APPLICATION OF BIG DATA IN TRANSPORTATION SAFETY ANALYSIS USING STATISTICAL AND DEEP LEARNING METHODS

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Ramin Arvin
May 2020

Copyright © 2020 by Ramin Arvin

All rights reserved.

This dissertation is dedicated to my wife, my parents, brothers, and beloved family who constantly encouraged me to pursue my dreams and finish my dissertation

ACKNOWLEDGMENT

First and most of all, I am thankful to God for His blessings during my endeavors. Without His continuous guidance, completing my dissertation would have been impossible and my dream of earning a doctoral degree would have remained mere dream.

I would like to thank everyone who has directly or indirectly supported, encouraged and guided me in my quest for knowledge. I wish to express my profound appreciation to Dr. Asad J. Khattak. His timely and powerful mentorship in these four years guided and supported me in the completion of my dissertation and research study. It was his leadership and mentorship that enabled me to undertake and contribute to research program at the University of Tennessee. It has been my honor working under his supervision and I always be proud of my association with you. A debt of gratitude is extended to my dissertation committee, Dr. Hamparsum Bozdogan, Dr. Hairong Qi, and Dr. Candace Brakewood. I am honored to have their tremendous guidance, help, and support in my dissertation. Also, special thanks to Dr. Bozdogan for supervising my M.S. statistics degree.

To my co-authors, Dr. Asad Khattak, Dr. Mohsen Kamrani, Dr. Jecheline Rios-Torres, Dr. Alexandra Boggs, Dr. Christopher Cherry, Dr. Lee Han, Dr. Hairong Qi, Dr. Asad Hoque, Dr. Amin Mohammadi, Nima Hoseinzadeh, Numan Ahmed, Sevin Mohammadi, Iman Mahdinia, Amin Mohamadnazar, and Salman Ahmed thank you for your valuable knowledge and insights in my research. I also want to thank all my lab mates for their continuous support and memorable experiences. Special thanks to the Department of

Civil and Environmental Engineering faculty and staff. My work has been greatly supported through National Science Foundation, Collaborative Sciences Center for Road Safety, US Department of Transportation, and Tennessee Department of Transportation.

My acknowledgement would be incomplete without thanking my biggest pillar of unconditional support and guidance, my wife, Razieh Kaviani Baghbaderani for her endless support in my life and career, my parents and family for their valuable supports and precious love for supporting me throughout these years. My Family's involvement, concern, forbearing, and understanding has been the prime support for me throughout these years and without their help I could not reach this stage in my life.

ABSTRACT

The emergence of new sensors and data sources provides large scale high-resolution big data from instantaneous vehicular movements, driver decision and states, surrounding environment, roadway characteristics, weather condition, etc. Such a big data can be served to expand our understanding regarding the current state of the transportation and help us to proactively evaluate and monitor the system performance. The key idea behind this dissertation is to identify the moments and locations where drivers are exhibiting different behavior comparing to the normal behavior. The concept of driving volatility is utilized which quantifies deviation from normal driving in terms of variations in speed, acceleration/deceleration, and vehicular jerk. This idea is utilized to explore the association of volatility in different hierarchies of transportation system, i.e.: 1) Instance level; 2) Event level; 3) Driver level; 4) Intersection level; and 5) Network level. In summary, the main contribution of this dissertation is exploring the association of variations in driving behavior in terms of driving volatility at different levels by harnessing big data generated from emerging data sources under real-world condition, which is applicable to the intelligent transportation systems and smart cities. By analyzing real-world crashes/near-crashes and predicting occurrence of extreme event, proactive warnings and feedback can be generated to warn drivers and adjacent vehicles regarding potential hazard. Furthermore, the results of this study help agencies to proactively monitor and evaluate safety performance of the network and identify locations where crashes are waiting to happen. The main objective of this dissertation is to integrate big data generated from emerging sources into safety analysis by considering different levels in the system. To this end, several data sources including Connected Vehicles data (with

more than 2.2 billion seconds of observations), naturalistic driving data (with more than 2 million seconds of observations from vehicular kinematics and driver behavior), conventional data on roadway factors and crash data are integrated.

TABLE OF CONTENTS

Abstract vi

Chapter 1 : Introduction 1

Summary 2

Problem justification 9

Literature review, gaps and contributions 11

Monitoring safety performance of the network 11

Analyzing crash frequency and types at intersections 15

The role of instability in driving on crash intensity 18

Association of driving impairment/distraction on crash risk 21

Real-time crash prediction 24

Methodological framework 27

Chapter 2 : Harnessing big data generated by connected vehicles to proactively monitor safety performance of the network: Application of Geographically Weighted Negative Binomial Regression 30

Abstract 31

Introduction 32

Data 34

Methodology 36

Conceptual framework 37

Measures of driving volatility 40

Calculation of volatility measures 43

Modeling Approach 47
Measures of Goodness of Fit 53
k-means clustering 54
Gaussian Mixture Model 55
Results 56
Descriptive Statistics 56
Concept illustration 59
Modeling Results 62
Hotspot identification 68
Limitations 72
Conclusion and future research 73
Chapter 3 : How Instantaneous Driving Behavior Contributes to Crashes at Intersections: Extracting Useful Information from Connected Vehicle Message Data 76
Abstract 77
Introduction 78
Methodology 81
Modeling Approach 81
Location-based volatility, measures and calculation 88
Measures of Goodness of Fit 90
Data 92
Results 97
Descriptive Statistics 97
Modeling Results 99
Limitations and future work 117

Conclusion 119
Chapter 4 : Examining the role of pre-crash driving volatility in contributing to crash intensity 122
Abstract 123
Introduction 124
Methodology 126
Modeling framework 127
Fixed parameter modeling of pathways 129
Random parameter modeling of pathways 131
Quantifying pathway by marginal effects 132
Measures of Volatility 134
Data 137
Exclusion of Evasive Maneuvers 139
Results 140
Descriptive Statistics 140
Modeling Results 142
Path analysis of driving volatility and crash intensity model 152
Limitations 155
Conclusion and future research 155
Chapter 5 : Driving impairments and Duration of distractions: Assessing Crash Risk by Harnessing Microscopic Naturalistic Driving Data 158
Abstract 159
Introduction 160
Data 163

Data description 163
Data Pre-processing 164
Methodology 167
Fixed parameter logit model 169
Random parameter (mixed) logit model 170
Results 171
Descriptive Statistics 171
Duration of distracted driving 174
Modeling Results 176
Discussion 179
Limitations 183
Conclusions 183
Chapter 6 : Real-time crash prediction through unified analysis of driver and vehicle volatilities: Application of 1D-Convolutional Neural Network - Long Short-Term Memory 186
Abstract 187
Introduction 188
Data description and pre-processing 191
Data description 191
Exclusion of Evasive Maneuvers 192
Measures of driving volatility 193
Concept illustration and descriptive statistics 196
Technical Approaches 199
Conceptual Framework 199

Problem formulation 200
Experimental evaluation 208
Training procedure 208
Evaluation metrics 211
Comparative results 212
Importance of volatility and distraction profile 214
Conclusion 215
Chapter 7 : Conclusion and implications 217
Contribution 224
Implications 225
References 228
VITA 262

LIST OF TABLES

Table 2.1 Algorithm for calculation of temporal driving volatility 46

Table 2.2 Descriptive Statistics (N=3007 grids) 58

Table 2.3 Measures of goodness of fit for the fitted model..... 63

Table 2.4 Modeling results of crash frequency model 66

Table 2.5 Descriptive statistics of volatility measures for each cluster 71

Table 3.1 Functions of volatility 90

Table 3.2 Descriptive Statistics of dependents and key variables (N=167)..... 98

Table 3.3 Measures of goodness of fit for the fitted model..... 101

Table 3.4 Modeling results for rear-end crashes (N=167 intersections)..... 105

Table 3.5 Modeling results for sideswipe crashes (N=167 intersections)..... 106

Table 3.6 Modeling results for angle crashes (N=167 intersections)..... 113

Table 3.7 Modeling results for head-on crashes (N=167 intersections) 114

Table 4.1 Descriptive statistics of variables 141

Table 4.2 Correlation of volatility measures with the crash intensity 144

Table 4.3 Tobit Modeling results for speed volatility (as dependent variable) 147

Table 4.4 Tobit Modeling results for deceleration volatility (as dependent variable) 148

Table 4.5 Modeling results for crash intensity 151

Table 4.6 Total marginal effect of random parameter model on crash intensity (in percent) 154

Table 5.1 Definition and list of recoded variables and their final Categories 165

Table 5.2 Descriptive statistics of the driver, vehicle, and roadway/environmental factors 173
Table 5.3 Descriptive statistics of the duration of distraction for 15 seconds of observations 175
Table 5.4 Fixed and random parameter modeling results 178
Table 6.1 Descriptive statistics of the baseline and critical events 198
Table 6.2 Structure of the LSTM model 206
Table 6.3 Structure of the CNN-LSTM model 208
Table 6.4 Test and train datasets 210
Table 6.5 Models performance evaluation 214
Table 6.6 Evaluation of feature importance in the 1DCNN-LSTM model 215

LIST OF FIGURES

Figure 1.1 Outline of the dissertation 8

Figure 1.2 Framework of the dissertation 29

Figure 2.1 Study area and generated map by connected vehicles 36

Figure 2.2 Conceptual framework 38

Figure 2.3 Workflow of the paper 40

Figure 2.4 Calculation of temporal driving volatility 45

Figure 2.5 Visualization of relationship between volatility and frequency of crashes (red points indicates higher volatility/crash frequency) 61

Figure 2.6 Estimated coefficient of the GWNBR model 67

Figure 2.7 Elbow method to find the optimal number of clusters 69

Figure 2.8 Association of volatility measures in low, medium and high volatility clusters 70

Figure 2.9 Crash frequency vs L2-Speed-V_f volatility for high-volatility cluster 72

Figure 3.1 Created map from BSM data (left), Data preparation framework (right) 95

Figure 3.2 Histogram of lateral acceleration 96

Figure 3.3 Location of selected intersections (N=167) 96

Figure 3.4 Local estimation of *Speed* – C_v (top), *Speed-2* S_{dev} (middle), and *Acceleration*$_x$ - Q_{cv} (bottom) on rear-end crashes 107

Figure 3.5 Local estimation of *Speed-2*$\boldsymbol{Sdev}$ *(top), Acceleration*$_x$ – C_v *(middle), and AccDec*$_y$ – $2S_{dev}$ (bottom) on sideswipe crashes 110

Figure 3.6 Local estimation of *Deceleration*$_y$ - C_v in angle crashes 112

Figure 3.7 Local estimation of $Acceleration_y$ – Q_{cv} (top), and $Deceleration_y$ – C_v in head-on crashes 117
Figure 4.1 Conceptual framework for the pathways modeled 128
Figure 4.2 Speed and acceleration profile of a randomly chosen crash 139
Figure 4.3 Pathway diagram of the model 144
Figure 5.1 Speed and acceleration profile of a randomly chosen crash 167
Figure 5.2 Conceptual framework of the study 168
Figure 5.3 Histogram of duration of key distraction types for extreme events 176
Figure 5.4 Probability of extreme event occurrence for different types of distraction 180
Figure 6.1 Speed and acceleration profile of a randomly chosen crash 193
Figure 6.2 Temporal speed volatility measure calculation 196
Figure 6.3 Boxplot of distracted driving, speed, longitudinal and lateral volatilities for the baseline and critical events 197
Figure 6.4 Conceptual framework of the study 199
Figure 6.5 Representation of 1D-CNN network used in this study 202
Figure 6.6 Illustration of the structure of the LSTM neural network 205
Figure 6.7 Structure of the CNN-LSTM model 207
Figure 6.8 Accuracy and loss for the training and test datasets 210

CHAPTER 1 : Introduction

Summary

The emergence of new sensors and data sources provides large scale high-resolution big data from instantaneous vehicular movements, driver decision and states, roadway characteristics, weather condition, etc. With mandating automobile companies to install communication equipment, enormous data will be available ranging from microscopic driver decisions to instantaneous traffic flow condition. Such a big data can be served to expand our understanding regarding the current state of the transportation and help us to proactively evaluate and monitor the system performance. Transportation safety is one of the main challenges with more than 37 thousand fatalities and more than 2 million injuries across the United States. The main question that might arises is whether the new large-scale data can be incorporated into safety analysis.

As such, this research attempts to answer this question from several perspectives. From the ***big data perspective***, this research developed a framework to pre-process unstructured raw data, assemble, extract additional engineering features (e.g. driving volatility, traffic exposure, roadway geometry features, traffic flow condition) and integrate this information with traditional transportation data sources. The data contains rich information on instantaneous driving behavior in naturalistic and connected environments, roadway/environmental characteristics, driver state, and biometrics (i.e. distraction). From the ***conceptual perspective***, this dissertation developed the concepts of temporal driving volatility and unintentional volatility in order to quantify variations in each driving instance. Furthermore, this study extends the concept of location-based

driving volatility by developing several volatility measures and incorporate lateral movements into analysis. From the ***methodological perspective***, this study is employed several innovative methods including heterogeneity-based simulation-assisted statistical models, Geographically Weighted Regression analysis, Machine Learning, and Deep Learning methods to model association of extracted features from the raw big data with crash risk (in terms of frequency of crashes, type of crashes, probability of occurrence, and crash propensity).

The key idea behind this dissertation is to identify the moments and locations where drivers exhibit different behavior comparing to the normal behavior. The concept of driving volatility is utilized which quantifies deviation from normal driving in terms of variations in speed, acceleration/deceleration, and vehicular jerk. This idea is utilized to explore the association of volatility in different hierarchies of transportation system, i.e.: 1) Instance level; 2) Event level; 3) Driver level; 4) Intersection level; and 5) Network level.

At the ***instance level***, the concept of temporal driving volatility is developed which quantify variations in each instance of driving and applied to the Connected Vehicle data and NDS data. This concept will help us to identify instances when drivers exhibit abnormal behavior and explore the factors associated with this behavior to reduce it. By matching micro information on driving behavior with roadway/environmental factors, driver state and biometrics, we explored their association with crash risk. This dissertation characterizes the probability of crash occurrence in real-time by applying rigorous deep learning methods. Referring to ***event level***, this study analyzed the association 15 seconds of pre-crash driving volatility with crash intensity, simultaneously modeling

association of driver state and roadway/environmental factors with event volatility itself. At the ***driver level***, the study explores all trips taken by each individual driver commuting in the study area and quantifies longitudinal and lateral driving volatilities. These measures are utilized to group system users to calm, normal, and aggressive drivers. The driver level volatility has several applications in real-life such as Advanced Driving Assistance Systems, scoring driver risk for insurance companies, and safety-based route guidance. At the ***network level***, the dissertation incorporates the concepts of location-based and temporal driving volatility to explore the association of variations in longitudinal and lateral vehicular movements with crash frequency and type at the intersections and network level. Large scale data from the Safety Pilot Model Deployment study is utilized and processed more than 2.2 billion of BSM observations to calculate several volatility measures. Also, additional features on traffic flow and roadway geometry are extracted from the CV data. This information is fused with crash and traditional data (e.g. roadway geometry, traffic volume) and association of driver behavior (in terms of driving volatility) with crash frequency and type is explored. The results are utilized to proactively identify hotspot locations in the network where driving volatility is high, while crash frequency is low and crashes are waiting to happen.

In summary, the main contribution of this dissertation is exploring the association of variations in driving behavior in terms of driving volatility at different levels by harnessing big data generated from emerging data sources under real-world condition, which is applicable to the intelligent transportation systems and smart cities. By analyzing real-world crashes/near-crashes and predicting occurrence of extreme event, proactive warnings and feedback can be generated to warn drivers and adjacent vehicles regarding

potential hazard. Furthermore, the results of this study help agencies to proactively monitor and evaluate safety performance of the network and identify locations where crashes are waiting to happen.

The analyses under this dissertation led to the following articles:

1. Arvin, R., Khattak, A. (2019). *Harnessing big data generated by connected vehicles to proactively monitor safety performance of the network: Application of Geographically Weighted Negative Binomial Regression*.
    - *Peer-reviewed conference paper:* Presented at the 98th Transportation Research Board Annual Meeting 2020, Washington DC.
    - *Journal article:* Under review in Accident Analysis and Prevention

2. Arvin, R., Kamrani, M., & Khattak, A. J. (2019). *How instantaneous driving behavior contributes to crashes at intersections: extracting useful information from connected vehicle message data.*
    - *Journal article:* Published in *Accident Analysis & Prevention*.
    - *Peer-reviewed conference paper:* Presented at the 97th Transportation Research Board Annual Meeting 2019, Washington DC.

3. Hosseinzadeh, N., Arvin, R.[1], Khattak, A., & Han, L. (2019). *Integrating safety and mobility for pathfinding using big data generated by connected vehicles.*
    - *Journal article:* Published in Journal of Intelligent Transportation Systems

[1] The contribution of the first and second author is equal

- *Peer-reviewed conference paper:* Presented at the 97th Transportation Research Board Annual Meeting 2018, Washington DC.

4. Arvin, R., Kamrani, M., & Khattak, A. J. (2019). *The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data.*
    - *Journal article:* Published in *Accident Analysis & Prevention*.
    - *Peer-reviewed conference paper:* Presented at the 97th Transportation Research Board Annual Meeting 2018, Washington DC.

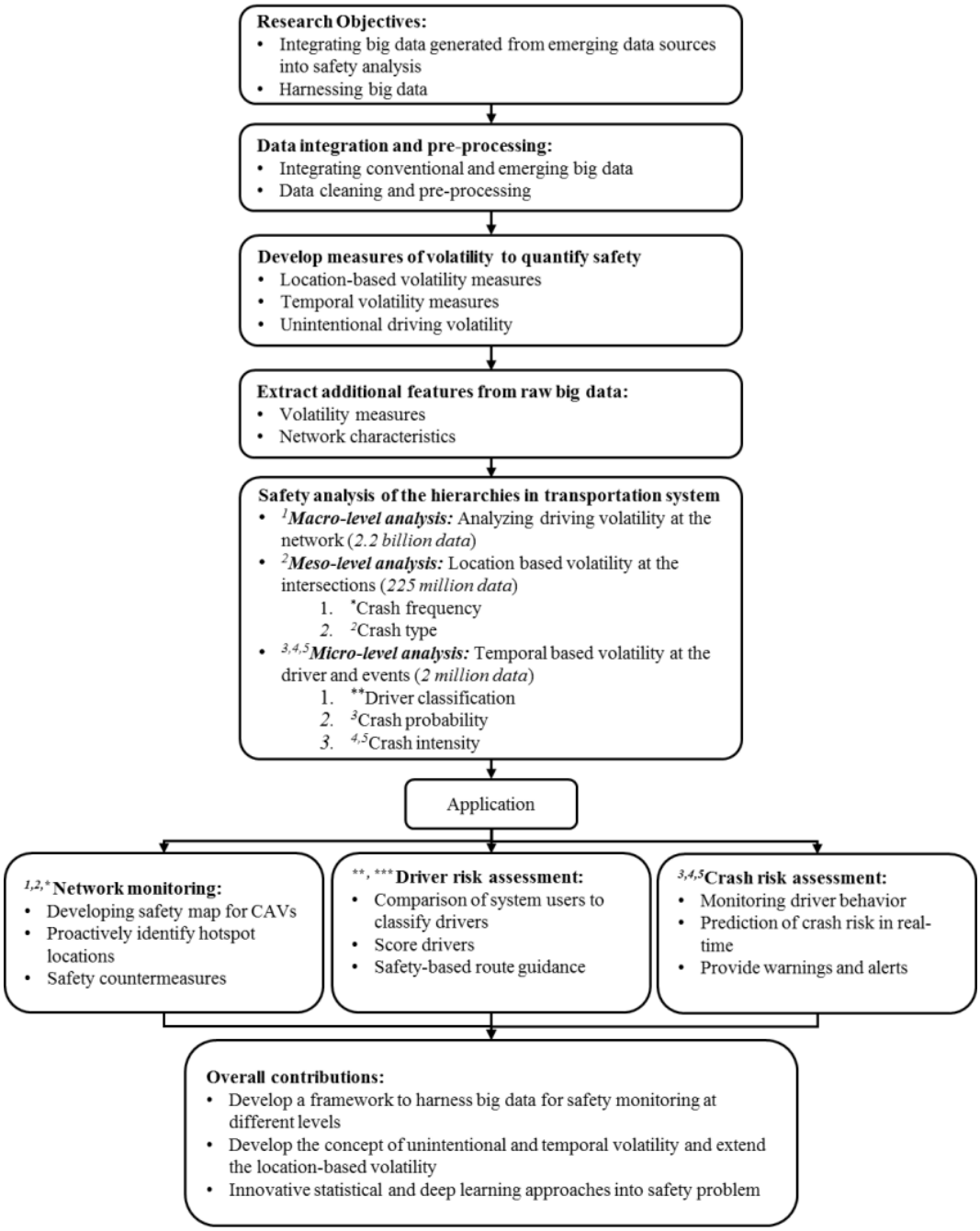
5. Arvin, R., Khattak, A. (2020). *Driving impairments and Duration of distractions: Assessing Crash Risk by Harnessing Microscopic Naturalistic Driving Data.*
    - *Peer-reviewed conference paper:* Presented at the 98th Transportation Research Board Annual Meeting 2020, Washington DC.
    - *Journal article:* Under second-stage review in Accident Analysis and Prevention

6. Arvin, R., Khattak, A., & Qi, H. (2020). *Real-time crash prediction through unified analysis of driver and vehicle volatilities: Application of 1D-Convolutional Neural Network - Long Short-Term Memory.*
    - *Journal article:* Under review in Engineering Application of Artificial Intelligence

The outline of the dissertation is provided in the Figure 1.1. The main objective of this dissertation is to integrate big data generated from emerging sources into safety analysis by considering different levels in the system. To this end, several data sources including

Connected Vehicles data (with more than 2.2 billion seconds of observations), naturalistic driving data (with more than 2 million seconds of observations from vehicular kinematics and driver behavior), conventional data on roadway factors and crash data are integrated. Data cleaning protocols are applied to remove erroneous data from the analysis. Next, this research develops several volatility indices such as temporal and unintentional driving volatility to quantify instantaneous variations in driving behavior. Furthermore, multiple location-based volatility measures are developed to explore association of driving behavior and crash risk at locations. Then, additional information is extracted from raw big data including volatility indices and network characteristics.

In terms of analysis, this dissertation performs analysis at different levels including macro-level (frequency of crashes at the network), meso-level (crash frequency and type at intersections), and micro-level (probability of a crash occurrence and severity). Different statistical, spatial analysis, machine learning, and deep learning methods are utilized to untangle the association of extracted features from the data and crash risk at different levels.



Note:
[1] Chapter 1: Relevant paper is under-review in the Transportation Research Board Conference
[2] Chapter 2: Relevant paper is published in the Accident Analysis and Prevention journal
[3] Chapter 3: Relevant paper is published in the Accident Analysis and Prevention journal
[4] Chapter 4: Relevant paper is under-review in the Transportation Research Board Conference
[5] Chapter 5: Relevant paper is under-review in the Transportation Research Board Conference
* Relevant paper is published in the Transportation Research Record journal
** Relevant paper is under-review in the Transportation Research Board Conference
*** Relevant paper under second review in the Journal of Intelligent Transportation Systems

Figure 1.1 Outline of the dissertation

Problem justification

It is estimated that the traffic incident costs across the US is more than $836 billion with more than 7.2 million crashes, 2.1 million injuries, and 35,000 fatalities across the country. Generally, driver behavior is identified as the main contributing factor in traffic crashes across the United States. It has shown that 94 percent of crashes involving some types of human error prior to the crash occurrence (Anon 2008). Therefore, it is obvious that further investigation is needed regarding driver behavior, specifically prior to crash occurrence. In the literature, researchers are studying the association of different factors on driver behavior and their correlation with safety outcome. However, their analysis is mainly relying on police-reported crash data. It is worth noting that based on the report by National Highway Traffic Safety Administration (NHTSA) (NHTSA 2009), 50% of property damage only crashes and 25% of minor injury crashes are not reported to the police and not recorded. Also, these crashes may be truncated due to states monetary threshold (Hauer 2006).

On the other hand, emergence of new data sources provides a new broad range of opportunities for researchers to think out of the box and apply new concepts and methods in the transportation. Transportation safety can greatly get benefits by incorporating big data to evaluate and monitor the performance of drivers and infrastructure. Big data in transportation might generated from different sources such as Basic Safety Messages generated by Connected Vehicles, naturalistic driving data, Bluetooth, cellular phones, traffic surveillance systems, etc. As an illustration, automotive companies will be mandated to equip their vehicles to be able to communicate with other vehicles and

infrastructures (NHTSA) and enormous data generated by CVs will be available. Furthermore, emergence of naturalistic driving data helps us to study microscopic decisions of drivers, vehicle state, and roadway/environmental condition prior to crash occurrence. On the contrary to the traditional police-reported data which suffering from unreported crashes and information, NDS data contains all crashes and near-crashes with rich information on driver behavior, vehicle movements and roadway condition.

Given such rich datasets and ongoing generation of data streams, there is a great need to incorporate this information in the transportation analysis. Currently, other traditional transportation data sources such as crash data, geometric characteristics, traffic volume, weather condition, sociodemographic factors, etc. extensively used in transportation safety studies. The main question that this dissertation is trying to answer is how we can develop methodological framework to harness big data generated by emerging sources and incorporate this information into transportation safety analysis.

Literature review, gaps and contributions

Monitoring safety performance of the network

Literature review

The majority of studies analyzed historical crash data, roadway and geometric factors, tried to suggest safety countermeasures by developing safety performance models (Farid *et al.* 2018, Wali *et al.* 2018d, Ahmad *et al.* 2019, Farid *et al.* 2019, Ulak *et al.* 2020). In order to model the frequency of crashes, mainly location characteristics are considered in the modeling, including density (Huang *et al.* 2010), skew angle (Nightingale *et al.* 2017), traffic volume and flow (Wang *et al.* 2009, Stipancic *et al.* 2017). On the other hand, although review of transportation safety literature suggest that driver behavior and human errors are the leading cause of crashes (Akamatsu *et al.* 2003, Curry *et al.* 2011, Dingus *et al.* 2016), human behavioral part received less attention. As an illustration, it has shown that aggressive driving is contributing to more than 50 percent of fatal crashes across the U.S. (AAA 2009). It can be inferred that the main limitation of these studies is ignorance of human behavioral side and reactively focusing on roadway and geometric factors which is mainly due to intrinsic data structure of traditional crash data which does not contain information on driver behavior and performance prior to crash involvement. As an alternative to traditional hotspot identification methods, several researchers utilized surrogate safety measures to quantify the crash risk (Essa and Sayed 2018, Rahman and Abdel-Aty 2018, Rahman *et al.* 2018), which are mainly rely on information of the subject and front vehicle. The main challenge in this context is limitation on information of front vehicle which needs to be obtained via computer vision techniques (Ismail *et al.* 2009, Xie *et al.* 2016) or other sensors (Xie *et al.* 2018). Furthermore, by emergence of various

sources of data (e.g. surveillance systems and Global Positioning System data), researchers tried to investigate the association of driving behavior and safety performance of intersections or segments. Quddus (Quddus 2013) investigated the association of speed and crash frequency at freeways using 1-hour average speed data. Another study by Pei et al (Pei *et al.* 2012) studied the association of travelling speed and crash frequency using GPS taxi data.

Emergence of connected vehicles potentially can help to alleviate the issues in the literature by targeting human factors elements and incorporate this information into safety analysis by analyzing large scale data. Recently, the concept of driving volatility is utilized as a surrogate safety measure to quantify variations in driving behavior and explored the association of driving behavior and crash risk at intersections (Kamrani *et al.* 2018b, Wali *et al.* 2018a, Arvin *et al.* 2019c). However, their effort is limited to intersections and their sample size in terms of study area is relatively small. Therefore, the results might not be generalizable to other locations.

From the methodological standpoint, due to complexity of traffic crashes and driving behavior and considering that we are only using CV data as a proxy of driver behavior and traffic condition, it is obvious that all factors that might affect occurrence of crashes are not observed. In addition, spatial data such as crash count typically (Mannering *et al.* 2016) are not independent, and spatial dependency needs to be taken into account (Hadayeghi *et al.* 2010b). While in the literature some methods including Conditional Autoregression (*CAR*), Simultaneous Autoregression (*SAR*), and Spatial Lag models are widely used (see (Aguero-Valverde *et al.* 2006, Wang *et al.* 2006, Hadayeghi 2009)) they

are not thought as local models (Hadayeghi *et al.* 2010b). On the other hand, in the recent crash modeling literature, *GWR* techniques are used to develop statistical models. This approach allows the parameters to vary across the space which accounts for spatial heterogeneity (Xu and Huang 2015). Several studies have utilized Geographically Weighted Poisson Regression to model crash frequency (Hadayeghi *et al.* 2010b, Xu and Huang 2015, Arvin *et al.* 2019c). While it has shown that *GWPR* model outperform the traditional Poisson and Negative Binomial methods (Fotheringham *et al.* 2003, Arvin *et al.* 2019c), the main limitation of this methodology is not accounting for overdispersion in the modeling crash frequency, which usually is not the case in crash frequency analysis.

Research gap

By reviewing the literature, several gaps are identified. First, the study area is mainly limited to segments or intersections, but a proactive network-based framework is not available. Second, the sample sizes are relatively small, and the results might not be widely generalizable. Third, these studies mainly considered longitudinal vehicular movements in order to quantify crash risk. Finally, the spatial heterogeneity among observations are ignored in the analysis.

Objectives and contribution

Given the gaps in the literature, the key objectives of this paper can be summarized in four main points:

1) To develop a fundamental method to quantify variations in instantaneous driving behavior in terms of speed, and longitudinal/lateral/vertical acceleration.

2) To understand the spatiotemporal driving behavior in real-time and study its correlation with traffic crashes.
3) To proactively monitor the network performance by understanding the correlation of driving behavior and traffic crashes to identify hotspots.
4) To account for spatial heterogeneity by developing Geographically Weighted Poisson and Negative Binomial Regression (*GWPR* and *GWNBR*).

The contribution is introducing the concept of temporal driving volatility and developed a methodological framework to process big data generated by more than 2800 CVs in real-world condition to explore the association of driving volatility and crash risk at the network level. Furthermore, variations in vehicular movements in three dimensions (longitudinal, lateral, and vertical) are explored. A systematic approach is proposed to monitor safety performance of the network and identify the hotspot locations for proactive treatment. In addition, from the methodological standpoint, this study utilized Geographically Weighted Negative Binomial Regression to address spatial heterogeneity and overdispersion in the data. In this paper, our main hypothesis is variations in vehicular movement in terms of driving volatility is associated with the crash frequency at the network level, and whether big data generated by CVs can be incorporated to proactively identify hotspot locations in the network. Considering emergence of CAVs and high-resolution big data generated in real-time, this study is timely and original by incorporating this data into safety management problem to proactively identify hotspot locations in the network where crashes are waiting to happen.

Analyzing crash frequency and types at intersections

Literature review

Numerous studies have focused on capturing associations between crash frequency and the geometric characteristics and traffic factors at intersections or road segments. The most favorable method for finding relationships between these variables are statistical count models due to the non-negative, discrete, and randomness nature of crashes.

Focusing on modeling, various methods were utilized for capturing the impact of explanatory variables on crash frequencies, among which fixed parameter models are the simplest. In this approach, the estimated parameters are not allowed to vary across the data (e.g., the effect of Average Annual Daily Traffic (AADT) is constant across all the intersections). However, due to presence of unobserved variations among intersections, one might expect that some of the estimated coefficients vary across intersections, elaborating the model estimation process (Anastasopoulos and Mannering 2009, Washington *et al.* 2010). To address this issue, different promising approaches were developed by researchers such as random-effect and random-parameter models that have been widely used in crash frequency modeling (El-Basyouny and Sayed 2009, Castro *et al.* 2012, Wu *et al.* 2013). The main objective of these approaches is to handle temporal and spatial correlations and account for unobserved heterogeneity among observations (Wali *et al.* 2018b, Wali *et al.* 2018c). However, models might not be transferrable to other datasets (Lord and Mannering 2010). Geographically Weighted Poisson Regression (GWPR) is another method for capturing the spatial variations across observations. This method has the same spirit and methodology as local generalized linear regression method, but there is a different process for determining the weights

(Loader 2006). It has shown that GWPR models outperformed traditional statistical models (i.e. the Poisson model) in terms of capturing spatial variations among crash counts and independent variables (Fotheringham *et al.* 2003). In the literature, most papers consider spatial variations in all of the predicting factors, while in some cases, degrees of variation for some parameters might be negligible. Therefore, it is necessary to apply semi-parametric Geographically Weighted Regression (S-GWPR) models in which some of the factors are global (Xu and Huang 2015). It should be noted that Random Parameter (RP) Poisson regression and GWPR methods are intrinsically different. The coefficients in RP Poisson models are drawn independently from a univariate distribution, disregarding the locations of the observations, while in GWPR the coefficients are derived from coordinates in the geographical space (Xu and Huang 2015).

In the literature, while various location characteristic were considered in crash frequency modeling such as intersection density (Huang *et al.* 2010), skew angle (Nightingale *et al.* 2017), congestion and traffic flow (Wang *et al.* 2009, Stipancic *et al.* 2017), traffic patterns (Noland and Quddus 2005), environmental and weather conditions (Lee and Abdel-Aty 2005, Ghasemzadeh and Ahmed 2018b), and signal characteristics (Agbelie and Roshandeh 2015), driver behavior factors received less attention. In the U.S., more than 50 percent of all fatal crashes were caused by aggressive driving behaviors such as speeding, reckless driving, and failure to yield the right of way (AAA 2009). In the literature, in order to quantify the variations in normal driving behavior, common vehicle kinematics are widely used (Ghasemzadeh and Ahmed 2017, Ahmed and Ghasemzadeh 2018, Ghasemzadeh and Ahmed 2018c). Recently, the term “driving volatility” was introduced (Wang *et al.* 2015a) which attempts to describe the driving behavior

performance. In order to define volatility, researchers have applied different measurements to the kinematic features of vehicles such as speed (Arvin *et al.,* Wang *et al.* 2015a, Kamrani *et al.* 2017, Kamrani *et al.* 2018b, a, Arvin *et al.* 2019b, Kamrani *et al.* 2019) , acceleration (Arvin *et al.,* Wang *et al.* 2015a, Kamrani *et al.* 2018b, Arvin *et al.* 2019b, Kamrani *et al.* 2019) and jerk (Wang *et al.* 2015a, Kamrani *et al.* 2018b). Moreover, some studies (Kamrani *et al.* 2017) have looked at the impact of volatility on the safety performance of traffic networks.

Research gap

In the previous studies several gaps exist. First, the aforementioned studies ignored the variations in lateral movement of the vehicle and only focused on longitudinal volatility. Second, the volatility measures that they were using were limited and might not truly represent the variations in driving behavior. Third, they modeled total number of crashes at intersections, while the impact of driving volatility might vary among different crash types. Finally, they ignored the spatial heterogeneity is not addressed.

Objectives and contribution

The main goals of this research are to:

1) Develop a framework for capturing and quantifying longitudinal and lateral driving volatilities using real-world instantaneous driving data.
2) Evaluate correlations between longitudinal and lateral volatilities with frequency of multiple crash types at intersections.
3) Account for unobserved heterogeneity by utilizing random parameter and semi-

parametric geographically weighted Poisson regression models.

The contribution of this paper is addressing the aforementioned gaps by extending the concept of volatility to longitudinal and lateral volatilities in order to quantify the variations in longitudinal and lateral control of the vehicle. By incorporating large scale Basic Safety Messages (BSM) data transmitted between CVs in real-world environment, 30 measures of volatilities were developed to explore the impact of these measures on the frequency of rear-end, sideswipe, angle and head-on crashes. Our hypothesis is variations in longitudinal and lateral vehicle movement is associated with the frequency of various crash types, controlling for other variables (e.g. traffic exposure, number of legs, number of lanes, etc.). To address the unobserved heterogeneity and spatial correlation, the random parameter and S-GWPR model was employed, and the performance of the models were compared with the fixed parameter Poisson regression.

The role of instability in driving on crash intensity

Literature review

Considerable studies in the literature focused on the investigation of speed, driver behavior, roadway, and environmental factors which are mainly based on police crash reports, which might not be precise and truly represents the crash circumstances. With the emergence of Naturalistic Driving Study (NDS) data, it enabled researchers to perform an in-depth analysis regarding the contributing factors just before a crash.

Various studies have investigated the human-errors and impact of driver behavior on the

severity outcome of a crash such as distracted driving (Neyens and Boyle 2008, Donmez and Liu 2015), aggressive driving (Paleti *et al.* 2010, Lambert-Bélanger *et al.* 2012), impaired driving (Behnood *et al.* 2014, Behnood and Mannering 2017), etc. In the United States, aggressive driving (such as speeding, failure to yield the right of way, and reckless) are accounted as contributing factor in more than 50 percent of fatal crashes (AAA 2009). On the other hand, the impact of distracted and aggressive driving on the driving stability performance is explored by different studies (Beede and Kass 2006, Horberry *et al.* 2006, Hamdar *et al.* 2008, Stavrinos *et al.* 2013). Various measurements are incorporated to explain stability performance of driving such as speed (Beede and Kass 2006, Ghasemzadeh *et al.* 2018), speed variability (Rakauskas *et al.* 2004, Beede and Kass 2006), lane position maintenance (Rakauskas *et al.* 2004), lateral control (Beede and Kass 2006), time to collision (Papazikou *et al.* 2017), reaction time (Rakauskas *et al.* 2004, Sheng *et al.* 2019), etc. In this study, the concept of "driving volatility" is utilized as an indicator for driving stability performance prior to a crash occurrence. In order to define driving volatility, various measures are applied to kinematics of vehicles such as speed (Kamrani *et al.* 2018b, a, Arvin *et al.* 2019c), acceleration and deceleration (Kamrani *et al.* 2018b, Arvin *et al.* 2019c), and vehicular jerk (Kamrani *et al.* 2018b). In addition, research has shown that driving volatility is highly correlated with the crash frequency (Kamrani *et al.* 2017, Kamrani *et al.* 2018b, Arvin *et al.* 2019c).

On the other hand, the association of roadway/environmental factors on the severity outcome of crashes are investigated by several studies. As an illustration, the impact of traffic flow (Theofilatos and Yannis 2014), weather condition (Ghasemzadeh and Ahmed

2018a, Jalayer *et al.* 2018), surface condition (Wang and Zhang 2017), roadway alignment (Wang and Zhang 2017, Haghighi *et al.* 2018), and time of day (Mokhtarimousavi *et al.* 2020) on the crash severity have studied. Furthermore, researchers have investigated the impact of these factors on driving stability such as traffic density (Shakouri *et al.* 2014, Teh *et al.* 2014), road geometry (Wang *et al.* 2015b, Hamdar *et al.* 2016), work zone (Shakouri *et al.* 2014, Mokhtarimousavi *et al.* 2019), adverse weather (Ghasemzadeh and Ahmed 2017, 2018c), surface condition (Kircher and Thorslund 2009), vehicle type (Rahimi *et al.*), etc.

Research gap

An obvious limitation in the literature is the vast majority of studies have not explored the impact of driving volatility on crash severity while investigating the association of driver behavior and roadway/environmental factors on both severity and driving stability. Driver behavior and roadway/environmental factors likely are contributing to the driving stability and might have both direct and indirect contribution to the intensity of crashes. In addition, most of the crash datasets are suffering from unreported property damaged only crashes, while this study takes advantage of the second Strategic Highway Research Program (SHRP 2) data contains detail information on extreme safety situations, including minor crashes leading us to investigate an in-depth analysis of PDO crashes.

Objectives and contribution

To summarize, the questions that this paper is trying to answer are:

- How can we extract useful information about enhancing safety from recently available microscopic vehicle kinematics data?
- How is crash intensity related to pre-crash driving volatility (or driving instability)?

In summary, the contributions of this study are:

1. Extract useful information by developing a framework for safe speed and movement, and by analyzing stability performance as a leading indicator prior to crash occurrence.
2. Exploring pathways that can intensify risky and unsafe events. This task is done by developing measures of driving volatility. The study explores the correlates of volatility itself and influence of volatility on crash intensity.
3. Instead of analyzing conventional police-reported crashes that do not contain microscopic vehicle kinematic information, this study analyzes pre-crash kinematic data and extracts a different set of contributing factors.

Association of driving impairment/distraction on crash risk

Literature review

The impact of distracted driving on driving performance has been widely studied in the literature. It has shown that deviation of attention from the driving task can lead to delay in reaction time (Horrey and Wickens 2004, Drews *et al.* 2009, Gao and Davis 2017), deteriorate vehicle control (Choi *et al.* 2013, Young *et al.* 2014, Arvin *et al.* 2019b, Kamrani *et al.* 2019), and miss events (Fitch *et al.* 2009, Hosking *et al.* 2009). The

availability of microscopic naturalistic driving data enabled researches to study driving behavior prior to critical events and study their associations. In the literature, several studies have investigated the association of distracted driving on crash risk (Dingus *et al.* 2011, Dingus *et al.* 2016, Kamrani *et al.* 2019, Nasr Esfahani *et al.* 2019) and its severity (Arvin *et al.* 2019b).

Recent studies are focusing on drivers secondary task in terms of removing eyes from the forward roadway, and established a relation between eye-off-road and crash risk (Klauer *et al.* 2006, Victor *et al.* 2015). Glance location can be utilized to infer whether the driver is fully engaged in the driving task or not (Wickens *et al.* 2003, Taylor *et al.* 2013). It has shown that drivers are not tending to hold their glances away from the roadway for more than 1.6-2 seconds (Sodhi *et al.* 2002, Liang *et al.* 2014), instead, they increase the number of times looking away from the road (Victor *et al.* 2005). Using safety surrogate measures, it has shown that higher percentage of the times that drivers have eyes off the road is associated with increase in probability of safety critical event (Ahlstrom *et al.* 2013).

Along with distracted and impaired driving, literature suggests that roadway and environmental factors such as weather condition (Ghasemzadeh and Ahmed 2016, Haghighi *et al.* 2018), road characteristics (Manan *et al.* 2017), surface condition (Wang and Zhang 2017), traffic flow (Theofilatos and Yannis 2014, Kamrani *et al.* 2019), etc. are associated with the crash risk.

Research gap

An obvious limitation in the literature is the vast majority of studies have not explored the impact of driving instability on crash severity while investigating the association of driver behavior and roadway/environmental factors on both severity and driving stability. Driver behavior and roadway/environmental factors likely are contributing to the driving stability and might have both direct and indirect contribution to the intensity of crashes. In addition, most of the crash datasets are suffering from unreported property damaged only crashes, while this study takes advantage of the second Strategic Highway Research Program (SHRP 2) data contains detail information on extreme safety situations, including minor crashes leading us to investigate an in-depth analysis of PDO crashes.

Objectives and contribution

This study contributes to the literature, by:

1- Developing an understanding regarding the influence of duration of distracted driving, categorized by different sources, on the probability of extreme event occurrence, while controlling for other driving behavior and roadway/environmental factors.
2- Providing in-depth analysis of impact of distraction duration by different secondary tasks during 15 seconds before crash and near-crash involvement.
3- Investigating the role of impaired (alcohol and drug) driving on crash risk.

Real-time crash prediction

Literature review

Numerous studies have explored the association of driving behavior, vehicle factors, and roadway/environmental characteristics on the probability of crash risk using statistical methods (add reference). Although most of these researchers are rely on police-reported data, they provide insightful inference regarding the association of driving behavior and crash risk. Emergence of naturalistic driving data and high-resolution driving decisions opened new area to explore microscopic driving behavior prior to crash occurrence. In our previous researches (insert references), we have shown that instability in driving not only increase the likelihood of a crash involvement but also severity of a crash. In order to quantify instability in driving, we have introduced the concept of driving volatility and we have shown that it can be served as a leading contributing factor.

On the other hand, deep learning methods recently have received lots of attention due to emergence of big data generated by multiple sources and availability of computational power. It has shown that deep learning methods are great tool for representation learning with little effort for manually feature extraction (Goodfellow *et al.* 2016). Referring to the transportation field, deep learning has applied to several fields including demand prediction (Lin *et al.* 2018, Xu *et al.* 2018a, Bao *et al.* 2019b), transportation safety (Li *et al.* 2018, Bao *et al.* 2019a), travel time prediction and reliability (Ghanim and Abu-Lebdeh 2015, Tang *et al.* 2019), driver behavior prediction (de Naurois *et al.* 2017, Liu and Shi 2019, Osman *et al.* 2019), signal control (Jeon *et al.* 2018, Xu *et al.* 2018b), driver impairment detection (Ye *et al.* 2017, de Naurois *et al.* 2018), vehicle classification (Nezafat *et al.* 2019), etc. the main advantage of deep learning architecture over

traditional statistical methods is modeling complex non-linear relations between associated factors and dependent variable by incorporating distributed and hierarchical features (Ma *et al.* 2015).

In terms of real-time crash prediction, we can identify two groups of studies attempted to address this issue in the literature. The first group, which real-time crash prediction mainly refers to, focuses on macro-level prediction of a crash in a network or segment (Shi and Abdel-Aty 2015, Basso *et al.* 2018, Yang *et al.* 2018, Parsa *et al.* 2019). In other words, these models are trying to predict the time and location of crashes that might occur in the network in order to support the monitoring the traffic data and network performance. Several studies have applied machine learning and deep learning methods including Bayesian network (Hossain and Muromachi 2012, Sun and Sun 2015), Support vector machine (Sun and Sun 2016, Wang *et al.* 2019b), CNN (Bao *et al.* 2019a), and LSTM (Ren *et al.* 2017, Bao *et al.* 2019a) to predict occurrence of a crash at the aggregate level.

Referring to micro-level analysis, few studies attempted to identify crash risk level in a real-time manner. Shi et al (Shi *et al.* 2019) performed discrete Fourier transform and performed XGBoost and K-mean clustering in order to detect critical events. Kluger et al. (Kluger *et al.* 2016) performed Discrete Fourier Transform and K-means on longitudinal acceleration to detect critical events on the 49 crashes and 42 near-crashes. Perez et al (Perez *et al.* 2017) utilized thresholds to identify boundaries for the detection of crash/near-crash events. Gao et al. (Gao *et al.* 2018) predict the longitudinal conflicts between vehicles with CNN using vehicle kinematics and front-camera videos. However, their analysis is only capturing a commercial truck fleet, and the results might not be

generalizable to other drivers and vehicle types. One of the few studies which attempted to classify the crash and near-crash events is performed by Osman et al. (Osman *et al.* 2018). They tried to predict the crash and near-crash events based on the vehicle kinematics data. They have tested multiple machine learning approaches including Random Forest, support Vector Machine, K Nearest Neighbor, Quadratic Discrimination Analysis and they reached 88 percent accuracy. However, they have not mentioned that whether they are excluding the seconds that the vehicles were involved in a crash or they are using the vehicle kinematics after the occurrence. On the other hand, from the methodological standpoint, it seems that their method cannot capture the complexity embedded in the data, which potentially can be improved by Deep Learning methods. Bugusa et al (Patil) tried to predict the real-time safety risk based on driver behavior and environment using Elastic Net regularized logistic regression. In this paper, they only discussed the possible framework that can be applied to the data and they did not discuss the modeling outcome.

Research gap

By reviewing the literature, it can be understood that there are several gaps. First, the previous studies mainly incorporate raw vehicular movements in the analysis, while driver behavior and instability in driving is mainly ignored. Second, the temporal nature of the dataset is ignored, and simple machine learning or neural network models are used, which might not fully address the time dependency between observations. Finally, the proposed models might not perfectly capture the non-linearity relationships between the input and output of the model.

Objectives and contribution

In this study, the main contribution is developing a deep learning framework to integrate multiple data streams including vehicular kinematics in terms of speed, longitudinal and lateral accelerations, driving stability, and driver behavior to predict the occurrence of a crash/near-crash. The developed framework has several advantages:

1- The architecture configuration of the model is compact, making the model easy to be implemented for real-time safety performance monitoring and failure detection.
2- Its ability to capture temporal variations in the input data generated from multiple sensors.
3- The capability of the model to efficiently train the model using limited training dataset and back-propagation iterations (Eren *et al.* 2019).

Methodological framework

The main goal of this study is to harness big data generated by emerging data sources and integrate this information with traditional transportation data in order to perform safety analysis at different levels. This dissertation develops a unique framework to integrate different big data sources and harness this information to perform safety analysis. The overall framework of the dissertation is provided in the Figure 1.2.

From the data perspective, three groups of data are considered in this framework: 1- conventional transportation data (including roadway inventory, traffic data, and historical

crash data), 2- emerging data sources (connected vehicle, naturalistic driving, and on-board unit data), and 3- driver biometrics data (distraction profile, heartrate, and brainwave).

These data sets are pre-processed, manipulated and integrated to create our final big data. In order to incorporate this data into analysis, first several features are extracted from raw data to gain more information regarding transportation system state and performance. These features represent geometric data, exposure of traffic, and traffic flow state. Next, the concept of driving volatility is extended by introducing the concepts of temporal and unintentional volatilities and expanding the location-based volatility measures. These measures aim to quantify instantaneous variations in driving behavior. Finally, initial analysis on the data is performed and correlation of safety metrics and extracted features (e.g. driving volatilities) are quantified.

From the safety analysis perspective, once the big data is established and pre-processed, the dependent and independent variables are identified. Referring to the dependent variables, this dissertation focused on frequency of crashes, type of crashes, risk of crash, severity of crashes and instability in driving. On the other hand, several independent variables are identified to present driving volatility, driver behavior, and roadway/environmental factors. Several modeling approaches including statistical modeling, spatial analysis, heterogeneity-based modeling, machine learning, and deep learning methods are utilized to explore the association of independent variables with dependent variables.

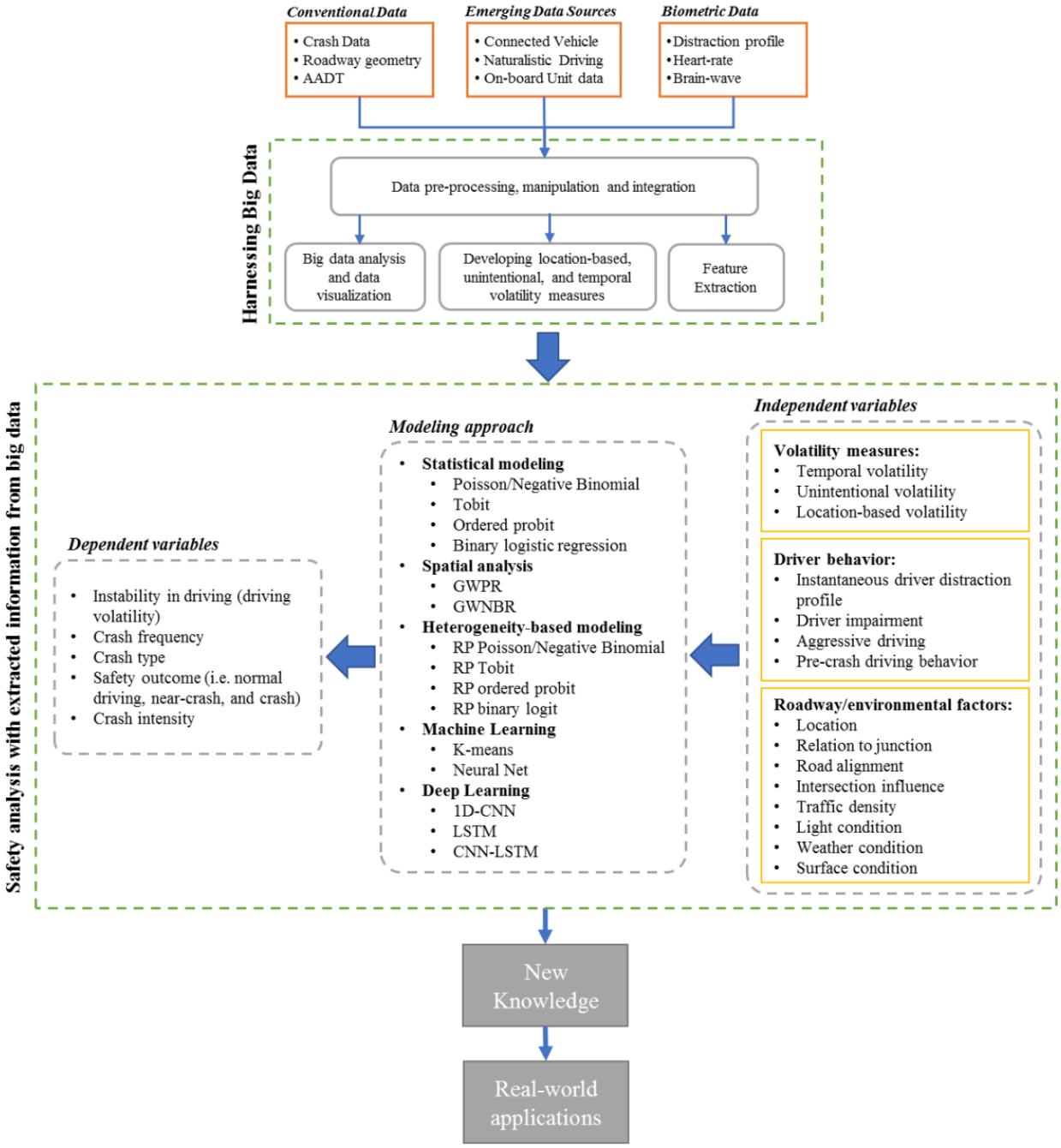


Figure 1.2 Framework of the dissertation

CHAPTER 2 : HARNESSING BIG DATA GENERATED BY CONNECTED VEHICLES TO PROACTIVELY MONITOR SAFETY PERFORMANCE OF THE NETWORK: APPLICATION OF GEOGRAPHICALLY WEIGHTED NEGATIVE BINOMIAL REGRESSION

This chapter is a modified version of a research article by Ramin Arvin and Asad J. Khattak. *"Harnessing big data generated by connected vehicles to monitor safety performance of network: Application of Geographically Weighted Negative Binomial Regression."* The manuscript presented at the 99th Annual Meeting of Transportation Research Board Conference at Washington DC, and it is currently under review in Accident Analysis and Prevention.

Abstract

The emergence of high-frequency and high-resolution big data generated by connected and automated vehicles provides promising opportunities to monitor and evaluate the transportation systems performance. This study develops a conceptual framework that harnesses such a big data to monitor the safety performance of the system by incorporating human behavior to identify high risk locations in the network. The main advantage of this framework is proactively monitoring system safety performance whereas traditional methods reactively identify high risk locations. More than 2.2 billion Basic Safety Messages transmitted between more than 2800 CVs collected in Ann Arbor, MI through the Safety Pilot Model Deployment are processed, analyzed and linked with crash data. This study captures the temporal dimension of driving volatility by quantifying variations in instantaneous driving behavior and decisions. Several measures of volatility are applied to vehicular speed, lateral, longitudinal, and vertical acceleration, and their correlations with observed crash frequency are explored. To address unobserved heterogeneity in safety performance and spatial correlations, Geographically Weighted Poisson and Negative Binomial models are estimated and their goodness of fit are

compared. Results reveal that driving volatility is positively and significantly correlated with frequency of crashes, and these associations vary substantially across space. Variations in longitudinal vehicle movements (speed and longitudinal acceleration volatility), and lateral movements (in terms of lateral acceleration) are associated with higher crash frequencies. In order to identify hotspot locations, k-means and Gaussian Mixture Model (GMM) clustering is performed, and the grids are clustered into low, medium and high volatility groups. Grids with high volatility and low crash frequency are potential hotspot locations. Further examinations are needed to identify reasons why drivers exhibit volatile driving behavior and to develop countermeasures that decrease crash risk by reducing driving volatility.

Introduction

In order to effectively allocate resources, precisely identifying hotspot locations in the transportation network is crucial. Traditional hotspot identification methods mainly rely on historical crash data by monitoring number of crashes to reach a sufficient threshold for further investigations and treatments. As an alternative to traditional methods, several studies utilized surrogate safety measures in order to assess and mitigate crash risk (Rahman and Abdel-Aty 2018, Rahman *et al.* 2018) by proactively applying countermeasures in advance. The emergence of Connected and Automated Vehicles (CAV), however, provides large scale high-resolution big data that was previously unavailable, ranging from macro-decisions such as origin and destination decision of a trip to micro-decisions including instantaneous driving behavior reflected in vehicular

movements. Useful information can be extracted from this big data to improve the performance of the network system in terms of mobility, operation, and safety. The main expected deliveries through the implementation of connectivity to vehicles and infrastructures are improvements in efficiency and safety performance (Lu *et al.* 2014).

In the near future, when automotive companies are mandated to equip their vehicles with technology that enables them to communicate with other vehicles and infrastructures (NHTSA), enormous data generated by CVs will be available. Therefore, there is an opportunity to harness this data in order to create innovative ways to monitor and improve the network performance. In this regard, the United States Department of Transportation (USDOT) conducted the Safety Pilot Model Deployment (SPMD) in order to advance the deployment of connected vehicles, which is known as one of the largest and most successful studies. The SPMD enabled vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication, involving more than 2800 vehicles and implemented Road Side Units (RSU) on more than 70 roadway miles in Michigan (Henclewood *et al.* 2014). The SPMD utilized Dedicated Short-Range Communication (DSRC) which enabled vehicles to communicate with other CVs and infrastructure at the frequency of 10 Hz to establish the largest communication testbed in the United States. It should be noted that collected data during the two months of this study contains more than 2.2 billion observations. In the future with implementation of connectivity between all vehicles in the network, an enormous amount of data will be available which can be harnessed and incorporated into safety management studies.

In recent literature, driving volatility, taken from the economic field, is one of the safety

surrogate measures that aimed to assess crash risk. Driving volatility captures and quantifies variations in driving behavior by measuring vehicular movements such as speed, acceleration, and jerk (Kamrani *et al.* 2018b, Arvin *et al.* 2019c). From a safety perspective, driving volatility has been shown to be a leading indicator of crash occurrence and crash severity (Arvin *et al.* 2019a, Kamrani *et al.* 2019). This concept can be utilized on big data collected by CVs and integrated with crash data to study the association of instantaneous driving behavior with frequency of crashes.

This paper develops a unique methodological framework that harnesses big data generated by CVs in order to monitor and evaluate the safety performance of transportation systems. We introduce the concept of temporal driving volatility, which quantifies variations in each instance of driving behavior, in order to extract useful information from large-scale raw data. From the methodological perspective, rigorous spatial modeling techniques are utilized to address spatial heterogeneity. Finally, a systematic approach is presented that proactively identifies hotspot locations in the system for further examination, treatment and the development of countermeasures that will decrease driving volatility.

Data

This study takes advantage of big data generated by two months of connected vehicle data from the SPMD dataset, contains information from about 2800 CVs and 30 roadside equipment (RSE) covering more than 73 lane-miles, traversing Ann Arbor, Michigan (Bezzina and Sayer 2014). The SPMD is known as one of the most comprehensive real-

world CV data collection efforts containing multimodal traffic and vehicles which were able to communicate to vehicles and infrastructures via V2V and V2I communication devices (Bezzina and Sayer 2014). The main objective of undertaking the SPMD by the USDOT was to support and advance the evaluation of DSRC for V2V safety applications (Bezzina and Sayer 2014). A subset of the data is publicly available which were collected on two months (October 2012 and April 2013) (N~2.2 billion observations) via the Intelligent Transportation system data hub of USDOT (https://www.its.dot.gov/data/). In this study, the full two-month Basic Safety Messages (BSM) transmitted between more than 2800 CVs is used and integrated with historical crash data. The CV data generates high-frequency and high-resolution information about location and motions of vehicle, and driving context factors. Figure 2.1 illustrates the study area and trajectory of vehicles passing the network. It can be inferred that data has high resolution and the trajectories are covering the entire city. The crash data is retrieved from the Michigan Data Poral (http://gis-mdot.opendata.arcgis.com/). Erroneous data observations were removed from the dataset using the procedure developed by Xie et al (Xie *et al.* 2018) and Kamrani et al (Kamrani *et al.* 2018b).

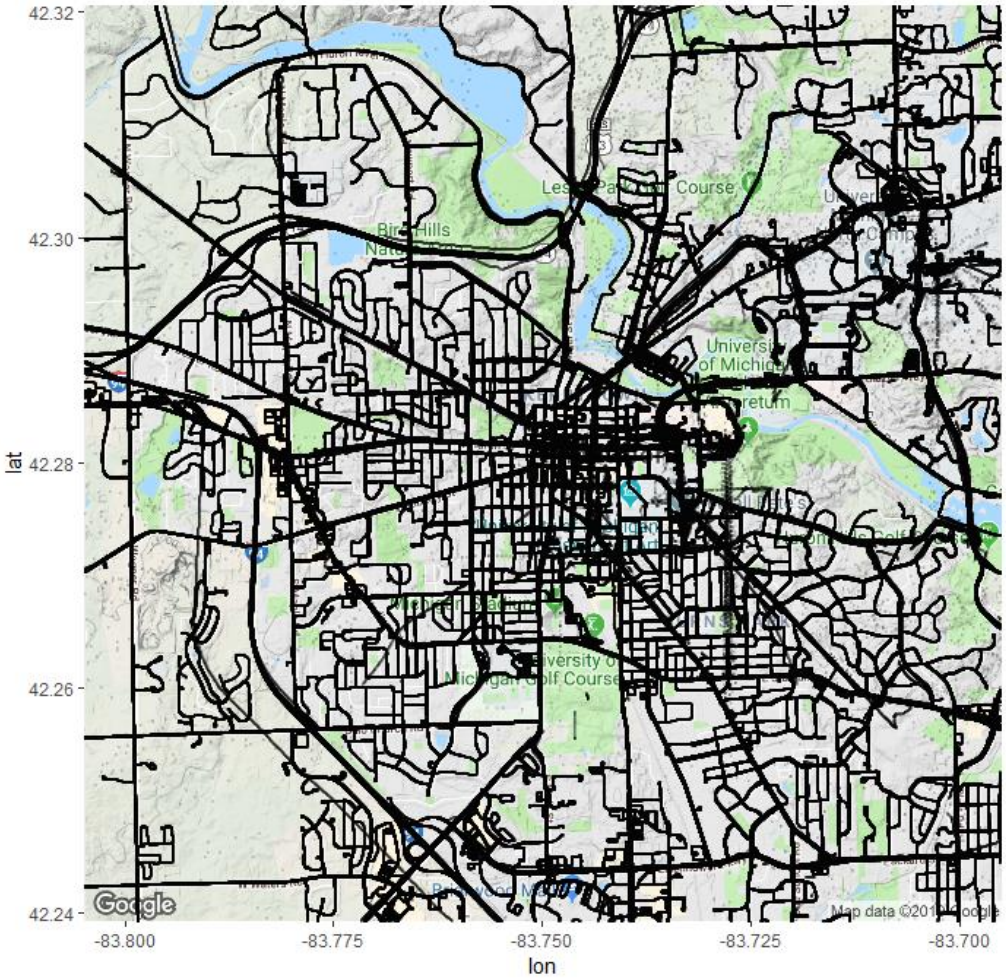


Figure 2.1 Study area and generated map by connected vehicles

Methodology

In this section, we will discuss the conceptual framework of the study, definition of volatility measures and the algorithm to calculate these measures, the modeling approach and model comparison, and finally unsupervised clustering approach for proactive hotspot identification.

Conceptual framework

In order to develop the methodological framework, high resolution microscopic vehicular movements of vehicles are required. In this study, this information is obtained from SPMD study. The conceptual framework of the paper is shown in the Figure 2.2. In this study, on the contrary to the traditional methods where we manually collect information of roadway, we are extracting features and information from the raw CV data. These features consist of network and volatility features. Network features focusing on general geometric information of the system (such as elevation, radius of the curve, number of BSM observations (as a proxy of AADT)), and vehicular movements in terms of average of speed, acceleration and yaw rate of vehicles passing each location. Focusing on volatility features, this study developed the concept of temporal volatility and coupled it with location-based volatility to add the drivers' behavioral aspect to the model. this information is added to the modeling framework, and the significant variables are fed to the unsupervised classifier (i.e. K-means and GMM) to identify hotspot locations where driving volatility is high, while number of crashes are low.

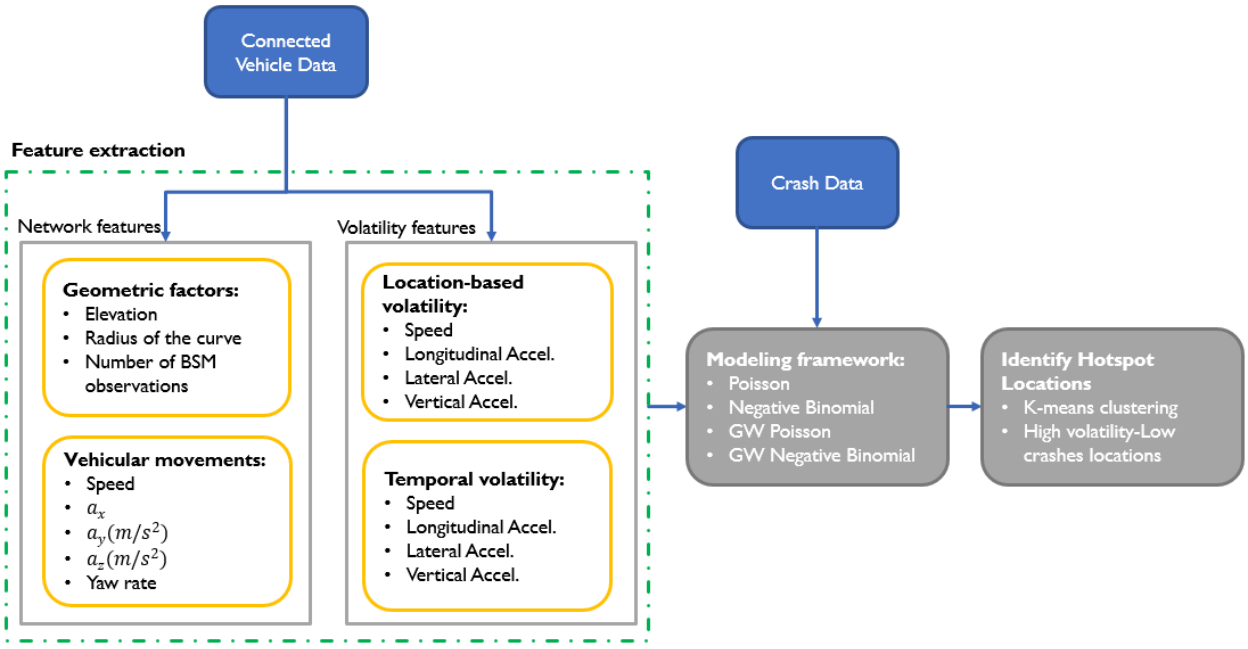


Figure 2.2 Conceptual framework

Figure 2.3 provides the workflow of the paper. There are five major steps:

1- *Data pre-processing:* The goal of this step is to pre-process raw data extracted from connected vehicles in order to prepare for further analysis. First, the data is filtered on the study area (i.e. Ann Arbor, MI). Next, errors and outliers are removed from the data. Finally, zero speeds are excluded from the analysis. These values potentially affect volatility measures, especially at intersections (Arvin *et al.* 2019c). The output of this step is a clean and processed dataset of vehicle trajectories and kinematics.

2- *Calculating temporal volatility:* In this step, we need to extract the trajectories of each individual trip taken by each driver and calculate temporal driving volatility (discussed in detail in section 4.3.2). This step outputs a dataset containing

information on instantaneous driving volatility in the longitudinal, lateral, and vertical directions for each instance of a trip.

3- *Mapping volatility on the network:* Given the temporal volatility measures for all drivers and trips, this information is averaged on each pair of *(x, y).* By defining a grid network in the study area, crash data and volatility indices are mapped on grids. Finally, for each grid, the location-based volatility measures are calculated (discussed in detail in section 4.3.1). The output of this step is the final dataset used in the modeling.

4- *Modeling framework:* After finalizing the dataset in the previous steps, the correlation of developed volatility measures with crash frequency is studied by developing fixed and Geographically Weighted Regression models to identify measures that have the highest correlation with crash frequency. Next, given these significant volatility measures, an unsupervised clustering approach is developed to identify grids with high driving volatility.

5- *Hotspot identification:* In the last step, after finding the contributing volatility measures, k-means and GMM clustering is performed. Next, locations with high volatility are identified and these locations where the number of crashes is also low are identified as potential hotspot locations.

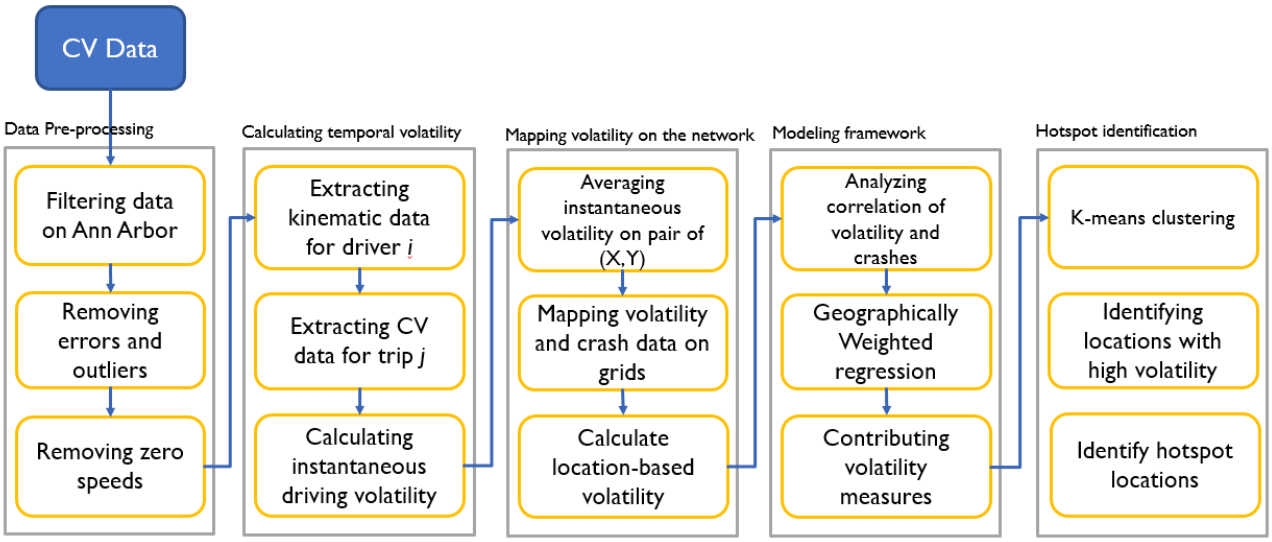


Figure 2.3 Workflow of the paper

Measures of driving volatility

In the literature, driving volatility quantifies variations in driving behavior from norm. It has been shown that these measures can represent the driving behavior of the majority of users in the study area (Arvin *et al.* 2019c). In previous studies, several functions are proposed to quantify variations in vehicular control including speed (Kamrani *et al.* 2018b, Arvin *et al.* 2019c, Kamrani *et al.* 2019), longitudinal acceleration (Kamrani *et al.* 2018b, Arvin *et al.* 2019c, Kamrani *et al.* 2019), lateral acceleration (Arvin *et al.* 2019c), and vehicular jerk (Kamrani *et al.* 2018b). This paper also investigates the association of vertical movements of vehicle in terms of vertical acceleration volatility. This study applies several mathematical functions on CV data to develop four groups of volatilities:

1- Speed volatility

2- Longitudinal acceleration volatility

3- Lateral acceleration volatility

4- Vertical acceleration volatility.

For each group of volatility, volatility measures are calculated at two levels: *Level 1*: Location-based volatility, and *Level 2*: Temporal driving volatility. In the section 4.3, the calculation procedure for each group will be discussed in detail. In the following., the formulation for each volatility function will be discussed in detail.

Time-varying stochastic volatility

This measure quantifies variations in vehicular movements by capturing changes in the ratio of observations. We can write (Figlewski 1994):

$$V_f = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})} \qquad from\ t = 1\ to\ n \tag{2.1}$$

where V_f is the time-varying stochastic volatility, *n* is number of observations, and r_i can be defined as:

$$r_i = ln\left(\frac{x_t}{x_{t-1}}\right) \tag{2.2}$$

where *ln* is the natural logarithm, x_t and x_{t-1} represent the observations at time t and $t-1$, respectively. Considering this volatility measure requires positive time-series observations, this function only applies to vehicle speed.

Coefficient of Variation

It quantifies variations by calculating the ratio of standard deviation over mean (Everitt and Skrondal 2002), and was applied to all four groups of volatility.

$$C_v = \frac{S_{dev}}{|\bar{x}|} \tag{2.3}$$

where S_{dev} is the standard deviation of the observations, and $|\bar{x}|$ is the mean of observations.

Quartile Coefficient of Variation

In cases where the data is not following the normal distribution, quartile coefficient of variation is one of the desirable measures (Zwillinger and Kokoska 2000), which can be written as (Bonett 2006):

$$Q_{CV} = \frac{Quart_3 - Quart_1}{Quart_3 + Quart_1} \tag{2.4}$$

where $Quart_1$ and $Quart_3$ represent the 25th and 75th percentiles of observations, respectively.

Mean absolute deviation

This measure quantify dispersion in the observations by calculating the distance between

each individual with central tendency (mean in this paper). We can write (Huber 2005):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{2.5}$$

Count of extreme values

This measure tries to count extreme observations that lie in the data by defining certain thresholds. We can write (Kamrani *et al.* 2018b):

$$Extreme = \frac{c > Threshold}{n} \times 100 \tag{2.6}$$

where c is the number of extreme points lying out of the threshold, and n is total number of observations. The threshold can be defined as (Kamrani *et al.* 2018b):

$$Threshold = \bar{x} \pm 2 \times S_{dev} \tag{2.7}$$

Calculation of volatility measures

In the previous section, four groups of volatility measures are defined (speed, longitudinal, lateral, and vertical acceleration) and mathematical functions that applied on each group is discussed. In this study, volatility measures for each volatility group are calculated at two levels.

1- Location-based volatility
2- Temporal driving volatility

In the following, each level will be discussed in detail.

Level 1: Location based volatility

In this approach, in order to calculate volatility measures, passings undertaken by each individual are disregarded and all the CV data in each grid is treated as a bulk. For each grid, CV message data are filtered, and volatility functions are applied on the data to obtain volatility indices. In terms of computational sources, this approach needs much lower processing units compared to temporal driving volatility measures. For further information, please refer to (Kamrani *et al.* 2018b).

1.4.3.2 Level 2: Temporal driving volatility

This study introduces the concept of temporal volatility, which attempts to quantify variations in instantaneous driving decisions at the micro (driver) level and creates a time-series data. The advantage of this approach is that it captures the time dependency between observations, which can help detect and identify the times that an individual driver is showing volatile behavior. The calculation has three main phases, which will be discussed in the following.

Phase 1 – Calculate temporal driving volatility: Similar to the concept of moving average, we considered a 3-second (30 deci-seconds) time-window to calculate temporal volatility measures, and the values are assigned to the subject time. Figure 2.4 illustrates the calculation of temporal volatility utilizing the moving window.

Phase 2 – Aggregate temporal volatility on points: In the previous step generates temporal driving volatility for each trip. Since the data is geocoded, this information can

be mapped on the road network. In this phase, given all the trips passing the location *(x, y)*, the temporal volatility measures are aggregated and averaged on location *(x, y)*. This procedure is performed for all the points on the network.

Phase 3 – Averaging volatility measures on grids: In the last phase, the information on grids is aggregated by averaging volatilities of locations that fall in the grid.

Detail information on calculation of temporal volatility measures is provided in Table 2.1.

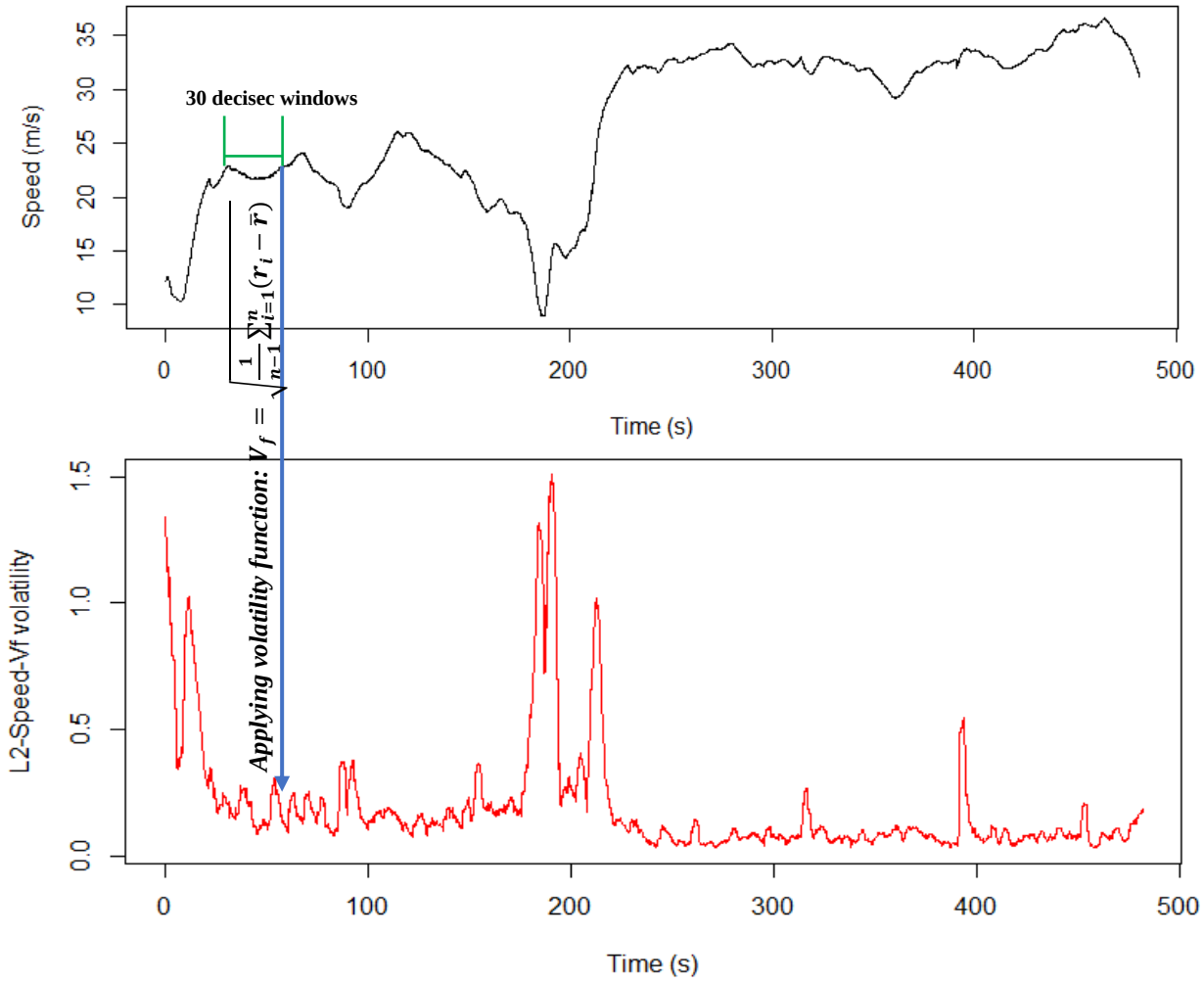


Figure 2.4 Calculation of temporal driving volatility

Inputs

Connected vehicle data

Geo-referenced coded crashes

Volatility functions

Polygon of grid network of the city

Outputs

Volatility indices on each grid

Crash frequency on each grid

Average kinematic information of passing vehicles on each grid

Table 2.1 Algorithm for calculation of temporal driving volatility

Phase 1

For each driver, $\mathrm{i} = (1, \mathrm{N})$

For each trip taken by driver *i*, $\mathrm{j} = (1, \mathrm{M_i})$

For each second of trip *j* taken by driver *i* $\mathrm{t_{k_{i,j}}} \in [30, \mathrm{T_{i,j}}]$

Step 1: Subset three seconds (30 deci-seconds) of data $[\mathrm{t_{k_{i,j}}} - 30, \mathrm{t_{k_{i,j}}}]$

Step 2: Record kinematic information of vehicle

Step 3: Apply volatility functions

Step 4: Assign calculated volatility measures and extracted kinematic information to time $\mathrm{t_{k_{i,j}}}$ and location $(\mathrm{x_{k_{i,j}}}, \mathrm{y_{k_{i,j}}})$ $\mathrm{Volatility_{v,t_{k_{i,j}},(x_{k_{i,j}},y_{k_{i,j}})}}$

Phase 2

For each location (x, y)

Step 5 Calculate mean $\mathrm{Volatility_{v,(x,y)}} = \frac{1}{\mathrm{n}} \sum_{(\mathrm{x,y})} \mathrm{Volatility_{v,t_{k_{i,j}},(x_{k_{i,j}},y_{k_{i,j}})}}$

Phase 3

For each grid, $\mathrm{l} = (1, \mathrm{G})$

Step 6: Subset from the processed data and crash data located on grid *l*, g_l

If number of observations > 0

Step 7: Calculate mean of volatility on grid *l* $Volatility_{v,g_l} = \sum_{(x,y)\in g_l} Volatility_{v,(x,y)}$

Else

Step 8: Remove grid *l*

Summary of notations

i: index of drivers

j: index of trips

M_i: number of trips taken by driver *i*

$T_{i,j}$: total travel time of trip j taken by driver *i*

$t_{k_{i,j}}$: time k of trip j taken by driver *i*

$(x_{k_{i,j}}, y_{k_{i,j}})$: location of driver i in trip j at time k

$Volatility_{v,t_{k_{i,j}},(x_{k_{i,j}},y_{k_{i,j}})}$: Volatility measure v at time k of trip j taken by driver *i*

(x,y): longitude and latitude

$Volatility_{v,(x,y)}$: Volatility measure v at location (x,y)

$Volatility_{v,g_l}$: Volatility measure v at grid gl

Modeling Approach

Once the temporal driving volatility and location-based volatility measures are calculated, we need to investigate the association of volatility measures and crashes. In the literature, considering non-negative integer values of the number of crashes in a specific period of time, different methods are utilized to model the dependent variable including Poisson regression, Negative Binomial, and Zero Inflated Models (Anastasopoulos and Mannering 2009, Azizi and Sheikholeslami 2012, Dong *et al.* 2014). In this study, the fixed parameter Poisson/Negative Binomial model, *GWPR*, and *GWNBR* were considered for modeling crash frequency as a function of extracted features from CV data.

Poisson Model

The Poisson regression model can be set up to estimate the probability of observing *n* crashes at grid *i* can be formulated as (Greene 2003):

$$P(n_i) = \frac{\lambda_i^{n_i} \, exp(-\lambda_i)}{n_i!} \tag{2.8}$$

where λ_i is the Poisson parameter (is equal to expected crash frequency for grid *i*, $E(n_i)$). To estimate the Poisson model, the λ_i parameter is written in the logarithm form of a set of explanatory factors (Greene 2003):

$$E(n_i) = ln(\lambda_i) = \beta X_i \tag{2.9}$$

where X_i is the matrix of the explanatory factors and β is a vector of the model parameters. In order to maximize the Poisson function, following maximum likelihood function is utilized (Washington *et al.* 2010):

$$L(\beta) = \prod_i \frac{\exp[-\exp(\beta X_i)] \, [\exp(\beta X_i)]^n}{n_i!} \tag{2.10}$$

Negative Binomial Model

The main limitation of Poisson regression model is that the variance and mean of crashes need to be equal. In the crash data, the variance of the data is generally larger than mean, known as presence of over-dispersion. Therefore, in order to address this limitation, the

Negative Binomial model is proposed to account for over-dispersion in the data. We can write (Washington *et al.* 2010):

$$\lambda_i = exp(\beta X_i + \varepsilon_i) \tag{2.11}$$

where exp(ε_i) follows a Gamma distribution with mean 1 and variance α. By adding this term, the variance can be different from the mean:

$$var(y_i) = E(y_i)\big(1 + \alpha E(y_i)\big) = E(y_i) + \alpha E(y_i)^2 \tag{2.12}$$

It is worth noting that by approaching α to zero, the model reduces to Poisson model. The distribution of negative binomial model can be written as (Washington *et al.* 2010):

$$P(y_i) = \frac{\Gamma\left(\left(\frac{1}{\alpha}\right) + y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right) y_i!}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}}\left(\frac{\lambda_i}{\frac{1}{\alpha} + \lambda_i}\right)^{y_i} \tag{2.13}$$

where $\Gamma(.)$ is a gamma function. The likelihood function can be written as (Washington *et al.* 2010):

$$L(\lambda_i) = \prod_i \frac{\Gamma\left(\left(\frac{1}{\alpha}\right) + y_i\right)}{\Gamma\left(\frac{1}{\alpha}\right) y_i!}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}}\left(\frac{\lambda_i}{\frac{1}{\alpha} + \lambda_i}\right)^{y_i} \tag{2.14}$$

Geographically Weighted Poisson Regression Model (GWPR)

The availability of recent geo-referenced crash data and increased computational power has provided opportunities to address spatial heterogeneity through rigorous geospatial statistical models (Xu and Huang 2015). One of the most well-known approaches is the *GWPR* model which is utilized to test whether the associations between the independent variables and dependent variable vary substantially across space (Fotheringham *et al.* 2003). We can write:

$$ln(\lambda_i) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)\, ln(E_{vi}) + \sum_{k=1}^{K} \beta_k(u_i, v_i) x_{ik} + \epsilon_i \qquad (2.15)$$

where (u_i, v_i) indicates the coordinates of grid *i.* In GWPR, $\beta_k(u_i, v_i)$ is a function of the location *i* and not randomly distributed. In order to estimate $\beta_k(u_i, v_i)$ we can write (Nakaya *et al.* 2005):

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y \qquad (2.16)$$

where $\hat{\beta}(u_i, v_i)$ is the $n \times 1$ vector of estimated coefficients at grid i, X denotes the matrix of explanatory variables, Y is the vector of crash frequency at each grid, and $W(u_i, v_i)$ is $n \times n$ spatial weight matrix, which can be written as:

$$W(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & w_{in} \end{bmatrix} \qquad (2.17)$$

where w_{ij} reflects the weight of variable *j* at grid *i*. In the *GWPR*, for each grid, a regression equation based on nearby observations is estimated. Each area is weighted

based on the distance from the subject point, where closer areas obtain higher weights than farther areas.

Geographically Weighted Negative Binomial Regression (GWNBR)

In the literature, the majority of studies used *GWPR* to address spatial heterogeneity and model count data. While the major limitation of Poisson models is overdispersion, and it has shown that the overdispersion in the crash data might not be taken into account by the *GWPR* method (Yu and Xu 2018). The negative binomial approach is one of the alternatives to traditional Poisson models because it incorporates the overdispersion factor in the modeling. The negative binomial is a generalization of the Poisson distribution where the dispersion parameter (α) equals 0 (Hilbe 2011). Da Silva and Rodrigues (da Silva and Rodrigues 2014) developed a procedure to estimate the GWNBR model by extending the Iteratively Reweighted Least Square (IRLS) and Newton-Raphson (NR) algorithm. The general form of the model can be written as (da Silva and Rodrigues 2014, Gomes *et al.* 2017):

$$y_i \sim Negative\ Binomial\left[t_j\, exp\left(\sum_k \beta_k(u_j, v_j)x_{jk}\right), \alpha(u_j, v_j)\right] \qquad (2.18)$$

where t_j is the offset variable, and α is the over-dispersion parameter. To estimate the model's parameter β_k and α, the modified version of IRLS with the maximum likelihood method using NR can be used (da Silva and Rodrigues 2014). The log-likelihood of the mode *GWNBR* model can be written as (da Silva and Rodrigues 2014):

$$L(\beta(u_i, v_j)|x_{jk}, y_j, \alpha_j)$$

$$= \sum_{j=1}^{n} \{y_j \, log(\alpha_j \mu_j) - \left(y_j + \frac{1}{\alpha_j}\right) * log(1 + \alpha_j \mu_j) + log\left[\Gamma\left(y_j + \frac{1}{\alpha_j}\right)\right]$$

$$- log\left[\Gamma\left(\frac{1}{\alpha_j}\right)\right]$$

$$- log[\Gamma(y_j + 1)] \qquad (2.19)$$

where

$$\mu_j = t_j \, exp\left(\sum_k \beta_k(u_j, v_j) x_{jk}\right) \qquad (2.20)$$

$$\alpha_j = \alpha(u_j, v_j) \qquad (2.21)$$

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \qquad (2.22)$$

For further details, please refer to (da Silva and Rodrigues 2014). In this procedure, it is crucial to define the optimum bandwidth through minimization of corrected Akaike Information Criteria (*AICc*) (AICc).

$$AIC_c = -2L(\beta, \alpha) + 2k + \frac{2k(k+1)}{n-k-1} \qquad (2.23)$$

where $L(\beta, \alpha)$ is the log likelihood of the model, and k is the effective number of parameters. It has shown that in cases where overdispersion parameter is equal to zero, the GWNBR will be reduced to GWPR (da Silva and Rodrigues 2014).

Measures of Goodness of Fit

In this study, in order to perform model selection and evaluate the performance of the fixed parameter and GWPR models, several criteria were used.

1- *Deviance of the model*: a goodness of fit measure which quantifies deviance of the fitted model from a saturated model. For the Poisson model, we can write:

$$D = 2\sum_{i=1}^{n}(Y_i\, log\left(\frac{Y_i}{\mu_i}\right) - (Y_i - \mu_i)) \qquad (2.24)$$

where Y_i, $\widehat{Y_i}$ and $\bar{y}$ are the observed, and predicted crash frequency at grid *i*, respectively.

2- *AIC*: measures the relative goodness of fit where a lower *AIC* value represents a better model fit (Bozdogan 1987). We can write:

$$AIC = -2LL + 2k \qquad (2.25)$$

where *LL* is the log-likelihood and *k* is the number of model's parameters. For the GWR model, we need to calculate the effective number of parameters. We can write (Nakaya *et al.* 2005):

$$K = trace(S) \qquad (2.26)$$

where S is the hat matrix. More details can be found (Nakaya *et al.* 2005).

3- *Mean Absolute Deviation (MAD):* quantifies the model performance in terms of deviation of predicted number of crashes from observed values. Smaller values imply a better goodness of fit. It can be written as:

$$MAD = \frac{\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right|}{N} \tag{2.27}$$

4- *Mean Squared Error (MSE):* Similar to *MAD*, *MSE* assess the model accuracy by calculating the distance between the observed and predicted values. It can be defined as:

$$MSE = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2}{N} \tag{2.28}$$

where *N* is the total number of grids.

k-means clustering

In this research, in order to identify the hotspot locations, k-means clustering is performed. K-means is one the most common unsupervised machine learning methods used to partition observations into *k* groups, where *k* denotes the number of clusters. The goal of this method is to define clusters in a way where observations in each cluster have high similarity with each other and are dissimilar with observations in other clusters as much as possible.

While the main idea behind k-means is to minimize the total within cluster variation, several methods for perform clustering have been proposed. This study utilizes the

method proposed by Kaufman and Rousseeuw (Kaufman and Rousseeuw 2009). In this method, the total variations within clusters are defined as the sum of squares of Euclidian distances among observations with the centroid of cluster. We can write:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{2.29}$$

where x_i is the observation falls in cluster C_k, and μ_k is the mean of observations in cluster C_k. The assignment of observations to clusters are performed such that the total sum of squares of distances are minimized. We can write:

$$total\ sum\ of\ squares = \sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{2.30}$$

where *K* is the number of clusters. One of the challenges in performing k-means clustering is finding the optimal number of clusters. One of the common methods to find the optimal number of clusters *(k)* is the elbow method. The main idea of this method is to choose a *k* value which minimize total sum of squares error. While by increase in *K*, the *SSE* tends to reach zero, the elbow represents the point where return of increasing *K* will be diminished.

Gaussian Mixture Model

A major drawback of k-means clustering arisen from selecting the cluster center using the mean value. This can be problematic when the means of clusters are close to each other. In other words, k-means approach can be considered as a special case of the GMM, since GMM is more expressive due to grouping the data into clusters irrespective of

cluster shape. The advantage of GMM is considering a Gaussian density function for each component. Considering an independent identically distributed sample of observations ($x = \{x_1, x_2, \dots, x_n\}$, the distribution of each observation can be specified using a probability density function through G components (Scrucca, 2016 #911):

$$f(x_i;\ \Psi) = \sum_{k=1}^{G} \pi_k f_k(x_i; \theta_k) \tag{2.31}$$

where G is the number of components, Ψ is the vector of mixture model parameters (which is unknown and need to be estimated), $f_k(x_i; \theta_k)$ is the density function of kth component for observation x_i, and θ_k is the vector pf mixing probabilities. Since GMM model assumes Gaussian distribution for the density function ($f_k(x_i; \theta_k)$), the clusters are ellipsoidal, and we can write:

$$f_k(x_i; \theta_k) \sim N(\mu_k, \Sigma_k) \tag{2.32}$$

Where μ_k denotes the center of cluster *k*, and Σ_k is the covariance matrix, determining the shape and geometry features of the cluster.

Results

Descriptive Statistics

Descriptive statistics of the independent and dependent variables is provided in Table 2.2. The table provides the descriptive information of the key variables to help conceptualize the variables distribution. The sample size, the number of grids in the study area, is 3007. Based on the descriptive statistics, the average number of crashes in 2013

on each grid was 0.75. Focusing on grid characteristics, the average speed in the grids was 12.96 m/s, and the average recorded curve radius was 850.98 m. The average elevation was 238.65 m, ranging from 2 to 2971 m.

Descriptive statistics of the two levels of driving volatility indices are also provided. There is substantial variation among the volatility measures between grids. As an illustration, time-varying stochastic volatility of speed at both location-based and temporal levels range substantially across grids. Further details can be found in Table 2.2.

Table 2.2 Descriptive Statistics (N=3007 grids)

Variable	Mean	S.D.	Min	Max
Crash frequency	0.75	2.36	0	27
Elevation (m)	238.652	58.596	2.885	2971.97
Speed (m/s)	12.956	6.392	0.180	32.374
$a_x(m/s^2)$	-0.016	0.160	-1.869	1.646
$a_y(m/s^2)$	-0.126	0.533	-9.784	3.442
$a_z(m/s^2)$	-0.730	0.997	-2.986	0.000
Yaw rate	0.011	4.067	-32.582	115.132
Radius of Curve (m)	850.981	912.925	-2422.05	3276.7
Number of BSM Observations	82408	199283.1	101	4562546
Level 1 - Location-based volatility				
L1-Speed-V_f	4.07	2.17	0.06	27.94
L1-Speed-MAD	0.25	0.13	0.01	1.6
L1-Speed-C_v	0.16	0.11	0.01	0.72
L1-Speed-Q_{CV}	0.05	0.02	0	0.19
*L1-Speed-2*S_{dev}	0.59	0.25	0.09	2.84
L1-$Acceleration_x$*-*C_v	0.99	0.28	0.05	3.2
L1-$Deceleration_x$*-*C_v	0.59	0.11	0	0.96
L1-$Acceleration_x$*-*Q_{CV}	0.59	0.11	0	0.97
L1-$Deceleration_x$*-*Q_{CV}	0.43	0.19	0.06	1.75
L1-$AccDec_x$*-MAD*	0.06	0.04	0	0.36
L1-$AccDec_x$*-2*S_{dev}	0.66	0.53	0.03	5.07
L1-$Acceleration_y$*-*C_v	1.35	0.72	0	6.49
L1-$Deceleration_y$*-*C_v	0.67	0.17	0	1
L1-$Acceleration_y$*-*Q_{CV}	0.65	0.16	0	0.99
L1-$Deceleration_y$*-*Q_{CV}	0.43	0.4	0.02	4.96
L1-$AccDec_y$*-MAD*	0.06	0.06	0	0.41
L1-$AccDec_y$*-2*S_{dev}	27.75	16.52	0.98	151.78

Table 2-2 Continued

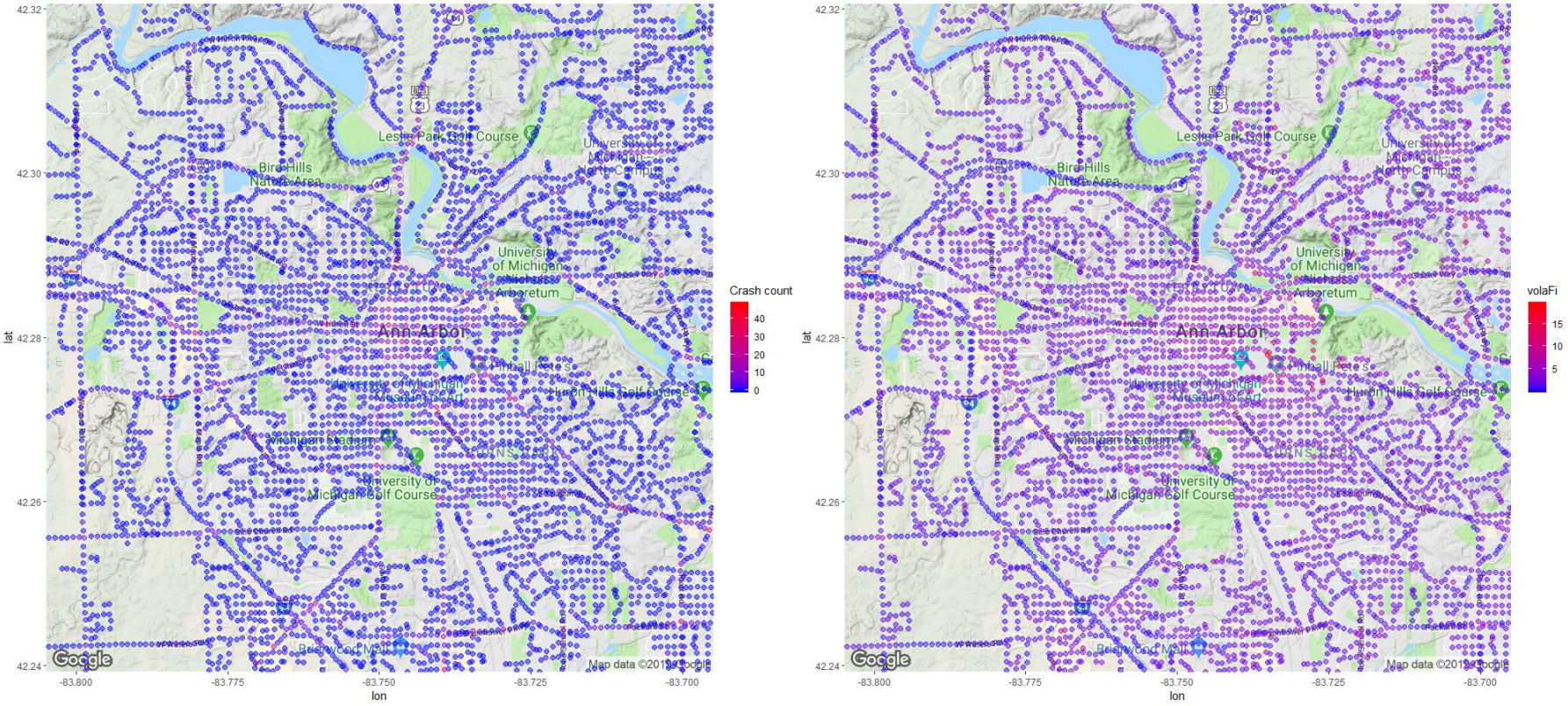
L1-$Deceleration_z$-C_v	1.06	0.55	0.0	4.05
L1-$AccDec_z$-MAD	0.58	0.77	0.0	4.98
L1-$AccDec_z$-$2S_{dev}$	2.92	0.99	0.0	15.61
Level 2 – Temporal volatility				
L2-Speed-V_f	1.86	1.43	0.1	19.5
L2-Speed-MAD	0.81	0.36	0.15	4.19
L2-Speed-C_v	0.07	0.05	0	0.48
L2-Speed-Q_{CV}	0.06	0.04	0	0.41
L2-Speed-$2S_{dev}$	0.02	0	0	0.05
L2-$Acceleration_x$-Q_{CV}	0.48	0.06	0.19	0.8
L2-$Deceleration_x$-Q_{CV}	0.48	0.06	0.21	0.68
L2-$AccDec_x$-MAD	0.37	0.13	0.1	1.35
L2-$AccDec_x$-$2S_{dev}$	0.04	0.01	0.01	0.1
L2-$Acceleration_y$-Q_{CV}	0.43	0.09	0	0.79
L2-$Deceleration_y$-Q_{CV}	0.43	0.09	0	0.7
L2-$AccDec_y$-MAD	0.18	0.11	0	1.61
L2-$AccDec_y$-$2S_{dev}$	0.03	0.01	0	0.08
L2-$Deceleration_z$-Q_{CV}	0.25	0.08	0	0.63
L2-$AccDec_z$-MAD	0.07	0.05	0	0.51
L2-$AccDec_z$-$2S_{dev}$	0.03	0.01	0	0.08

**L1: Location-based volatility measure; L2: Temporal volatility measure; ($2S_{dev}$): % of extreme points beyond mean ± two standard deviation; C_v: coefficient of variation; Q_{cv}: quartile coefficient of variation; D_{mean}: mean absolute deviation; $Acceleration_x$: longitudinal acceleration; $Deceleration_x$: longitudinal deceleration; $AccDec_x$:both longitudinal acceleration and deceleration; $Accleration_y$: lateral acceleration; $Deceleration_y$: lateral deceleration; $AccDec_y$: both lateral acceleration and deceleration; $Accleration_z$: vertical acceleration; $Deceleration_z$: vertical deceleration; $AccDec_z$: both vertical acceleration and deceleration;*

Concept illustration

This section provides a visualization of the relationship between driving volatility and crash frequency at each grid. As mentioned before, the main hypothesis of this research

is that driving volatility is associated with crash frequency. For illustration purpose, Figure 5 depicts the heatmap of crash frequency and speed volatility in terms of *L2-Speed-*V_f on Ann Arbor's transportation network. The volatility values range from 0.079 to 19.502 (Figure 2.5(a)), illustrating a substantial difference across the network. Figure 2.5(b) depicts the crash frequency at the grid level, which ranges from zero to 48. In the figures, blue bubbles represent low volatility/crash frequency, while red bubbles indicate high volatility/crash frequency. Generally, grids with higher volatility values have a higher crash frequency, which supports the initial hypothesis that locations with higher volatility have a higher number of crashes. As an illustration, in the downtown area where the speed volatility in terms of *L2-Speed-*V_f is high, the crash frequency is also high, suggesting a positive correlation between crash frequency and driving volatility.



(a) (b)

Figure 2.5 Visualization of relationship between volatility and frequency of crashes (red points indicates higher volatility/crash frequency)

Modeling Results

This paper explores the association between driving volatility and crash frequency using fixed-parameter Poisson, Negative Binomial, *GWPR*, and *GWNBR* models. The *GWPR* and *GWNBR* models help us address unobserved heterogeneity and spatial correlation among observations. In order to estimate the models, factors extracted from CV data in each grid including vehicle kinematics, two levels of driving volatility, and geographic information, are utilized. While the goal of this study is not to compare modeling approaches, we provided model comparison results to shed light on the goodness of fit performances of the estimated models. Finally, the modeling results of the fixed-parameter and geographically weighted Poisson and Negative Binomial regression models will be discussed.

Model comparison

Several measures can be used to compare non-spatial modeling techniques and geographically weighted regression models (i.e. Poisson and Negative Binomial): $R^2_{poisson}$, *AIC*, *MAD* and *MSE* (Table 2.3). By taking into account the overdispersion parameter, it can be inferred that the non-spatial negative binomial model improved the performance of the Poisson model. However, the *GWPR* model outperformed the *NB* model. The best performance is obtained by the *GWNBR* model.

Table 2.3 Measures of goodness of fit for the fitted model

	Goodness of fit	Poisson	Negative Binomial	GWPR	GWNBR
Crash Frequency	*Deviance*	3246.8	1747.2	2110.5	1676.9
	AIC	5136.58	4495.18	5055.97	4391.66
	MAD	0.672	0.811	0.487	0.313
	MSE	2.754	9.827	1.189	1.065

Model estimation

As the descriptive statistics and concept illustration sections earlier depict the meaningful relationship between driving volatility and crash frequency, this section aims to quantify the association between volatility measures and frequency of crashes in the network. Given the count nature of crash data, this paper considers fixed parameter and geographically weighted regression models that model the number of crashes at each grid (Table 2.4), which are the level of analysis in these models (N=3007 grids). The model selection procedure is performed based on intuition, statistical significance, and model parsimony. Since the *GWNBR* model outperformed other models in terms of goodness of fit, this section only discusses that model's coefficients. The local estimation of coefficients in the *GWNBR* model are mapped across the city by applying the Inverse Distance Weighted (*IDW*) interpolation (Figure 2.6).

In the model estimation, three sets of variables were considered: (1) geometric information collected by CVs of each grid (average of elevation, radius, and observations as a proxy for traffic exposure), (2) temporal volatility measures, and (3) location-based volatility measures. Referring to geometric information, modeling results suggest that

grids in higher elevations experience more crashes. The coefficient of elevation is negative in the Poisson and Negative Binomial models, while the *GWNBR* and *GWPR* models suggest that its association is positive at some grids (7.88 percent). However, they are not significant in these models. As mentioned before, the number of BSM observation variable can be a proxy for AADT by assuming that locations with more passings by CVs have higher AADT. The results suggest that grids with higher numbers of observations have more crashes, which is consistent with the findings of previous studies (Chen and Xie 2016, Xie *et al.* 2018, Arvin *et al.* 2019c). The modeling results suggest that grids with higher average speed have lower number of crashes, which is in line with the literature (Imprialou *et al.* 2016, Yu *et al.* 2018).

Referring to temporal volatility measures, speed volatility measures in terms of *L2-Speed-*V_f and *L2-Speed-*D_{mean} are significantly correlated with frequency of crashes, implying that locations with higher speed volatility have higher crash frequency. The results are intuitive in a sense that locations with higher variations in vehicular speed have higher risk of crash compared to other locations, which is consistent with the findings of previous studies (Vadeby and Forsman 2017, Kamrani *et al.* 2018b). Modeling results of *GWNBR* suggest that the strength of this association varies significantly across the study area, being higher south of the city. Although 4.55 percent of *L2-Speed-*V_f and 7.51 percent of *L2-Speed-*D_{mean} estimations are negative, they are not significant. Furthermore, the results reveal that the vertical volatility measure *L2-* $AccDec_z$ *-* D_{mean} is negatively associated with crash frequency (4.12 percent of the *GWNBR* estimations are positive but are not significant). Since vertical volatility is a function of vertical alignment, this association may be because drivers at these locations might be more cautious, leading

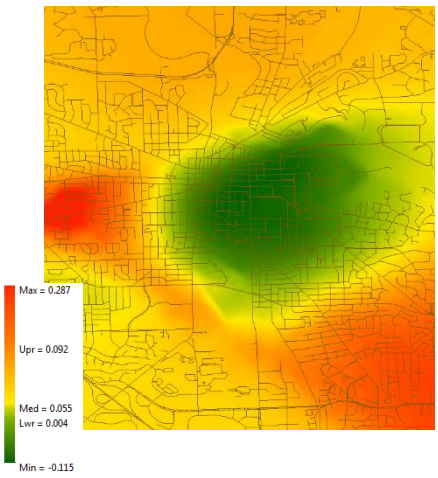
to a decrease in the number of crashes.

Focusing on location-based volatility measures, modeling results reveal that speed volatility measures such as *L1-Speed-*V_f and *L1-Speed-*D_{mean} are significantly correlated with the number of crashes, and these correlations vary significantly and substantially across the city. The results are in line with literature which had shown that higher speed volatility is correlated with crash counts (Kamrani *et al.* 2018b, Arvin *et al.* 2019c). In addition, variations in the location-based longitudinal deceleration (*L1-*$Deceleration_x$*-2*S_{dev}) and the location-based lateral deceleration volatility measure (*L1-*$Deceleration_y$*-*C_v) are highly correlated with crash frequency. It can be inferred that this association is higher north of the city where the roadway transportation network has several sharp curves, contrary to the structure of the downtown area (Figure 2.6).

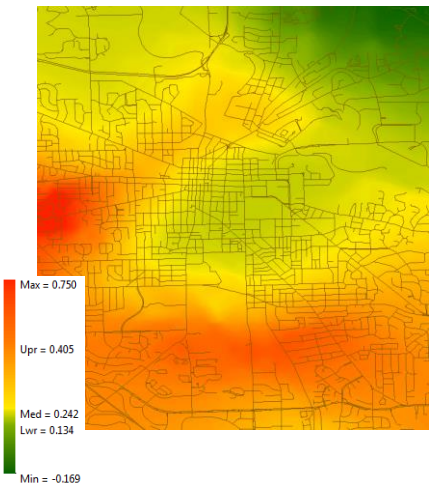
Table 2.4 Modeling results of crash frequency model

	Poisson		Negative Binomial		Geographically Weighted Poisson model						Geographically Weighted Negative Binomial model					
Variable	β.[1]	Std. Err.	β.	Std. Err.	Mean	Min	1st Q	Med	3rd Q	Max	Mean	Min	1st Q	Med	3rd Q	Max
Intercept	-2.2728***	0.337	-3.321***	0.511	-3.017	-33.890	-6.788	-3.413	-0.103	42.427	-3.147	-8.551	-4.856	-2.746	-1.406	1.160
Speed	0.015*	0.008	0.0263*	0.012	0.105	-0.389	-0.028	0.105	0.205	0.554	0.048	-0.115	0.004	0.055	0.092	0.287
Elevation	-0.0056***	0.001	-0.0049***	0.002	-0.011	-0.191	-0.025	-0.003	0.002	0.115	-0.007	-0.025	-0.012	-0.005	-0.002	0.006
# of BSM Observations	0.823***	0.178	1.014***	0.034	0.942	0.299	0.757	0.912	1.086	1.800	0.001	0.001	0.001	0.001	0.001	0.001
L1-Speed-V_f	0.0035*	0.002	0.0028	0.003	0.0005	-0.060	-0.012	0.000	0.013	0.065	0.003	-0.021	-0.003	0.002	0.007	0.024
L1-Speed-MAD	0.0609***	0.009	0.0815***	0.018	0.1122	-0.300	0.043	0.103	0.191	0.576	0.111	-0.009	0.079	0.115	0.143	0.213
L1-$AccDec_x$-$2S_{dev}$	1.9234***	0.512	2.3452**	0.839	0.606	-21.548	-2.345	1.573	4.383	17.163	2.093	-5.486	0.786	2.092	3.834	8.570
L1-$Deceleration_y$-C_v	0.1591***	0.027	0.1768***	0.047	-0.025	-1.010	-0.157	0.005	0.144	0.609	0.126	-0.130	0.066	0.107	0.196	0.328
L2-Speed-V_f	0.1948***	0.022	0.2259***	0.038	0.300	-0.967	0.049	0.282	0.530	1.745	0.259	-0.169	0.134	0.242	0.405	0.750
L2-Speed-MAD	0.8***	0.088	0.661***	0.144	0.875	-3.168	0.254	0.943	1.517	4.455	0.678	-0.213	0.353	0.725	1.031	1.533
L2-$AccDec_z$-MAD	-9.2759***	0.085	-5.6988***	1.300	-5.969	-54.981	-10.197	-5.625	0.711	18.538	-6.007	-16.277	-8.284	-5.440	-2.886	2.138
Disp. Param.	-	-	0.907***	0.101	-	-	-	-	-	-	0.785	0.510	0.649	0.760	0.884	1.193
Null Deviance	8850.4		8850.4		8850.4						8850.4					
Model Deviance	3246.8		1747.2		2110.5						1676.9					
Explained Dev.	0.633		0.802		0.762						0.810					
AIC	5136.58		4495.18		5055.9						4391.6					

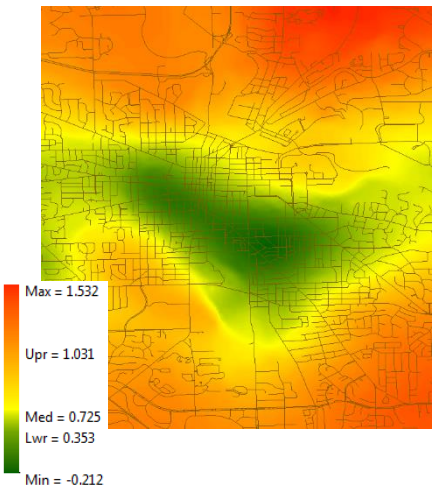
[1] Significance at *** 1%, ** 5% and * 10%



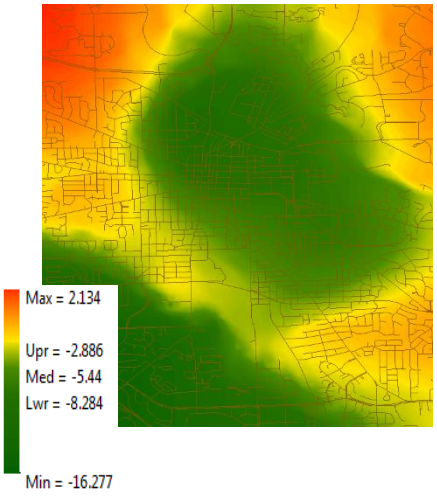
Speed



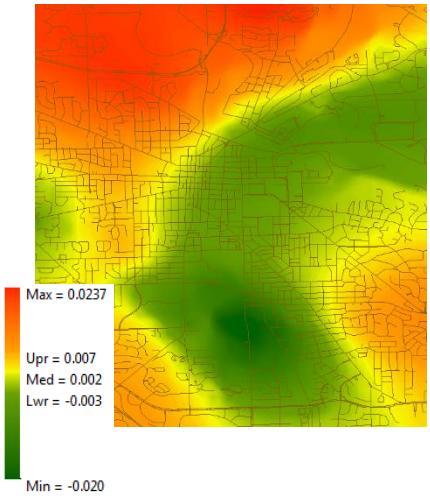
L2-Speed-V_f



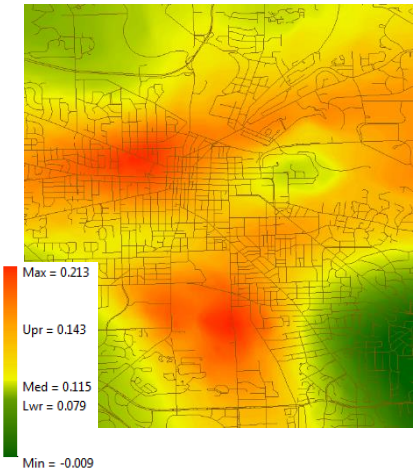
L2-Speed-MAD



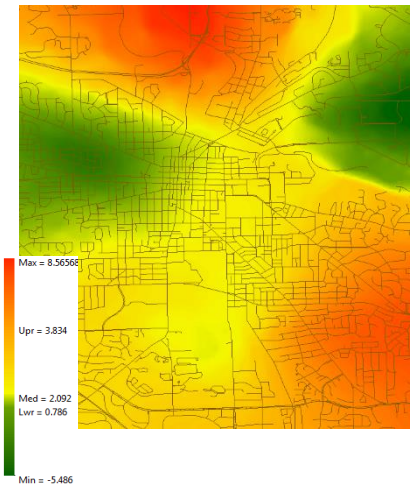
L2-Acc-Az-MAD



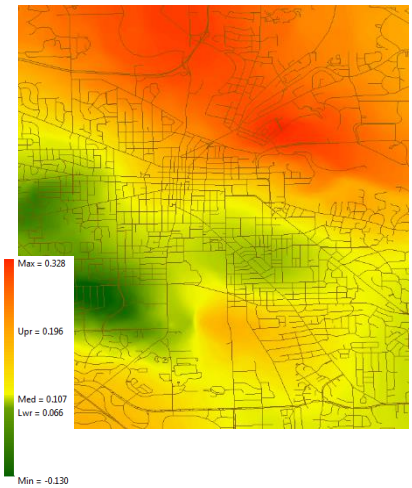
L1-Speed-V_f



L1-Speed-MAD



L1-Acc-Ax-2S_{dev}



L1-Acc-Ay- Cv

Figure 2.6 Estimated coefficient of the GWNBR model

Hotspot identification

In the modeling section, we investigated the association of driving volatility and crash risk in the Ann Arbor city network. The results revealed that there is a positive and significant correlation between driving volatility indices and crash frequency. These significant measures (that are highly contributing with the crash frequency) can be used to group the locations and to proactively identify locations where driving volatility is high but crash frequency is low. These locations may not receive the same amount of attention because of historic crash frequency statistics, but their high levels of driving volatility indicate possible higher crash frequencies in the future. The first step is to group the grids considering those volatility measures that are significant in the model. To reach this goal, an unsupervised learning technique (i.e. k-means and GMM clustering) is performed by taking the following steps:

1. Find the optimum number of clusters (*k*)
2. Perform k-means and GMM clustering with the optimal number of *k* groups
3. Identify clusters with high volatility and find locations with low crash frequency

The optimal number of clusters which minimize total intra-cluster variation in clusters need to be determined. In order to identify the optimal number of clusters, the elbow method is utilized. It can be observed that the optimal number of clusters in which the grids can be grouped is 3 (*k*=*3*) (Figure 2.7).

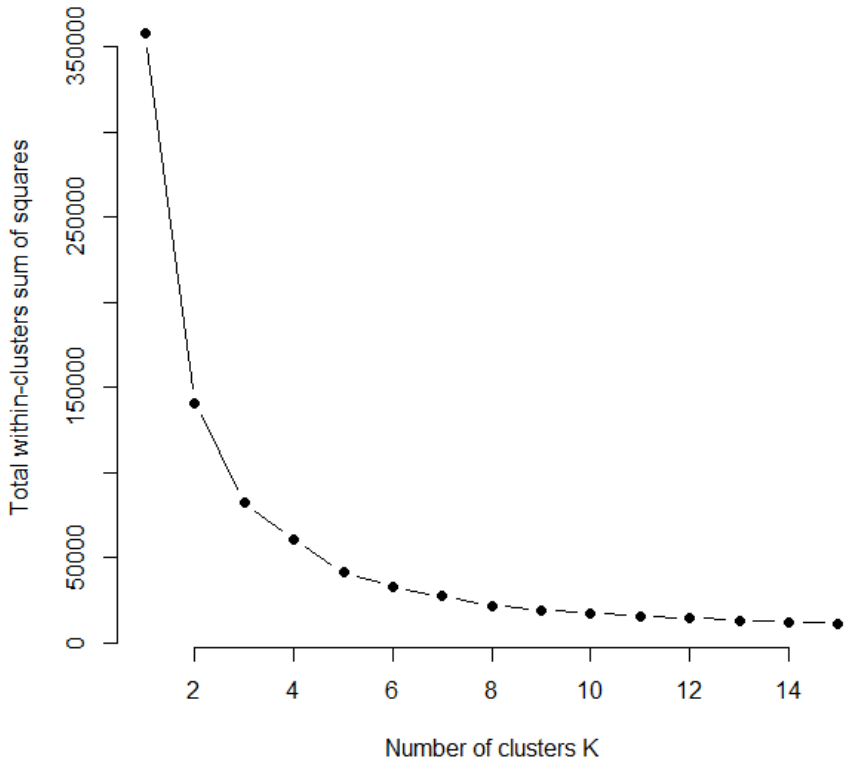


Figure 2.7 Elbow method to find the optimal number of clusters

The optimal number of clusters which minimize total intra-cluster variation in clusters need to be determined. In order to identify the optimal number of clusters, the elbow method is utilized. It can be observed that the optimal number of clusters in which the grids can be grouped is 3 (*k=3*) (Figure 2.7).

Given that optimal number of clusters is equal to three (*k=3*), the within clusters sum of squares for k-means and GMM are compared and the results revealed that GMM model performs better. Therefore, we have focused on the GMM clustering method. Utilizing the GMM, clustering analysis is performed and the grids are grouped into three categories. A radar plot is used to assess the level of each volatility measure in each cluster's locations (Figure 2.8). Based on the results, the clusters reflect "*low volatility*", "*normal*", and "*high*

volatility" locations. It can be observed that the association of volatility measures are substantially higher in the "*high volatility*" group compared to others. Table 2.5 provides descriptive statistics of the volatility measures for the three extracted clusters. The descriptive statistics show that there is a substantial difference in the volatility measures among the three extracted clusters.

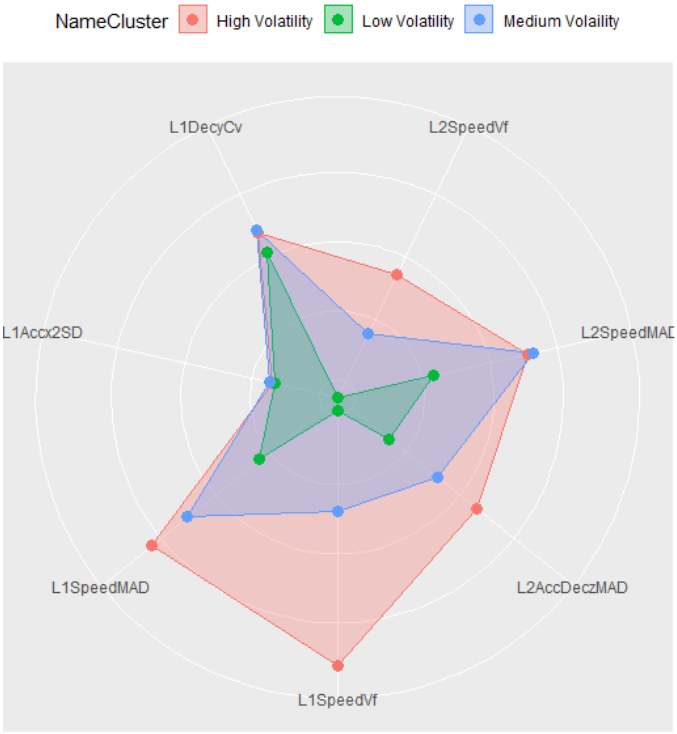


Figure 2.8 Association of volatility measures in low, medium and high volatility clusters

Table 2.5 Descriptive statistics of volatility measures for each cluster

Variable	Low volatility		Medium volatility		High Volatility	
	Mean	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
L2-Speed-V_f	0.85	0.71	1.8	1.19	2.76	1.99
L2-Speed-MAD	0.58	0.33	0.86	0.37	0.85	0.33
L2-$AccDec_z$*-MAD*	0.04	0.04	0.07	0.04	0.08	0.06
L1-Speed-V_f	12.57	5.35	31.82	6.66	61.04	16.94
L1-Speed-MAD	3.09	1.98	5.08	2.42	5.93	2.99
L1-$AccDec_x$*-2*S_{dev}	0.05	0.04	0.06	0.05	0.06	0.05
L1-$Deceleration_y$*-*C_v	1.2	0.78	1.36	0.72	1.33	0.73

After identifying locations with high volatility, the next step is to identify hotspot locations where the number of accidents is low while the driving volatility measures are high. We can focus on the “High volatility” grids to identify potential hotspots. The plot of crash count vs driving volatility for the “High volatility” cluster is provided in Figure 2.9. Given that all of these grids have significantly higher volatility indices comparing to other grids, we can identify hotspot locations (shown with the red eclipse) where crashes are waiting to happen despite historically low crash frequencies.

Limitations

In this study, driving volatility is utilized as the surrogate safety measure, but the literature has previously explored multiple different safety surrogate measures in the modeling process. Although different criteria were used to check for erroneous data, some errors might remain during the data collection. Furthermore, while police reports are the main source of crash data, they have a tendency to under-report certain types of crashes. Specifically, the National Highway Traffic Safety Administration report (NHTSA 2009) states that 50% of property damage only crashes and 25% of minor injury crashes are unreported. In addition, drivers in this study might not truly represent the driving behavior of population.

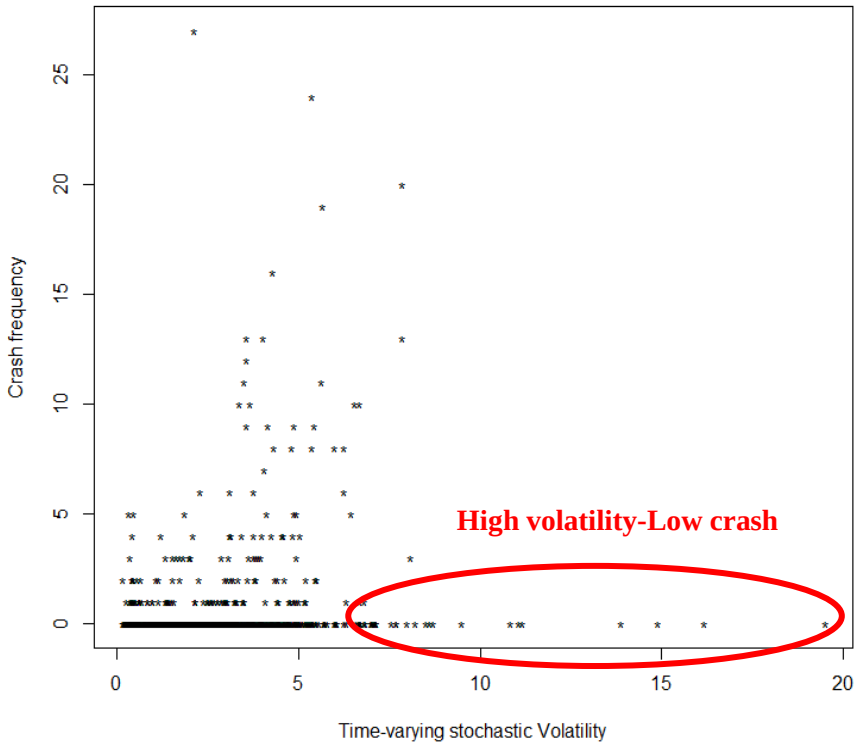


Figure 2.9 Crash frequency vs L2-Speed-V_f volatility for high-volatility cluster

Conclusion and future research

While driving behavior is known as a leading cause of crashes, it has received a relatively small amount of attention in evaluations of transportation network safety. Due to the unavailability of real-world high-resolution data, historical approaches mainly consider exposure and geometric information and largely ignore the human behavior side. The emergence of big data generated by connected and automated vehicles and powerful computational resources have helped researchers incorporate proactive methods to identify hotspot locations. This paper develops a methodological framework for proactively monitoring the safety performance of the network by harnessing big data generated by CVs to bring in the human-behavior side of crash occurrence through the concept of driving volatility. By generating additional features from CV data, correlation of extracted features and crash frequency is explored.

The Safety Pilot Model Deployment (SPMD) study data, which collected data on more than 2800 connected vehicles and contains more than 2.2 billion BSM observations, is utilized. Significant effort and time were taken to process and analyze such a big data to extract useful information and link it with crash data. To study the drivers' behavior, this paper introduces the concept of temporal driving volatility which generates temporal volatility indices quantifying spatiotemporal variations in driving behavior. Two levels of driving volatility are considered in order to model crash frequency: *Level 1* - location-based volatility, and *Level 2* - temporal driving volatility.

From the methodological perspective, Geographically Weighted Negative Binomial

(*GWNBR*) and Poisson Regression (*GWPR*) models are estimated to address unobserved heterogeneity and spatial correlation in the data. This study overcomes the limitation of the *GWPR* model that ignores overdispersion in the crash data. Modeling results reveal that the *GWNBR* model, by incorporating spatial overdispersion, has a significantly better goodness of fit than other models, which is consistent with the findings of previous studies (da Silva and Rodrigues 2014, Gomes *et al.* 2017). This study is one of the first studies to apply the *GWNBR* model on crash data in order to address overdispersion. Modeling results also reveal that driving volatility both at the temporal and location-based levels is highly correlated with crash frequency. Variations in longitudinal control is highly correlated with crash frequency, and volatile lateral and vertical movements also increase crash risk. The results suggest that the magnitude of this association varies significantly and substantially across space.

Finally, hotspot identification is performed by applying an unsupervised classification approach. Given the association of driving volatility and crash frequency, k-means and GMM clustering identified locations in the network with high levels of driving volatility. This study defined hotspots where driving volatility is high, but crash frequency is low. The results identified grids where the behavior of drivers significantly differed from those at other locations, and this difference might lead to higher levels of crash risk previously unnoticed due to low crash frequencies. Further examinations are needed at these locations to find the correlates of driving volatility, such as roadway geometry design, traffic conflict, signal timing, etc.

While this study explores the association of total number of crashes with driving volatility,

future studies can investigate the association of crash types with volatility in longitudinal, lateral, and vertical directions. Furthermore, with increases in CV penetration rates, researches can utilize other surrogate safety measures such as time-to-collision (TTC), and the rear-end crash risk and potential index (RCRI and CPI) (Essa and Sayed 2018, Rahman and Abdel-Aty 2018, Rahman *et al.* 2018, Zhao and Lee 2018) and integrate this information into modeling process. These measures require instantaneous information of the lead vehicle, which is not available in the SPMD data.

CHAPTER 3 : HOW INSTANTANEOUS DRIVING BEHAVIOR CONTRIBUTES TO CRASHES AT INTERSECTIONS: EXTRACTING USEFUL INFORMATION FROM CONNECTED VEHICLE MESSAGE DATA

A version of this chapter is presented at the 98th Transportation Research Board Annual Meeting and published in the Accident Analysis and Prevention journal.

Arvin, R., Kamrani, M., & Khattak, A. J. (2019). How instantaneous driving behavior contributes to crashes at intersections: extracting useful information from connected vehicle message data. *Accident Analysis & Prevention*, *127*, 118-133.

Abstract

Connected and automated vehicles have enabled researchers to use big data for development of new metrics that can enhance transportation safety. Emergence of such a big data coupled with computational power of modern computers have enabled us to obtain deeper understanding of instantaneous driving behavior by applying the concept of "driving volatility" to quantify variations in driving behavior. This paper brings in a methodology to quantify variations in vehicular movements utilizing longitudinal and lateral volatilities and proactively studies the impact of instantaneous driving behavior on type of crashes at intersections. More than 125 million Basic Safety Message data transmitted between more than 2800 connected vehicles were analyzed and integrated with historical crash and road inventory data at 167 intersections in Ann Arbor, Michigan, USA. Given that driving volatility represents the vehicular movement and control, it is expected that erratic longitudinal/lateral movements increase the risk of crash. In order to capture variations in vehicle control and movement, we quantified and used 30 measures of driving volatility by using speed, longitudinal and lateral acceleration, and yaw-rate. Rigorous statistical models including fixed parameter, random parameter, and geographically weighted Poisson regressions were developed. The results revealed that

controlling for intersection geometry and traffic exposure, and accounting unobserved factors, variations in longitudinal control of the vehicle (longitudinal volatility) are highly correlated with the frequency of rear-end crashes. Intersections with high variations in longitudinal movement are prone to have higher rear-end crash rate. Referring to sideswipe and angle crashes, along with speed and longitudinal volatility, lateral volatility is substantially correlated with the frequency of crashes. When it comes to head-on crashes, speed, longitudinal and lateral acceleration volatilities are highly associated with the frequency of crashes. Intersections with high lateral volatility have higher risk of head-on collisions due to the risk of deviation from the centerline leading to head-on crash. The developed methodology and volatility measures can be used to proactively identify hotspot intersections where the frequency of crashes is low, but the longitudinal/lateral driving volatility is high. The reason that drivers exhibit higher levels of driving volatility when passing these intersections can be analyzed to come up with potential countermeasures that could reduce volatility and, consequently, crash risk.

Introduction

The need for a safer and more sustainable transportation system has pushed the public and private sectors to improve the performance of the network. Connected Vehicles (CV) provide enriched data such as instantaneous driving behavior, maneuvers, trajectory, individual origin and destination, and traffic data which previously were not obtainable. These data can be transmitted via vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication which can be incorporated to gain precise information to monitor and evaluate the performance of the system (Ghiasi *et al.* 2017, Nezafat *et al.* 2018). The National Highway Traffic Safety Administration (NHTSA) has announced that

communication between vehicles will become mandatory in the near future. In order to advance V2V and V2I technology, the U.S. Department of Transportation developed the Safety Pilot Model Deployment (SPMD) study. The SPMD is one of the most successful studies to implement V2V and V2I communication in the real-world environment (Henclewood *et al.* 2014), and is one of the largest vehicle communication test-bed by incorporating more than 2800 instrumented vehicles and more than 70 miles of roadway instrumented with Road Side Units (RSU) in Ann Arbor, MI (Henclewood *et al.* 2014). In this experiment, CVs and RSUs were capable of communicating via Dedicated Short-Range Communication (DSRC) at a frequency of 10 Hz (Henclewood *et al.* 2014). The emergence of Big Data provided by CVs, RSUs, and other sources of information provides opportunities for researchers to innovate and implement new concepts aiming to increase safety, mobility and moves toward sustainability.

From the safety perspective, previous studies reveal that rear-end and sideswipe crashes are the most frequent type of crash at signalized intersections (Wang and Abdel-Aty 2006). On average, rear-end and sideswipe crashes are the least dangerous type of collision, while head-on and angle crashes are the most dangerous ones (Paleti *et al.* 2010). According to U.S. traffic safety facts for the year 2015, while 4.1% of all crashes were head-on collisions, they contribute to 10.2% of fatal crashes (NHTSA 2015). As a result, researcher pay a great amount of attention to decrease the frequency and severity of head-on crashes.

Given the importance of type of crashes, this study explores the impact of instantaneous driving behavior on multiple crash types at intersections. The study utilizes “driving

volatility", a newly developed concept in transportation (Wang *et al.* 2015a, Kamrani *et al.* 2018b, Kamrani *et al.* 2018c) which captures variations in vehicular movements, as an indicator for driving behavior at intersections. This study extends the concept of driving volatility to longitudinal and lateral volatilities and explores the correlation between volatilities with rear-end, sideswipe, angle and head-on crashes. The main goals of this research are to:

1. Develop a framework for capturing and quantifying longitudinal and lateral driving volatilities using real-world instantaneous driving data.
2. Evaluate correlations between longitudinal and lateral volatilities with frequency of multiple crash types at intersections.
3. Account for unobserved heterogeneity by utilizing random parameter and semi-parametric geographically weighted Poisson regression models.

Since human-error contributes to 94 percent of crashes in the U.S (Anon 2008), findings from this study can help agencies proactively identify hazardous intersections where there is a substantial variations in driving behavior by utilizing the concept of driving volatility. Proactively countermeasures might apply to reduce driving volatilities to prevent future crashes.

Methodology

Modeling Approach

Traditionally, to model the crash frequency, the count-data models such as Poisson, Negative Binomial and Zero Inflated Models are commonly utilized (Abdel-Aty and Radwan 2000, Azizi and Sheikholeslami 2012, Jamali and Wang 2017) due to the fact that crash counts are non-negative integer values in a specific period of time (Anastasopoulos and Mannering 2009). In this study, fixed parameter Poisson regression model, the random parameter Poisson regression model, and the geographically weighted Poisson regression model (GWPR) were used to model crash frequency.

Poisson Model

In the Poisson regression model, the probability of occurrence of *n* crashes at intersection *i* can be written as (Greene 2003):

$$P(n_i) = \frac{\lambda_i^{n_i} \, exp(-\lambda_i)}{n_i!} \tag{3.1}$$

where λ_i (Poisson parameter) is the expected number of crashes for intersection *i*, $E(n_i)$. In order to fit the regression model, the Poisson parameter, λ_i, is written in the logarithm form (Greene 2003):

$$ln(\lambda_i) = \beta X_i \tag{3.2}$$

where X_i is the matrix of the independent variables and β is a vector of the estimated coefficients. The Poisson function defined in Equation 1 and 2 is maximized by the

maximum likelihood with the following function (Washington *et al.* 2010):

$$L(\beta) = \prod_i \frac{exp[-exp(\beta X_i)]\,[exp(\beta X_i)]^n}{n_i!} \tag{3.3}$$

It should be noted that in cases where the mean and the variance of the dependent variable are not equal, applying the Poisson regression might lead to misleading results. Therefore, in order to test the over-dispersion existence in the Poisson model, the Lagrange multiplier method was performed (Greene 2003). We can write:

$$LL = \left(\frac{\sum_{i=1}^{N}((y_i - \mu_i)^2 - y_i)}{2\sum_{i=1}^{N}\mu_i^2}\right)^2 \tag{3.4}$$

where y_i and μ_i are the observed and predicted crash frequency at the intersection *i*, and *N* is the number of intersections.

Random Parameter Poisson Model

In this approach, unobserved heterogeneity, arising from unobserved contributing factors, is addressed by developing a random parameter model using simulated maximum likelihood estimation (Greene 2003). The RP Poisson regression model is an important method because it accounts for heterogeneity arising from factors relating to traffic characteristics, vehicle types, road geometry, pavement conditions, time of day and other unobserved factors (Anastasopoulos and Mannering 2009). The formulation for

estimating the coefficients of the RP Poisson model is (Greene 2003):

$$\beta_i = \beta + \varphi_i \tag{3.5}$$

where φ_i is a randomly distributed term with a specified distribution. The log-likelihood function is (Anastasopoulos and Mannering 2009):

$$LL = \sum_i ln \int_{\varphi_i}^{i} g(\varphi_i) P(n_i|\varphi_i) d\varphi_i \tag{3.6}$$

where g(.) is the pre-specified distribution of φ_i. In this study, the Halton draws simulation approach is utilized, which is the most popular simulation approach as it provides a more efficient distribution than other methods (Train 2000a, Bhat 2003).

Geographically Weighted Poisson Regression Model

The availability of geo-referenced crashes coupled with computational power has enabled researchers to develop rigorous geospatial models that account for spatial heterogeneity by allowing parameters to vary across space (Xu and Huang 2015). The Geographically Weighted Poisson Regression (GWPR) can be used to test whether the relationship between the explanatory variables and the dependent variable substantially varies across space (Fotheringham *et al.* 2003, Liu *et al.* 2017). The model can be written as:

$$ln(\lambda_i) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i) ln\,(E_{vi}) + \sum_{k=1}^{K} \beta_k(u_i, v_i) x_{ik} + \epsilon_i \quad (3.7)$$

where (u_i, v_i) denotes the coordinates of *i.* It should be noted that in GWPR, $\beta_k(u_i, v_i)$ is not randomly distributed, but rather is a function of the location *i*. The following equation can be used to estimate $\beta_k(u_i, v_i)$:

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y \quad (3.8)$$

where $\hat{\beta}(u_i, v_i)$ is the vector of estimated coefficients at location *i*, *X* is the matrix of independent variables, *Y* is the *n×1* vector of the number of crashes at each intersection, and $W(u_i, v_i)$ is *n×n* spatial weight matrix:

$$W(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & w_{in} \end{bmatrix} \quad (3.9)$$

where w_{ij} is the weight of variable *j* at location *i*. In this approach, based on observations at nearby areas, a regression equation is estimated for each location. Based on the distance from the regression point each area is weighted (areas that are closer have a higher weight than ones that are farther). The W matrix can be estimated using the adaptive Gaussian Kernel function:

$$w_{ij} = exp\left(-\frac{d_{ij}^2}{{\theta_{i(N)}}^2}\right) \quad (3.10)$$

where d_{ij} is the Euclidean distance between area *i* and *j*, $\theta_{i(N)}$ is the adaptive bandwidth defined by the *N*th nearest neighbor. In this formulation, the Gaussian Kernel bandwidth is adaptive, meaning that the weight function magnitude varies across all intersections.

In this study, along with the adaptive Gaussian kernel, adaptive bi-square kernel was considered, which can be written as:

$$w_{ij}\begin{cases}\left(1-\left({}^{d_{ij}}\!/_{d_{iN}}\right)^2\right)^2 & if\ d_{ij} < d_{iN} \\ 0 & otherwise\end{cases} \qquad (3.11)$$

where d_{iN} denotes the distance to the *N*th nearest neighbor of intersection *i*.

It is worth mentioning that applying fixed bandwidth kernel, the local coefficients in areas with sparse intersections is estimated with limited points, leading to high standard error in estimation and unreliable results. Thus, in this study adaptive kernel was employed which tries to overcome this issue by letting the bandwidth vary based on the data's sparsity. To determine the bandwidth of the adaptive kernel, the corrected Akaike Information Criteria (AICc) (Hurvich *et al.* 1998) was used. The best model is the one with the lowest AICc score (Fotheringham *et al.* 2003, Hadayeghi *et al.* 2010a).

As previously mentioned, there was a probability that some of the coefficients in the model do not significantly vary across space. In this case, the semi-parametric GWPR (S-GWPR) is ideal where some of the parameters vary spatially, while others are held fixed. We can write (Nakaya *et al.* 2005, Xu and Huang 2015):

$$ln(\lambda_i) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)\, ln(E_{vi}) + \sum_{j=2}^{l} \beta_j x_{ij} + \sum_{k=1}^{k} \beta_k(u_i, v_i) x_{ik} + \epsilon_i \quad (3.12)$$

where β_j is the *j*th estimated global variable. In order to evaluate the existence of variation in the estimated coefficients across space (spatial variation), the non-stationarity test was performed. Given 167 intersections, the GWPR model suggests specific coefficients for each observation. The non-stationarity test calculates the difference between the upper and lower quartile of the estimated coefficients from GWPR and performs the evaluation. We can write:

$$Delta = \beta_{upper} - \beta_{lower} \quad (3.13)$$

$$\begin{cases} Delta > 1.96 * SE \text{ and } Delta > max(|t_i|) & Pass\ the\ test\ (local\ coefficient) \\ if\ not & failed\ to\ pass\ (global\ coefficient) \end{cases} \quad (3.14)$$

where *SE* is the standard error of the coefficient in the global Poisson model, and $|t_i|$ is the significance t-value of the GWPR model at intersection *i* which can be calculated as $\left|\frac{\beta(u_i,v_i)}{SE(u_i,v_i)}\right|$. If *Delta* is greater than *1.96*SE* and max of $|t_i|$ is greater than 1.96, then the test is passed and there are substantial variations among the estimated coefficients across the space. Otherwise, the test failed and the coefficient is considered as the global coefficient. Obviously, if all the variables are estimated as local coefficients, the S-GWPR model is equivalent to the GWPR model. For further details regarding the S-GWPR calibration, please refer to (Nakaya *et al.* 2005).

It should be noted that GWPR provides a set of local coefficients at each intersection. To

map the GWPR results across space, the Inverse Distance Weighted (IDW) method was applied (Bartier and Keller 1996). The goal of this approach is to create a continuous coefficient surface that interpolates and maps the results across the space. IDW assigns the value to unknown locations based on the estimated coefficients for the nearby areas. The assigned value obtained by weighting the nearby coefficients based on their distance from the unknown point. We can write:

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i) \tag{3.15}$$

where $\hat{Z}(s_0)$ is the predicted coefficient at location s_0, *N* is number of known sample points surrounding the location s_0, λ_i are the assigned weights to each measured coefficient, and $Z(s_i)$ is the observed coefficient at location s_i. To determine the weights, we can write:

$$\lambda_i = \frac{d_{i0}^{-2}}{\sum_{i=1}^{N} d_{i0}^{-2}} \tag{3.16}$$

where d_{i0} represents the distance between point *i* and *o*. It can be inferred that by increasing the distance between the unknown coefficient and observed coefficient, the weight of the observed point will decrease.

In order to estimate S-GWPR model, GWR4.0 software which is developed by Nakaya et al. (Nakaya *et al.* 2012) was used.

Location-based volatility, measures and calculation

The concept of location-based volatility attempts to develop a meaningful process on instantaneous driving behavior and decisions in order to generate driving volatility measures at intersection/segment level (Wali *et al.* 2018a). These volatility measures can potentially be representative of the driving behavior of majority of drivers passing the study area (Wali *et al.* 2018a). Such volatility indices can be utilized to identify locations that driving behavior is different compared to driving behavior of same drivers at other locations. In addition, the correlation between volatility measures and frequency of various crash types can be investigated.

Multiple volatility measures were used by researchers to capture variations in longitudinal control of the vehicle (Arvin *et al.,* Wang *et al.* 2015a, Kamrani *et al.* 2017, Kamrani *et al.* 2018b, Arvin *et al.* 2019b, Kamrani *et al.* 2019) and have been applied to speed, acceleration, and jerk. However, one of the main drawbacks of the previous studies is ignorance of lateral movement of vehicles which potentially could be contributing to crash frequency. Therefore, in this study, volatility functions were applied to speed, longitudinal, lateral acceleration, and yaw-rate at the level of intersections, which were available from connected vehicle BSM data from SPMD. The data is representative of 3-4 percent of total driving in Ann Arbor, MI (Shou and Di 2018). The data provides high-resolution microscopic driving decisions and vehicle motions in terms of position, speed, acceleration, and yaw-rate with a frequency of 10 Hz. Given that, three groups of volatilities are identified and calculated for the selected 167 intersections in Ann Arbor, MI (discussed later in details):

- Speed volatility
- Longitudinal acceleration volatility
- Lateral acceleration volatility.
- Yaw-rate volatility

In order to process and calculate volatility indices at intersection level, 150-ft polygons were established from the center of intersections,[2] and BSM data was assigned to each intersection by processing more than 220 million BSM data. It should be noted that due to the difference in speed profile of vehicles at signalized and unsignalized intersections, and signal timing of signalized intersections, zero speeds were removed from the data prior to volatility calculation. For selected intersection, multiple volatility functions are applied on the speed, longitudinal and lateral acceleration, which presented in Table 3.1. For more details on volatility functions, please refer to (Kamrani *et al.* 2018b).

[2] Although 250-ft threshold from the center of intersection is a common threshold as an intersection influence area, in In this study we chose 150-ft threshold due to two main reasons. First, the network of Ann Arbor city is dense, and intersections are close, and using 250-ft threshold leads to overlapping territories. Therefore, 150-ft represents the intersection influence area. Second, the crash and road inventory data of Ann Arbor, which obtained from MPO of Ann Arbor, is identified based on 150-ft threshold.

Table 3.1 Functions of volatility

Measures of volatility	***Formulation***
Standard Deviation	$S_{dev} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Coefficient of Variation	$C_v = \frac{SD}{\bar{x}} * 100$
Mean Absolute Deviation	$D_{mean} = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$
Quartile Coefficient of Variation	$Q_{cv} = \frac{Q_3 - Q_1}{Q_3 + Q_1} * 100$
Percent of extreme values	$\%T = \frac{c > Threshold}{n} * 100$ $Threshold = \bar{x} \pm z * s$

Finally, 30 measures of volatility at the aggregate level of intersections were defined among which six measures capture speed volatility, sixteen measures quantify longitudinal and lateral acceleration volatility, and 8 measures capture yaw-rate driving volatility.

Measures of Goodness of Fit

In order to evaluate and compare the performance of traditional Poisson regression, RP

Poisson, and GWPR, four statistics were utilized to measure estimation accuracy.

1- *R-squared for Poisson model*: this statistic assesses the overall goodness of fit of model based on standardized residuals. Larger values of $R^2_{Poisson}$ (max is 1) indicate better fit. It is defined as (Cameron and Windmeijer 1996):

$$R^2_{Poisson} = 1 - \frac{\sum_{i=1}^{n} \frac{\left(Y_i - \widehat{Y}_i\right)^2}{\widehat{Y}_i}}{\sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{\bar{Y}}} \quad (3.17)$$

where Y_i and $\widehat{Y}_i$ are the observed and predicted number of crashes at location *i* respectively, and $\bar{y}$ is the average number of crashes.

2- *AIC*: a lower AIC represents a better goodness of fit (Bozdogan 1987). A three point decrease in an AIC value indicates a significant improvement in the goodness of fit (Bozdogan 1987). We can write:

$$AIC = D + 2k \quad (3.18)$$

where D denotes the model deviance, and k is the number of parameters. In the S-GWPR, due to the non-parametric framework of the model, the number of parameters is meaningless. Therefore, an effective number of parameters should be calculated which can be written as (Nakaya *et al.* 2005):

$$K = trace(S) \quad (3.19)$$

where S is the hat matrix. For more details, please see (Nakaya *et al.* 2005).

3- Bozdogan's Consistent AIC (CAIC): While the AIC criteria often leads the model

to overfit the data, CAIC almost always select the correct model size (Bozdogan 1987). We can write:

$$CAIC = D + ln(N) * p \tag{3.20}$$

where *N* is the sample size, and *p* is the number of parameters.

4- *Mean Absolute Deviation:* a smaller value of MAD implies a better model estimation. It can be defined as:

$$MAD = \frac{\sum_{i=1}^{n} \left| \hat{Y}_i - Y_i \right|}{N} \tag{3.21}$$

5- *Mean Squared Error:* assess the estimation accuracy of the model by measuring the distance between the observations and the estimated model. We can write:

$$MSE = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - Y_i \right)^2}{N} \tag{3.22}$$

where Y_i and $\hat{Y}_i$ are the actual number and estimated number of crashes, and N is the number of intersections. The MAD measure provides the average of misprediction in the method, while the MSE measure is used to assess the error associated in the estimation.

Data

In this study, three data sources were integrated: (1) Basic Safety Messages (BSM) data exchanged by connected vehicles obtained from the SPMD, (2) road inventory data and

(3) historical crash data. Figure 3.1 (right) shows the data process steps. BSM data were obtained via the Research Data Exchange website (https://www.its.dot.gov/data/). The data provides high-frequency information regarding vehicle location, motion, and driving context factors. Data were collected on October 2012 and April 2013 (N~225 million observations) using standard protocols by UMTRI at university of Michigan. In this paper, full two-months of publicly available CV data is processed. Due to the error made by developers during data transfer process from DSRC devices to comma separated value (CSV) files, 45.4% of lateral acceleration data are stored as either -9.81, 9.81, and 19.62 m/s^2 which are equivalent to "*-g"*, "*g"* and "*2g"* in the dataset (the histogram of the lateral acceleration is shown in Figure 3.2). However, these values belong to 1,048 vehicles out of the 2,544 vehicles that passed the selected intersections. Therefore, we did not include the erroneous data (shown in Figure 3.2 via red eclipses) and the final dataset contains the information of 1,496 vehicles passing the intersections.

In order to evaluate the correlation of driving volatility and crashes, we should account for the effect of traffic and geometric characteristics of intersections. Therefore, significant effort was undertaken in order to obtain road inventory data including AADT for major and minor approaches, speed limit, number of lanes in each direction, etc. Data were collected from Google Maps and the Metropolitan Planning Organization Website (http://semcog.org/). Among the intersections in Ann Arbor, 167 intersections were selected (Figure 3.3), considering AADT information availability and the availability of BSM data that can calculate 30 measures of driving volatility. To extract the BSM data at intersections, 150-ft. threshold from the center of intersections was established, and by processing 230 million BSMs, CV data for each intersection is extracted and linked to

selected 167 intersections. Next, by applying volatility functions to extracted BSMs at aggregate intersection level, speed, longitudinal acceleration, lateral acceleration, and yaw-rate volatility measures are calculated for each intersection.

The historical crash data were obtained from the Metropolitan Planning Organization Website. One of the main challenges in this study is describing 2-month connected vehicle data with historical crash data. In this paper, we are assuming that drivers of CVs that passing the selected intersections are representative of the majority of drivers. The ideal approach is comparing the speed distribution of connected vehicles with the distribution of speed obtained from non-at-fault drivers at study area by conducting quasi-induced method and evaluate whether the difference is acceptable (Lyles *et al.* 1991, Stamatiadis and Deacon 1997, Chandraratna and Stamatiadis 2009). However, the speed of vehicles prior to crash involvement is not available in the crash data. To mitigate this issue, we filtered the historical crash data from October 2012 to 2013 (1-year period) to obtain accurate inference regarding the correlation of intersection volatility and frequency of crashes. It is worth noting that 2-month CV data lies between the selected period.

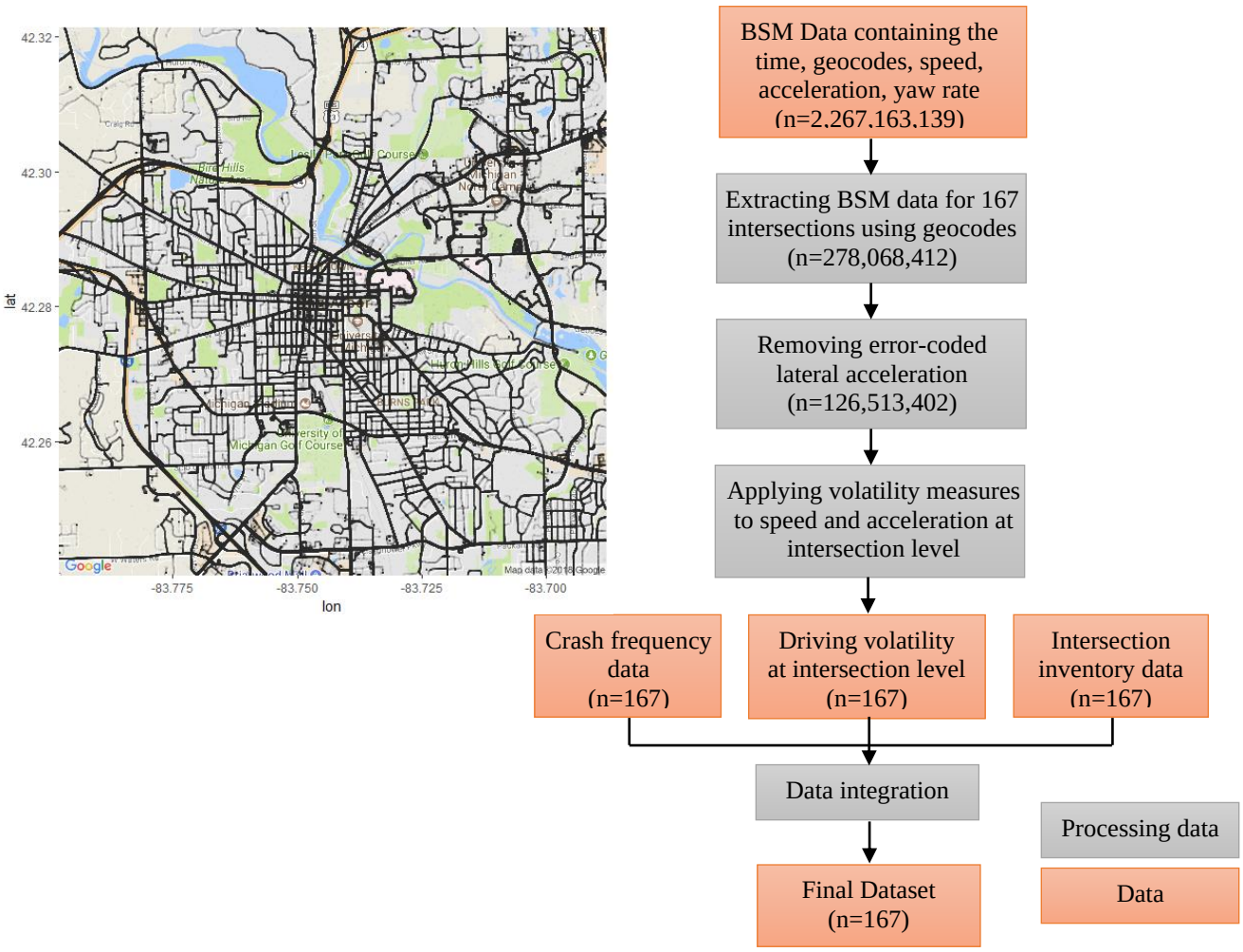


Figure 3.1 Created map from BSM data (left), Data preparation framework (right)

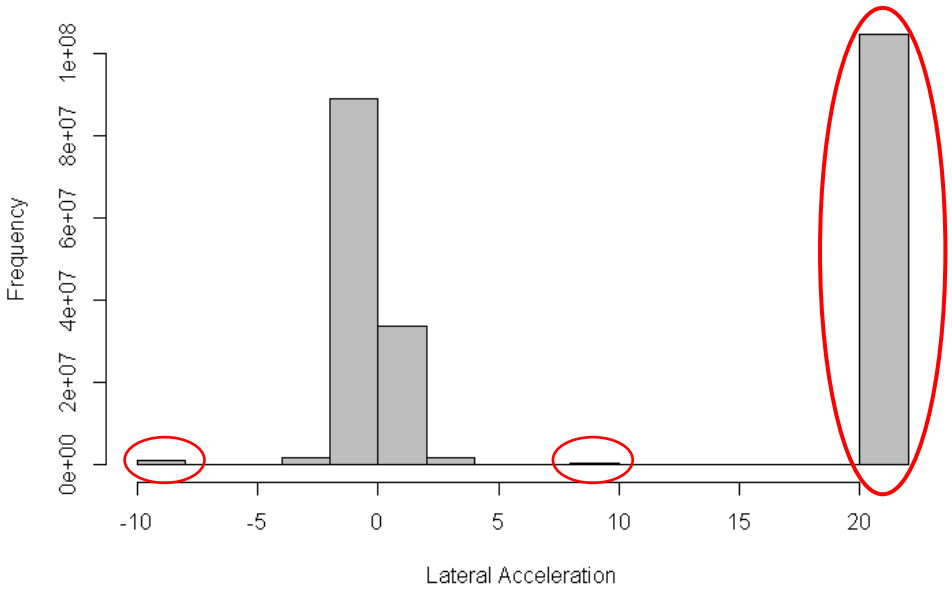


Figure 3.2 Histogram of lateral acceleration

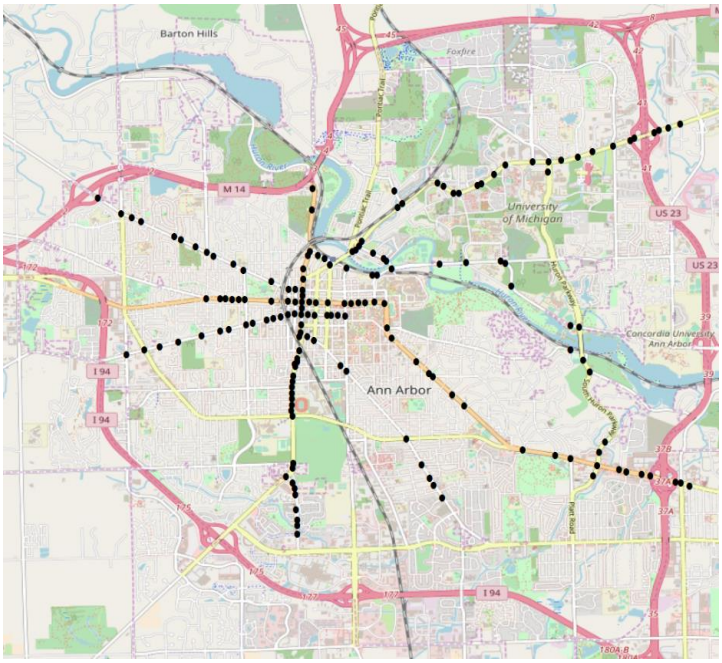


Figure 3.3 Location of selected intersections (N=167)

Finally, to ensure the accuracy of the manually collected data, 20% of the data was randomly checked and verified. In addition, the plot of the data in Figure 3.1 (left) indicates the high precision of the BSM data.

Results

Descriptive Statistics

In this section, the descriptive statistics of dependent variables, calculated intersection-based driving volatilities, and intersection related variables are shown in Table 3.2. As discussed before, the two-month CV data is used to calculate the volatility measures that attempt to capture the variations of speed, longitudinal/lateral acceleration, and yaw-rate. In order to help conceptualize the distribution of variables, the mean, standard deviation, minimum and maximum of the variables are provided for 167 intersections.

Table 3.2 Descriptive Statistics of dependents and key variables (N=167)

Variable	Mean	S.D.	Min	Max
Dependent variables				
Rear end crashes	3.51	4.79	0	28
Sideswipe crashes	1.54	2.22	0	13
Angle crashes	1.38	1.77	0	9
Head-on crashes	0.61	1.23	0	6
Intersection related variables				
AADT major road (1000)	18.47	8.60	2.53	45.40
AADT minor road (1000)	8.85	3.87	1.10	27.40
Speed limit of major road (mph)	34.52	6.52	25	45
Speed limit of minor road (mph)	29.28	4.37	15	45
Signalized intersection (yes = 1)	0.49	0.50	0	1
4-legged intersection (yes = 1)	0.47	0.50	0	1
Total through lanes	4.25	1.38	2	8
Total left turn lanes	1.38	1.37	0	6
Total right turn lanes	0.84	0.80	0	4
Intersection-based volatility measures				
Speed Volatility measures				
$Speed - S_{dev}$ (m/s)	10.88	2.57	4.83	16.78
$Speed - C_v$ (%)	44.48	16.00	12.34	80.81
$Speed - Q_{cv}$ (%)	31.67	16.74	6.17	66.74
$Speed - D_{mean}$ (m/s)	7.56	2.07	3.18	12.33
$Speed - 1S_{dev}$ (%)	28.74	13.30	11.35	60.28
$Speed - 2S_{dev}$ (%)	3.63	2.90	0.00	11.31
Longitudinal acceleration volatility measures				
$AccDec_x - S_{dev}$ (m/s^2)	0.76	0.18	0.33	1.42
$Acceleration_x - C_v$ (%)	58.49	5.53	42.53	74.11
$Deceleration_x - C_v$ (%)	65.44	8.54	52.30	120.13
$Acceleration_x - Q_{cv}$ (%)	38.47	5.64	21.84	50.00
$Deceleration_x - Q_{cv}$ (%)	43.20	7.42	22.58	59.62
$AccDec_x - D_{mean}$ (m/s^2)	0.39	0.09	0.15	0.54
$AccDec_x - 1S_{dev}$ (%)	23.44	4.71	6.50	35.87
$AccDec_x - 2S_{dev}$ (%)	6.45	1.83	1.61	11.09
Lateral acceleration volatility measures				
$AccDec_y - S_{dev}$(m/s^2)	1.05	0.36	0.12	2.14
$Acceleration_y - C_v$ (%)	87.34	38.25	28.64	225.62
$Deceleration_y - C_v$ (%)	128.25	34.18	57.45	221.63
$Acceleration_y - Q_{cv}$ (%)	45.80	14.02	10.00	93.01
$Deceleration_y - Q_{cv}$ (%)	57.41	20.39	15.09	92.79
$AccDec_y - D_{mean}$ (m/s^2)	0.77	0.62	0.07	4.65
$AccDec_y - 1S_{dev}$ (%)	5.80	6.10	0.0	39.57
$AccDec_y - 2S_{dev}$ (%)	1.92	2.17	0.0	13.16

Table 3.2 Continued

Yaw rate volatility measures				
$YawRate - S_{dev}$ (degree/s)	3.64	1.45	0.418	8.88
$YawRate - C_v$ (%)	1.64	0.54	0.22	3.13
$YawRate - C_v$ (%)	1.64	0.51	0.33	3.21
$YawRate - Q_{cv}$ (%)	0.57	0.20	0.08	0.92
$YawRate - Q_{cv}$ (%)	0.55	0.20	0.10	0.92
$YawRate - D_{mean}$ (degree/s)	2.99	1.70	0.18	6.93
$YawRate - 1S_{dev}$ (%)	0.10	0.08	0.00	0.40
$YawRate - 2S_{dev}$ (%)	0.04	0.03	0.00	0.13

** S_{dev}: standard deviation; ($1S_{dev}$): % of extreme points beyond mean ± one standard deviation; ($2S_{dev}$): % of extreme points beyond mean ± two standard deviation; C_v: coefficient of variation; Q_{cv}: quartile coefficient of variation; D_{mean}: mean absolute deviation; $Acceleration_x$: longitudinal acceleration; $Deceleration_x$: longitudinal deceleration; $AccDec_x$:both longitudinal acceleration and deceleration; $Accleration_y$: lateral acceleration; $Deceleration_y$: lateral deceleration; $AccDec_y$: both lateral acceleration and deceleration;*

Modeling Results

According to the aforementioned methods the Poisson regression, RP Poisson regression, and GWPR models were developed to explain the observed variations in frequency of rear-end, sideswipe, angle and head-on collisions given road inventory and intersection-based driving volatility factors. Although the aim of this study is not to compare different methodological approaches for modeling crash counts, we provide a model comparison in the following sections to illustrate more insights regarding their performance. In the following, the models performance on estimating the frequency of various crash types is compared. Next, the developed models are presented and discussed.

Model comparison

In order to estimate the fixed parameter Poisson regression models, the intersection related factors and driving volatility measures were incorporated.

In order to estimate the RP Poisson regression, 200 Halton draws was applied considering multiple functional form of the coefficients such as normal, lognormal, triangular, and uniform. Similar to previous studies (Anastasopoulos and Mannering 2009, El-Basyouny and Sayed 2009, Xu and Huang 2015, Kamrani *et al.* 2018b), for the random-held parameters the normal distribution had the best fit to the data, in all the crash type models.

To estimate the S-GWPR, the study considered bi-square and Gaussian fixed and adaptive kernels. In all crash type models, the adaptive bi-square kernel showed the best fit to the data based on their AIC score. In addition, all variables are significantly varied across the space for rear-end crashes, leading to the basic GWPR model. On the other hand, the S-GWPR model performed better for sideswipe, angle and head-on crashes by reducing the model complexity.

As discussed in the methodology section, to compare performance of the models, the $R^2_{poisson}$, AIC, MAD and MSE statistics are quantified. Table 3.3 shows the results for rear-end, sideswipe, angle, and head-on crashes. Based on the results, the RP Poisson regression outperformed the fixed parameter and GWPR models in all types of crashes. It should be noted that, both the RP and the S-GWPR models improved the fit for the fixed parameter models.

Table 3.3 Measures of goodness of fit for the fitted model

	Goodness of fit	Fixed Parameter	Random Parameter	S-GWPR
Rear-End	$R^2_{poisson}$	0.674	0.908	0.754
	AIC	376.99	215.96	338.6
	CAIC	401.934	250.258	335.144
	MAD	2.014	0.885	1.723
	MSE	8.477	1.322	6.001
Sideswipe	$R^2_{poisson}$	0.473	0.791	0.582
	AIC	279.67	190.05	258.05
	CAIC	304.614	221.230	269.704
	MAD	1.144	0.685	1.039
	MSE	2.991	0.843	2.346
Angle	$R^2_{poisson}$	0.391	0.675	0.457
	AIC	247.59	198.37	237.61
	CAIC	275.652	229.550	254.512
	MAD	0.994	0.739	0.938
	MSE	1.851	0.949	1.644
Head-on	$R^2_{poisson}$	0.446	0.584	0.509
	AIC	169.46	152.51	165.68
	CAIC	197.522	185.690	157.922
	MAD	0.539	0.459	0.508
	MSE	0.798	0.538	0.711

Model estimation

In order to estimate the fixed parameter models, intersection related variables were used, and the significant ones were kept in the model, then measures of driving volatility were added into the model. For model selection, the AIC, log-likelihood values, and variable significance were used. As discussed in the methodology section, the Lagrange Multiplier test was conducted to test for the over-dispersion existence (Greene 2003). Based on the results, the LM values for rear-end, sideswipe, angle and head-on crashes were lower than the critical Chi-square value for the 95 percent confidence interval, which is 3.84. Therefore, for all the crash type models the null hypothesis failed to reject, and it is appropriate to use the Poisson regression models (Washington *et al.* 2010).

After developing the fixed parameter model, significant variables in the models were used to develop RP Poisson and GWPR models. The estimated parameters for the RP Poisson and S-GWPR are presented by the minimum, lower quartile, median, upper quartile, and maximum estimated coefficients. In order to check for the multicollinearity, a common rule of thumb suggests that if the variance of Inflation (VIF) is higher than 5, multicollinearity might be an issue. VIF values for included variables were checked and all of them were below 5. The following sections discuss modeling results for rear-end, sideswipe, angle and head-on collisions.

Rear-end crashes

The modeling result for frequency of rear-end crashes in the selected time period is shown in Table 3.4. As discussed before, it is evident that the RP Poisson model outperformed the fixed Poisson and GWPR. The models suggest that three measures of driving volatility are highly correlated with the number of rear-end crashes at intersections: Coefficients of variation in speed (*Speed-*C_v), number of speed points lying beyond two standard deviations (*Speed-2*S_{dev}), and coefficient of variation volatility of positive longitudinal acceleration ($Acceleration_x - Q_{cv}$). The fixed parameter Poisson model states that the associations of driving volatility on rear-end crashes are fixed across the intersections. However, based on the RP Poisson model results, the effects of coefficients of some volatilities significantly vary across intersections with normal distribution. The number of speed points lying beyond two standard deviations (*Speed*-2 S_{dev}), are positively associated with number of rear-end crashes. They indicate that intersections with higher speed volatility are prone to have a higher number of rear-end crashes. Referring to partial

effects, it can be observed that a one percent increase in Speed-C_v and Speed-2S_{dev} increase the average number of rear-end crashes for 0.17 and 0.03, respectively. In addition, quartile coefficient of variation volatility of positive longitudinal acceleration ($Acceleration_x - Q_{cv}$) is significant in the model with positive sign reveals that increase in the variation of longitudinal control of the vehicle in terms of acceleration, increases the expected number of rear-end crashes. Controlling for other variables, a one percent increase in $Acceleration_x - Q_{cv}$ increases the rear-end number of crashes, on average, for 0.1. Considering the high variations in these volatilities, they have a substantial impact on the number of crashes. It is worth mentioning that all of the lateral volatilities are tested in the model but none of them was significant. From the model, it can be inferred that intersections with higher longitudinal volatility expected to have a higher number of rear end crashes. Based on intuition, we expect that failure in longitudinal control of the vehicle lead to rear-end crashes which is consistent with the results.

Other factors used in the model as control variables are significant and show the expected sign. According to the Table 3.4, a one thousand increase in AADT in major and minor streets contributes to a 0.1 and 0.11 increase in the number of rear-end crashes, respectively. Based on the results, on average, signalized intersections have 0.88 more rear-end crashes than un-signalized intersections. Four-legged intersections have 0.91 more crashes than T-intersections.

Referring to the GWPR model, as shown in Table 3.4, non-stationary test and the results show that there is a non-stationary spatial pattern and significant variation in all of the estimated coefficients across space. According to the results, volatility measures are

positively correlated with the number of rear-end crashes in almost all locations. It should be noted that presence of over-dispersion in data could lead to negative coefficient signs at some intersections (Xu and Huang 2015). In addition, volatilities with unexpected signs might be insignificant in the model. Focusing on volatility measures, an estimated coefficient for *Speed-* C_v varies from -0.012 to 0.029. Based on the results, 17 intersections have negative values among which none of them are significant at a 95% confidence level. The Estimated coefficients for *Speed-*$2S_{dev}$ vary from -0.008 to 0.156, and 11 intersections (6.5%) have negative signs. However, none of them was significant in the model. Along with volatilities, as shown in Table 3.4, intersection related variables vary across space significantly. Although the coefficients vary from negative values to positive, none of the negative estimates is significant in the model. By applying IDW interpolation, the coefficients are mapped in the space and the results of GWPR model for the local estimation of volatility measures are shown in Figure 3.4.

Table 3.4 Modeling results for rear-end crashes (N=167 intersections)

Variable	Poisson		Random Parameter							GWPR						
	β.[1]	ME	Mean	ME	Min	1st Q	Med	3rd Q	Max	Mean	Min	1st Q	Med	3rd Q	Max	Test[2]
Constant	-4.135***	-	-4.837***	-						-3.518	-5.126	-3.969	-3.453	-2.864	2.369	Yes
AADT MAJOR (1000)	0.054***	0.19	0.053***	0.1	0.052	0.059	0.062	0.066	0.079	0.063	0.021	0.059	0.071	0.073	0.075	Yes
Std. AADT Major			0.016***	-												
AADT MINOR (1000)	0.057***	0.2	0.06***	0.11						0.049	0.036	0.041	0.046	0.051	0.097	Yes
SIGNALIZED (yes=1)	0.464***	1.46	0.468***	0.88						0.529	0.02	0.217	0.299	0.495	1.16	Yes
4 legged intersection	0.630***	2.15	0.494***	0.91						0.401	0.143	0.43	0.538	0.596	1.087	Yes
*Speed-2*S_{dev}	0.066***	0.23	0.093***	0.17						0.066	-0.008	0.054	0.059	0.079	0.156	Yes
Speed-C_v	0.009***	0.03	0.015***	0.03	0.0184	0.0186	0.0187	0.0188	0.0201	0.015	-0.012	0.007	0.020	0.022	0.029	Yes
Std. Speed-C_v			0.002***	-												
$Acceleration_x - Q_{cv}$	0.058***	0.2	0.063***	0.1	0.014	0.035	0.046	0.053	0.093	0.036	0.003	0.018	0.029	0.059	0.078	Yes
Std. $Acceleration_x - Q_{cv}$			0.011***	-												
Null Deviance	878.86		878.86							878.86						
Model Deviance	360.99		193.96							294.2						
Explained Deviance	0.589		0.779							0.665						
AIC	376.99		215.96							338.6						

[1] Significance at *** 1%, ** 5%, and * 10%
[2] Non-stationary test

Table 3.5 Modeling results for sideswipe crashes (N=167 intersections)

Variable	Poisson		Random Parameter							GWPR						
	β.[1]	ME	Mean	ME	Min	1st Q	Med	3rd Q	Max	Mean	Min	1st Q	Med	3rd Q	Max	Test [2]
Constant	-6.869***	-	-6.737***	-						-4.971	-8.014	-6.459	-4.176	-3.945	-3.758	Yes
AADT MAJOR (1000)	0.002	0.00	0.004	0.00						0.001						
AADT MINOR (1000)	0.023*	0.03	0.011	0.01						0.019	-0.054	-0.007	0.034	0.043	0.049	Yes
SIGNALIZED (yes=1)	2.337***	3.26	2.317***	1.98						2.067						
Speed $-2S_{dev}$	0.104***	0.16	0.103***	0.1						0.079	0.026	0.055	0.068	0.108	0.146	Yes
$Acceleration_x - C_v$	0.062***	0.09	0.061***	0.06	0.063	0.106	0.109	0.112	0.171	0.05	0.029	0.033	0.037	0.074	0.1	Yes
Std. $Acceleration_x - C_v$			0.01***	-												
$Deceleration_y - C_v$	0.009***	0.01	0.008***	0.01						0.004						
$AccDec_y - 2S_{dev}$	0.114***	0.17	0.11***	0.10	0.051	0.057	0.059	0.063	0.085	0.071	-0.006	0.012	0.022	0.09	0.348	Yes
Std. $AccDec_y - 2S_{dev}$			0.07***	-												
Null Deviance	461.86		461.86							461.86						
Model Deviance	263.67		170.05							228.76						
Explained Deviance	0.429		0.632							0.505						
AIC	279.67		190.05							258.05						

[1] Significance at *** 1%, ** 5%, and * 10%
[2] Non-stationary test

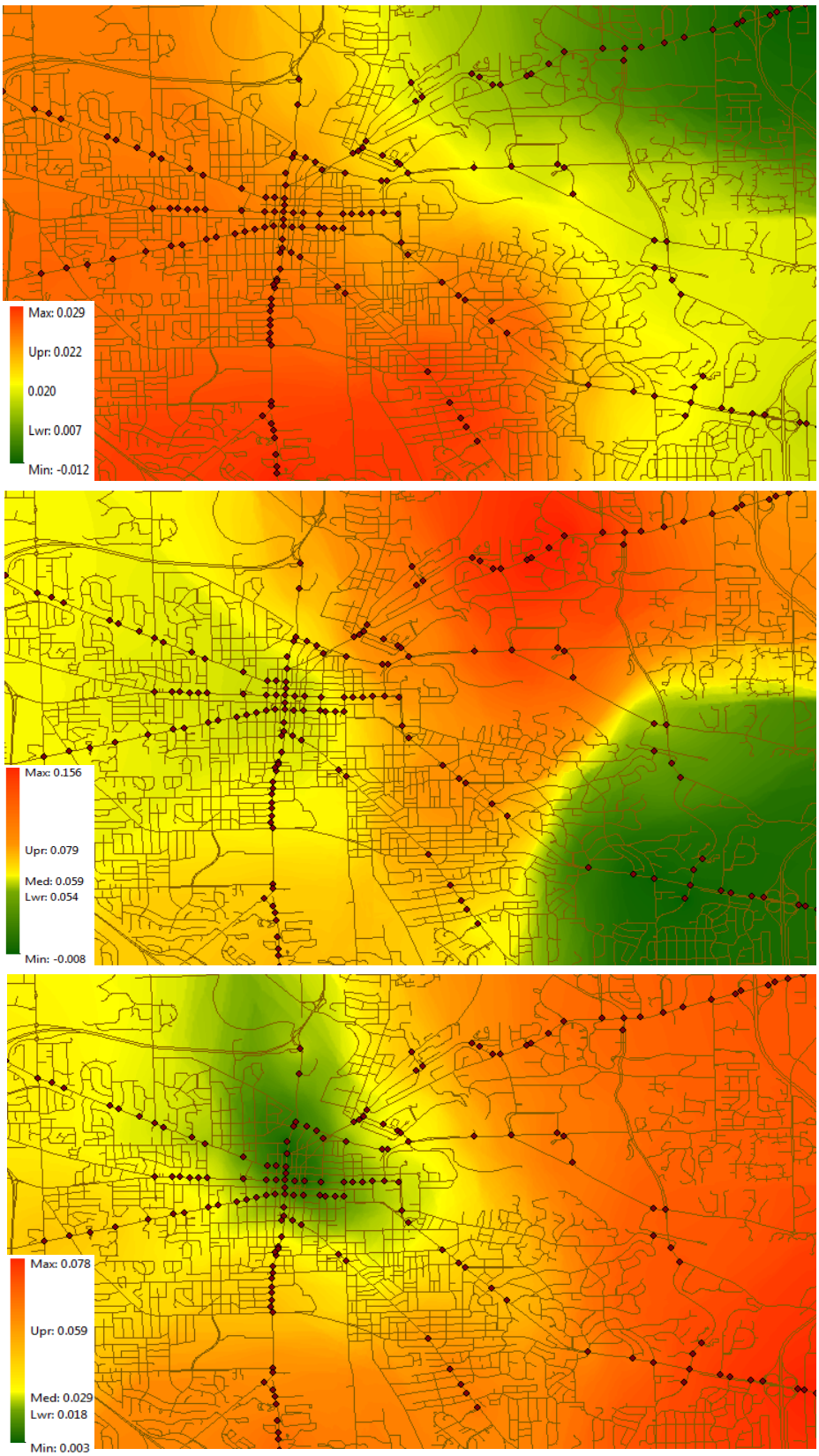


Figure 3.4 Local estimation of *Speed* – C_v (top), *Speed-2* S_{dev} (middle), and *Acceleration*$_x$ - Q_{cv} (bottom) on rear-end crashes

Sideswipe crashes

In this section, the association of intersection-based volatilities on the frequency of sideswipe crashes is discussed. Table 3.5 summarizes the modeling results for fixed parameter, random parameter, and S-GWPR models on such crashes. In terms of goodness of fit, by capturing unobserved heterogeneity with RP Poisson and S-GWPR models, the model fits improved significantly. All the models suggest that intersection volatilities in terms of speed, longitudinal and lateral acceleration volatilities are highly associated with frequency of sideswipe crashes. That said, four intersection-based volatility measures are highly contributing to crash frequency: number of speed points lying beyond two standard deviations (*Speed-2*S_{dev}), coefficient of variation volatility of positive longitudinal acceleration ($Acceleration_x - C_v$), coefficient of variation of negative lateral acceleration ($Deceleration_y - C_v$), and number of lateral acceleration points lying beyond two standard deviations ($AccDec_y - 2S_{dev}$). However, RP Poisson and S-GWPR suggest that the impacts of intersection-based volatility measures are not fixed across the intersections.

The marginal effect of the RP Poisson model reveals that a one percent increase in *Speed-2*S_{dev} is correlated with 0.16 increase in sideswipe crashes, on average. The model also indicates that the effect of $Acceleration_x - C_v$ is normally distributed across the intersections so that one percent increase in $Acceleration_x - C_v$, on average, is associated with a 0.06 increase in sideswipe crashes. Referring to lateral acceleration volatilities, a one percent increase in $Deceleration_y - C_v$ is correlated to 0.01 increase in the frequency of sideswipe crashes. In addition, the effect of $AccDec_y - 2\,S_{dev}$ on sideswipe crashes is normally distributed across the selected intersections contributing

to 0.1 increase in the frequency of sideswipe crashes, on average, by one percent increase in its magnitude. It is worth noting that in order to control for intersection-related variables, traffic exposure and type of the signal is used in the model, which is summarized in Table 3.5.

Coming to S-GWPR model, the results suggest that along with $Acceleration_x - C_v$ and $AccDec_y - 2\ S_{dev}$ volatilities, the impact of *Speed-2* S_{dev} is not fixed across the intersections. The distribution of estimated coefficients is shown in Table 3.5. One might notice that for some intersections, the estimation of $AccDec_y - 2S_{dev}$ is negative, while these observations are 5.9 percent of the intersections and they are not statistically significant. The local estimation plots of the volatility measures are shown in Figure 3.5. The estimated local coefficients suggest that intersection-based volatilities are an issue in eastern region of the city, comparing to the west side.

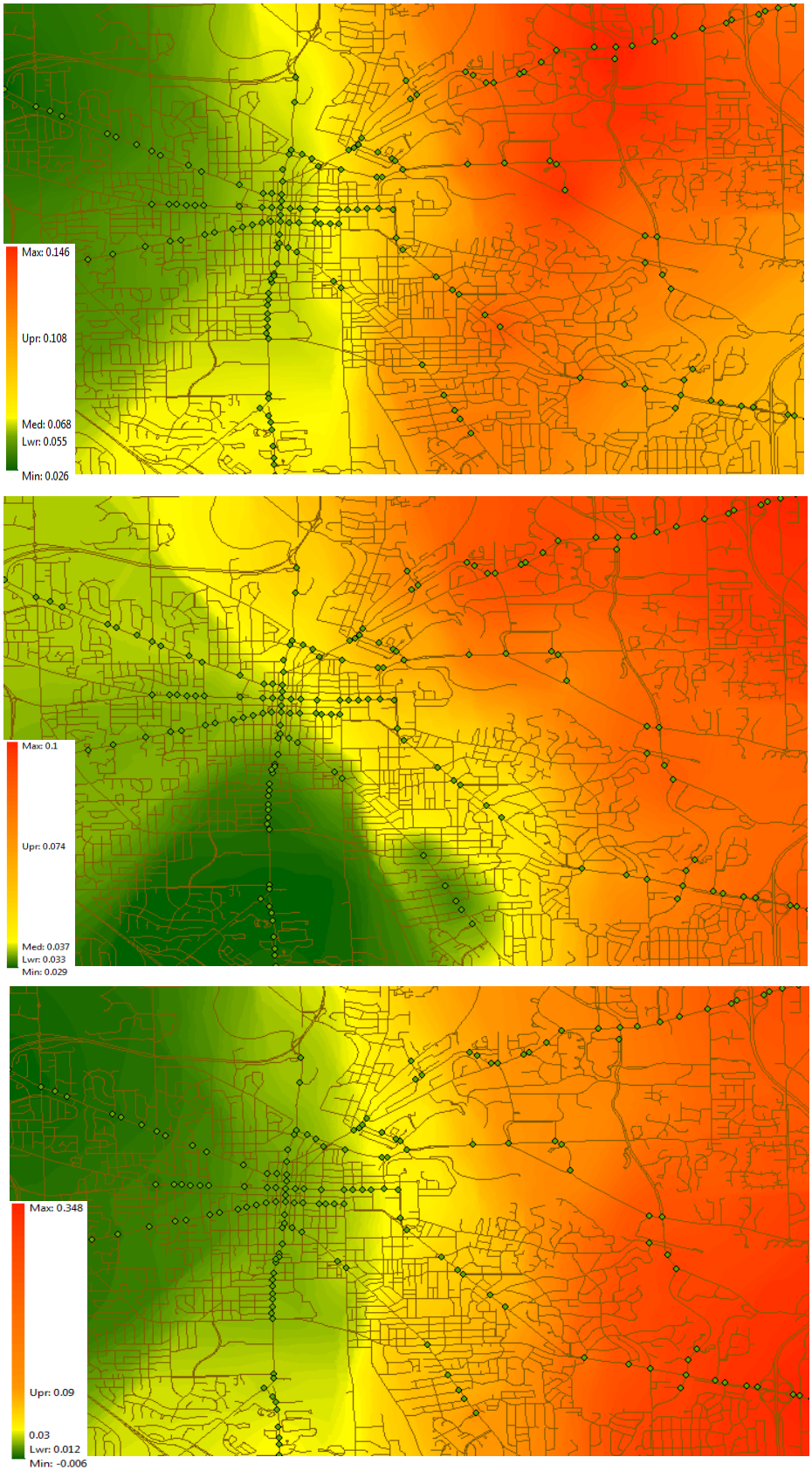


Figure 3.5 Local estimation of *Speed-2S_{dev} (top), Acceleration$_x$ – C_v (middle), and AccDec$_y$ – 2S$_{dev}$* (bottom) on sideswipe crashes

Angle Crashes

Modeling results for angle crashes are summarized in Table 3.6. Compared to the fixed parameter model, S-GWPR and RP Poisson models fit better, indicating that unobserved heterogeneity is captured. It can be observed that RP Poisson outperformed S-GWPR in terms of AIC value.

The developed models suggest that frequency of angle crashes is associated with four intersection-based volatility measures including speed, longitudinal, and lateral acceleration volatilities. In terms of speed volatility, quartile coefficient of variation in speed (*Speed-*Q_{cv}), is significantly correlated with angle crashes. On average, a one percent increase in *Speed-*Q_{cv} is associated with a 0.02 increase in angle crashes. Along with speed volatility, intersections with higher longitudinal volatilities experienced a higher number of angle crashes. One percent increase in the coefficient of variation of positive longitudinal accelerations ($Acceleration_x - C_v$) is associated with a 0.08 increase in number of angle crashes, on average. Intersections with higher lateral volatility are prone to have higher number of angle crashes. The coefficient of variation of negative lateral acceleration ($Deceleration_y - C_v$), and number of lateral acceleration points lying beyond two standard deviations ($AccDec_y - 2S_{dev}$) are statistically associated with angle crashes. In particular, the model suggests that coefficients of $Deceleration_y - C_v$ are normally disturbed across the intersections. On average, one percent increase in $Deceleration_y - C_v$ and $AccDec_y - 2S_{dev}$ is associated with a 0.01 and 0.11 increase in angle crashes respectively.

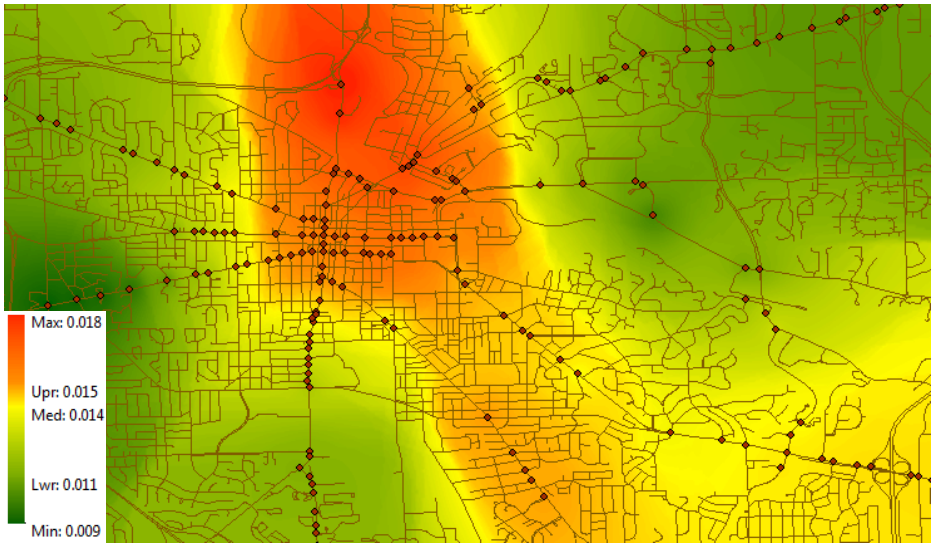


Figure 3.6 Local estimation of *Deceleration*$_y$ - C_v in angle crashes

The S-GWPR model shows a better fit than fixed parameter model and its results suggest that there is a spatial variation regarding the impact of lateral volatility across the intersections. Figure 3.6 depicts the heatmap of estimated coefficients for the lateral volatility ($Deceleration_y - C_v$). The estimated coefficients range from 0.009 to 0.018 and 85 percent of the estimated coefficients are statistically significant.

In terms of intersection-specific variables, the AADT of major road is contributing to frequency of angle crashes, while the AADT of minor approach is not statistically significant in the model. In addition, signalized and four-legged intersections have 0.55 and 0.67 higher angle crashes compared to unsignalized and three-legged intersections, on average.

Table 3.6 Modeling results for angle crashes (N=167 intersections)

Variable	Poisson Regression		Random Parameter							S-GWPR						
	β.[1]	ME	Mean	ME	Min	1st Q	Med	3rd Q	Max	Mean	Min	1st Q	med	3rd Q	max	Test[2]
Constant	-8.794***	-	-8.988***	-						-8.891						
AADT MAJOR (1000)	0.024***	0.03	0.026***	0.02						0.036						
AADT MINOR (1000)	0.018	0.02	0.018	0.01						0.022						
SIGNALIZED (yes=1)	0.726***	0.90	0.651***	0.55						0.602						
4 legged intersection	0.791***	1.01	0.803***	0.67						0.783						
Speed-Q_{cv}	0.022***	0.03	0.024***	0.02						0.0177						
$Acceleration_x - C_v$	0.094***	0.13	0.096***	0.08						0.078						
$Deceleration_y - C_v$	0.007**	0.01	0.006**	0.01	0.003	0.005	0.006	0.007	0.012	0.013	0.009	0.011	0.014	0.015	0.018	Yes
Std. $Deceleration_y - C_v$			0.003***	-												
$AccDec_y - 2S_{dev}$	0.133***	0.18	0.129***	0.11						0.178						
Null Deviance	368.50		368.50							368.5						
Model Deviance	229.59		178.37							208.45						
Explained Deviance	0.377		0.516							0.434						
AIC	247.59		198.37							237.61						

[1] Significance at *** 1%, ** 5%, and * 10%
[2] Non-stationary test

Table 3.7 Modeling results for head-on crashes (N=167 intersections)

Variable	Poisson Regression		Random Parameter							S-GWPR						
	β.[1]	ME	Mean	ME	Min	1st Q	Med	3rd Q	Max	Mean	Min	1st Q	med	3rd Q	max	Test[2]
Constant	-13.772***	-	-13.100***	-						-13.182	-14.852	-13.961	-13.127	-12.653	-11.773	No
AADT MAJOR (1000)	-0.006	0.00	-0.005	0.00												
AADT MINOR (1000)	0.042*	0.03	0.047	0.01												
SIGNALIZED (yes=1)	0.892**	0.42	0.952**	0.22						1.079	0.162	0.622	1.227	1.461	1.820	Yes
4 legged intersection	0.694***	0.39	0.707***	0.17												
Speed-Q_{cv}	0.072***	0.05	0.065***	0.03	0.061	0.064	0.065	0.066	0.076							
Std. Speed-Q_{cv}			0.007***	-												
$Acceleration_x - C_v$	0.1***	0.06	0.105**	0.02												
$Acceleration_y - Q_{cv}$	0.023**	0.02	0.018*	0.01						0.024	0.002	0.015	0.02	0.033	0.056	Yes
$Deceleration_y - C_v$	0.009**	0.01	0.008*	0.005						0.008	0.001	0.005	0.008	0.012	0.016	Yes
Null Deviance	279.70		279.70							279.70						
Model Deviance	151.46		134.51							137.45						
Explained Deviance	0.458		0.519							0.508						
AIC	169.46		152.51							165.68						

[1] Significance at *** 1%, ** 5%, and * 10%
[2] Non-stationary test

Head-on crashes

Table 3.7 shows estimated fixed parameter Poisson, RP Poisson, and S-GWPR models for head-on crashes. Comparing goodness of fit, in terms of AIC value, the RP Poisson model outperformed the fixed and GWPR models. In the following, the estimated parameters in the RP Poisson model will be discussed.

Based on the results, four measures of driving volatility are significantly associated with the number of head-on crashes. The coefficient of variation for speed (*Speed-*C_v), which represents variations in vehicle speeds, is significant in the model, suggesting that intersections with higher speed volatility have higher numbers of head-on crashes. A one percent increase in *Speed-*C_v, on average, increases the number of head-on crashes by 0.03. The coefficient of variation for positive longitudinal acceleration ($Acceleration_x - C_v$), which represents variations in longitudinal control of the vehicle, is significant in the model with a positive sign. Based on the model, on average, a one percent increase in $Acceleration_x - C_v$ is associated with an increase in the number of head-on collisions for 0.02. Two volatility measures capturing the variation in lateral movement of the vehicle ($Acceleration_y - Q_{cv}$, and $Deceleration_y - C_v$) are significant with a positive sign. Controlling for other variables, a one percent increase in quartile coefficient of variation of positive lateral acceleration ($Acceleration_y - Q_{cv}$), and coefficient of variation of negative lateral acceleration ($Deceleration_y - C_v$) increases the number of head-on crashes by 0.01 and 0.005.

According to the results, not only variations in longitudinal control of the vehicle (in terms of speed and longitudinal acceleration) are positively significant but also intersections with

greater lateral volatility are prone to experience a higher number of head-on crashes. Deviation from the centerline of the road is a major cause of head-on collisions (Gårder 2006), which is more probable at intersections with greater variations in lateral acceleration. In addition, higher variations in the longitudinal control of a vehicle might lead to deviations from the lane in order to avoid a crash (e.g. rear-end), leading to head-on collisions with vehicles approaching from opposite direction.

Intersection related variables are used in the model as control variables. Based on the results, AADT in major approaches is not significant in the model. However, it was kept in the model as a control variable. Based on the results, a 1000 increase in AADT of minor approach increase the frequency of head-on crashes for 0.01. Controlling for other variables, signalized intersections have 0.22 more head-on crashes compared to un-signalized intersections. In terms of intersection geometry, four-legged intersections on average have 0.17 more head-on crashes than T-intersections.

Referring to S-GWPR model, by considering the spatial variation of the coefficients, the model improved the AIC and explained deviance compared to the fixed parameter Poisson model. As shown in Table 3.7, non-stationary test was conducted on all variables and those that failed to pass the test are considered a global variable in the model. In the final model, the signalized intersection and measures of lateral acceleration volatilities ($Acceleration_y - Q_{cv}$, and $Deceleration_y - C_v$) passed the non-stationary test and significantly vary across the space. By applying IDW interpolation, we mapped estimated lateral acceleration volatilities. Figure 3.7 displays the results. Focusing on measures of driving volatility, measures positively contribute to the number of head-on crashes in all

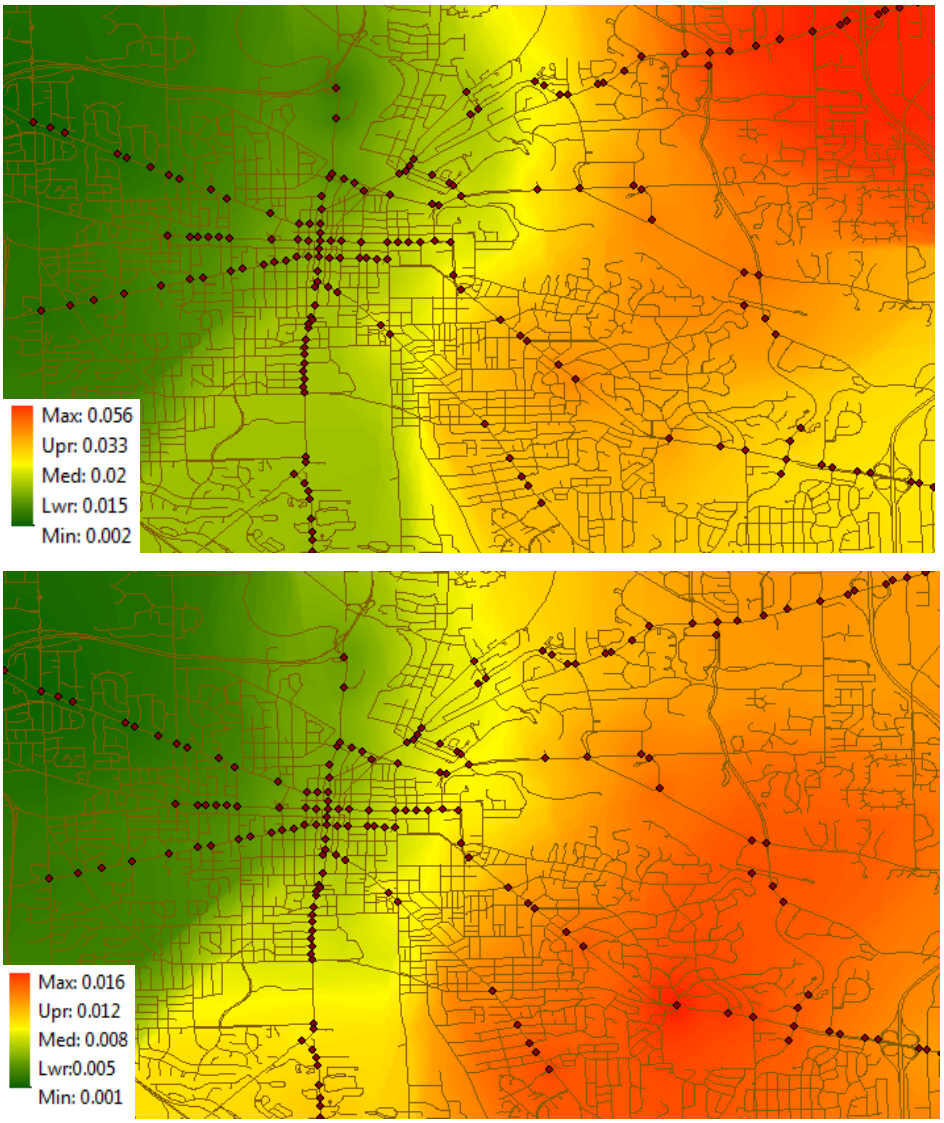


Figure 3.7 Local estimation of *Acceleration*$_y$ – Q_{cv} (top), and *Deceleration*$_y$ – C_v in head-on crashes

areas. Results revealed that in downtown areas, the estimated coefficient of driving volatility measures have a lower correlation with the number of head-on crashes. In the east side of the city, lateral acceleration volatilities have a greater contribution in crashes.

Limitations and future work

Because of the error in decoding the CV data from DSRC to csv, around 45 percent of

the lateral volatility of trips were voided. However, these errors came from specific device IDs, which were removed from the dataset during the cleaning process. In addition, during the intersection selection procedure, there might be a sample selection issue due to the unavailability of AADT and speed limit information for minor roads in the city. Furthermore, drivers in the study might not represent the population. Also, vehicles whose data was used to obtain driving volatilities might not be representative of the ones who were involved in crashes at intersections. Finally, although the data was error-checked, some errors might still remain from the data collection process.

This study investigates the association of longitudinal and lateral volatilities on the frequency of rear-end and head-on crashes at intersections. The future study would explore the impact of volatility on other crash types, such as sideswipe, angle, and single-vehicle crashes extending the model to multivariate random parameter and geographically weighted Poisson regression models. Furthermore, future studies should investigate contributing factors such as geometric design, traffic conditions, signal timing, etc., that might increase driving volatility at intersections. While in the literature, there are multiple surrogate safety measures such as time-to-collision (TTC), exposed time-to-collision, time integrated time-to-collision (TIT), and rear-end crash risk index (RCRI) aiming to quantify the crash risk (Essa and Sayed 2018, Rahman and Abdel-Aty 2018, Rahman *et al.* 2018), calculation of such measures need relative distance and kinematic information of front vehicle, which was not available in the data. In future, with higher penetration rate of CVs and availability of data, this information can be integrated in the model.

Conclusion

This study evaluated the impact of variations in longitudinal and lateral vehicular control on the frequencies of rear-end, sideswipe, angle and head-on crashes at intersections using the driving volatility which quantifies the degree of variations in instantaneous driving behavior. The goal of this study is to develop a fundamental framework to conceptualize and quantify variations in longitudinal and lateral control of vehicles (using speed, longitudinal/lateral acceleration, and yaw-rate volatilities), and explore the association of volatilities with type of crash.

To reach these goals, the Basic Safety Messages (BSMs) data exchanged by connected vehicles in real-world environments obtained from the Safety Pilot Model Deployment (SPMD) study conducted by the US Department of Transportation in Ann Arbor, MI is used. Such a big and precise dataset is available, which could be incorporated with historical crash data in order to understand the safety performance of the system prior to crash occurrences. This study creates a unique dataset by integrating BSM data, historical crash, and road inventory data. More than 2,225,000,000 BSMs obtained from two months of experiments in Michigan is processed and observations (n ~ 125,000,000) from 167 intersections are extracted. In order to capture the variations in vehicle control, 30 measures of driving volatility at the intersection level are calculated using speed, longitudinal/lateral acceleration and yaw-rate. Crash data from October 2012 to 2013 is linked with road inventory data including AADT of major and minor approaches, speed limits, and number of lanes, integrated with BSM data. Significant efforts were made to clean, process, and link the datasets.

From a methodological standpoint, rigorous modeling techniques including fixed parameter, random parameter (RP), and semi-parametric geographically weighted Poisson regression (S-GWPR) are developed to explore the impact of the measured volatilities on the frequency of several crash types. RP Poisson and S-GWPR allows us to consider the unobserved heterogeneity in the data coming from multiple unobserved factors. It is worth noting that RP Poisson model outperformed S-GWPR and Fixed Poisson models in all of the crash type models.

Referring to rear-end crashes, the RP Poisson model fitted better to the data compared to the fixed parameter and GWPR. Based on the random parameter and GWPR results, variations in longitudinal control of the vehicle in terms of longitudinal acceleration and speed are highly correlated with the number of rear-end crashes, and the estimated coefficients significantly vary across intersections. None of the lateral volatilities is significant in the model.

Focusing on sideswipe and angle crashes, modeling results suggest that along with longitudinal volatilities, in terms of longitudinal acceleration and speed, lateral volatility is highly associated with the frequency of frequency of such crashes. The results indicate that there is a substantial variation among the estimated coefficients for volatility indices at intersections level.

When it comes to head-on crashes, both longitudinal and lateral volatilities are positively associated with the number of crashes. Based on the results, variations in speed and

longitudinal and lateral acceleration are statistically significant and increase the frequency of head-on crashes. Deviation from the centerline of the road is the main reason of head-on crashes, and vehicles passing the intersections with higher lateral volatility are prone to deviate from their lane leading to head-on collision. In the S-GWPR model, contributions of lateral acceleration volatility vary across space. In downtown areas, it has a lower contribution while in the east side of the city the association is higher.

Given the calculated measures of volatilities, researchers can proactively identify hotspot intersections where crashes are waiting to happen. These hotspots are intersections where the frequency of crashes is low while the driving volatility is high (Kamrani *et al.* 2017). We can identify at-risk intersections where the driving behavior differs compared to other intersections by evaluating the driving volatility measures. In order to treat the intersection proactively, further examinations are needed to identify the contributing factors that increase the volatility of intersections such as inappropriate geometric designs, traffic conflicts, limited sight distances, inappropriate signal timing, etc. In addition, utilizing V2I communication, proactive warnings could be generated and transmitted by RSUs at these locations that inform drivers about potential hazards. This information could potentially enhance drivers' situational awareness, leading to a decrease in their driving volatility (Arvin *et al.* 2018).

CHAPTER 4 : EXAMINING THE ROLE OF PRE-CRASH DRIVING VOLATILITY IN CONTRIBUTING TO CRASH INTENSITY

A version of this chapter is presented at the 98th Transportation Research Board Annual Meeting and published in the Accident Analysis and Prevention journal.

Arvin, R., Kamrani, M., & Khattak, A. J. (2019). The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. *Accident Analysis & Prevention*.

Abstract

While the cost of crashes exceeds $1 Trillion a year in the U.S. alone, the availability of high-resolution naturalistic driving data provides an opportunity for researchers to conduct an in-depth analysis of crash contributing factors, and design appropriate interventions. Although police-reported crash data provides information on crashes, this study takes advantage of the SHRP2 Naturalistic Driving Study (NDS) which is a unique dataset that allows new insights due to detailed information on driver behavior in normal, pre-crash, and near-crash situations, in addition to trip and vehicle performance characteristics. This paper investigates the role of pre-crash driving instability, or driving volatility, in crash intensity (measured on a 4-point scale from a tire-strike to an injury crash) by analyzing microscopic vehicle kinematic data. NDS data are used to investigate not only the vehicle movements in space but also the instability of vehicles prior to the crash and their contribution to crash intensity using path analysis. A subset of the data containing 617 crash events with around 0.18 million temporal trajectories are analyzed. To quantify driving instability, microscopic variations or volatility in vehicular movements before a crash are analyzed. Specifically, nine measures of pre-crash driving volatility are

calculated and used to explain crash intensity. While most of the measures are significantly correlated with crash intensity, substantial positive correlations are observed for two measures representing speed and deceleration volatilities. Modeling results of the fixed and random parameter probit models revealed that volatility is one of the leading factors increasing the probability of a severe crash. Additionally, the speed prior to a crash is highly correlated with intensity outcomes, as expected. Interestingly, distracted and aggressive driving are highly correlated with driving volatility and have substantial indirect effects on crash intensity. With volatile driving serving as a leading indicator of crash intensity, given the crashes analyzed in this study, early warnings and alerts for the subject vehicle driver and proximate vehicles can be helpful when volatile behavior is observed.

Introduction

According to the National Highway Traffic Safety Administration, there were 7.27 million automobile crashes leading to more than 2.17 million injuries and 37,914 fatalities across the United States in 2016 (Anon 2018). It has shown that human-error was contributing in 94 percent of crashes across the U.S. (Anon 2008). These statistics suggest a great need and attention for research to explore the role of driving behavior on the severity outcome of crashes. Severities of accidents are the outcomes of complex interactions between multiple factors such as driver behavior, roadway and environmental factors, and vehicle defect. The main goal of injury severity model is to elucidate the association of severity outcome and these factors. Developing an understanding regarding the extent of contribution for each factor is the effective approach to improve the safety.

In the literature, a majority of methods are frequentist-based where a population is sampled (Savolainen *et al.* 2011). The data sources include police-reported crashes and road inventory. While the police reported data are the major source of crashes, certain types of crashes are under-reported in such databases. Specifically, a National Highway Traffic Safety Administration report (NHTSA 2009) has shown that 50% of no-injury crashes and 25% of minor injury crashes are unreported. This study focuses on developing a more complete picture of such unreported crashes by examining crash intensity using a unique database. Additionally, the crashes that may be truncated due to monetary thresholds imposed by states (Hauer 2006) are also captured in the analysis. The focus of this study is on crash intensity with the full range of mostly unreported crashes all the way to severe crashes.

The emergence of high-resolution naturalistic driving data provides a great opportunity for researchers to develop an in-depth analysis of crashes and investigate the crash contributing factors by analyzing microscopic driving performance and behavior prior to crash involvement. It helps us to investigate not only the movement of the vehicle in space but also the variations of movements prior to a crash and their contribution to severity. A new opportunity is offered when this information is coupled with the driver behavior and roadway/environmental characteristics at the crash time.

In this study, an in-depth analysis of crash intensity is performed by exploring driving instability and coupled that with driving behavior and roadway/environmental factors to investigate their association with crash intensity (which is the extent of harm to a person or property in a crash). In order to quantify instability in driving, the concept of driving

volatility is utilized, which captures the variations in instantaneous driving behavior. In the literature, the term “crash severity” is widely used to describe the severity level of a crash, while commonly it reflects the level of injury in the data. In this paper, the SHRP2 NDS data is used, which contains detailed information on extreme safety situations including minor crashes leading us to investigate an in-depth analysis of PDO crashes. Therefore, we used the term “crash intensity”. Crash intensity is characterized by several categories of property damage only crashes that include tire-strike, minor, police-reportable, and severe crashes. To summarize, the questions that this paper is trying to answer are:

- How can we extract useful information about enhancing safety from recently available microscopic vehicle kinematics data?
- How is crash intensity related to pre-crash driving volatility (or driving instability)?

By analyzing driving stability, driver behavior, and surrounding environment, proactive warnings could be transmitted regarding potential hazards via infrastructure-to-vehicle (I2V) and vehicle-to-vehicle (V2V) communication.

Methodology

In this study, by incorporating a unique naturalistic driving data obtained from SHRP2, 617 crash events containing around 0.18 million temporal trajectory observations are analyzed. The goal of this study is exploring the impact of driving instability (in terms of various aspects of driving volatility) on crash intensity, controlling for other factors. Here, we have investigated whether greater volatility prior to a crash occurrence increases the

probability of an intense crash. Multiple measures of driving volatility are calculated using vehicle kinematics (i.e., speed, acceleration and deceleration). This study differentiates between speed, acceleration, and deceleration volatilities expecting that speed and deceleration volatilities have a higher contribution on the crash intensity. Furthermore, the impact of driving behavior and roadway/environmental factors on driving volatility and its indirect correlation with crash intensity is explored using path analysis. In the following, the modeling approach and measures of volatility are discussed.

Modeling framework

The common approach in modeling the crash severity is linking the associated factors directly to the safety outcome. However, some of these associations might be more complicated and need further analysis to be investigated. In this study, the goal is to explain the associated factors of crash intensity through the influence of speed, driving behavior, and surrounding environment on the stability performance of the vehicle in terms of driving volatility. In other words, we are trying to investigate the impact of driving volatility on the intensity of crashes, while driving volatility is influenced by driving behavior and the surrounding environment. Path analysis is one of the well-known methods widely used by researchers to explain direct and indirect association of factors (Loehlin 2004, Şimşekoğlu *et al.* 2013, Yu *et al.* 2019). The conceptual framework of this study is shown in Figure 4.1. The associated factors include driver behavior, roadway/environmental factors, and vehicle-specific factors. These factors can be directly associated with the safety outcome, i.e., crash intensity. Furthermore, these factors can indirectly affect crash intensity through driving instability. Although the vehicle-specific factors can also potentially affect crash intensity, due to unavailability of such information in the available

subset of SHRP2 NDS data, relevant variables could not be included in the analysis. The structure of the path model can be written as:

$$Y_1 = F_{volatility}(\alpha_1 + \beta_1 X_1) \quad (4.1)$$

$$Y_2 = F_{intensity}(\alpha_2 + \beta_2 X_2 + \gamma Y_1 + \beta_3 V) \quad (4.2)$$

where $F_{volatility}$ is the driving volatility function, $F_{intensity}$ is the crash severity model, α_1 and α_2 are the model intercept, β_1 is the estimated coefficients, X_1 is the matrix of covariates including driver behavior and roadway/environmental factors, β_2 is the estimated coefficients for explanatory variables X_2, γ is the association of driving volatility

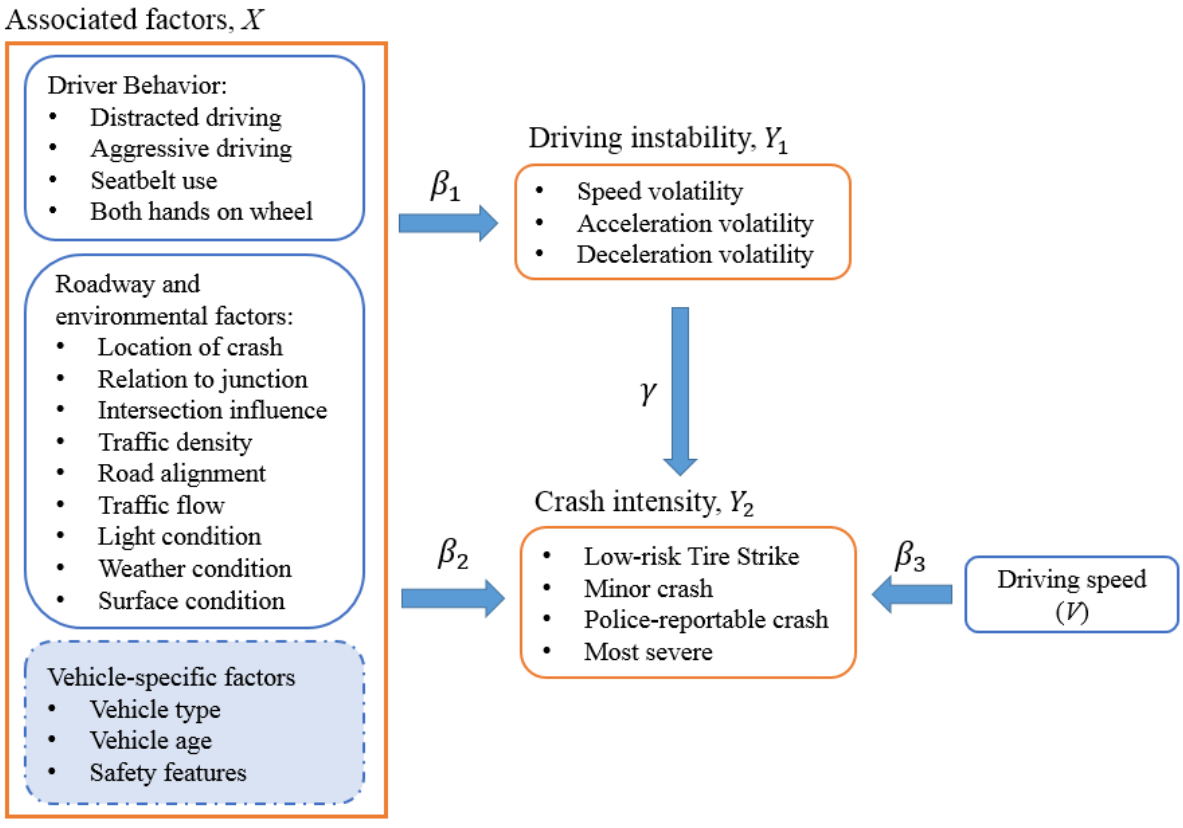


Figure 4.1 Conceptual framework for the pathways modeled

on crash intensity, V is the vehicle speed just before the crash, β_3 is the estimated coefficients for speed.

While driver behavior and roadway/environmental factors can affect the driving speed of vehicles, in which several studies have investigated these associations (Gargoum and El-Basyouny 2016, Huang *et al.* 2018, Sadia *et al.* 2018, Wang *et al.* 2019a), the focus of this study remains on the investigation of these factors on driving instability and crash intensity.

Fixed parameter modeling of pathways

Fixed parameter model estimates one set of coefficients, which are stationary across all the observations. Fixed parameter model helps us to understand the whole picture of the correlation between driving volatility and crash intensity. In this paper, two fixed parameter models are estimated:

1) Tobit models are used to investigate the correlations of driving instability with driving behavior and roadway/environmental factors. Notably, the volatility measures are generally left-censored at zero. Tobit model, originally proposed by Tobin (Tobin 1958), is an appropriate modeling framework to analyze such censored variables; such models are widely used in the literature to analyze crash rates (Anastasopoulos and Mannering 2009, Anastasopoulos *et al.* 2012, Zeng *et al.* 2017a, Zeng *et al.* 2017b). Given the left-censored limit of zero, we can write (Anastasopoulos *et al.* 2012):

$$Y_{1,i}^{*} = \alpha_1 + \beta_1 X_1 + \varepsilon_i \, , i = 1, 2, .., N$$

$$Y_{1,i} = Y_{1,i}^{*} \quad if \;\; Y_{1,i}^{*} > 0 \tag{4.3}$$

$$Y_{1,i} = 0 \quad if \;\; Y_{1,i}^{*} \leq 0$$

where Y_1 is the driving volatility, α_1 is the model intercept, β_1 is the estimated coefficients, X_1 is the matrix of covariates including driver behavior and roadway/environmental factors, N is the number of observations, and ε_i is the error term normally distributed with mean zero and variance σ^2. It is worth noting that latent variable, Y_1^* is observed when it is positive. The likelihood function for the Tobit model can be written as following (Anastasopoulos *et al.* 2012):

$$Likelihood = \prod_{0}[1 - \Phi\left(\frac{\beta X}{\sigma}\right)] \prod_{1} \sigma^{-1} \Phi[\frac{Y_i - \beta X}{\sigma}] \tag{4.4}$$

where Φ represents the normal density function.

2) Crash intensity model, which is an ordered probit model to explore the direct association of driving volatility, speed, driver behavior, and roadway/environmental factors to the crash intensity. It should be noted that due to the ordinal nature of crash intensity, multiple studies have suggested using ordered-response models (Huang *et al.* 2011, Savolainen *et al.* 2011, Huang *et al.* 2014, Nickkar *et al.* 2019a, Azimi *et al.* 2020, Rahimi *et al.* 2020). Crash intensity is considered as a four-level ordinal response variable ranging from low-risk tire strike crashes to most severe crashes (will be discussed in detail in section 4). Crash intensity function can be defined as

(Washington *et al.* 2010):

$$y^* = \alpha_2 + \beta_2 X_2 + \gamma Y_1 + \beta_3 V + \varepsilon \quad (4.5)$$

where y^* is the latent variable, β_2 is the estimated coefficients for explanatory variables X_2, γ is the association of driving volatility on crash intensity, *V* is the vehicle speed just before the crash , β_3 is the estimated coefficients for speed and, ε is error term. It should be noted that the explanatory variables in the first equation are indirectly associated with the crash intensity through γ. The latent variable (y^*) can be transformed to the observed ordinal response, *y*, as (Greene 2003):

$$y = 1 \quad if \;\; y^* \leq 0 \quad (4.6)$$
$$y = 2 \quad if \;\; 0 < y^* \leq \mu_1$$
$$y = 3 \quad if \;\; \mu_1 < y^* \leq \mu_2$$
$$y = 4 \quad if \;\; y^* > \mu_2$$

where μ_1 and μ_2 (also known as thresholds) needs to be estimated by the model.

Random parameter modeling of pathways

While traditional methods model crash severity under the assumption of fix effect of each parameter across all of the observations, in this study we applied random parameter model to address unobserved heterogeneity which arising from unobserved contributing factors in crashes. Although SHRP2 NDS data contains rich information on vehicle kinematics, driver pre-crash behavior, and driving environment, still there are some

factors that cannot easily be captured, such as drivers' risk perception, cautiousness, and situational awareness.

Random parameter model extends the fix parameter model by allowing coefficients to vary across observations.

$$\beta_i = \beta + \omega_i \tag{4.7}$$

where ω_i is distributed randomly. In the literature, several studies (Train 2000b, Bhat 2003) suggested Halton sequence method (Halton 1960) and simulated maximum likelihood approach to estimate model parameters. While several random parameter density functions, including normal, uniform, triangle, and log-normal are tested in this study, the normal distribution revealed the superior outcomes. It is worth noting that we considered random variables with significant mean and/or standard deviation.

Quantifying pathway by marginal effects

Crash intensity model explores the direct association between driving volatility, driving behavior, and environmental factors with intensity outcome of the crash. On the other hand, the driving volatility models uncover the correlation between driving instability and associated factors, which are indirectly associated with the crash intensity. In this context, path analysis is helpful in discovering direct and indirect relationship between the crash intensity and contributing factors. In this study, in order to quantify the direct and indirect association between factors and dependent variables (driving volatilities and crash intensity), marginal effect is used. Marginal effect represents the change in probability of

occurrence of dependent variable once the independent variable increase by one unit. The main advantage of marginal effect is providing intuitive interpretation regarding the associated factors and dependent variables (driving volatilities and crash intensity).

- ***Direct marginal effect***: One of the practical issues in the ordered probit models is the interpretation of dependent variables, where the association of positive/negative sign is not clear (Washington *et al.* 2010). Thus, the marginal effect analysis, which is a common approach in the literature, is used to uncover the association of independent variables with the model outcome. In order to calculate the direct marginal effect of a continuous factor, we can write (Jalayer *et al.* 2018, Zeng *et al.* 2019):

$$\frac{\partial P_{i,1}}{\partial x} = \phi(-(\beta_2^t x^t + \beta_2' X' + \gamma Y_1 + \beta_3 V))\beta_2 \qquad (4.8)$$

$$\frac{\partial P_{i,2}}{\partial x} = [\phi(-(\beta_2^t x^t + \beta_2' X' + \gamma Y_1 + \beta_3 V)) - \phi(\mu_1 - (\beta_2^t x^t + \beta_2' X' + \gamma Y_1 + \beta_3 V))]\beta_2$$

$$\frac{\partial P_{i,3}}{\partial x} = [\phi(\mu_1 - (\beta_2^t x^t + \beta_2' X' + \gamma Y_1 + \beta_3 V)) - \phi(\mu_2 - (\beta_2^t x^t + \beta_2' X' + \gamma Y_1 + \beta_3 V))]\beta_2$$

$$\frac{\partial P_{i,4}}{\partial x} = [\phi(\mu_2 - (\beta_2^t x^t + \beta_2' X' + \gamma Y_1 + \beta_3 V))]\beta_2$$

where $P_{i,j}$ is the probability of intensity level of j for observation i, $\phi(.)$ is the cumulative standard normal function, β_2^t is the estimated coefficient for the subjected factor, x^t is the subjected factor, $\beta_2^{'}$ is the estimated coefficients for other independent variables in the model, and $X^{'}$ are is the other associated factors, Y_1 is the driving volatility, and γ the estimated coefficient for the driving volatility.

- ***Indirect marginal effect***: In order to obtain the indirect association of the factors, estimated coefficients of the two models are integrated by calculating marginal effects on driving volatilities and crash intensity. The marginal effect on driving volatility can be written as:

$$Direct\ ME\ on\ driving\ volatility = f_{volatility}(\beta_1^t(x^t + 1) + \beta_1' X') - f_{volatility}(\beta_1^t x^t + \beta_1' X') \quad (4.9)$$

where $f_{volatility}$ is the estimated Tobit model for the driving volatility, β_1^t is the estimated coefficient for the subjected factor, x^t is the subjected factor, $\beta_1^{'}$ is the estimated coefficients for other independent variables in the model, and $X^{'}$ are is the other associated factors. The indirect marginal effect on severity through driving volatility can be written as:

$$Indirect\ ME\ on\ crash\ intensity = f_{intensity}(\beta_2^t X + \gamma(\beta_1^t x^t + 1 + \beta_1' X') + \beta 3V) - f_{intensity}(\beta_2^t X + \gamma\beta_1^t x^t + \beta 3V) \quad (4.10)$$

The total marginal effect on the crash severity is:

Total ME on crash intensity: Direct ME + Indirect ME

Measures of Volatility

In recent literature, the concept of driving volatility is defined and utilized to capture the

variations in instantaneous driving behavior. Multiple driving volatility measures have been used by researchers (Kamrani *et al.* 2017, Kamrani *et al.* 2018b, Arvin *et al.* 2019c), which can be applied to the vehicle kinematics, including speed, acceleration/deceleration, and jerk. In the following, the volatility functions used in this study and applied to vehicle speed, acceleration, and deceleration are discussed. Further details are available in (Kamrani *et al.* 2018b). Utilizing these functions, nine volatility measures are calculated using vehicle kinematic data prior to crash involvement. In the following, the applied functions on vehicle kinematics is discussed.

Standard deviation

The first function is standard deviation, which is desirable for capturing the data variations. We can write:

$$S_{dev} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.11}$$

where x_i is the observed value *i*, $\bar{x}$ is the mean of observations, and *n* is the total number of observations. This function is applied on speed and acceleration/deceleration.

Time-varying stochastic volatility

The time-varying stochastic volatility measure is widely used in the econometric field, which can be written as (Figlewski 1994):

$$V_f = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})} \qquad from\ t = 1\ to\ n \tag{4.12}$$

where

$$r_i = ln\left(\frac{x_t}{x_{t-1}}\right) \tag{4.13}$$

where x_t and x_{t-1} are the observations at time t and $t-1$, respectively, and *ln* is the natural logarithm. Since this measure needs time-series observations with positive values, only vehicle speed is used (acceleration/deceleration have negative values).

Coefficient of Variation

This measure obtained by dividing the standard deviation by the mean (Everitt and Skrondal 2002), which applied to speed, acceleration, and deceleration, and can be written as:

$$C_v = \frac{S_{dev}}{|\bar{x}|} \tag{4.14}$$

Quartile Coefficient of Variation

This measure is desirable when the data is not following a normal distribution (Zwillinger and Kokoska 2000), which can be defined as (Bonett 2006):

$$Q_{CV} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \tag{4.15}$$

where Q_1 and Q_3 are the 25th and 75th percentiles of data, respectively.

Data

In this study, the second Strategic Highway Research Program (SHRP 2) data are used. About 4 petabytes of data were collected under this program and it is known to be the most comprehensive naturalistic driving study. Referring to the effort that went in to the data collection, this is a high-quality data which contains information of more than 3500 drivers participating from six states in the United States including Washington, New York, Pennsylvania, Florida, North Carolina, and Indiana, during three years (2010 to 2013), travelling more than 50 million vehicles miles and 5 million trips (Hankey *et al.* 2016). In order to collect data, onboard data acquisition system (DAS) is installed on vehicles. Along with DAS, various sensors (camera, alcohol sensor, forward sensor, accelerometers) are used to record information including vehicle kinematics (speed, acceleration, steering position) at 10-Hz frequency, video views, vehicle controls, offset from lane center, etc. (Hankey *et al.* 2016).

The SHRP2 NDS data used in this study is a subset of data containing 617 crashes. For each crash-involved trip, 30 seconds of vehicle kinematics data is available. The data contains the seconds of evasive maneuver (taken by the driver to avoid the crash) and after crash occurrence. Since we are studying the impact of pre-crash behavior on crash intensity, we need to only include unintentional driving behavior and exclude intentional volatility arising from the drivers to avoid crashes in the analysis. Therefore, we need to exclude these seconds from our analysis, which will be discussed in the next section. These extracted seconds were used to calculate aforementioned measures of driving volatility. The final dataset is formed by linking the measured volatilities with the summary

of trip. The trip summary file contains the driver behavior and roadway/environmental characteristics which recorded via camera and recoded by data reductionist. In the dataset, the crash is defined as “any contact that the subject vehicle has with an object, either moving or fixed, at any speed in which kinetic energy is measurably transferred or dissipated”, and grouped into four categories:

1- *Level 1 Severe Crash*: These crashes include any injury, airbag deployment, vehicle rollover, or a high-delta V.
2- *Level 2 Crash Moderate Severity:* Not a level 1 crash. *These c*rashes have a minimum $1500 damage worth. Also, the crashes where acceleration reaches ±1.3 g are included.
3- *Level 3 Crash Minor Severity:* Not a level 1 or 2 crash. In these crashes the vehicle contacts other objects, or they are crashes where a vehicle departs from the road and sustains some (minimal) damage.
4- *Level 4 Crash Tire Strike:* Not a level 1, 2 or 3 crash. These crashes are the ones where the tire strikes an object, but there is little damage or risk element compared with the other categories.

It should be noted that the reported intensity levels in the original SHRP2 NDS dataset is in descending order ranging from level 1 with the highest severity to level 4 corresponding to lowest intensity crashes. Therefore, in this research, the order of the crash intensity is reversed into ascending order, which helps us to easily interpret the sign of the estimated coefficients (variables with positive sign increase the probability of an intense crash and vice versa).

Exclusion of Evasive Maneuvers

While the goal of this study is examining the role of driving stability on crash intensity, it is crucial to exclude the seconds of vehicle trajectories that drivers are attempting to avoid crashes, in order to isolate unintentional volatility (when drivers are performing normal driving) from intentional volatility (that drivers trying to avoid the crash) (Kamrani *et al.* 2019). To shed more light on this issue, Figure 4.2 illustrates the speed and acceleration profile of a randomly chosen crash in which the crash occurred in the 23rd second of the video and the driver started to react to the situation at 22nd second of the sequence. To exclude the irrelevant data, we have excluded the seconds that the driver is reacting. Therefore, this study only uses the seconds of the data up to the moment that the drivers started to react. The speed and acceleration of these seconds are used to calculate nine

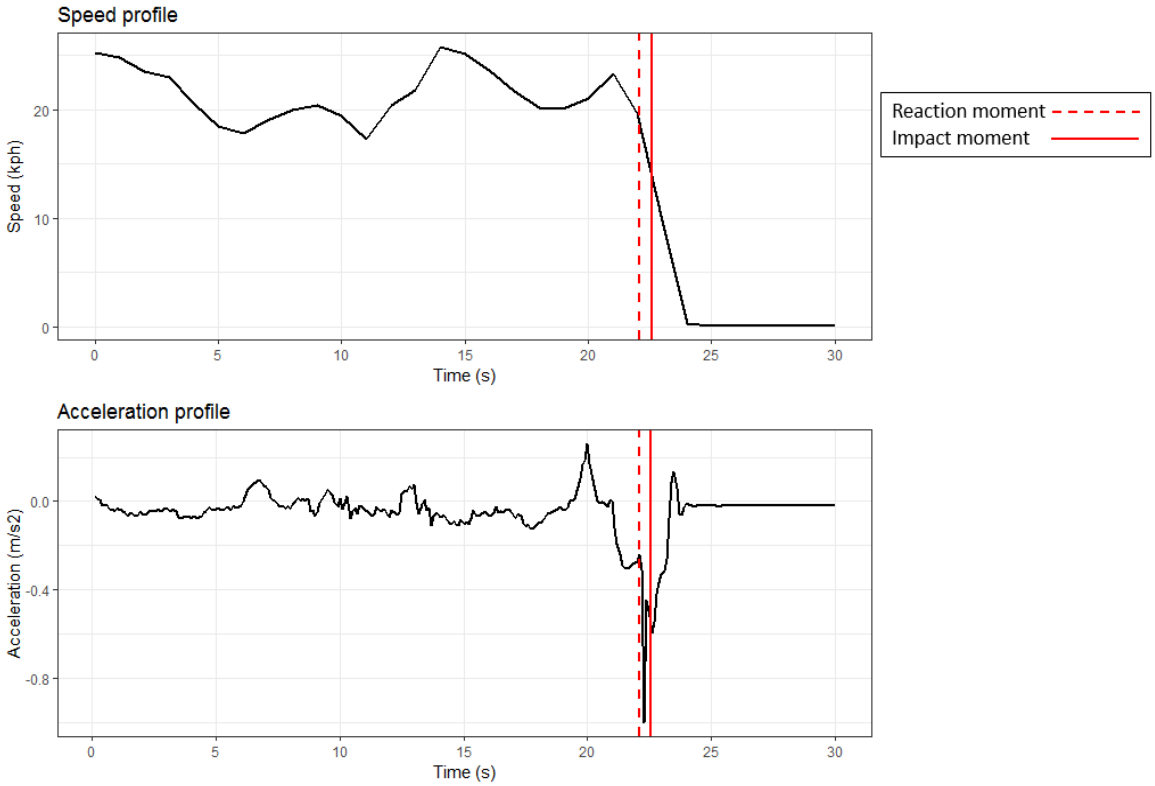


Figure 4.2 Speed and acceleration profile of a randomly chosen crash

measures of driving volatility to quantify driving stability. It should be noted that there are crashes in which the driver did not react, or he/she reacted after the impact time. Thus, we used either the reaction time or impact time (whichever comes first).

Results

Descriptive Statistics

The descriptive statistics of the variables used in the modeling are shown in Table 4.1. The table presents the statistics regarding the crash intensity, measures of stability performance (driving volatility), driving behavior, and roadway/environmental factors. Based on the descriptive statistics, 40.2% of crashes are low-risk tire strikes, 36.8% are minor intensity crashes, 13.6% are moderate intensity crashes, and 9.4% are severe crashes. As mentioned in the methodology, nine measures of driving volatility are utilized and calculated using seconds of the vehicle data prior to crash involvement, and the summary of the descriptive statistics is shown in Table 4.1.

Focusing on key variables of driving behavior, the average speed of the vehicles is 8.2 m/s, ranging from 0.13 to 34.4 m/s. prior to the crash, 9.72% of drivers are observed while showing the aggressive behavior, and 64.67 percent are distracted with a secondary task. In terms of environmental factors, most crashes happened in business/industrial (46.8%), and moderate residential areas (19.8%).

Table 4.1 Descriptive statistics of variables

Variable	Description	Mean/ Percent	S.D./ frequency	Min	Max
Crash intensity					
	Low-risk Tire Strike	40.19%	248	0	1
	Minor Crash	36.79%	227	0	1
	Moderate Crash	13.61%	84	0	1
	Severe Crash	9.4%	58	0	1
Measures of performance stability					
Speed	*Speed prior to the crash (m/s)*	8.2	6.2	0.13	34.4
$Speed - S_{dev}$ (m/s)	*Standard deviation of speed*	3.9	2.35	0.15	12.43
$Speed - V_f$ (m/s)	*Time varying stochastic volatility*	0.6	0.43	0.01	3.11
$Speed - C_v$ (m/s)	*Coefficient of variation of speed*	0.66	0.4	0.01	2.68
$Speed - Q_{cv}$ (m/s)	*Quartile coefficient of variation of speed*	0.51	0.3	0.01	1.00
$AccDec - S_{dev}$ (m/s^2)	*Standard deviation of acceleration and deceleration*	1.01	0.5	0.07	3.77
$Accel - C_v$ (m/s^2)	*Coefficient of variation of acceleration*	0.91	0.32	0.31	2.51
$Decel - C_v$ (m/s^2)	*Coefficient of variation of deceleration*	1.04	0.37	0	2.72
$Accel - Q_{cv}$ (m/s^2)	*Quartile coefficient of variation of acceleration*	0.67	0.19	0	1.00
$Decel - Q_{cv}$ (m/s^2)	*Quartile coefficient of variation of deceleration*	0.71	0.18	0	1.00
Driving behavior					
Hand on wheel	*Two hands on wheel*	46.52%	287	0	1
	Other	53.48%	330	0	1
Aggressive	*Aggressive driving*	9.72%	60	0	1
	None	90.28%	557	0	1
Distracted	*Distracted driving*	64.67%	399	0	1
	None	35.33%	218	0	1
Seatbelt	*Seatbelt used*	90.6%	559	0	1
	No	9.4%	58	0	1
Legal Maneuver	*Yes*	82.82%	511	0	1
	No	17.18%	106	0	1
Roadway/Environmental factors					
Locality	*Business/Industrial*	46.84%	289	0	1
	Bypass/Divided Highway with traffic signals	2.59%	16	0	1
	Church	2.11%	13	0	1
	Bypass/Divided Highway with no traffic signal	6.65%	41	0	1
	Moderate residential	19.77%	122	0	1
	Open country	1.13%	7	0	1
	Open residential	5.19%	32	0	1
	Playground	0.81%	5	0	1
	School	7.78%	48	0	1
	Urban	7.13%	44	0	1

Table 4.1 Continued

Relation to Junction	*Relation to junction (base: non-junction)*	27.07%	167	0	1
	Driveway, alley access, etc.	5.67%	35	0	1
	Entrance/Exit ramp	2.11%	13	0	1
	Interchange area	3.4%	21	0	1
	Intersection	19.77%	122	0	1
	Intersection-related	11.35%	70	0	1
	Other	0.49%	3	0	1
	Parking lot entrance/exit	13.94%	86	0	1
	Parking lot, within boundary	16.21%	100	0	1
Density	*Traffic density (base: LOS A)*	73.42%	453	0	1
	LOS B	18.31%	113	0	1
	LOS C and Below	8.27%	51	0	1
Road Alignment	*Straight*	85.74%	529	0	1
	Curve	14.26%	88	0	1
Roadway type	*Divided (median strip or barrier)*	22.69%	140	0	1
	No lanes	17.18%	106	0	1
	Not divided - center 2-way left turn	5.51%	34	0	1
	Not divided - simple 2-way traffic way	48.30%	298	0	1
Surface condition	*One-way traffic*	6.32%	39	0	1
	Dry	74.39%	459	0	1
	Ice/snow	3.24%	20	0	1
	Other	0.32%	2	0	1
	Wet	22.04%	136	0	1
Weather	*Weather (base: no adverse condition)*	85.58%	528	0	1
	Adverse Conditions	8.59%	53	0	1
	Mist/Light Rain	5.83%	36	0	1
Light condition	*Daylight*	71.64%	442	0	1
	Darkness, lighted	19.77%	122	0	1
	Darkness, not lighted	5.02%	31	0	1
	Dawn/Dusk	3.57%	22	0	1

Modeling Results

In this study, seconds of vehicle kinematic data prior to crash involvement containing unintentional driving behavior are used, and 9 measures of driving volatility are calculated by applying defined functions on vehicles speed and acceleration/deceleration. Table 4.2 provides the Spearman's correlation matrix between the volatility measures and crash intensity. Based on the results, it can be observed that six measures of volatility

significantly correlated with crash intensity, among which $Speed - S_{dev}$, $AccDec_x - S_{dev}$ and $Decel - C_v$ have the highest correlation. For model parsimony, two driving volatility measures are selected as a proxy for the stability performance measure. Therefore, $Speed - S_{dev}$ is selected as a measure for speed volatility, and $Decel - C_v$ as a measure for deceleration volatility. It should be mention that $Speed - S_{dev}$ and $AccDec_x - S_{dev}$ measures are highly correlated with each other (0.733). From the point forward, the terms "speed volatility" for $Speed - S_{dev}$, and "deceleration volatility" for $Decel - C_v$ are used.

In the following, the modeling results for three models, including speed volatility, deceleration volatility, and crash intensity, will be discussed. The first two models explore the impact of driving behavior and roadway/environmental factors on driving stability. In the third model, speed and deceleration volatilities along with behavioral and roadway/environmental variables used in the crash intensity model. Figure 4.3 illustrates the structure of the final model, obtained by a forward stepwise model selection, considering intuition, statistical significance, model fit, and parsimony. In the model estimation, several interactions among key variables including driving volatility, distracted driving, and aggressive driving were considered, but none of them were statistically significant at the 5% level. In the following, the modeling results for speed and deceleration volatilities will be discussed, and finally, their contribution on crash intensity will be explored.

Table 4.2 Correlation of volatility measures with the crash intensity

		Measures of driving volatility								
		$Speed - S_{dev}$	$Speed - V_f$	$Speed - C_v$	$Speed - Q_{cv}$	$AccDec - S_{dev}$	$Accel - C_v$	$Decel - C_v$	$Accel - Q_{cv}$	$Decel - Q_{cv}$
Intensity	*Corr.*	0.257	0.113	-0.073	-0.152	0.289	0.041	0.351	0.005	0.210
	Sig.	*0.000*	*0.005*	*0.071*	*0.000*	*0.000*	*0.315*	*0.000*	*0.906*	*0.000*

Speed volatility

As mentioned, we used $Speed - S_{dev}$ volatility measure to capture the variations in instantaneous vehicle speed prior to the crash. The Tobit model is developed to assess the correlated of speed volatility in terms of driver behavior and roadway/environmental before the crash occurrence. The modeling results for the fixed and random parameter

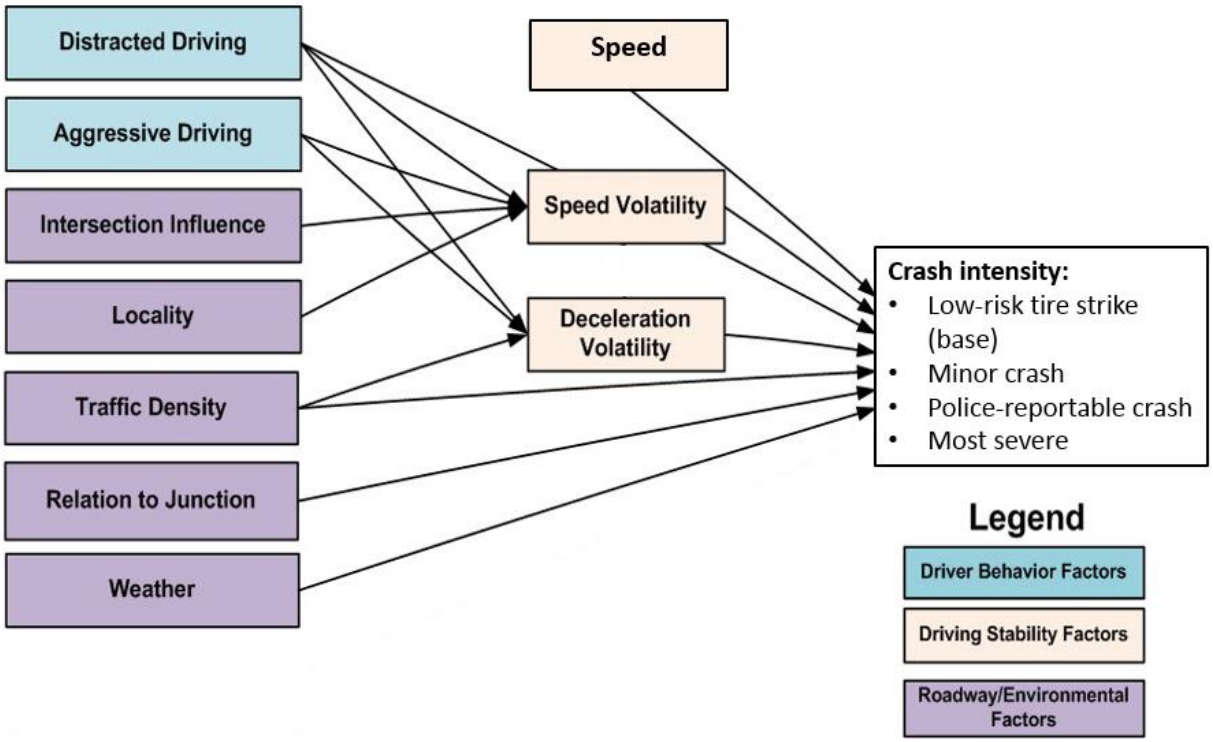


Figure 4.3 Pathway diagram of the model

linear regression model is shown in Table 4.3. For the model selection, forward step-wise variable selection is performed. Based on the results, driver behavior factors are highly associated with speed volatility. Comparing AIC values, the random parameter model outperformed fixed parameter model. Focusing on the random parameter model, distracted driving is associated with 0.41 unit increase in the speed volatility comparing to non-distracted driving, which is consistent with previous studies which found distracted driving impairs the driving stability performance (Beede and Kass 2006, Hanowski *et al.* 2006, Stavrinos *et al.* 2013). The possible reason is when the driver is distracted, the driver workload increases, leading to a decrease in the reactions (Horberry *et al.* 2006). Distractions can divert the drivers' attention from monitoring the speed of the vehicle (Young and Salmon 2012), which can increase the driving volatility. Furthermore, it has shown that a higher workload and distraction level of drivers increase the variations in speed and deceleration (Rakauskas *et al.* 2004). Furthermore, controlling for other variables, aggressive driving is associated with a 2.18 units increase in the speed volatility. The results are consistent with the previous studies which had shown that aggressive driving impairs the driving stability (Shinar and Compton 2004, Hamdar *et al.* 2008). It has shown that aggressive driving is highly associated with variations in acceleration, in terms of vehicular jerk (Feng *et al.* 2017). Aggressive driving will increase the workload, and due to aggressiveness, the driver performs harder accelerations and brakes (Liu and Lee 2005).

Referring to roadway/environmental characteristics, an increase in the number of lanes is correlated with an increase in speed volatility. On average, one unit increase in the number of lanes is associated with a 0.24 units increase in the speed volatility. Controlling

for other variables, driving on the divided highways with traffic signals, on average is correlated with increases the speed volatility for 2.09 units comparing to the business areas. Also, driving volatility in divided highways without traffic signal is associated with increases the speed volatility for 1.73. Speed volatility in moderate and open residential are 0.48 and 1.36 units higher than the business area, controlling for other variables. Driving on locations that are influenced by the intersection, on average is correlated with a 0.54 units increase in the speed volatility. Overall, the main underlying reason might be the complexity in the driving environment and traffic flow condition. These locations not only increase the drivers' workload but also in a more congested area, there is a higher oscillation in driving speed, which potentially can increase the driving volatility. Other location factors are not significant in the model.

Table 4.3 Tobit Modeling results for speed volatility (as dependent variable)

Variable Description	Fixed-parameter			Random parameter		
	β	S.E.	p-value	β	S.E.	p-value
Number of the lanes	0.222	0.055	< 0.001	0.238	0.054	< 0.001
Std. Number of lanes	-	-	-	0.222	0.027	< 0.001
Distracted with secondary task *(Yes=1, No=0)*	0.434	0.175	0.016	0.415	0.161	0.010
Std. Distracted with secondary task	-	-	-	0.409	0.092	< 0.001
Aggressive driving *(Yes=1, No=0)*	2.261	0.294	< 0.001	2.185	0.241	< 0.001
Intersection influence *(Yes=1, No=0)*	0.531	0.181	0.002	0.542	0.161	< 0.001
Locality *locality (base: business area)*						
Bypass/Divided Highway with traffic signals	2.052	0.533	< 0.001	2.095	0.427	< 0.001
Bypass/Divided Highway with no traffic signal	1.592	0.356	< 0.001	1.733	0.226	< 0.001
Moderate residential	0.469	0.226	0.033	0.48	0.212	0.024
Open residential	1.321	0.397	0.001	1.357	0.377	0.003
School	-0.250	0.324	0.454	-0.216	0.369	0.557
Urban	-0.244	0.335	0.402	-0.240	0.375	0.521
Other	0.107	0.433	0.805	0.159	0.352	0.649
Model intercept	2.280	0.226	< 0.001	2.241	0.221	< 0.001
Disturbance standard deviation	2.067	0.058	< 0.001	1.855	0.051	< 0.001
Number of observations	617			617		
AIC	2672.4			2651.4		
LL at the model	-1323.6			-1310.7		
LL at the null	-1401.1			-1401.1		

Deceleration volatility

The modeling results of the Tobit model for the deceleration volatility ($Decel - C_v$) is shown in Table 4.4. This measure attempts to capture the variations in the vehicle deceleration values prior to the crash occurrence. In order to explore the contributing factors on the deceleration volatility, fixed and random parameter Tobit models are developed. Comparing the AIC value, the random parameter model outperformed the fixed parameter model. Consistent with speed volatility model, driver behavior factors are

significantly associated with driving volatility. Distracted driving is associated with a 0.05 units increase in the deceleration volatility, controlling for other variables. Furthermore, aggressive driving is correlated with a 0.09 units increase in the deceleration volatility. Focusing on traffic density, congested locations are positively correlated the deceleration volatility. Driving in locations with a level of service *B* and *C* is correlated with the driving volatility on average for 0.13 and 0.27 comparing to a level of service *A*, respectively.

Crash intensity model

In the previous sections, the correlation of driving behavior and roadway/environmental characteristics with the driving volatilities is explored. This section investigates the direct impact of driving volatilities, along with speed, driver behavior, and

Table 4.4 Tobit Modeling results for deceleration volatility (as dependent variable)

Variable	Description	Fixed parameter			Random parameter		
		β	S. E.	p-value	β	S. E.	p-value
Distracted with secondary task *(Yes=1, No=0)*		0.059	0.030	0.051	0.055	0.025	0.036
Std. Distracted with secondary task		-	-	-	0.211	0.013	< 0.001
Aggressive driving *(Yes=1, No=0)*		0.116	0.049	0.020	0.091	0.040	0.009
Level of Service	*(Base: LOS A)*						
	LOS B	0.137	0.038	< 0.001	0.132	0.031	< 0.001
	Std deceleration LOS B	-	-	-	0.226	0.028	< 0.001
	LOS C and Below	0.260	0.054	< 0.001	0.267	0.044	< 0.001
	Std deceleration LOS C and Below	-	-	-	0.291	0.035	< 0.001
Intercept		0.942	0.027	< 0.001	0.953	0.022	< 0.001
Sigma		0.366	0.010	< 0.001	0.287	0.007	< 0.001
Summary statistics	*Number of observations*	617			617		
	Deviance at null model	-269.8			-269.8		
	Deviance at model	-261.2			-230.3		
	AIC	534.5			478.6		

roadway/environmental factors on the crash intensity. While several studies have analyzed the association of driving speed, pre-crash behavior, roadway/environmental characteristics with crash intensity, this study also investigates the impact of driving instability in terms of variations in instantaneous driving behavior just prior to crash occurrence. The fixed-parameter and random parameter ordered probit model is developed based on the intuition, and significance of the variables. The modeling results are shown in Table 4.5. The random parameter model performed better in terms of AIC and pseudo R-squared.

Focusing on driving instability, while previous studies examined the impact of speed dispersion among vehicles and its impact on the crash rate (Taylor *et al.* 2000, Qu *et al.* 2014), this study investigates the variations in speed and deceleration of the subject vehicle prior to a crash and their association with crash intensity. Based on the results, both speed and deceleration volatilities are highly associated with crash intensity. Higher variations in driving speed and deceleration in terms of speed and deceleration volatilities is correlated with increasing the probability of a severe crash. Higher volatility indicates the inability of the driver to control the vehicle before the crash, which potentially increases the severity of the crash. Based on the results, the vehicle speed is significantly contributing to the intensity outcome, as expected. Similar to previous studies (Kockelman and Kweon 2002, Das and Abdel-Aty 2011), by increasing vehicle speed prior to the collision, the likelihood of an intense crash is increased. With an increase in speed, the vehicle has higher kinematic energy, and this released energy in a crash can increase the likelihood of serious severity (Hauer 2009, Pei *et al.* 2012). Among the driver behavior factors, distracted driving is significantly associated with crash intensity.

Focusing on roadway/environmental factors, adverse weather conditions are associated with crash intensity comparing to no adverse condition. Also, previous studies found similar results (Abdel-Aty 2003). Crashes that occurred in congested areas are more severe, which might be due to the fact that these locations have higher variations in speed, leading to a higher number of severe crashes (Vadeby and Forsman 2017). Comparing to non-junction locations, crashes which happened at entrance/exit ramps and interchange areas are more severe. While intersection and parking related crashes are less severe than non-junction crashes. Crashes that happened in traffic condition with LOS B and C and below are more severe than LOS A.

Table 4.5 Modeling results for crash intensity

Variable	Description	Fixed parameter			Random parameter		
		β	S.E.	p-value	β	S.E.	p-value
Intercept		-1.293	0.208	<0.001	-1.582	0.217	<0.001
Speed	*Speed prior to crash occurrence*	0.021	0.010	0.044	0.032	0.011	0.004
Speed volatility	*Standard deviation of speed*	0.053	0.025	0.038	0.050	0.028	0.074
Deceleration volatility	*coefficient of variation of deceleration*	1.021	0.131	<0.001	1.261	0.147	<0.001
	Std deceleration volatility	-	-	-	0.470	0.048	<0.001
Distracted	*Distracted with secondary task (Yes=1, No=0)*	0.55	0.102	<0.001	0.677	0.106	<0.001
Density	*Level of Service (base: LOS A)*						
	LOS B	0.530	0.125	<0.001	0.597	0.137	<0.001
	Std deceleration LOS B				1.132	0.048	<0.001
	LOS C and Below	0.951	0.179	<0.001	1.223	0.190	<0.001
	Std deceleration LOS C and Below				0.672	0.173	<0.001
Relation to Junction	*Relation to junction (base: non-junction)*						
	Driveway, alley access, etc.	-0.321	0.216	0.137	-0.335	0.212	0.114
	Entrance/Exit ramp	0.653	0.322	0.042	0.814	0.357	0.022
	Interchange area	0.519	0.265	0.049	0.767	0.319	0.016
	Intersection	-0.362	0.143	0.011	-0.417	0.154	0.007
	Intersection-related	-0.353	0.165	0.032	-0.423	0.181	0.019
	Other	-0.259	0.659	0.693	-0.405	0.991	0.682
	Parking lot entrance/exit	-0.813	0.172	<0.001	-1.001	0.191	<0.001
	Parking lot, within boundary	-0.706	0.171	<0.001	-0.760	0.176	<0.001
Weather	*Weather (base: no adverse condition)*						
	Adverse Conditions	0.365	0.198	0.065	0.429	0.191	0.024
	Mist/Light Rain	0.138	0.169	0.414	0.199	0.190	0.293
Crash intensity	μ_1	1.355	0.065	<0.001	1.652	0.098	<0.001
	μ_2	2.189	0.088	<0.001	2.753	0.131	<0.001
Summary Statistics	*Number of observations*	617			617		
	AIC	1260.8			1250.8		
	Deviance at null model	-757.91			-757.91		
	Deviance at model	-611.39			-603.4		
	Pseudo-R^2	0.19			0.20		

Path analysis of driving volatility and crash intensity model

One of the common methods in addressing direct and indirect association of factors is path analysis. As shown in Table 4.5, speed and deceleration volatilities are highly associated with an increase in the probability of severe crashes. The marginal effect is provided in Table 4.6, which illustrates the direct effect of driving volatilities on crash intensity. On the other hand, the contributing factors that are associated with the speed and deceleration volatilities are indirectly associated with the intensity outcome of the crash. Although some factors are not significant in the intensity model, they are significantly associated with driving volatilities and indirectly correlated with the intensity outcome. As an illustration, aggressive driving is not significant in the severity model, and one might conclude that it is not correlated with crash intensity, while it is significant in both speed and deceleration volatility models and indirectly increase the likelihood of a severe crash. In the following, the marginal effect analysis for severe crashes is discussed, and the results for other severity categories can be found in Table 4.6.

As discussed in the previous section, instability in driving prior to a crash occurrence significantly increases the probability of a severe crash. Referring to volatility measures, results revealed that one-unit increase in the speed volatility is associated with a 0.4 percent chance of severe crashes. Considering a wide range of speed volatility, its impact can be substantial. Furthermore, higher deceleration volatility positively and significantly associates with an increase the probability of a severe crash. A one-unit increase in deceleration volatility is associated with an increase in the chance of severe crash for 10.9 percent. In addition, the vehicle speed is directly associated with the crash intensity and 1 m/s increase in the speed of the vehicle is associated with a 0.3 percent increase

the chance of a severe crash, which is in line with previous studies (O'donnell and Connor 1996, Yasmin *et al.* 2014).

Previous studies investigated the association of distracted driving on the crash intensity, and it was shown that distracted driving increases the probability of a severe crash (Neyens and Boyle 2008, Donmez and Liu 2015). Modeling results revealed that distracted driving increases the probability of severe crash by 11.1 percent. On the other hand, although aggressive driving is not significant in the crash intensity model, the indirect association through speed and deceleration volatilities increase the probability of a severe crash by 1.3 percent.

Referring to the crash location, comparing to the non-junction, entrance/exit ramps and interchange areas increase the likelihood of a severe crash by 9 and 8.3 percent, respectively. On the other hand, parking lot crashes are less severe than at non-junction areas. Speed volatility at intersections is higher than non-intersections, indirectly increasing the probability of a severe crash.

Table 4.6 Total marginal effect of random parameter model on crash intensity (in percent)

Variable	Direct Marginal Effect			Indirect Marginal Effect via speed volatility			Indirect Marginal Effect via deceleration volatility			Total Marginal Effect		
	Minor	Moderate*	Severe**	Minor	Mod.	Severe	Minor	Mod.	Severe	Minor	Mod.	Severe
Speed prior to the crash	0.3	0.3	0.3							0.3	0.3	0.3
Speed volatility	0.5	0.4	0.4							0.5	0.4	0.4
Deceleration volatility	11.7	11.1	10.9							11.7	11.1	10.9
Aggressive driving *(Yes=1, No=0)*				0.3	0.9	1.2	0.7	0.1	0.1	1.0	1.0	1.3
Distracted *(Yes=1, No=0)*	2.2	8.6	9.9	0.0	0.2	0.3	0.0	0.6	0.9	2.2	9.4	11.1
Traffic density *(base: LOS A)*												
LOS B	1.6	4.7	6.9				1.1	1.5	1.6	2.7	6.2	8.5
LOS C and Below	0.0	11.6	15.8				2.1	3.6	3.2	2.1	15.2	19.0
Relation to junction												
Driveway, alley access, etc.	-3.8	-2.7	-2.4							-3.8	-2.7	-2.3
Entrance/Exit ramp	2.5	8	9							2.5	8	9
Interchange area	2.7	7.5	8.3							2.7	7.5	8.3
Intersection	-4.9	-3.3	-2.9							-4.9	-3.3	-2.9
Intersection-related	-4.9	-3.4	-3.0							-4.9	-3.4	-3.0
Other	-4.7	-3.2	-2.9							-4.7	-3.2	-2.9
Parking lot entrance/exit	-13.1	-6.9	-5.7							-13.1	-6.9	-5.7
Parking lot, within boundary	-9.7	-5.6	-4.7							-9.7	-5.6	-4.7
Weather												
Adverse Conditions	2.7	4.1	4.2							2.7	4.1	4.2
Mist/Light Rain	1.6	1.8	1.8							1.6	1.8	1.8
Number of the lanes				0.0	0.8	1.8				0.0	0.8	1.8
Intersection influence *(Yes=1,No=0)*				0.1	0.2	0.3				0.1	0.2	0.3
Locality *(base: business area)*												
Bypass/Divided Highway with traffic signals				0.3	0.9	1.2				0.3	0.9	1.2
Bypass/Divided Highway with no traffic signal				0.3	0.7	1.0				0.3	0.7	1.0
Moderate residential				0.0	0.2	0.3				0.0	0.2	0.3
Open residential				0.2	0.6	0.7				0.2	0.6	0.7
School				0.0	-0.1	-0.1				0.0	-0.1	-0.1
Urban				0.0	-0.1	-0.1				0.0	-0.1	-0.1
Other				0.0	0.0	0.0				0.0	0.0	0.0

* Police reportable crash (base is tire-strike)
** Most severe crash

Limitations

Although the NDS data is one of the richest datasets, it has some limitations. The drivers might not be representative of all drivers since they were hired with monetary incentive. The subset of the data used in this study does not include sociodemographic information of the participants, and vehicle characteristics such as vehicle type, make, year, etc. Such variables can potentially enhance the explanatory power of the models. In addition, the intensity of the crashes does not include high injury severity crashes. While there might be human errors and personal judgments in coding the data, the descriptive statistics seem reasonable. The parametric and distributions assumptions of the frequentist models are acknowledged.

Conclusion and future research

In general, driving behavior is known as a key contributing factor to traffic crashes. The emergence of high-resolution naturalistic driving data provides a promising opportunity for researchers to investigate the association of pre-crash behavior with crash intensity. This study attempted to answer the research questions by utilizing the concept of driving volatility to extract useful information and investigate the association of driving stability with crash intensity. Although previous studies have explored the impact of speed on crash intensity (Kockelman and Kweon 2002, Aarts and Van Schagen 2006, Das and Abdel-Aty 2011), the correlation of crash intensity with the instability prior to the crash occurrence remains largely unexplored. The key finding of this study is that instability in driving is associated with an increasing likelihood of more intense crashes. Specifically,

speed and deceleration volatilities are positively associated with crash intensity. In addition, results revealed that driver distraction and driving speed are positively correlated with the intensity outcome of the crash, as expected. Such results would not be possible without the availability of microscopic vehicle kinematics and driving behavior prior to crash involvement.

The pathway framework further explored how driving volatility is influenced by driver behavior and roadway/environmental factors. A subset of SHRP2 NDS data containing 617 crash events with around 0.18 million temporal observations of microscopic vehicle kinematics is processed and analyzed. Modeling results revealed that aggressive and distracted driving are highly correlated with both speed and deceleration volatilities prior to the crash. In other words, those drivers who were involved in aggressive or distracted driving showed higher variations in vehicular speeds and decelerations. Along these lines, correlation of other roadway factors (e.g., intersections, roadway type, and level of service) are also explored. While lower levels of service, increase the pre-crash driving volatility, drivers at intersections showed higher volatility compared with non-intersection areas.

Focusing on driver behavior, distracted driving has emerged as a variable that both directly and indirectly is associated with intense crashes. Distracted driving directly and indirectly increases the probability of severe crash for 9.9 and 1.2 percent, respectively. Additionally, aggressive driving was indirectly associated with crash intensity through volatility, but not directly, and increase the likelihood of severe crash involvement for 1.3 percent.

Given the association of driving volatility with crash intensity, and its potential to serve as a leading indicator, by detecting high risk driving behaviors, alerts and warnings can be generated and transmitted to the subject vehicle driver to encourage lowering his/her driving volatility, and surrounding vehicles can be warned of a potential hazard via vehicle-to-vehicle communication. While this study explores the association of crash intensity with speed, driving volatility, driver behavior, and surrounding environment, future research can focus on studying the correlates of driving stability in normal conditions, and how providing warnings might enhance driving stability. From the methodological standpoint, this research can be further extended by jointly modeling their speed and deceleration volatilities to account for the potential correlation utilizing the multivariate random-parameter Tobit model (Anastasopoulos 2016, Zeng *et al.* 2017b). Furthermore, the fixed thresholds estimated by the probit model may lead to biased estimates (Eluru *et al.* 2008), which can be addressed by developing generalized ordered probit models (Eluru *et al.* 2008, Balusu *et al.* 2018, Zeng *et al.* 2019). In the literature, several methods are proposed to generalize the probit model by overcoming the limitation of a fixed threshold. This study also tried to use proportional odds ratio and random threshold ordered probit models. However, the models' performance in terms of AIC and BIC did not improve. Future studies can further extend this research from a methodological standpoint by applying other generalized ordered probit model, e.g., Bayesian spatial generalized ordered logit (Zeng *et al.* 2019).

CHAPTER 5 : DRIVING IMPAIRMENTS AND DURATION OF DISTRACTIONS: ASSESSING CRASH RISK BY HARNESSING MICROSCOPIC NATURALISTIC DRIVING DATA

This chapter is a modified version of a research article by Ramin Arvin and Asad J. Khattak. *"Driving Impairments and Duration of Distractions: Assessing Crash Risk by Harnessing Microscopic Naturalistic Driving Data."* The manuscript presented at the 99th Annual Meeting of Transportation Research Board Conference at Washington DC, and it is currently under second-stage of review in Accident Analysis and Prevention.

Abstract

Distracted and impaired driving is a key contributing factor in crashes, leading to about 35% of all transportation-related deaths in recent years. Along these lines, cognitive issues like inattentiveness can further increase the chances of crash involvement. Despite the prevalence and importance of distracted driving, little is known about how the duration of distractions is associated with critical events, such as crashes or near-crashes. With new sensors and increasing computational resources, it is possible to monitor drivers, vehicle performance, and roadways to extract useful information, e.g., eyes off the road, indicating distraction and inattention. Using high-resolution microscopic SHRP2 naturalistic driving data, this study conducts in-depth analysis of both impairments and distractions. The data has more than 2 million seconds of observations of 7394 baselines (no event), 1237 near-crashes, and 617 crashes. The event data was processed and linked with driver behavior and roadway factors. The interval of distracted driving during the period of observation (15 seconds) were calculated; next, rigorous fixed and random parameter logistic regression models of crash risk was estimated. The results reveal that alcohol and drug impairment is associated with a substantial increase in extreme event involvement of 29.7%, and the highest correlations with crash risk include duration of

distraction by cellphones, driver reading or writing, and atypical distraction, e.g. insect in the car. Using detailed pre-crash data from instrumented vehicles, the study contributes by quantifying crash risk vis-à-vis detailed driving impairment and streams of secondary task involvement and discusses implications of the results.

Introduction

Human-based errors such as distracted driving, alcohol/drug impairment, fatigue driving, and speeding are commonly known as the main contributing cause of fatal crashes (Pietrasik 2018). In particular, distracted and impaired driving contributes to about 35% of all transportation-related deaths, e.g., 10,497 fatalities in 2016, based on US Traffic Safety Facts (NHTSA 2017). While the driving task requires execution of several cognitive, sensory, and psychomotor skills (Young *et al.* 2007), it is common to observe drivers under impairment (Fan *et al.* 2019) and engaged in various non-driving tasks such as using a cellphone, interacting with other passengers, listening to music, and even writing and reading (Stutts *et al.* 2005, Dingus *et al.* 2016, Kamrani *et al.* 2019). Impaired and distracted driving allocate fewer available attention of driver to driving tasks such as controlling vehicle position and maintaining speed (Martin *et al.* 2013, Verstraete *et al.* 2014, Paolo Busardo *et al.* 2018). Distracted driving can be defined as “a diversion of attention from driving, because the driver is temporarily focusing on an object, person, task or event not related to driving, which reduces driver’s awareness, decision making ability and/or performance, leading to an increased risk of corrective actions, near-crashes, or crashes” (Regan *et al.* 2008). Distracted driving is known as a prominent contributing factors in traffic crashes (Lee *et al.* 2008). It is estimated that driver inattention is contributing to around 23 percent of police reported crashes (Klauer *et al.* 2006). In

addition, the introduction of cellphones worsened the situation and became widely common (Anon 2011, Engelberg *et al.* 2015, Arvin *et al.* 2017), especially among young drivers (Anon 2011). While a majority of drivers are aware of the associated risks with distracted driving, more than 25 percent still frequently use their cellphone while driving (Motamedi and Wang 2016). Cellphone distracted driving is one of the great challenges in the transportation field, as it contributes to 18 percent of fatal and 5 percent of injury crashes across the U.S. based on the police-reported crash data (Overton *et al.* 2015). However, these crash databases are deficient due to unreported crashes (around 50% of no-injury and 25% of minor-injury crashes were not reported to the police (NHTSA 2009)). Furthermore, such datasets under-report prevalence of distracted driving and does not have information on distraction duration.

On the other hand, impaired driving, in terms of alcohol/drug impairment, fatigue, emotional state, is widely common among drivers, Although share of alcohol-related traffic fatalities significantly dropped in last decades (from 48 percent in 1982 to 28% in 2016), still it remains the main contributing factor in fatal crashes. It is estimated that prevalence of alcohol related impaired driving among drivers aged 16 years and older is 11.6 percent (Lipari *et al.* 2016). Impairment substantially affect drivers' ability to control vehicle and increase driver-risk taking (Laude and Fillmore 2015). In terms of driver performance, impaired driving significantly increases number of errors (Verster *et al.* 2009) and driver reaction time (Deery and Love 1996, Verster *et al.* 2009), worsen lateral (Hartman *et al.* 2015) and longitudinal vehicle control (Hartman *et al.* 2016). While these studies mainly investigated the correlation of distracted and impaired driving with driving performance using a driving simulator (Rumschlag *et al.* 2015, Saifuzzaman *et al.* 2015,

Li *et al.* 2016), it has been shown that driving simulator sickness may affect the validity and reliability of results (Nickkar *et al.* 2019b). Crash datasets suffer from unreported crashes and near-crashes, and lack of detail information on pre-crash driver state and behavior. While crash only databases can only be used for frequency and prevalence of specific factors with crashes (Shinar and Gurion 2019), naturalistic driving study (NDS) data provides an opportunity to analyze the associated risk with these factors. Emergence of high-resolution microscopic NDS data compensates for this limitation by collecting real data on real-world condition. The second Strategic Highway Research Program (SHRP2), which is sponsored by National Academy of Science, is the largest naturalistic driving data collection by collecting data on more than 3500 drivers (Dingus *et al.* 2015). It provides an opportunity for researchers to gain insight into factors leading to an extreme event, especially actual driver state, behavior, and performance (Dingus 2003, Dingus *et al.* 2011). Such a dataset helps researchers to overcome limitations of traditional datasets and explore not only minor crashes but also pre-crash driver state and behavior, specifically impairment and distraction profile.

Overall, the goal of this study is to harness microscopic big data from multiple sources and link this information with driver behavior, roadway, and environmental factors in order to evaluate impaired driving and the association of duration of different distraction types on the probability of occurrence of crashes and near-crashes. Given that distracted driving and human error are the key contributing factors in crashes (Kludt *et al.* 2006, Arvin *et al.* 2017, Shinar 2017), the findings of this research identify the role of impairments and distraction types that are highly associated with crash risk, and explore how impairment and duration of distraction affect driving performance and risk of a crash.

Data

Data description

This study utilized the second Strategic Highway Research Program (SHRP2) data. More than 4 petabytes of various information were collected in the original data, which makes the SHRP2 the most comprehensive naturalistic driving study. The high-quality and high-resolution data was captured from 2010 to 2013 via multiple sensors including a camera, accelerometer, alcohol sensor, forward sensor, and a data acquisition system (DAS) with a 10 Hz frequency (Hankey *et al.* 2016). The study has information on more than 3500 drivers from six states (Washington, New York, Pennsylvania, North Carolina, Florida, and Indiana) across the U.S., with more than five million trips covering more than 50 million miles travelled (Hankey *et al.* 2016). The NDS data includes vehicular movement data (e.g., speed, acceleration), along with information regarding the drivers' behavior, roadway factors, and environmental factors from the videos coded by the data reductionist using the appropriate protocols to ensure consistency and high quality.

This study considers a subset of original SHRP2 data, containing 9,239 trips taken by 1,546 drivers with 7394 baseline events, 1228 near-crashes, and 617 crashes. In the data, the definition of a crash is "any contact that the subject vehicle has with an object, either moving or fixed, at any speed in which kinetic energy is measurably transferred or dissipated". Even though near-crashes did not result in a crash, the data for crash events and near-crash events were combined in this study and defined as extreme events. For each extreme event, 30 seconds of vehicular movement data is available. It is worth noting that the data contains time in which the driver uses evasive maneuvers to avoid

the extreme event and also the seconds after the occurrence of an extreme event. Since this paper examines the association of distracted driving before extreme event occurrence, the seconds after the crash should be excluded which will be further discussed. Moreover, since we are investigating the association of distraction duration with crash risk, the information on driver distraction needs to be linked with vehicle kinematics.

Data Pre-processing

The data contains detailed information on baseline and extreme events coded as categorical variables. Furthermore, for baseline events, we have 20 seconds of vehicle kinematics and a distraction profile, while for extreme events 30 seconds of vehicle kinematics was collected. However, these 30 seconds contain the time that the vehicle is involved in the crash (near-crash). Therefore, the data was pre-processed to remove time in which the crash occurs. This is discussed in detail below.

Data recoding

As mentioned, the data contains rich and detailed information on driver behavior, roadway condition, and environment condition, etc., and the variable “secondary task” was coded into 62 different groups. However, in some groups there are similarities that allow the data to be merged into more intuitive and cleaner variables. To do this, the coding approach developed by Kamrani et al. (Kamrani *et al.* 2019) was used to aggregate the categories when considering similarity of variables and number of observations in each group. The original and recoded variables for the distracted driving are provided in the Table 5.1. It should be noted that the full description of the other variables is provided in section 5.1.

Table 5.1 Definition and list of recoded variables and their final Categories

Variable	Affected Original Values in the Variable*	Coded as
Driving Behavior	Aggressive driving, other/ Aggressive driving, specific, directed menacing actions/ Cutting in, too close behind other vehicle/ Cutting in, too close in front of other vehicle/ Following too closely	Aggressive
	Did not see other vehicle during lane change or merge/ Driving in other vehicle's blind zone/ Driving without lights or with insufficient lights/ Failed to signal/ Improper backing, did not see/ Improper backing, other/ Improper start from parked position/ Improper turn, cut corner on left/ Improper turn, cut corner on right/ Improper turn, other/ Improper turn, wide left turn/ Improper turn, wide right turn/ Making turn from wrong lane/ Other improper or unsafe passing/ Parking in improper or dangerous location/ Passing on right/ Sudden or improper braking/ Sudden or improper stopping on roadway	Improper Action
	Driving slowly in relation to other traffic: not below speed limit/ Driving slowly: below speed limit	Low Speed
	Exceeded safe speed but not speed limit/ Exceeded speed limit/ Speeding or other unsafe actions in work zone	Speeding
	Illegal passing/ Other sign (e.g., Yield) violation, apparently did not see sign/ Other sign violation/ Signal violation, apparently did not see signal/ Signal violation, intentionally disregarded signal/ Signal violation, tried to beat signal change/ Stop sign violation, apparently did not see stop sign/ Stop sign violation, intentionally ran stop sign at speed/ Wrong side of road, not overtaking/ Stop sign violation, "rolling stop"	Violation
	Apparent general inexperience driving/ Apparent unfamiliarity with roadway/ Avoiding animal/ Avoiding other vehicle/ Avoiding pedestrian	Other
Secondary Tasks	Adjusting/monitoring climate control/Adjusting/monitoring other devices integral to vehicle/Adjusting/monitoring radio/Inserting/retrieving CD (or similar)/ Moving object in vehicle/Object dropped by driver/Object in vehicle, other/Reaching for object, other	Object Distraction
	Applying make-up/Biting nails/cuticles/Brushing, flossing teeth/Combing, brushing, fixing hair/Other personal hygiene/Reaching for personal body-related item/Removing/adjusting clothing/Removing/adjusting jewelry/Removing, inserting, adjusting contact lenses or glasses	Body Related Distraction
	Cell phone, Browsing/Cell phone, Dialing hand-held/Cell phone, Dialing hand-held using quick keys/Cell phone, Dialing hands-free using voice-activated software/Cell phone, Holding/Cell phone, Locating, reaching, answering/Cell phone, other/Cell phone, Talking/listening, hand-held/Cell phone, Talking/listening, hands-free/Cell phone, Texting/Tablet device, locating, reaching/Tablet device, Operating	Cell Phone
	Child in adjacent seat – interaction/ Child in rear seat – interaction/ Passenger in adjacent seat – interaction/ Passenger in rear seat – interaction	Interaction
	Distracted by construction/ Looking at an object external to the vehicle/ Looking at animal/ Looking at pedestrian/ Looking at previous crash or incident/ Other external distraction/	External
	Drinking from open container/ Drinking with lid and straw/ Drinking with lid, no straw/ Drinking with straw, no lid/ Eating with utensils/ Eating without utensils/ Extinguishing cigar/cigarette/ Lighting cigar/cigarette/ Reaching for cigar or cigarette/ Reaching for food-related or drink-related item/ Smoking cigar or cigarette	Drink/Eat/ Smoke
	Cognitive, other/Dancing/ Insect in vehicle/ Other known secondary task/ Pet in vehicle/ Reading/ Unknown/ Unknown type (secondary task present)/ Writing/Other non-specific internal eye glance	Other

Exclusion of Evasive Maneuvers

The aim of this study is to investigate the association of duration of distracted driving on probability of crash occurrence utilizing microscopic data. Therefore, it is vital to consider only the seconds of driving that contain typical driver behavior instead of the seconds that drivers are reacting to a crash stimulus. In other words, we need to exclude the seconds that the driver is reacted to the crash and the time after the crash occurrence. To further demonstrate the time exclusion used, a speed profile, an acceleration profile, and a distraction profile of a random crash event are provided in Figure 5.1. In this event, the crash happened at the 24th second of the data stream, while the driver reacted to the stimulus at 23th second of the data. Therefore, the observations after second 23 need to be excluded for the purpose of this study. In other words, only the seconds of the data up to the second that the driver starts to react to the extreme event was considered in this study. It is worth noting that the data contains crashes in which the driver did not react to the event, or the reaction occurred after impact. Therefore, either impact time or reaction time was used (whichever occurred first). Next, in order to be consistent in all the events (both baselines and extreme events), only 15 seconds of data stream was considered in the analysis. For the extreme events, these 15 seconds were selected from the second that the driver reacted to the crash (or near crash). For example, for the example shown in Figure 5.1, the data from second 8 to 23 was considered in the analysis.

After coding, the data was error-checked for outliers or unexpected distributions, and no major issues were found. The data is of good quality, given that specific protocols were followed in data collection and data processing, and error-checking was completed.

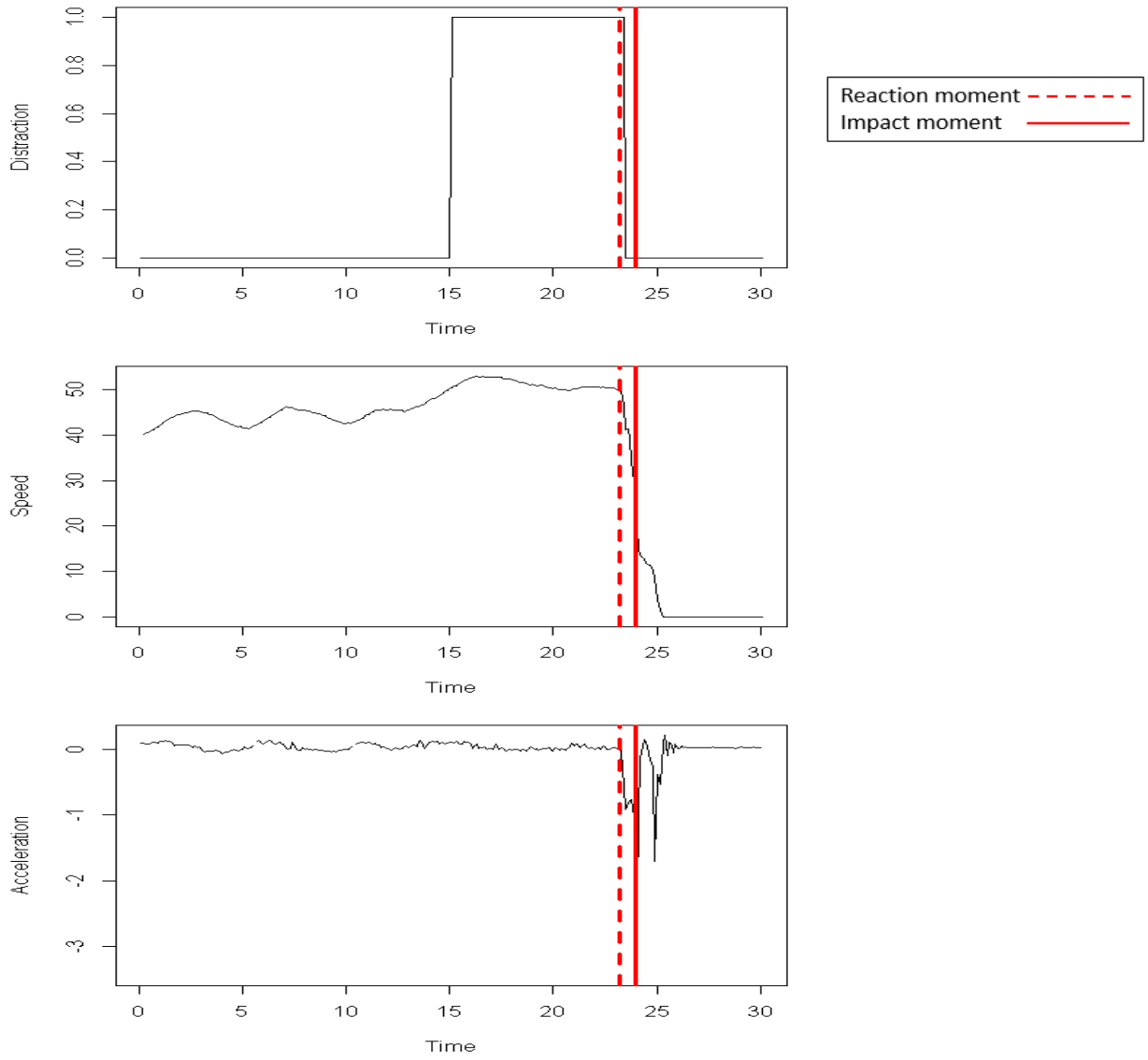


Figure 5.1 Speed and acceleration profile of a randomly chosen crash

Methodology

The main motivation of this study is to explore the association of the duration of distracted driving and impairment on the probability of extreme events. While significant literature exists on the investigation of correlation of distracted driving on crash risk and severity, the association of the duration of distracted driving on the probability of extreme events remains unknown. In order to untangle this problem, binary logistic regression approach

was utilized for modeling. This method is widely used in the literature in cases where the variable of interest has a binary nature (Dingus *et al.* 2016, Mokhtarimousavi *et al.* 2019, Nazari *et al.* 2019). Along with distraction duration, vehicular movements, driver behavior, roadway/environmental factors were considered as control variables. The study framework is shown in the Figure 5.2.

Upon linking the events with other factors, descriptive statistics are provided to gain initial insights. Next, fixed and random parameter binary logit models are developed to quantify the correlation of distracted duration on crash risk. In the following, more details on the modeling framework is provided.

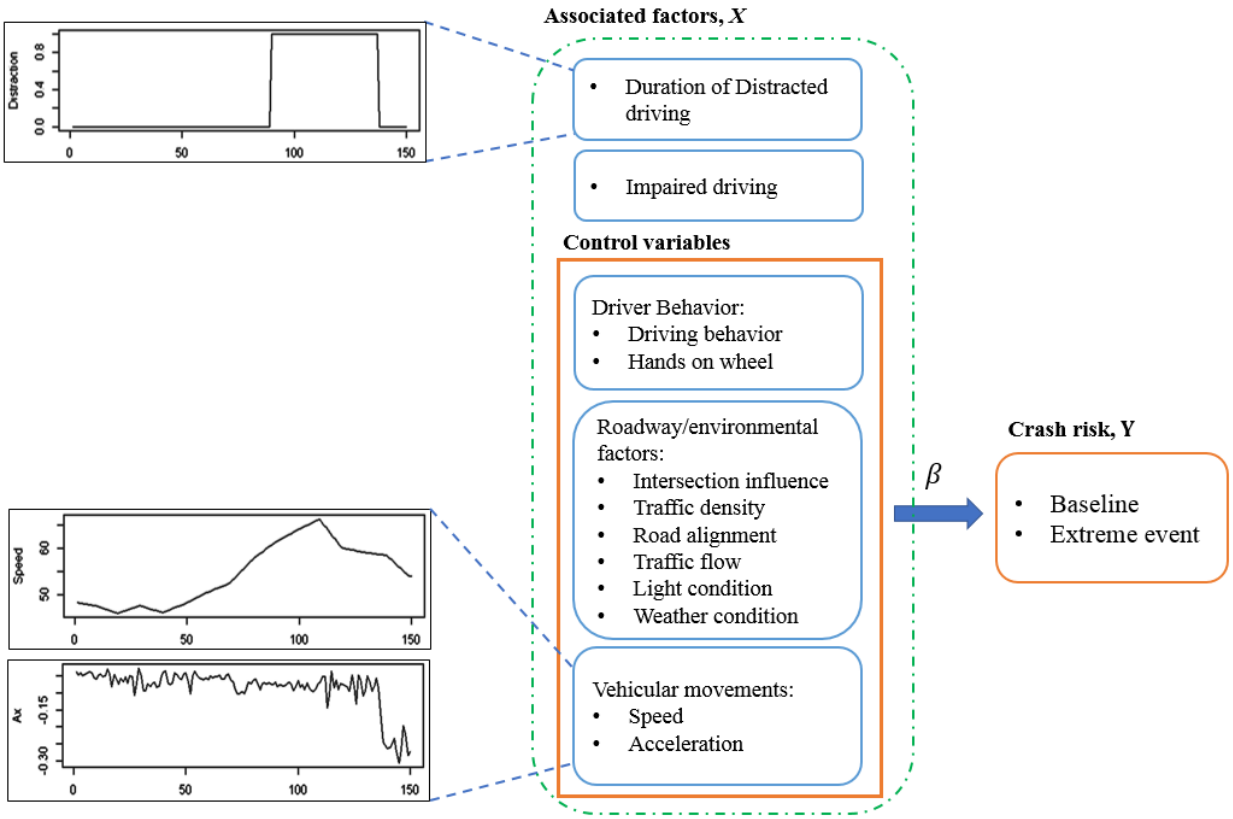


Figure 5.2 Conceptual framework of the study

Fixed parameter logit model

In the approach, the estimated parameters are fixed across the observations, and the estimations are not allowed to vary. Assuming that P_i is the probability of the occurrence of an extreme event in observation *i*, the association of the response variable and explanatory factors can be written as (Washington *et al.* 2010):

$$S_{in} = \beta_i X_{in} + \varepsilon_{in} \tag{5.1}$$

where β_i is the estimated coefficient for event i; X_{in} is the vector of independent variables; and ε_{in} is the error term following extreme value distribution. The probability of involvement in an extreme event can be written as (Washington *et al.* 2010):

$$logit(P_i) = log\left[\frac{P_i}{1 - P_i}\right] = \alpha + \beta X \tag{5.2}$$

where, P_i denotes the probability that event i is an extreme event; β is a vector of estimated parameters, X is a vector of independent variables; and, α is the model intercept. The likelihood can be written as (Washington *et al.* 2010):

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1 - \pi(x_i)\right)^{i - y_i} \tag{5.3}$$

where y_i is the outcome of observation *i*, and *n* is the number of observations. Accordingly, the log-likelihood function is:

$$LL(\beta) = \ln\left(l(\beta)\right) = \sum_{i=1}^{n}\left\{y_i \ln\left(\pi(x_i)\right) + (i - y_i)\ln\left(1 - \pi(x_i)\right)\right\} \tag{5.4}$$

In order to ease the understanding the association of explanatory variables on the probability of dependent variable, marginal effects are calculated. It can be defined as increase in the probability of occurrence of extreme event (*y=1*) by one-unit change in the variable of interest (*X*). We can write (Greene 2002):

$$\frac{\partial E[P(y_i)]}{\partial X_i} = \frac{dF(\beta' X)}{d(\beta' X)}\beta = F'(\beta X)\beta = f(\beta X)\beta \quad (5)$$

where $E[P(y_i)]$ is the expected value of the probability, $F(\beta' X)$ and $f(\beta X)$ are the density and probability functions of $E(y_i|X)$, respectively (Greene 2002).

Random parameter (mixed) logit model

As discussed, the fixed parameter model assumes that the variation of coefficients across the observations is fixed, which might not be the case. This issue must be addressed due to heterogeneity among events and drivers. In order to account for heterogeneity, random parameter models are widely used in the literature (Ukkusuri *et al.* 2011, Wali *et al.* 2017, Wali *et al.* 2018d, Azimi *et al.* 2019, Esfahani and Song 2019, Wali *et al.* 2019). This can be written as (Train 2009):

$$S_{in} = \beta_i X_{in} + \varepsilon_{in} + \eta_{in} \quad (5.4)$$

where η_{in} denotes the random term with pre-specified distribution and a mean of zero. Depending on the assumption of random term distribution, the outcome probability can be written as (Train 2009):

$$P_{in} = \int \frac{\exp(\beta_i X_{in})}{\sum_l \exp(\beta_i X_{in})} f(\beta|\varphi) d\beta \quad (5.5)$$

where $f(\beta|\varphi)$ is the density function of β, and φ is a parameter vector of density distribution also referred to mixing distribution (Washington *et al.* 2010). Several functional forms for the parameter density are assumed including normal, log-normal, uniform, Weibull, and triangular distributions. To evaluate and compare the developed models, Akaike Information Criteria (AIC) was utilized. It should be noted that a lower AIC denotes a model with a better fit to the data and a three-point reduction in AIC represents a significant improvement in the model fit (Bozdogan 1987).

Results

This section provides an in-depth analysis of the impact of distracted driving on the probability of crash occurrence and the role of impaired driving. First, the descriptive statistics of variables for the baseline and extreme events are provided and discussed. Next, the modeling results are provided. Finally, the impact of distraction on the probability of crash occurrence probability is described in detail in the discussion section.

Descriptive Statistics

Table 5.2 provides the descriptive statistics of the key variables. The table consists of three sections, driver related variables, roadway/environmental factors, and vehicular movements. The driver variables include distraction type, driver behavior, and impairment. The considered roadway/environmental factors including light and weather

condition, density of traffic, road alignment, construction zone, intersection influence, and roadway type are provided in the table. The results are also separated for the baseline and extreme events. Descriptive statistics for the baseline and extreme events can be observed to have a substantial difference, especially in terms of driver factors. This indicates that further analysis is needed to explore the association of these factors on the probability of being involved in a crash/near-crash event.

As shown in Table 5.2, there is substantial variation among the prevalence of secondary tasks when comparing extreme events to baseline events. As an illustration, the prevalence of cellphone use in extreme events is almost twice the usage observed in baseline events (15.06% vs 7.84%). A similar trend can be observed for distraction by objects inside the vehicle, where the drivers were nearly distracted two times than the baseline (7.80% vs 3.86%). Meaningful differences can be observed among baseline and extreme events for other types of distractions, emphasizing the importance of further investigation of distracted driving. Furthermore, driving impairments are generally associated with extreme events.

Similar trends can be observed for some categories of behavioral factors. For example, aggressive driving is substantially higher in extreme events compared to the baselines. This indicates a potential positive association between aggressive driving and the chance of involvement in an extreme event. It can be observed that the prevalence of aggressive behavior in extreme events (3.09%) is considerably greater than baselines (0.7%). Additionally, improper action, speeding, and traffic violation illustrate the same trend. Further details can be retrieved in Table 5.2.

Table 5.2 Descriptive statistics of the driver, vehicle, and roadway/environmental factors

Variable	Category	Total (N = 9239)		Baseline (N = 7394)		Extreme Event (N = 1845)	
		Perc.	Freq.	Perc.	Freq.	Perc.	Freq.
Distraction	Cellphone	9.29%	858	7.84%	580	15.06%	278
	Drink/Eat/Smoke	2.89%	267	2.88%	213	2.93%	54
	External	9.43%	871	9.44%	698	9.38%	173
	Interaction	12.88%	1190	13.23%	978	11.49%	212
	None	46.76%	4321	49.24%	3641	36.86%	680
	Object Distraction	4.64%	430	3.86%	286	7.80%	144
	Other	5.01%	463	4.6%	338	14.85%	274
	Talking/singing	5.88%	544	5.88%	435	5.91%	109
	Body Related	3.19%	295	3.04%	225	3.79%	70
Driving Behavior	Aggressive	0.70%	65	0.11%	8	3.09	57
	Drowsy, sleepy	1.27%	118	1.26%	93	1.35	25
	Improper Action	4.96%	459	2.7%	200	14.03	259
	Low Speed	0.95%	88	1.13%	84	0.21	4
	None	85.32%	7885	89.99%	6654	66.72	1231
	Other	0.63%	58	0.39%	29	1.57	29
	Speeding	4.20%	388	3.03%	224	8.89	164
	Traffic violation	1.93%	178	1.38%	102	4.12	76
Impairment	Emotional state	0.50	46	0.28	21	1.36	25
	Drowsy/Fatigue	1.40	129	1.23	91	2.06	38
	Alcohol/Drug	0.24	22	0.05	4	0.98	18
	No impairment	97.60	9016	98.24	7264	94.96	1752
	Other	0.28	26	0.19	14	0.65	12
Light	Darkness, lighted	13.77%	1272	13.01%	962	16.80%	310
	Darkness, not lighted	5.57%	515	6.10%	451	3.47%	64
	Dawn/Dusk	4.54%	419	4.76%	352	3.63%	67
	Daylight	76.12%	7033	76.12%	5629	76.10%	1404
Weather	Adverse Conditions	6.13%	567	5.91%	437	7.05%	130
	Mist/Light Rain	4.09%	378	3.85%	285	5.04%	93
	No Adverse Conditions	89.77%	8294	90.24%	6,672	87.91%	1622
Density (Level-of-service)	A1	40.23%	3717	42.51%	3,143	31.11%	574
	A2	30.15%	2786	32.31%	2,389	21.52%	397
	B	20.16%	1863	18.49%	1,367	26.88%	496
	C	6.07%	561	4.56%	337	12.14%	224
	D	2.10%	194	1.27%	94	5.42%	100
	E	1.02%	94	0.72%	53	2.22%	41
	F	0.25%	23	0.14%	10	0.70%	13
	Unknown	0.01%	1	0.01%	1	0.0%	0
Road Alignment	Curve	13.60%	1256	13.97%	1034	12.03%	222
	Straight	86.40%	7983	86.03%	6,360	87.97%	1623

Duration of distracted driving

As discussed, this study utilized a unique method to investigate the effect of the duration of distracted driving on the probability of crash occurrence by analyzing the time that drivers were disengaged from driving and performing tasks other than driving. While section 5.1 presents descriptive statistics on the prevalence and presence of distraction among baseline and extreme events (whether it was present or not), the correlation of each distraction duration with the resulting crash risk is discussed here. Table 5.3 provides the duration of distracted driving for each distraction category for both baseline and extreme events. Comparing the two groups, there is a significant difference between the duration of distraction in extreme events compared to baseline events. When considering overall distraction by disregarding the distraction type, drivers were distracted on average for 1.85 seconds within baseline events, while in the extreme events the distraction duration was 3.12 seconds. This time difference implies that the prevalence of distraction is higher, and the duration of the distraction is longer in extreme events. Focusing on the distraction types, a similar pattern can be observed in all the distraction types.

Drivers were distracted by cellphones for 0.37 seconds on average, with an average duration of 1 second in extreme events. Distraction by objects inside the vehicle follows a similar pattern, indicating that on average drivers were distracted for longer compared to baselines (0.26 vs 0.13 second). The duration of interaction with other passengers is slightly higher in extreme events. Furthermore, distraction duration of the category “atypical” is substantially higher in extreme events compared to baseline events (0.19 vs

0.06 seconds). These statistics suggest that there is a significant correlation between the duration of distracted driving and the risk of a crash. Statistical modeling will provide more insights on the significance of these variables and their association with near-crash and crash risks, which will be discussed in the next section.

In order to shed more light on the duration of distraction among drivers, the histograms for key distraction types are provided in Figure 5.3 below. It can be observed that there is a significant difference in distraction duration among different duration types. Cellphone use appears to have normal distribution, and the histogram of duration of external distraction is right skewed. However, object and atypical types of distractions can be observed to have a bimodal distribution.

Table 5.3 Descriptive statistics of the duration of distraction for 15 seconds of observations

Variable	Baseline (N=7394)				Extreme event (N = 1845)			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Total duration of distraction	1.85	2.2	0	14	3.12	3.27	0	13.9
Body Related	0.11	0.66	0	5	0.17	0.99	0	9.5
Cellphone	0.37	1.29	0	5	1	2.5	0	13.9
Drink/Eat/Smoke	0.13	0.78	0	5	0.19	1.17	0	13
External	0.18	0.72	0	8.9	0.27	1.1	0	13
Interaction	0.58	1.55	0	14	0.64	1.99	0	13.5
Object Distraction	0.13	0.72	0	5	0.26	1.22	0	11.6
Other	0.11	0.61	0	5	0.27	1.27	0	12.4
Talking/singing	0.24	1.03	0	10.6	0.31	1.36	0	11.9

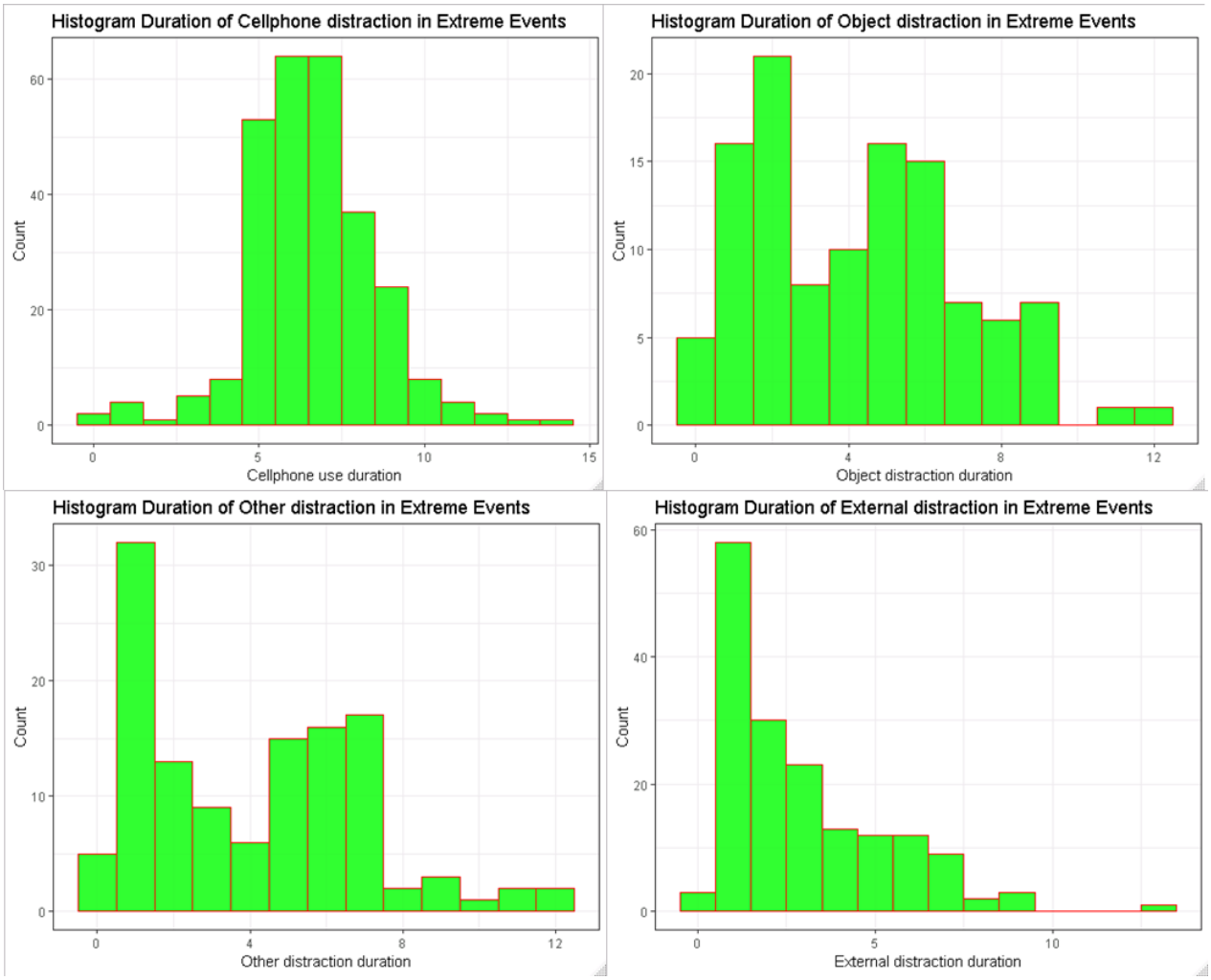


Figure 5.3 Histogram of duration of key distraction types for extreme events

Modeling Results

The descriptive statistics of the data presented in the previous section revealed meaningful relationships between duration of distraction and crash risk. However, without controlling for other factors such as driving behavior and roadway/environmental factors, these relations might not be generalizable and conclusive. As discussed in the methodology section, this study utilized a fixed and a random parameter binary logistic regression model to explore the association of the duration of distracted driving with the

probability of crash occurrence. The random parameter model addresses unobserved heterogeneity, and a parameter is considered to be random in two different conditions: first, only standard deviation is significant; second, both mean and standard deviation are significant. Along with duration of distraction and impaired driving factors, driver behavior and roadway environmental variables are considered in the model as the control variables. To perform the model selection, intuition, variable significance, and model parsimony were considered and AIC was used to score model performance. The modeling results for the fixed and random parameter models are provided in Table 5.4. Also, the measured marginal effect is provided in the table for the fixed and random parameter models to ease the understanding the effect of each variable on the probability of an extreme event. The marginal effect can be defined as the effect of one unit increase in a factor on the probability of occurrence of an extreme event, with all other factors controlled at their mean values.

According to the model summary, the random parameter model outperformed the fixed parameter model in terms of AIC, and McFadden R-Square statistics by capturing heterogeneity among observations. The McFadden R-square value for the random-parameter model is 0.241, which is an acceptable value considering the sample size. All variables in the model are significant at the 90 percent confidence interval. Due to better fit of the random parameter model, only the results of the random parameter model are discussed.

Table 5.4 Fixed and random parameter modeling results

Variable	Fixed parameter				Random parameter			
	β	Std. Err.	p-value	ME (%)	β	Std. Err.	p-value	ME (%)
Intercept	-1.628	0.081	<0.001	-	-1.121	0.058	<0.001	-
Duration of distraction								
Body related	0.239	0.036	<0.001	2.80	0.185	0.027	<0.001	2.88
Cellphone	0.275	0.017	<0.001	3.23	0.219	0.013	<0.001	3.42
Eating/Drinking/Smoking	0.172	0.032	<0.001	2.02	0.130	0.023	<0.001	2.02
External (e.g. looking outside)	0.256	0.033	<0.001	3.01	0.198	0.025	<0.001	3.08
Interaction (with passengers)	0.147	0.018	<0.001	1.73	0.119	0.013	<0.001	1.85
Object distraction (e.g. GPS, climate control, audio control)	0.242	0.032	<0.001	2.83	0.191	0.027	<0.001	2.98
Singing/talking	0.163	0.026	<0.001	1.91	0.126	0.019	<0.001	1.97
Other (e.g. reading, writing, insect in the vehicle	0.329	0.034	<0.001	3.85	0.263	0.027	<0.001	4.11
Driver impairment								
Emotional driving (Angry, sadness, etc.)	0.772	0.375	0.040	10.49	0.555	0.302	0.066	8.65
Drowsy, Fatigue	1.451	0.500	0.004	21.69	1.411	0.477	0.003	22.02
Alcohol/Drugs	2.445	0.640	0.000	40.09	1.906	0.439	0.000	29.74
Other	1.138	0.470	0.016	16.37	0.927	0.361	0.010	14.47
Driving behavior								
Aggressive driving	3.589	0.405	0.000	59.08	2.854	0.273	0.000	44.53
Drowsy or fatigued	-0.436	0.541	0.420	-4.67	-0.661	0.506	0.192	-10.3
Improper action	1.918	0.111	0.000	30.55	1.416	0.080	0.000	22.09
Low speed driving	-1.414	0.533	0.008	-12.12	-1.005	0.346	0.004	-15.7
Other	1.772	0.306	0.000	27.60	1.293	0.177	0.000	20.18
Speeding	2.466	0.129	0.000	40.02	1.819	0.110	0.000	28.38
Std. Speeding				-	1.786	0.181	0.000	-
Violation	1.260	0.170	<0.001	18.38	0.588	0.193	0.002	9.17
Std. Violation				-	3.868	0.535	0.000	-
Weather condition								
Adverse condition	0.148	0.122	0.224	1.79	0.126	0.093	0.175	1.96
Rain	0.258	0.143	0.071	3.18	0.173	0.109	0.111	2.71
Traffic density (base:								
LOS A2	0.387	0.083	<0.001	4.68	0.305	0.062	0.000	4.75
LOS B	1.193	0.083	<0.001	15.92	0.891	0.063	0.000	13.91
LOS C	1.664	0.115	<0.001	25.22	1.220	0.088	0.000	19.04
Std LOS C	-	-	-	-	0.624	0.112	0.000	-
LOS D	1.954	0.171	<0.001	30.96	1.456	0.154	0.000	22.72
Std LOS D	-	-	-	-	2.505	0.322	0.000	-
LOS E	1.272	0.241	<0.001	18.63	0.790	0.209	0.000	12.32
Std LOS E	-	-	-	-	2.020	0.384	0.000	-
LOS F	1.456	0.474	0.002	21.89	0.975	0.376	0.010	15.21
Vehicular movement								
Average Speed over 15 sec	-0.025	0.001	0.000	-0.29	-0.022	0.001	<0.001	-0.34
Std Speed	-	-	-	-	0.010	0.001	<0.001	-
Model Summary								
Number of observations	9239				9239			
Null Deviance	-3494.21				-3480.8			
Model Deviance	-4619.3				-4619.3			
McFadden R Square	0.243				0.246			
AIC	7046.4				7031.7			

Discussion

Distracted driving

While several studies have explored the correlation of distracted driving and crash risk by analyzing the presence of distraction on the crash (Dingus *et al.* 2016, Gao and Davis 2017, Arvin *et al.* 2019b), this study considers the duration of distraction by different types of distractions to explore their association with crash risk. The results suggest that duration of all types of distracted driving are positively and significantly associated with the probability of the occurrence of an extreme event (i.e. near-crash and crash events). Figure 5.4 provides the plot of the probability of an extreme event with increasing duration of distraction for all types of distraction. It can be observed that atypical distraction types (such as reading, writing, and distraction by insect) and cellphone related distraction has the highest impact on extreme event occurrence probability.

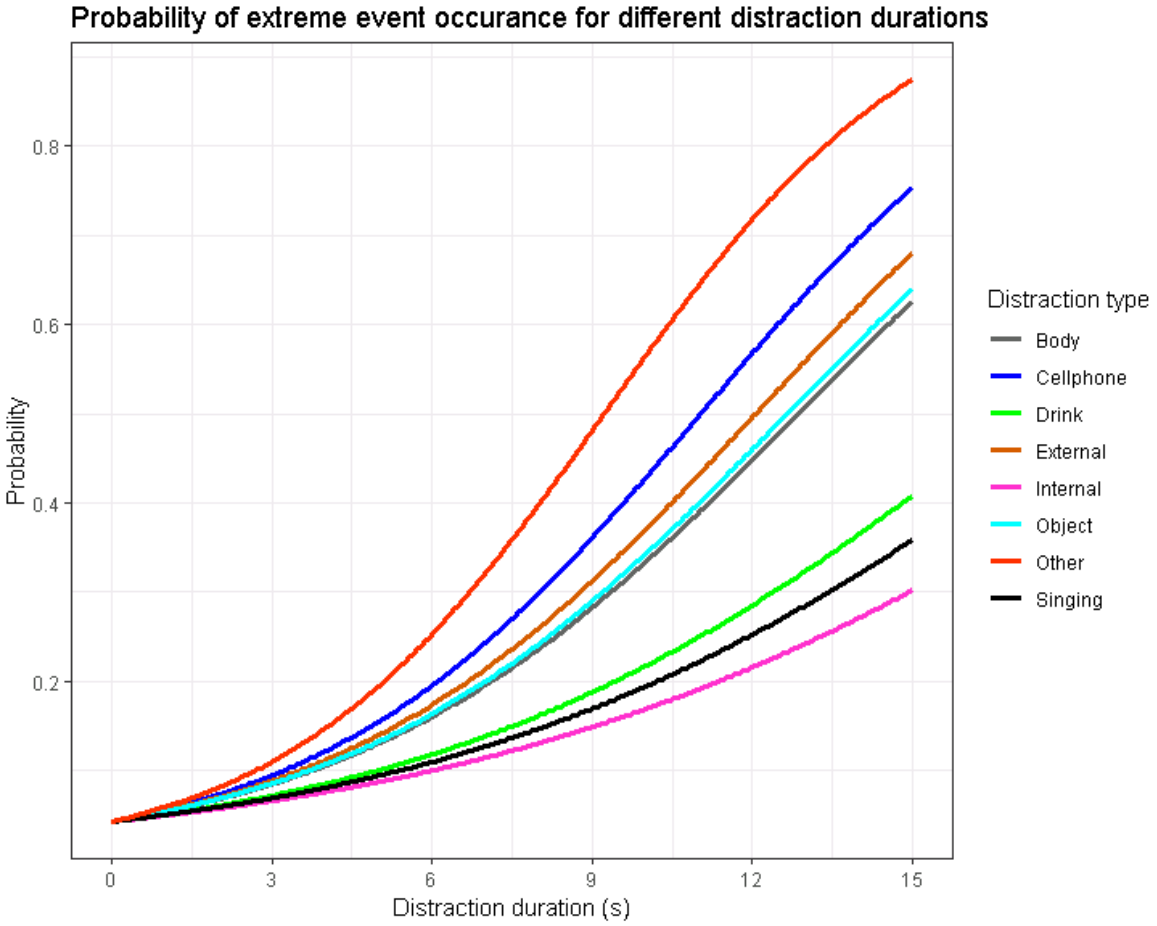


Figure 5.4 Probability of extreme event occurrence for different types of distraction

Duration of other types of distraction has the highest impact on the probability of crash occurrence. The marginal effect results indicate that, keeping other variables at their mean, one second increase in the duration of other distraction, increase the probability of a crash for 4.21 percent, on average. The reason might be this category contains unusual distraction types such as insect in the vehicle, reading, writing, etc. Therefore, it is expected that these types of distraction have higher impact compared to other types of distraction. As literature suggests (Dingus *et al.* 2011, Dingus *et al.* 2016, Kamrani *et al.* 2019), distracted driving is highly correlated with the crash risk. The results revealed that

duration of cellphone use is substantially and significantly increasing the probability of extreme event involvement. One second increase in the distracted driving with cellphone increased the likelihood of crash/near-crash for 3.51 percent. Focusing on duration of distraction by external stimulus, it is significantly associated with the extreme event occurrence. One second increase in the duration of external distraction, increase the probability of extreme event involvement for 3.12 percent. Seconds of distraction by objects inside the vehicle (such as navigation system, climate control, radio adjusting) is significantly associated with the probability of crash involvement. Marginal effect analysis indicates that one second increase in the duration of object distraction increase the probability of an extreme event for 3.03 percent. In line with the literature (Dingus *et al.* 2011),Eating, drinking, and smoking in the car is increasing the probability of crash/near-crash involvement in a manner that one second increase in such distractions, increases the extreme event probability for 2.15 percent. Singing and talking with him/herself and interaction with other passengers also increase the likelihood of extreme event involvement. One second increase in the duration of singing/talking and interaction with others is associated with increase in the extreme event probability for 2.04 and 1.92 percent, respectively. Duration of interaction with passengers has less negative effect on driving performance since the responsibility of monitoring environment could be shared with passengers (Overton *et al.* 2015).

Driver impairment and behavior

The results of modeling reveal that all types of impairment increase the likelihood of extreme events. Specifically, alcohol/drug related impairments are associated with a 29.7 percent increase in the probability of crash/near-crash involvement. The results are

consistent with the findings of Dingus et al (Dingus *et al.* 2016) who found that alcohol and drug impairment increases the crash risk 35.9 times. Furthermore, drowsy and fatigued driving are associated with increased probability of extreme event by 22 percent, which is in line with the literature (Klauer *et al.* 2006, Lee *et al.* 2016). In line with previous studies (Dingus *et al.* 2016), emotional driving increased the probability of involvement in an extreme event by 8.65 percent. Atypical impairment types are associated with 14.5 percent higher crash risk.

The results revealed that driving behavior is substantially and significantly correlated with the likelihood of extreme event occurrence. As an illustration, in line with the literature (Zhang and Chan 2016), aggressive driving was found to increase the likelihood of extreme event by 44.53 percent. Speeding behavior is significantly correlated with the likelihood of crash involvement, by increasing its likelihood by 28.4 percent. Improper driving action increased the likelihood of crash involvement by 22.1 percent. Violation of traffic law is another significant driver behavior that is positively and significantly associated with crash risk. Traffic violation is correlated with an increase in the likelihood of extreme event occurrence for 9.2 percent.

Roadway/environmental factors

Roadway and environmental factors are included in the model to control for other contributing factors. The modeling results suggest that higher traffic density in terms of level of service increase the likelihood of crash involvement. The results are in line with previous studies where the chance of a crash or near crash in congested traffic is higher compared to the free-flow state (Kamrani *et al.* 2019). Furthermore, driving with a higher

speed decreases the likelihood of an extreme event, since the vehicle has lower conflict with other vehicles and surrounding environment which decreases the probability of crash involvement. Moreover, while the fixed-parameter model suggests that weather condition is significantly associated with crash risk, the random parameter model suggests that it is marginally correlated with the increase in the likelihood of extreme event. Further details are provided in the Table 5.4.

Limitations

The drivers participating in SHRP2 NDS might not represent the driving population, since they are voluntarily hired with monetary incentives, i.e., they are self-selected. Although the data are collected professionally with federal support and specific protocols are used for data collection and coding, there still might be some human error in coding the information, especially from the recorded videos. The proportion of crashes and near crashes compared to baselines are not truly reflective of real-world conditions, as extreme events are relatively rare, and this fact might affect the results.

Conclusions

Generally, human error is known to be the key contributing factor in traffic crashes. Availability of microscopic information collected through instrumented vehicles on instantaneous driving behavior and instantaneous decisions of drivers has enabled the exploration of the association of driver behavior with the occurrence of crashes and near-crash events. This study sheds light on the association of distracted driving on the

probability of a crash/near-crash by performing an in-depth analysis on pre-crash driving that lead to extreme event involvement. The main contributions of this paper are, first, linking large-scale data on instantaneous driver distraction and vehicular movements with the driving behavior, roadway, and environmental factors, and, second, using rigorous methods for exploring the association of impairments and duration of distracted driving by different secondary tasks on the likelihood of involvement in an extreme event—a topic that is very lightly investigated in the literature. The SHRP2 NDS data is used, containing 9239 baselines and extreme events, in terms of crashes and near-crashes. A unique database was created by analyzing more than 1.8 million observations and creating time-series profile of distracted driving, and linking it to the vehicle kinematics, driving behavior, and roadway/environmental factors. The seconds that drivers were reacting to the crash stimuli and the period after a crash were removed from the analysis. In this research, 15 seconds of the data was considered for each event.

The descriptive analysis shows that there is a substantial difference in prevalence of impaired and distracted driving between baselines and extreme events, as expected. Moreover, the analysis of duration of distraction revealed that the duration of distraction is also significantly different among these two groups. The fixed and random parameter binary logistic regression model is estimated to model the association of distraction duration on the probability of the extreme event occurrence. The modeling results revealed that duration of all types of distractions is one of the leading indicators of an extreme event occurrence and longer durations significantly and substantially increase the crash risk. Based on the results, cellphone distraction and atypical distraction types

(in terms of reading and writing) have the highest association with crash risk compared to other distraction types. One second increase in an atypical and cellphone distraction will increase chance of a crash for 4.1 and 3.4 percent, respectively. In terms of impaired driving, alcohol/drugs substantially increase the chances of extreme event involvement by 29.7 percent. It is worth noting that driving behavior and roadway/environmental factors are also modeled as the controlling factors. Overall, the results point to exploring and evaluating countermeasures that can reduce the most dangerous types of impaired and distracted driving.

CHAPTER 6 : REAL-TIME CRASH PREDICTION THROUGH UNIFIED ANALYSIS OF DRIVER AND VEHICLE VOLATILITIES: APPLICATION OF 1D-CONVOLUTIONAL NEURAL NETWORK - LONG SHORT-TERM MEMORY

This chapter is a modified version of a research article by Ramin Arvin, Asad J. Khattak, and Hairong Qi. *"Crash prediction through unified analysis of driver and vehicle volatilities: Application of 1D-Convolutional Neural Network - Long Short-Term Memory."* The manuscript is currently under review in Engineering Application of Artificial Intelligence.

Abstract

Transportation safety is highly correlated with driving behavior, especially human error playing a key role in a large portion of crashes. Modern instrumentation and computational resources allow for the monitorization of driver, vehicle, and roadway/environment to extract leading indicators of crashes from multi-dimensional data streams. To quantify variations that are beyond normal in driver behavior and vehicle kinematics, the concept of volatility is applied. The study measures driver-vehicle volatilities using the Naturalistic Driving Study (NDS) data. By integrating and fusing multiple data streams, i.e., driver distraction, vehicle kinematics, and driving stability in real-time, this study aims to generate useful feedback to drivers and warnings to surrounding vehicles regarding hazards. The NDS data is used which contains detailed information on more than 3500 drivers (7589 normal driving events, and 2004 severe events i.e., crash and near-crash) in addition to vehicle kinematics and driver behavior. In order to capture the local dependency and volatility in time-series data 1D-Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and 1DCNN-LSTM are applied. Vehicle kinematics, driving volatility, and impaired driving (in terms of distraction) are used as the input parameters. The results reveal that the 1DCNN-LSTM model provides the best performance, with 92.36% accuracy and prediction of 71% of

crashes with a precision of 93%. Additional features are extracted with the CNN layers and temporal dependency between observations is addressed. The model can be used to monitor driving behavior in real-time and provide warnings and alerts to drivers in low-level automated vehicles, reducing their crash risk.

Introduction

In 2016, around 7.27 million vehicle crashes are recorded across the United States which leads to 37,914 fatalities and more than 2.17 million injuries (Administration 2018), while human-error is the leading cause of crashes with contribution in 94 % (Anon 2008). Although occurrence of a crash in an outcome of several factors, these statistics suggests that researchers need to provide a great attention to the human behavioral side of crashes, while the literature mainly focuses on the roadway and infrastructure factors. On the other hand, conventional data sources including police-reported crashes, are the major source of the literature, which suffers from the under-reported crashes. Based on the report by National Highway Traffic Safety Administration (NHTSA) (NHTSA 2009), 50% of property damage only crashes and 25% of minor injury crashes are not reported to the police and not recorded. Also, these crashes may be truncated due to states monetary threshold (Hauer 2006). Given all these limitations, the police-reported data has limited information on the pre-crash events, vehicular movements, driver state and decision, and maneuvers.

By the emergence of new sources for data collection, high-resolution naturalistic driving data is emerged which provides a great opportunity to investigate in-depth crash analysis

by incorporating microscopic driving performance and behavior prior to crash involvement. Furthermore, these datasets contain detail information not only on PDO and minor crashes, but also near-crash events where critical event happened but did not lead to a crash. This information can be coupled with driver behavior and vehicular movements and help us for real-time prediction of occurrence of crash or near-crash in order to potentially prevent their occurrence. In this context, since the prediction accuracy is vital, our desirable is a model that can accurately predict crash risk before its occurrence.

Referring to the methodological standpoint, several studies attempted to study the correlates of pre-crash driving behavior and roadway/environmental factors on the crash risk and severity using traditional statistical approach. While these methods are very beneficial by providing insights regarding the association of factors, usually they suffer from low accuracy on the out of sample data and prediction. Therefore, in the context of real-time warning generation for the crash risk prediction, other supervised methods are needed to perform better in terms of accuracy. Furthermore, due to high dimensionality of data, traditional statistical methods might not be appropriate in this context.

Deep learning methods have gained significant attention in the recent literature due to their promising performance in several fields. In this context, the convolutional neural networks (CNNs) (Hinton and Salakhutdinov 2006) and Recurrent Neural Networks (RNNs) (add reference) are mainly utilized to process visual-related and time-series problems. With the recent improvements in the CNN (LeCun *et al.* 1998, Simard *et al.* 2003, Ciresan *et al.* 2011) and RNN (add reference), and emergence of large-scale data integrated with efficient implementation of computational powers (i.e. graphics processing

units (GPUs)), they outperformed not only conventional methods but also human performance (Sermanet and LeCun 2011).

In this study, the main contribution is developing a deep learning framework to integrate multiple data streams including vehicular kinematics in terms of speed, longitudinal and lateral accelerations, driving stability, and driver behavior to predict the occurrence of a crash/near-crash. The developed framework has several advantages:

1- The architecture configuration of the model is compact, making the model easy to be implemented for real-time safety performance monitoring and failure detection.
2- Its ability to capture temporal variations in the input data generated from multiple sensors.
3- The capability of the model to efficiently train the model using limited training dataset and back-propagation iterations (Eren *et al.* 2019).

With the emergence of new data sources, this study is timely and original by harnessing this big data and incorporate it in the instantaneous driving behavior analysis by developing a deep learning framework to warn driver regarding the risk of crash involvement. The compact configuration of the developed model help agencies to easily implement it in real time applications.

Data description and pre-processing

Data description

This study utilized the second Strategic Highway Research Program (SHRP 2) naturalistic driving study data. The original data contains more than 4 petabytes of information, which known as the most comprehensive driving study. The data collection is performed from 2010 to 2013 and contains high-quality and high-resolution data of more than 3500 drivers travelling more than 50 million vehicles miles and 5 million trips from six states in the United States including Washington, New York, Pennsylvania, Florida, North Carolina, and Indiana (Hankey *et al.* 2016).

For the data collection, onboard data acquisition system (DAS) along various sensors (camera, alcohol sensor, forward sensor, accelerometers) are used to record information including vehicular movements (speed, acceleration, steering position) at 10-Hz frequency, video views, vehicle controls, offset from lane center, etc. (Hankey *et al.* 2016). The dataset that we used in this study is a subset of SHRP2 NDS data containing 7566 baseline, 1307 near-crash and 617 crashes. For each crash and near-crash-involved trip, 30 seconds of vehicle kinematics is available. The data contains evasive maneuver seconds (taken by the driver to avoid the crash or near-crash) and after its occurrence.

Since this paper is predicting the critical event before its occurrence, we need to only include unintentional driving decisions and exclude intentional behavior arising from drivers' behavior to avoid these events. Thus, these seconds need to be removed from our analysis (which will be discussed in the next section). Furthermore, since we are using

the information on driver distraction, we need to extract the seconds that driver was distracted, which was obtained from the summary of each trip. In the dataset, crashes are defined as “any contact that the subject vehicle has with an object, either moving or fixed, at any speed in which kinetic energy is measurably transferred or dissipated”. This study combines the crashes, and near-crashes events in a sense that both events are critical, while near-crashes was eventually become a crash due to appropriate response of the drivers to avoid collision at the last second.

Exclusion of Evasive Maneuvers

While the goal of this study is real-time critical event detection using vehicular movements and distraction information, it is crucial to exclude the seconds of vehicle trajectories that drivers are attempting to avoid crashes. To elaborate more, speed and acceleration profiles of a randomly chosen crash are provided in the Figure 6.1 in which the crash happened at 23^{th} second of the video, while the driver started to react to the situation at 22^{th} second of the data. Thus, we need to not only exclude the seconds after the crash occurrence but also exclude the seconds that driver is reacting to the stimuli. In other words, only the seconds of the data up to the moment that the drivers started to react is used in this study. The speed, longitudinal and lateral acceleration of 15 seconds before the reaction time are used to calculate measures of driving volatility (which will be discussed in the next section) which help us to quantify driving instability. It is worth noting that there are crashes in which there is no reaction by driver, or the reaction happened after the impact time. Thus, either the reaction time or impact time is used.

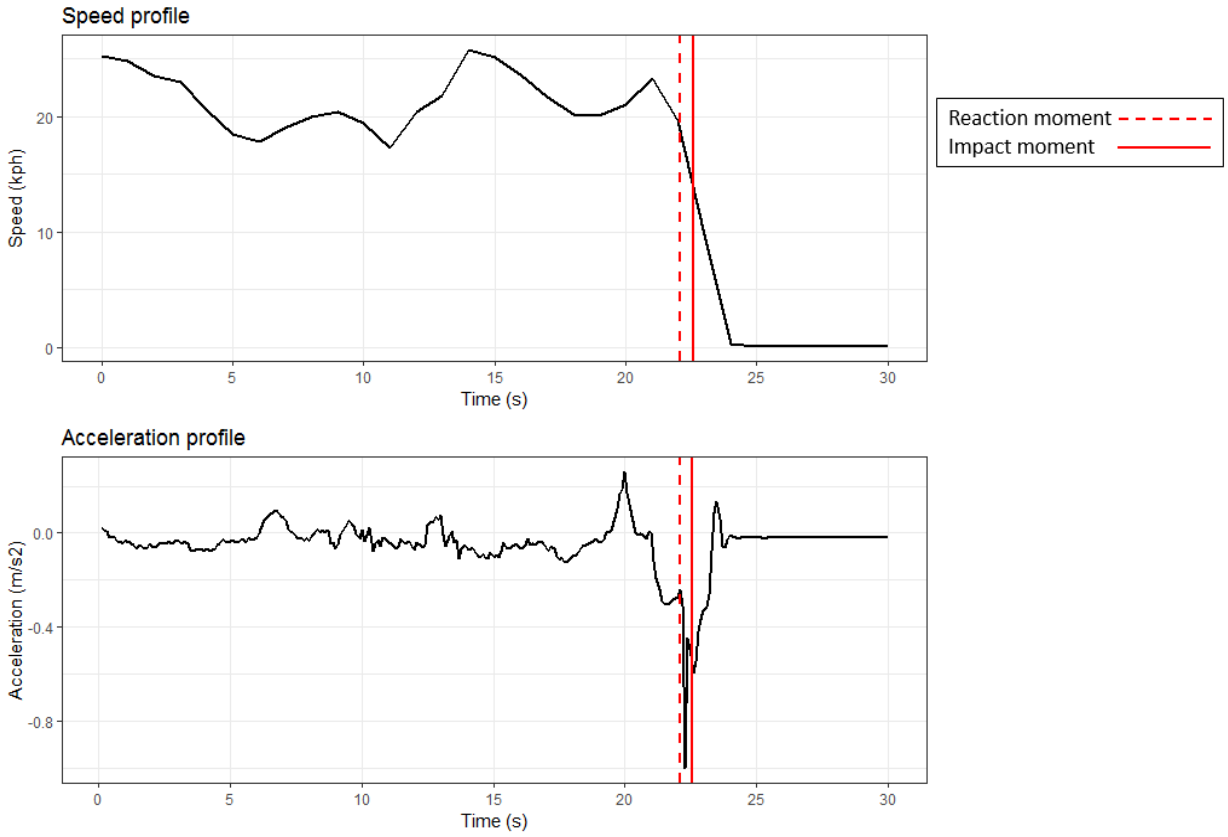


Figure 6.1 Speed and acceleration profile of a randomly chosen crash

Measures of driving volatility

The concept of driving volatility is introduced to extract useful information and features from microscopic vehicular kinematics to quantify variations in instantaneous driving decisions. I the literature, several functions are proposed to quantify these variations and applied to vehicle speed (Kamrani *et al.* 2018b, Arvin *et al.* 2019c, b, Kamrani *et al.* 2019, Hoseinzadeh *et al.* 2020), longitudinal acceleration (Kamrani *et al.* 2018b, Arvin *et al.* 2019c, b, Kamrani *et al.* 2019), and lateral acceleration (Arvin *et al.* 2019c). This study applied several volatility functions to extract additional features from the data. in general, three volatility groups are extracted:

1- Speed volatility

2- Longitudinal acceleration volatility

3- Lateral acceleration volatility

In the following, the mathematical formulation of each volatility function is provided and discussed in detail.

Exponentially Weighted Moving Average Volatility (EWMA): This measure was introduced by RiskMetrics in 1996 (Longerstaey and Spencer 1996) which considers volatility as a weighted average of volatility observations over time. We can write (Longerstaey and Spencer 1996):

$$EWMA = \sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)\epsilon_{t-1}^2 \tag{6.1}$$

where ϵ_{t-1} is the return at time *t-1*, σ_{t-1}^2 is the EWMA volatility at time *t-1* and lambda is user defined weight (assumed 0.94 in this paper).

Time-varying stochastic volatility: which quantify dispersion in the vehicular movements by considering changes in the ratio of observations. We can write (Figlewski 1994):

$$V_f = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})} \qquad \text{from } t = 1 \text{ to } n \tag{6.2}$$

where V_f denotes the time-varying stochastic volatility, *n* is number of observations, and r_i is:

$$r_i = \ln\left(\frac{x_t}{x_{t-1}}\right) \tag{6.3}$$

where *ln* is natural logarithm, x_t and x_{t-1} are the observations at t and $t-1$, respectively. Since this measure requires positive time-series input, it only applied to vehicle speed.

Mean absolute deviation: which quantifies variations in the data by measuring the distance between observations and their central tendency (mean in this paper). We can write (Huber 2005):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{6.4}$$

Standard Deviation: which is the most common and basic approach to quantify dispersion in the data. We can write:

$$S_{dev} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.5}$$

where x_i and $\bar{x}$ denotes the observed value *i* and the mean of observations, and *n* is the total number of observations.

Next, each volatility measure is calculated at two levels: Level 1: Event-based volatility, Level 2: Temporal driving volatility. Event-based volatility applies the functions on 150

deci-second observations and returns one value as a volatility measure. On the other hand, temporal volatility utilizes the concept of moving average and applies 3-seconds (30 deci-seconds) time-window to calculate temporal volatility for each second of driving. Thus, the output will be temporal volatility. As an illustration, Figure 6.2 provides the L2-Speed-Vf volatility measure for one of the critical events.

Concept illustration and descriptive statistics

This section provides some statistical analysis in order to illustrate the positive association of driving volatility and distracted driving on the crash risk. The previous sections have discussed the procedure for calculating event-based and temporal driving volatility for speed, longitudinal and lateral acceleration. Here, the association of the driving volatility on the crash risk is shown using boxplot analysis. The results are provided in Figure 6.3. It can be observed that there is a substantial difference in these two groups. In critical

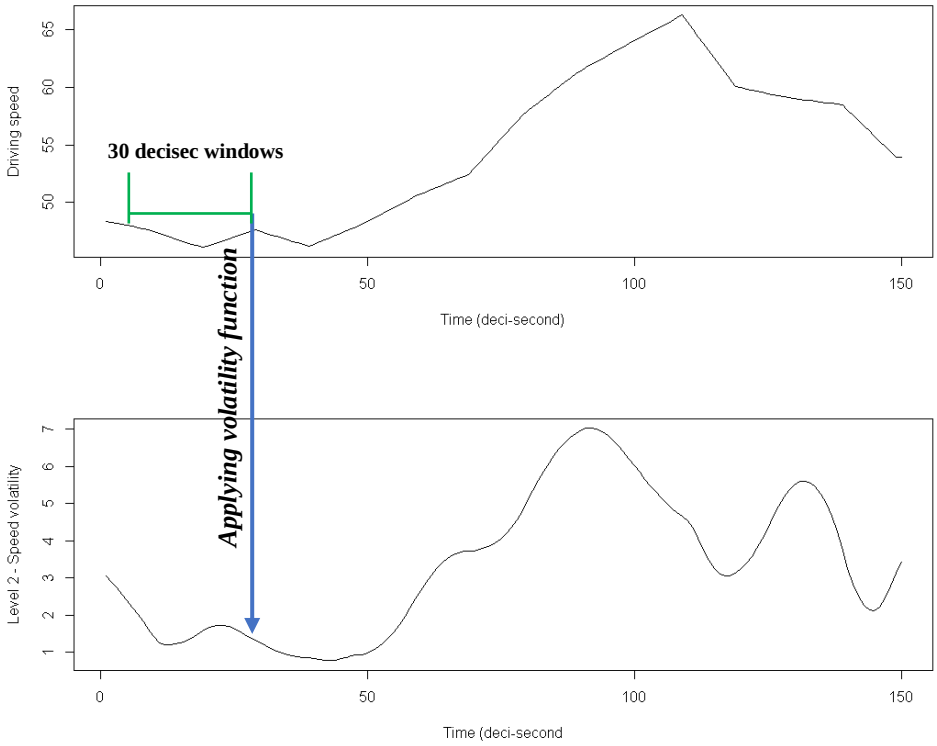


Figure 6.2 Temporal speed volatility measure calculation

events, drivers were more distracted and volatile compared to the baseline events. Therefore, the question that might arise is whether this information can be used to predict the chance of a crash/near-crash before its occurrence.

As discussed, there is a substantial difference between the baseline and critical events. This section provides descriptive statistics of the variables used as the feature in the

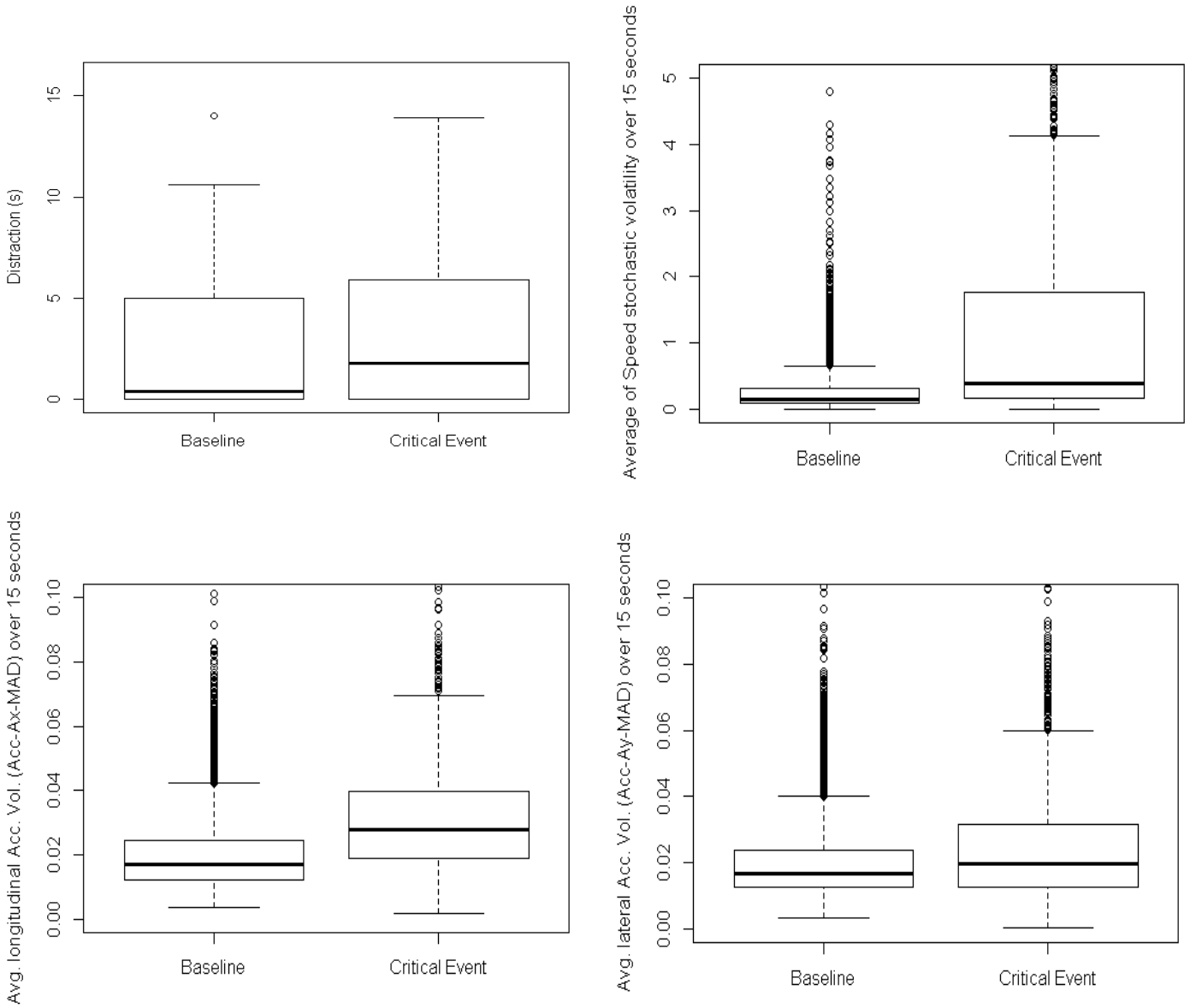


Figure 6.3 Boxplot of distracted driving, speed, longitudinal and lateral volatilities for the baseline and critical events

models (Table 6.1). The feature space contains information on three dimensions of vehicular movements, seconds that driver was distracted with a secondary task, event-based and temporal driving volatility indices for speed, longitudinal, and lateral accelerations. It can be observed that not only seconds of distraction but also driving volatility is significantly higher in the critical events compared with baselines.

Table 6.1 Descriptive statistics of the baseline and critical events

	Baseline events (N=7566)				Critical events (N=1925)			
Variable (feature)	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Speed (mph)	62.36	31.22	0	125.81	41.23	30.12	0	116.74
$Acceleration_x\ (m/s^2)$	-0.01	0.04	-0.23	0.25	-0.01	0.06	-0.87	0.26
$Acceleration_x\ (m/s^2)$	0	0.04	-0.2	0.33	0	0.04	-0.2	0.24
Seconds of distraction	1.852	2.19	0	14.00	3.11	3.26	0	13.90
L1-Speed-S_{dev}	1.51	1.46	0	31.88	2.2	1.76	0	12.12
L1-Speed-D_{mean}	1.28	1.27	0	27.05	1.88	1.58	0	11.6
L1-$Accleration_x$*-*S_{dev}	0.05	0.03	0.01	0.2	0.08	0.04	0.01	0.28
L1-$Accleration_x$*-*D_{mean}	0.04	0.03	0	0.18	0.06	0.03	0	0.22
L1-$Accleration_y$*-*S_{dev}	0.04	0.04	0.01	0.24	0.06	0.05	0	0.4
L1-$Accleration_y$*-*D_{mean}	0.03	0.03	0	0.21	0.04	0.04	0	0.31
L2-Speed-V_f	0.01	0.04	0	0.68	0.03	0.06	0	0.6
L2-Speed-S_{dev}	2.68	2.32	0	96.22	3.72	2.52	0	20.06
L2-Speed-D_{mean}	2.28	1.98	0	81.91	3.16	2.15	0	16.89
L2-Speed-C_v	0.04	0.07	0	1.15	0.13	0.17	0	1.16
L2-Speed-EWMA	0.01	0.04	0	0.68	0.03	0.06	0	0.6
L2-$Accleration_x$*-*S_{dev}	0.02	0.01	0	0.12	0.04	0.02	0	0.15
L2-$Accleration_x$*-*D_{mean}	0.02	0.01	0	0.1	0.03	0.02	0	0.13
L2-$Accleration_y$*-*S_{dev}	0.03	0.01	0	0.15	0.03	0.02	0	0.23
L2-$Accleration_y$*-*D_{mean}	0.02	0.01	0	0.13	0.02	0.02	0	0.19

**L1: Event-based volatility measure; L2: Temporal volatility measure;* S_{dev}*: Standard deviation;* V_f*: Time-varying stochastic volatility;* C_v*: coefficient of variation;* D_{mean}*: mean absolute deviation;* $Acceleration_x$*: longitudinal acceleration;* $AccDec_x$*:both longitudinal acceleration and deceleration;* $Accleration_y$*: lateral acceleration; EWMA: Exponentially Weighted Moving Average Volatility*

Conceptual Framework

The conceptual framework of the study is provided in Figure 6.4. It has three main phases. The first phase is sensing which collect driver information (i.e. in terms of distraction) and vehicular movements (i.e. speed, longitudinal and lateral acceleration). As discussed in the previous section, the data is preprocessed and cleaned by excluding the evasive maneuvers of critical events and considering 15 seconds for each event. Next, the raw data is fed to the feature extraction phase in order to obtain volatility indices at the event and temporal levels. Seventeen volatility indices are extracted o quantify speed, longitudinal and lateral acceleration variations. Finally, the raw data and extracted features are fed to the deep-learning phase. Deep NN, 1D-CNN and LSTM RNN models are developed to classify events to baseline and critical event and the performance of the models are evaluated.

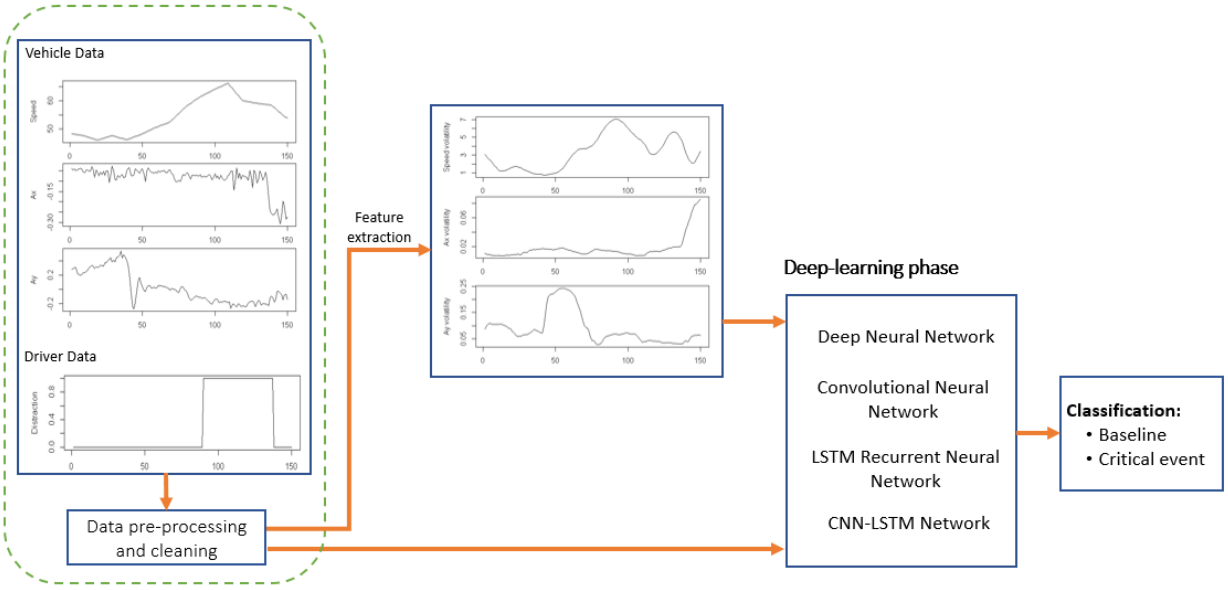


Figure 6.4 Conceptual framework of the study

Problem formulation

As discussed, three deep-learning methods are utilized to classify events: Deep Neural Network, 1D-Convolutional Neural Network (1D-CNN), Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN), and CNN-LSTM model. While the multi-layer deep neural network process the input data and information through interconnected neurons, it suffers from a main limitation where it assumes that all inputs are independent from each other, which is not the case in many fields, such as image classification, language processing, and time-series problems. Therefore, several methods are proposed to address the dependency in the input of the network (in this paper time dependency) by including local information (temporal information) in the input data.

Deep NN

A DNN model is known as a feed-forward artificial neural network, which has more than one hidden layer between the input and output layers (Hinton *et al.* 2012a). These models are processing the information through a series of fully connected layers and associated with other layers through weighted connections. Each node is called a neuron which transforms the input with a non-linear function to create a decision boundary. Each neuron can be considered as a non-linear computational unit which applies activation function (e.g. sigmoid function). The neurons can be defined as:

$$a^{l+1} = f(W^l a^l + b^l) \tag{6.6}$$

where a^l and a^{l+1} denotes the activation value in level l and l+1, respectively, W^l is a weight matrix, b^l is the bias, and $f(.)$ represents the activation function. The especial

case is *l=1* which denotes the input layer, and we denote it by $a^l = x$. The last layer of the DNN is a softmax classifier and the output is the two classes (i.e. Baseline and Critical events). Here, we considered a fully connected network with four dense layers with 200 nodes on each layer, following a softmax layer. While DNN models are prone to overfit the data, dropout regularization is utilized to penalize the weights. Dropout is known as one of the powerful techniques for improvement in the generalization error of large NNs, introduced by Hinton et al. (Hinton *et al.* 2012b). The training procedure applies forward and backward propagations. While forward propagation computes actual classifications based on the input data, the backward propagation aim is to update the parameters to minimize the discrepancy between the predicted and observed values.

1D CNN classifier

Comparing to simple NN models that perform feature extraction by only taking a vector of inputs to the model, 1D CNNs allow us to operate in a multi-scale manner and further investigating the time-series dependency between the observations. The structure of the 1D-CNN model used in this study is shown in the Figure 6.5. The time-series motions of vehicles, driver distraction profile, and driving volatility measures are the model input, and the output layer classified the output event. In the time-series data analysis, we can treat the input as a picture with the size of (n,1) pixels with v bands where v is number of input streams. The convolutional layers in the model are extracting additional features from the data. In the convolutional layer, the model applies the convolution operation on the local input data in order to generate the corresponding 1D features, while applying different convolutions will generate several features from the input data. In each convolutional layer of the model, 1D forward propagation is performed which can be formulized as:

$$x_k^l = b_k^l + \sum_{i=1}^{N_{l-1}} conv_{1D}(w_{ik}^{l-1}, s_i^{l-1}) \tag{6.7}$$

where x_k^l and b_k^l represent the k^{th} neuron at layer *l* input and bias, respectively. w_{ik}^{l-1} is the kernel function in layer *l-1.*

Since the feature map resulted from each convolutional layer is sensitive to the location of the features in the input and number of feature maps increase the dimensionality of data, a pooling layer is utilized to down-sample the feature map. Among two common pooling methods which are average-pooling and max-pooling, the latter is used to summarize the most activated node of a feature. At the end, the high-level features are flattened and fed into a fully connected layer to perform classification.

As earlier discussed, one of the drawbacks of Deep NN is ignorance of local dependency between observations. On the other hand, our data is streams of time-series data which have one dimension with multiple channels (i.e. Speed, longitudinal, and lateral

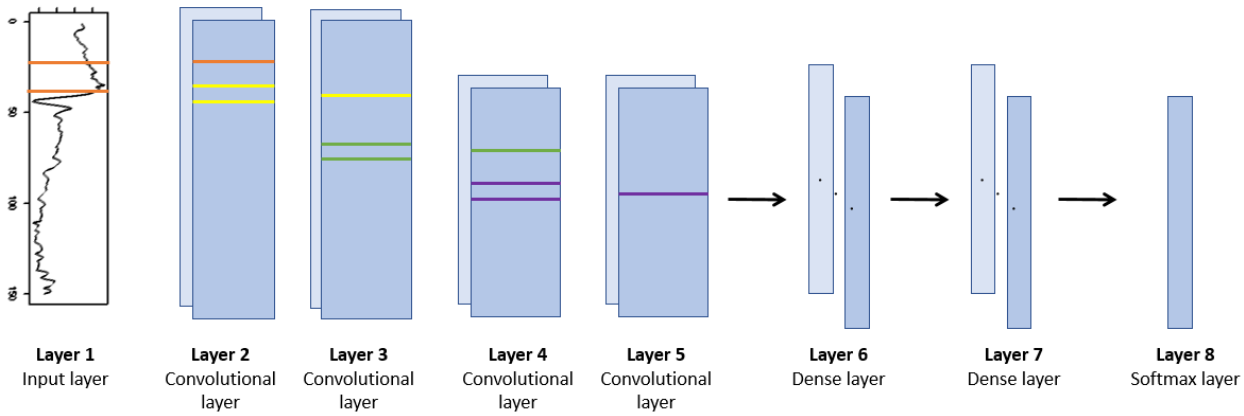


Figure 6.5 Representation of 1D-CNN network used in this study

acceleration). Thus, we used 1D-CNN to extract the features and feed the fully connected NN. Figure 6.5 depicts the structure of the model.

While in the conventional deep neural network, a neuron is fully connected to neurons of a next layer, CNN structure is different in a sense that neurons are sparsely connected to each other based on their relative position. Therefore, in a DNN, values of the next layer (hidden neuron *i* in layer *j*, $h_{i,j}$) is obtained by multiplying all the neurons of the previous layer (h_{j-1}). On the other hand, the CNN model computes each hidden activation by multiplying a subset of local input to the matrix of weights (W). It should be noted that weight matrix is shared across the entire layer which helps the model to reduce the number of estimated parameters and efficient training. A max-pooling layer frequently used after a convolutional layer.

LSTM-RNN

Recurrent Neural Network models (RNN) models are aimed to tackle this issue by using recurrent connection in every neuron to include temporal information and feeding back this information to itself with a unit of time delay (Ordóñez and Roggen 2016). This helps the model to learn the temporal dynamics of time-series input.

The RNN model requires an input sequence $a^l = \left(a_1^l, a_2^l, \dots, a_T^l\right)$ where a_i^l is the activation of unit *i* at layer *l*, and *T* is the length of input sequence. Be performing the following recursive equation, a sequence of activations of the next layer, $a^{l+1} = \left(a_1^{l+1}, a_2^{l+1}, \dots, a_T^{l+1}\right)$, will be obtained:

$$h_i^l = \sigma\left(W_{xh}^l a_t^l + h_{t-1}^l W_{hh}^l + b_h^l\right) \tag{6.8}$$

where σ is the activation function, W_{xh}^l and W_{xh}^l are the input-hidden and hidden-hidden weight matrix, respectively, and b_h^l is the vector of bias. The a_t^l can be obtained as following:

$$a_t^{l+1} = h_t^l W_{ha}^l + b_a^l \tag{6.9}$$

where W_{ha}^l is the hidden-activation weight matrix, and b_a^l is the vector of activation bias. Although RNN is designed to deal with time-series data, they are suffering from the problem of vanishing and exploding gradient which affect the model fit for the long-time lag models (Hochreiter and Schmidhuber 1997, Zhao *et al.* 2017) and face several challenges in real-world sequence modeling (Gers *et al.* 2002).

The LSTM models are attempting to extend the conventional RNN models which are capable of learning long-term time dependency in the input, introduced in 1997 by Hochreiter and Schmindhuber (Hochreiter and Schmidhuber 1997). Similar to RNN models, LSTM have the chain structure, while the difference is in the design of a neurons (Figure 6.6). While RNNs have a single learning neuron (e.g. tanh), there are three gates interacting with each other. The LSTM utilizes the concept of gating to provide a mechanism that defines the behavior of each memory cell in the network. the cell state is updated according to the gates' activations. The input data of the LSTM is fed into the write gate (input), read gate (output), and reset gate (forget). The LSTM layer can be

written as:

$$i_t = \sigma_i(W_{ai}a_t + W_{hi}h_{t-1} + b_i) \tag{6.10}$$

$$f_t = \sigma_i\left(W_{af}a_t + W_{hf}h_{t-1} + b_f\right) \tag{6.11}$$

$$c_t = f_t c_{t-1} + i_t \sigma_c(W_{ac}a_t + b_c) \tag{6.12}$$

$$o_t = \sigma_o(W_{ao}a_t + W_{ho}h_{t-1} + b_o) \tag{6.13}$$

$$h_t = o_t \sigma_h(c_t) \tag{6.14}$$

where i_t, f_t, and o_t denotes the input, forget, and output gates, respectively, c_t represents cell activation vectors, and σ is the activation function.

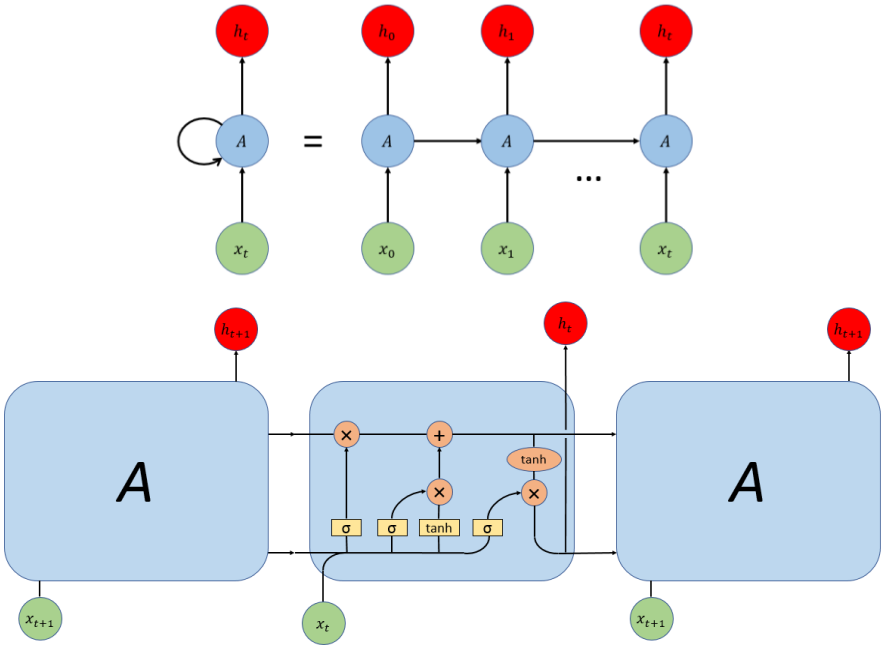


Figure 6.6 Illustration of the structure of the LSTM neural network

Table 6.2 Structure of the LSTM model

Layer type	Output shape	Number of parameters
LSTM 1	(None, 150, 100)	48400
LSTM 2	(None, 150, 100)	80400
LSTM 3	(None, 150, 100)	80400
Dense 1	(None, 100)	10100
Dense 2	(None, 100)	10100
Dense 3	(None, 2)	202

The LSTM classifier commonly followed by dense fully connected hidden layers, and softmax layer. The structure of the LSTM model is provided in the Table 6.2.

CNN-LSTM

Typically, DNN and RNN models receive the raw input data, while it has shown that by applying feature derived layers, their accuracy can be improved significantly (Palaz and Collobert 2015). Convolutional layers have been suggested in order to extract additional features from the raw data (Yang *et al.* 2015). A convolution layer extracts features from the input data by applying a kernel (filter). By applying these kernels to different regions of input data, possibly additional patterns are recognized and captured by these kernels. It is worth noting that these kernels are optimized during the training process.

The application of convolution layer mainly relies on the input dimension. Considering images that has two dimensions, 2D kernels typically applied in the convolution layer (Baghbaderani and Hairong 2019), while in a time-series analysis, 1D kernels are the most common approach (Ordóñez and Roggen 2016). In the 1D context, a kernel can be

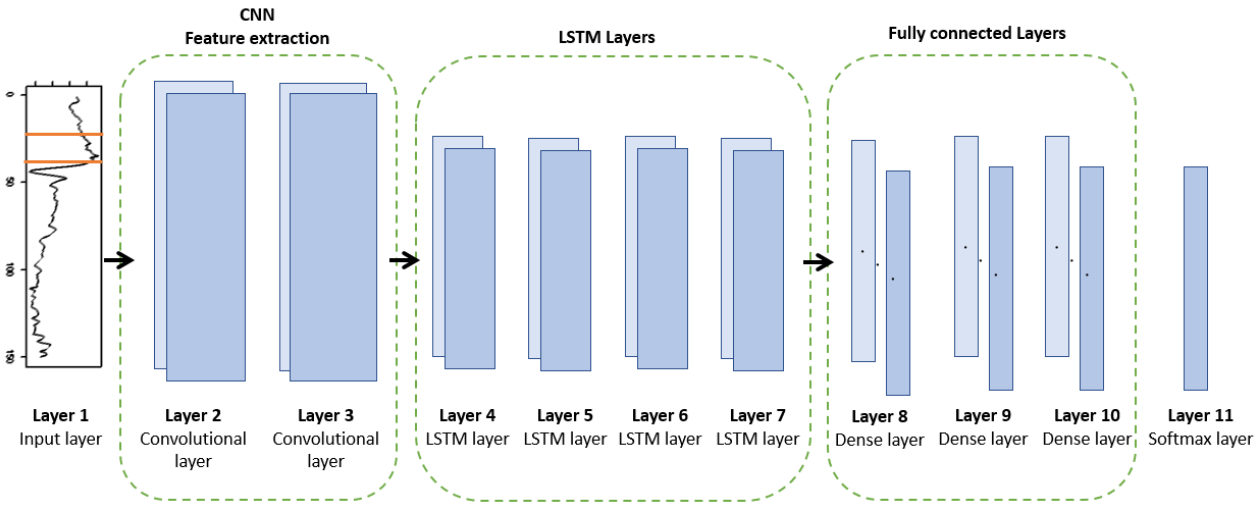


Figure 6.7 Structure of the CNN-LSTM model

considered as a filter which can removes the outliers, feature detector, and data filtering (Ordóñez and Roggen 2016). Feature map extraction using one-dimensional convolution can be written as:

$$a_j^{l+1}(\tau) = \sigma\left(b_j^l + \sum_{f=1}^{F^l} K_{jf}^l(\tau) * a_f^l(\tau)\right) \tag{6.15}$$

where a_j^{l+1} is the feature map *j* in layer *l+1*, σ is the kernel non-linear function, F^l denotes the feature maps of layer *l*, K_{jf}^l is the kernel mapping *f* to feature map *j* in layer *l+1*.

The structure of the 1DCNN-LSTM model is provided in the Table 6.3. The input data is fed to the two convolution layers and max-pooling layer is applied and the outputs are flattened. Next, the features are fed into three layers of LSTM with 100 nodes, and finally a dense fully connected network with a softmax classifier performs the classification.

Table 6.3 Structure of the CNN-LSTM model

Layer type	Output shape	Number of parameters
1D-CNN 1	(None, None, 75, 12)	492
1D-CNN 2	(None, None, 75, 12)	300
LSTM 1	(None, None, 50)	190200
LSTM 2	(None, None, 50)	20200
LSTM 3	(None, None, 50)	20200
LSTM 4	(None, 50)	20200
Dense 1	(None, 100)	5100
Dense 2	(None, 100)	10100
Dense 3	(None, 100)	10100
Dense 4	(None, 2)	202

Experimental evaluation

Training procedure

The dataset is randomly divided to the train and test datasets. A 20 percent of the training data set is randomly separated for the validations. In order to initialize the training, all the weight matrices and bias vectors were initialized randomly. The dropout and L2 regularization approaches are utilized to prevent overfitting of the training sample. Optimal dropout value for each model is obtained in a manner to prevent the risk of overfitting while getting the highest accuracy. To improve the efficiency, the data is segmented to mini-batches with the size of 256.

All the models are trained by the Adam optimization approach (Kingma and Ba 2014) which is known for its efficient stochastic optimization approach with only requiring the

first-order derivatives and low memory for analyzing. Moreover, it has the advantage of high computational efficiency, low memory requirement, straightforward implementation, and invariant feature to diagonal gradients rescaling, and also it is appropriate for problems that the data and parameters are large scale (Kingma and Ba 2014). Furthermore, it has the advantage of suitability problems with noisy and sparse derivatives and non-stationary objectives (Kingma and Ba 2014). The Adam optimizer is utilized to minimize the loss of objective function, which we used cross entropy function, by finding the optimal weights and bias terms.

In order to perform training, the dataset is randomly divided to train and test datasets, and the break-down of the events can be found in Table 6.4. This study took advantage of using the Keras and TensorFlow deep learning tools. Keras and TensorFlow are open-source python machine learning libraries developed by the Google Brain Team (Chollet 2015, Abadi *et al.* 2016), which widely used in the deep learning context by several studies (Baghbaderani and Hairong 2019, Baghbaderani *et al.* 2019, Nezafat *et al.* 2019, Parsa *et al.* 2019). The model training and classification are run on a workstation computer with the TITAN RTX graphical processing unit (GPU) with 4608 cores, 1770 MHz clock speed and 24 GB RAM.

Figure 6.8 illustrates the accuracy and loss for the training and validation data vs training epoch for each model. It can be observed that the performance of the CNN-LSTM model in terms of training and test accuracy and loss is better than other models. The results will be discussed further in section 5.3.

Table 6.4 Test and train datasets

	Training sample size	Test sample size
Baseline	5623	1941
Crash/Near crash	1376	549
Total	6999	2490

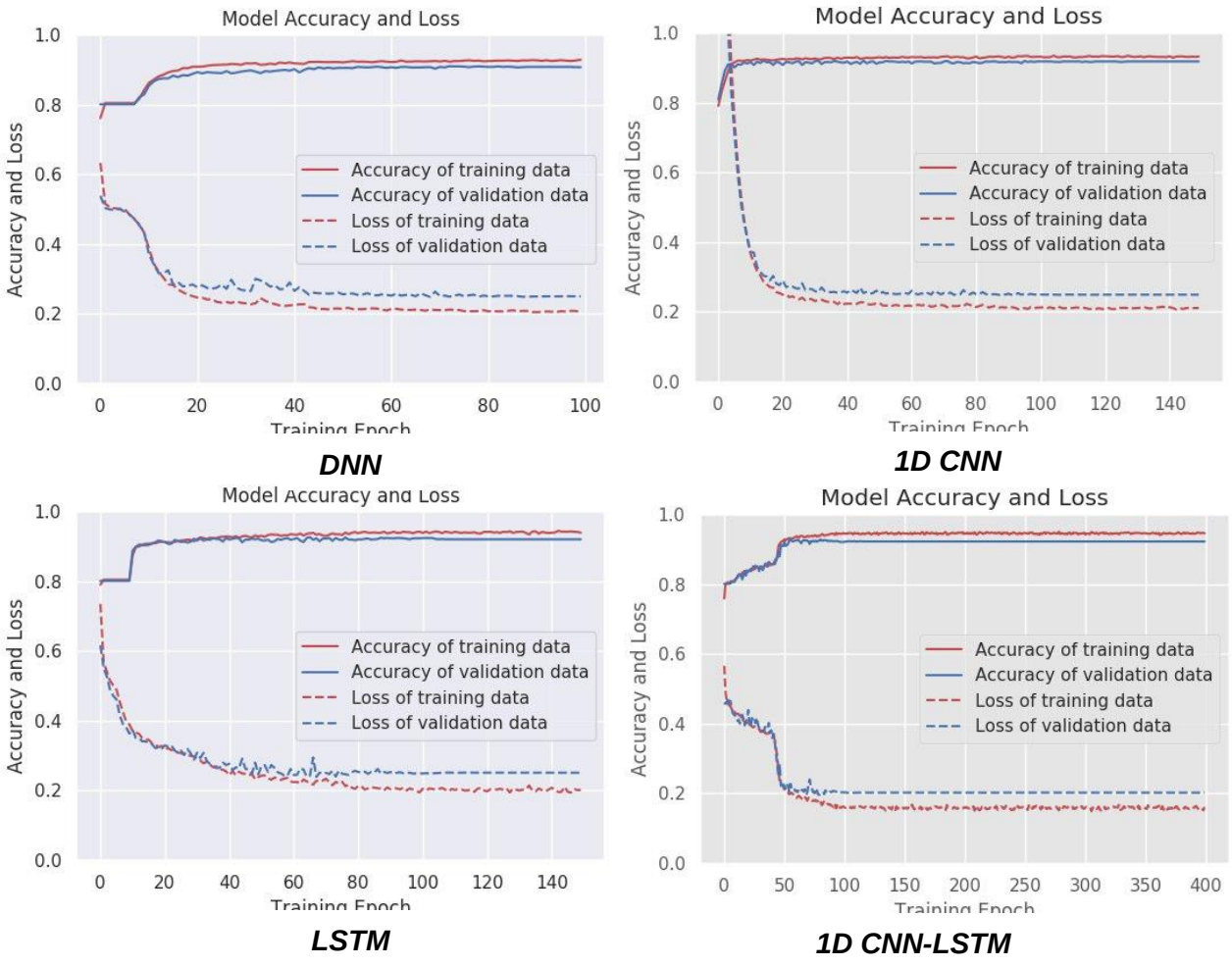


Figure 6.8 Accuracy and loss for the training and test datasets

Evaluation metrics

In order to score the performance of the models, four common metrics are utilized including accuracy, precision, recall, and F-measure (F_1). The definitions are provided in the following:

1- Accuracy which measures the overall performance of the model by quantifying the proportion of correct predictions over all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.16)$$

where TP, TN, FP, and FN are the total true-positive, true-negative, false-positive, and false-negative predictions.

2- Precision which measures the number of true positive over total predicted positive. The precision of class *c* can be obtained by:

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \quad (6.17)$$

where TP_c and FP_c are the number of true-positive and false-positive of class *c,* respectively.

3- Recall which measures number of corrected classified observations over the total observations in the class *c.*

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad (6.18)$$

where FN_C is the number of false-negative of class *c*.

4- F_1 score which applies weighted harmonic average to the precision and recall.

$$F_1 = \sum_i 2 * w_i \frac{Precision_i * Recall_i}{Precision_i + Recall_i} \quad (6.19)$$

where *i* represents the class index (i.e. baseline and critical events), w_i is the proportion of class *i* observation ($w_i = \frac{n_i}{N}$) where n_i is the number of observations in class *i* and *N* is the total number of observations. The advantage of F_1 is its suitability for imbalanced data where the proportion of one of the classes is lower comparing to others. This measure combines precision and recall metrics and weighting the classes based on the proportion.

Comparative results

The comparative results of the DNN, CNN, LSTM, and CNN-LSTM models are provided in the Table 6.5. As discussed, several metrics are utilized to evaluate the performance of the models. In terms of overall goodness of fit, Accuracy and Loss metrics are utilized, while for each class, precision, recall, and F1 measure is used.

Focusing on overall performance on the test dataset, results suggest that the accuracy of the DNN model is 88.51 percent. On the other hand, LSTM, 1D-CNN, and 1DCNN-LSTM models substantially improved the accuracy to 91.61%, 90.76%, and 92.36%. Furthermore, the total loss is significantly dropped from 0.31 to 0.24. It can be observed

that the LSTM, 1D CNN, and 1DCNN-LSTM models improved the fit by incorporating temporal dependency between observations. Therefore, it can be inferred that there is a great need to consider local dependency of time-series input in this context. Also, the results revealed that combining 1D-CNN and LSTM models can improve the fit by extracting additional features from the input data and incorporate those to a time-series analyzer model (i.e. LSTM model).

Since the goal of this paper is prediction of a critical event occurrence, we want to assess the models' performance on crash/near-crash events. Precision, recall, and F_1-score metrics are suggesting that consideration of time dependency in the analysis improved the performance comparing with DNN model. While DNN model is predicting 68 percent of crashes with the precision of 0.77, CNN-LSTM model improved it to 71 percent with the precision of 0.93. Moreover, F_1-score substantially improved from 0.72 to 0.80.

Referring to the baseline events, it can be observed that CNN-LSTM model improved the precision of the predictions from 83 percent to 92 and 93 percent, respectively. The results suggest that the 1D-CNN model slightly performed better than the LSTM and 1D-CNN models.

Table 6.5 Models performance evaluation

		Train Data				Test Data			
	Metric	**DNN**	**1D CNN**	**LSTM**	**CNN-LSTM**	**DNN**	**1D CNN**	**LSTM**	**CNN-LSTM**
test time (millisec)		-	-	-	-	0.181	0.194	19.65	0.345
Overall	*Accuracy (%)*	93.07	94.96	93.48	93.84	88.51	91.61	90.76	**92.36**
	Loss	0.21	0.15	0.20	0.21	0.31	0.26	0.25	**0.24**
Baseline	*Precision*	0.92	0.94	0.92	0.94	0.90	**0.92**	0.91	**0.92**
	Recall	0.99	0.99	0.99	0.99	0.94	**0.98**	**0.98**	**0.98**
	F_1*-Score*	0.96	0.97	0.96	0.97	0.93	**0.95**	0.94	**0.95**
Critical Event	*Precision*	0.96	0.97	0.97	0.97	0.77	0.90	0.89	**0.93**
	Recall	0.67	0.76	0.66	0.76	0.68	0.69	0.66	**0.71**
	F_1*-Score*	0.79	0.85	0.76	0.85	0.72	0.78	0.76	**0.80**

Importance of volatility and distraction profile

As showed in the previous section, the 1DCNN-LSTM model performed the best in terms of predicting extreme events occurrence comparing to the other discussed models. In this section, the feature space is divided into three main blocks and contribution of adding each block to the model is discussed by evaluating the 1DCNN-LSTM model performance. Three sets of features are considered: 1- Vehicle kinematics, 2- Driving volatility, and 3- Distraction profile. Initially, the model is trained with vehicle kinematics, next driving volatility features are added to the model, and finally distraction profile is added to the feature space. The results for the 1DCNN-LSTM model is summarized in the Table 6.6. According to the results, by feeding the model only with vehicle kinematics, the model accuracy on the test dataset will reach 78.15 percent. By adding the extracted volatility features to the model, the model accuracy reaches 86.75 percent, which indicates 8.6 percent improvement comparing to the vehicle kinematics only. Finally, adding the distraction profile of the driver to the model will enhance the accuracy to 92.36 percent.

Overall, it can be inferred that driving volatility and distraction profile substantially improve the prediction accuracy of the 1DCNN-LSTM model.

Conclusion

Emergence of high-resolution big data generated by connected and automated vehicles and new sensors coupled with availability of high-performance computational resources, enabled the application of new concepts and methods. This study develops a framework to quantify the real-time crash occurrence risk by integrating multiple data sources and applying deep learning methods. The kinematic movement of vehicle and information on driver impairment, in terms of distraction, is obtained from second Strategic Highway Research Program (SHRP2) and volatility functions are employed to extract additional features from vehicular movements to quantify variations in instantaneous driving decisions. The initial statistical analysis revealed that impaired driving, in terms of distraction, and instability in driving can be served as the leading crash occurrence

Table 6.6 Evaluation of feature importance in the 1DCNN-LSTM model

		Vehicle Kinematics		Kinematics & Volatility		Kinematics & Volatility & Distraction	
Performance		**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Overall	*Accuracy (%)*	84.19	78.15	87.09	86.75	93.84	92.36
	Loss	0.42	0.50	0.36	0.35	0.21	0.24
Baseline	*Precision*	0.87	0.86	0.88	0.87	0.94	0.92
	Recall	0.85	0.81	0.98	0.98	0.99	0.98
	F_1-*Score*	0.86	0.84	0.93	0.92	0.97	0.95
Critical Event	*Precision*	0.50	0.49	0.86	0.85	0.97	0.93
	Recall	0.53	0.49	0.49	0.46	0.76	0.71
	F_1-*Score*	0.52	0.43	0.62	0.60	0.85	0.80

indicator.

In order to perform real-time critical event prediction, several deep learning models including 1D-convolutional neural network (1D-CNN), Long Short-Term Memory (LSTM), and CNN-LSTM approaches are utilized to compare their performance with the Deep Neural Network (DNN) model as the baseline. Based on the results, by capturing the time dependency of the input data, the model performance can be improved significantly. The results revealed that extra features extracted by the CNN model coupled with the LSTM model can help us to achieve 92.36% accuracy on the test data and predicting 71 percent of the crashes correctly with the 93 percent precision. Furthermore, the analysis revealed that by adding driving volatility features and driver distraction profile, the model accuracy can enhance for 14.21% comparing to the LSTM CNN model feeding with vehicle kinematics only.

The developed model in this study can be used to proactively and in real-time monitor the driving performance of the drivers and provide warning at the times that they exhibit volatile and distracted driving. This study utilized driving instability, driver distraction, and vehicle kinematics as the inputs of the network. In future research, other streams of data including roadway condition and traffic state, and information of the surrounding vehicles can be incorporated to improve the model performance by providing additional information regarding the surrounding environment.

CHAPTER 7 : CONCLUSION AND IMPLICATIONS

Emergence of new data sources opens a new window to the researchers to think out of the box and apply new methods into the transportation field. This dissertation aims to harness emerging large-scale microscopic data generated by new technologies including connected and automated vehicles, and naturalistic driving data, and integrate this information into transportation safety analysis from micro to macro level, focusing on driver behavior as the leading crash cause. It should be noted that driver behavior is the most critical and unpredictable factor in the transportation system, which has shown that contributing to more than 93% of crashes. Therefore, by in-depth analysis of instantaneous driver behavior and decision, we can apply countermeasures and strategies to reduce crash risk.

New technologies and transportation modes ranging from connected vehicles, automated vehicles, roadside units, crowdsource data, and camera surveillance are generating enormous data which can be utilized to perform in-depth analysis of driver behavior. This dissertation take advantage of several data sources including real-world connected vehicle data collected in Safety Pilot Model Deployment study (with more than 2.2 billion observation), naturalistic driving and biometrics data collected via SHRP2 study (with more than 2 million observation), roadway inventory, and crash data. This dissertation develops a unique and systematic framework to integrate these multidimensional datasets and incorporate them into the transportation safety analysis. The developed framework is used to incorporate such a big data into the analysis at different levels, i.e. 1) instance level, 2) event level, 3) Location level, and 4) Network level in order to explore the association of driver behavior with crash risk. A data processing approach is

generated to harness this big data and extract useful information to improve traditional safety analysis.

This dissertation utilized the concept of driving volatility, which aims to quantify variations in driving behavior. We introduced the concepts of temporal and unintentional driving volatility to quantify instantaneous variations in driving behavior in terms of speed, acceleration/deceleration, and jerk. In addition, the concepts of intentional and unintentional driving volatility are developed to quantify instability in driving prior and during safety critical events (i.e. crashes and near-crashes). Furthermore, the concept of location-based driving volatility is expanded by developing several volatility measures to identify locations where drivers exhibit erratic behaviors. While the literature just focused on longitudinal vehicular movements, this dissertation extends the concept into 3D dimension by incorporating lateral and vertical movements into the analysis.

The chapter 2 of the dissertation developed a methodological and systematic framework to harness and integrate big data generated by connected vehicles into safety analysis at the network level. The concept of temporal driving volatility is introduced to capture and quantify the extent of variations in each instance of driving. The driving volatility concept is extended to the 3D dimension by incorporating lateral and vertical volatilities into the analysis and utilized to quantify spatiotemporal variations in driving decisions. The CV data collected in the SPMD study at Ann Arbor, MI is used to illustrate the framework. The initial analysis revealed that there is a positive correlation between developed temporal driving volatility measures and crash risk. To quantify this correlation, several spatial modeling approaches (i.e. Geographically Weighted Poisson and Negative

Binomial Regressions) are developed to account for spatial heterogeneity. The modeling results revealed that extracted volatility measures from the CV data are significantly associated with crashes. Based on the results, using only CV data we can explain around 70% of the model deviance. Given the contributing volatility measures, the unsupervised learning methods (i.e. k-means and Gaussian Mixture Models) are used to identify hotspot locations where crash frequency is low while driving volatility is high. This analysis will help agencies to proactively identify potential hotspots locations in the network and treat them by applying countermeasures to reduce driving volatility.

The Chapter 3 of the dissertation focuses on the integrating CV data into the safety analysis of intersections. This chapter explores the association of longitudinal and lateral driving volatility with different crash types, i.e. rear-end, sideswipe, angle, and head-on crashes. The CV data collected in the SPMD study is utilized to calculate several volatility measures capturing variations in longitudinal and lateral movements of more than 2800 CVs passing the intersections. In this study, 167 intersections of Ann Arbor, MI is selected, and several volatility measures are calculated by analyzing more than 125 million Basic Safety Messages transmitted between CVs. In order to capture variations in vehicle control and movement, 30 measures of driving volatility are calculated by using speed, longitudinal and lateral acceleration, and yaw-rate. Also, intersection inventory and historical crashes are manually extracted for the study area. In terms of modeling, fixed parameter, random parameter, and geographically weighted Poisson regression models are utilized to quantify correlation of developed volatility measures and crash risk at intersections. The results revealed that the developed volatility measures are significantly associated with crash risk and they can substantially improve model performance

comparing to the conventional safety methods, which only use traffic exposure and intersection inventory data. According to the results, controlling for intersection geometry and traffic exposure, and accounting unobserved factors, variations in longitudinal control of the vehicle (longitudinal volatility) are highly correlated with the rear-end crash frequency. Intersections with high variations in longitudinal movement are prone to have higher rear-end crash rate. Referring to sideswipe and angle crashes, along with speed and longitudinal volatility, lateral volatility is substantially correlated with the frequency of crashes. When it comes to head-on crashes, speed, longitudinal and lateral acceleration volatilities are highly associated with the frequency of crashes. Intersections with high lateral volatility have higher risk of head-on collisions due to the risk of deviation from the centerline leading to head-on crash. The developed methodology and volatility measures can be used to proactively identify hotspot intersections where the frequency of crashes is low, but the longitudinal/lateral driving volatility is high. The reason that drivers exhibit higher levels of driving volatility when passing these intersections can be analyzed to come up with potential countermeasures that could reduce volatility and, consequently, crash risk.

Chapter 4 of the dissertation focuses on event-level analysis in order to perform in-depth analysis of crash contributing factors using high-resolution naturalistic driving data. This chapter studies the role of pre-crash instability in driving, in terms of driving volatility, on crash intensity. The crash intensity in this dataset is measured on a four-point scale from a tire-strike to an injury crash. The SHRP2 NDS data are used to investigate the movements and instability of vehicles in space prior to involvement in a crash and their contribution to crash intensity using path analysis. The data containing 617 crash events

with around 0.18 million temporal trajectories which were used to quantify instability in driving. To quantify driving instability, microscopic variations or volatility in vehicular movements before a crash are analyzed. Specifically, nine measures of pre-crash driving volatility are calculated and used to explain crash intensity. The fixed and random parameter probit models are applied to model intensity of crashes. The results revealed that unintentional volatility is one of the leading factors increasing the chance of a severe crash. Interestingly, distracted and aggressive driving are highly correlated with driving volatility and have substantial indirect effects on crash intensity. With volatile driving serving as a leading indicator of crash intensity, given the crashes analyzed in this study, early warnings and alerts for the subject vehicle driver and proximate vehicles can be helpful when volatile behavior is observed.

Chapter 5 focuses on distracted and impaired driving as one of the key contributing factors in roadway crashes, leading to about 35% of all transportation-related deaths in recent years. Despite the prevalence and importance of distracted driving, little is known about how the duration of distractions is associated with critical events. With new sensors and increasing computational resources, it is possible to monitor drivers, vehicle performance, and roadways to extract useful information, e.g., eyes off the road, indicating distraction and inattention. Using high-resolution microscopic SHRP2 naturalistic driving data, this chapter conducts in-depth analysis of both impairments and distractions. The data has more than 2 million seconds of observations of 7394 baselines (no event), 1237 near-crashes, and 617 crashes. The interval of distracted driving during the period of observation (15 seconds) were calculated and fixed and random parameter logistic regression models of crash risk was estimated. The results reveal that alcohol

and drug impairment is associated with a substantial increase in extreme event involvement of 29.7%, and the highest correlations with crash risk include duration of distraction by cellphones, driver reading or writing, and atypical distraction, e.g. insect in the car. Using detailed pre-crash data from instrumented vehicles, the study contributes by quantifying crash risk vis-à-vis detailed driving impairment and streams of secondary task involvement and discusses implications of the results.

Chapter 6 take an advantage of emerging vehicle instrumentation and computational power which allow for the real-time monitorization of driving environment in terms of driver, vehicular movement, roadway, and vehicle's surrounding to identify leading indicators of crashes from multi-dimensional data streams. In order to extent variations in driving decisions that are beyond normal in driver behavior and vehicle kinematics, the concepts of temporal and unintentional volatility are applied. By integrating and fusing multiple streams of data, this study aims to predict the probability of safety critical event and generate useful feedback to drivers and warnings to surrounding vehicles regarding hazards. To capture the local dependency and volatility in time-series data, several deep learning techniques including 1D-Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and 1DCNN-LSTM are applied, and vehicle kinematics, driving volatility, and impaired driving (in terms of distraction) are used as the input parameters. The results reveal that the 1DCNN-LSTM model provides the best performance, with 92.5% accuracy and prediction of 71% of crashes with a precision of 93%. Additional features are extracted with the CNN layers and temporal dependency between observations is addressed. The model can be used to monitor driving behavior in real-time and provide warnings and alerts to drivers in low-level automated vehicles, reducing their crash risk.

Contribution

The contribution of this dissertation is develop a unique and systematic framework to harness big data generated by emerging data sources into the transportation safety analysis at the different levels including: 1) system-level, 2) location-level, 3) driver-level, 4) evet-level, and 5) instance-level. The developed framework can integrate multidimensional data generated from different sources including 1) emerging data sources (i.e. connected and automated vehicles, naturalistic driving, roadside units), 2) roadway inventory data (i.e. geometric information, traffic exposure, historical crashes), and 3) driver biometrics (i.e. profile of driver distraction). In order to harness this data, the concepts of temporal driving volatility and unintentional driving volatility are developed to quantify instantaneous variations in driving behavior. Furthermore, this study contributes the literature by extending the longitudinal driving volatility to lateral and vertical volatilities in order to capture vehicular movements in the 3D dimensions. Furthermore, a full review of volatility measures is conducted in order to find the best measures quantifying crash risk[3].

In response to each level of analysis, a safety index is utilized and the associations of extracted features from the big data in terms of driving volatility, network features, and driver distraction profile are explored. Different modeling techniques including frequentist approach, heterogeneity-based approach, spatial models, machine learning, and deep learning techniques are utilized, and the correlation of extracted features and crash risk

[3] This research is published in the transportation research record journal.
Kamrani, M., Arvin, R., & Khattak, A. J. (2018). Extracting useful information from Basic Safety Message Data: an empirical study of driving volatility measures and crash frequency at intersections. *Transportation Research Record*, *2672*(38), 290-301.

is studied. The results revealed that the extracted features from the big data at different level can significantly and substantially enhance the current literature. While currently the human driver, as the key contributing factor in traffic crashes, is neglected in the analysis, this dissertation provided the framework to quantify instantaneous driver decisions and incorporate it into the analysis. Developing such a framework is crucial due to disruptive developments in emerging technologies, especially connected and automated vehicles, that generates enormous data each day. This dissertation highlights the value of such a data and provides a framework to harness and incorporate it into safety analysis at different levels.

Implications

Overall, this study utilized the concept of driving volatility and extended it in several aspects to contribute to the analysis of real-world big data in order to produce applicable knowledge for intelligent transportation systems and smart cities. The developed framework has wide range of application in different contexts, especially in connected and automated vehicles. While the micro level analysis of this study can be adopted in low levels of automated vehicles for driver monitoring, macro level analysis can be used to improve the navigation capability of CAVs. In the following, some of the potential applications are discussed.

As discussed in the chapter 2, this dissertation developed a safety map quantifying driving volatility across the study area. The created map using CV data can be utilized to better navigate the CAVs in a mixed traffic with conventional vehicles and adjust its behavior in

different locations based on the safety level across the space. Furthermore, additional map based on the CAV data can be generated which mainly models driving behavior of CAVs across the study area. This will help manufacturer to improve the controlling algorithm at unsafe locations to prevent any potential conflicts with surrounding environment. On the other hand, government agencies (e.g. Department of Transportation) can use the developed methodology to identify hotspot locations in the network where crashes are waiting to happen. The algorithm will help them to find locations where further investigations and improvements are needed to reduce driving volatility as a leading indicator of crashes.

Focusing on conventional human-driven vehicles, the developed framework is capable to classify drivers based on the developed spatiotemporal driving profile for each driver in chapter 2 using their historical driving across the study area[4]. As an illustration, the CV data collected in the SPMD study contains information of more than 2800 drivers, which was used to classify drivers in different spatial and temporal contexts using unsupervised machine learning based on their historical driving records at different neighborhoods, i.e. commercial, residential, and highways. This score will help both insurance companies and drivers to monitor their driving performance and reduce their crash risk.

The developed safety map can also be utilized to perform route choice modeling for the system users. As part of this dissertation, we have published a paper which suggests a

[4] Mohammadnazar, A., Arvin, R. & Khattak, A. (2020). “Categorizing Driving Style using Connected Vehicle Data: Application of unsupervised learning”. Presented at the 99th Transportation Research Board Conference.

framework to integrate mobility and safety for the pathfinding problem[5]. The developed model is capable to perform route choosing based on the historical driver behavior (in terms of driving volatility), safety of the alternative routes, and travel time. To illustrate the model, the framework is applied to the CV data in Ann Arbor, MI and the implications of the approach widely discussed. Further application of this approach can be safety-based traffic assignment. Currently, the literature has mainly considered travel time, travel cost, and fuel consumption to perform traffic assignment, while due to lack of data, there is no study that focuses on safety-based assignment. Using the developed volatility at the network, safety-based traffic assignment can be performed in order to increase the overall safety of the system.

Referring to the micro-level implications of this research, instantaneous crash risk of drivers is quantified considering vehicular movements, driver behavior, and roadway/environmental factors. The developed framework is capable to monitor driving environment and predict crash occurrence in real-time. The deep learning framework can be used in Advanced Driving Assistant Systems of low level of automated vehicles, where the human driver still controls the vehicle but receiving some additional information from the system, in order to generate feedback and warnings to the drivers while their risk of crash is high to take safe maneuver and prevent an incident. Furthermore, the developed statistical models can be used to develop strategies and countermeasures to reduce driving volatility.

[5] Hoseinzadeh, N., Arvin, R., Khattak, A. J., & Han, L. D. (2020). Integrating safety and mobility for pathfinding using big data generated by connected vehicles. *Journal of Intelligent Transportation Systems*, 1-17.

REFERENCES

Aaa, 2009. Aggressive driving: Research update. In: Association, A.A. ed. American Automobile Association Foundation for Traffic Safety.

Aarts, L., Van Schagen, I., 2006. Driving speed and the risk of road crashes: A review. Accident Analysis & Prevention 38 (2), 215-224.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Year. Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265-283.

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of safety research 34 (5), 597-603.

Abdel-Aty, M.A., Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention 32 (5), 633-642.

Administration, F.H., 2018. Highway statistics, 2016. Federal Highway Administration.

Agbelie, B.R., Roshandeh, A.M., 2015. Impacts of signal-related characteristics on crash frequency at urban signalized intersections. Journal of Transportation Safety & Security 7 (3), 199-207.

Aguero-Valverde, J., Jovanis, P.P.J.a.A., Prevention, 2006. Spatial analysis of fatal and injury crashes in pennsylvania. 38 (3), 618-625.

Ahlstrom, C., Kircher, K., Kircher, A., 2013. A gaze-based driver distraction warning system and its effect on visual behavior. IEEE Transactions on Intelligent Transportation Systems 14 (2), 965-973.

Ahmad, N., Ahmed, A., Wali, B., Saeed, T.U., Year. Exploring factors associated with crash severity on motorways in pakistan. In: Proceedings of the Proceedings of the Institution of Civil Engineers-Transport, pp. 1-10.

Ahmed, M.M., Ghasemzadeh, A., 2018. The impacts of heavy rain on speed and headway behaviors: An investigation using the shrp2 naturalistic driving study data. Transportation research part C: emerging technologies 91, 371-384.

Akamatsu, M., Sakaguchi, Y., Okuwa, M., Year. Modeling of driving behavior when approaching an intersection based on measured behavioral data on an actual road. In: Proceedings of the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 1895-1899.

Anastasopoulos, P.C., 2016. Random parameters multivariate tobit and zero-inflated count data models: Addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. Analytic methods in accident research 11, 17-32.

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

Anastasopoulos, P.C., Mannering, F.L., Shankar, V.N., Haddock, J.E., 2012. A study of factors affecting highway accident rates using the random-parameters tobit model. Accident Analysis & Prevention 45, 628-633.

Anon, 2008. National highway traffic safety administration. National motor vehicle crash causation survey: Report to congress. National Highway Traffic Safety Administration Technical Report DOT HS 811, 059.

Anon, 2011. Mobile phone use: A growing problem of driver distraction. In: Organization, W.H. ed.

Anon, 2018. Federal highway administration. Highway Statistics, 2016.

Arvin, R., Kamrani, M., Khattak, A., 2019a. The role of pre-crash driving instability in contributing to crash intensity using naturalistic driving data. Accident Analysis & Prevention 132.

Arvin, R., Kamrani, M., Khattak, A.J., 2019b. Examining the role of speed and driving stability on crash severity using shrp2 naturalistic driving study data. Transportation Research Board 98th Annual Meeting. Washington DC.

Arvin, R., Kamrani, M., Khattak, A.J., 2019c. How instantaneous driving behavior contributes to crashes at intersections: Extracting useful information from connected vehicle message data. Accident Analysis & Prevention 127, 118-133.

Arvin, R., Kamrani, M., Khattak, A.J., Rios-Torres, J., 2018. Safety impacts of automated vehicles in mixed traffic. Transportation Research Board 97th Annual Meeting Washington DC.

Arvin, R., Khademi, M., Razi-Ardakani, H., 2017. Study on mobile phone use while driving in a sample of iranian drivers. International journal of injury control and safety promotion 24 (2), 256-262.

Arvin, R., Khattak, A.J., Rios-Torres, J., Evaluating safety with automated vehicles at signalized intersections: Application of adaptive cruise control in mixed traffic. Transportation Research Board 98th Annual Meeting. Washington DC.

Azimi, G., Asgari, H., Rahimi, A., Jin, X., 2019. Investigation of heterogeneity in severity analysis for large truck crashes. Transportation Research Board. Washington D.C.

Azimi, G., Rahimi, A., Asgari, H., Jin, X., 2020. Severity analysis for large truck rollover crashes using a random parameter ordered logit model. 135, 105355.

Azizi, L., Sheikholeslami, A., 2012. Safety effect of u-turn conversions in tehran: Empirical bayes observational before-and-after study and crash prediction models. Journal of transportation engineering 139 (1), 101-108.

Baghbaderani, R.K., Hairong, Q., 2019. Incorporating spectral unmixing in satellite imagery semantic segmentation. IEEE International Conference on Image processing.

Baghbaderani, R.K., Wang, F., Stutts, C., Qu, Y., Qi, H., Year. Hybrid spectral unmixing in land-cover classification. In: Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 3009-3012.

Balusu, S.K., Pinjari, A.R., Mannering, F.L., Eluru, N., 2018. Non-decreasing threshold variances in mixed generalized ordered response models: A negative correlations approach to variance reduction. Analytic methods in accident research 20, 46-67.

Bao, J., Liu, P., Ukkusuri, S.V., 2019a. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accident Analysis & Prevention 122, 239-254.

Bao, J., Yu, H., Wu, J., 2019b. Short-term ffbs demand prediction with multi-source data in a hybrid deep learning framework. IET Intelligent Transport Systems.

Bartier, P.M., Keller, C.P., 1996. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (idw). Computers & Geosciences 22 (7), 795-799.

Basso, F., Basso, L.J., Bravo, F., Pezoa, R., 2018. Real-time crash prediction in an urban expressway using disaggregated data. Transportation Research Part C: Emerging Technologies 86, 202-219.

Beede, K.E., Kass, S.J., 2006. Engrossed in conversation: The impact of cell phones on simulated driving performance. Accident Analysis & Prevention 38 (2), 415-421.

Behnood, A., Mannering, F.L., 2017. The effects of drug and alcohol consumption on driver injury severities in single-vehicle crashes. Traffic injury prevention 18 (5), 456-462.

Behnood, A., Roshandeh, A.M., Mannering, F.L., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Analytic Methods in Accident Research 3, 56-91.

Bezzina, D., Sayer, J.J.R.N.D.H., 2014. Safety pilot model deployment: Test conductor team report. 812, 171.

Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences. Transportation Research Part B: Methodological 37 (9), 837-855.

Bonett, D.G., 2006. Confidence interval for a coefficient of quartile variation. Computational Statistics & Data Analysis 50 (11), 2953-2957.

Bozdogan, H., 1987. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. Psychometrika 52 (3), 345-370.

Cameron, A.C., Windmeijer, F.A., 1996. R-squared measures for count data regression models with applications to health-care utilization. Journal of Business & Economic Statistics 14 (2), 209-220.

Castro, M., Paleti, R., Bhat, C.R., 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. Transportation research part B: methodological 46 (1), 253-272.

Chandraratna, S., Stamatiadis, N., 2009. Quasi-induced exposure method: Evaluation of not-at-fault assumption. Accident Analysis & Prevention 41 (2), 308-313.

Chen, C., Xie, Y., 2016. Modeling the effects of aadt on predicting multiple-vehicle crashes at urban and suburban signalized intersections. Accident Analysis & Prevention 91, 72-83.

Choi, J.-S., Kim, H.-S., Kang, D.-W., Choi, M.-H., Kim, H.-S., Hong, S.-P., Yu, N.-R., Lim, D.-W., Min, B.-C., Tack, G.-R., 2013. The effects of disruption in attention on driving performance patterns: Analysis of jerk-cost function and vehicle control data. Applied ergonomics 44 (4), 538-543.

Chollet, F., 2015. Keras.

Ciresan, D.C., Meier, U., Masci, J., Gambardella, L.M., Schmidhuber, J., Year. Flexible, high performance convolutional neural networks for image classification. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence.

Curry, A.E., Hafetz, J., Kallan, M.J., Winston, F.K., Durbin, D.R., 2011. Prevalence of teen driver errors leading to serious motor vehicle crashes. Accident Analysis & Prevention 43 (4), 1285-1290.

Da Silva, A.R., Rodrigues, T.C.V., 2014. Geographically weighted negative binomial regression—incorporating overdispersion. Statistics Computing 24 (5), 769-783.

Das, A., Abdel-Aty, M.A., 2011. A combined frequency–severity approach for the analysis of rear-end crashes on urban arterials. Safety Science 49 (8-9), 1156-1163.

De Naurois, C.J., Bourdin, C., Bougard, C., Vercher, J.-L., 2018. Adapting artificial neural networks to a specific driver enhances detection and prediction of drowsiness. Accident Analysis & Prevention 121, 118-128.

De Naurois, C.J., Bourdin, C., Stratulat, A., Diaz, E., Vercher, J.-L., 2017. Detection and prediction of driver drowsiness using artificial neural network models. Accident Analysis & Prevention.

Deery, H.A., Love, A.W., 1996. The effect of a moderate dose of alcohol on the traffic hazard perception profile of young drink-drivers 1. Addiction 91 (6), 815-827.

Dingus, T.A., Year. Human factors applications in surface transportation. In: Proceedings of the Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2002 NAE Symposium on Frontiers of Engineering, pp. 39.

Dingus, T.A., Guo, F., Lee, S., Antin, J.F., Perez, M., Buchanan-King, M., Hankey, J., 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences 113 (10), 2636-2641.

Dingus, T.A., Hankey, J.M., Antin, J.F., Lee, S.E., Eichelberger, L., Stulce, K.E., Mcgraw, D., Perez, M., Stowe, L., 2015. Naturalistic driving study: Technical coordination and quality control.

Dingus, T.A., Hanowski, R.J., Klauer, S.G., 2011. Estimating crash risk. Ergonomics in Design 19 (4), 8-12.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B.J.a.A., Prevention, 2014. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. 70, 320-329.

Donmez, B., Liu, Z., 2015. Associations of distraction involvement and age with driver injury severities. Journal of safety research 52, 23-28.

Drews, F.A., Yazdani, H., Godfrey, C.N., Cooper, J.M., Strayer, D.L., 2009. Text messaging during simulated driving. Human factors 51 (5), 762-770.

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis & Prevention 41 (5), 1118-1123.

Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis & Prevention 40 (3), 1033-1054.

Engelberg, J.K., Hill, L.L., Rybar, J., Styer, T., 2015. Distracted driving behaviors related to cell phone use among middle-aged adults. Journal of Transport Health 2 (3), 434-440.

Eren, L., Ince, T., Kiranyaz, S., 2019. A generic intelligent bearing fault diagnosis system using compact adaptive 1d cnn classifier. Journal of Signal Processing Systems 91 (2), 179-189.

Esfahani, H.N., Song, Z., 2019. A new method for microsimulation model calibration: A case study of i-710.

Essa, M., Sayed, T., 2018. Traffic conflict models to evaluate the safety of signalized intersections at the cycle level. Transportation research part C: emerging technologies 89, 289-302.

Everitt, B., Skrondal, A., 2002. The cambridge dictionary of statistics Cambridge University Press Cambridge.

Fan, A.Z., Grant, B.F., Ruan, W.J., Huang, B., Chou, S.P., 2019. Drinking and driving among adults in the united states: Results from the 2012–2013 national epidemiologic survey on alcohol and related conditions-iii. Accident Analysis & Prevention 125, 49-55.

Farid, A., Abdel-Aty, M., Lee, J., 2018. A new approach for calibrating safety performance functions. Accident Analysis & Prevention 119, 188-194.

Farid, A., Abdel-Aty, M., Lee, J., 2019. Comparative analysis of multiple techniques for developing and transferring safety performance functions. Accident Analysis & Prevention 122, 85-98.

Feng, F., Bao, S., Sayer, J.R., Flannagan, C., Manser, M., Wunderlich, R., 2017. Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data. Accident Analysis & Prevention 104, 125-136.

Figlewski, S., 1994. Forecasting volatility using historical data.

Fitch, G., Lee, S., Klauer, S., Hankey, J., Sudweeks, J., Dingus, T., 2009. Analysis of lane-change crashes and near-crashes. US Department of Transportation, National Highway Traffic Safety Administration.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2003. Geographically weighted regression: The analysis of spatially varying relationships John Wiley & Sons.

Gao, J., Davis, G.A., 2017. Using naturalistic driving study data to investigate the impact of driver distraction on driver's brake reaction time in freeway rear-end events in car-following situation. Journal of safety research 63, 195-204.

Gao, Z., Liu, Y., Zheng, J., Yu, R., Wang, X., Sun, P., Year. Predicting hazardous driving events using multi-modal deep learning based on video motion profile and kinematics data. In: Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 3352-3357.

Gårder, P., 2006. Segment characteristics and severity of head-on crashes on two-lane rural highways in maine. Accident Analysis & Prevention 38 (4), 652-661.

Gargoum, S.A., El-Basyouny, K., 2016. Exploring the association between speed and safety: A path analysis approach. Accident Analysis & Prevention 93, 32-40.

Gers, F.A., Schraudolph, N.N., Schmidhuber, J., 2002. Learning precise timing with lstm recurrent networks. Journal of machine learning research 3 (Aug), 115-143.

Ghanim, M.S., Abu-Lebdeh, G., 2015. Real-time dynamic transit signal priority optimization for coordinated traffic networks using genetic algorithms and artificial neural networks. Journal of Intelligent Transportation Systems 19 (4), 327-338.

Ghasemzadeh, A., Ahmed, M.M., 2016. Crash characteristics and injury severity at work zones considering adverse weather conditions in washington using shrp 2 roadway information database. Transportation Research Board.

Ghasemzadeh, A., Ahmed, M.M., 2017. Drivers' lane-keeping ability in heavy rain: Preliminary investigation using shrp 2 naturalistic driving study data. Transportation Research Record: Journal of the Transportation Research Board (2663), 99-108.

Ghasemzadeh, A., Ahmed, M.M., 2018a. Exploring factors contributing to injury severity at work zones considering adverse weather conditions. IATSS Research.

Ghasemzadeh, A., Ahmed, M.M., 2018b. Exploring factors contributing to injury severity at work zones considering adverse weather conditions. IATSS Research.

Ghasemzadeh, A., Ahmed, M.M., 2018c. Utilizing naturalistic driving data for in-depth analysis of driver lane-keeping behavior in rain: Non-parametric mars and parametric logistic regression modeling approaches. Transportation research part C: emerging technologies 90, 379-392.

Ghasemzadeh, A., Hammit, B.E., Ahmed, M.M., Young, R.K., 2018. Parametric ordinal logistic regression and non-parametric decision tree approaches for assessing the impact of weather conditions on driver speed selection using naturalistic driving data. Transportation research record, 0361198118758035.

Ghiasi, A., Hussain, O., Qian, Z.S., Li, X., 2017. A mixed traffic capacity analysis and lane management model for connected automated vehicles: A markov chain method. Transportation Research Part B: Methodological 106, 266-292.

Gomes, M.J.T.L., Cunto, F., Da Silva, A.R., 2017. Geographically weighted negative binomial regression applied to zonal level safety performance models. Accident Analysis & Prevention 106, 254-261.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning MIT press.

Greene, W.H., 2002. Nlogit: Version 3.0: Reference guide Econometric software.

Greene, W.H., 2003. Econometric analysis Pearson Education India.

Hadayeghi, A., 2009. Use of advanced techniques to estimate zonal level safety planning models and examine their temporal transferability Citeseer.

Hadayeghi, A., Shalaby, A., Persaud, B., 2010a. Development of planning-level transportation safety models using full bayesian semiparametric additive techniques. Journal of Transportation Safety & Security 2 (1), 45-68.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010b. Development of planning level transportation safety tools using geographically weighted poisson regression. Accident Analysis & Prevention 42 (2), 676-688.

Haghighi, N., Liu, X.C., Zhang, G., Porter, R.J., 2018. Impact of roadway geometric features on crash severity on rural two-lane highways. Accident Analysis & Prevention 111, 34-42.

Halton, J.H.J.N.M., 1960. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. 2 (1), 84-90.

Hamdar, S.H., Mahmassani, H.S., Chen, R.B., 2008. Aggressiveness propensity index for driving behavior at signalized intersections. Accident Analysis & Prevention 40 (1), 315-326.

Hamdar, S.H., Qin, L., Talebpour, A., 2016. Weather and road geometry impact on longitudinal driving behavior: Exploratory analysis using an empirically supported acceleration modeling framework. Transportation research part C: emerging technologies 67, 193-213.

Hankey, J.M., Perez, M.A., Mcclafferty, J.A., 2016. Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets. Virginia Tech Transportation Institute.

Hanowski, R.J., Olson, R.L., Hickman, J.S., Dingus, T.A., 2006. The 100-car naturalistic driving study: A descriptive analysis of light vehicle-heavy vehicle interactions from the light vehicle driver's perspective. United States. Federal Motor Carrier Safety Administration.

Hartman, R.L., Brown, T.L., Milavetz, G., Spurgin, A., Pierce, R.S., Gorelick, D.A., Gaffney, G., Huestis, M.A., 2015. Cannabis effects on driving lateral control with and without alcohol. Drug and alcohol dependence 154, 25-37.

Hartman, R.L., Brown, T.L., Milavetz, G., Spurgin, A., Pierce, R.S., Gorelick, D.A., Gaffney, G., Huestis, M.A., 2016. Cannabis effects on driving longitudinal control with and without alcohol. Journal of Applied Toxicology 36 (11), 1418-1429.

Hauer, E., 2006. The frequency–severity indeterminacy. Accident Analysis & Prevention 38 (1), 78-83.

Hauer, E., 2009. Speed and safety. Transportation Research Record 2103 (1), 10-17.

Henclewood, D., Abramovich, M., Yelchuru, B., 2014. Safety pilot model deployment-one day sample data environment data handbook. USDOT Research and Technology Innovation Administrations 1.

Hilbe, J.M., 2011. Negative binomial regression Cambridge University Press.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., 2012a. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal processing magazine 29.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. science 313 (5786), 504-507.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012b. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9 (8), 1735-1780.

Horberry, T., Anderson, J., Regan, M.A., Triggs, T.J., Brown, J., 2006. Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. Accident Analysis & Prevention 38 (1), 185-191.

Horrey, W.J., Wickens, C.D., 2004. Driving and side task performance: The effects of display clutter, separation, and modality. Human factors 46 (4), 611-624.

Hoseinzadeh, N., Arvin, R., Khattak, A.J., Han, L.D., 2020. Integrating safety and mobility for pathfinding using big data generated by connected vehicles. Journal of Intelligent Transportation Systems, 1-17.

Hosking, S.G., Young, K.L., Regan, M.A., 2009. The effects of text messaging on young drivers. Human factors 51 (4), 582-592.

Hossain, M., 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.

Huang, H., Abdel-Aty, M., Darwiche, A., 2010. County-level crash risk analysis in florida: Bayesian spatial modeling. Transportation Research Record: Journal of the Transportation Research Board (2148), 27-37.

Huang, H., Hu, S., Abdel-Aty, M., 2014. Indexing crash worthiness and crash aggressivity by major car brands. Safety science 62, 339-347.

Huang, H., Siddiqui, C., Abdel-Aty, M., 2011. Indexing crash worthiness and crash aggressivity by vehicle type. Accident Analysis & Prevention 43 (4), 1364-1370.

Huang, Y., Sun, D.J., Zhang, L.-H., 2018. Effects of congestion on drivers' speed choice: Assessing the mediating role of state aggressiveness based on taxi floating car data. Accident Analysis & Prevention 117, 318-327.

Huber, P.J., 2005. Robust statistics John Wiley & Sons.

Hurvich, C.M., Simonoff, J.S., Tsai, C.L., 1998. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60 (2), 271-293.

Imprialou, M.-I.M., Quddus, M., Pitfield, D.E., Lord, D., 2016. Re-visiting crash–speed relationships: A new perspective in crash modelling. Accident Analysis & Prevention 86, 173-185.

Ismail, K., Sayed, T., Saunier, N., Lim, C., 2009. Automated analysis of pedestrian–vehicle conflicts using video data. Transportation research record 2140 (1), 44-54.

Jalayer, M., Shabanpour, R., Pour-Rouholamin, M., Golshani, N., Zhou, H., 2018. Wrong-way driving crashes: A random-parameters ordered probit analysis of injury severity. Accident Analysis & Prevention 117, 128-135.

Jamali, A., Wang, Y., 2017. Estimating pedestrian exposure for small urban and rural areas. Transportation Research Record: Journal of the Transportation Research Board (2661), 84-94.

Jeon, H., Lee, J., Sohn, K., 2018. Artificial intelligence for traffic signal control based solely on video images. Journal of Intelligent Transportation Systems 22 (5), 433-445.

Kamrani, M., Arvin, R., Khattak, A.J., 2018a. Analyzing highly volatile driving trips taken by alternative fuel vehicles. Transportation Research Board 97th Annual Meeting Washington DC, United States.

Kamrani, M., Arvin, R., Khattak, A.J., 2018b. Extracting useful information from basic safety message data: An empirical study of driving volatility measures and crash frequency at intersections. Transportation Research Record, 0361198118773869.

Kamrani, M., Arvin, R., Khattak, A.J., 2019. The role of aggressive driving and speeding in road safety: Insights from shrp2 naturalistic driving study data. Transportation Research Board 98th Annual Meeting. Washington DC.

Kamrani, M., Khattak, A.J., Li, T., 2018c. A framework to process and analyze driver, vehicle and road infrastructure volatilities in real-time. Transportation Research Board 97th Annual Meeting. Washington DC.

Kamrani, M., Wali, B., Khattak, A.J., 2017. Can data generated by connected vehicles enhance safety? Proactive approach to intersection safety management. Transportation Research Record: Journal of the Transportation Research Board (2659), 80-90.

Kaufman, L., Rousseeuw, P.J., 2009. Finding groups in data: An introduction to cluster analysis John Wiley & Sons.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv.

Kircher, K., Thorslund, B., 2009. Effects of road surface appearance and low friction warning systems on driver behaviour and confidence in the warning system. Ergonomics 52 (2), 165-176.

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J., 2006. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data.

Kludt, K., Brown, J., Richman, J., Campbell, J., 2006. Human factors literature reviews on intersections, speed management, pedestrians and bicyclists, and visibility.

Kluger, R., Smith, B.L., Park, H., Dailey, D.J., 2016. Identification of safety-critical events using kinematic vehicle data and the discrete fourier transform. Accident Analysis & Prevention 96, 162-168.

Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: An application of ordered probit models. Accident Analysis & Prevention 34 (3), 313-321.

Lambert-Bélanger, A., Dubois, S., Weaver, B., Mullen, N., Bedard, M., 2012. Aggressive driving behaviour in young drivers (aged 16 through 25) involved in fatal crashes. Journal of safety research 43 (5-6), 333-338.

Laude, J.R., Fillmore, M.T., 2015. Simulated driving performance under alcohol: Effects on driver-risk versus driver-skill. Drug alcohol dependence 154, 271-277.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.J.P.O.T.I., 1998. Gradient-based learning applied to document recognition. 86 (11), 2278-2324.

Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in florida. Accident Analysis & Prevention 37 (4), 775-786.

Lee, J.D., Young, K.L., Regan, M.A., 2008. Defining driver distraction.

Lee, M.L., Howard, M.E., Horrey, W.J., Liang, Y., Anderson, C., Shreeve, M.S., O'brien, C.S., Czeisler, C.A., 2016. High risk of near-crash driving events following night-shift work. Proceedings of the National Academy of Sciences 113 (1), 176-181.

Li, X., Yan, X., Wu, J., Radwan, E., Zhang, Y., 2016. A rear-end collision risk assessment model based on drivers' collision avoidance process under influences of cell phone use and gender—a driving simulator based study. Accident Analysis & Prevention 97, 1-18.

Li, Y., Ma, D., Zhu, M., Zeng, Z., Wang, Y., 2018. Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. Accident Analysis & Prevention 111, 354-363.

Liang, Y., Lee, J.D., Horrey, W.J., Year. A looming crisis: The distribution of off-road glance duration in moments leading up to crashes/near-crashes in naturalistic driving. In: Proceedings of the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 2102-2106.

Lin, L., He, Z., Peeta, S., 2018. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. Transportation Research Part C: Emerging Technologies 97, 258-276.

Lipari, R.N., Hughes, A., Bose, J., 2016. Driving under the influence of alcohol and illicit drugs. The cbhsq report. Substance Abuse and Mental Health Services Administration (US).

Liu, B.-S., Lee, Y.-H., 2005. Effects of car-phone use and aggressive disposition during critical driving maneuvers. Transportation Research Part F: Traffic Psychology and Behaviour 8 (4-5), 369-382.

Liu, J., Khattak, A.J., Wali, B., 2017. Do safety performance functions used for predicting crash frequency vary across space? Applying geographically weighted regressions to account for spatial heterogeneity. Accident Analysis & Prevention 109, 132-142.

Liu, M., Shi, J., 2019. A cellular automata traffic flow model combined with a bp neural network based microscopic lane changing decision model. Journal of Intelligent Transportation Systems 23 (4), 309-318.

Loader, C., 2006. Local regression and likelihood Springer Science & Business Media.

Loehlin, J.C., 2004. Latent variable models: An introduction to factor, path, and structural equation analysis Psychology Press.

Longerstaey, J., Spencer, M., 1996. Riskmetricstm—technical document. Morgan Guaranty Trust Company of New York: New York 51, 54.

Lord, D., 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.

Lu, N., Cheng, N., Zhang, N., Shen, X., Mark, J.W., 2014. Connected vehicles: Solutions and challenges. IEEE internet of things journal 1 (4), 289-299.

Lyles, R.W., Stamatiadis, P., Lighthizer, D.R., 1991. Quasi-induced exposure revisited. Accident Analysis & Prevention 23 (4), 275-285.

Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies 54, 187-197.

Manan, M.M.A., Várhelyi, A., Çelik, A.K., Hashim, H.H., 2017. Road characteristics and environment factors associated with motorcycle fatal crashes in malaysia. IATSS research.

Mannering, F.L., Shankar, V., Bhat, C.R.J.a.M.I.a.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. 11, 1-16.

Martin, T.L., Solbeck, P.A., Mayers, D.J., Langille, R.M., Buczek, Y., Pelletier, M.R., 2013. A review of alcohol-impaired driving: The role of blood alcohol concentration and complexity of the driving task. Journal of forensic sciences 58 (5), 1238-1250.

Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A., Hadi, M., 2019. Improved support vector machine models for work zone crash injury severity prediction and analysis. Transportation research record.

Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A., Hadi, M., 2020. Factors affecting injury severity in vehicle-pedestrian crashes: A day-of-week analysis using random parameter ordered response models and artificial neural networks. International Journal of Transportation Science Technology.

Motamedi, S., Wang, J.-H., 2016. The impact of text driving on driving safety. International Journal for Traffic Transport Engineering 6 (3).

Nakaya, T., Charlton, M., Lewis, P., Fortheringham, S., Brunsdon, C., 2012. Windows application for geographically weighted regression modeling. Ritsumeikan University, Kyoto, Japan.

Nakaya, T., Fotheringham, A.S., Brunsdon, C., Charlton, M., 2005. Geographically weighted poisson regression for disease association mapping. Statistics in medicine 24 (17), 2695-2717.

Nasr Esfahani, H., Arvin, R., Song, Z., Sze, N.-N., 2019. Prevalence of cell phone use while driving and its impact on driving performance, focusing on near-crash risk: A survey study in tehran. Journal of Transportation Safety & Security.

Nazari, F., Rahimi, E., Mohammadian, A.K., 2019. Simultaneous estimation of battery electric vehicle adoption with endogenous willingness to pay. eTransportation 1, 100008.

Neyens, D.M., Boyle, L.N., 2008. The influence of driver distraction on the severity of injuries sustained by teenage drivers and their passengers. Accident Analysis & Prevention 40 (1), 254-259.

Nezafat, R.V., Beheshtitabar, E., Cetin, M., Williams, E., List, G.F., 2018. Modeling and evaluating traffic flow at sag curves when imposing variable speed limits on connected vehicles. Transportation Research Record, 0361198118784169.

Nezafat, R.V., Sahin, O., Cetin, M., 2019. Transfer learning using deep neural networks for classification of truck body types based on side-fire lidar data. Journal of Big Data Analytics in Transportation 1 (1), 71-82.

Nhtsa, 2009. Traffic safety facts: Motorcycles. National Highway Traffic Safety Association, pp. 159.

Nhtsa, 2013. Preliminary statement of policy concerning automated vehicles. Washington, D.C.

Nhtsa, 2015. Traffic safety facts 2015: A compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system.

Nhtsa, 2017. Traffic safety facts 2016 data: Alcohol-impaired driving. In: Administration, N.H.T.S. ed. National Highway Traffic Safety Administration.

Nickkar, A., Jeihani, M., Sahebi, S., 2019a. Analysis of driving simulator sickness symptoms: Zero-inflated ordered probit approach. Transportation Research Record, 0361198119841573.

Nickkar, A., Jeihani, M., Sahebi, S., 2019b. Analysis of driving simulator sickness symptoms: Zero-inflated ordered probit approach. Transportation Research Record 2673 (4), 988-1000.

Nightingale, E., Parvin, N., Seiberlich, C., Savolainen, P.T., Pawlovich, M., 2017. Investigation of skew angle and other factors influencing crash frequency at high-speed rural intersections. Transportation Research Record: Journal of the Transportation Research Board (2636), 9-14.

Noland, R.B., Quddus, M.A., 2005. Congestion and safety: A spatial analysis of london. Transportation Research Part A: Policy and Practice 39 (7-9), 737-754.

O'donnell, C., Connor, D., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accident Analysis & Prevention 28 (6), 739-753.

Ordóñez, F., Roggen, D., 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors 16 (1), 115.

Osman, O.A., Hajij, M., Karbalaieali, S., Ishak, S., 2018. Crash and near-crash prediction from vehicle kinematics data: A shrp2 naturalistic driving study.

Osman, O.A., Hajij, M., Karbalaieali, S., Ishak, S., 2019. A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. Accident Analysis & Prevention 123, 274-281.

Overton, T.L., Rives, T.E., Hecht, C., Shafi, S., Gandhi, R.R., 2015. Distracted driving: Prevalence, problems, and prevention. International journal of injury control safety promotion 22 (3), 187-192.

Palaz, D., Collobert, R., 2015. Analysis of cnn-based speech recognition system using raw speech as input. Idiap.

Paleti, R., Eluru, N., Bhat, C.R., 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. Accident Analysis & Prevention 42 (6), 1839-1854.

Paolo Busardo, F., Pichini, S., Pellegrini, M., Montana, A., Fabrizio Lo Faro, A., Zaami, S., Graziano, S., 2018. Correlation between blood and oral fluid psychoactive drug concentrations and cognitive impairment in driving under the influence of drugs. Current neuropharmacology 16 (1), 84-96.

Papazikou, E., Quddus, M.A., Thomas, P., 2017. Detecting deviation from normal driving using shrp2 nds data.

Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A.K., 2019. Real-time accident detection: Coping with imbalanced data. Accident Analysis Prevention 129, 202-210.

Patil, S., Prediction of driving outcomes based on driver behaviour and roadway information.

Pei, X., Wong, S., Sze, N.-N., 2012. The roles of exposure and speed in road safety analysis. Accident Analysis & Prevention 48, 464-471.

Perez, M.A., Sudweeks, J.D., Sears, E., Antin, J., Lee, S., Hankey, J.M., Dingus, T.A., 2017. Performance of basic kinematic thresholds in the identification of crash and near-crash events within naturalistic driving data. Accident Analysis & Prevention 103, 10-19.

Pietrasik, T., 2018. Road traffic injuries. In: Organization, W.H. ed. World Health Organization, Geneva Switzerland.

Qu, X., Kuang, Y., Oh, E., Jin, S., 2014. Safety evaluation for expressways: A comparative study for macroscopic and microscopic indicators. Traffic injury prevention 15 (1), 89-93.

Quddus, M., 2013. Exploring the relationship between average speed, speed variation, and accident rates using spatial statistical models and gis. Journal of Transportation Safety & Security 5 (1), 27-45.

Rahimi, A., Azimi, G., Asgari, H., Jin, X., 2019. Clustering approach toward large truck crash analysis. Transportation Research Record.

Rahimi, E., Shamshiripour, A., Samimi, A., Mohammadian, A.K., 2020. Investigating the injury severity of single-vehicle truck crashes in a developing country. Accident Analysis & Prevention 137, 105444.

Rahman, M.S., Abdel-Aty, M., 2018. Longitudinal safety evaluation of connected vehicles' platooning on expressways. Accident Analysis & Prevention 117, 381-391.

Rahman, M.S., Abdel-Aty, M., Wang, L., Lee, J., 2018. Understanding the highway safety benefits of different approaches of connected vehicles in reduced visibility conditions. Transportation Research Record, 0361198118776113.

Rakauskas, M.E., Gugerty, L.J., Ward, N.J., 2004. Effects of naturalistic cell phone conversations on driving performance. Journal of safety research 35 (4), 453-464.

Regan, M.A., Lee, J.D., Young, K., 2008. Driver distraction: Theory, effects, and mitigation CRC Press.

Ren, H., Song, Y., Liu, J., Hu, Y., Lei, J., 2017. A deep learning approach to the prediction of short-term traffic accident risk. arXiv preprint arXiv:.09543.

Rumschlag, G., Palumbo, T., Martin, A., Head, D., George, R., Commissaris, R.L., 2015. The effects of texting on driving performance in a driving simulator: The influence of driver age. Accident Analysis & Prevention 74, 145-149.

Sadia, R., Bekhor, S., Polus, A., 2018. Structural equations modelling of drivers' speed selection using environmental, driver, and risk factors. Accident Analysis & Prevention 116, 21-29.

Saifuzzaman, M., Haque, M.M., Zheng, Z., Washington, S., 2015. Impact of mobile phone use on car-following behaviour of young drivers. Accident Analysis & Prevention 82, 10-19.

Savolainen, P.T., Mannering, F.L., Lord, D., 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.

Sermanet, P., Lecun, Y., Year. Traffic sign recognition with multi-scale convolutional networks. In: Proceedings of the IJCNN, pp. 2809-2813.

Shakouri, M., Ikuma, L.H., Aghazadeh, F., Punniaraj, K., Ishak, S., 2014. Effects of work zone configurations and traffic density on performance variables and subjective workload. Accident Analysis & Prevention 71, 166-176.

Sheng, S., Pakdamanian, E., Han, K., Kim, B., Tiwari, P., Kim, I., Feng, L., 2019. A case study of trust on autonomous driving. arXiv preprint arXiv:.11007.

Shi, Q., 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 58, 380-394.

Shi, X., Wong, Y.D., Li, M.Z.-F., Palanisamy, C., Chai, C., 2019. A feature learning approach based on xgboost for driving assessment and risk prediction. Accident Analysis & Prevention 129, 170-179.

Shinar, D., 2017. Traffic safety and human behavior Emerald Publishing Limited.

Shinar, D., Compton, R., 2004. Aggressive driving: An observational study of driver, vehicle, and situational variables. Accident Analysis & Prevention 36 (3), 429-437.

Shinar, D., Gurion, B., 2019. Crash causes, countermeasures, and safety policy implications. Accident Analysis & Prevention 125, 224-231.

Shou, Z., Di, X., 2018. Similarity analysis of frequent sequential activity pattern mining. Transportation Research Part C: Emerging Technologies 96, 122-143.

Simard, P.Y., Steinkraus, D., Platt, J.C., Year. Best practices for convolutional neural networks applied to visual document analysis. In: Proceedings of the Icdar.

Şimşekoğlu, Ö., Nordfjærn, T., Zavareh, M.F., Hezaveh, A.M., Mamdoohi, A.R., Rundmo, T., 2013. Risk perceptions, fatalism and driver behaviors in turkey and iran. Safety science 59, 187-192.

Sodhi, M., Reimer, B., Llamazares, I., 2002. Glance analysis of driver eye movements to evaluate distraction. Behavior Research Methods, Instruments, Computers 34 (4), 529-538.

Stamatiadis, N., Deacon, J.A., 1997. Quasi-induced exposure: Methodology and insight. Accident Analysis & Prevention 29 (1), 37-52.

Stavrinos, D., Jones, J.L., Garner, A.A., Griffin, R., Franklin, C.A., Ball, D., Welburn, S.C., Ball, K.K., Sisiopiku, V.P., Fine, P.R., 2013. Impact of distracted driving on safety and traffic flow. Accident Analysis & Prevention 61, 63-70.

Stipancic, J., Miranda-Moreno, L., 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), 43-54.

Stutts, J., Feaganes, J., Reinfurt, D., Rodgman, E., Hamlett, C., Gish, K., Staplin, L., 2005. Driver's exposure to distractions in their natural driving environment. Accident Analysis & Prevention 37 (6), 1093-1101.

Sun, J., 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.

Sun, J., Sun, J., 2016. Real-time crash prediction on urban expressways: Identification of key variables and a hybrid support vector machine model. IET intelligent transport systems 10 (5), 331-337.

Tang, K., Chen, S., Khattak, A.J., Pan, Y., 2019. Deep architecture for citywide travel time estimation incorporating contextual information. Journal of Intelligent Transportation Systems, 1-17.

Taylor, M.C., Lynam, D., Baruya, A., 2000. The effects of drivers' speed on the frequency of road accidents Transport Research Laboratory Crowthorne.

Taylor, T., Pradhan, A., Divekar, G., Romoser, M., Muttart, J., Gomez, R., Pollatsek, A., Fisher, D.L., 2013. The view from the road: The contribution of on-road glance-monitoring technologies to understanding driver behavior. Accident Analysis & Prevention 58, 175-186.

Teh, E., Jamson, S., Carsten, O., Jamson, H., 2014. Temporal fluctuations in driving demand: The effect of traffic complexity on subjective measures of workload and driving performance. Transportation research part F: traffic psychology and behaviour 22, 207-217.

Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accident Analysis & Prevention 72, 244-256.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica: journal of the Econometric Society, 24-36.

Train, K., 2000a. Halton sequences for mixed logit. Department of Economics, UCB.

Train, K., 2000b. Halton sequences for mixed logit.

Train, K.E., 2009. Discrete choice methods with simulation Cambridge university press.

Ukkusuri, S., Hasan, S., Aziz, H.A., 2011. Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Transportation research record 2237 (1), 98-106.

Ulak, M.B., Ozguven, E.E., Karabag, H.H., Ghorbanzadeh, M., Moses, R., Dulebenets, M., 2020. Development of safety performance functions for restricted crossing u-turn intersections. Journal of Transportation Engineering, Part A: Systems 146 (6), 04020038.

Vadeby, A., Forsman, Å., 2017. Changes in speed distribution: Applying aggregated safety effect models to individual vehicle speeds. Accident Analysis & Prevention 103, 20-28.

Verster, J.C., Wester, A.E., Goorden, M., Van Wieringen, J.-P., Olivier, B., Volkerts, E.R., 2009. Novice drivers' performance after different alcohol dosages and placebo in the divided-attention steering simulator (dass). Psychopharmacology 204 (1), 127-133.

Verstraete, A.G., Legrand, S.-A., Vandam, L., Hughes, B., Griffiths, P., 2014. Drug use, impaired driving and traffic accidents Publications Office of the European Union.

Victor, T., Dozza, M., Bärgman, J., Boda, C.-N., Engström, J., Flannagan, C., Lee, J.D., Markkula, G., 2015. Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk.

Victor, T.W., Harbluk, J.L., Engström, J.A., 2005. Sensitivity of eye-movement measures to in-vehicle task difficulty. Transportation Research Part F: Traffic Psychology Behaviour 8 (2), 167-190.

Wali, B., Ahmed, A., Ahmad, N., Year. An ordered-probit analysis of enforcement of road speed limits. In: Proceedings of the Proceedings of the Institution of Civil Engineers-Transport, pp. 225-234.

Wali, B., Khattak, A.J., Ahmad, N., 2019. Examining correlations between motorcyclist's conspicuity, apparel related factors and injury severity score: Evidence from new motorcycle crash causation study. Accident Analysis & Prevention 131, 45-62.

Wali, B., Khattak, A.J., Bozdogan, H., Kamrani, M., 2018a. How is driving volatility related to intersection safety? A bayesian heterogeneity-based analysis of instrumented vehicles data. Transportation Research Part C: Emerging Technologies 92, 504-524.

Wali, B., Khattak, A.J., Khattak, A.J., 2018b. A heterogeneity based case-control analysis of motorcyclist's injury crashes: Evidence from motorcycle crash causation study. Accident Analysis & Prevention 119, 202-214.

Wali, B., Khattak, A.J., Waters, J., Chimba, D., Li, X., 2018c. Development of safety performance functions: Incorporating unobserved heterogeneity and functional form analysis. Transportation Research Record, 0361198118767409.

Wali, B., Khattak, A.J., Waters, J., Chimba, D., Li, X., 2018d. Development of safety performance functions: Incorporating unobserved heterogeneity and functional form analysis. Transportation research record 2672 (30), 9-20.

Wang, F., Chen, Y., Guo, J., Yu, C., Stevenson, M., Zhao, H., 2019a. Middle-aged drivers' subjective categorization for combined alignments on mountainous freeways and their speed choices. Accident Analysis & Prevention 127, 80-86.

Wang, L., Abdel-Aty, M., Lee, J., Shi, Q., 2019b. Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors. Accident Analysis & Prevention 122, 378-384.

Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident Analysis & Prevention 38 (6), 1137-1150.

Wang, X., Abdel-Aty, M., Almonte, A., Darwiche, A., 2009. Incorporating traffic operation measures in safety analysis at signalized intersections. Transportation Research Record: Journal of the Transportation Research Board (2103), 98-107.

Wang, X., Abdel-Aty, M.J.a.A., Prevention, 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. 38 (6), 1137-1150.

Wang, X., Khattak, A.J., Liu, J., Masghati-Amoli, G., Son, S., 2015a. What is the level of volatility in instantaneous driving decisions? Transportation Research Part C: Emerging Technologies 58, 413-427.

Wang, X., Wang, T., Tarko, A., Tremont, P.J., 2015b. The influence of combined alignments on lateral acceleration on mountainous freeways: A driving simulator study. Accident Analysis & Prevention 76, 110-117.

Wang, Y., Zhang, W., 2017. Analysis of roadway and environmental factors affecting traffic crash severities. Transportation research procedia 25, 2119-2125.

Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. Statistical and econometric methods for transportation data analysis CRC press.

Wickens, C.D., Goh, J., Helleberg, J., Horrey, W.J., Talleur, D.A., 2003. Attentional models of multitask pilot performance using advanced display technology. Human factors 45 (3), 360-380.

Wu, Z., Sharma, A., Mannering, F.L., Wang, S., 2013. Safety impacts of signal-warning flashers and speed control at high-speed signalized intersections. Accident Analysis & Prevention 54, 90-98.

Xie, K., Li, C., Ozbay, K., Dobler, G., Yang, H., Chiang, A.-T., Ghandehari, M., Year. Development of a comprehensive framework for video-based safety assessment. In: Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), pp. 2638-2643.

Xie, K., Yang, D., Ozbay, K., Yang, H., 2018. Use of real-world connected vehicle data in identifying high-risk locations based on a new surrogate safety measure. Accident Analysis & Prevention.

Xu, C., Ji, J., Liu, P., 2018a. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. Transportation research part C: emerging technologies 95, 47-60.

Xu, M., Wu, J., Huang, L., Zhou, R., Wang, T., Hu, D., 2018b. Network-wide traffic signal control based on the discovery of critical nodes and deep reinforcement learning. Journal of Intelligent Transportation Systems, 1-10.

Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. Accident Analysis & Prevention 75, 16-25.

Yang, J., Nguyen, M.N., San, P.P., Li, X.L., Krishnaswamy, S., Year. Deep convolutional neural networks on multichannel time series for human activity recognition. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence.

Yang, K., Wang, X., Quddus, M., Yu, R., 2018. Deep learning for real-time crash prediction on urban expressways.

Yasmin, S., Eluru, N., Pinjari, A.R., Tay, R., 2014. Examining driver injury severity in two vehicle crashes–a copula based approach. Accident Analysis & Prevention 66, 120-135.

Ye, M., Osman, O.A., Ishak, S., Hashemi, B., 2017. Detection of driver engagement in secondary tasks from observed naturalistic driving behavior. Accident Analysis & Prevention 106, 385-391.

Young, K., Regan, M., Hammer, M., 2007. Driver distraction: A review of the literature. Distracted driving 2007, 379-405.

Young, K.L., Rudin-Brown, C.M., Patten, C., Ceci, R., Lenné, M.G., 2014. Effects of phone type on driving and eye glance behaviour while text-messaging. Safety science 68, 47-54.

Young, K.L., Salmon, P.M., 2012. Examining the relationship between driver distraction and driving errors: A discussion of theory, studies and methods. Safety science 50 (2), 165-174.

Yu, C.-Y., Xu, M., 2018. Local variations in the impacts of built environments on traffic safety. Journal of Planning Education Research 38 (3), 314-328.

Yu, R., Quddus, M., Wang, X., Yang, K., 2018. Impact of data aggregation approaches on the relationships between operating speed and traffic safety. Accident Analysis & Prevention 120, 304-310.

Yu, R., Zheng, Y., Abdel-Aty, M., Gao, Z., 2019. Exploring crash mechanisms with microscopic traffic flow variables: A hybrid approach with latent class logit and path analysis models. Accident Analysis & Prevention 125, 70-78.

Zeng, Q., Gu, W., Zhang, X., Wen, H., Lee, J., Hao, W., 2019. Analyzing freeway crash severity using a bayesian spatial generalized ordered logit model with conditional autoregressive priors. Accident Analysis & Prevention 127, 87-95.

Zeng, Q., Wen, H., Huang, H., Abdel-Aty, M., 2017a. A bayesian spatial random parameters tobit model for analyzing crash rates on roadway segments. Econometrica: journal of the Econometric Society 100, 37-43.

Zeng, Q., Wen, H., Huang, H., Pei, X., Wong, S., 2017b. A multivariate random-parameters tobit model for analyzing highway crash rates by injury severity. Accident Analysis & Prevention 99, 184-191.

Zhang, T., Chan, A.H., 2016. The association between driving anger and driving outcomes: A meta-analysis of evidence from the past twenty years. Accident Analysis & Prevention 90, 50-62.

Zhao, P., Lee, C., 2018. Assessing rear-end collision risk of cars and heavy vehicles on freeways using a surrogate safety measure. Accident Analysis & Prevention 113, 149-158.

Zhao, Z., Chen, W., Wu, X., Chen, P.C., Liu, J., 2017. Lstm network: A deep learning approach for short-term traffic forecast. IET Intelligent Transport Systems 11 (2), 68-75.

Zwillinger, D., Kokoska, S., 2000. Standard probability and statistical tables and formula. Chapman & Hall, Boca Raton.

VITA

Ramin Arvin was born in Isfahan, Iran. After obtaining his diploma from National Organization for Development of Exceptional Talents, he started his Bachelor of Science in Civil Engineering at Isfahan University of Technology. After completing his BSc, in the same year he admitted to Sharif University of Technology for his Master of Science program in Civil Engineering-Transportation and finished his study in 2016. After completing his master thesis titled “Identification and Assessment of Nomadic Transportation and Analysis of its Supply and Demand for Mainly Farmer Nomads”, Mr. Arvin went to the United States to pursue his PhD in Civil and Environmental Engineering-Transportation and Master in Statistics at The University of Tennessee, Knoxville. His research covers a wide range of topics related to Intelligent Transportation Systems, transportation safety, mobility, energy, and sustainable transportation. Specifically, Specifically, his research interests are big data science and its application in transportation domain, advanced statistical and econometrics models, and machine learning and deep learning methods and their application in ITS and safety. To date, Mr. Arvin has authored/co-authored 10 journal articles and 16 peer-reviewed conference papers.

As a doctoral student, Mr. Arvin received several scholarships and awards including Drs. Greg and Kay Reed Scholarship, Lifesavers Scholarship, ITS Tennessee Scholarship, TSITE Student Paper Competition Award, and Graduate Student Senate Travel Award. Mr. Arvin is an active ITE student member. He served as the president and secretory of ITE Student Chapter at The University of Tennessee, when awarded the second Student

Chapter in the Southern District and the first Student Chapter in the Tennessee District for the 2018-2019 school year. Mr. Arvin also served as a Senator for the Civil and Environmental Engineering Department in the Graduate Student Senate. Mr. Arvin also served as a reviewer for several academic journals and conferences.