.006	-5.611	2.010E-08	0.955	0.978
up_05_510_sample_size	0.081	1.084	0.027	3.004	0.003	1.029	1.143
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	47.12		2		6.00E-11		
Wald	43.70		2		3.00E-10		
Score (logrank)	45.75		2		1.00E-10		
AIC	1118.5						
n	1799						
n_{crash}	457						

Table 0-20 Model Results for Dataset 0.5 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #1

Multi-Veh Crashes Only, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	0.5 mi		**Distance Downstream**			0.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_05_510_avg_speed	-0.032	0.969	0.006	-5.080	3.78E-07	0.957	0.981
up_05_510_sample_size	0.092	1.096	0.027	3.355	0.001	1.039	1.157
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	43.390		2		4.00E-10		
Wald	40.450		2		2.00E-09		
Score (logrank)	42.210		2		7.00E-10		
AIC	1118.5						
n	1705						
n_{crash}	426						

Table 0-21 Model Results for Dataset 0.5 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #2

All Crashes, Controls=Def. #2 (same location, random time)							
Distance Upstream	0.5 mi		**Distance Downstream**			0.0 mi	
MODEL RESULTS							
Variable Name	**Coefficien**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_05_510_avg_speed	-0.044	0.957	0.004	-10.110	<2e-16	0.949	0.965
up_05_510_sample_size	0.064	1.066	0.024	2.640	0.008	1.017	1.117
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	158.30		2		<2e-16		
Wald	140.80		2		<2e-16		
Score (logrank)	158.90		2		<2e-16		
AIC	1255.93						
n	2221						
n_{crash}	469						

Table 0-22 Model Results for Dataset 0.5 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #2

Multi-Veh Crashes Only, Controls=Def. #2 (same location, random time)							
Distance Upstream	0.5 mi		**Distance Downstream**			0.0 mi	
MODEL RESULTS							
Variable Name	**Coefficien**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_05_510_avg_speed	-0.046	0.955	0.005	-9.974	<2e-16	0.947	0.964
up_05_510_sample_size	0.080	1.083	0.025	3.212	0.001	1.032	1.137
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	165.400		2		<2e-16		
Wald	143.900		2		<2e-16		
Score (logrank)	165.100		2		<2e-16		
AIC	1149.088						
n	2065						
n_{crash}	436						

Table 0-23 Model Results for Dataset 1.0 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #1

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.0 mi		**Distance Downstream**			0.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_10_510_avg_speed	-0.039	0.962	0.006	-6.618	0.000	0.951	0.973
up_10_1015_max_speed	0.013	1.013	0.005	2.506	0.012	1.003	1.023
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	50.35		2		1.00E-11		
Wald	47.28		2		5.00E-11		
Score (logrank)	49.50		2		2.00E-11		
AIC	1402.145						
n	2145						
n_{crash}	518						

Table 0-24 Model Results for Dataset 1.0 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #1

Multi-Veh Crashes Only, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.0 mi		**Distance Downstream**			0.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_10_510_avg_speed	-0.038	0.963	0.006	-6.316	2.68E-10	0.952	0.974
up_10_1015_max_speed	0.014	1.014	0.005	2.611	0.009	1.003	1.024
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	46.57		2		8.00E-11		
Wald	43.82		2		3.00E-11		
Score (logrank)	45.82		2		1.00E-10		
AIC	1315.88						
n	2015						
n_{crash}	482						

Table 0-25 Model Results for Dataset 1.0 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #2

All Crashes, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.0 mi		**Distance Downstream**		0.0 mi		
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_10_510_avg_speed	-0.054	0.947	0.004	-13.920	<2e-16	0.940	0.955
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	213.90		1		<2e-16		
Wald	193.60		1		<2e-16		
Score (logrank)	220.80		1		<2e-16		
AIC	1557.796						
n	2954						
n_{crash}	520						

Table 0-26 Model Results for Dataset 1.0 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #2

Multi-Veh Crashes Only, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.0 mi		**Distance Downstream**		0.0 mi		
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_10_510_avg_speed	-0.057	0.944	0.004	-14.070	<2e-16	0.937	0.952
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	223.20		1		<2e-16		
Wald	198.00		1		<2e-16		
Score (logrank)	229.20		1		<2e-16		
AIC	1424.129						
n	2746						
n_{crash}	483						

Table 0-27 Model Results for Dataset 1.5 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #1

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.5 mi		**Distance Downstream**			0.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.042	0.959	0.006	-7.162	7.92E-13	0.948	0.970
up_15_1015_max_speed	0.012	1.012	0.005	2.374	0.018	1.002	1.022
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	58.73		2		2.00E-113		
Wald	55.16		2		1.00E-12		
Score (logrank)	58.29		2		2.00E-13		
AIC	1485.361						
n	2290						
n_{crash}	540						

Table 0-28 Model Results for Dataset 1.5 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #1

Multi-Veh Crashes Only, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.5 mi		**Distance Downstream**			0.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.041	0.960	0.006	-6.871	6.37E-12	0.949	0.971
up_15_1015_max_speed	0.013	1.013	0.005	2.405	0.016	1.002	1.023
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	54.36		2		2.00E-12		
Wald	51.17		2		8.00E-12		
Score (logrank)	54.02		2		2.00E-12		
AIC	1380.925						
n	2131						
n_{crash}	498						

Table 0-29 Model Results for Dataset 1.5 mi Upstream, 0.0 mi downstream, ALL Crashes, Control Definition #2

All Crashes, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.5 mi		**Distance Downstream**		0.0 mi		
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.056	0.946	0.004	-14.660	<2e-16	0.939	0.953
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	236.10		1		<2e-16		
Wald	215.10		1		<2e-16		
Score (logrank)	246.80		1		<2e-16		
AIC	1680.482						
n	3275						
n_{crash}	540						

Table 0-30 Model Results for Dataset 1.5 mi Upstream, 0.0 mi downstream, MV Crashes, Control Definition #2

Distance Upstream	1.5 mi		**Distance Downstream**		0.0 mi		
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.060	0.942	0.004	-14.917	<2e-16	0.935	0.949
up_15_1015_max_speed	0.010	1.010	0.005	2.002	0.045	1.000	1.021
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	255.10		2		<2e-16		
Wald	224.90		2		<2e-16		
Score (logrank)	263.90		2		<2e-16		
AIC	1515.927						
n	3023						
n_{crash}	498						

Table 0-31 Model Results for Dataset 1.0 mi Upstream, 1.0 mi downstream, ALL Crashes, Control Definition #1

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.0 mi		**Distance Downstream**			1.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_10_510_avg_speed	-0.034	0.966	0.006	-5.589	2.28E-08	0.955	0.978
dn_10_510_sample_size	0.054	1.056	0.020	2.757	0.006	1.016	1.097
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	47.11		2		6.00E-11		
Wald	44.02		2		3.00E-10		
Score (logrank)	46.10		2		1.00E-10		
AIC	1281.54						
n	1958						
n_{crash}	485						

Table 0-32 Model Results for Dataset 1.0 mi Upstream, 1.0 mi downstream, MV Crashes, Control Definition #1

Multi-Veh Crashes Only, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.0 mi		**Distance Downstream**			1.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_10_510_avg_speed	-0.033	0.967	0.006	-5.325	1.01E-07	0.955	0.979
dn_10_510_sample_size	0.055	1.057	0.020	2.736	0.006	1.016	1.099
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	43.61		2		3.00E-10		
Wald	40.84		2		1.00E-09		
Score (logrank)	42.71		2		5.00E-10		
AIC	1212.557						
n	1853						
n_{crash}	456						

Table 0-33 Model Results for Dataset 1.0 mi Upstream, 1.0 mi downstream, ALL Crashes, Control Definition #2

All Crashes, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.0 mi		**Distance Downstream**		1.0 mi		
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
dn_10_510_avg_speed	-0.052	0.949	0.004	-12.605	<2e-16	0.941	0.957
dn_10_1015_hbe_count	0.007	1.007	0.003	2.154	0.031	1.001	1.014
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	187.10		2		<2e-16		
Wald	165.40		2		<2e-16		
Score (logrank)	189.00		2		<2e-16		
AIC	1375.343						
n	2513						
n_{crash}	491						

Table 0-34 Model Results for Dataset 1.0 mi Upstream, 1.0 mi downstream, MV Crashes, Control Definition #2

Multi-Veh Crashes Only, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.0 mi		**Distance Downstream**		1.0 mi		
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
dn_10_510_avg_speed	-0.055	0.946	0.004	-12.694	<2e-16	0.939	0.955
dn_10_1015_hbe_count	0.008	1.008	0.003	2.223	0.026	1.001	1.015
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	192.40		2		<2e-16		
Wald	167.30		2		<2e-16		
Score (logrank)	193.30		2		<2e-16		
AIC	1274.398						
n	2360						
n_{crash}	460						

Table 0-35 Model Results for Dataset 1.5 mi Upstream, 1.0 mi downstream, ALL Crashes, Control Definition #1

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.5 mi		**Distance Downstream**			1.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.037	0.964	0.006	-6.027	1.67E-09	0.952	0.975
dn_10_510_sample_size	0.053	1.054	0.020	2.708	0.007	1.015	1.096
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	53.78		2		2.00E-12		
Wald	49.98		2		1.00E-11		
Score (logrank)	52.65		2		4.00E-12		
AIC	1330.167						
n	2045						
n_{crash}	500						

Table 0-36 Model Results for Dataset 1.5 mi Upstream, 1.0 mi downstream, MV Crashes, Control Definition #1

Multi-Veh Crashes Only, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.5 mi		**Distance Downstream**			1.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.037	0.964	0.006	-5.785	7.25E-09	0.952	0.976
dn_10_510_sample_size	0.055	1.056	0.020	2.721	0.007	1.015	1.098
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	50.50		2		1.00E-11		
Wald	46.97		2		5.00E-11		
Score (logrank)	49.44		2		2.00E-11		
AIC	1253.947						
n	1929						
n_{crash}	469						

Table 0-37 Model Results for Dataset 1.5 mi Upstream, 1.0 mi downstream, ALL Crashes, Control Definition #2

Distance Upstream	1.5 mi		**Distance Downstream**			1.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.054	0.947	0.004	-13.310	<2e-16	0.940	0.955
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	197.40		1		<2e-16		
Wald	177.10		1		<2e-16		
Score (logrank)	201.30		1		<2e-16		
AIC	1443.934						
n	2681						
n_{crash}	504						

Table 0-38 Model Results for Dataset 1.5 mi Upstream, 1.0 mi downstream, MV Crashes, Control Definition #2

Multi-Veh Crashes Only, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.5 mi		**Distance Downstream**			1.0 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.057	0.945	0.004	-13.370	<2e-16	0.937	0.953
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	202.50		1		<2e-16		
Wald	178.80		1		<2e-16		
Score (logrank)	205.60		1		<2e-16		
AIC	1340.06						
n	2521						
n_{crash}	473						

Table 0-39 Model Results for Dataset 1.5 mi Upstream, 1.5 mi downstream, ALL Crashes, Control Definition #1

All Crashes, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.5 mi		**Distance Downstream**			1.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.039	0.961	0.006	-6.490	0.000	0.950	0.973
dn_15_1015_sample_size	0.036	1.036	0.017	2.091	0.037	1.002	1.071
up_15_1015_max_speed	0.011	1.011	0.005	2.065	0.039	1.001	1.022
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	60.68		3		4.00E-13		
Wald	56.31		3		4.00E-12		
Score (logrank)	59.61		3		7.00E-13		
AIC	1387.707						
n	2142						
n_{crash}	515						

Table 0-40 Model Results for Dataset 1.5 mi Upstream, 1.5 mi downstream, MV Crashes, Control Definition #1

Multi-Veh Crashes Only, Controls=Def. #1 (same time, day of week, and location)							
Distance Upstream	1.5 mi		**Distance Downstream**			1.5 mi	
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.037	0.964	0.006	-5.883	0.000	0.952	0.976
dn_15_510_sample_size	0.047	1.048	0.017	2.741	0.006	1.014	1.084
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	53.16		2		3.00E-12		
Wald	49.50		2		2.00E-11		
Score (logrank)	52.09		2		5.00E-12		
AIC	1305.296						
n	2014						
n_{crash}	481						

Table 0-41 Model Results for Dataset 1.5 mi Upstream, 1.5 mi downstream, ALL Crashes, Control Definition #2

All Crashes, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.5 mi		**Distance Downstream**		1.5 mi		
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
dn_15_510_avg_speed	-0.052	0.949	0.004	-11.797	<2e-16	0.941	0.958
dn_15_1015_min_speed	-0.008	0.992	0.004	-2.138	0.033	0.985	0.999
dn_15_510_max_speed	0.010	1.010	0.005	2.027	0.043	1.000	1.020
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	217.90		3		<2e-16		
Wald	188.70		3		<2e-16		
Score (logrank)	217.20		3		<2e-16		
AIC	1528.027						
n	2887						
n_{crash}	520						

Table 0-42 Model Results for Dataset 1.5 mi Upstream, 1.5 mi downstream, MV Crashes, Control Definition #2

Multi-Veh Crashes Only, Controls=Def. #2 (same location, random time)							
Distance Upstream	1.5 mi		**Distance Downstream**		1.5 mi		
MODEL RESULTS							
Variable Name	**Coefficient**	**exp(coeff)**	**SE(coeff)**	**Z stat**	**p-val**	**Lower 95%**	**Upper 95%**
up_15_510_avg_speed	-0.052	0.950	0.004	-11.755	<2e-16	0.942	0.958
dn_15_1015_min_speed	-0.012	0.988	0.004	-2.987	0.003	0.980	0.996
MODEL FIT							
Test Name	**Test Statistic**		**DF**		**p-val**		
Likelihood Ratio	224.50		2		<2e-16		
Wald	190.90		2		<2e-16		
Score (logrank)	222.00		2		<2e-16		
AIC	1406.331						
n	2702						
n_{crash}	485						

DERIVATION OF CI-PIS

Here, derivations for the 95% CI for M and 95% PI Y are shown for each of the five types of mixed-Poisson models considered in the study. Additionally, the derivation for the Var(M) is shown. In all cases, the following is assumed:

$$m = \mu v$$

Where, the distribution of ν is the mixture distribution of interest.

Negative Binomial (NB) Model

Variance of M

$$\begin{aligned}
&Var(M) = Var(\mu * \upsilon) \\
&= E(\mu^2\upsilon^2) - E(\mu * \upsilon)^2 \\
&= E(\mu^2)E(\upsilon^2) - E(\mu)^2E(\upsilon)^2 \\
&= [Var(\mu) + E(\mu)^2] * [Var(\upsilon) + E(\upsilon)^2] - E(\mu)^2E(\upsilon)^2 \\
&= [\sigma_0^2 + \mu_0^2] * \left[\frac{1}{\varphi} + 1\right] - \mu_0^2 * 1 \\
&= \alpha * (\sigma_0^2 + \mu_0^2) + \sigma_0^2 \ (where, \alpha = \frac{1}{\varphi} = dispersion\, parameter)
\end{aligned}$$

95% CI for m

$$\begin{aligned}
&\hat{\mu} \pm 1.96 * \sqrt{Var(m)} \\
&\hat{\mu} \pm 1.96 * \sqrt{\hat{\alpha}(\hat{\sigma}_0^2 + \hat{\mu}_0^2) + \hat{\sigma}_0^2} \\
&\hat{\mu} \pm 1.96 * \sqrt{\hat{\alpha}(\hat{\mu}^2 Var(\hat{\eta}) + \hat{\mu}^2) + \hat{\mu}^2 Var(\hat{\eta})} \\
&\hat{\mu} \pm 1.96 * \sqrt{\hat{\mu}^2[\hat{\alpha}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})]} \\
&\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2[\hat{\alpha}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})]}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2[\hat{\alpha}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})]}\right]
\end{aligned}$$

95% PI for y

$$\begin{aligned}
&\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + Var(m)} \right\rfloor\right] \\
&\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2[\hat{\alpha}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})]} \right\rfloor\right]
\end{aligned}$$

Poisson-Inverse-Gaussian (PIG) Model

Variance of M

$$\begin{aligned}
&Var(M) = Var(\mu * \upsilon) \\
&= E(\mu^2\upsilon^2) - E(\mu * \upsilon)^2
\end{aligned}$$

$$= E(\mu^2)E(\upsilon^2) - E(\mu)^2E(\upsilon)^2$$
$$= [Var(\mu) + E(\mu)^2] * [Var(\upsilon) + E(\upsilon)^2] - E(\mu)^2E(\upsilon)^2$$
$$= [\sigma_0^2 + \mu_0^2] * \left[\frac{1}{\lambda} + 1\right] - \mu_0^2 * 1 \; (since\; Var[\nu] = \frac{\mu^3}{\lambda} = \frac{1}{\lambda})$$
$$= \frac{1}{\lambda} * (\sigma_0^2 + \mu_0^2) + \sigma_0^2$$

95% CI for m

$$\hat{\mu} \pm 1.96 * \sqrt{Var(m)}$$
$$\hat{\mu} \pm 1.96 * \sqrt{\tfrac{1}{\hat{\lambda}}(\hat{\sigma}_0^2 + \hat{\mu}_0^2) + \hat{\sigma}_0^2}$$
$$\hat{\mu} \pm 1.96 * \sqrt{\tfrac{1}{\hat{\lambda}}(\hat{\mu}^2 Var(\hat{\eta}) + \hat{\mu}^2) + \hat{\mu}^2 Var(\hat{\eta})}$$
$$\hat{\mu} \pm 1.96 * \sqrt{\hat{\mu}^2 \left[\tfrac{1}{\hat{\lambda}}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})\right]}$$
$$\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2 \left[\tfrac{1}{\hat{\lambda}}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})\right]}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2 \left[\tfrac{1}{\hat{\lambda}}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})\right]}\right]$$

95% PI for y

$$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + Var(m)} \right\rfloor\right]$$
$$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2 \left[\tfrac{1}{\hat{\lambda}}(Var(\hat{\eta}) + 1) + Var(\hat{\eta})\right]} \right\rfloor\right]$$

Sichel (SI) Model

Variance of M

$$Var(M) = Var(\mu * \upsilon)$$
$$= E(\mu^2\upsilon^2) - E(\mu * \upsilon)^2$$
$$= E(\mu^2)E(\upsilon^2) - E(\mu)^2E(\upsilon)^2$$
$$= [Var(\mu) + E(\mu)^2] * [Var(\upsilon) + E(\upsilon)^2] - E(\mu)^2E(\upsilon)^2$$
$$= [\sigma_0^2 + \mu_0^2] * \left[\left(\frac{2\sigma(\nu_{GIG} + 1)}{c} + \frac{1}{c^2} - 1\right) + 1\right] - \mu_0^2 * 1$$
$$= [\sigma_0^2 + \mu_0^2] * \left(\frac{2\sigma(\nu_{GIG} + 1)}{c} + \frac{1}{c^2}\right) - \mu_0^2$$

95% CI for m

$$\hat{\mu} \pm 1.96 * \sqrt{Var(m)}$$
$$\hat{\mu} \pm 1.96 * \sqrt{[\hat{\sigma}_0^2 + \hat{\mu}_0^2] * \left(\tfrac{2\hat{\sigma}_{GIG}(\hat{\nu}+1)}{c} + \tfrac{1}{c^2}\right) - \hat{\mu}_0^2}$$
$$\hat{\mu} \pm 1.96 * \sqrt{[\hat{\mu}^2 Var(\hat{\eta}) + \hat{\mu}^2] * \left(\tfrac{2\hat{\sigma}_{GIG}(\hat{\nu}+1)}{c} + \tfrac{1}{c^2}\right) - \hat{\mu}^2}$$

$$\hat{\mu} \pm 1.96 * \sqrt{\hat{\mu}^2 \left\{[Var(\hat{\eta}) + 1] * \left(\frac{2\hat{\sigma}_{GIG}(\hat{\nu}+1)}{c} + \frac{1}{c^2}\right) - 1\right\}}$$

$$\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2 \left\{[Var(\hat{\eta}) + 1] * \left(\frac{2\hat{\sigma}_{GIG}(\hat{\nu}+1)}{c} + \frac{1}{c^2}\right) - 1\right\}}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2 \left\{[Var(\hat{\eta}) + 1] * \left(\frac{2\hat{\sigma}_{GIG}(\hat{\nu}+1)}{c} + \frac{1}{c^2}\right) - 1\right\}}\right]$$

95% PI for y

$$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + Var(m)} \right\rfloor\right]$$

$$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2 \left\{[Var(\hat{\eta}) + 1] * \left(\frac{2\hat{\sigma}_{GIG}(\hat{\nu}+1)}{c} + \frac{1}{c^2}\right) - 1\right\}} \right\rfloor\right]$$

Poisson-Lognormal (PLN) Model
Variance of M

$$\begin{aligned}
&Var(M) = Var(\mu * \upsilon) \\
&= E(\mu^2\upsilon^2) - E(\mu * \upsilon)^2 \\
&= E(\mu^2)E(\upsilon^2) - E(\mu)^2E(\upsilon)^2 \\
&= [Var(\mu) + E(\mu)^2] * [Var(\upsilon) + E(\upsilon)^2] - \ E(\mu)^2E(\upsilon)^2 \\
&= [\sigma_0^2 + \mu_0^2] * \left[\left(e^{\sigma_{LN}^2} - 1\right) * e^{2\left(-\frac{\sigma_{LN}^2}{2}\right) + \sigma_{LN}^2} + 1\right] - \mu_0^2 * 1 \\
&= [\sigma_0^2 + \mu_0^2] * \left[e^{\sigma_{LN}^2}\right] - \mu_0^2 \\
&= e^{\sigma_{LN}^2} * [\sigma_0^2 + \mu_0^2] - \mu_0^2
\end{aligned}$$

95% CI for m

$$\hat{\mu} \pm 1.96 * \sqrt{Var(m)}$$
$$\hat{\mu} \pm 1.96 * \sqrt{e^{\hat{\sigma}_{LN}^2} * [\hat{\sigma}_0^2 + \hat{\mu}_0^2] - \hat{\mu}_0^2}$$
$$\hat{\mu} \pm 1.96 * \sqrt{e^{\hat{\sigma}_{LN}^2} * [\hat{\mu}^2 Var(\hat{\eta}) + \hat{\mu}^2] - \hat{\mu}^2}$$
$$\hat{\mu} \pm 1.96 * \sqrt{\hat{\mu}^2[e^{\hat{\sigma}_{LN}^2}(Var(\hat{\eta}) + 1) - 1]}$$

$$\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2[e^{\hat{\sigma}_{LN}^2}(Var(\hat{\eta}) + 1) - 1]}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2[e^{\hat{\sigma}_{LN}^2}(Var(\hat{\eta}) + 1) - 1]}\ \right]$$

95% PI for y

$$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + Var(m)} \right\rfloor\right]$$

$$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2[e^{\hat{\sigma}_{LN}^2}(Var(\hat{\eta}) + 1) - 1]} \right\rfloor\right]$$

Poisson-Weibull (PW) Model

Variance of M

$$\begin{aligned}
&Var(M) = Var(\mu * v) \\
&= E(\mu^2 v^2) - E(\mu * v)^2 \\
&= E(\mu^2)E(v^2) - E(\mu)^2 E(v)^2 \\
&= [Var(\mu) + E(\mu)^2] * [Var(v) + E(v)^2] - E(\mu)^2 E(v)^2 \\
&= [\sigma_0^2 + \mu_0^2] * \left[\frac{1}{\mu_{WEI}^{2/\sigma}}\left[\Gamma\left(\frac{2}{\sigma} + 1\right) - \left(\Gamma\left(\frac{1}{\sigma} + 1\right)\right)^2\right] + 1^2\right] - \mu_0^2 \\
&= [\sigma_0^2 + \mu_0^2] * \left[\frac{\Gamma\left(\frac{2}{\sigma}+1\right)}{\left(\Gamma\left(\frac{1}{\sigma}+1\right)\right)^2} - 1 + 1\right] - \mu_0^2 \\
&= [\sigma_0^2 + \mu_0^2] * \left[\frac{\Gamma\left(\frac{2}{\sigma}+1\right)}{\left(\Gamma\left(\frac{1}{\sigma}+1\right)\right)^2}\right] - \mu_0^2
\end{aligned}$$

95% CI for m

$$\hat{\mu} \pm 1.96 * \sqrt{Var(m)}$$

$$\hat{\mu}_0 \pm 1.96 * \sqrt{[\hat{\sigma}_0^2 + \hat{\mu}_0^2] * \left[\frac{\Gamma\left(\frac{2}{\sigma}+1\right)}{\left(\Gamma\left(\frac{1}{\sigma}+1\right)\right)^2}\right] - \hat{\mu}_0^2}$$

$$\hat{\mu} \pm 1.96 * \sqrt{[\hat{\mu}^2 Var(\hat{\eta}) + \hat{\mu}^2] * \left[\frac{\Gamma\left(\frac{2}{\hat{\sigma}}+1\right)}{\left(\Gamma\left(\frac{1}{\hat{\sigma}}+1\right)\right)^2}\right] - \hat{\mu}^2}$$

$$\hat{\mu} \pm 1.96 * \sqrt{\hat{\mu}^2\left([Var(\hat{\eta}) + 1] * \left[\frac{\Gamma\left(\frac{2}{\hat{\sigma}}+1\right)}{\left(\Gamma\left(\frac{1}{\hat{\sigma}}+1\right)\right)^2}\right] - 1\right)}$$

$$\left[\max\left(0, \hat{\mu} - 1.96 * \sqrt{\hat{\mu}^2\left([Var(\hat{\eta}) + 1] * \left[\frac{\Gamma\left(\frac{2}{\hat{\sigma}}+1\right)}{\left(\Gamma\left(\frac{1}{\hat{\sigma}}+1\right)\right)^2}\right] - 1\right)}\right), \hat{\mu} + 1.96 * \sqrt{\hat{\mu}^2\left([Var(\hat{\eta}) + 1] * \left[\frac{\Gamma\left(\frac{2}{\hat{\sigma}}+1\right)}{\left(\Gamma\left(\frac{1}{\hat{\sigma}}+1\right)\right)^2}\right] - 1\right)}\right]$$

95% PI for y

$$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + Var(m)} \right\rfloor\right]$$

$$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu} + \hat{\mu}^2\left([Var(\hat{\eta}) + 1] * \left[\frac{\Gamma\left(\frac{2}{\hat{\sigma}}+1\right)}{\left(\Gamma\left(\frac{1}{\hat{\sigma}}+1\right)\right)^2}\right] - 1\right)} \right\rfloor\right]$$