INTEGRATING AND ANALYZING DRIVER, VEHICLE AND ROAD INFRASTRUCTURE VOLATILITIES USING CONNECTED AND INSTRUMENTED VEHICLES TECHNOLOGY

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Mohsen Kamrani
December 2018

Copyright © 2018 by Mohsen Kamrani
All rights reserved.

DEDICATION

This dissertation
is dedicated to my beloved wife,
Fedra, whose sacrificial care for me made
it possible for me to complete this work, and
to my parents and siblings, who have been a source of
encouragement and support to undertake and earn my Ph.D.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Professor Asad Khattak for his support, guidance and mentorship during my Ph.D. I also would like to thank Professors Hamparsum Bozdogan, Christopher Cherry and Dr. Chakraborty for serving on my dissertation committee.

I would like to express my appreciation to the U.S. National Science Foundation for funding my work through grant 1538139 mainly, as well as U.S. Department of Transportation through the Collaborative Sciences Center for Road Safety, a consortium led by The University of North Carolina at Chapel Hill in partnership with The University of Tennessee for supporting my research partially.

ABSTRACT

This dissertation proposes a framework on how to process and analyze the data available from the driver, the vehicle and the road infrastructure i.e. data streams in real-time. Particularly, it conceptualize measures of driver, vehicle and road infrastructure performance and process the volatilities in data streams from sensors. It also provides a framework for real-time identification of anomalies, linking them with alerts, warnings and control assists. We explore different measures of driving volatility used to explain crash frequencies at intersections through developing a unique database that integrates intersection crash and inventory data with real-world Basic Safety Messages logged by connected vehicles. We introduce location-based volatility (LBV) as a proactive safety measure, quantifying variability in instantaneous driving decisions at intersections. Such an analysis is fundamental towards proactive intersection safety management. In addition, Markov Decision Process (MDP) framework is used to learn observed behavior by analyzing instantaneous driving decisions of acceleration, deceleration, and maintaining constant speed. Moreover, the developed measures of volatilities are applied to speed profiles from the Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS) to come up with the most accurate crash-prediction model with used for real-time driving assist warning generation. Finally**,** by incorporating the data from the driver, vehicle and infrastructure into the analysis, the impact of detailed pre-crash driving behavior and recently developed measures of driving volatility on crash and near-crash risks is investigated. The knowledge gained from studying individual driving behaviors can be used to generate alerts and warnings for the driver of the host vehicle and to be passed via connected vehicle technology with the purpose of improving safety. The methods applied in this dissertation can form a foundation for human driver behavior prediction and personally revealed choice extraction. They also can help proactively identify locations with high levels of driving volatility (i.e., hot spots where crashes are waiting to happen) as candidates for safety improvements. Proactive warnings and alerts can be generated about potential hazards and transmitted to drivers via connected vehicle technologies such as road-side equipment, increasing drivers' situational and safety awareness.

TABLE OF CONTENTS

INTRODUCTION 1
CHAPTER I A Framework to Process and Analyze Driver, Vehicle and Road Infrastructure Volatilities in Real-time 4
Abstract 5
Introduction 5
Literature Review 6
Methodology 8
Vehicle Volatility 8
Driver Volatility 10
Road Infrastructure Volatility 11
Exploring Relationships 13
Real-Time Processing of Data and Events 13
Data 14
Initial Results 16
Discussion 16
Advanced Processing of the Events 19
Limitation 19
Conclusion 19
Acknowledgement 20
CHAPTER II Extracting Useful Information from Basic Safety Message Data: An Empirical Study of Driving Volatility Measures and Crash Frequency at Intersections 21
Abstract 22
Introduction 22
Literature Review 23
Methodology 24
Measures of Driving Volatility 25
Two Levels for Calculating Volatility 28
Modeling Approach 30
Data 32
Results 32
Descriptive Statistics 32
Correlations 32
Modeling Results and Discussion 36
Limitations 37
Conclusion 38
Acknowledgment 39

CHAPTER III Can Data Generated by Connected Vehicles Enhance Safety? A Proactive Approach to Intersection Safety Management 40
Abstract 41
Introduction 41
Literature Review 42
Research Objective and Contribution 44
Methodology 44
Conceptual Framework 44
Location Based Volatility 45
Calculation of LBV 46
Modeling Approach 47
Data 48
Data Accuracy 49
Results 49
Descriptive Statistics 49
Modeling Results 51
Discussion 53
Limitations 57
Conclusions 57
Acknowledgement 59
CHAPTER IV Applying Markov Decision Process to Understand Driving Decisions Using Basic Safety Messages Data 60
Abstract 61
Introduction 61
Methodology 63
Markov Decision Processes 63
MDP and Reinforcement Learning Distinction 64
MDP in Driving Maneuvers Context 66
Data 73
Results 73
Discrete Choice Model 73
Personally Revealed Choices (PRC's) 75
Comparison of PRC's and Real-world Observed Choices 80
Limitations 81
Conclusion 82
Acknowledgement 82
CHAPTER V Harnessing Instrumented Vehicles Data for Crash Prediction: A Comparison of Machine Learning Classifiers 83
Abstract 84

Introduction 84
Methodology 86
Measures of Driving Volatility 86
Types of Driving Volatility 88
Classification Methods 89
Data 94
Results 96
Descriptive Statistics 96
Misclassification Rate 96
Ga Logistic 99
Contributing Measures of Driving Volatility to Crashes 100
Limitations 102
Conclusion 102
Acknowledgement 103
CHAPTER VI The Role of Aggressive Driving and Speeding in Road Safety: Insights from SHRP2 Naturalistic Driving Study Data 104
Abstract 105
Introduction 105
Literature Review 106
Methodology 107
Measures of Driving Volatility 107
Exclusion of Evasive Maneuver Duration 108
Modeling Approach 109
Marginal Effects 110
Data 111
Data Pre-processing and Recoding 111
Results 111
Descriptive Statistic 111
Modeling Results 116
Limitations 126
Conclusions 126
Acknowledgement 127
CONCLUSION 128
REFERENCES 131
VITA 144

LIST OF TABLES

Table 1. ECG parameters with their normal ranges 12
Table 2. Descriptive statistics 15
Table 3. Driver and vehicle volatility on four trips 16
Table 4. Summary of measures for driving volatility quantification 26
Table 5. Descriptive statistics of measures of volatilities (n = 116) 34
Table 6. Fixed and random parameters Poisson model results 36
Table 7. Description of key variables and descriptive statistics 52
Table 8. Modeling results of fixed- and random-parameter Poisson regressions 56
Table 9. Descriptive statistics of the selected trips (n=120) 74
Table 10. Descriptive statistics of MNL estimated coefficients (n=120) 76
Table 11. Measures of driving volatility descriptive statistics (n = 9372 trips) 97
Table 12. Normal confidence interval of the error mean for different methods (n training = 8,000 & n testing = 1,153) 97
Table 13. Genetic Algorithm settings for logistic regression subset selection 100
Table 14. Fixed and random parameter logistic model for crash prediction 101
Table 15. List of recoded variables and their final Categories 112
Table 16. Descriptive statistics of the categorical variables in NDS data 114
Table 17. Descriptive statistics of calculated variables from NDS data 117
Table 18. The speed volatility model 120
Table 19. Deceleration volatility model 121
Table 20. Crash and near-crash model 122
Table 21. Marginal effects 125

LIST OF FIGURES

Figure 1. Chapters' theme and overall scope 3
Figure 2. Schematic diagram from data origination to alert and warning 9
Figure 3. A normal electrocardiogram (ECG) waveform 12
Figure 4. Element-based sliding window with size 100 for streaming data 15
Figure 5. On-going data collection for the study 15
Figure 6. Plots of incoming streaming data from driver (left: blood pressure and heart ECG) and the vehicle (right: acceleration and speed) 17
Figure 7. Graphical queries used in this study (in Odysseus software) 18
Figure 8. Speed bins for calculating acceleration thresholds at various speeds using BSMs data 28
Figure 9. Measures of driving volatility at intersections 29
Figure 10. Left: Ann Arbor map created from BSM data, Right: Data preparation steps 33
Figure 11. Correlations between crash frequency and measures of volatilities 35
Figure 12. Expected-actual number of crashes for fixed and random parameter models 37
Figure 13. a) Four quadrants used to calculate coefficients of variation (=σ/μ) for each intersection, b) Plot of used data (left)/ Histogram of lateral acceleration (right) 50
Figure 14. Mean-expected over actual number of crashes for fixed and random-parameter Poisson models (Green: fixed parameter models; Red: random parameter models) 54
Figure 15. Known hotspots and spots where crashes are waiting to happen 57
Figure 16. Framework of Markov Decision Process (A), reinforcement learning (B) and the paper (C) 65
Figure 17. Examples of the driving states and actions taken by drivers 68
Figure 18. Definition of states based on number of objects and distance to the front object 69
Figure 19. MDP structure used in the study 69
Figure 20. Grid world toy example 72
Figure 21. Transition probabilities of the toy example (the transition probability of action "stay" is an identity matrix) 72
Figure 22. Safety Pilot Model Deployment Site, Ann Arbor, Michigan *(129)* 74
Figure 23. The data preparation steps 74
Figure 24. Personally Revealed Choices given specific states (left: native approach, right: MDP) (n=120 trips). Example of reading the figure (right): in state one, 49% of drivers' PRC was acceleration, 23% was deceleration and 28% constant speed. . 77
Figure 25. Distribution of average number of objects surrounding the host vehicle (n=120 trips) 78

Figure 26. Personally revealed choices made on non-crowded trips (n=32) 79
Figure 27. Personally revealed choices made on semi-crowded trips (n=44) 79
Figure 28. Personally revealed choices in crowded trips (n=44) 80
Figure 29. Percentage distribution of PRC among drivers (n=120) 81
Figure 30. Comparison of baseline with crash/near-crash driving profiles 89
Figure 31. Left: classification of events in the data; Right: Data preparation steps 95
Figure 32. Distribution of second used for volatilities calculation 95
Figure 33. Misclassification rate of different methods (Tr: train, N = 8,000, Ts: test, N = 1,153) 98
Figure 34. Misclassification rate of crashes (recall) for different methods (Tr: train, N = 8,000, Ts: test, N = 1,153) 98
Figure 35. Receiver Operating Curves for test dataset 99
Figure 36. Misclassification error for different fitness criteria (tr: train, n = 8,000, ts: test, n = 1,153) 101
Figure 37. Comparison of baseline with crash/near crash driving profiles 109
Figure 38. Conceptual framework of the model 110
Figure 39. Distribution of evasive maneuver (left) and second used for volatilities calculation 117
Figure 40. Final model structure 118

INTRODUCTION

The term system might have different meanings in different disciplines. In particular, any *physical* system whose output comes from the interaction of its elements can fail, if one of its elements does not perform normally. The output of a system is the ultimate physical goal for which the system has been created. For instance, if we consider the human body as a system, well-being or health can be considered as its ultimate goal. All the organs or subsystems in the body interact with each other to keep the whole system work in a healthy manner. Although the body has its own ways of monitoring and generating alarms (i.e. symptoms) to notify the person of the problem, physicians are able to find out about the organs' malfunctions by performing preventive and periodic tests. Another example is a nuclear reactor system whose ultimate goal is to produce energy in a safe manner. As opposed to human body system, this system is built by the human and therefore, the builder should have devised ways to ensure safe performance of its elements. Due to the criticality of the failure in a nuclear reactor, the operators and computers constantly monitor its components.

A transportation system, with the ultimate goal of moving people and goods from point A to point B safely, has the subsystems of the infrastructure, the vehicle (or any element with the task of moving) and the driver. If any of these subsystems fails to be in the safe status, the consequence is an incident or a crash. It is ironic that while in majority of crashes, the driver is the main contributing factor (i.e. fails to act in a safe manner), there have not been done enough to monitor them, and the focus of transportation engineers, practitioners and researchers have been more on the infrastructure and the vehicles. That being said, the proposed framework and the research pursued and presented in this dissertation try to fill a portion of the existing gap. In particular, each chapter presents the results of peer-reviewed papers addressing and revolving around different ways of extracting useful information from streams of data with the ultimate goal of reducing crashes proactively. Of course, due to the limitation in the data, making simplifying assumptions or deviation from presented framework is sometimes inevitable. In summary:

Chapter I proposes a framework on how to process and analyze the data available from the driver, the vehicle and the road infrastructure i.e. data streams in real-time. Particularly, it conceptualize measures of driver, vehicle and road infrastructure performance and process the volatilities in data streams from sensors. It also provides a framework for real-time identification of anomalies, linking them with alerts, warnings and control assists.

Chapter II explores different measures of driving volatility used to explain crash frequencies at intersections. Given that driver behavior is the main contributing factor in crashes, findings from this study are beneficial in two ways. First, they can help

proactively identify locations with high levels of driving volatility but might not have many crashes (i.e., hot spots where crashes are waiting to happen) as candidates for safety improvements. Second, reduction of driving volatility at high crash locations can reduce future crashes.

Chapter III, by developing a unique database that integrates intersection crash and inventory data with more than 65 million real-world Basic Safety Messages logged by 3,000 connected vehicles, provides a more complete picture of operations and safety performance of intersections. The chapter introduces location-based volatility (LBV) as a proactive safety measure and a leading indicator of safety, which quantifies variability in instantaneous driving decisions at intersections. The study explores how variability in driving can be mapped to historical safety outcomes such as crashes at specific locations. Such an analysis is fundamental towards proactive intersection safety management. Proactive warnings and alerts can be generated about potential hazards at specific intersections and transmitted to drivers via connected vehicle technologies such as road-side equipment; these can in turn increase drivers' situational and safety awareness, and help them pursue safer driving at dangerous intersections.

Chapter IV, by studying high-resolution connected vehicle data available from the Michigan Safety Pilot Program, applies Markov Decision Process (MDP) framework to learn observed behavior by analyzing instantaneous driving decisions. Personally Revealed Choices are explored in terms of acceleration, deceleration, and maintaining constant speed decisions through analyzing detailed data from 120 trips. The methods applied in the chapter can potentially form a foundation for human driver behavior prediction and personally revealed choice extraction using field-collected empirical data. Moreover, knowledge from studying individual driving behaviors can be used to generate alerts and warnings for the driver of the host vehicle and to be passed via connected vehicle technology with the purpose of improving safety.

Chapter V applies the developed measures from chapter II to speed profiles from the Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS) to come up with the most accurate crash-prediction model at the disaggregate level (i.e., individual drivers) with the purpose of being used for real-time driving assist warning generation. Given the criticality of accuracy in this context, several machine learning methods including neural networks are attempted and the results are compared to select the best predicting model. The importance of results lies in the fact that the features used to classify normal (baseline) and crash events are solely extracted from speed profiles. It means such features can be quantified constantly in real-time using sliding timed windows as inputs in the model for crash prediction. Proper warnings can be issued to the driver for behavior correction and crash avoidance purposes.

Chapter VI, by harnessing the rich information available from naturalistic driving study data, this paper studies the impact of detailed pre-crash driving behavior and

recently developed measures of driving volatility on crash and near-crash risks. Building on previous efforts in developing of driving volatility measures, highly correlated measures with crash risk were identified. Then driving behaviors contributing to both driving volatility and crash risk are explored. This chapter incorporates the data from the driver, vehicle and infrastructure into the analysis.

Figure 1 shows the theme of each chapter in term of used data and the overall scope.

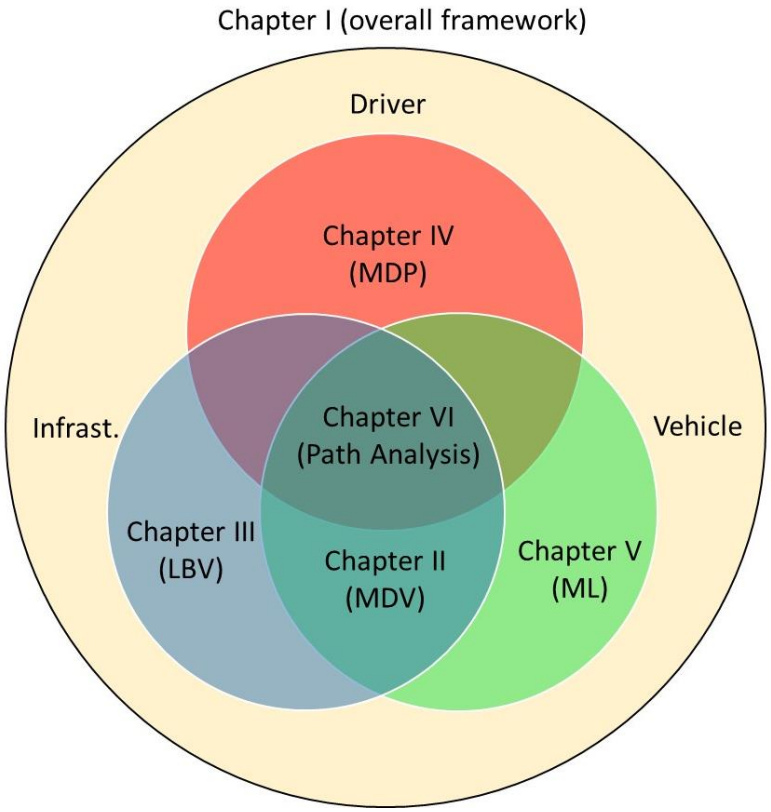


Figure 1. Chapters' theme and overall scope

CHAPTER I
A FRAMEWORK TO PROCESS AND ANALYZE DRIVER, VEHICLE AND ROAD INFRASTRUCTURE VOLATILITIES IN REAL-TIME

A version of this chapter was originally submitted by Mohsen Kamrani, Tenlong Li and Asad J. Khattak to Transportation Research Board 97th Annual Meeting for Presentation:

Kamrani, M., Khattak, A. J., & Li, T. (2018). *A Framework to Process and Analyze Driver, Vehicle and Road infrastructure Volatilities in Real-time* (No. 18-00979).

Abstract

Volatile driving characterized by hard accelerations and braking, sharp turns and weaving through traffic, can result in crashes and higher energy/environmental costs. With the ubiquity of sensors and increasing computational resources, it has become easier to observe volatility and extract useful information from data streams coming in from diverse sources. Sensors can now be used to monitor the condition of the driver (e.g., biometrics, brain waves, and eye movements), the performance of the vehicle (e.g., speed, acceleration, and turning movements) and the changes in the roadway environment (e.g., road surface and lighting conditions). Specifically, real-time driving event processing is proposed for learning from the streams and taking proactive measures when needed to help drivers avoid crashes and save fuel. Examples of data streams along with several parameters from which alerts and warnings can be generated are presented and discussed, with a focus on helping drivers or system operators identify anomalies, spot red flags, provide timely alerts and early warnings and tackle issues as they arise. For example, abnormal heart rhythm of a driver may indicate a heart attack or an eye off the road may indicate distraction and inattention. These situations will pose a high risk to the driver, occupants, surrounding walkers, bicyclists, and vehicles. The objectives of this paper are to conceptualize measures of the driver, vehicle, and road infrastructure volatility and to propose a framework for processing and implementing an early warning system that is based on the concept of volatility in data streams emanating from the driver, vehicle, and roadway.

Introduction

The emergence of Connected and Automated Vehicles (CAVs) has created the opportunity to dramatically improve safety, mobility, fuel use and the environment. Part of the opportunity lies in analyzing detailed data from various sensors and provide actionable information to drivers and transportation system managers. Driving volatility is reflected in decisions such as hard acceleration and braking, sharp lane changing, and weaving through traffic. To this end, researchers have analyzed the observations from the vehicle kinematics (1-3). Increasingly, there is an opportunity to expand the concept of volatility to individual drivers (by monitoring their biometrics, brain activity, and eye movements) as well as roadway environment.

Although the conventional reactive methods provide beneficial information on the different correlates of crash frequency and injury severity, preventing crashes requires comprehensive real-time monitoring of the drivers, vehicles, and roadways. Specifically, development of proactive alerts, warnings and control assists, demands two important tasks. First, to consider information from the triangle of drivers, vehicles, and roadways simultaneously. Second, the fusion and processing of the data streams available from sensors in real-time (1).

The task of simultaneous consideration of the driver, the vehicle and road infrastructure in the analyses, requires defining key measures. Monitoring the right measures can help detect volatility from the normal state, i.e., anomalies. For instance, strong and frequent fluctuations in the acceleration of a vehicle can be monitored and processed for the purpose of alerting or warning the driver, if the volatility in acceleration exceeds a threshold. In previous studies (2-4), we use the term “driving volatility” for the fluctuation, dispersion, variation and similar elements observed from vehicle kinematics. In this study, we extend the concept of volatility to the driver and the roadway, i.e. referring to them as “driver volatility” and “road infrastructure volatility” respectively.

Apart from the conceptualization of the driver and road infrastructure volatilities, this paper proposes a framework on how to process and analyze the data available from the driver, the vehicle and the road infrastructure i.e. data streams in real-time. In this framework, the anomalies are considered as events. A single event or a combination of events will trigger alerts, warnings or control assists for the CAV. We take advantage of developments in processing and learning from data streams coming from the vehicle or the driver. This study has two main objectives:

1. To conceptualize measures of driver, vehicle and road infrastructure and process the volatilities in data streams.
2. To provide a framework for real-time identification of anomalies, linking them with alerts, warnings and control assists that can help drivers improve safety, energy efficiency and environmental impacts.

Literature Review

Since there is a considerable number of papers in the data stream and event processing in disciplines others than transportation, we will discuss some of the ones that explored the processing of data related to drivers’ bio-signals, vehicle, and the road infrastructure.

Stress level detection of the drivers has captured a significant amount of attention. For instance, Rigas et al. (*4*) applied a Bayesian networks method to detect the driver’s stress level through processing the data from the driver and the vehicle. In another study (*5*), the level of drivers’ stress was detected using a combination of physiological data, questionnaire, and video coding. A continuous calculation of the correlation between the stress level and extracted features revealed that heart rate and skin conductivity could be

used as the drivers' stress level metrics. Apart from the detection accuracy, fast and non-intrusive detection of the stress level are crucial in making the detection methods practical in the real world. First, they need to be fast (*6*) and second they have to be nonintrusive (*7; 8*).

Drivers' stress level, fatigue, and drowsiness are among the factors that affect driving performance. Several studies are focused on detecting the level of these factors. Sun et al. (*8*) proposed a system to detect drivers' alertness using electrocardiogram (ECG), heart rate variability (HRV) and eye activities. According to their study results low and high frequency component of HRV and the duration of eyes blinking are significant measures for drowsiness and alertness. In a different approach, Ji et al. (*9*) presented a nonintrusive vision based system to predict drivers' fatigue from only the visual features such as eye, head and hand movements. Similarly, Lee et al. (*10*) developed an Android based system to monitor drivers' level of alertness (drowsiness) by fusing bio-signals and facial features and using the dynamic Bayesian network. Rigas et .al (*11*), however, used some data from the environment along with bio-signal data from the physiological signals and facial expression to detect both stress and fatigue. Their experiment found an association between fatigue and driving performance.

Drivers' emotional states and distraction are also among the studied factors. A speed-based recognition system was proposed by Tawari and Trivedi (*12*) to classify the driver's emotions solely to positive, neutral and negative. Likewise, five emotional states of being happy, sad, angry, fatigue and neutral were studied by Wang et al. (*13*) for multiple drivers through applying a latent variable algorithm. However, considering fatigue as one of the emotional states is arguable.

Driver's danger level, distraction, and unsafe behaviors are explored in the number of studies as well. Jabon et al. (*14*) attempted to predict unsafe driving behavior through facial expression analysis. In this study, time and frequency statistics were used to extract and select facial features to predict accidents before they happen. The assumed causal relationship between the facial expression and incidents is arguable. When an incident is about to happen, it affects several elements including the facial expression. In other words, an abnormal facial expression before a crash seems to be happening at the perception time when it might be late for the driver to act properly to avoid the crash. Wu et al. (*15*) used only vehicle kinematics to detected driving decisions (rather than behavior) in terms of acceleration, brake, lane change and turns, while Lee et al. (*16*) fused the data from both vehicle kinematics and the driver's bio-signal to generate safety warnings to the driver. Eye movement patterns combined with the standard deviation of the steering wheel and lane positions are used to detect driver's distraction in the study. In addition, it was shown that support vector machine (SVM) outperformed the logistic regression for classification of the driver distraction. Speaking of methods, apart from classification approaches such as SVM and logistic regressions, other learning

approaches such as Hidden Markov Model (HMM) and Reinforcement Learning (RL) methods are used to detect the deviation of driving behavior from the normal state. For instance, Lee et al. (*17*) used HMM as a pattern recognition tool to monitor the correlation between the driving path and the drivers' sight line through imaging from the road and the drivers' face. HMM also was used in (*18*) to predict driving danger level by processing the driver's physiological data and vehicle dynamic behaviors. The study compared the performance of HMM with RL and concluded that the RL method has better prediction capability.

From the reviewed literature, it appears that simultaneous processing of data from the triangle of the driver, the vehicle and the road infrastructure is underexplored. The study, in addition to proposing a framework to consider the road infrastructure data along with the driver and the vehicle data, conceptualizes the driver, vehicle, and road infrastructure volatilities to detect anomalies in a more objective manner.

Methodology

Vehicle volatility has been explored more widely in the literature, and it is directly related to unsafe driving behavior, fuel consumption, and emissions. The volatility in road infrastructure comes from changes in road surfaces, weather, lighting conditions, etc. and it is often not controllable by transportation system managers. Such volatility can affect drivers' decisions and consequently vehicle performance. Also, the state of the driver can impact vehicle volatility. For example, the emotional state of a driver (e.g., depressed, angry, disturbed) can influence how they drive a vehicle. In this sense, vehicle performance and volatility is often impacted by the state of the driver while the road infrastructure (and the environment) is not affected by the actions of the driver and the performance of the vehicle. That said, vehicle and driver volatilities go hand in hand, and they can be highly correlated. Figure 2 shows the overall flow of the information originating from different devices and sensors to last step which is the generation of warnings and alerts. In the following, the main elements of the figure are discussed.

Vehicle Volatility

Vehicle volatility can be captured by monitoring vehicles kinematics that includes acceleration, speed and vehicular jerk (both in longitudinal and lateral directions). Different measures of vehicle volatility can be used for capturing the range, variation, and dispersion in the streams of data coming from the vehicle. Specifically, range, standard deviation, the coefficient of variation, mean absolute deviation, the proportion of outliers (beyond mean plus predefined quantiles) and even the function used in finance to calculate the stock price volatility (*19*), are the examples of functions that can be used (*20*).

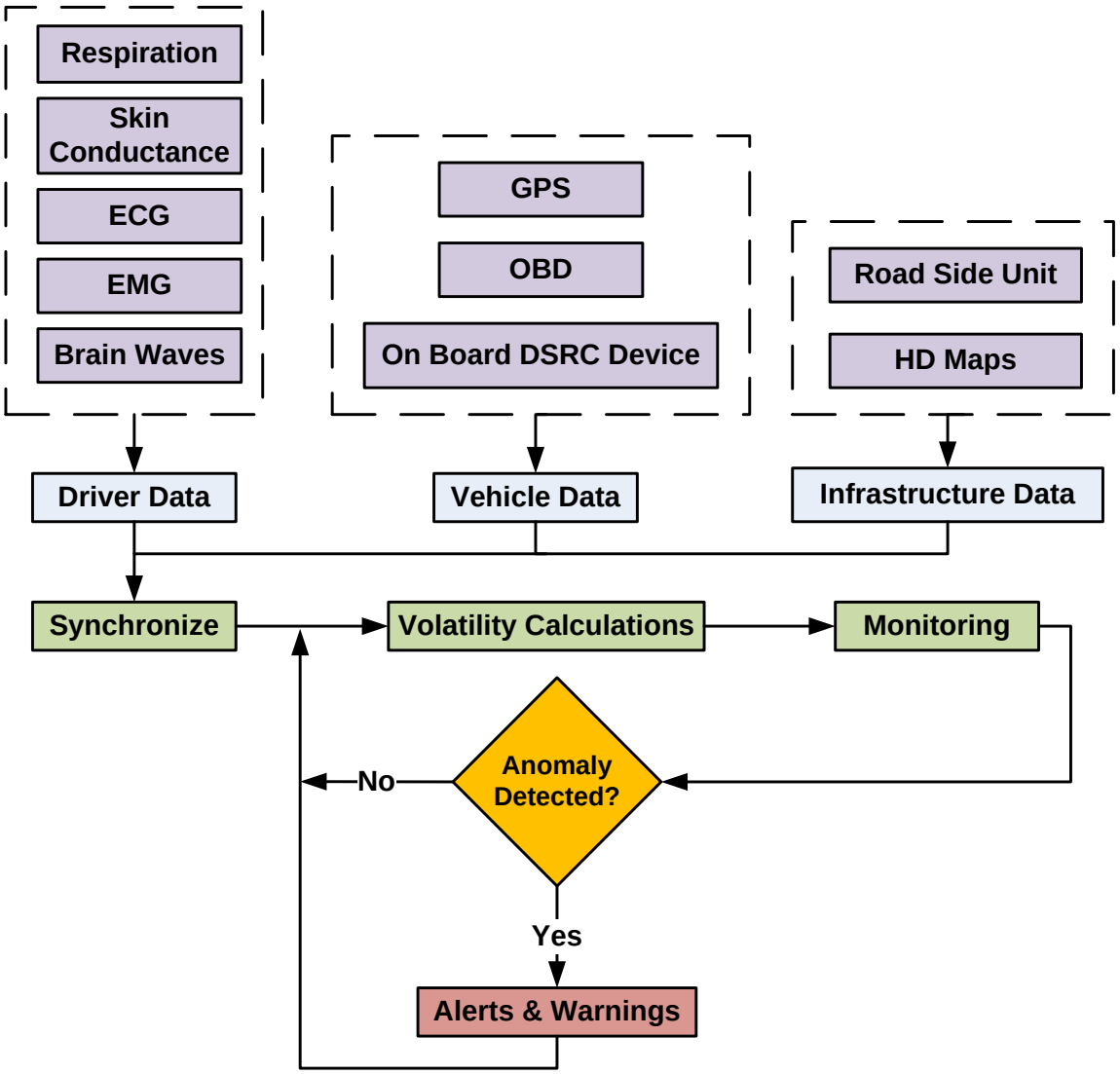


Figure 2. Schematic diagram from data origination to alert and warning

The question is what measures better capture the vehicle volatility (which is a reflection of driver's short-term driving decisions) that are associated with safety outcomes, energy consumption, and emissions. For demonstration, we use the historical financial volatility (HFV) as a measure of volatility. This measure is appropriate because time-series nature of the data is accounted for in calculating real-time vehicle volatility. Specifically, HFV of time-series data can be obtained using the following equation:

$$V_f = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{r_i} - \bar{r})^2} \quad \text{from } i = 1 \text{ to } n \tag{1}$$

where

$$\mathrm{r_i} = \ln\left(\frac{x_i}{x_{i-1}}\right) * 100 \tag{2}$$

and x_i and x_{i-1} are the current and previous observations (in this study instantaneous vehicular speeds) respectively and *ln* is the natural logarithm.

Driver Volatility

Driver volatility relates to deviations from the norm for drivers' biometrics (such as their heartbeat rate and blood pressure), facial expressions, and eye movements. (These measures are not the same as driving volatility, which, for the most part, can be considered equivalent to vehicle volatility.) A key aspect of these measures is to identify volatility or anomalies in a driver's state. For instance, a driver who is under time pressure or stressed out may exhibit volatile driving compared with when the driver is not stressed out. Drowsiness, fatigue, stress, being overloaded with multiple tasks, being drunk or under special medical, emotional or psychological conditions are examples that can increase the volatility of the driver, some of which, can be detected by processing the physiological data stream, facial expressions, and eye movements through image processing. With modern sensors, a variety of data streams such as skin conductance (EDA), electrocardiogram (ECG), electromyogram (EMG), heartbeat rate, blood pressure, body temperature, etc. can be obtained unobtrusively. After fusing the data, statistical methods can be applied to these streams of the data to calculate driver volatility and monitor anomalies.

Figure 3 depicts a normal ECG waveform (*21*). Small squares are 0.04 seconds intervals, and thus large squares length would be 0.2 seconds long. The vertical axis shows the voltage in millivolt (mV) so that small squares are 0.1 mV and big squares are 0.5 mV. The figure shows three waves of P, QRS Complex and T. The atria contraction generates the P wave. The general range for P wave is either one small tall or one small box wide. If the P wave ranges more than one box in both vertical and horizontal axes, it might be an indication of atria hypertrophy (enlargement). The heart rate can be obtained by the following formula (*22*):

$$HR\ (bpm) = \frac{60}{R\ to\ R\ interval\ (sec.)} \quad (2)$$

Other important parameters to be monitored is QT and QT_c that have following relationship (*21*):

$$QT_c = QT + 1.75(HR - 60) \quad (3)$$

where *QT* is the activation and recovery of the ventricular muscle shown in Figure 3, and *HR* is the heart rate. By calculating and monitoring different parameters of the ECG waves, a comparison can be made between them and normal ranges to detect anomalies. Table 1 provides some of the parameters along with their reference ranges (*21*).Brain waves are also helpful in evaluating the level of the driver's brain activity and distraction. A functional near infrared spectroscopy (NIRS) device can be used to monitor the brain tissue oxygenation level. Analysis of the streams from NIRS is beneficial to assess the driver's brain functionalities, e.g., attention and distraction (*23*). For this study, we demonstrate detection of anomalies from bio-signals with simulated streams of heartbeat and blood pressure data, while realizing that other measures related to physiological data can also be added.

Road Infrastructure Volatility

Road infrastructure volatility relates to deviations related to roadway geometry, surface, and the surrounding environment. Notably, road infrastructure volatility is different from the vehicle and driver volatilities in the sense that it is not affected but likely affects them. Another difference is that although road infrastructure can become volatile sometimes, overall, the movement of the drivers through the road makes them experience the changes in attributes of the road infrastructure while those attributes are often fixed in specific locations. For instance, during a short trip, the road infrastructure can be considered "volatile" in terms of lane widths (if they vary substantially) as a driver passes over them. Similarly, a road without a shoulder would have higher volatility compared with one with a shoulder. Lane width and existence of a shoulder at a specific segment are mostly fixed in the context of the time frames of trips. However, some attributes may change more frequently. For instance, road surface can become slippery when it is raining. Another example is the number of driveways per mile for a segment that affects the safety. The variation in lane width might also affect the driving style and safety. Therefore, the mean and standard deviation of the lane width for a one-mile segment can be used to quantify the infrastructure volatility.

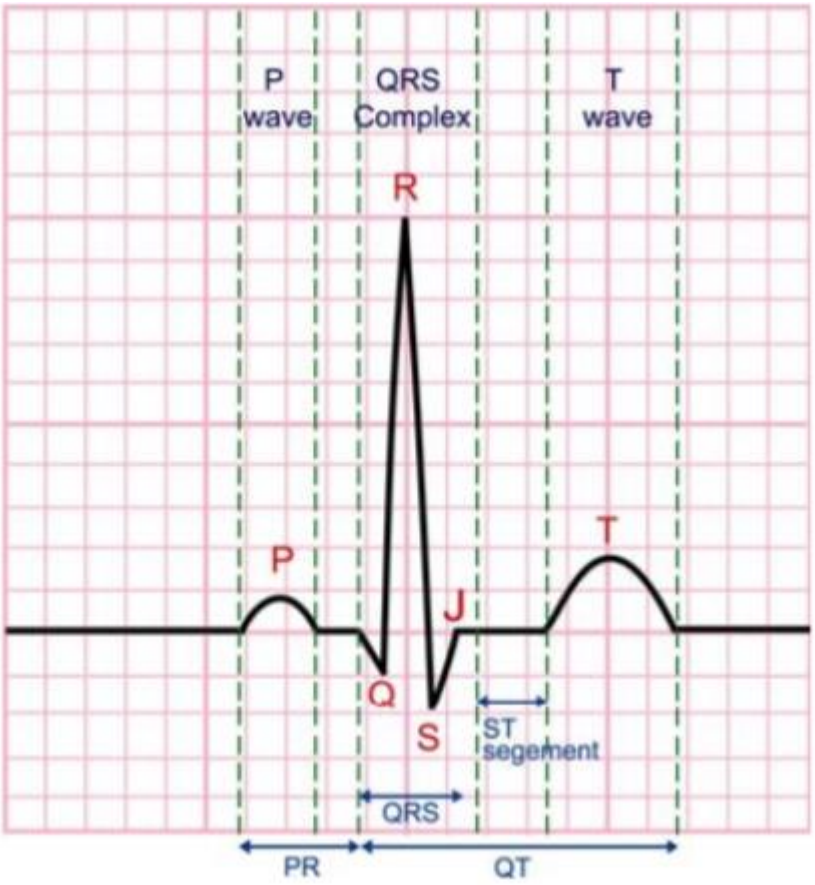


Figure 3. A normal electrocardiogram (ECG) waveform

Table 1. ECG parameters with their normal ranges

Parameters	Normal Range
Heart Rate (HR)	60 –100 bpm
QRS Complex	0.08 – 0.12 sec.
PR	0.12 – 0.20 sec.
QT	0.32 – 0.44 sec.
QT_c	0.41 – 0.44 sec.

Depending on the road infrastructure volatility, the defined thresholds for the vehicle volatility that trigger the warnings will need to be adjusted. For example, a 10% driving volatility might be acceptable where the shoulder exists, but with the absence of a shoulder, vehicle volatility above 5% can substantially increase the risk of running off the road and hence should result in a driver alert. Data collection for road infrastructure volatility demands several sensors in the vehicle along with high definition real-time maps that are used for CAVs. For this study, we use the simulated streaming data that relates to lane width, speed limit, and the existence of a slippery surface. Note that these data will change as a vehicle traverses the roadway.

Exploring Relationships

There may be relationships between different data streams that help in identifying anomalies. Specifically, it is important to note that detected anomalies from the driver's bio-signals can be due to the fluctuations in the roadway environment. These fluctuations represent "road infrastructure volatility." For example, volatility in the road infrastructure (such as a slippery road) can make a driver anxious. Such anxiety can be detected from the bio-signal, e.g., blood pressure and skin conductance. In this situation, first, the driver can become volatile and then the driver volatility, in turn, can affect the vehicle volatility. Therefore, a chain of volatilities from the environment to the driver and the vehicle can be observed from the data streams.

Driver and vehicle volatilities may be correlated, but with a lag. Typically, vehicle volatility will be a reflection of the driver's state (i.e. driver volatility). Therefore, it is very likely, unless we have defined the measures inappropriately, that they are correlated. However, such correlations need to be explored with real-life streaming data and more broadly a knowledge base of how relationships vary in real-time.

Real-Time Processing of Data and Events

A key issue is how to process the streams of the data available from the vehicle, driver and the road infrastructure in real-time with the purpose of identifying anomalies, spotting red flags, providing timely alerts and early warnings, even triggering vehicle control assists and tackling issues as they arise. With available computational capabilities, we can implement the concept of real-time data stream processing using software, e.g., Odysseus (24). The software provides the basic functions for quick development of event stream processing, and at the same time, it has the flexibility of extending and modifying data streams depending on users' preference (24). Odysseus uses Procedural Query Language (PQL) and Continuous Query Language (CQL) that are similar to SQL rather than a more complex language (25). Focusing on PQL, it offers several operators from basic (e.g., aggregate, filter, sort, join) to advanced (e.g., classification, clustering, regression, and anomaly detection). In addition, Odysseus can develop user-defined

functions and operators. The basic “Aggregate” operator allows users to obtain attributes such as minimum, maximum, count, mean, and standard deviation. Other mathematical operations can be applied using the “Map” operator. This operator is used to obtain vehicle volatility based on Equation 1.

To process the data streams in real-time, it is necessary to define a sampling rate and a time window. To demonstrate event data processing, we use a representative trip from the Michigan Safety Pilot Program DAS (data acquisition system) with 10 Hz sampling rate (10 observations per second). For consistency, the simulated driver and road infrastructure data are also at 10 Hz. The time window can be sliding or tumbling (26). In a sliding window, if the step is one, with one new observation coming in the window during the time unit, then the “first” observation is discarded, as shown in Figure 4. In a tumbling window, however, the consecutive windows do not overlap. Therefore, each observation is processed once. In this type of the window, in the case of window size 10, observations received during ten milliseconds are processed, and the system waits for the arrival of observations related to the next ten milliseconds. Once received, the system discards all observations from the previous window. Both sliding and tumbling windows are time-based windows. This means that if the rate of incoming data is changing, then the number of observations during each time window is also changing. For this study, a sliding window is used with the size of 100. In this setting, when the 101st observation enters, the first observation is discarded from the other side. Figure 4 shows how 100 observations are processed.

Data

Data collection under different driving conditions was performed in this study (Figure 5). Table 2 shows the descriptive statistics of one of the trips. The data contains vehicles’ instantaneous geocodes, speed, longitudinal acceleration, drivers’ electrocardiographs, and 4-second and 30-second heart rate average at 20 Hz. The data also include a timed video capturing the front view of the vehicle. Using the time of the video and the data, more information such as the type of the road and traffic condition are added to the dataset.

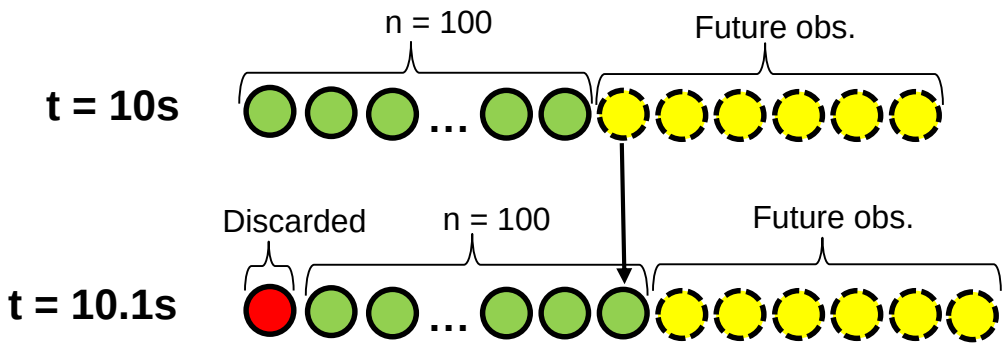


Figure 4. Element-based sliding window with size 100 for streaming data

Table 2. Descriptive statistics

Variable	Mean	SD	Min	Max
Speed (m/s)	3.05	1.53	0.00	11.88
Acceleration (m/s^2)	0.31	0.45	0	3.76
Deceleration (m/s^2)	-0.38	0.55	-5.45	-0.01
Heart rate 4-sec. average (bpm)	91.04	5.3	67	107



Figure 5. On-going data collection for the study

Initial Results

Table 3 shows vehicle performance (speeds and accelerations) and driver volatilities (heart rate) using four trips by the same driver taken in Xian, China in October 2017. Heart rate mean and standard deviation varies across different trips. Additionally, the volatility of the driver's heart rate varies considerably across trips, with Trip 3 showing a lower heart rate. However, the vehicle volatility results show that speed and acceleration volatilities are high for Trip 3. Furthermore, the results show that although speed volatility of vehicles can be similar, the acceleration volatility, which is a reflection of hard acceleration and hard braking, can be quite different. For instance, while the volatility of speeds in Trips 1 and 2 are close to speed volatility in Trip 4, the latter trip acceleration volatility (13.3%) is considerably higher than the other trips.

Given that the data only reflect four trips undertaken by a driver, the results are from a pilot experiment and may not be the same for other drivers or even for the same driver on other trips. As more data becomes available, rigorous statistical methods will be applied to investigate the correlation between vehicle volatility and driver volatility under diverse traffic conditions. Moreover, it is important to process streams of data coming from the vehicle, driver and the road infrastructure in real-time to identify anomalies, spot red flags and provide timely alerts and early warnings. To process the data streams in real-time, it is necessary to define a sampling rate and a time windows. Then statistical methods (e.g., time-series) can be applied to defined time windows.

Discussion

Figure 6 shows the screenshots of plots from the streams coming in from the vehicle and the driver. On the left-hand side, ECG and systolic blood pressure streams are shown at top and bottom respectively. On the right-hand side, the top and bottom figures show the streams of vehicular acceleration and the speed data. Given these incoming streams of data from the driver, the vehicle, and road infrastructure, we need to filter and join the required elements for further processing.

Table 3. Driver and vehicle volatility on four trips

Variable	Trip 1	Trip 2	Trip 3	Trip 4
Heart rate (4-sec. average) mean/SD	91.37/4.1	89.74/5.4	89.61/5.02	92.33/5.26
Heart rate (4-sec. average) volatility* (%)	2.63	3.27	1.17	1.49
Speed Volatility* (%)	4.66	4.81	5.31	4.42
Acceleration Volatility (%) **	0	0	27.89	13.3

* using $V_f = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{r_i} - \bar{r})^2}$ where $\mathrm{r_i} = \ln\left(\frac{x_i}{x_{i-1}}\right)$*100 see ref (*20*)

** using $\%T = \frac{c>Threshold}{n} * 100$ where $Threshold = \bar{x} \pm z * s$ see ref (*20*)

Figure 7 shows the graphical query regarding the sequence of processing steps to generate appropriate alerts or warnings, when necessary. Starting out at the bottom of the figure, having selected the required element from each source by filtering, the streams are fused to form the dataset. On the left side of the figure, real-time speed of the vehicle is compared with the speed limit stream from the road infrastructure. The “speed limit alert” is for the normal road conditions.

An alert is generated when the speed of the vehicle exceeds the speed limit (this can be changed to when the speed limit is exceeded by a certain threshold). The “speed limit wet” alert considers 5 mph lower than the posted speed limit as the threshold to issue an advisory, by simultaneously processing the “speed” from the vehicle and “speed limit” and “slippery” variables from the road infrastructure. For vehicle acceleration, we use the built-in “Anomaly Detection.” This operator needs the learning information from a built-in operator called “Deviation Learn.” This operator learns the mean and the standard deviation of a tuple and sends it to the “Anomaly Detection” operator. The operator uses received deviation information from the “Deviation Learn” operator and compares the current values of the tuple to the deviation information without the current tuple (24). If the value of the tuple is beyond the normal range around the mean, it is considered as an anomaly an alert can be generated.

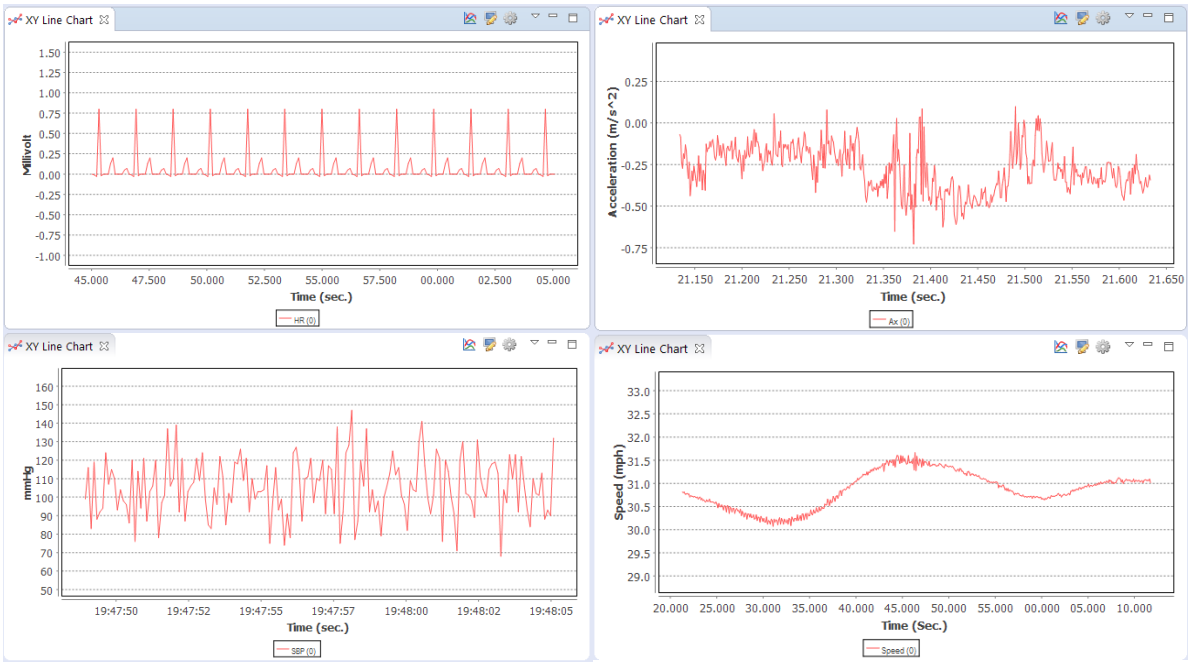


Figure 6. Plots of incoming streaming data from driver (left: blood pressure and heart ECG) and the vehicle (right: acceleration and speed)

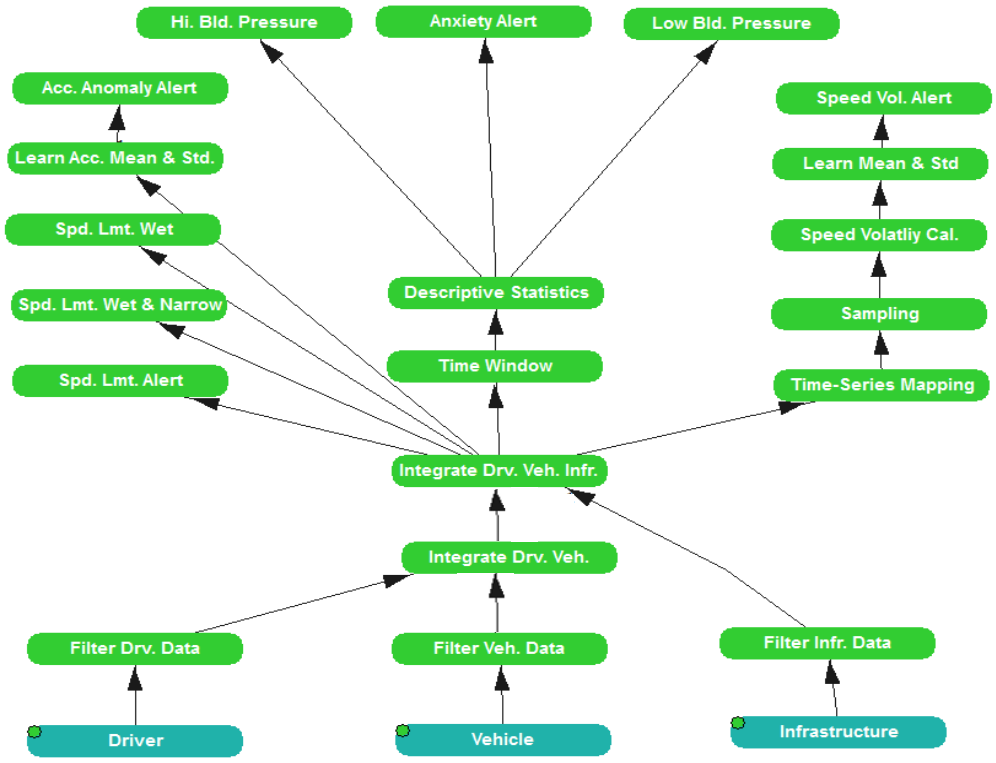


Figure 7. Graphical queries used in this study (in Odysseus software)

The right side of the Figure 7 further shows the sequence of calculating driving volatility (Equation 1) which is then used to generate driving volatility alerts. Previously discussed speed alerts were based on a simple comparison of the vehicle speed and posted speed limits given the road conditions (wet or slippery) and lane width. However, speed volatility alert is triggered when any calculated speed volatility falls beyond the normal ranges using the anomaly detection operator discussed above.

In the center of Figure 7, there is a “Descriptive Statistics” from which the “High Blood Pressure”, “Anxiety Alert” and “Low Blood Pressure” alerts are generated. Calculation of descriptive statistics requires specifying a normal range which is done partly through the definition of a time window that was discussed earlier (see Figure 4). While the significance of real-time processing is emphasized in this study, monitoring blood pressure and heart rate for generating alerts, needs historical data on those variables for individuals. For instance, a heart rate of 70 bps might be a sign of anxiety for one person, but it may be considered normal for another. Similarly, there are some normal ranges for blood pressure, but the common range (not necessarily normal for drivers with medical conditions) of blood pressure for different drivers are different. Given the variation of these parameters even in one person, the thresholds not only needs to be calibrated but also the alerts should be generated based on sufficient observations (and their average value) over a (predefined) time window rather than single values.

Advanced Processing of the Events

Correlation and Regressions

There might be correlations and relationships between the streams that can help us identify anomalies, spot red flags, provide timely alerts and early warnings and tackle issues as they arise. Therefore, processing streams from the vehicle, the driver and the road infrastructure in isolation, may miss opportunities to identify anomalous situations. For example, let us assume that a driver shows a higher heart rate than normal for any reason. In this case, the system can issue an alert, if the threshold is exceeded. However, it will be further helpful to observe if the higher heartbeat rate is also affecting the driver's behavior in terms of vehicle volatility. The simplest way to do this is to perform a linear correlation test on the driver's heart rate and the vehicle's speed, acceleration or one of the calculated volatilities. An alert can be generated if there exists a high positive or even high negative correlation (in some cases). It should be noted that in critical situations such as a heart attack, the correlation criterion has to be ignored because it is possible that a person is about to have a heart attack, but it is not reflected in the vehicle volatility yet. Therefore, waiting for vehicle volatility anomaly may delay an important early warning.

Most event streaming software can fit a linear or logistic regression to the incoming data in real-time. However, care is needed when interpreting the modeling results because of the issues such as sample size and model goodness of fit. Another appropriate use of regression model is if a previously fitted model that evaluates the risk of crashes is available. It will be possible to insert the values of the required variables coming from the data streams in the model to make the prediction. Afterward, we can evaluate the risk of a crash in real-time to generate warnings if the risk reaches a predefined threshold.

Limitation

The main limitation of the proposed framework is the difficulty of the data collection for the road infrastructure. It is likely that this limitation has made the analysis of the infrastructure data underexplored. Many commercial and open-source software provide real-time processing of the data streams. However, a challenge would be to implement the user defined and more complicated anomaly detection algorithm in a timely fashion.

Conclusion

Volatile driving characterized by hard accelerations and braking, sharp turns and weaving through traffic can result in crashes and higher energy/environmental costs. With the availability of new sensors data and increasing computational resources, it has become easier to observe volatility and extract useful information from data streams coming in from diverse sources. Sensors can be used to monitor the condition of the driver (e.g., biometrics, brain waves, and eye movements), the performance of the vehicle (e.g.,

speed, acceleration, and turning movements) and the changes in the roadway environment (e.g., road surface and lighting conditions). This paper conceptualized measures of driver, vehicle, and road infrastructure volatility and proposed a framework for monitoring them by fusing and processing streams of data. Specifically, real-time driving event processing is suggested for learning from the streams and taking proactive measures at critical moments to avoid crashes. Examples of data streams along with several parameters from which alerts and warnings can be generated were presented and discussed, with a focus on helping drivers or system operators identify anomalies, spot red flags, provide timely alerts and early warnings and tackle issues as they arise. The proposed event processing framework provides a way to bring the data to life and gives us a greater situational awareness. A proposed framework is easy to implement (as we showed the implementation of simple scenarios in an open source software). Therefore, in practice, a CAV or even a conventional vehicle can be retrofitted with sensors and a laptop to generate warnings to the driver. The conceptualization of the vehicle, the driver and the infrastructure volatilities presented in this paper would be beneficial to the researchers as well. The study attempted to present a fundamental platform to provide enabling technology to be implemented in lower levels of automation (level 0 to 3) with a focus on timely alerts, early warnings, and control assists.

Acknowledgement

Support was provided by the US Department of Transportation through the Collaborative Sciences Center for Road Safety, a consortium led by The University of North Carolina at Chapel Hill in partnership with The University of Tennessee. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

CHAPTER II
EXTRACTING USEFUL INFORMATION FROM BASIC SAFETY MESSAGE DATA: AN EMPIRICAL STUDY OF DRIVING VOLATILITY MEASURES AND CRASH FREQUENCY AT INTERSECTIONS

A version of this chapter was originally published by Mohsen Kamrani, Ramin Arvin and Asad J. Khattak:

Kamrani, Mohsen, Ramin Arvin, and Asad J. Khattak. "Extracting useful information from Basic Safety Message Data: An empirical study of driving volatility measures and crash frequency at intersections." *Transportation Research Record: Journal of the Transportation Research Board* (2018)

Abstract

With the emergence of high-frequency connected and automated vehicle data, analysts have become able to extract useful information from them. To this end, the concept of “driving volatility” is defined and explored as deviation from the norm. Several measures of dispersion and variation can be computed in different ways using vehicles’ instantaneous speed, acceleration, and jerk observed at intersections. This study explores different measures of volatility, representing newly available surrogate measures of safety, by combining data from the Michigan Safety Pilot Deployment of connected vehicles with crash and inventory data at several intersections. The intersection data was error-checked and verified for accuracy. Then, for each intersection, 37 different measures of volatility were calculated. These volatilities were then used to explain crash frequencies at intersection by estimating fixed and random parameter Poisson regression models. Given that volatility reflects the degree to which vehicles move, erratic movements are expected to increase crash risk. Results show that an increase in three measures of driving volatility are positively associated with higher intersection crash frequency, controlling for exposure variables and geometric features. More intersection crashes were associated with higher percentages of vehicle data points (speed & acceleration) lying beyond threshold-bands. These bands were created using mean plus two standard deviations. Furthermore, a higher magnitude of time-varying stochastic volatility of vehicle speeds when they pass through the intersection is associated with higher crash frequencies. These measures can be used to locate intersections with high driving volatilities, i.e., hot-spots where crashes are waiting to happen. Therefore, a deeper analysis of these intersections can be undertaken and proactive safety countermeasures considered at high volatility locations to enhance safety.

Introduction

High-frequency connected vehicle (CV) data offers an opportunity to detect dispersions in vehicular speeds, accelerations, and jerks. Measures of dispersion attempt to quantify the spread of data. Commonly used dispersion measures include variance, range,

minimum, and maximum values. In this paper, we expand the concept of "driving volatility," defined as deviation from the norm.

Volatility in driving reflects the degree to which a vehicle moves in three dimensions. If the vehicle's movements are erratic, then the risk of a crash is higher. Higher driving volatility is associated with higher safety risks, more fuel consumption, and increased emissions (*27*). The focus of this paper is to explore different measures of driving volatility, which have not yet been explored systematically in a spatial context.

CVs transmit high-frequency data between vehicles and road infrastructure. Widespread deployment of communication technologies has provided an unprecedented amount of data. Such "Big Data" combined with new tools can help researchers study, monitor, and evaluate transportation network performance in real-time (*28; 29*). This study takes advantage of the big data provided by the Safety Pilot Model Deployment (SPMD). SPMD is a field test in Ann Arbor, Michigan that offers detailed and relevant data demonstrating real-world vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication. In this program, around 3000 vehicles, equipped with Dedicated Short Range Communications (DSRC) devices, communicate with roadside equipment (*30*). The SPMD test provides rich information packages transmitted as Basic Safety Messages (BSMs) through V2V and V2I communication. BSMs contains the vehicles position and motion information, their component status, and other information (*30; 31*). To explore the relationship between volatility and crash frequency, this study has created a new and unique database that integrates BSMs, crashes, and inventory data to extract critical information from large-scale BSM data.

This study defines measures to quantify the driving volatility in a spatial context. Then we explore correlations between the measures of driving volatilities and crash frequencies at 116 intersections in Ann Arbor, MI, where sufficient instrumented vehicles' movements were recorded. The objectives of the study are to:

1) Define and calculate several measures of volatility using vehicles' kinematic characteristics.
2) Identify measures of driving volatility (if any) that are strongly associated with crashes at intersections.

Given that driver behavior is the main contributing factor in crashes (*32; 33*), findings from this study are beneficial in two ways. First, they can help proactively identify locations with high levels of driving volatility but might not have many crashes as candidates for safety improvements. Second, reduction of driving volatility at high crash locations can reduce future crashes.

Literature Review

There are various definitions for aggressive driving in the literature, but there is little agreement among them. In the current literature, researchers often use the term

“aggressive” for describing behaviors that threaten the safety of drivers and occupants in the host vehicle and other vehicles. In the U.S., aggressive driving such as speeding, failure to yield the right of way, and reckless driving account for more than 50 percent of fatal crashes (*34*). Different definitions of aggressive driving have been presented in the literature. Lajunen et al. (*35*) defined driver aggression as “any form of driving behavior that is intended to injure or harm other road users physically or psychologically.” These behaviors vary from less aggressive forms such as flashing lights, verbal threats, tailgating, and cutting other vehicles off, to more extreme behaviors such as physical attacks (*36*). When it comes to instantaneous driver behavior, aggressive driving can be described using different aspects of vehicle kinematics such as speed, acceleration, and vehicular jerk.

Many previous studies used common vehicle kinematics to quantify aggressive behavior or deviation from normal behavior (*37-39*). One of the more favorable variables for describing aggressiveness is maximum acceleration/deceleration of the vehicle. In the urban driving environment, Kim et al. (*40*) suggested the threshold of 1.47 m/s^2 and 2.28 m/s^2 for aggressive and extreme aggressive acceleration. De Vlieger (*41*) defined different thresholds for different driving styles in urban areas e.g., a range of 0.85 – 1.10 m/s^2 as an aggressive driving. Han et al. (*42*) quantified variations in driving behaviors under different driving conditions by providing different acceleration thresholds that vary with speed of the vehicle. Vehicular jerk, change in acceleration rate with respect to the time, is another element that can evaluate the aggressiveness of drivers. Vehicular jerk has been used to classify drivers’ style of aggressiveness (*43*) by using the ratio of standard deviation to the mean of jerk within a time span or identifying accident-prone drivers (*44*). Feng et al. (*45*) showed that there are unique characteristics of the vehicular jerk in the gas pedal operations. Also, aggressive drivers are found to be associated with significantly higher values of vehicular jerk (*45*).

More recently, a new term “driver volatility” was introduced to describe the performance of driving behavior. The difference between “volatility” and “aggressiveness” terms is similar to the “crash” and “accident” (*46*). Different measures for driving volatilities have been used in the previous studies (*1; 47*). Kamrani et al. (*1*) defined volatility score as the coefficient of variation (ratio of standard deviation to the mean) of acceleration and deceleration. To the best of authors’ knowledge, different measures of driving volatility have not been explored systematically, especially in the transportation context. Therefore, this study comprehensively explores several measures of driving volatility (applied to BSM data) and investigates their associations with intersection crash frequency.

Methodology

Various instantaneous driving measures can be used to quantify driving volatilities such as acceleration, brake position, and steering angle. Volatility in instantaneous driving

behavior should be measured by considering both longitudinal and lateral acceleration. Considering speed, acceleration, or jerk solely as the measure of volatility might ignore the importance of information embedded in the data. However, given a significant questionable error in the lateral acceleration data (*1*), only longitudinal acceleration, speed, and jerk are used in this paper. It should be noted that excluding lateral acceleration does not affect the results drastically for two reasons. First, the lateral acceleration is more critical where there is a noticeable amount of curvature in the travelers' trip, while the territory of the intersection in this study is limited to 150 ft from the center toward each approach. Second, in the area of an intersection, the traveled distance is short (called "passing" in this paper), and the geometry of the intersection does not allow drivers to have considerable lane changing space.

One hundred sixteen intersections were selected in the city of Ann Arbor, MI to extract BSM data consisting of speed, longitudinal acceleration (hereafter acceleration), time and geocodes. For each intersection, appropriate polygons are drawn based on 150 feet from the center of intersection toward all approaches. These polygons are used to filter the BSM data based on the longitude and latitude values available in the data. After the filtration, out of nearly 2,500,000,000 BSMs, 215,000,000 were found to be at the selected intersections. Data at this level are used for "level 1" calculations of driving volatilities (discussed later). The time and device ID variables of the BSMs are used to identify passings taken by each vehicle. Around 3,300,000 passings have been taken by more than 900 vehicles. Data from this step are used to do "level 2" calculations of driving volatilities (discussed later) at intersections. Crash and inventory data were also collected for individual intersections. The driving volatility and intersection related data are integrated to form the final dataset. The study uses rigorous modeling techniques that are suitable for the analysis of newly available volatility data.

Measures of Driving Volatility

While some of the measures used for volatility are common, as shown in Table 4, other measures presented are relatively new in the transportation field. Variations in longitudinal control of a vehicle are reflected in speed, acceleration, and vehicular jerk. The values of vehicular speed and acceleration are available directly from BSM data while the jerk values are calculated from the acceleration values, since it is the rate of change of acceleration.

Standard Deviation

A key measure for quantifying volatility is the standard deviation (S_{dev}) which is a simple and desirable statistic used for expressing variation in data:

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

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

Coefficient of Variation

A basic measure of dispersion is the coefficient of variation which is obtained from the division of the standard deviation by the mean (*1; 48*), providing a relative measure of dispersion shown in Equation (2).

$$C_v = \frac{S_{dev}}{|\bar{x}|} * 100 \quad (2)$$

where S_{dev} and $\bar{x}$ are the standard deviation and the mean respectively.

Mean absolute deviation around central point

This measure is defined as the average distance between each observation and the central tendency of the dataset (here mean) which is defined as (*49*):

Table 4. Summary of measures for driving volatility quantification

Measure of Driving Volatility	Formula	Applied to vehicular						
		Speed	Acceleration			Jerk		
			+	-	both	+	-	both
Standard Deviation	$S_{dev} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$	√			√			√
Coefficient of Variation	$C_v = \frac{S_{dev}}{\|\bar{x}\|} * 100$	√	√	√		√	√	
Mean Absolute Deviation	$D_{mean} = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \bar{x}\|$	√			√			√
Quartile Coefficient of Variation	$Q_{cv} = \frac{Q_3 - Q_1}{Q_3 + Q_1} * 100$	√	√	√		√	√	
Percent of extreme values	$\%T = \frac{c > Threshold}{n} * 100$ $Threshold = \bar{x} \pm z * S_{dev}$	√			√			√
Time-varying stochastic volatility	$r_i = \ln\left(\frac{x_i}{x_{i-1}}\right)$*100 $V_f = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2}$	√						

$$D_{mean} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \quad (3)$$

where x_i is the observation *i*, $\bar{x}$ is the mean, and *n* are the number of observations.

Quartile Coefficient of Variation

Another measure for describing dispersion of a dataset is the Quartile Coefficient of Variation, especially when the sample has non-normal distribution. The quartile coefficient of variation is defined as (*50*):

$$Q_{CV} = \frac{Q_3 - Q_1}{Q_3 + Q_1} * 100 \quad (4)$$

where Q_1 and Q_3 are the sample 25th and 75th percentiles respectively.

Count of extreme values

This measure captures driving volatility by counting the number of observations beyond a defined threshold-band. Equation 5 is showing the function (*51*) :

$$\%T = \frac{c > Threshold}{n} * 100 \quad (5)$$

where *c* is the count of observations beyond the threshold and *n* is the total number of observations. The threshold-band can be defined as (*51*):

$$Threshold = \bar{x} \pm z * S_{dev} \quad (6)$$

where $\bar{x}$ is the mean, and S_{dev} is the standard deviation; *z* represents the distance between a mean and a point in units of standard deviations, i.e., z = 1, 2, 3, etc. Application of this measure takes into account the magnitude vehicular speed, when calculating volatility of acceleration (*1*). Figure 8 shows how the speed bin concept is applied to the real-world acceleration data obtained from the BSMs. Notably, the ability of a vehicle to accelerate declines with higher speeds. Therefore, instead of having a fixed pair of upper and lower bounds to count the number of acceleration and deceleration extreme points, speed bins of 5 mph are used in this study. The upper and lower bound for each bin are calculated using its mean and standard deviation. Similarly, vehicular jerk is classified based on corresponding speed bins.

Time-varying stochastic volatility

The time-varying stochastic volatility which is commonly used in finance is computed by (*52; 53*):

$$V_f = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{r_i} - \bar{r})^2} \quad \text{from } i = 1 \text{ to } n \qquad (7)$$

where $\mathrm{r_i} = \ln\left(\frac{x_i}{x_{i-1}}\right) * 100$ (8)

and x_i and x_{i-1} are the current and previous observations (in this study instantaneous vehicular speeds) respectively and *ln* is the natural logarithm. This measure requires positive time-series observations; therefore, it is not applicable to the acceleration and jerk values due to their negative values. Using only the positive values of acceleration and jerk will be inconsistent with the time-series nature of data required by this measure. That said, this measure is applied to speed at the vehicle passing level (level 2), which is discussed next.

Two Levels for Calculating Volatility

Volatility measures can be applied in two ways to obtain driving volatility at intersections as shown in Figure 9.

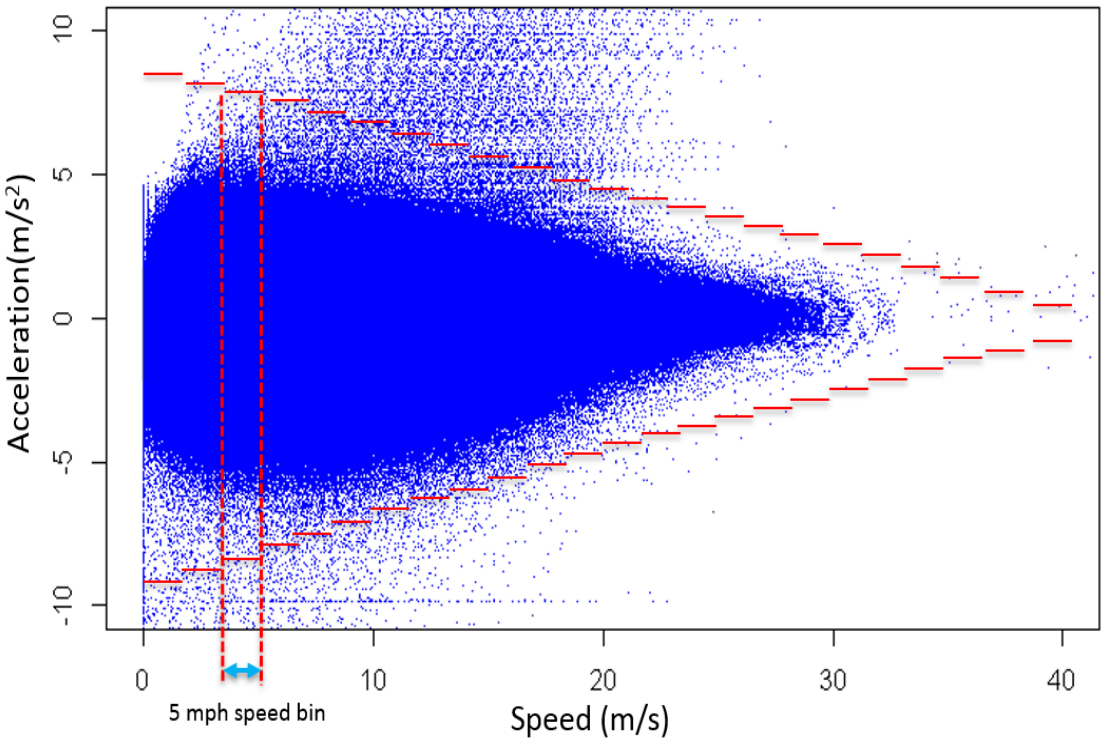


Figure 8. Speed bins for calculating acceleration thresholds at various speeds using BSMs data

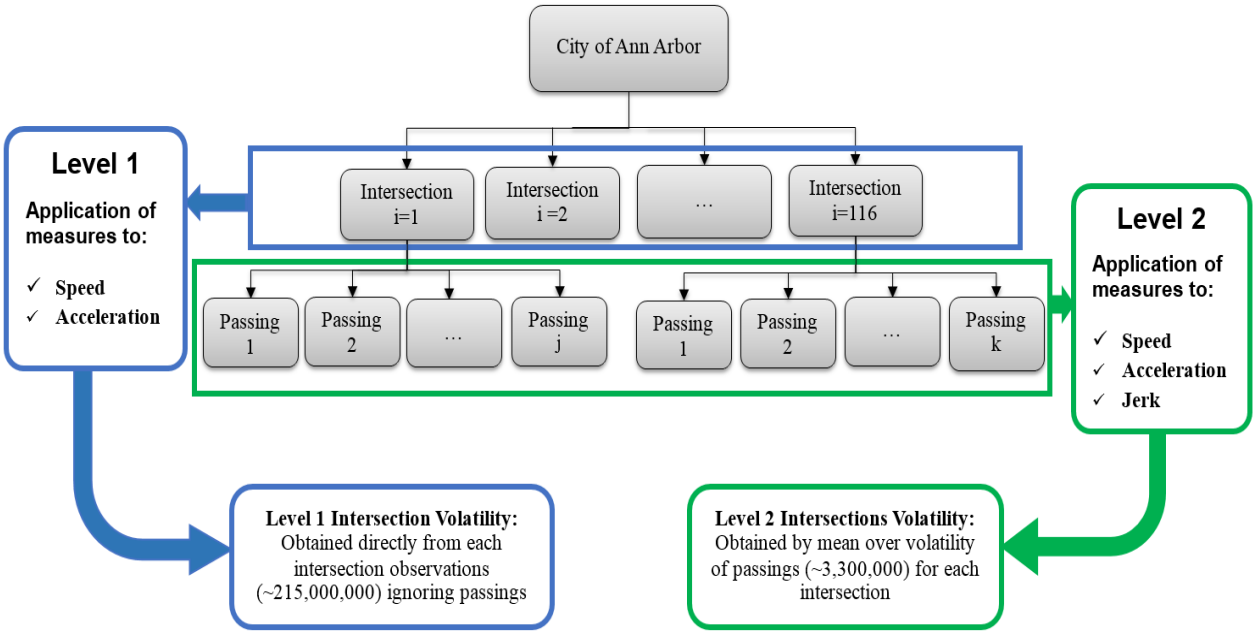


Figure 9. Measures of driving volatility at intersections

Level 1 calculation of volatility

The level 1 calculation of driving volatility disregards the individual passings (vehicles trips crossing the intersection) and treats all data for each intersection as bulk (at the aggregate level, N~215,000,000). Compared with Level 2, this calculation is simpler, easier, and faster to perform.

Level 2 calculation of volatility

In this method, volatility of each passing at the intersections is calculated separately. For this, the time and device ID available in BSMs are used to identify the passings. The averages of calculated volatilities for all passings are reported as measures of volatility for each intersection. Nearly 3,300,000 passings were identified for 116 intersections during the two-month period taken by around 900 unique device IDs.

Notation of Variables

Applying each of the measures to the speed, acceleration, and jerk at two levels results in 37 driving volatility values for each intersection. To distinguish them, a notation system is used where the volatilities have three terms in their names separated by dash "-".

- The first term is either "L_1" for "Level 1" or "L_2" for "Level 2" which indicates the method of calculation.

- The second term indicates the element to which the volatility measure is applied. Since some of the measures of volatilities necessitate the separation of positive and negative values, the second term can have the following notation:
    - ***Speed:*** vehicular speed
    - ***AccDec:*** both positive and negative values of acceleration
    - ***Accel:*** positive values of acceleration
    - ***Decel:*** negative values of acceleration
    - ***Jerk:*** vehicular jerk calculated from acceleration
    - ***PosJerk:*** positive values of jerk
    - ***NegJerk*****:** negative values of jerk
- The last term shows what measure was applied to obtain the volatility. For example, if **standard deviation** is applied to the **acceleration (both positive and negative values)** for individual **passings (level 2)**, the variable will be named: ***"L2-AccDec-*S_{dev}*"***.

Modeling Approach

Count-data models are commonly used for modeling accident frequency since the number of crashes on a roadway or intersection is per unit of time and is a non-negative integer (*54*). Count data are usually modeled using Poisson or its derivatives Negative Binomial and zero-inflated models (*55; 56*). For the Poisson regression model, the probability of having *n* crashes at intersection *i* is (*57*):

$$P(n_i) = \frac{\lambda_i^{n_i} \exp(-\lambda_i)}{n_i!} \quad (9)$$

where $P(n_i)$ is the probability of having *n* crashes at intersection i, λ_i is the Poisson parameter for the intersection *i*. These are the expected number of crashes for the intersections in each year. In order to fit the model, λ_i can be expressed in the logarithm form as the function of a set of independent variables (*57*):

$$\ln(\lambda_i) = \beta X_i \quad (10)$$

where X_i is a vector of explanatory variables; β is a vector of estimated coefficients. The Poisson function defined in Equation (9) and (10) can be maximized by standard maximum likelihood procedures.

Applying Poisson regressions to the data while the mean and variance are not equal, might lead to inappropriate results. To address over-dispersion ($E(n_i) < VAR(n_i)$), or under-dispersion ($E(n_i) > VAR(n_i)$) in the data, the Negative Binomial model can be derived as:

$$\lambda_i = \exp(\beta X_i + \varepsilon_i) \quad (11)$$

where error term, $\exp(\varepsilon_i)$, is a gamma-distributed with mean 1 and variance α. The additional term, allows variance to be different from the mean:

$$Var\ (n_i) = E(n_i) + \alpha E(n_i)^2 \tag{12}$$

where $Var(n_i)$ and $E(n_i)$ are the variance and the expected number of crashes respectively.

Choosing between Poisson and Negative Binomial regression depends on the estimated α parameter. If α significantly does not differ from zero, Poisson regression model should be used. Otherwise, the Negative Binomial model is appropriate (*58*). Although the presence of over-dispersion can be evaluated by the mean and variance of crash data (*58*), a Lagrange multiplier (LM) can be used to statistically test the existence of overdispersion in Poisson model (*57*).

On the other hand, it is possible that associations between independent variables and the dependent variable is not consistent across all observations. Several observed and unobserved factors associated with crash frequency might lead to unobserved heterogeneity *(59-63)*. To address the heterogeneity with random parameters, using simulated maximum likelihood estimation, Greene (*57*) developed an approach to model random parameters in the Poisson model. Equation (13) indicates the formulation of estimated coefficients:

$$\beta_i = \beta + \varphi_i \tag{13}$$

where φ_i is a randomly distributed term with any specified distribution (e.g., normal distribution with mean zero and standard deviation σ). The Negative Binomial parameter in Equation (10) can be written as:

$$\lambda_i|\varphi_i = e^{(\beta X_i + \varepsilon_i)} \tag{14}$$

The log-likelihood function for the random-parameter model can be written as (*54*):

$$LL = \sum_i ln \int_{\varphi_i}^{i} g(\varphi_i) P(\ n_i|\varphi_i) d\varphi_i \tag{15}$$

where g(.) is the pre-specified probability density function for φ_i. In order to maximize the log-likelihood function, a simulation-based approach using Halton draws can be used. Different studies (*64; 65*) have shown that Halton draws provide a more efficient distribution for numerical integration in comparison with random draws. Further details on random parameter models can be found in (*54*).

Data

The data used in this study are the result of integrating BSMs from the Michigan Safety Pilot with intersection crash and inventory data. The steps for data preparation are shown in Figure 10 (right). The BSMs data were collected, under real-world conditions, at the Ann Arbor test site by equipping around 3,000 vehicles with DSRC devices enabling them to log different variables including their instantaneous speed, acceleration heading, coordinates, etc. at usually 10 Hz. The data is accessible via ITS Public Data Hub (https://www.its.dot.gov/data/), maintained by the Federal Highway Administration under US DOT. Speed, acceleration, longitude, and latitude values of the complete two-month data (October and April 2012) were utilized in this study. The data examination and error-checking process shows high accuracy in the variables used in this study. For instance, the accuracy of the map created from BSMs shown in Figure 10 (left) is a good indication of data precision.

Intersection specific data such as the average number of crashes (2010-2014), annual average daily traffic (AADT), and speed limits for all approaches were collected. The dataset was error checked (via randomly double checking 10% of the data by a third person) and verified. The data can be obtained via Metropolitan Planning Organization website: http://semcog.org/Crash-and-Road-Data. Among intersections in the Ann Arbor area, 116 intersections are identified keeping in view that enough BSM data should be available for calculation of different measures of driving volatility. Finally, appropriate geocodes are used to filter out BSMs data for each intersection. These BSMs were used to calculate 37 different measures of driving volatilities. The final dataset was created by integrating intersection inventory data, crash data and computed driving volatilities.

Results

Descriptive Statistics

Table 5 presents the descriptive statistics. For all intersections, the five-year mean of crashes is 7.56 with a standard deviation of 7.64. About 46% of the intersections are signalized, 40% of the intersections are 4-legged, and the rest are T-intersections. Table 5 also presents the descriptive statistics of variables calculated from BSM data i.e. measures of volatilities. Please note that the unit of analysis is the intersection.

Correlations

Given the number of computed volatilities, correlation analysis may shed some light on relationships between crash frequency and driving volatilities (Figure 11). Bars in the figure are sorted based on the value of positive correlation. Blue bars show volatilities with a positive correlation between average crashes while the red ones indicate negatively correlated volatilities. This figure was used as a guide to insert variables in the

model specification and to examine their associations and improvements in model fit. . As expected, there is a high level of correlation among some of the computed volatilities.

For instance, two highly correlated volatilities at the bottom of the figure (L_2-AccDec-1S_{dev} and L_2-AccDec-2S_{dev}) are calculated in a similar way, with the only difference being in the number of standard deviations from the mean. If such highly correlated variables are used simultaneously in estimation, then the model may suffer from multicollinearity. Using engineering judgment and Variance Inflation Factor (VIF>5), multicollinearity was addressed in the model specification.

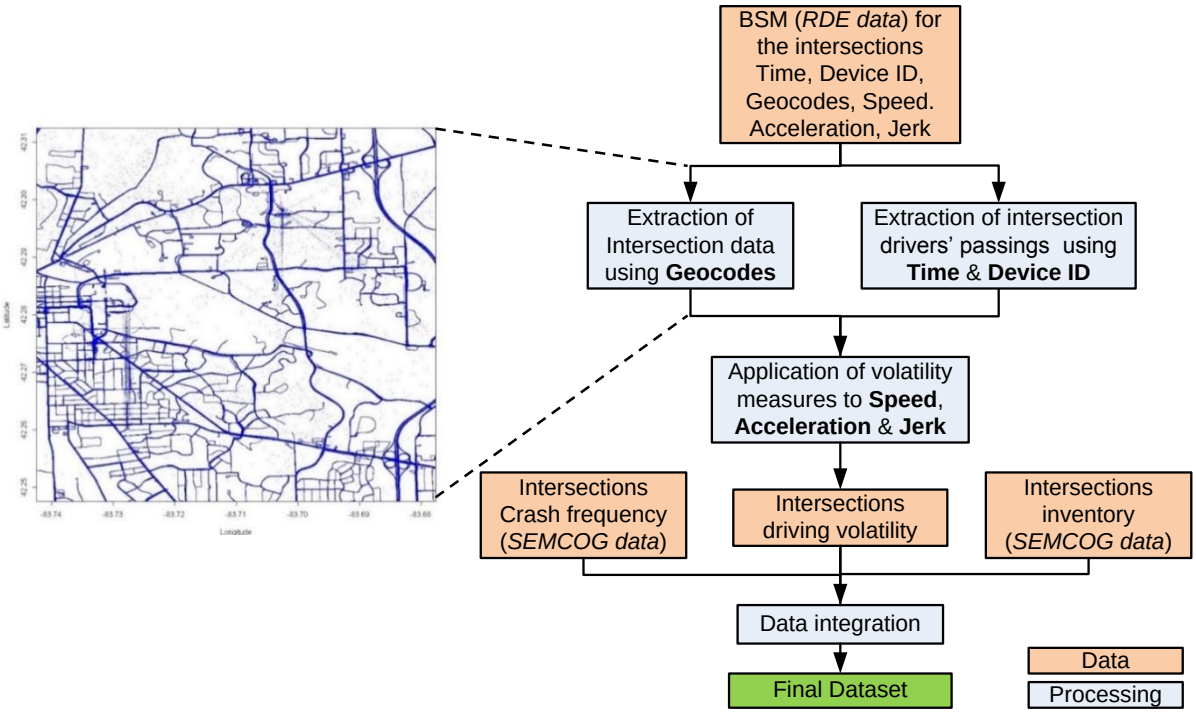
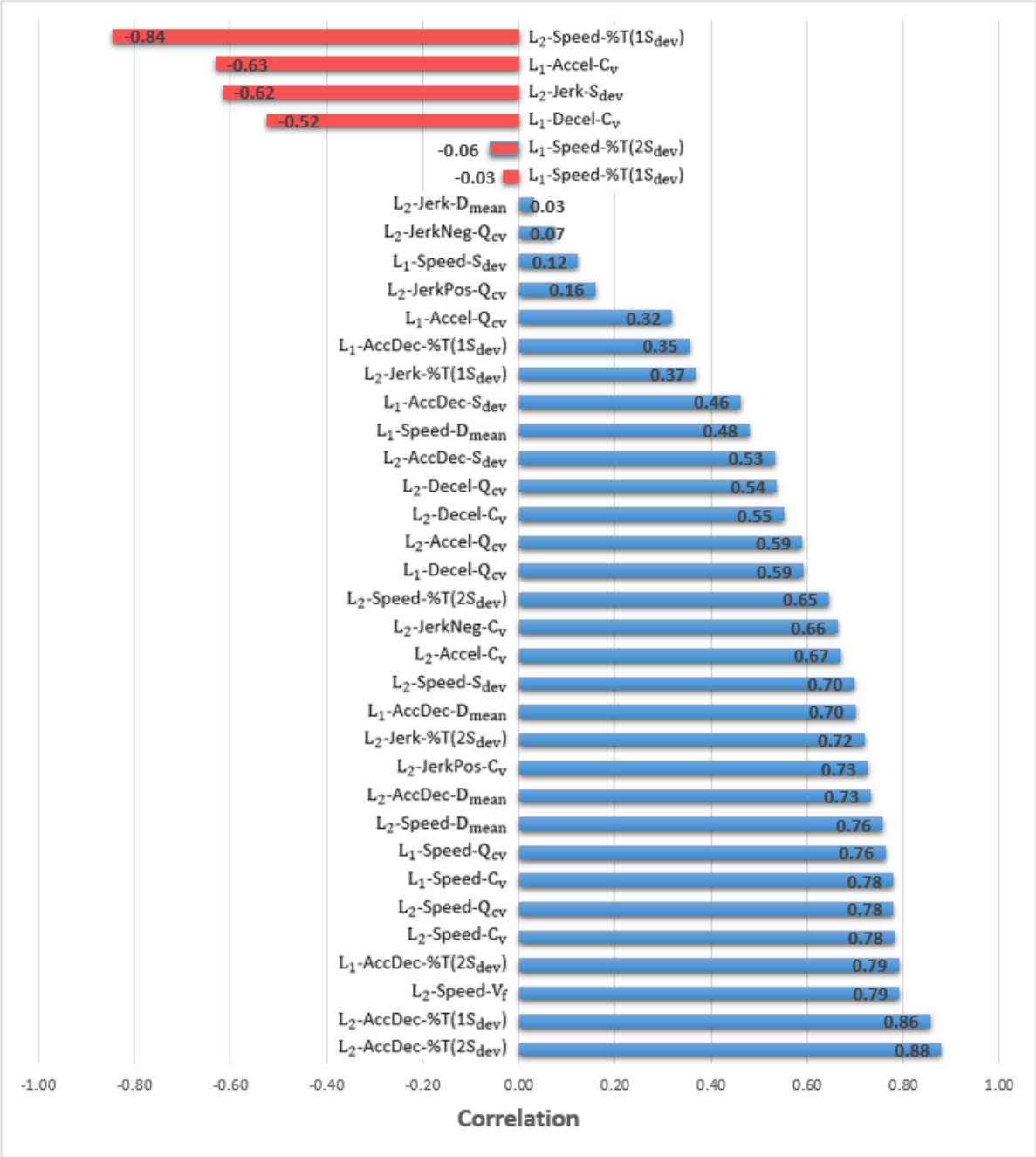


Figure 10. Left: Ann Arbor map created from BSM data, Right: Data preparation steps

Table 5. Descriptive statistics of measures of volatilities (n = 116)

Variables	Mean	S_{dev}	Min	Max
Intersection-related variable				
Average crashes (5 years)	7.56	7.64	0	44
AADT major road	20805	8326	3100	45400
AADT minor road	9396	4138	1100	27400
Speed limit major road	35.34	7.24	25	45
Speed limit minor road	30.47	3.95	25	45
Signalized intersection (yes = 1)	0.46	0.5	0	1
4-legged intersection (yes = 1)	0.4	0.49	0	1
Total through lanes	4.45	1.28	2	8
Total left turn lanes	1.53	1.32	0	6
Total right turn lanes	0.93	0.78	0	4
*Volatility of Level 1 variables (ignoring individual vehicle passings)**				
L_1-Speed-S_{dev} (m/s)	11.35	2.4	4.92	16.41
L_1-Speed-C_v (%)	45	16	13	71
L_1-Speed-Q_{cv} (%)	32	16	6	61
L_1-Speed-D_{mean} (m/s)	7.85	1.96	3.21	12.32
L_1-Speed-$\%T(1S_{dev})$ (%)	28	13	11	59
L_1-Speed-$\%T(2S_{dev})$ (%)	4	3	0	11
L_1-AccDec-S_{dev} (m/s^2)	0.75	0.17	0.34	1.43
L_1-Accel-C_v (%)	59	6	44	73
L_1-Decel-C_v (%)	65	9	51	103
L_1-Accel-Q_{cv} (%)	39	6	23	51
L_1-Decel-Q_{cv} (%)	44	7	23	59
L_1-AccDec-D_{mean} (m/s^2)	0.4	0.09	0.15	0.52
L_1-AccDec-$\%T(1S_{dev})$ (%)	23	4	14	36
L_1-AccDec-$\%T(2S_{dev})$ (%)	7	1	3	9
*Volatility of Level 2 variables (averaged over passings)**				
L_2-Speed-S_{dev} (m/s)	2.02	0.95	0.41	5.28
L_2-Speed-V_f (%)	2	2	0	6
L_2-Speed-C_v (%)	15	10	1	40
L_2-Speed-Q_{cv} (%)	10	7	1	26
L_2-Speed-D_{mean} (m/s)	1.49	0.7	0.3	3.47
L_2-Speed-$\%T(1S_{dev})$ (%)	34	2	29	39
L_2-Speed-$\%T(2S_{dev})$ (%)	2	1	1	4
L_2-AccDec-S_{dev} (m/s^2)	0.4	0.13	0.17	1.18
L_2-Accel-C_v (%)	27	6	15	43
L_2-Decel-C_v (%)	29	5	16	44
L_2-Accel-Q_{cv} (%)	18	4	10	28
L_2-Decel-Q_{cv} (%)	20	4	12	29
L_2-AccDec-D_{mean} (m/s^2)	0.16	0.06	0.05	0.35
L_2-AccDec-$\%T(1S_{dev})$ (%)	36	4	27	49
L_2-AccDec-$\%T(2S_{dev})$ (%)	4	1	2	8
L_2-Jerk-S_{dev} (m/s^3)	1.37	0.15	1.04	1.78
L_2-JerkPos-C_v (%)	59	3	52	65
L_2-JerkNeg-C_v (%)	59	3	52	64
L_2-JerkPos-Q_{cv} (%)	44	3	32	48
L_2-JerkNeg-Q_{cv} (%)	44	3	32	47
L_2-Jerk-D_{mean} (m/s^3)	0.81	0.11	0.56	1.09
L_2-Jerk-$\%T(1S_{dev})$ (%)	26	1	23	28
L_2-Jerk-$\%T(2S_{dev})$ (%)	7	1	4	10

** L_1: level 1 calculation; L_2: level 2 calculation; S_{dev}: standard deviation; $\%T(1S_{dev})$: % of extreme points beyond mean ± one standard deviation; $\%T(2S_{dev})$: % of extreme points beyond mean ± two standard deviation; C_v: coefficient of variation; Q_{cv}: quartile coefficient of variation; D_{mean}: mean absolute deviation; V_f: time-varying stochastic volatility; Accel: acceleration; Decel: deceleration; AccDec: both acceleration & deceleration; JerkPos: positive jerk; JerkNeg: negative jerk.*



** L_1: level 1 calculation; L_2: level 2 calculation; S_{dev}: standard deviation; $\%T(1S_{dev})$: % of extreme points beyond mean ± one standard deviation; $\%T(2S_{dev})$: % of extreme points beyond mean ± two standard deviation; C_v: coefficient of variation; Q_{cv}: quartile coefficient of variation; D_{mean}: mean absolute deviation; V_f: stochastic time-varying volatility; Accel: acceleration; Decel: deceleration; AccDec: both acceleration & deceleration; JerkPos: positive jerk; JerkNeg: negative jerk.*

Figure 11. Correlations between crash frequency and measures of volatilities

Modeling Results and Discussion

Table 6 provides the results for fixed and random parameter Poisson regression. Fixed-parameter model is estimated for crash frequency as a function of intersection-related variables and measures of driving volatility. Starting out with intersection-related variables and keeping the significant ones in the model, measures of volatility variables were inserted in the model based on correlations from Figure 11. The models fit were compared using AIC and log-likelihood.

The random-parameter Poisson model is estimated (using simulated maximum likelihood) assuming a normal distribution for random parameters (*54*). Compared to fixed-parameter model, the random-parameter model shows a better fit based on log-likelihood, AIC, and McFadden ρ^2 (*58*). As Figure 12 shows, the random-parameter model outperforms the fixed-parameter in terms of crash frequency prediction.

The marginal effects are shown in Table 6. These effects are the average increases in crash frequencies of intersections given one unit increase in the respective independent variable. For instance, a one-percent increase in the time-varying stochastic volatility of speed (L_2-Speed-V_f) is associated with a 0.55 increase in average crash frequency.

Table 6. Fixed and random parameters Poisson model results

Variables	Fixed Parameter			Random Parameter		
	β [a]	**z**	**M.E.**	***β*** [a]	**z**	**M.E.**
Constant	-1.497***	-4.73	--	-1.852***	-5.42	--
Intersection-related						
AADT major approach (1000)	0.033***	7.39	0.25	0.033***	7.84	0.17
Std. dev.	--	--	--	0.007***	5.36	--
AADT minor Approach (1000)	0.023***	3.55	0.17	0.024***	3.70	0.12
Signalized intersection (yes = 1)	0.789***	6.01	5.21	0.704***	5.77	3.58
Four-legged intersection (yes = 1)	0.260**	3.11	1.95	0.248***	2.93	1.26
Measures of volatility [b]						
L_1-Speed-$\%T(2S_{dev})$	0.050***	3.57	0.38	0.041***	2.97	0.21
Std. dev.	--	--	--	0.065***	8.53	--
L_1-AccDec-$\%T(2S_{dev})$	0.225***	4.38	1.70	0.260***	4.63	1.32
L_2-Speed-V_f	0.061 ·	1.92	0.47	0.109***	3.47	0.55
Summary Statistics						
AIC	609.65			585.6		
Log-likelihood at Zero *L(0)*	-578.32			-578.32		
Log-likelihood at Convergence *L(β)*	-296.83			-282.81		
McFadden ρ^2	0.487			0.517		
Sample Size (N)	116			116		

[a] *Significance codes: *** 0.01%, **1%, * 5%, · 10%*

[b] L_1*: level 1 calculation ;* L_2*: level 2 calculation;* $\%T(2S_{dev})$*: % of extreme points beyond mean ± two standard deviation;* V_f*: time-varying stochastic volatility*; *AccDec*: both acceleration & deceleration.

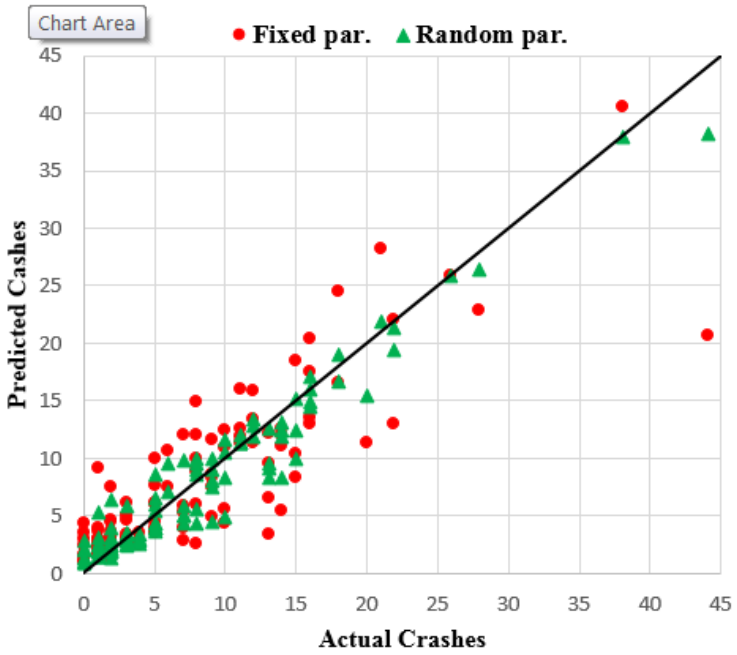


Figure 12. Expected-actual number of crashes for fixed and random parameter models

That means a higher magnitude of time-varying stochastic volatility of vehicle speeds when they pass through the intersection is associated with higher crash frequencies, as expected. In addition, more intersection crashes were associated with higher percentages of vehicle data points (speed & acceleration) lying beyond threshold-bands created using mean plus two standard deviations at intersections (L_1-Speed-$\%T(2S_{dev})$ and L_1-AccDec-$\%T(2S_{dev})$ variables).

Other variables which are used as controls in the model show the expected signs and magnitudes. According to Table 6, 1000 more vehicles per day on the major approach are associated with a 0.17 increase in crash frequency. As expected, the association of the minor approach AADT is less than the major approach ADDT. One-thousand more vehicles on the minor road are correlated with a 0.12 increase in crash frequency. According to the model, signalized intersections on average have 3.58 more crashes than un-signalized ones. Likewise, 4-legged intersections on average have more crashes than 3-legged intersections.

Limitations

The sample data used in this study does not come from representative drivers. This study did not consider volatility in the lateral direction, which could result in a sideswipe crash. Given that lane change frequency is generally relatively small at intersections, the results might not be considerably different. Furthermore, the data used in this study is the product

of averaging 5-year crashes and using two-month BSMs data. In other words, a short period of instantaneous driving behavior was used to explore correlations with 5-year average crash frequencies. The authors have used all available data to make the results as accurate as possible, even though handling and processing such large-scale data was difficult. Although the data was error-checked, it is possible that some errors, made during collection of data, remain. This paper considers only crash frequency while it is worthwhile to investigate the associations of driving volatility with crash severity. Finally, it should be noted that only the means of calculated volatilities for passings (level 2 volatility) were used to model volatility at each intersection, while the between-passings variation could also be used as measures of volatility.

Conclusion

This study discusses a way to extract useful information in the form of driving volatility from newly available BSM data. Such data are increasingly becoming available, providing a valuable resource for studying vehicle kinematics and microscopic behaviors of drivers, e.g., instantaneous vehicle speed, acceleration, and jerk. This study creates a new and unique database (BSM data integrated with crash and inventory data) and mined critical information from large-scale BSM data. More than 2,500,000,000 BSMs were processed along with crash and inventory data from 116 intersections in the city of Ann Arbor, Michigan. Volatilities of vehicles passing within 150 feet from the center of each intersection are calculated. Using nearly 215,000,000 observations for nearly 3,300,000 passings, 37 measures of driving volatility were calculated. To explore relationships between measures of driving volatility and crash frequency at intersections, rigorous statistical models were estimated. The models account for unobserved heterogeneity associated with crashes at intersections.

Three measures of driving volatilities show positive and statistically significant association with crash frequencies at the intersections. More intersection crashes are found to be associated with higher percentage of BSM data points of speed and acceleration lying beyond the threshold-bands created using mean plus two standard deviations at intersections. Furthermore, a higher magnitude of time-varying stochastic volatility of vehicle speeds when they pass through the intersection is associated with higher crash frequency. The findings are significant in the sense that they can be used to identify intersections with high levels of driving volatility. In particular, intersections where crash frequency may be low, but the volatility is high, may be good candidates for further study and future safety treatments. These are likely to be intersections where crashes are waiting to happen due to higher driving volatility. Such intersections can be proactively examined to find the causes of driving volatility to prevent crashes. Higher levels of driving volatility might be due to outdated signal timing, higher speed limits, limited line of sight, inappropriate signal timing, etc. In practice, depending on the detected reasons, proactive

countermeasures can be taken to reduce drivers' volatility. In addition, appropriate alerts can be given to vehicle drivers when they are approaching locations (*66*) with a high level of driving volatility.

Acknowledgment

This paper is based upon work supported by National Science Foundation (Award number: 1538139). Additional support was provided by the US Department of Transportation through the Collaborative Sciences Center for Road Safety, a consortium led by The University of North Carolina at Chapel Hill in partnership with The University of Tennessee. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect views of the sponsors.

CHAPTER III
CAN DATA GENERATED BY CONNECTED VEHICLES ENHANCE SAFETY? A PROACTIVE APPROACH TO INTERSECTION SAFETY MANAGEMENT

A version of this chapter was originally published by Mohsen Kamrani, Behram Wali and Asad J. Khattak:

Kamrani, Mohsen, Behram Wali, and Asad J. Khattak. "Can data generated by connected vehicles enhance safety? Proactive approach to intersection safety management." *Transportation Research Record: Journal of the Transportation Research Board* 2659 (2017): 80-90.

Abstract

Traditionally, evaluation of intersection safety has been largely reactive, based on historical crash frequency data. However, the emerging data from Connected and Automated Vehicles (CAVs) can complement historical data and help in proactively identify intersections which have high levels of variability in instantaneous driving behaviors prior to the occurrence of crashes. Based on data from Safety Pilot Model Deployment in Ann Arbor, Michigan, this study developed a unique database that integrates intersection crash and inventory data with more than 65 million real-world Basic Safety Messages logged by 3,000 connected vehicles, providing a more complete picture of operations and safety performance of intersections. As a proactive safety measure and a leading indicator of safety, this study introduces location-based volatility (LBV), which quantifies variability in instantaneous driving decisions at intersections. LBV represents the driving performance of connected vehicle drivers traveling through a specific intersection. As such, by using coefficient of variation as a standardized measure of relative dispersion, LBVs are calculated for 116 intersections in Ann Arbor. To quantify relationships between intersection-specific volatilities and crash frequencies, rigorous fixed- and random-parameter Poisson regression models are estimated. While controlling for exposure related factors, the results provide evidence of statistically significant (5% level) positive association between intersection-specific volatility and crash frequencies for signalized intersections. The implications of the findings for proactive intersection safety management are discussed in the paper.

Introduction

There is considerable evidence about vehicle conflicts at intersections resulting in crashes, making them among the most dangerous locations on roadways (*67; 68*). Traditionally, intersection safety evaluations are done based on historical data and they are largely reactive i.e. the state-of-the-art methods characterize unsafe intersections based on historical and expected crash frequencies (*68; 69*). Safety treatments can then be applied to intersections based on historical crash data methodology. Variability in instantaneous driving behaviors can be leading indicators of occurrence of unsafe outcomes such as crashes/incidents. In this study, we posit that expanding the concept

of driving volatility (*2; 3; 70*) to specific locations (termed as Location-Based Volatility) by using real-world large-scale connected vehicle data has a significant potential in unveiling critical relationships between extreme driving behaviors (and its fluctuations) and safety outcomes at specific intersections.

The Safety Pilot Model Deployment (SPMD) offers detailed and relevant data. This pilot is underway in Ann Arbor, Michigan, intended to demonstrate vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication in a real-world environment. Within SPMD, Basic Safety Messages (BSMs) contain rich information packets (exchanged at the frequency of 10 Hz) that describe a vehicle's position, motion, its component status, and other relevant information exchanged between vehicles/infrastructure through V2V and V2I applications (*71*). Such emerging data has been used for creating trip-based driving volatilities for drivers, capable of identifying abnormal or extreme behaviors prior to unsafe outcomes such as crashes/incidents (*2*). Important in this aspect is the concept of "driving volatility" that captures the extent of variations in driving, especially hard accelerations/braking, jerky maneuvers, and frequent switching between different driving regimes (*70*). Specifically, Wang et al. (*3*) and Liu and Khattak (*2*) examined the relationships between trip-based driving volatility and several factors such as demographics, trip purpose, duration, and detailed vehicle characteristics (*2; 3*). The potential of driver-specific trip-based volatilities for developing advanced traveler information systems, driving feedback devices, and alternative fuel vehicle purchase decision tools were concluded (*2; 3*).

This study focuses on developing an analytic methodology to examine instantaneous driving behaviors at specific locations, and its variability. The paper explores how variability in driving can be mapped to historical safety outcomes such as crashes at specific locations. Such an analysis is fundamental towards proactive intersection safety management.

Literature Review

There are different branches of ongoing research topics in the connected vehicles (CV) area. Several major directions of research can be identified. Topics such as network robustness and information propagation efficiency (*72*) are still under investigation in order to establish a better vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) connection (*72*). Another is the systems and algorithms whose ultimate goal are the reduction of the gap between vehicles in order to increase roads capacity and reduction in fuel consumption through different method such as speed harmonization (*73*), trajectory optimization (*74*) and platooning as discussed in Bergenhem et al. (*75*).

Also, there are a number of studies (not necessarily in CV area) trying to characterize aggressive, reckless or risky driving style (*76*). Among them, speed limits are usually the threshold that determines a driver's performance (*77; 78*). While

characterizing driver's performance, the important finding is that risky driving behaviors have been found to be positively correlated with the likelihood of crashes or near-crash events (*79*). This said, a broad spectrum of studies related to connected vehicle systems have proposed mechanisms for warnings or alerts to drivers using the CV applications and their effect on safety. For instance, Chrysler *et al.* (*80*) investigated the effect of warning messages on drivers' ability to handle primary and secondary threats. The results showed an improved detection time for the primary threat while increased reaction time to the secondary threat which was placed after the primary threat. In another study (*81*), the impacts of dynamic route guidance on work zone safety under different market penetration of CV were explored. Per the interesting results, 40% penetration of CV and below improves safety while above that leads to decreased safety of work zones. However, these benefits are dependent on the information dissemination delay (*82*). Although, positive effects of warning messages have been investigated, the way those warning should be created from BSMs is still under explored.

One approach is trying to link the generation of warning messages to drivers' behavior. In some recent studies, the authors have initiated efforts to extract useful information from BSMs to understand the drivers' behavior. For instance, a measure of driving performance in connected vehicles network has been defined as "driving volatility" (*83*). As such, trip-based driving volatility was introduced (*83*) to account for the variation of driving behaviors under different conditions using objective driving performance evaluation matrix i.e. vehicular jerk. More succinctly, Liu *et al.* (*84*) studied extreme driving behaviors (trip-based volatility) using exhaustive high frequency connected vehicle data, and the analysis demonstrated framework for the generation of warnings/alerts for connected vehicles informing drivers about potential hazards. Also another study (*85*) proposed a way to identify abnormal or extreme behaviors (i.e., hard acceleration and decelerations) from BSMs, and warn drivers through the V2V, V2I, or other connected vehicle applications. In this paper, the authors believe that expanding the concept of driving volatility in connected vehicles environment to specific locations has significant potential in identifying hazardous roadway segments. Such a perspective of location-specific driving behavior in connected vehicle systems has not been identified and analyzed. Therefore, this paper is aimed at developing the new concept of location-based driving volatility (LBV) via using BSMs exchanged between connected vehicles in real-world and linking it to historical crash data with the purpose of identifying hazardous spots proactively. Although the novelty of this study is in using high volume and high velocity connected vehicle data, the significance of works done by other researchers on crash frequency cannot be overlooked, given the emergence of new approaches, e.g., see Lord & Mannering (*86*). Also random parameter and/or varying coefficient models have become popular as opposed to fixed parameter for their capability to address unobserved heterogeneity (*54; 87; 88*).

Research Objective and Contribution

The objectives of this study are to:

1) Quantify instantaneous driving decisions and its variability in intersection-specific Basic Safety Messages (BSMs).
2) Understand the relationship between intersection-specific volatility with crash frequencies, while controlling for other variables, using rigorous statistical tools.

The present study contributes by analyzing real-world large-scale connected vehicle data to extract critical driving behavior information embedded in raw BSMs. Such an analysis is important because driving actions and behaviors are believed to be the main cause of traffic crashes, and understanding the relationship between location-based volatility and historical crash data can provide fundamental knowledge regarding proactive safety countermeasures. A unique aspect of this study is that significant efforts have been undertaken to integrate large-scale connected vehicle data (more than 65 million BSMs) with intersection crash and inventory data in order to provide providing a more complete picture of operations and safety performance of intersections. The assembled database allows investigation of correlations between potentially leading indicator of safety (location-based volatility) and historical crash frequencies. By taking the first step towards proactive safety using large-scale connected vehicle data, the current study is original and timely in sense that real-world data has been processed and used to understand the phenomena under discussion.

Methodology

Conceptual Framework

The two-month connected vehicle data from Safety Pilot Model Deployment (SPMD) (https://www.its-rde.net/home) contains rich information (i.e., basic safety messages in 10 Hz) that was exchanged between vehicles/infrastructure through vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) applications. Such data provide us with an opportunity to scrutinize the mechanisms that lead to unsafe events on roadways. However, the methods of making a good use of such high-volume and high-resolution data need further development. SPMD collects Basic Safety Messages (BSMs) that describe a vehicle's position, motion, its component status, and other relevant travel information (*89*). However, BSMs are not informative to drivers when they need to make decisions based on information received through V2V or V2I applications. Most BSMs describe normal driver behaviors while abnormal and highly fluctuating driver behaviors determine the safety of driving in the short-term.

This study is focused on developing an innovative methodology for estimating location-based volatility for specific intersections and comparing it with their observed crashes. We hypothesized that the nature of extreme instantaneous driving behaviors at

intersections can be correlated with their crash history. Such correlations can help us understand instantaneous driving behaviors and how they relate to transportation safety. Location-based volatility (LBV) represents the driving performance of a substantial number of users traveling through a specific location. LBV may play a critical role in highway safety management, as it will highlight locations where many drivers behave differently from other locations. Proactive countermeasures can be considered in such locations. If many drivers make extreme driving behaviors or if driving behaviors are highly fluctuating at certain locations, the reasons of such extreme behaviors may be related to factors such as the road conditions. Such information can be disseminated to connected vehicle drivers through roadside equipment (RSE) which are able to send information to vehicles, and thus drivers may be alerted about potential hazards (e.g. conflicts/intersection sight distance) while traveling through certain intersections.

First, the connected vehicle data consisting of geo codes and longitudinal acceleration were cleaned. In the next step, 116 intersections were identified in Ann Arbor, Michigan (discussed later). Crash data along with other geometric elements were collected. Then, four different coefficients of variation ($C_{V_{AL}}, C_{V_{AH}}, C_{V_{DL}}, C_{V_{DH}}$) are calculated and used as measures of location-based volatility (LBV) for each intersection (150 ft. from the center of each intersection). Given the hypothesis that higher LBV is likely to be positively correlated with historical crash data at intersections, appropriate statistical models are developed to investigate the correlation between LBV (among other traffic exposure factors) and crash frequency. The knowledge generated from the modeling results can identify intersections where drivers, on average, show higher volatility in their instantaneous driving decisions (e.g. longitudinal acceleration), and where such volatilities are found to be correlated with crash frequency. By carefully analyzing high-resolution real-world data transmitted between connected vehicles and applying appropriate statistical methods, we can ultimately generate proactive (rather than the traditionally reactive safety approach) alerts and warnings given to vehicle drivers at intersections. Such proactive warning and alerts can be disseminated through roadside equipment to vehicles approaching specific intersections to warn them regarding the chance or ranking of intersection in terms of crash occurrence. In the next section, the computation of LBV is discussed.

Location Based Volatility

Understanding instantaneous driving volatility at specific intersections is one of the most challenging aspects of the current study. To calculate location-based volatility, different instantaneous driving measures can be used such as accelerations, steering angles or position of brakes (*2*). As explicitly discussed in Liu and Khattak (*2*), volatility in trip-based instantaneous driving decisions should be captured by considering both longitudinal and lateral accelerations. Considering longitudinal acceleration as the only measure of driving

volatility can mask important information embedded in instantaneous driving data. For instance, at moments longitudinal acceleration can be low and thus considered normal, but the driver could still be volatile due to large magnitudes of lateral accelerations.

To calculate LBV, the authors intended to use longitudinal and lateral acceleration as they are direct outcomes of vehicle maneuvering. However, due to a considerable amount of questionable lateral acceleration data (see Data Accuracy section), only longitudinal acceleration data were used. The longitudinal acceleration data is reasonable and available for all BSMs and has been error checked by estimating accelerations from speed trajectories of the vehicles. Given the data limitation, this study only focuses on capturing location-based volatility by using longitudinal accelerations. There are two reason for this decision: First, excluding lateral acceleration does not seem to be affecting the results drastically since lateral acceleration is more informative in trip based volatility calculation where curvature of the road changes and where the length of the trip allows several lane changes. Second, using the data with removed lateral acceleration reduces the amount of data for several intersections leading to reduction of sample size i.e. number of intersections.

Calculation of LBV

The present study uses a standardized measure of dispersion called Coefficient of Variation (C_V) (also known as the ratio of relative standard deviation) for quantifying the fluctuations in longitudinal acceleration /decelerations at a specific intersection. Note that different measures such as range, interquartile range, variance or standard deviation can be used for capturing variability in longitudinal accelerations. Although standard deviation and variance are preferable as whole information embedded in the data is used for calculation of variability, both measures are insensitive to magnitude of acceleration values in the data. Thus, we prefer the relative measure of dispersion (Coefficient of Variation), where the dispersion in accelerations or decelerations can be quantified as the proportion of their means. This approach can capture the variability (e.g. standard deviation) in instantaneous driving decisions with respect to the mean accelerations or decelerations undertaken by different drivers at a specific intersection.

To compute volatility for each intersection, two speed bins (see Figure 13a), one from minimum observed speed to the mean and one from the mean to maximum speed were considered. The rationale behind considering speed bins is that the acceleration capability of a vehicle depends on current vehicle speed i.e. at larger speeds the capability to accelerate decrease as compared to acceleration capability at lower speeds. For each bin within an intersection, acceleration and deceleration values are separated, and the means and standard deviations are computed. Finally, C_V as a measure of LBV is obtained by dividing standard deviations of accelerations to the mean, i.e., σ/μ. For each intersection, four C_Vs are reported as shown in see Figure 13a. The calculated C_Vs for a

specific intersection provide the relative measure of dispersion of longitudinal accelerations with respect to their means, and thus different intersections can be compared based on their C_Vs.

Modeling Approach

After quantification of volatility for each intersection, we investigate the correlations between location-based volatility (for each intersection), crash data, and other traffic related factors. Appropriate modeling can provide an empirical evidence as of how intersection location-based volatility relates to historical crash data. Given the count nature of crashes, Poisson and/or Poisson-gamma models (Negative Binomial) can be estimated depending on the mean and variance of crash data.

For a Poisson model, the probability of having a specific number of crashes "*n*" at intersection "*i*" can be written as (*57*):

$$P(n_i) = \frac{\exp(-\lambda_i)\,\lambda_i^n}{n_i!} \tag{1}$$

where $P(n_i)$ is probability of crash occurring at intersection "*i*", "*n*" times per specific time-period; and λ_i is Poisson parameter for intersection "*i*" which is numerically equivalent to intersection "*i*" expected crash frequency per year $E(n_i)$. The regression can be fitted to crash data by specifying λ_i as a function of explanatory variables such as location-based volatility, Annual Average Daily Traffic, and speed limits on major and minor approach. Formally, λ_i can be viewed as a log link function of a set of independent variables (*57*):

$$\ln(\lambda_i) = \beta(X_i) \tag{2}$$

where X_i is a vector of explanatory variables and β is a vector of estimable parameters.

Application of Poisson regression to over-dispersed crash data can result in inappropriate results. If mean and variance of crash data are not equal, corrective measures are applied to Equation 2 by adding an independently distributed error term $\in$. While presence of over-dispersion can be indicated by the mean and variance of crash data (*57*), formally a Lagrange multiplier can be performed to statistically test the existence of over- dispersion in Poisson model (*57*). The test statistic is defined as:

$$LM = \left[\frac{\sum_{i=1}^{n}[(y_i - \mu_i)^2 - y_i]}{2\sum_{i=1}^{n}{\mu_i}^2}\right]^2 \tag{3}$$

where: y_i are actual crash frequency for intersection "*i*", μ_i is expected crash frequency for intersection "*i*" as predicted by Poisson model, and n are number of observations. The null hypothesis is that Poisson regression is appropriate for the crash data at hand. Under this hypothesis, the LM test statistic should have chi-square distribution with degree of freedom equal one. If the asymptotic chi-square distribution obtained from Equation 3 is

less than critical chi-square of 3.84 at 95% level of confidence, Poisson regression should be favored, otherwise Negative Binomial regression can be more appropriate (*57*).

Finally, it is likely that the associations between key explanatory variables and crash frequency may not be consistent across intersections. The intrinsic unobserved heterogeneity can arise due to several observed and unobserved factors related to intersection crash frequency, which may not be available in the data at hand. This is referred to omitted variable bias in safety literature (*57*). Furthermore, if key variables are omitted from analysis and too few variables are included in the model, it is likely that location-based volatility (explanatory factor) can capture those effects and may not be the true association between location-based volatility and crash frequency. One way to address this issue is to allow parameter estimates to vary across observations (*57*). As such, random parameters can be included in the estimation framework as:

$$\beta_i = \beta + \varphi_i \tag{4}$$

where φ_i is randomly distributed term with any pre-specified distribution such as normal distribution with mean zero and variance σ^2.With Equation 4, the Poisson parameter in Equation 2 becomes:

$$\lambda_i|\varphi_i = e^{BX} \tag{5}$$

And, the Poisson parameter in Equation 2 in Poisson-Gamma model becomes:

$$\lambda_i|\varphi_i = e^{BX+\epsilon} \tag{6}$$

Finally, the following likelihood function for random-parameter model can be maximized through maximum simulated likelihood technique (*54*):

$$LL = \sum_i ln \int_{\varphi_i}^{i} g(\varphi_i) P(n_i|\varphi_i) d\varphi_i \tag{7}$$

where g(.) is the probability density function of randomly distributed term with pre-specified distribution such as normal distribution with mean zero and variance σ^2. More details on random parameter models can be found in (*54*).

Data

The data used in this study (retrieved from https://www.its-rde.net/home) are BSMs from vehicles participating the SPMD in Ann Arbor, Michigan. SPMD is a comprehensive data collection effort, under real-world conditions, at Ann Arbor test site with multimodal traffic hosting approximately 3,000 connected vehicles equipped with V2V and V2I communication devices. BSMs are frequently transmitted messages (usually at 10Hz)

that is meant to increase vehicle's situational awareness. At its core, the dataset contains vehicle's instantaneous driving statuses of vehicle's position (latitude, longitude, and elevation) and motion (heading, speed, accelerations).

To examine correlations, location-based volatility (LBV) data for each intersection (as explained earlier) are linked with historical crash data, annual average daily traffic (AADT) data for major and minor approaches, speed limits on major and minor approaches, and number of approaches at each intersection. Such data are publicly available at the website of the Metropolitan Planning Organization: http://semcog.org/Data-and-Maps. Out of all intersections in the Ann Arbor area, 116 intersections are identified for which connected vehicle data are available, i.e. connected vehicles pass through such intersections and generating enough data for calculation of LBV. Finally, five years of crashes (2011-2015) along with geometric factors and flows were extracted and linked to LBV for each intersection. Note that the data are not available in spreadsheet format, and thus significant efforts went into carefully extracting data manually and linking it to LBV for 116 intersections.

Data Accuracy

Based on the distributions of key variables provided in Table 7, the data seems to be of reasonable quality. To assure the accuracy of intersection data, after initial collection, another person checked 10% of intersection data randomly and no discrepancies were observed. Also, the descriptive statistics of intersection data in Table 7 provide reasonable difference between signalized and un-signalized intersections. The major inaccuracy of data is from the lateral acceleration as it is shown in see Figure 13b. Since 27,240,788 data points (42% of the data) had the maximum allowable value that can be recorded in DSRC devices (2g), lateral acceleration data was not used in the analysis.

Results

Descriptive Statistics

Table 7 presents the descriptive statistics of key variables used in modeling. The mean, standard deviation, minimum and maximum values are given for each variable which can help conceptualizing the distributions. Descriptive statistics are given for all the intersections (N=116) as well as for signalized intersections (N=53) and un-signalized intersections (N=63) separately. For all intersections, signalized, and un-signalized intersections, the mean five-year crash frequency is 7.56, 12.94, and 3.04. As expected, signalized intersections have significantly higher crash frequency (on average) than un-signalized intersections. This finding is in agreement with Abdel Aty and Keller (*90*) who found approximately 9.6 crashes per year at signalized intersections as opposed to only 2 crashes per year on un-signalized intersections (*90*).

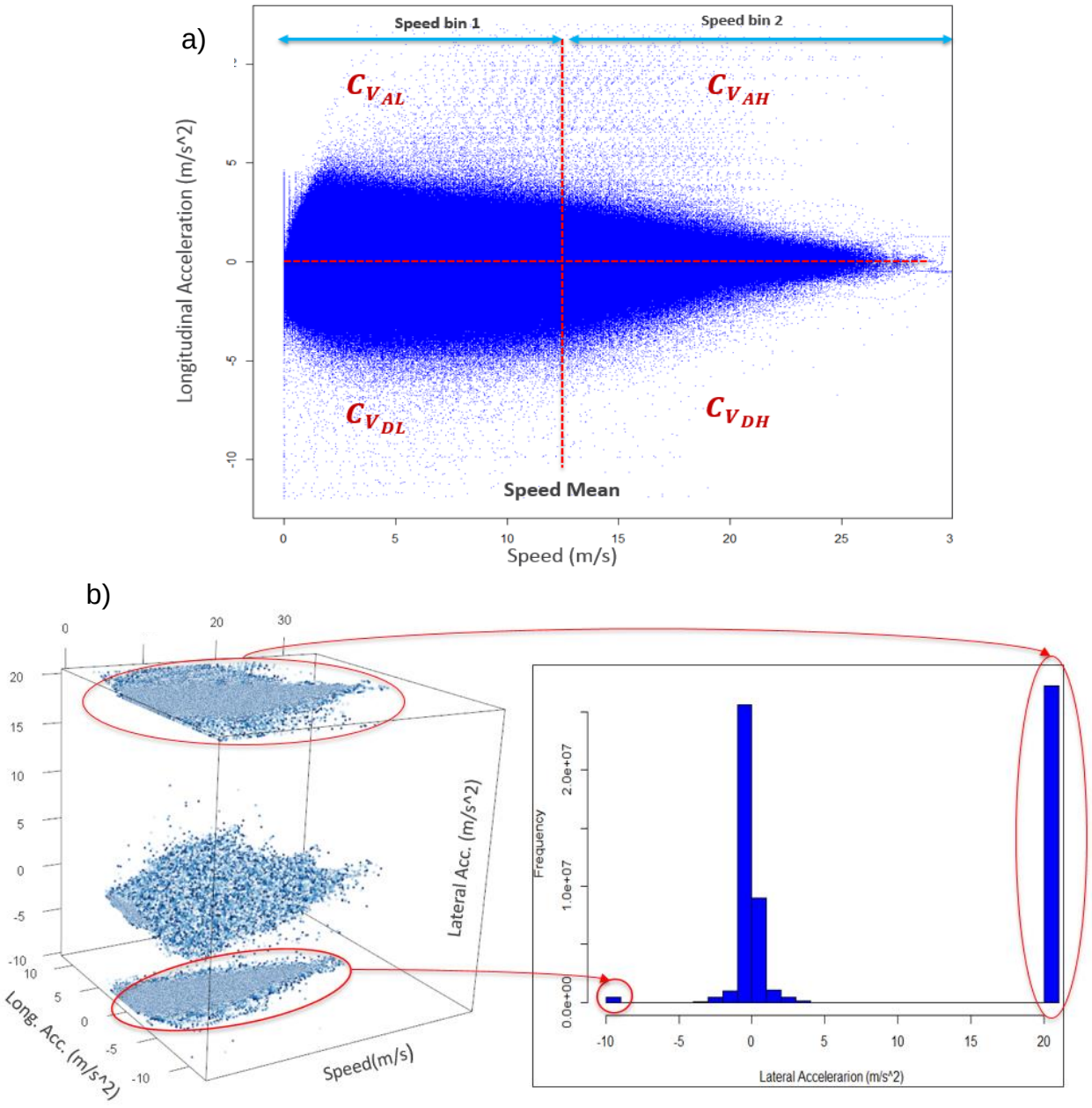


Figure 13. a) Four quadrants used to calculate coefficients of variation (=σ/μ) for each intersection, b) Plot of used data (left)/ Histogram of lateral acceleration (right)

There can be several factors which may contribute to occurrence of crashes at signalized intersections such as conflicting movements as well as different intersection-specific design variables (*90*). This said, investigating instantaneous driving actions at such locations, and higher volatility (if any) may help us design appropriate proactive strategies from preventing an "accident waiting to happen" (*91*).

Regarding location-based volatility, all C_V statistics suggest that signalized intersections on average have higher variability in longitudinal accelerations/decelerations compared with un-signalized intersections, and thus can be more volatile (this is the case for all C_V's except $C_{V_{AH}}$). One reason for higher $C_{V_{AH}}$ (volatility of acceleration above mean speed) of un-signalized intersections as compared to signalized intersections can be due to uninterrupted traffic of un-signalized intersections.

In order to avoid omitted variable bias in modeling (*92*), data on other variables such as five-year average AADT (major and minor approach), speed limits (major and minor approaches), and number of approaches were collected. Regarding the number of approaches, 40% of all intersections, 62.2% of signalized intersections, and 22% of un-signalized intersections are four-legged intersections (Table 7). In terms of exposure on major and minor roads, signalized intersections have higher (on average) AADT than un-signalized intersections (22,747 vs. 19,171 for major roads and 9,994 vs. 8,893 for minor roads). Regarding number of lanes, number of through and left turns for signalized intersection are considerably higher as compared to un-signalized intersections.

Modeling Results

For examining the correlations between crash frequency and location-based volatility (as measured by C_Vs), count data models are estimated given the count nature of crash frequency. Separate count data regression models are estimated for all intersections, signalized intersections and un-signalized intersections. Specifically, fixed-parameter Poisson regressions are estimated for total crash frequency as a function of location based volatility, major and minor road AADT, major and minor road speed limits, and total number of through lanes. It should be noted that the descriptive statistics for crash frequencies in Table 7 apparently reveal the existence of over-dispersion in the data where Negative Binomial model should be preferred over Poisson model (58).Thus, statistical tests are conducted to confirm the existence of over-dispersion (57). As explained in methodology section, Lagrange Multiplier tests were conducted for all three Poisson models. By using Equation 3, the Lagrange Multiplier (LM) values were 0.05, 0.031, and 0.15 for all intersections, signalized intersections, and un-signalized intersections respectively. The LM values are much smaller than critical Chi-square value of 3.84 for one degree of freedom at 95% confidence level.

Table 7. Description of key variables and descriptive statistics

Variables	All Intersections (N = 116)			Signalized (N = 53)			Un-signalized (N=63)		
	Mean	SD	Min/Max	Mean	SD	Min/Max	Mean	SD	Min/Max
Average crashes (5 years)	7.56	7.64	0/44	12.94	8.03	1/44	3.04	2.95	0/14
Average rear-end crashes (5 years)	4.28	4.56	0/24	7.07	5.24	1/24	1.93	1.79	0/9
$C_{V_{AL}}$ (In percent)	143.71	56.03	69/239	182.44	57.58	83/329	111.13	26.12	69/191
$C_{V_{AH}}$ (In percent)	84.9	13.76	56/121	77.93	12.7	59/113	90.77	11.8	57/121
$C_{V_{DL}}$ (In percent)	137.51	43	71/287	168.67	41.15	87/287	111.29	21.94	71/181
$C_{V_{DH}}$ (In percent)	96.29	12.9	57/155	99.44	14.86	76/155	93.64	10.39	57/115
AADT major road	20805	8326	3100/45400	22747	8209	3600/45400	19171	8131	3100/38900
AADT minor road	9396	4138	1100/27400	9994	5706	3100/27400	8893	1972	1100/13400
Ln (AADT major road)	9.84	0.49	8.03/10.72	9.96	0.39	8.18/10.72	9.74	0.54	8.03/10.56
Ln (AADT minor road)	9.05	0.47	7/10.21	9.07	0.52	8.03/10.21	9.03	0.42	7/9.50
Speed limit major	35.34	7.24	25/45	35.94	7.34	25/45	34.84	7.18	25/45
Speed limit minor	30.47	3.95	25/45	30.84	5.16	25/45	30.15	2.53	25/40
4-legged intersection	0.4	0.49	0/1	0.622	0.489	0/1	0.22	0.41	0/1
Total through lanes	4.45	1.28	2/8	5.13	1.35	2/8	3.38	0.9	2/6
Total left turn lanes	1.53	1.32	0/6	2.26	1.4	0/6	0.92	0.88	0/3
Total right turn lanes	0.93	0.78	0/4	1.11	1.01	0/4	0.79	0.48	0/2

Notes:$C_{V_{AL}}$: Coefficient of variation of acceleration below mean speed of intersection;$C_{V_{AH}}$: Coefficient of variation of acceleration above mean speed of intersection;$C_{V_{DL}}$: Coefficient of variation of deceleration below mean speed of intersection;$C_{V_{DH}}$: Coefficient of variation of deceleration above mean speed of intersection; AADT: Annual Average Daily Traffic; SD is standard deviation; Min is minimum value; Max is maximum value.

Thus, the null hypothesis that Poisson regressions are more appropriate is failed to reject, and it would be more appropriate to use Poisson regressions (*58*).

Due to the likely presence of unobserved heterogeneity in crash data (*54*) which may arise due to several unobserved factors, random-parameter Poisson models are also estimated. Fixed parameter models are estimated with standard maximum likelihood whereas random parameter models are estimated through simulated maximum likelihood with 200 Halton draws used for random-held parameters (*54*). Regarding functional form of random-parameters, log-normal, Weibull, uniform, and triangular distributions are tested with normally distributed random parameters giving the best fit and shown in this study.

The results obtained from fixed and random parameter Poisson model are presented in Table 8. Marginal effects are also provided for the random parameter models that translate unit change in crash frequency with unit change in explanatory variable. Compared to fixed-parameter models, random-parameter models resulted in better fit as of improved log-likelihood at convergence and McFadden's ρ^2 (Table 8) (*58*). While this study does not focus on methodological approaches for modeling intersection crash data, the predicted vs. actual values of crashes (Figure 14) are plotted and reveal statistical superiority of random parameter models in fitting the data at hand.

Discussion

Coming to the fixed-parameter estimation results for all intersections (Table 8), the results provide evidence that $C_{V_{AL}}$, $C_{V_{DL}}$, and $C_{V_{DH}}$ are positively associated (statistically significant at 95% confidence level) with crash frequency. However, $C_{V_{AH}}$ is negatively associated with crash frequency (at 90% confidence interval).It can be concluded, overall, volatility of deceleration regardless of speed range is positively associated with crash frequency. However, when it comes to acceleration, volatility at lower speed is more a significant factor as compared to volatility at higher speeds.

At signalized intersections, the association between $C_{V_{AL}}$, $C_{V_{AH}}$ and $C_{V_{DH}}$ and crash frequency is also positive and statistically significant. $C_{V_{DL}}$ for signalized intersection; however, it is negatively correlated with crash frequency. Referring to marginal effects for random parameter model in Table 8, on average one-percent increase in $C_{V_{DH}}$ is associated with 0.11 increase in crash frequency for all intersections and 0.089 increase in crash frequency for signalized intersections. These findings have implications for proactive intersection-related safety strategies. In addition, it is interesting to note the significantly higher marginal effect of acceleration C_V s for signalized intersections, implying that higher variability in acceleration at signalized intersections may potentially result in more crashes. Given that signalized intersections are typically observed to have more crashes (*90*), proactive intersection-customized strategies can be designed.

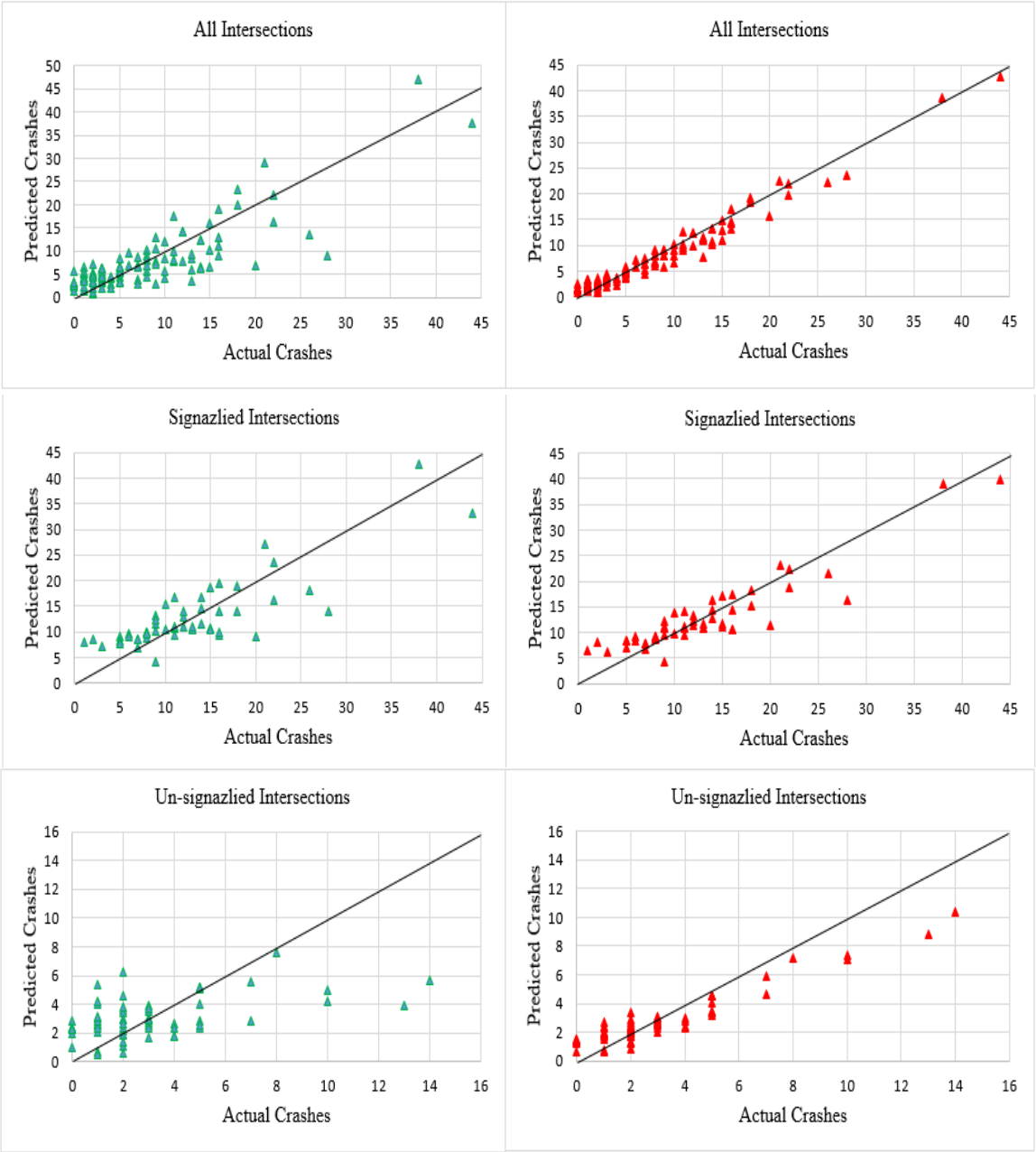


Figure 14. Mean-expected over actual number of crashes for fixed and random-parameter Poisson models (Green: fixed parameter models; Red: random parameter models)

For instance, proactive warnings and alerts can be generated about potential hazards at specific intersections and transmitted to drivers via connected vehicle technologies such as road-side equipment. This can in turn increase drivers' situational and safety awareness, and help drivers in undertaking safer driving behaviors.

Regarding un-signalized intersections, as shown in Table 8, $C_{V_{AL}}$ and $C_{V_{DL}}$ are statistically significant. We found negative association between $C_{V_{AL}}$ and crash frequency. This finding is seemingly counter intuitive and needs further investigation. Possibly, for un-signalized intersection, due to their uninterrupted traffic in major approach (78% of them are T-intersections), separation of 3-leg and 4-leg intersection might shed more clarification in future studies. However, the finding that $C_{V_{DL}}$ (Coefficient of variation of deceleration below mean speed of intersection) is positively associated with crash frequency is intuitive i.e. larger the volatility/variation in decelerations at low speeds, the more crash frequency at a particular intersection.

The estimation results quantify associations between major and minor road AADT and crash frequency. Referring to marginal effects from the random-parameter model, one-log unit increase in major road AADT is associated with 2.69, 6.57, and 1.82-unit increase in crash frequency for all intersections, signalized intersections, and un-signalized intersections, respectively. Minor road AADT is statistically significant in the random-parameter model for signalized intersections, but the relationships are not statistically significant for un-signalized intersections (Table 8). Speed limit on major roads is negatively associated with crash frequency for all intersections. These findings are consistent with past studies on this topic (*67; 93*). Notably, the total number of through lanes is positively associated with crash frequency. From Table 8, it can be observed that one added through lane is correlated with 0.547 more crashes.

Figure 15 illustrates how the study results can assist in proactive intersection safety management. The black, green and red circles in the figure are scaled crashes, volatility of acceleration, and volatility of deceleration at lower speeds, respectively. The intersection in the center is a known hotspot because it has more crashes and proportionately high levels of volatility. However, two other intersections shown in dashed ellipses have relatively low crashes but high volatility levels ($C_{V_{AL}}, C_{V_{DL}}$). In such locations (hotspots), although crash frequencies are low, drivers show proportionately more volatile driving behavior. In other words, at such locations crashes may be waiting to happen. Proactive countermeasures can be taken in those locations depending on the real cause of driving volatility, e.g., by studying speed limits, signal timing, geometric design, dilemma zone, and lines of sight.

Table 8. Modeling results of fixed- and random-parameter Poisson regressions

Variables	Signalized and Un-signalized					Signalized Intersections					Un-signalized Intersections				
	Fixed Par.		Random Par.			Fixed Par.		Random Par.			Fixed Par.		Random Par.		
	β	t-stat	β	t-stat	M.E.	β	t-stat	β	t-stat	M.E.	β	t-stat	β	t-stat	M.E.
Constant	-7.752	-6.6	-7.786	-7.237	---	-7.21	-4.975	-7.35	-6.958	---	-10	-3.574	-9.61	-3.235	---
*Standard deviation**	---	---	---	---	---	---	---	---	---	---	---	---	0.488	6.155	---
$C_{V_{AL}}$	0.006	4.152	0.004	2.902	0.025	0.009	3.434	0.01	5.346	0.125	-0.014	-2.831	-0.016	-2.911	-0.035
Standard deviation	---	---	---	---	---	---	---	0.0002	1.991	---	---	---	---	---	---
$C_{V_{AH}}$	-0.003	-0.776	-0.007	-1.983	-0.038	0.009	1.453	0.01	1.959	0.118	0.005	0.683	0.004	1.28	0.01
Standard deviation	---	---	0.005	11.856	---	---	---	---	---	---	---	---	---	---	---
$C_{V_{DL}}$	0.002	1.243	0.005	2.827	0.027	-0.003	-1.541	-0.004	-2.222	-0.057	0.015	2.698	0.0153	3.186	0.036
Standard deviation	---	---	---	---	---	---	---	0.0009	4.363	---	---	---	---	---	---
$C_{V_{DH}}$	0.02	6.449	0.021	6.33	0.11	0.008	1.872	0.007	1.981	0.089	-0.0007	-0.09	0.0001	0.05	0.0004
Standard deviation	---	---	0.0007	2.182	---	---	---	---	---	---	---	---	---	---	---
Ln (Major Road AADT)	0.547	4.899	0.527	5.322	2.694	0.55	3.716	0.565	5.561	6.575	0.866	4.801	0.757	4.106	1.823
Standard deviation	---	---	0.011	3.376	---	---	---	---	---	---	---	---	0.488	6.155	---
Ln (Minor Road AADT)	0.123	1.656	0.15	1.97	0.767	0.191	2.083	0.207	2.03	2.413	0.231	1.004	0.292	1.25	0.704
Standard deviation	---	---	0.006	2.152	---	---	---	---	---	---	---	---	---	---	---
Speed limit major road	-0.009	-1.736	-0.014	-2.497	-0.073	0.004	0.576	0.008	1.227	0.097	---	---	---	---	---
Speed limit minor road	---	---	---	---	---	-0.016	-1.444	-0.023	-1.62	-0.271	---	---	---	---	---
Total through lanes	0.61	1.733	0.107	3.223	0.547	---	---	---	---	---	---	---	---	---	---
Summary Statistics															
Log-lik. at Zero *L(0)*	-578.31		-578.31			-226.73		-226.73			-158.18		-158.18		
Log-lik. at Convergence *L(β)*	-336.72		-305.02			-159.43		-154.91			-138.26		-130.44		
McFadden ρ^2	0.417		0.831			0.31		0.893			0.125		0.59		
Sample Size (N)	116					53					63				

*Notes: ME: Average Marginal Effects from Random Parameter Model. $C_{V_{AL}}$: Coefficient of variation of acceleration below mean speed of intersection; $C_{V_{AH}}$: Coefficient of variation of acceleration above mean speed of intersection; $C_{V_{DL}}$: Coefficient of variation of deceleration below mean speed of intersection; $C_{V_{DH}}$: Coefficient of variation of deceleration above mean speed of intersection; AADT: Annual Average Daily Traffic; *Standard deviation of normally distributed random parameters.*

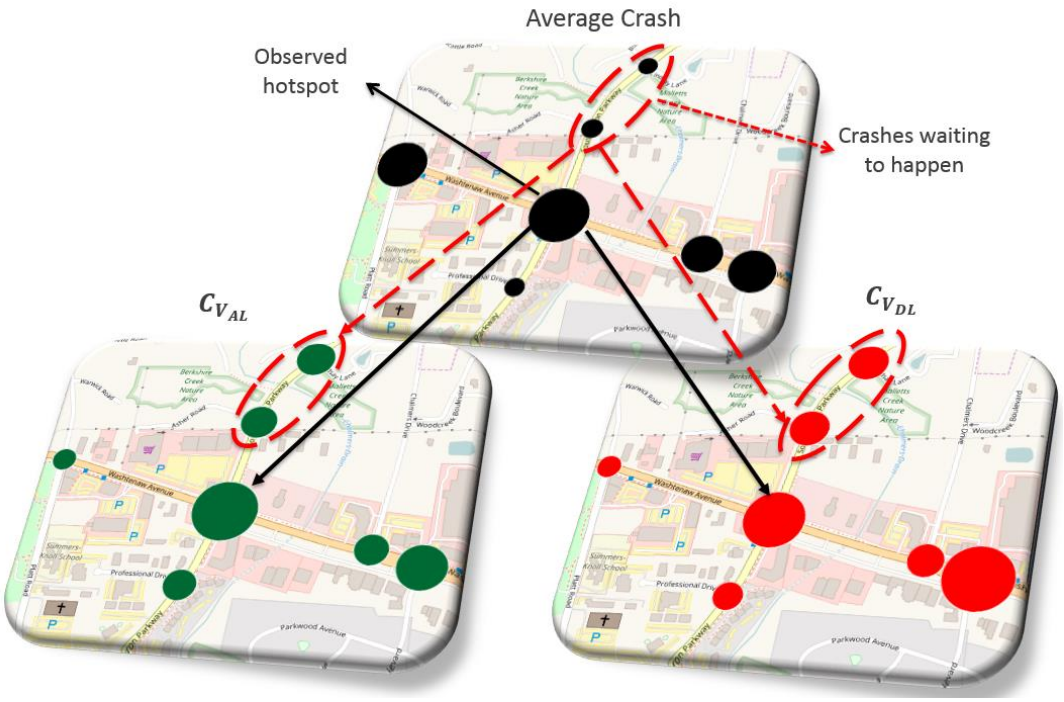


Figure 15. Known hotspots and spots where crashes are waiting to happen

Limitations

The study captures variability in longitudinal acceleration/deceleration as a measure of intersection-specific volatility, which only partially capture the true volatility exhibited by drivers. As explained in the methodology section, due to data limitations, the study could not incorporate lateral acceleration/deceleration in estimation of intersection-specific volatility. While the results from this study provide evidence between crash frequency and intersection-specific volatility, more robust measures such as vehicular jerk and combination of longitudinal and lateral accelerations can be used in future studies for quantifying volatility at specific intersections. Also, the results and conclusions of this study are dependent on the sample-size. Another limitation is that one month data were used to explain 5-year average crash. While the current sample size may not be enough to draw robust conclusions, the authors have used all available data for 116 intersections.

Conclusions

This study contributes by developing and demonstrating a proactive intersection safety methodology using real-world large-scale connected vehicle data. The study quantifies volatility in instantaneous driving decisions using intersection-specific Basic Safety Messages (BSMs) and its relationship with observed crash frequencies, while controlling

for other variables. Such a method can complement the state-of-the-art in evaluating intersection safety, which is largely reactive, based on observed and expected crash frequencies. The emerging data from Connected and Automated (CAVs) are increasingly becoming available, which can help us understand the detailed nature of instantaneous driving behaviors prior to the occurrence of unsafe outcomes such as crashes/incidents. This study proposes the concept of location-based volatility that captures the extent of variations in instantaneous driving decisions.

A unique database that provides a more complete picture of operations and safety performance was created by combining more than 65 million Basic Safety Messages transmitted between connected vehicles and roadside units at 116 intersections in Ann Arbor, Michigan, with crash and inventory data. The geo-coded raw BSMs were allocated to each intersection and the connected vehicles trajectories extracted from raw BSMs were plotted, revealing reasonable data precision and coverage. A simple and standardized measure of dispersion called Coefficient of Variation (C_V) (also known as the ratio of relative standard deviation) was used to quantify the fluctuations in longitudinal acceleration and/or decelerations at specific intersections. Five-year crash frequencies, AADT, speed limits, and number of approaches for all intersections are extracted and linked with location-based volatilities. Significant efforts went into data processing, collection, and linkage.

Rigorous fixed and random parameter Poisson regression models are estimated that allow consideration of unobserved heterogeneity in crash data. The modeling results reveal that most of computed C_Vs (as measures of volatilities) are positively associated with crash frequency. The study has implications for proactive intersection safety management. Importantly, the magnitude of association between location-based volatility and crash frequency is significantly higher for signalized intersections, implying that higher variability in instantaneous driving decisions at signalized intersections may potentially result in more crashes. This finding is important in the sense that if many drivers behave in a volatile manner at a specific intersection (exhibit higher variability in longitudinal accelerations), then such intersections can be identified before accidents happen. Of course, the reasons for volatile behaviors may be related to intersection and environmental conditions, vehicles' and drivers' conditions. Given that signalized intersections are typically observed to have more crashes (*90*), intersection-customized strategies can be designed to improve safety. Proactive warnings and alerts can be generated about potential hazards at specific intersections and transmitted to drivers via connected vehicle technologies such as road-side equipment; these can in turn increase drivers' situational and safety awareness, and help them pursue safer driving at dangerous intersections.

Acknowledgement

This paper is based upon work supported by the US National Science Foundation under grant No. 1538139. Additional support was provided by the US Department of Transportation through the Collaborative Sciences Center for Road Safety, a consortium led by The University of North Carolina at Chapel Hill in partnership with The University of Tennessee. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

CHAPTER IV
APPLYING MARKOV DECISION PROCESS TO UNDERSTAND DRIVING DECISIONS USING BASIC SAFETY MESSAGES DATA

A version of this chapter was originally submitted by Mohsen Kamrani, Aravinda Ramakrishnan Srinivasan, Subhadeep Chakraborty and Asad J. Khattak to *Transportation Research Part C: Emerging Technologies* and is under review at the time of this document submission. The paper also was presented at Transportation Research Board 96th Annual Meeting lectern session.

Abstract

While a number of studies have focused on investigating driving behaviors, only recently, detailed microscopic driving data has become available for analysis. By studying Basic Safety Messages (BSM) data available from the Michigan Safety Pilot Program, this study applies Markov Decision Process (MDP) framework to learn driving behavior in terms of acceleration, deceleration and maintaining constant speed decisions. Personally Revealed Choices (PRC) that maximize the expected sum of rewards for individual drivers are obtained by analyzing detailed data from 120 trips and use of MDP. Specifically, eight states are defined based on the number of objects around the host vehicle and distance to the front object. Individual drivers' reward functions given the states are estimated using multinomial logistic regression to be used in the MDP framework. Value iteration algorithm is used obtain optimal policies (i.e. PRC). The results show, with increasing objects around the host vehicle, drivers would rather accelerate to escape the crowdedness around them. In addition, segmenting trips based on the level of crowdedness indicated that with increased level of trip crowdedness, fewer drivers choose acceleration as their PRC's because keeping constant speed or deceleration are imposed on them by the traffic condition. One potential application of this study is to generate short-term predictive information about driver decisions, which can be used to warn the host vehicle driver when he/she deviates substantially from their own PRC based on their own historical driving performance. Yet another is to disseminate preferred actions of a driver to surrounding vehicles, enabling surrounding vehicles to foresee the states and actions of other drivers' potentially avoiding collisions.

Introduction

Availability of detailed data about driving performance has provided opportunities to investigate new aspects of driving behavior. Vehicle motion (e.g., speed, acceleration) can be used to explain some aspects of driving behaviors. Since driving behavior is a general term and can be interpreted differently, for the purpose of this paper, we define it as instantaneous driving decisions in terms of acceleration, deceleration and maintaining constant speed. Driver behaviors are expected to vary based on contexts. Detailed models are available for the dynamics of vehicle components (*94; 95*). In some models, the driver is modeled as a feedback controller that seeks to achieve a particular control goal, such as tracking a reference (*96; 97*). In other cases, the driver is represented by

an autonomous system, often driven by a random process. For instance, a study (*98*) proposes a linear model with additional non-linearity such as actuator saturation, slew-rate, and time delays, (*99*) suggests a hidden Markov model and (*100*) proposes nonlinear ARMAX models. In (*101*), a hybrid driver model is proposed, which consists of discrete modes and continuous control functions.

In general, driving behavior models (*102*) describe vehicle movements under different traffic conditions. These models include speed/acceleration and lane changing models, which are critical components of microscopic traffic simulators. Other application areas, such as safety and capacity analysis where aggregate traffic flow characteristics may be extracted from the behavior of individual drivers could also benefit from such models.

Early driving behavior models focused on car-following theory. These models explain the behavior of a vehicle following the lead vehicle where the subject vehicle is assumed to react to the leader's actions (*103; 104*). Recently, the development of microscopic traffic simulation models has led to the development of general acceleration models and lane changing behavior studies. General acceleration models (*105; 106*) define multiple driving regimes (e.g., free-flow, emergency) while considering different behaviors in each regime as well as various car-following types (e.g., reactive and non-reactive). For instance, drivers in the free-flow acceleration regime may focus on attaining their desired speed. In lane changing models (*107; 108*) there are typically two components: the decision to consider and then to execute the lane change maneuver. More recently, car following models driven by trajectory data, incorporating other contributing factors (e.g., distraction, reaction time) and addressing the heterogeneity among the drivers, have been developed (*102; 109-118*).

Finally yet importantly, there exist some studies using Inverse Reinforcement Learning to recover drivers' reward function and consequently their driving styles. For instance, Shimosaka et al. (*119)* tried to predict driving behavior by considering multiple reward functions and using maximum entropy method. Similarly, Kuderer, Gulati and Burgard (*120)*, studied the possibility of learning driving styles and navigation behavior of the driver through demonstration. In another study drivers' car following behavior are studied using Continuous Inverse Optimal Control (CIOC) (*121*). This study benefits from the availability of BSM data, where the authors explore driving behaviors from a different perspective. The aim is to extract the drivers' personally revealed choices (PRC's) from their acceleration/deceleration profiles. By treating instantaneous driving decisions as the realization of an optimal policy in an MDP framework, it is possible to define the states over time in terms of objects surrounding a vehicle. Actions are taken by the driver (i.e. acceleration, braking or maintaining speed) and values of the decisions are derived, which can be quantified in terms of accumulated discounted rewards. When the expected sum of the rewards is maximized for drivers, their personally revealed

choices can be inferred, given different states. MDP is well suited for this study because of the Markov property of instantaneous driving decisions as well as the stochasticity of their outcomes. A stochastic process has the Markov property if the probability of future states (conditioned on both past and present states) depends only upon the present state, not on the sequence of events that led to it. The methods applied in this paper can potentially form a foundation for human driver personally revealed choice extraction using field-collected empirical data. Moreover, knowledge from studying individual driving behaviors can be used to generate alerts and warnings for the driver of the host vehicle and to be passed via with the purpose of improving safety.

From the methodological standpoint, the paper contributes by using high volume and diverse driving data to learn driving decisions. Specifically, the reward and states are defined theoretically and real-life driving decisions are analyzed to explore their correlations with contextual factors i.e., proximity to surrounding objects. The study reveals personal preferences of drivers in each state. Since the states are defined based on the number of objects around the host vehicle and distance to the front object, the MDP reveals PRC's based on the level of the crowdedness around the host vehicle by accounting for the Markov property of drivers' decisions and the stochasticity of their outcomes.

Methodology

To understand short-term behavior, driver decisions of acceleration, deceleration, and maintaining constant speed are conceptualized using the Markov Decision Process (MDP) (*122*). A specific structure needs to be imposed on the MDP framework to model drivers' behaviors in terms of different maneuvers, which is briefly explained below.

Markov Decision Processes

The MDP (*122*) is defined by a set of states ($s \in S$), a set of all possible actions ($\mathrm{a} \in \mathrm{A}$), a transition function ($T(s, a, s')$), a reward function (R(s)), and a discount factor (γ). To make the model mathematically tractable, the discount factor is restricted to $0 < \gamma < 1$. The tuple $\{S, A, T, R, \gamma\}$ is used to represent an MDP. The basic idea is that the agent (e.g. decision maker, driver, etc.) takes an action based on the current state and an intuitively determined policy. The outcome of the action is stochastic and is parameterized as the transition probability function. Having transitioned to a new state, the decision maker accumulates some reward associated with that state. The process of decision-making continues infinitely. Generally, the term "Markov" implies that given the present state, the future and the past are independent. Specifically, for an MDP, this means that the future outcomes depend only on the current state and action performed. A policy $\pi(s)$ is a prescription for action to be taken, given a current state. The total expected discounted reward for the agent following the policy is defined as (*123*):

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} T(s, \pi(s), s')\, V^{\pi}(s') \quad (1)$$

The equation quantifies how valuable is the state (s) under the policy($\pi(s)$).
Likewise, the value or utility of a state-action pair $\left(Q^{\pi}(s,a)\right)$ is given by the following equation (*123*):

$$Q^{\pi}(s,a) = R(s) + \gamma \sum_{s'} T(s,a,s')\, V^{\pi}(s') \quad (2)$$

Policy $\pi(s)$ is referred to as the optimal policy $\pi^{*}(s)$ *(*hereafter PRC*)*, if it satisfies the Bellman's optimality equation (*122*) given by:

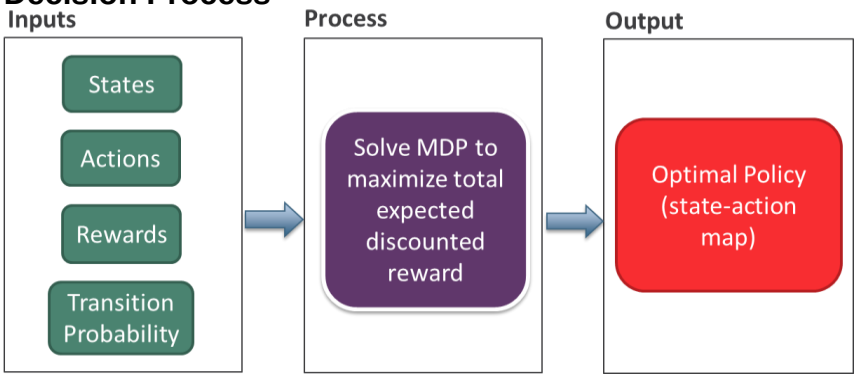
$$\pi^{*}(s) = \underset{a}{\operatorname{argmax}}\, Q^{*}(s,a) \quad (3)$$

A PRC can be defined as a policy that maximizes the expected cumulative reward an agent can achieve for the given MDP. In other words, no other choice can provide a more expected cumulative reward for all defined state-action pairs of the given MDP. Later, more details on how to solve the MDP problem and find the optimal policy is discussed.

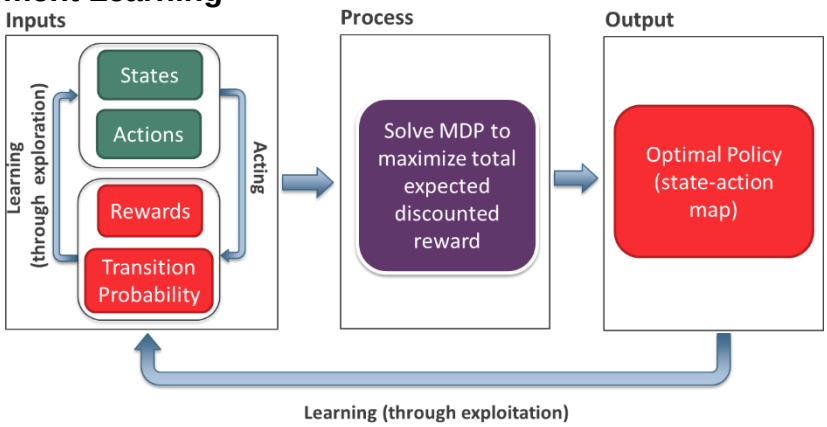
MDP and Reinforcement Learning Distinction

Before moving to MDP in driving context, discussion of the distinction between MDP and reinforcement learning (RL) as well as this paper's approach is useful. In MDP framework shown in Figure 16.A, states, actions, rewards and transition probabilities are given and are used to obtain the optimal policy under each state. In RL however, as shown in Figure 16.B, either reward or transition probabilities or both are unknown (actions and states are given). Therefore, in RL, the agent learns the reward and transition probability through acting and experiencing i.e. exploration. That said, a typical RL algorithm comprises two steps. First is to learn the missing MDP model elements, which can be reward, transition probability or both. Second is to solve the learned MDP to obtain the optimal policy given the states. Figure 16.B depicts the framework of an RL where the learning comes from a combination of exploration and exploitation. A learning solely through exploration might not maximize the reward because the agent tries to obtain more information about the reward by trying out things with unknown reward. On the other hand, a pure exploitation (making the best actions given the state obtained from the previously solved MDP), might lead to being stuck with policies with small amount of reward. Typically, a trade-off between exploration and exploitation is used to benefit from both.

A. Markov Decision Process



B. Reinforcement Learning



C. This paper approach

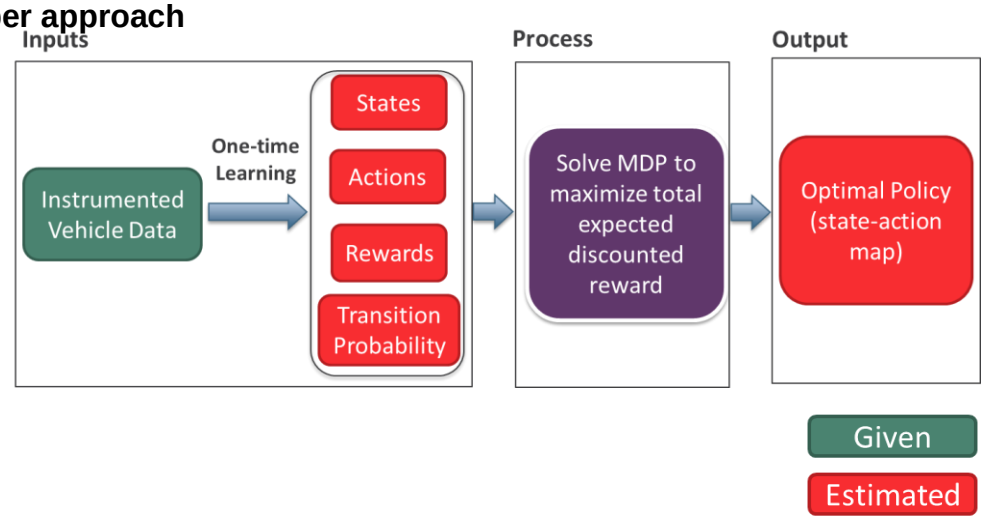


Figure 16. Framework of Markov Decision Process (A), reinforcement learning (B) and the paper (C)

Figure 16.C presents the approach used by this paper. Using instrumented vehicle data, all the inputs for the MDP framework are obtained and used to find the optimal policy. In particular, by defining the states and actions, the transition probabilities and the rewards are learned from the data. This approach puts the paper somewhere between MDP and RL; after learning the reward and transition probability from observed data (RL), all inputs become available to obtain the optimal policy (MDP). The obtained reward, however, is different from the rewards of typical RL practices. The reward in RL is unknown to the agent only (known to the trainer, researcher, designer etc.) and once the agent acts and lands in a specific state, it realizes the reward associated with landed state and contributes that knowledge to the next time decision-making. The reward is in this paper however, is unknown to both the driver (agent) and researcher and the drivers did not receive feedback in terms of rewards to improve their actions to maximize their expected rewards. We estimated the rewards using discrete choice model. Therefore, individual drivers have different rewards. Ideally, to train drivers with the purpose of making better driving decisions, or to train autonomous vehicle computers with the purposes of behaving similar to human drivers or making safe decisions, RL is preferred where the rewards can be given to the agent (driver, computer) in real-time (*51*) in terms of distance to the front objects, some calculated driving volatility (*1; 124*), risk factors etc.

MDP in Driving Maneuvers Context

The MDP applied to driving comprises 1) the driver as the decision maker or agent, 2) states, which can be defined by surrounding traffic conditions, current speed of the vehicles, state of traffic along the route, and 3) a set of actions such as acceleration, deceleration, maintaining speed, lane change. Depending on the defined states and actions, MDP can provide a transition probability matrix and a reward function which is the numerical measures of payoff associated with the transition from one state to another. The state transition probability can depend on (but not limited to) the vehicle dynamics, current speeds of surrounding vehicles, minimum stopping distance, maximum acceleration or deceleration rate, the distance between vehicles, reaction time, time to collision etc.

Actions

The variability of drivers' actions can be attributed to several factors, including aggressiveness, gap acceptance, psychological states, mindsets, attitudes, and preferences. The preferable MDP set of actions for the explained setting is: accelerate, brake, change lane to the left, change lane to the right, and maintain speed. Among different driving decisions mentioned earlier, lane changing is excluded due to the limitations of variables in the available dataset. Therefore, the actions are:

- *A*: Acceleration; increasing the speed at the rate more than a specified threshold

- *C*: Maintaining constant speed; zero or any acceleration within a specific threshold
- *D*: Deceleration; decreasing the speed at the rate more than a specific threshold

The necessity for defining a threshold in high-resolution data and especially when the decisions are defined based on acceleration values is to overcome subtle deviations in speed (that may be noise and hard to attribute to drivers' decision). As an example, let us say that a driver has activated the cruise control or decided to maintain constant speed by not changing the acceleration pedal displacement. If the magnitude of the acceleration during that time is recorded at ten times/second, then zero acceleration values will rarely be observed. To remove noise, the data is aggregated over one second and a threshold is used to characterize drivers' decision of maintaining a constant speed. Values of accelerations that fall between -2.5% to 2.5% of the data around zero are considered constant speed. Accordingly, positive and negative values of acceleration are defined as acceleration and deceleration decisions respectively.

States

Figure 17 (left) shows an example to illustrate the states in instantaneous driving decisions. The red car (host) is in the state (*s*) where there are four vehicles around it. The driver decides to change lanes and overtake the vehicle in front of it (the yellow car). By taking the action of changing lanes, among all different outcomes (stochasticity), he/she ends up in the new state, denoted by s' (which also depends on other drivers' maneuvers and decisions). It should be noted in a given state, different drivers would make different decisions (i.e. take different actions). In the example, the driver of the red car decided to change lanes while even maintaining constant speed can change someone's state due to the dynamics of the surrounding traffic.

To show the states formally, a spatial layout of proximate vehicles can be used as shown in Figure 17 (right). The central cell of the matrix represents the host vehicle. The surrounding cells represent the surrounding space, which can be empty or occupied by a vehicle or other objects (0 for empty and 1 for occupied). There are eight cells around the host vehicle; each can be either 0 or 1 forming 28 = 256 different states. The number of possible states will decline if a vehicle is in the left or right lanes or where there are only two lanes in the same direction (32 states).

Determining the current state of the vehicle based on the structure shown in Figure 17 requires a complete awareness of the surroundings (360 degrees). The technology to provide such awareness and related data is already available in certain instrumented vehicles (IV). IVs use technologies such as cameras and laser scanners to obtain information about their surroundings to create vision-based high-density maps. In the data at hand, only the total number of objects around the host vehicle, without their spatial positions, and distance to the front object are available. Thus, the states are defined based on them. Due to one-second data aggregation, the average number of objects and

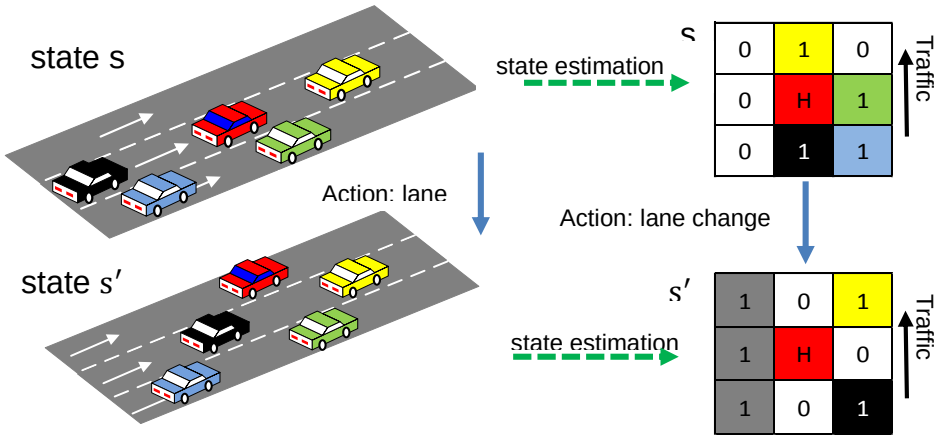
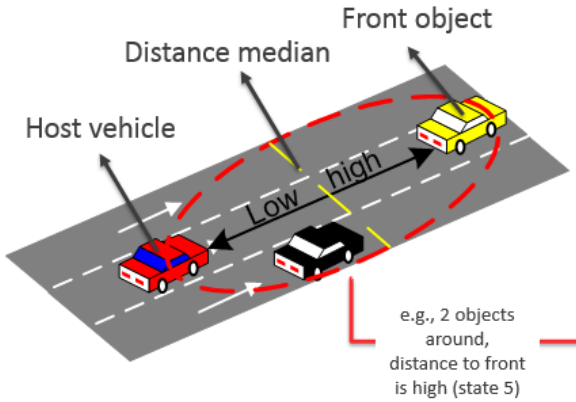


Figure 17. Examples of the driving states and actions taken by drivers

average distance to the front object (over one second) are used to define the states as shown in Figure 18. Figure 18 also provides an example of how to determine the host vehicle state in each second. In the figure, there are two objects around the host vehicle; therefore, according to the table on the right, rows 5 and 6 that correspond to the number of objects between two and three, [2, 3) should be referred to. Since the front vehicle (yellow car) is beyond the distance median, the current state of the host vehicle is determined as five.

Transition Probabilities

Figure 19 depicts an example of the MDP structure for this study where the states are shown by circles, decisions by dashed arrows (A: acceleration, D: deceleration, C: constant speed) and transition probabilities as solid arrows. For each set of initial states, action and landed state (s, a, s'), a transition probability is defined as $T(s, a, s')$. This quantifies the probability of ending up in state s' given that the agent's initial state was s and it took action *a* i.e. $P(s'|s, a)$. For each trip, having determined the states (according to Figure 18) and actions (according to acceleration values), the probabilities can be specified by counting the occurrence of each set (s, a, s') and dividing them by total number of occurrences. Therefore, there are three matrices (for acceleration, deceleration and constant speed) each 8 by 8, where rows are initial states (1 to 8) and columns are landed state (1 to 8). For example, the element (1, 2) in acceleration matrix indicates the probability of transitioning from state 1 to state 2 if the driver has accelerated.



State	Distance to front object	Ave. No. of objects (during 1 sec.)
1	High	[0 , 1)
2	Low	[0 , 1)
3	High	[1 , 2)
4	Low	[1 , 2)
5	High	[2 , 3)
6	Low	[2 , 3)
7	High	[3 , ...)
8	Low	[3 , ...)

Increased level of crowdedness

***Note**: Low and high in the table are different for each trip.*
***High**: Distance to the front object >= median distance*
***Low**: Distance to the front object < than median distance*

Figure 18. Definition of states based on number of objects and distance to the front object

***Note**: To avoid clutter, transitions from state 1, given acceleration decision and related rewards are shown.*
***A**: acceleration*
***D**: deceleration*
***C**: constant speed*
***Circles** (1 to 8): states*
***T(s, A, s')**: probability of transition from sate s to s' given acceleration decision*

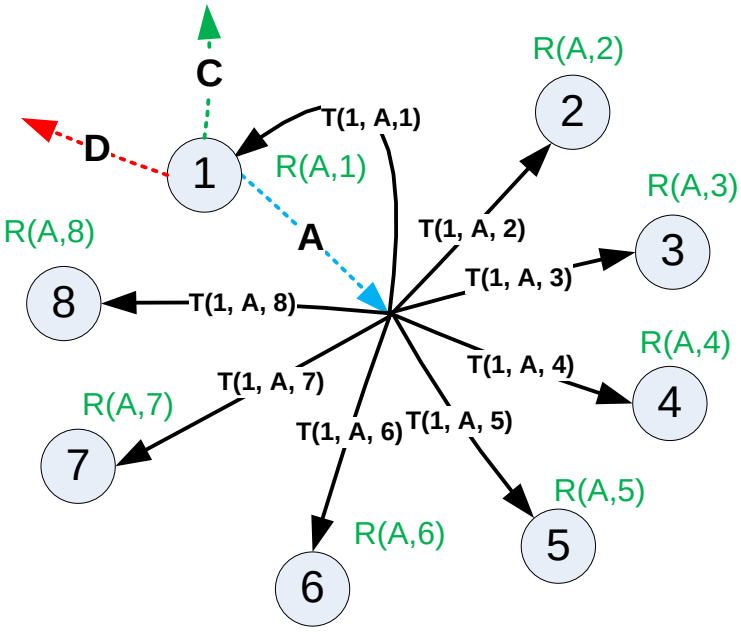


Figure 19. MDP structure used in the study

Reward

Consider an aggressive driver who tends to reduce the speed later than a calm driver does, or may drive in close proximity of other drivers. Such behaviors are likely to be personally optimal for him/her. Therefore, the reward in driving context is unknown, and it varies from one driver to another. Even for the same driver, it probably changes due to several factors such as the trip purpose, traffic condition, time of the trip etc. The goal is to find a reward structure that uncovers the PRC's based on the personal reward function, which the driver uses to make instantaneous driving decisions. In an MDP structure, the reward can be a function of states, actions taken, or both. In this study, we assumed the reward function to be dependent on both the action taken and the landed state.

Utility functions have long been used in transportation to represent and explain travel behavior (e.g. mode choice) (*125-128*). In such cases, there are clear economic arguments that go towards specifying the utility function. For example, a choice between using a bus vs. car for daily commutes is likely made by considering the respective cost and time, the comfort and convenience of the modes for specific trip purposes. To estimate the reward for each state-action pair, multinomial logistic regression (MNL) is applied separately for each trip. In other words, 120 MNL models are estimated by treating each second of individual trips as one observation. In this setting, observations are assumed independent from each other. The dependent variable (discrete choice) is the driver's action and current state is considered as independent categorical variable. Therefore, utility of choices are defined as follow:

$$U_{acc} = \beta_{0,a} + \beta_{1,a} state\ 1 + \cdots + \beta_{8,a} state\ 8 \quad (4)$$

$$U_{dec} = \beta_{0,d} + \beta_{1,d} state\ 1 + \cdots + \beta_{8,d} state\ 8 \quad (5)$$

$$U_{cons} = \beta_{0,c} + \beta_{1,c} state\ 1 + \cdots + \beta_{8,c} state\ 8 \quad (6)$$

where

$\beta_{0,j}$: estimated intercept for choice j, (*j = acceleration, deceleration, and constant speed*)

$\beta_{i,j}$: parameter estimates for state-specific dummies (*i = 1 to 8, j = acceleration, deceleration, constant speed)*

In the context of MNL, if a driver chooses acceleration in a given state, the utility of acceleration is highest among the other choices. The basic equation defining an MNL model is:

$$P_j = \frac{\exp\{U_j\}}{\sum \exp\{U_k\}} \quad (7)$$

where P_j is the probability of choosing j among k choices and U is the utility defined earlier. Equation 7 indicates if a choice has higher utility, its probability of being chosen

is higher. That said, by estimating the coefficients of the MNL, the probability of actions given the state are used as the reward for the MDP:

$$R(i,j) = \frac{\exp\{U_{ij}\}}{\sum \exp\{U_{ik}\}} \tag{8}$$

Using the probabilities obtained from the MNL as respective rewards is based on the assumption that the ***reward of an action given the state is proportional to the probability of that action given the state***. Now, we have all necessary inputs to solve the MDP problem, i.e. states, actions, transition probability and a reward function. Next, the solution approach called value iteration is discussed.

Value Iteration

In the context of this study, solving an MDP is equivalent to finding the PRC's (optimal action for individual states) for a given reward structure along with a transition probability matrix and a discount factor. The value iteration approach is used in this paper to find the PRC for each state. The steps of the value iteration algorithm are as follow:

Step 1: For each state (s ∈ S) initialize V(s) = 0
Step 2: Set a threshold (θ) as stopping condition
Step 3: For each state (s ∈ S) – (loop over states):
3.1 $\Delta \leftarrow 0$
Repeat - (loop over actions):
3.2 v ← V(s)
3.3 V(s) ← $max\ \sum_{s'} T(s,\pi(s),s')[R(\pi(s),s') + \gamma V(s')]$
3.4 $\Delta \leftarrow max\ (\Delta, |v - V(s)|)$
Until Δ < θ
Step 4: PRC = $\pi^*(s) = \underset{a}{argmax}\ \sum_{s'} T(s,\pi(s),s')[R(\pi(s),s') + \gamma V(s')]$

Given each state, the algorithm gives one of the actions (either acceleration, deceleration or constant speed) as the optimal policy. As discussed earlier, since reward function estimates are different from one driver to another, we call it Personally Revealed Choice (PRC) as opposed to optimal policy.

A toy example was used to demonstrate the performance of the algorithm. Specifically, Figure 20 depicts a grid world where an agent can either take actions of going up, going right or staying. The cells of the grid are indeed different states. The agent's action outcome is stochastic which is defined based on the transition probability matrix. It is assumed, if the agent decides to go up, with a 70% chance it ends up in the upper cell and has equal chances of 10% of either staying in its current state, or moving to left or right cells. Similarly, if it decides to go right, 70% of the time it ends up to the

right cell and there are equal chances of 10% to end up in its current cell, or above or below cells. In case the agent decides to stay, the agent will always remain in its current cell. According to possible defined actions (up, right, and stay), some of their outcomes are impossible in some states. For example, if the agent is in state two and decides to "right," given that agent's right and bottom sides are blocked, the probability of landing in the right cell (70%) and landing below (10%) will be distributed to other available outcomes equally (therefore, in this example, the probability of landing in 4 or staying in 2 become 50%, i.e. 10+(70+10)/2). The transition probabilities for each action are shown in Figure 21.The rewards for states (shown in parentheses) are zero except for states five and eight where their rewards are -1 and +1 respectively. This reward structure implies that the optimal policy is the one that moves the agent toward state eight to maximize cumulative discounted reward. Therefore, the optimal policy (shown in arrows) for the agent is to go up when in states 1 to 5, go right in states 6 and 7, and stay when in state 8. This toy example was introduced to the algorithm to assure that the algorithm also produces optimal policy shown as arrow in Figure 20. By changing the rewards, we also checked if the algorithm yields the respective intuitive optimal policies.

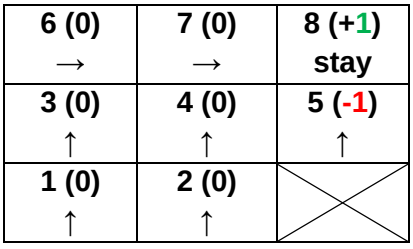


Figure 20. Grid world toy example
(***Note****: states from 1 to 8, reward are shown in parentheses, arrows represent optimal policies*)

Go up		Landed state (s')							
		1	2	3	4	5	6	7	8
Initial state (s)	1	0.15	0.15	0.7	0	0	0	0	0
	2	0.15	0.15	0	0.7	0	0	0	0
	3	0	0	0.15	0.15	0	0.7	0	0
	4	0	0	0.1	0.1	0.1	0	0.7	0
	5	0	0	0	0.15	0.15	0	0	0.7
	6	0	0	0	0	0	0.5	0.5	0
	7	0	0	0	0	0	0.33	0.34	0.33
	8	0	0	0	0	0	0	0.5	0.5

Go right		Landed state (s')							
		1	2	3	4	5	6	7	8
Initial state (s)	1	0.15	0.7	0.15	0	0	0	0	0
	2	0	0.5	0	0.5	0	0	0	0
	3	0.1	0	0.1	0.7	0	0.1	0	0
	4	0	0.1	0	0.1	0.7	0	0.1	0
	5	0	0	0	0	0.5	0	0	0.5
	6	0	0	0.15	0	0	0.15	0.7	0
	7	0	0	0	0.15	0	0	0.15	0.7
	8	0	0	0	0	0.5	0	0	0.5

Figure 21. Transition probabilities of the toy example (the transition probability of action "stay" is an identity matrix)

Data

The data used in this study is from Basic Safety Messages (BSM) sent/received by vehicles that participated in the Safety Pilot Model Deployment (SPMD) in Ann Arbor (Figure 22), Michigan (*129*). The data is stored in Department of Transportation ITS JPO (Joint Program Office) Data website (https://data.transportation.gov/Automobiles/Safety-Pilot-Model-Deployment-Data) maintained by U.S. Department of Transportation. The data was collected for two months (April and October 2012) in the real-world condition from more than 2800 vehicles equipped with DSRC devices that transmitted the instantaneous vehicle geocodes and kinematics such as speed, acceleration, heading, yaw rate, etc. The dataset used in this study is a subset of SPMD data, collected by vehicles equipped with Data Acquisition Systems (DAS). It includes host vehicle position (altitude, latitude, and longitude), motion (speed and acceleration), major components' status (accelerator, brakes, lights, cruise control, and wipers), and instantaneous driving contexts (surrounding objects and distance to front object with approximate detection range of 256 ft. /78 meters). There are 259 trips undertaken by 71 unique vehicles. Since we are interested in decision-making mechanism of drivers, the 10 Hz data (10 observations per second) were aggregated to one observation per second. Aggregation also helps remove some noise in acceleration values as discussed earlier. Having explored the data, 120 trips were found to be appropriate for the purpose of this study because to obtain a reasonable transition probability, the driver has to encounter different states during his/her trip. Figure 23 shows how the data needed for the analysis were prepared. From the value of the acceleration, drivers' decisions are determined (step 1). In addition, looking at the average number of objects and distance to the front object and referring to Figure 18, the current state of the vehicles are specified (step 2). Landed state, which is the stochastic outcome of the driver decision under given state, is the one-second lag of current state (step 3). Red arrows on Figure 23 show the process.

Results

The descriptive statistics of the 120 processed trips are provided in Table 9. The results seem reasonable (e.g., average speed is about 40 mph and average trip duration is 18 minutes). The results indicated that the data accuracy for these trips is reasonably high. On average, at the trip level, there is an object in front of the host vehicle in 77% of the time. The distance to front object on average, is 44.2 ft. with standard deviation of 20.43 ft.

Discrete Choice Model

MNL is used to estimate the rewards, by assuming that rewards are proportional to the probability of actions given the state.

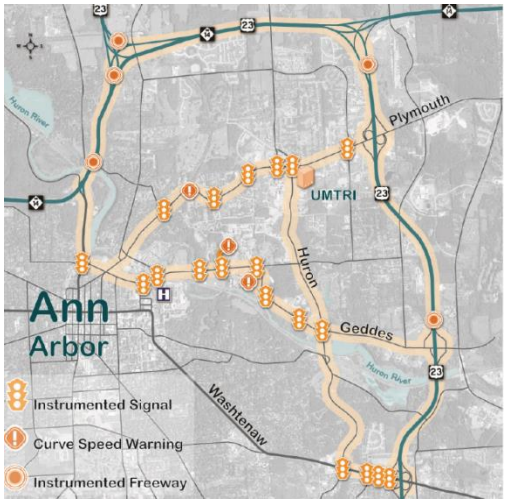


Figure 22. Safety Pilot Model Deployment Site, Ann Arbor, Michigan *(129)*

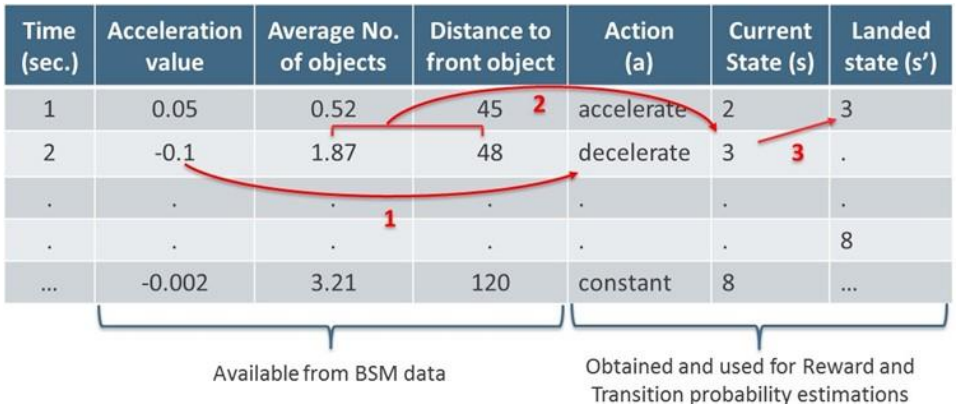


Figure 23. The data preparation steps

Table 9. Descriptive statistics of the selected trips (n=120)

Variable	Explanation	Mean	SD	Min	Max
Speed (mph)	Average vehicle speed during the trip	40.49	35.12	3.10	70.61
Acceleration (ft/s^2)	Vehicle acceleration during the trip	1.05	0.62	0.26	4.46
Deceleration (ft/ s^2)	Vehicles deceleration during the trip	-1.48	0.95	-6.20	-0.29
Duration (min)	Duration of the trip	18.16	32.96	1.32	217.53
Average Number of Objects	Average number of objects around the host vehicle during the trip	1.63	0.88	0.03	3.53
Object present (yes)	Binary variable indicating if there is an object in front of the host vehicle	0.77	0.25	0.03	1
Distance to Front Object* (ft.)	Average distance from the head of host vehicles to the front object during the trip	44.20	20.43	11.38	94.57

**Only for cases when a front object is present.*

For each of 120 trips, an MNL model is estimated where the choices are acceleration, deceleration and constant speed (as the base), and the independent variable is the categorical variable of state. Table 10 presents the descriptive statistics of the MNL estimates. The table indicates, on average, in states 2, 3 and 4, acceleration decision is less likely to be taken than constant speed (compared with state 1). From state 5 to 8, however, acceleration decision is more probable than constant speed. Similar interpretation can be done for deceleration decision using the mean of respective coefficients in the Table 10. The results show a wide range of estimates with high standard deviation across the trips encompassing positive and negative coefficients, making the results about overall drivers' preferences non-conclusive. Although the estimated coefficients of state-action pairs for individual trips could be used as the respective rewards, they all are zero for the base choice (constant speed) and they change if the we consider acceleration or deceleration as the base. Therefore, the probability of choices given the states, which is invariant to the base choice selection, is used as per Equation 8. Having estimated the rewards, the question is what action given state maximizes total discounted reward. Be noted that although drivers make decisions, they do not have control on the landing states because of the stochastic nature of driving environment and decisions outcome. Using the state-action transition probability coupled with the state-action rewards, the optimal policy (i.e. PRC) for individual drivers are obtained.

Personally Revealed Choices (PRC's)

For each trip, value iteration algorithm yields a PRC given the state. The estimated rewards along with other inputs are introduced to value iteration algorithm to obtain the optimum policies that maximizes total expected accumulated discounted reward. It is worth mentioning, there is no predetermined goodness for each state. This means a driver might be OK with other vehicles around while another driver would rather avoid them. These personal preferences are captured (some of them) through the reward values.

Importantly, as we assumed that reward is proportional to estimated MNL probabilities, one possible way of maximizing total accumulated reward is to select the actions with highest probabilities (i.e. highest reward) as the PRC's. With this naïve approach however, the stochasticity of decisions outcomes and sequence of decisions and the future expected rewards are not accounted for. For comparison purposes, we have provided the results based on this approach in Figure 24, left. Figure 24, right presents the PRC's results from MDP. We have shown the proportion of PRC's, which add up 100% for each state. Figure 24 caption provide an example of how to read the figure. The peaks and valleys in both figures are intuitive because for example, the average number of objects are the same for states 1 and 2 and the only difference is the distance to the front object i.e. in state 2 the front object is closer to the host vehicle.

Table 10. Descriptive statistics of MNL estimated coefficients (n=120)

Choice	Variable	Category	Estimate Mean	Estimate Std. Dev.	Min	Max
Constant Speed (base)		--	--	--	--	--
Acceleration	Intercept	--	0.97	3.86	-2.56	20.15
	State	1	--	--	--	--
		2	-0.68	4.8	-19.34	19.31
		3	-0.9	4.33	-18.77	14.53
		4	-0.39	5.05	-20.15	18.37
		5	0.15	5.53	-32.13	17.5
		6	0.58	7	-19.87	18.38
		7	0.44	6.1	-20.31	18.29
		8	0.23	7.07	-34.53	18.98
Deceleration	Intercept	--	0.21	4.48	-18.6	20.15
	State	1	--	--	--	--
		2	0.21	6.05	-20.15	36.68
		3	-0.43	5.22	-19.52	18.04
		4	-0.11	5.17	-19.25	18.93
		5	1.3	6.27	-19.72	19.73
		6	0.92	7.07	-20.15	19.61
		7	0.98	6.23	-19.81	20.15
		8	0.73	7.92	-35.22	18.98

This holds true for state pairs of (3, 4), (5, 6) and (7, 8). That naïve approach of PRC determination results does not show much trend in the proportion of PRC's. However, the MDP results show an ascending trend in the proportion of acceleration. Overall, acceleration is the most probable PRC throughout the states (ranges from 35% in state 2 to 58% in state 7) while proportion of deceleration does not change significantly across the states.

Drivers have different level of calmness and aggressiveness. An aggressive driver' PRC could be acceleration even when they are surrounded by three or more vehicles (i.e. state 7 or 8). Moreover, the layout of objects around the host vehicle also affects the PRC. If the objects were mostly in front of the vehicle throughout a trip, an intuitive PRC would suggest maintaining constant speed or deceleration. However, if the objects are in rear or sides of the host vehicle, then perhaps acceleration might be the PRC to avoid crash and crowdedness. That said, even the same layout of objects for two different drivers can result in different PRC's due to their differences in perceptions, information processing, situational factors, preferences, and experiences. Given all the complex factors contributing to PRC results, still the overall outcomes are reasonable and intuitive.

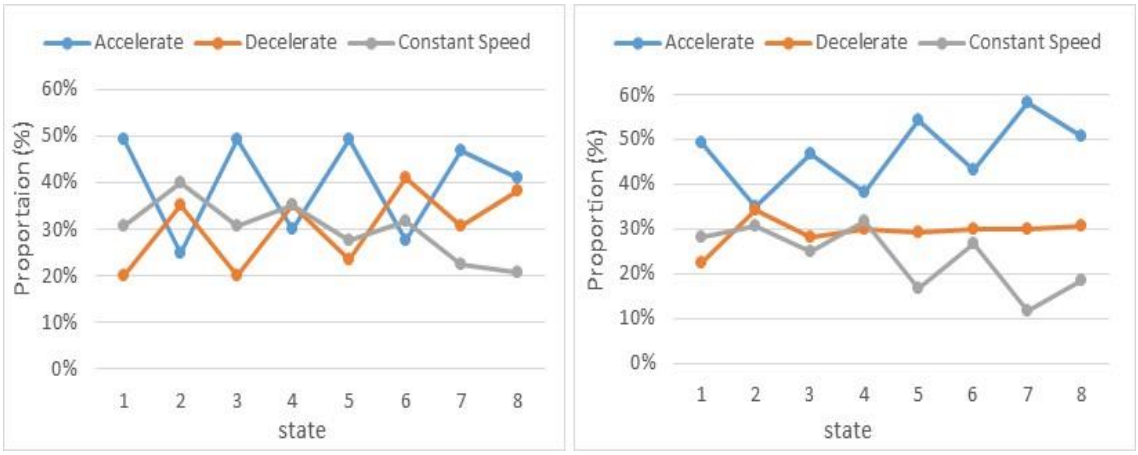


Figure 24. Personally Revealed Choices given specific states (left: native approach, right: MDP) (n=120 trips). Example of reading the figure (right): in state one, 49% of drivers' PRC was acceleration, 23% was deceleration and 28% constant speed.

PRC's at Different Levels of Crowdedness

To supplement information on PRC with respect to changes in the number of objects, trips were segmented based on the level of crowdedness. The idea is that on a crowded roadway where the average number of objects around the vehicle throughout the trips is higher, the distribution of PRC's may be different from a trip taken during less crowded conditions. Figure 25 presents the distribution of the average number of objects. Based on the distribution, three levels of crowdedness are defined: non-crowded, semi-crowded and crowded. Any trip with the average number of objects between 0 and 1 was classified as non-crowded. Likewise, trips with the average number of objects between 1 and 2 are categorized as semi-crowded and trips with the average number of objects more than 2 are considered as crowded trips.

In non-crowded trips (Figure 26), an ascending trend of acceleration proportion from state 2 to 8 is obvious. This observation is very interesting because as the surrounding of the host vehicle becomes more crowded, the majority of drivers' PRC is acceleration. The intuitive interpretation of this behavior is that they prefer to avoid (or escape from) crowdedness by increasing their speeds. How successful will be that decision? It depends on many factors such as instantaneous traffic conditions, surrounding drivers' decision, etc. It should be noted, although a PRC may not yield a drivers' desired outcome, the driver feels the highest personal satisfaction by making that decision as compared to other available options.

In semi-crowded trips results (Figure 27), state 4 is where that it there is a balance between the proportions of decisions. After that, the proportion of deceleration decision

remains almost the same and around 10% of constant speed decision is replaced by acceleration decision with increased level of crowdedness around the host vehicle. Comparing non-crowded with semi-crowded trips shows that the proportion of acceleration have almost shrunk to half in the latter which indicates more drivers considered either deceleration or constant speed as their PRC's.

In crowded trips (Figure 28) where drivers on average have more than 2 objects around them (i.e. a higher level of crowdedness compared to semi-crowded trip), the effect of trip crowdedness is more distinguishable. In crowded conditions, the acceleration decision is only dominant in states where the front object is far from the host vehicle. Be noted according to our state definition, the pair states of (1, 2), (3, 4), (5, 6) and (7, 8) have similar number of objects while in the higher state, the front object is closer to the host vehicle. The impact of front object can be easily seen by paying attention to the oscillations of acceleration deceleration and constant speed proportions across the states. In other words, closer front object increased the reward of deceleration and constant speed for some drivers making them PRC, which is intuitive. This change in the reward can be due to two reasons: first, some drivers decided intentionally to replace acceleration with deceleration and constant speed because of more crowded traffic conditions. Second, even though some drivers have not decided to do so, the traffic conditions imposed the above-mentioned replacement on them. In other words, the drivers may be forced to follow (or be more constrained) rather than having the option of making decisions freely.

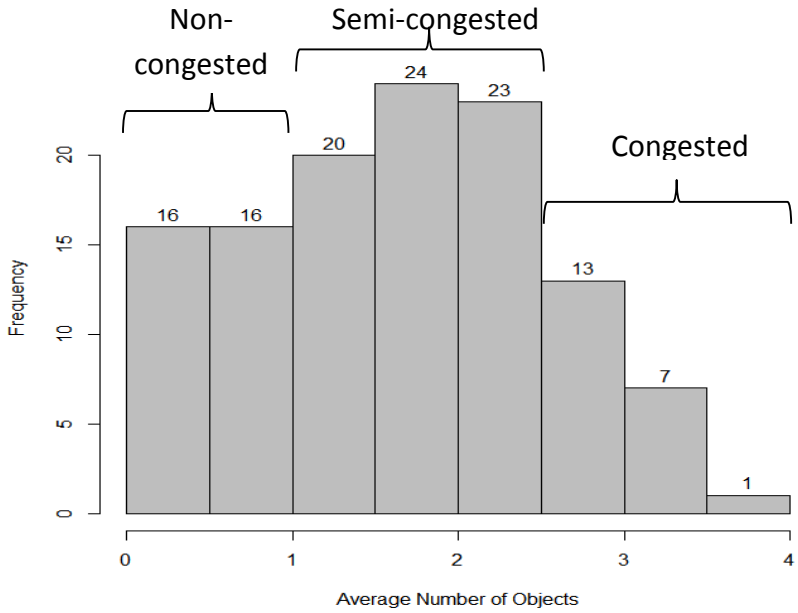


Figure 25. Distribution of average number of objects surrounding the host vehicle (n=120 trips)

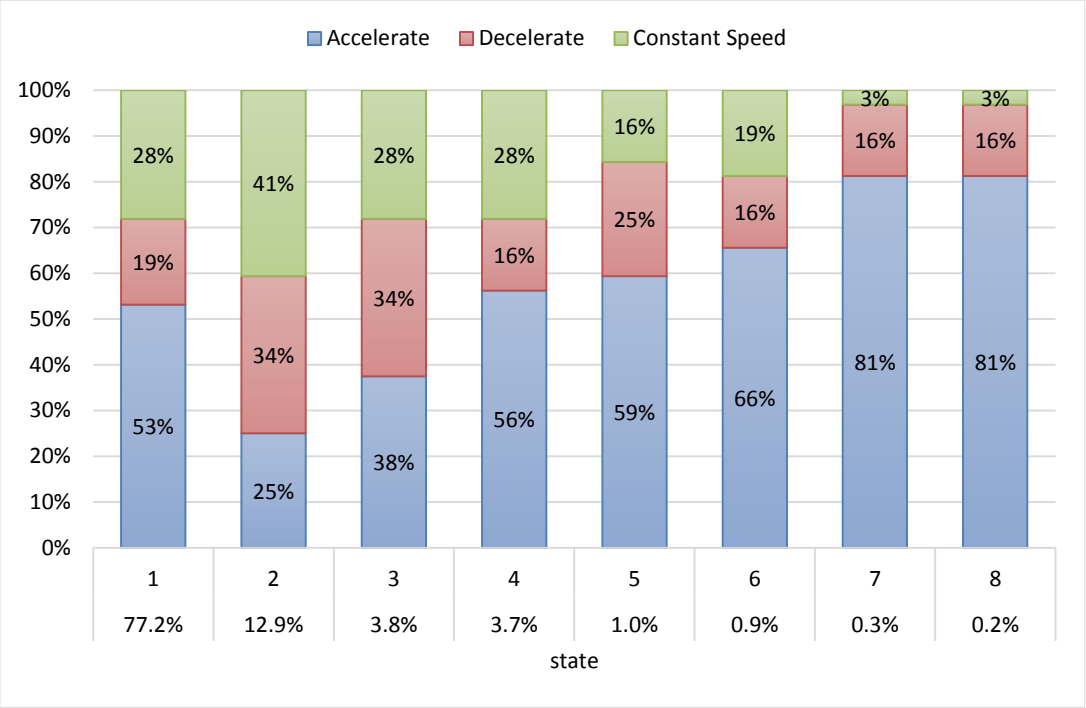


Figure 26. Personally revealed choices made on non-crowded trips (n=32)

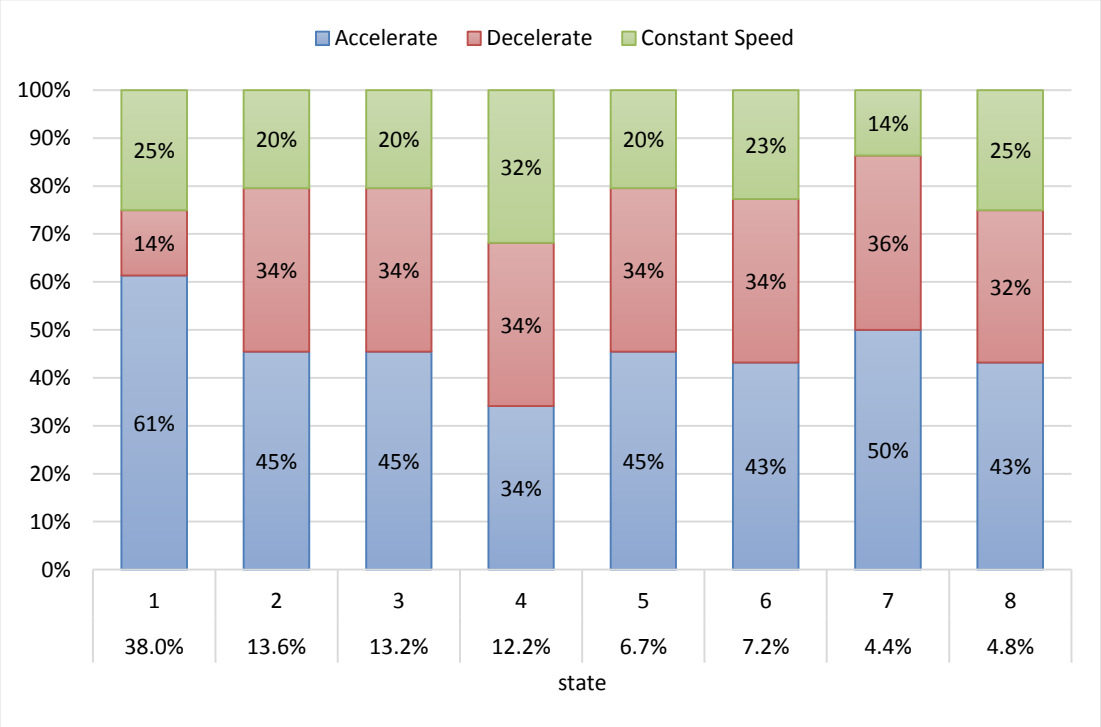


Figure 27. Personally revealed choices made on semi-crowded trips (n=44)

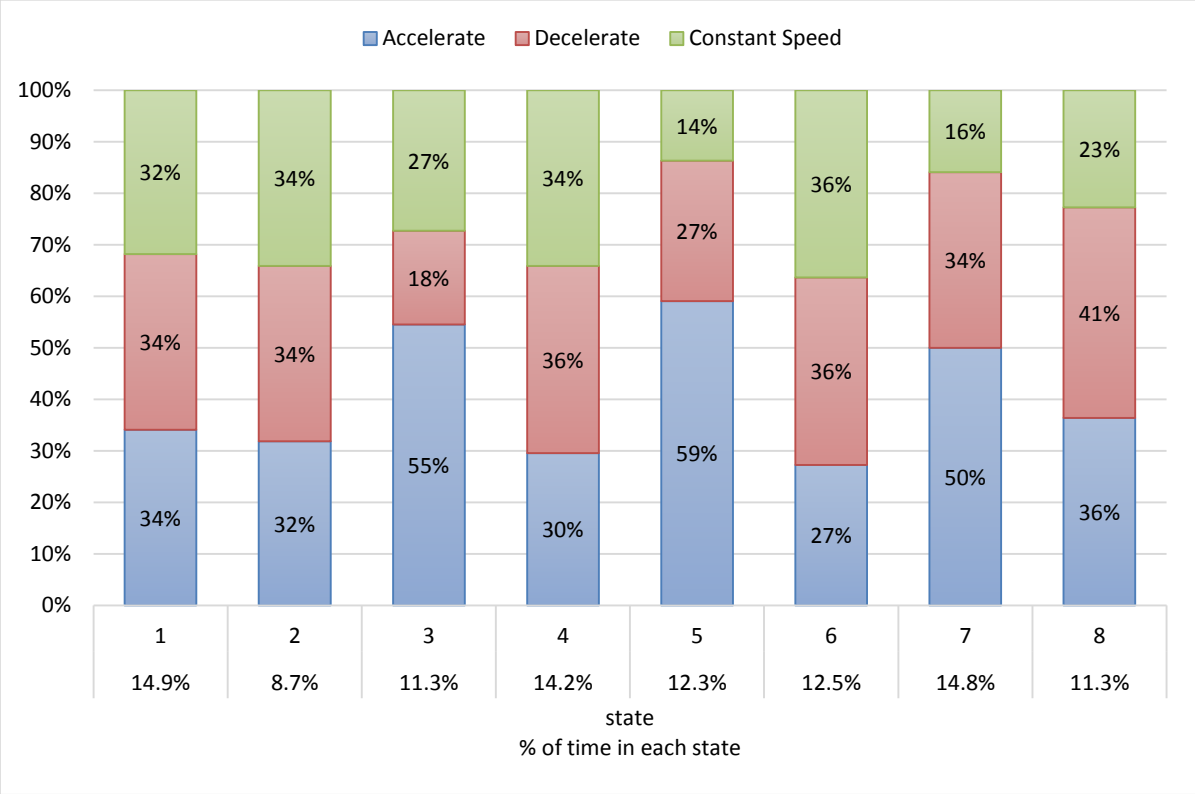


Figure 28. Personally revealed choices in crowded trips (n=44)

Comparison of PRC's and Real-world Observed Choices

Having found the PRC for each trip, it is beneficial to observe how long each driver could follow his/her PRC. It is evident that even though a PRC is obtained by the MDP method, drivers are not able to follow their PRC's all the time. For instance, a driver would prefer to accelerate in a particular state and that decision satisfice him/her, but the traffic condition (and many other hidden reasons) makes the driver compromise his/her PRC at that moment i.e. the driver fails to behave optimally. In fact, investigating the level of following PRC's can help us to see to what extent of drivers' behavior match their PRC's (i.e. how much of observed behavior are explained by PRC's). Figure 29 presents the distribution of following PRC's among 120 trips. The numbers above the bars indicate the number of trips and each bar label on the horizontal axis shows the percentage range of the time drivers followed their PRC's. For example, the bar shown in the red box in Figure 29 indicates, in two trips out of 120, the driver could follow their PRC's, 70% of the time (the bar midpoint). Likewise, if we look at top three bars in Figure 29, cumulatively, in 71% of trips (85 trips), the level of following PRC is at least 30%. The weighted average (based on trip duration) of following PRC is 36% which means the proposed PRC's are followed across the trips 36% of the time. Before we discuss the potential application this study in the conclusion section, the limitations are discussed next.

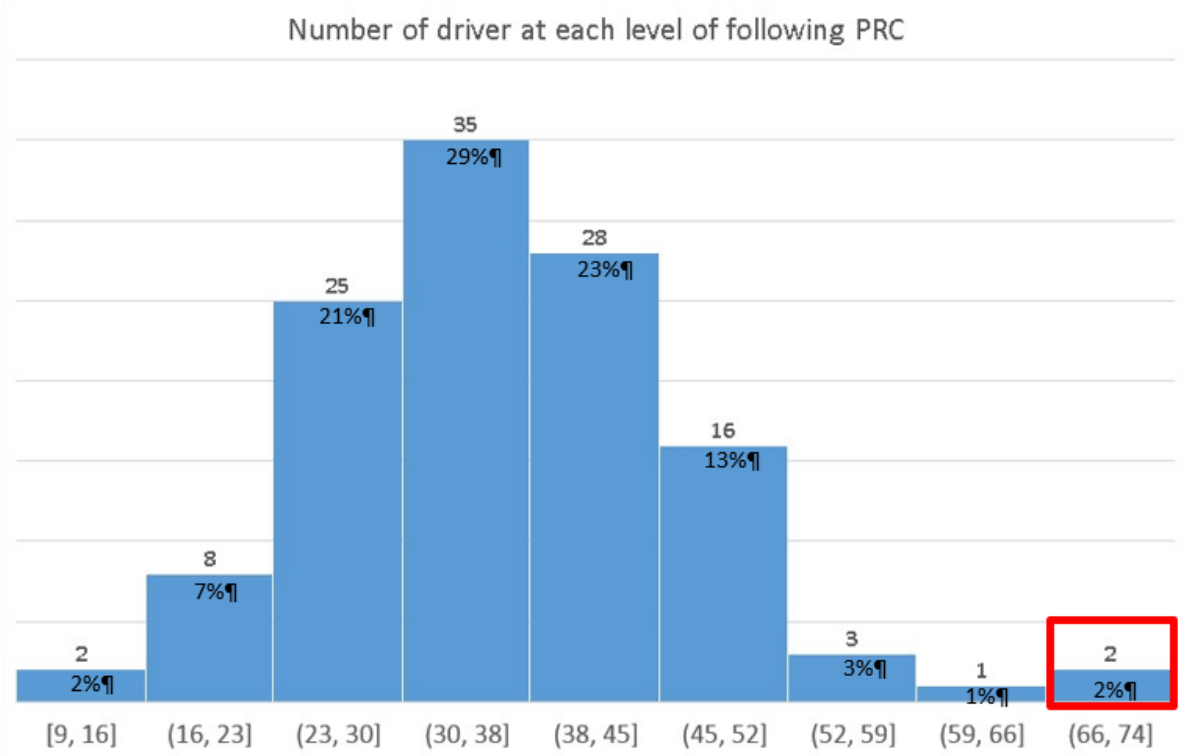


Figure 29. Percentage distribution of PRC among drivers (n=120)

Limitations

The significant limitation of this study lies in the data. Although the DAS dataset provides relatively accurate data for acceleration and speed, the amount of information on objects surrounding the host vehicle is limited. Only the number of objects and the distance to the front object are provided in the dataset. As it was noted in the methodology section, the authors had to deviate from the complete method of defining the states due to lack of information on objects position around the host vehicle. The data limitation also constrained the possibility of defining more states and consequently, more sophisticated reward functions. For the purpose of this study, to simplify the computation, the reward function was assumed dependent on the decision made and the landed state. Indeed, complex driving decisions are influenced by many factors including the attributes of the driver and characteristics of the context. More information about the biometrics of the driver and characteristics of the situation (vehicle kinematics and environmental variability) can paint a more complete picture of driving events. Clearly having more variables can help with more in-depth analysis of driving behavior and hence deeper understanding of driving decisions. Similar analysis with richer datasets containing the layout, speed and distances of surrounding vehicles as well as information about the weather, lighting, surface condition would probably yield better and more reliable results.

Also, note that the results are obtained by aggregating different driver behaviors and comparison of proportions at trip level. However, the MDP method provides more value when investigating individuals' behavior separately. Therefore, as more data from the same person becomes available, the proposed method enables deeper learning about the person's driving behavior.

Conclusion

This study contributes by using high velocity and diverse driving data to learn driving decisions. Specifically, the reward and states are defined theoretically and real-life driving decisions are analyzed to explore correlations with contextual factors i.e., proximity to surrounding objects. MDP is applied to learn personally revealed choices of drivers in terms of acceleration, deceleration and maintaining a constant speed. Eight states are defined based on the number of objects surrounding the host vehicle and distance to the front object. Individual drivers' reward functions are estimated using multinomial logistic regression. The results show that with increasing objects around a host vehicle, drivers would rather accelerate to avoid crowdedness around them.

Segmenting trips based on the level of crowdedness indicated that with increased level of crowdedness, fewer drivers choose acceleration as their PRC's because they are constrained to either keep constant speed or decelerate due to traffic condition. In addition, the level of following PRC results showed that the obtained PRC's match drivers' real behavior 36% of the time.

One potential application of this study is to generate short-term predictive information about driver decisions, which can be used to warn the driver when they deviate substantially from their own PRC's based on their own historical driving performance. The other is to use policies from expert drivers for training novice drivers or drivers who are more prone to near-misses and crashes. Yet another is to disseminate preferred actions of a driver to surrounding vehicles, enabling surrounding vehicles to foresee the states and actions of other drivers' potentially avoiding collisions. To develop applications, a considerable amount of trip data will be needed for each driver under different traffic conditions, which should not be a problem as IVs big data is emerging rapidly.

Acknowledgement

This paper is based upon work supported by the U.S. National Science Foundation under Grant No. 1538139. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CHAPTER V
HARNESSING INSTRUMENTED VEHICLES DATA FOR CRASH PREDICTION: A COMPARISON OF MACHINE LEARNING CLASSIFIERS

Abstract

Harnessing new sources of data provides new opportunities for improving transportation safety. While police-reported crash data provides useful information about crashes, new sources of data shed further light on some of the unobserved contributing factors such as pre-crash driver behaviors. Volatile driving behavior, characterized by hard accelerations and braking is hypothesized to be a key correlate of crash occurrence. This paper applies several statistical and machine learning classification methods to a number of recently developed “measures of driving volatility” in order to predict crash risks. Using a unique database, speed profiles from the Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS), the paper estimates models for real-time crash prediction purposes. Since different methods, depending on the data, can perform differently in terms of training and test misclassification rates, the following methods are compared to come up with most accurate crash risk prediction model: decision trees, bagging, boosting, random forest, logistic regression (also lasso and genetic algorithm logistic regression) and artificial neural networks. The results show that the neural networks and random forest outperform other methods, in terms of training and test errors of 4.47% and 5.19% respectively. The paper also provides insights about what measures of driving volatility are correlated with higher risks of crashes. As such, among the developed measures of driving volatilities, four speed volatilities found to be correlated with higher crash risks. The importance of these results lies in the fact that the features used to classify normal (baseline) and crash events are solely extracted from speed profiles. It means such features can be quantified constantly in real-time through the use of sliding windows as inputs in the models for crash prediction. Proper warnings can be issued to the driver for crash avoidance purposes.

Introduction

Crashes are wicked problems that burdens modern society. Emergence of new data sources has created new opportunities to explore factors contributing to vehicle crashes that were previously not possible to study. Although the traditional crash reports by police officers provide clues about the probable causes, they cannot provide concrete information on detailed instantaneous pre-crash behaviors. Instantaneous driving behavior information provided by new sources can be utilized for real-time prediction of crashes, potentially preventing them. In this context, however, the accuracy of prediction is vital. Therefore, a model that can predict crash risk accurately is highly desirable. Among supervised learning methods, statistical models are beneficial for providing insights about the magnitude and direction of the independent variables on the response variable. Specifically, when it comes to predicting crashes, these models can help practitioners take necessary countermeasures that reduce crashes. However, predicting crashes in real-time for the purpose of generating alerts to the driver, demands models

with high level of accuracy where inference from the model is not a high priority. As a result, other supervised learning methods might yield more accurate results. Another challenge of real-time crash prediction is that traditional variables used for crash prediction might not be useful. For example, in a reactive manner, historical data (e.g., average daily traffic, segment length, number of the lanes, average speed, demographics etc.) are used to build the crash prediction models. Such information is not helpful for real-time incident prediction because they are not driving context-related variables, i.e., they do not have enough variations during the course of a trip (e.g., drivers' age and gender do not change during a trip). In addition, given the contribution of drivers' instantaneous decisions and driving styles to crashes, there exists a need for more personalized measures that can be collected, processed and used in a short time for crash prediction.

Although numerous studies have investigated driving behaviors and their effects on crashes, the emergence of detailed data on vehicle dynamics has opened up new windows for studying driving behaviors from a different perspective. Extracting useful information from such data is a lightly explored area of research. For instance, the measures of driving volatility that were developed and discussed in our previous works (*1; 20*) could not be quantified effectively if instantaneous and high resolution data on vehicles speed and acceleration were not available. Driving volatility tries to capture and quantify driving variations in order to provide timely alerts to drivers if their driving volatility surpasses some pre-specified threshold. It should be noted that previous studies have found that driving volatility (the higher the driving volatility, the higher driving aggressiveness) is correlated with a higher number of crashes (*1*). Therefore, the objective of this study is to harness large-scale data, using machine learning models to accurately predict crash risk. The main advantage of such models is that they can be used in real-time crash prediction because their inputs are solely different measures of driving volatility, which can be calculated in real-time from vehicle kinematics. In this context, prediction accuracy is of higher priority than model interpretability. Therefore, several statistical and machine learning classifiers will be attempted to find the best crash risk predicting model in terms of training and test misclassification errors.

In the literature, the term "real-time crash prediction" have been mostly used by researchers for crash prediction over some segments using 5 to 20 minutes traffic data. Hossain and Muromachi (*130*) built a Bayesian belief network able to predict dangerous traffic conditions on a specific segment for the next 4 to 9 minutes with 66% accuracy and 20% false positive. By real-time, as noted, the authors meant next 4 to 9 minutes. In another study (*131*), dynamic Bayesian network was used to predict crashes from the data collected using loop detector on the traffic conditions. The model reached to 76.4% accuracy with 23.7% false alarm rate. The data used in the study were collected within 5 to 20 minutes before the crashes. Wang et al. (*132*), presented a Bayesian logistic

regression for real-time crash prediction of expressway weaving segments using 5 to 10 minutes data from traffic detectors. The authors noted that along with traffic conditions, wet surface increases the crash odds ratio by 77%. As previous studies were focused on the crash risk, Xu et al. (*133*) included crash severity levels in their study for real-time crash severity and likelihood prediction on freeways using the data from loop detectors. Similar to previous studies, a 5-minute time window was used for the prediction. Other studies, although similar in terms of source of the data, have used other methods such as support vector machine (*134*) , matched case–control logistic regression (*135; 136*), genetic programing (*137*) for crash risk evaluation. As noted, all mentioned studies were focused on crash-prediction over a segment and using the timed window data ranging from 5 to 20 minutes. In this study however, the focus is on the real-time crash prediction for individual drivers using their roughly 20 to 22-second speed profiles.

In summary the paper contributes by unveiling potential of the information residing in large-scale instrumented vehicle data to build models able to predict the vehicle crash risks with high level of accuracy. The study is original in terms of estimating models whose inputs are solely derived from vehicle kinematics making them applicable to real-time crash risk prediction for the purpose of crash avoidance and warning assists. Moreover, the models can be used for real-time crash prediction as the stream of data from vehicle kinematic are processed.

Methodology

In previous studies, we have discussed measures of driving volatility and applied them to intersection crash frequency data (*20*). For review, they are discussed briefly below.

Measures of Driving Volatility

Standard Deviation

Researchers commonly use standard deviation (S_{dev}) to capture variation in data. It is calculated as follows (*58*):

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

where x_i denotes ith observation, $\bar{x}$ is the mean and *n* is the sample size.

Mean Absolute Deviation

This measure averages the deviation of each observation from a central tendency (here mean) of the data (*49*):

$$D_{mean} = \frac{1}{n}\sum_{i}^{n}|\bar{x} - x_i| \quad (2)$$

where $\bar{x}$ is the mean, x_i is i[th] observation and *n* is sample size.

Coefficient of Variation

This measure is the ratio of standard deviation (S_{dev}) to the absolute value of mean ($|\bar{x}|$) (*138*), provides a relative measure of dispersion:

$$C_v = \frac{S_{dev}}{|\bar{x}|} * 100 \quad (3)$$

Quartile Coefficient of Variation

This measure is especially useful when the sample is not normally distributed (*139*). It is defined as (*50*):

$$Q_{CV} = \frac{Q_3 - Q_1}{Q_3 + Q_1} * 100 \quad (4)$$

where Q_1 and Q_3 denote the sample's first and third quartiles respectively.

Time-varying Stochastic Volatility

The time-varying stochastic volatility (*52*) quantifies the dispersion in a positive value time-series data and is defined as:

$$V_f = \sqrt{\frac{1}{n-1}\sum_{i}^{n}(\mathrm{r_i} - \bar{r})^2} \quad i = 1 \dots n \quad (5)$$

and

$$\mathrm{r_i} = \ln\left(\frac{x_i}{x_{i-1}}\right) * 100 \quad (6)$$

where *ln* is the natural logarithm and x_i is the ith observation (in this study speed).

Count of Extreme Values

The ratio of extreme valued observations to the total number of observation captures driving volatilities in the case of extreme driving behavior (*20*):

$$\%T = \frac{c > Threshold}{n} * 100 \quad (7)$$

where *c* is the number of observations beyond the *thresholds* and *n* is the sample size. The upper and lower thresholds are defined as:

$$Threshold = \bar{x} \pm z * S_{dev} \quad (8)$$

where $\bar{x}$ and S_{dev} are the mean and standard deviation respectively; *z* determines how wide or narrow the thresholds should be in terms of standard deviation. When applying this measure to acceleration and jerk, the sample is portioned according to the speed values in order to account for reduction in vehicles' acceleration ability at higher speed (*1; 20*). Then for each speed category, separate upper and lower limits are established to obtain number of extreme observations. Please see (*20*) for more details.

The complexity of instantaneous pre-crash behavior and driving volatility demands application of several measures so that if some aspects of volatile driving behaviors are not capture by one measure, they can be covered and quantified by other measures.

Types of Driving Volatility

Since the data consists of both baseline and crash driving profiles, it is important to include only normal driving volatility (intentional volatility attributable to the driving style) in the analysis and exclude the volatilities during evasive maneuver to avoid a crash (unintentional volatility). For illustration, Figure 30 shows two driving profiles of a baseline and a crash from NDS data. For baseline profiles (e.g., Figure 30, left), all 20-second speed and acceleration values are used to obtain the driving volatilities. However, for crash profiles, only seconds up to the driver's reaction moment are used to calculate driving volatilities. For example, in Figure 30 (right), 21 seconds of the data before the crash are used. In 17% of the cases (combined crash and near-crash profiles) drivers did not react or reacted after the impact. For these cases seconds up to the moment of incident are used for the calculation of driving volatilities. In summary, either the incident moment or the driver reaction moment whichever comes first, was used to obtain the number of seconds from speed and acceleration profiles for volatility calculations.

It is important to note that the exclusion of evasive maneuver duration from the calculation of driving volatilities is appropriate because the volatilities are far enough removed from the event that pre-crash volatility can be appropriately used for real-time crash prediction. Notably, it will be too late to warn a driver who has already started taking evasive actions. Hence, we have excluded the duration of evasive maneuver from the explanatory variables i.e. volatilities.

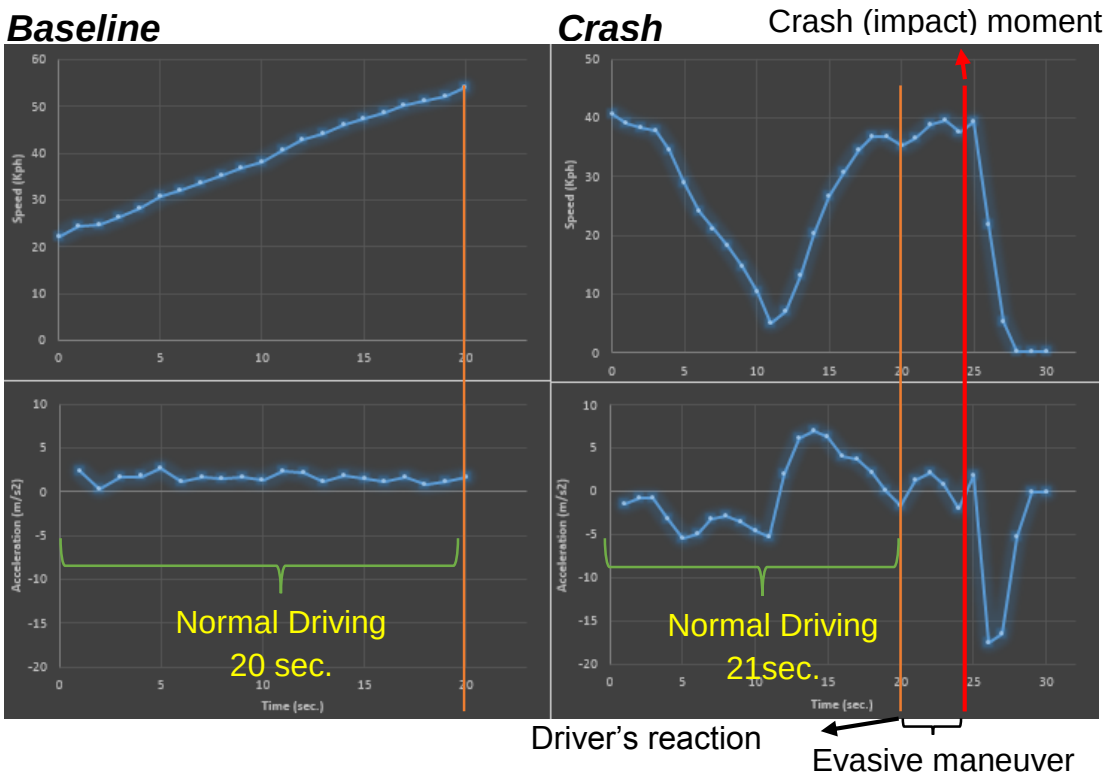


Figure 30. Comparison of baseline with crash/near-crash driving profiles

Classification Methods

Given the binary nature of the response variable in the dataset (i.e. crash or non-crash), several classification methods can be applied. This study examines and compares the prediction capability of the following methods whose approaches for classifications differ:

Decision Trees

In the decision tree approach (*140*), the training set is divided into *m* regions using binary splits based on the independent variables. Typically, researchers use cross entropy or the Gini index to quantify the quality of the splits and to avoid over-fitting. The Gini index measures the total variance of a K-class problem through:

$$G = \sum_{k=1}^{K} \hat{P}_{mk}(1 - \hat{P}_{mk}) \tag{9}$$

where K is the number of classes, m is the number of regions and $\hat{P}_{mk}$ is the proportion of observations in the mth region from the kth class. When the majority of observations at a node are from the same class, the Gini index for that node nears zero. Therefore, the decision tree algorithm tries to minimize Eq. 9 and stops once the difference between two consecutive attempted trees reaches a predetermined threshold. Decision trees can be

very sensitive to changes in data and show higher levels of variance when they are introduced to a testing dataset. In decision trees, the variables at the top are more significant than lower ones.

Bagging

To reduce the variance discussed in the decision tree method, the bagging method (*141*) samples *n* times from the training set and makes a separate tree for each sample and reports the mean (over the built trees) in the final model. In this approach, if each of the trees has variance σ^2, the mean of the final tree will have the reduced variance i.e. σ^2/n. In practice, because the training data is limited, samples are drawn repeatedly from the training dataset (bagging). Therefore, one observation might be used in several samples. For the prediction, the model uses the majority vote given by the trees as the final prediction for each observation. Although this method reduces variance, compared to a decision tree the averaging step will compromise the interpretation capability of variable importance.

Random Forest

The random forest (*142*) mechanism is similar to bagging except, instead of using all of the independent variables to make the trees, the method randomly selects a subset of them. As a rule of thumb, researchers use the rounded square root of the number of variables at each node (binary split). When the data is high dimensional, even a smaller number of predictors than square root of predictors is recommended. Selecting a subset of predictors results in the creation of trees that are less correlated with each other when compared to the highly correlated trees built in bagging. Since averaging highly correlated values in bagging will not lead to a reduced variance significantly, the random forest method should theoretically return a lower variance than bagging by trying more decorrelated trees.

Boosting

Similar to bagging, boosting (*143*) works by sampling from the training sets to create decision trees. The difference between bagging and boosting is that in bagging the trees are built in parallel (even though they are correlated due to samples overlapping) while in boosting a tree is built upon the previous tree in a sequential manner without the bootstrapping. In other words, the decision trees grown one by one on the original training dataset so that each tree is a modification of the previous tree by working on the residuals. This procedure makes the boosting method a slow learner with the tendency of over-fitting the data as opposed to the other tree-based approaches discussed earlier. Therefore, the number of trees, the regularization term and the order of interactions need to be tuned which is done commonly through cross-validation. The pseudo code for the algorithm is as follow (*140*):

1. Initialize: $f(x) = 0$ and $r_i = y_i$ (*i* is the size of training dataset)
2. For j from 1 to number of trees:
    a. fit tree $f^j(x)$ to training dataset (X, r)
    b. update the model: $f(x) = f(x) + \lambda f^j(x)$
    c. update the residual vector: $r = r - \lambda f^j(x)$
3. Return the final model: $f(x)$

Logistic Regression

Among the machine learning tools applied in this paper, logistic regression (*144*) is probably the most frequent method used in the literature. The *logit* is defined as the odds of the dependent variable being 1 through:

$$logit(P) = \ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_i X_i \qquad (10)$$

where β_0 is the intercept and β_i is the parameter estimate for the predictor X_i. The probability of individual observations in both training and test datasets are calculated using the estimated parameters. The observations with probability of more than 0.5 are classified as class 1 while all others are 0. Finally, the misclassification error rate is reported as the ratio of the number of misclassified observations to the total number of observations.

Lasso Logistic Regression

Lasso regression (*145*) is one of the shrinkage methods with the capability of subset selection of the independent variables through setting some of the independent variable estimates to zero depending on value of the regularization parameter. In general, if the regression model estimates provide low bias and high variance, the prediction accuracy might be improved by applying the lasso method to it. In this way, some of the model bias is compromised by having a smaller number of predictors in order to reduce the variance (*145*). Another advantage is that having a parsimonious model is more beneficial in terms of interpretation of the contributing factors to the response variable. In lasso, the objective is to minimize:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \qquad s.t. \quad |\beta_j| \le t \qquad (11)$$

where y_i is the dependent variable, β_0 is the intercept, β_j is the parameter estimate for the predictor j and t is the amount of budget (or penalty) we would like to spend in the estimation. Choosing a large t reduces the problem to the logistic regression while the smaller values of t will lead to models that are more parsimonious.

Genetic Algorithm Logistic Regression

Although the initial goal of the Genetic algorithm (GA) by *(146)* was to study and transfer the phenomenon of adaptation in nature to computer systems (*147*), researchers in different disciplines have used the algorithm to find sub-optimal solutions for optimization problems. Concisely, a typical GA has the following elements: population of chromosomes, operators of selection, crossover and mutation, and a fitness function. *Chromosomes* can considered as a string of zero and ones. A *selection* operator selects chromosomes based on the fitness function so that the chromosomes with better fit function values are more likely to appear in the selection. The *crossover* randomly exchanges the sequences after a given bit in chromosomes to create offspring. For example, if we have two chromosomes: **11011111** and **10000100** and want to crossover them according to their fourth bits (shown in red), the results would be **11010100** and **10001111**. A *mutation* that commonly happens with small probability is when a randomly selected bit in a given chromosomes flips. For instance, if a mutation is to happen to **11011111** on the fourth bit, the result is **11001111**. The *fitness function* quantifies the quality of created chromosomes. The steps of a simple GA are as follows (*147*):

1. Generate *n* chromosomes (population) with length *l*
2. Obtain the quality of chromosomes in the population
3. Repeat until *n* offspring are generated:
    a. ***select*** a pair of chromosomes (parents) with replacement from the population
    b. ***crossover*** the parents at the random bit to generate two offspring
    c. ***mutate*** the two offspring and place them in the new population pool
4. Replace the old population with the new generated population
5. Go to step 2

Every time we go from step 2 to 5, a new population is created containing *n* chromosomes which is called *generation.* Researchers generally obtain several generations (e.g., 100) and pick the most frequent chromosome that appears. The GA structure makes it

applicable to model selection in statistical models (*148; 149*) where each bit in the chromosomes represents one of the predictors. Therefore, the length of chromosomes is equal to the number of predictors plus the intercept. If a variable is included in the model, the relative bit is 1 while all others are 0. The fitness function can be goodness of fit score such as AIC (*150*), CAIC (*151*), ICOMP (*152-157*) and ICOMP IFIM (*148; 149*).

Akaike information criteria (AIC) is obtained by:

$$AIC = -2LL + 2K \tag{12}$$

where *LL* is the maximum log likelihood and *K* is the number of estimated parameters in model.

Consistent AIC (CAIC) can is calculated as following:

$$CAIC = -2LL + K(\log(n) + 1) \tag{13}$$

where *n* is the sample size from which the model is estimated.
The entropic statistical complexity criterion, ICOMP is defined by:

$$ICOMP = -2LL + 2C\left(\hat{\Sigma}_{Model}\right) \tag{14}$$

where $\hat{\Sigma}_{Model} = \widehat{Cov}(\hat{\beta}_K)$ denotes the estimated covariance matrix of the parameter vector of the model and C represents a real-value complexity measure given by:

$$C\left(\widehat{Cov}(\hat{\beta}_K)\right) = \frac{K}{2}\log\left(\frac{tr\left(\widehat{Cov}(\hat{\beta}_K)\right)}{K}\right) - \frac{1}{2}\log\left|\widehat{Cov}(\hat{\beta}_K)\right| \tag{15}$$

The entropic statistical complexity criterion, ICOMP using inverse information matrix is defined by:

$$ICOMP_{IFIM} = -2LL + 2C_1\left(\hat{F}^{-1}\right) \tag{16}$$

where C_1 is the maximal information complexity of the estimated inverse Fisher information matrix $\left(\hat{F}^{-1}\right)$ of the model defined by:

$$C_1\left(\hat{F}^{-1}\right) = \frac{K}{2}\log\left(\frac{tr\left(\hat{F}^{-1}\right)}{K}\right) - \frac{1}{2}\log\left|\hat{F}^{-1}\right| \tag{17}$$

Neural Networks

Although the initial goal of the neural networks (*158*) was to mimic the problem solving ability of the human brain, it eventually evolved into solving a variety of problems such as speak recognition, translation, computer vision and image processing, and classification. Backpropagation algorithm (*159*) is used in neural networks to distribute errors from outputs all the way *back* to the input through layers. Advancements in computational powers have led to bigger neural networks with many hidden layers and the advent of the term “deep learning.” In addition, the emergence of big data in many disciplines has made neural networks an attractive tool for researchers. One reason is that as the amount of training examples increase, the neural networks usually does a better job in terms of prediction as compared to traditional classification methods, even though its mechanisms are similar (a neural networks with a sigmoid activation function without hidden layers is the same as a logistic regression). However, increased prediction accuracy in neural networks comes at the expense of losing the ability for interpretation of the relationships between predictors and response variable. Therefore, in cases where the accuracy of the prediction outweighs the interpretation of the model, one might prefer neural networks.

Data

In this paper, the Naturalistic Driving Study (NDS) data (*160; 161*) from the Strategic Highway Research Program 2 (SHRP2) is used to obtain individual driving volatilities. The NDS data includes time-series records from eight different sensors installed on vehicles as well as videos recorded by installed cameras. The study obtained the data from 3100 volunteer drivers in six selected cities in the United States. The data were error-checked and erroneous cases were excluded from the analysis. Out of the 9,153 driving speed profiles, 1865 were coded as crash (Figure 31, left). For each profile, vehicular acceleration and jerk data were calculated from the vehicle speed value. Application of the measures of driving volatility to the speed, acceleration and jerk produced 23 variables (Figure 31, right). In the data, baselines are 20 second and crash and near-crash case are 30 second long. Therefore, to exclude the driving volatility caused by drivers to avoid crashes, the seconds up to either the drivers’ reaction moment or crash/near-crash moment whichever came first in the speed profiles were used. The moment of drivers’ reaction and crash/near-crash are available in the data. Figure 32 shows the distribution of seconds used for calculation of driving volatilities. According to the figure, for each case we used different pre-crash durations to calculate driving volatilities.

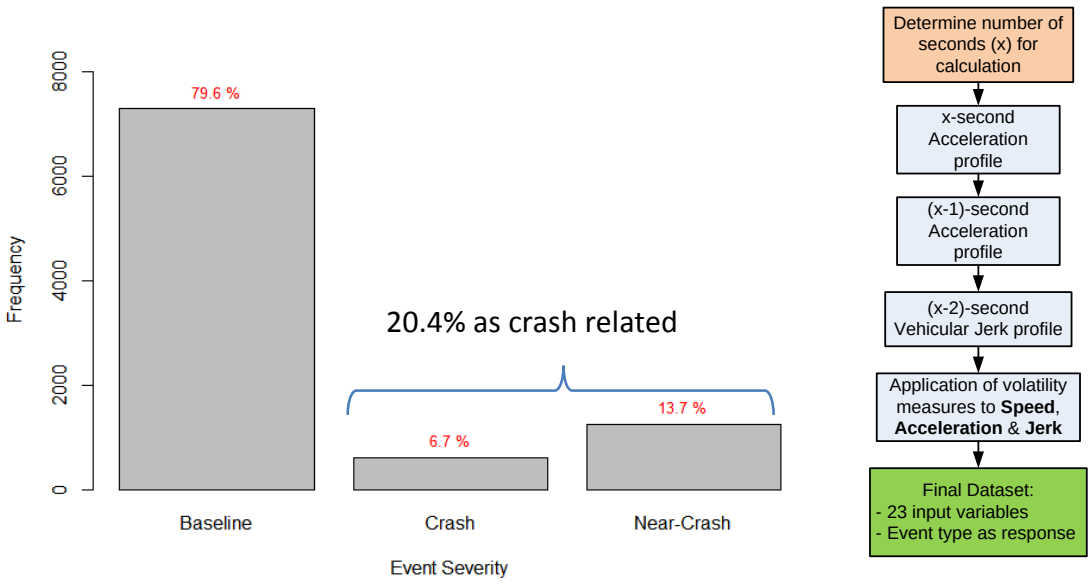


Figure 31. Left: classification of events in the data; Right: Data preparation steps

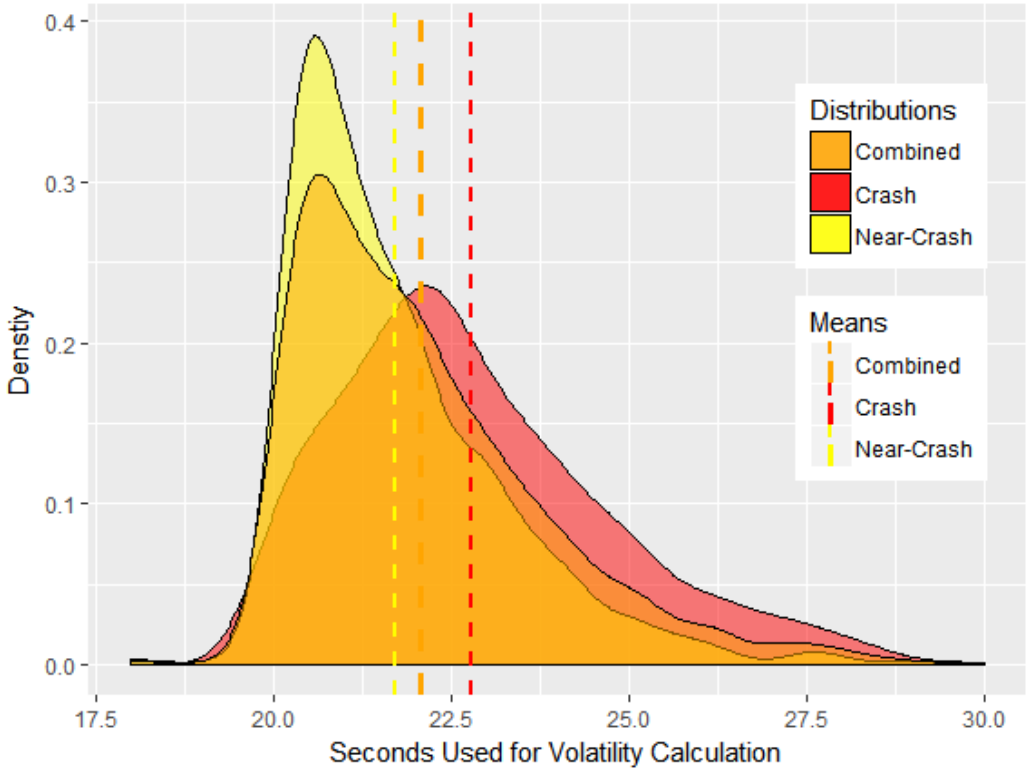


Figure 32. Distribution of second used for volatilities calculation

Results

Descriptive Statistics

Table 11 provides descriptive statistics for each variable. The initial descriptive statistics resulted in the identification of some erroneous trips where the speed values were not reasonable. These trips were excluded from the analysis.

Misclassification Rate

We used 8000 (87%) of the observations for training the predictive models training and the other 1153 (13%) observations for testing their out-of-sample prediction capability. Fifty iterations were tried. In each iteration, the training observations were randomly sampled from the dataset and the remaining observations kept for testing the models performance (i.e. out-of-sample prediction). During the implementation of the methods, when an assignment of a parameter was necessary (e.g., lasso regularization parameter), the value was obtained using cross-validation in the relative iteration. The mean and 95% normal confidence intervals are also provided in Table 12. Figure 33 and Figure 34 show the performance of each method in terms of overall and recall misclassification rates for training and out-of-sample testing dataset with 50 runs. The centerline in each box in the figures represent the median of the misclassification rates. The figures and table show that neural networks yields the lowest training and random forest gives the lowest test errors.

The logistic-based methods (i.e. logistic, lasso logistic and GA logistic), lasso logistic have similar performance. Among the tree-based methods (i.e. decision tree, boosting, bagging and random forest), random forest was the best. The boxplots provided in Figure 33 and Figure 34 are also useful for showing the over-fitting issue. Methods whose median test error are more than their median training error have likely over-fitted the data. Thus, decision tree, boosting, logistic GA and neural networks suffer from over-fitting. The means of training and test errors provided in Table 12 also lead to the same conclusion. A higher value of test error minus training error means a higher level of over-fitting. Therefore, the single-layer neural networks has over-fitted the data the most (which of course is fixable to some extent). Misclassification rate, especially for real-time prediction of crashes is an important metric. Large misclassification rate of non-crash events will generate false alarm to the driver making the driver insensitive to the warning or disable the system. On the other hand, large misclassification of crash events affect the reliability of the system adversely and could lead to severe safety outcomes in terms of crash frequency and injury severity.

Table 11. Measures of driving volatility descriptive statistics (n = 9372 trips)

Variable	Mean	S_{dev}	Min	Max
Speed-S_{dev} (m/s)	2.12	1.94	0	33.52
Speed-V_f (%)	0.18	0.34	0	4.81
Speed-C_v (%)	0.23	0.30	0	3.05
Speed-Q_{cv} (%)	0.18	0.23	0	1.00
Speed-D_{mean} (m/s)	1.74	1.63	0	26.25
Speed-$\%T(1S_{dev})$ (%)	0.32	0.11	0	0.71
Speed-$\%T(2S_{dev})$ (%)	0.03	0.04	0	0.50
AccDec-S_{dev} (m/s^2)	0.65	0.54	0	22.48
Accel-C_v (%)	0.76	0.26	0	3.14
Decel-C_v (%)	0.84	0.32	0	2.88
Accel-Q_{cv} (%)	0.59	0.18	0	1.00
Decel-Q_{cv} (%)	0.62	0.18	0	1.00
AccDec-D_{mean} (m/s^2)	0.27	0.32	0	10.77
AccDec-$\%T(1S_{dev})$ (%)	0.35	0.15	0	0.95
AccDec-$\%T(2S_{dev})$ (%)	0.11	0.12	0	0.94
Jerk-S_{dev} (m/s^3)	0.66	0.61	0	32.94
JerkPos-C_v (%)	0.90	0.33	0	2.94
JerkNeg-C_v (%)	0.88	0.34	0	3.00
JerkPos-Q_{cv} (%)	0.63	0.17	0	1.00
JerkNeg-Q_{cv} (%)	0.61	0.17	0	1.00
Jerk-D_{mean} (m/s^3)	0.34	0.38	0	19.61
Jerk-$\%T(1S_{dev})$ (%)	0.31	0.12	0	0.83
Jerk-$\%T(2S_{dev})$ (%)	0.08	0.09	0	0.60
Y (dependent variable)	1.20	0.40	1	2.00

Table 12. Normal confidence interval of the error mean for different methods (n training = 8,000 & n testing = 1,153)

Method	Overall Error (%)						Recall (crash) Error (%)					
	Training			Test			Training			Test		
	Mean	LL	UL	Mean	LL	UL	Mean	LL	UL	Mean	LL	UL
Decision Tree	7.74	7.68	7.80	8.22	8.02	8.42	22.36	21.95	22.76	23.43	22.60	24.25
Logistic	6.72	6.68	6.75	6.83	6.64	7.02	22.63	22.51	22.74	22.67	21.92	23.43
Logistic GA*	6.82	6.78	6.86	7.05	6.87	7.23	23.18	22.99	23.36	23.74	23.02	24.46
Logistic Lasso	6.76	6.72	6.79	6.88	6.69	7.07	22.87	22.75	22.99	23.07	22.31	23.83
Bagging	5.35	5.32	5.39	5.26	5.10	5.41	17.64	17.50	17.77	17.46	16.74	18.17
Random Forest	5.30	5.27	5.33	**5.19**	5.04	5.35	17.58	17.45	17.70	**17.18**	16.47	17.89
Boosting	6.19	6.16	6.22	6.51	6.32	6.70	21.47	21.37	21.58	22.12	21.40	22.85
Neural Networks**	**4.74**	4.68	4.80	5.84	5.64	6.05	**15.50**	15.27	15.73	18.01	17.21	18.82

** Results with ICOMP criteria; LL: Lower Limit; UL: Upper Limit*

*** Single layer neural networks with 5 neurons in the hidden layer and sigmoid activation function*

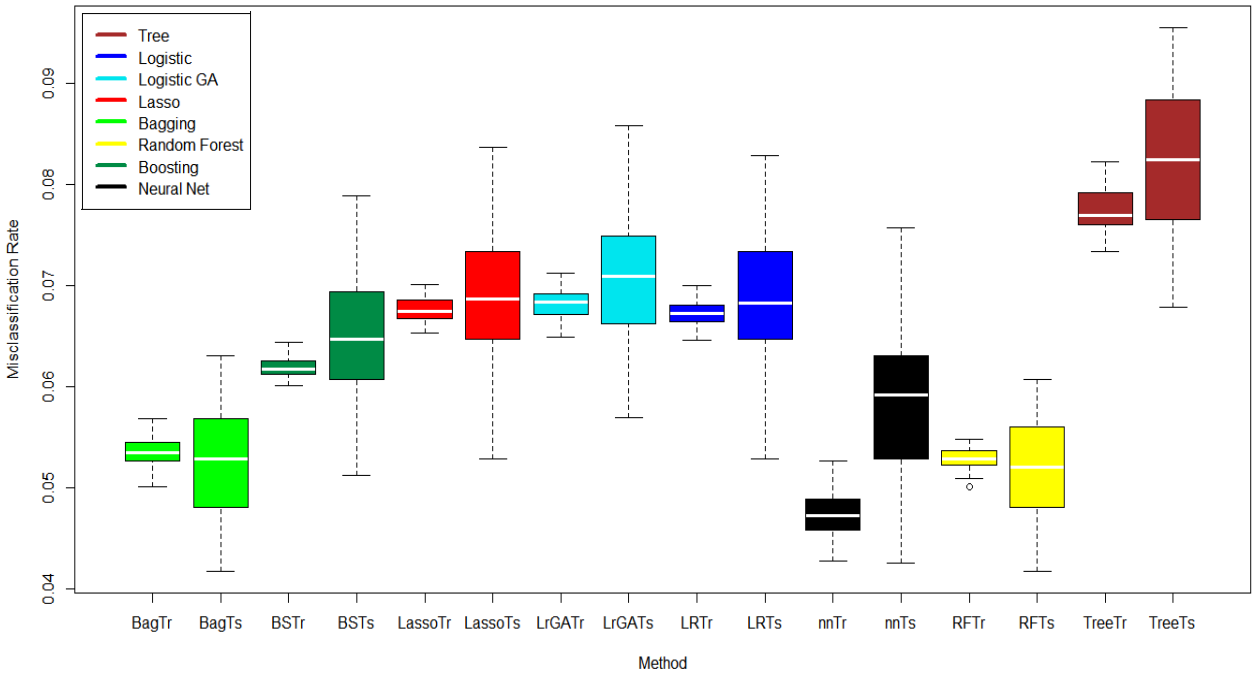


Figure 33. Misclassification rate of different methods (Tr: train, N = 8,000, Ts: test, N = 1,153)

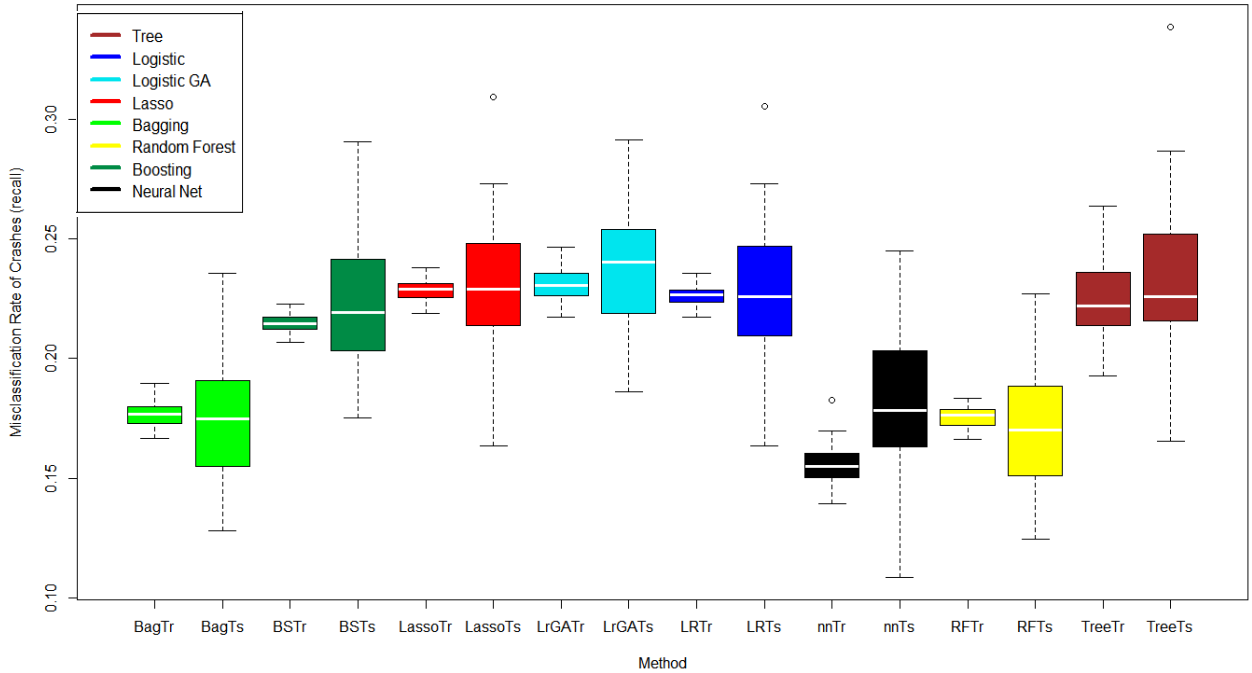


Figure 34. Misclassification rate of crashes (recall) for different methods (Tr: train, N = 8,000, Ts: test, N = 1,153)

True Positive and False Positive

Receiver Operating Curves (ROC) shown in Figure 35 are commonly used to investigate or compare the detection ability of binary classifiers in terms of true positive (sensitivity) and one minus false positive (specificity). The accuracy of a classifier is measured by the calculation of the Area Under Curve (AUC). If a classifier predicts the classes as accurate as a random guess, its accuracy or AUC (the probability of detection) will be around 0.5. The ROC of such classifier is shown as a 45 degree grey line in Figure 35. According to the figure, bagging and random forest method outperforms the other methods. However, it should be note the differences between the methods are small excluding decision tree method whose AUC is smaller than other classifiers.

Ga Logistic

The saturated logistic model has 23 variables. However, a better predictive model does not have to use all the variables for prediction. Therefore, the best model should be selected from 2^{23} (8,388,608) possible models. GA moves forward to better models heuristically and stops based on the threshold criteria rather than trying all possible models. Thus, it can reach suboptimal answers in a shorter period of time instead of searching for the true optimal model, which can be time consuming.

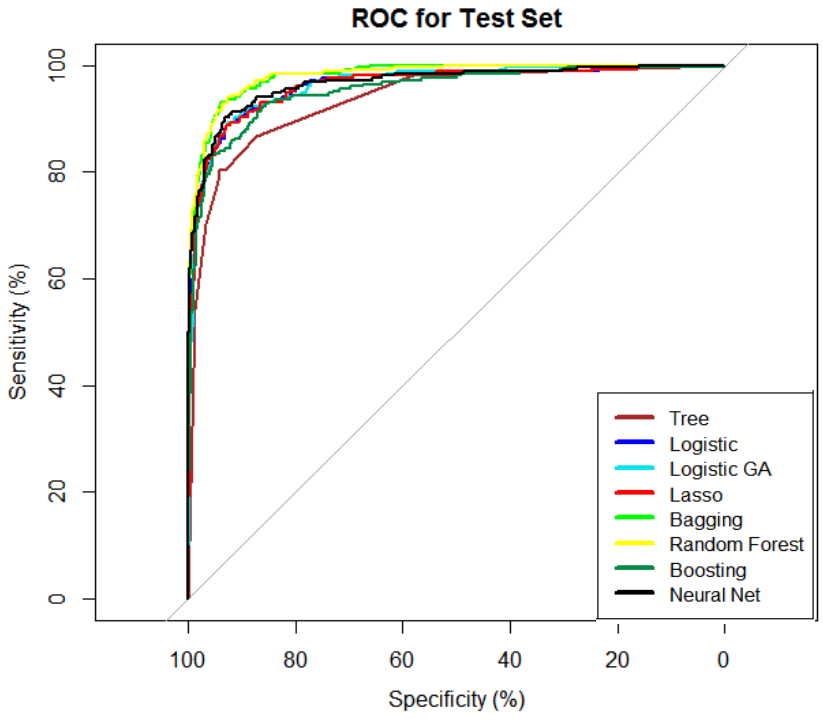


Figure 35. Receiver Operating Curves for test dataset

Table 12 presented the results of the logistic regression for the saturated model (i.e. 23 variables and intercept all included in the model), but the GA logistic might suggest models with a subset of independent variables. Table 13 shows the GA algorithm settings. Figure 36 provides results for 50 iterations of logistic regression using the variables suggested by GA (ran 100 times) for different fitness criteria: AIC, CAIC, ICOMP and ICOMP IFIM. ICOM and ICOM IFIM were the same and yielded lower amount of error than AIC and CAIC. Therefore, we used ICOMP suggested model results to compare the performance of logistic GA with other classifiers (shown in Figure 33 and Figure 34).

Contributing Measures of Driving Volatility to Crashes

The classification procedures presented in the previous sections included all measures of driving volatility for predicting crash risks. However, the best crash-predicting model (*best statistical model*) is not necessarily the best intuitive model (*best engineering model*). That said, if we are looking for the best engineering model to identify crash-contributing factors in terms of driving volatility, we should include those variables in the model which are positively associated with crashes. Generally, a stepwise approach where variables are inserted in the model in each step and their effects are investigated, is applied. Table 14 shows the results of the estimated fixed and random parameter logistic models (to account for unobserved heterogeneity). The random parameter model is estimated using simulated maximum likelihood method with 500 Halton draws. Among the 23 measures of driving volatility, eight measures contributed to higher crash risk. The marginal effects provided in Table 14 presents the amount of increase in crash probability with one unit increase in the respective independent variable. For example, according to the random parameter model, one unit increase in Speed-V_f is correlated with a 48.5% increase in crash risk. The results of random parameter model indicate that the effects of driving volatility are different among the trips. For example, the fixed parameter model suggests a coefficient of 0.642 for Speed-V_f, while the mean random parameter model estimates a coefficient with mean 4.018 and standard deviation of 4.557. Overall, the random parameter model is a better fit when compared to the fixed parameter model.

Table 13. Genetic Algorithm settings for logistic regression subset selection

GA Parameter	Value
Number of Generation	100
Population Size	40
Crossover Probability	0.8
Mutation Probability	0.1
Elitism	Yes

Note: Fitness Criteria used include ICOMP (IFIM), AIC, CAIC, ICOMP

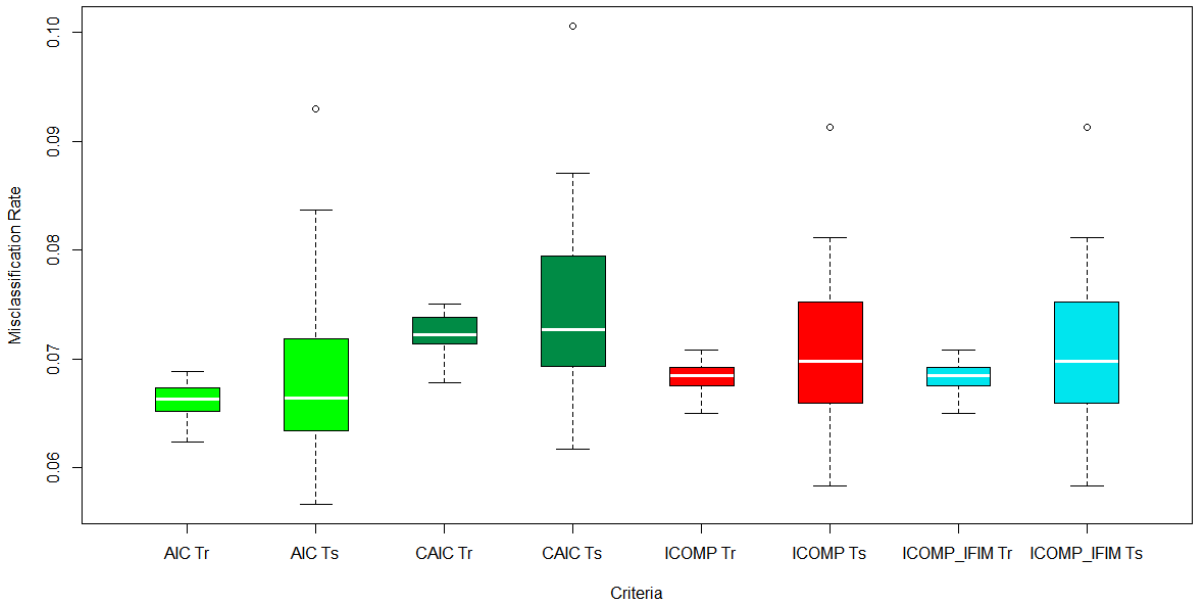


Figure 36. Misclassification error for different fitness criteria (tr: train, n = 8,000, ts: test, n = 1,153)

Table 14. Fixed and random parameter logistic model for crash prediction

Variables	Fixed Parameter			Random Parameter		
	β^a	z	M.E.	β^a	z	M.E.
Constant	-8.888***	-38.05	--	-8.077***	-35.26	--
Speed-S_{dev}	0.319***	10.62	0.020	.22748***	7.6	0.027
Speed-V_f	0.642***	4.16	0.041	4.018***	16.25	0.485
Std. dev.	--	--	--	4.557***	19.8	--
Speed-Q_{cv}	4.714***	15.9	0.302	2.553***	8.98	0.308
Speed-%T($2S_{dev}$)	6.035***	5.01	0.387	2.657**	2.45	0.321
Acceleration -S_{dev}	0.430***	3.63	0.028	0.602***	6	0.073
Std. dev.	--	--	--	1.230***	20.82	--
Acceleration-Q_{cv}	0.559**	2.51	0.036	0.613***	2.92	0.074
Deceleration-C_v	3.460***	20.82	0.222	3.236***	21.04	0.391
Jerk-Negative-C_v	1.342***	8.86	0.086	1.222***	9.02	0.148
Summary Statistics						
AIC	4118.6			3699.8		
Log-likelihood at Zero *L(0)*	-4657.6			-4657.6		
Log-likelihood at *L(6)*	-2050.3			-1838.9		
McFadden ρ^2	0.559			0.605		
Sample Size (N)	9153			9153		

[a] *Significance codes: *** 0.01%, **1%, * 5%, · 10%*

[b] *V_f: time-varying stochastic volatility; D_{mean}: mean absolute deviation around mean; Q_{cv}: quartile coefficient of variation*; *$\%T(2S_{dev})$: % of extreme points beyond mean ± two standard deviation*; *C_v: coefficient of variation;*

Limitations

The results from the previous sections clearly indicate the criticality of information from the speed profiles. Vehicle kinematics are the physical realization of driver decisions. These decisions are affected by many factors. Thus, vehicle kinematics alone are not enough and it should be complimented with other data such as distance to the front object, relative speed of the vehicle, driver biometric information as well as information from roadway infrastructure (*51*).

Another limitation is the difference between the proportion of crash events to non-crash event in reality and data. In real world, crashes are rare events while in the data in hand, 21% of the observations are crash or crash related events. The difference might have affected the results in this paper. In particular, if customized driving alerts for individual drivers are preferred, there might not be enough instances of crash and crash related events to train models with high accuracy. Although the discussed model can serve as a base artificial intelligent for driving assist systems, there is still a need for future training of models for certain and rare events. Finally, the NDS data might not be representative of driver population because they participated the program voluntarily with monetary incentives.

Conclusion

This paper used data from the Naturalistic Driving Study (NDS) to estimate several crash risk models. Several statistical and machine learning methods were used including decision trees, bagging, boosting, random forest, logistic regression, lasso logistic regression, genetic algorithm logistic regression and artificial neural networks. The predictions of these models were assessed in terms of misclassification rate. Twenty three measures of driving volatility were applied to vehicles speed, acceleration and jerk and used as predictors in the models. In each of the fifty iterations, the data were randomly portioned to training and test datasets and the mean of their misclassification rates were reported. The results show that neural networks and random forest outperformed the other methods with overall (crash and non-crash) training and test errors of 4.74% and 5.19%, respectively which means that these methods are able to correctly classify crash and non-crash events roughly 95% of the time. The accuracy out of sample prediction of crash events (recall) is approximately 85%. The analysis of ROCs indicate that random forest and bagging are the best classifiers in terms of accuracy (true and false positive rates) for training and test, respectively. The importance of these results lies in the fact that the features used to classify normal (baseline) and crash events are solely extracted from speed profiles. It means such features can be quantified constantly in real-time through the use of sliding windows as inputs to the models for crash prediction. Proper warnings can be issued to the driver for crash avoidance purposes. Notably, increased computational power is enabling researchers to test different methods.

Specially, if predictability with high accuracy is more important than interpretability, then machine learning methods can be used quickly to obtain the best possible results.

Acknowledgement

This paper is based upon work supported by National Science Foundation (Award number: 1538139). Additional support was provided by the US Department of Transportation through the Collaborative Sciences Center for Road Safety, a consortium led by The University of North Carolina at Chapel Hill in partnership with The University of Tennessee. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect views of the sponsors. Special thanks to Dr. Tom Karnowski at Oak Ridge National Laboratory for provided the data. Dr. Haileab Hilafu is also appreciated for his comments.

CHAPTER VI
THE ROLE OF AGGRESSIVE DRIVING AND SPEEDING IN ROAD SAFETY: INSIGHTS FROM SHRP2 NATURALISTIC DRIVING STUDY DATA

Abstract

By harnessing the rich information available from naturalistic driving study data, this paper studies the impact of detailed pre-crash driving behavior and recently developed measures of driving volatility on crash and near-crash risks. Building on previous efforts in developing of driving volatility measures, highly correlated measures with crash risk were identified. Then driving behaviors contributing to both driving volatility and crash risk were explored. The paper contributes by incorporating driver, vehicle and infrastructure data collected in a naturalistic setting into the analysis as well as studying the near-crash risks. In particular, both direct and indirect effects of aggressive driving and speeding on crash and near-crash risks were investigated through path analysis. According to the results, aggressive driving is associated with increased risk of near-crash and crashes by 35% and 6% respectively. Speeding also was found to be correlated to increased risk of near-crash and crash events by 16% and 9% respectively. The study of near-crashes and its correlates in addition to traditional study of crash risks is a unique aspect of this investigation. The findings are beneficial in two ways. First, they are helpful in identifying dangerous pre-crash driving behaviors that are not generally available in conventional databases. Second, by avoiding such behaviors, drivers' crash risks can be reduced.

Introduction

Reducing traffic speed is the core principle of many Vision Zero approaches. Specifically, driving speed variation is associated with driving volatility, which is characterized by hard braking, hard accelerations, and jerky movements. New high resolution data has provided an unprecedented opportunity to capture the movements of vehicles in space and quantify the degree of driving stability. Specifically, if such data are enriched with videos of the driver behaviors, useful information can be included in the analysis to better investigate the driving factors to crash risk.

The absence of variables related to pre-crash driving behavior in crash risk models will impact the results due to the fact that human driver errors is the main contributing factor. An aggressive driver for example, increases the crash risk directly and indirectly. The driver might not be able to make proper instantaneous driving decisions due to aggressiveness increasing the crash risk, i.e. direct effect. They also might not pay attention to the vehicle speed and drive more volatilely increasing the crash risk through driving volatility, i.e. indirect effect. Therefore, pre-crash driving behavior related variables along with other common variables should be included in crash risk estimates to draw more concrete conclusions.

Given above arguments, the overall goal of this study is to revisit the study of crash risk factors by harnessing the rich information available from naturalistic driving data to study the impacts of detailed pre-crash driving behaviors and recently developed measures of driving volatility (*162*) on crash risks. In this framework, both direct and

indirect effects of pre-crash driving behavior on crash risk will be investigated. Since driving volatility is associated with higher crash risks, more fuel consumption, and increased emissions (*27*), identifying behaviors contributing to higher driving volatility is of importance. Thus, building upon on previous efforts in developing of driving volatility measures, first, highly correlated measures with crash occurrence will be identified and then driving behaviors contributing to driving volatilities and crash risks will be explored.

Given that driver errors is the main contributing factor in crashes (*32; 33*), the paper findings are beneficial as it identifies dangerous pre-crash driving behavior contributing to driving volatility as well as crashes.

Literature Review

The availability of naturalistic driving data in recent years, has made it possible to study detailed driving behaviors prior to crashes and their contributions. Driving behaviors which are mainly risky behaviors related to driving, have been of interests due to their relations with crashes and injury severity. In particular, aggressive driving and speeding have been found to be correlated with higher injury severity (*79; 163-165*). A meta-analysis of fifty one studies by Zhang and Chan (*165*) concluded that driving anger was a significant predictors of aggressive driving behaviors and was correlated with higher risk of involvement with accident related situations.

In the literature, many studies have explored the impact of phone use on reaction time and crash risks (*166-169*). According to Dingus et al. study, crash causation has shifted to driver related factors in recent years (*170*) among which distraction with handheld devices is the one of the main factors. A latent variable model done by Márquez et al. (*167*) on surveyed data revealed that persons with higher education are more likely to use phone while driving. Also, the traffic speed (*167*) and age (*168*) were found to be contributing to the phone use tendency while driving. Generally, phone use is considered a subgroup of secondary tasks e.g., reading, eating (*171*), cognitive and passenger related tasks (*172*). National Highway Traffic Safety Administration (NHTSA), through analysis on 100-car Naturalistic Driving Study (NDS) data, concluded that although drowsiness is associated with four to six times increase in near-crash and crash risks, involvement in complex and moderate secondary tasks are associated with increased risk of crashes by three and two times respectively compared with baseline driving (*173*).

The vehicle kinematics available from NDS data, has also opened windows for the researchers to explore how for instance, secondary task affects drivers' speed choice. Intuitively, unless vehicle is set on cruise control, secondary task should leave its trace on vehicle speed profile. The lateral movement of the vehicle can also be affected. However, it is difficult to conclude if the changes in vehicle kinematics because of the secondary task is a compensation strategy or is due to the inattention of the driver to the driving task. By studying SHRP2 NDS data, Schneidereit et al. (*174*) concluded that some

drivers adjust their speeds when texting, and the level of adjustment is considerably different among drivers. Other secondary tasks were found not to affect the speed behavior significantly. Ye et al. (*175*) was able to detect drivers' passenger interaction, call and texting by using vehicles speed, longitudinal and lateral acceleration, yaw rate and throttle position as inputs to a neural network with average accuracy of 99%.

Given the literature and the fact that the impact of pre-crash driving behaviors on driving volatility is lightly explored, this study investigates the effects of pre-crash driving behaviors on near-crash and crash events as well as driving volatility. As driving volatility was found to be correlated with higher risk of crashes (*1; 162; 176*), it is beneficial to investigate their overall impacts on near-crash and crash events. Specifically, direct and indirect effects of pre-crash driving behavior on near-crashes and crashes will be explored path analysis where measures of driving volatility are included as endogenous and exogenous variables.

Methodology

While the data used in this study is rich in terms of collected variables and attributes of the drivers, the paper adds more value to the data by using the instantaneous speed profiles of drivers. As such, measures of driving volatility are quantified to capture the amount of dispersion and fluctuation in vehicles speed and acceleration. Given the scope of this study and previously obtained results from other studies, nine measures of volatility are included in this paper. A comprehensive study of different measures of driving volatility can be found in (*162*). To make this paper stand-alone, a brief discussion of the measures of driving volatility that are use is provided below.

Measures of Driving Volatility

Standard Deviation

The standard deviation is probably the most common measure defined as:

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

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

Coefficient of Variation

By quantifying the ratio of the standard deviation to the mean, this measure is found to be correlated with higher crash risk (*1; 48*):

$$C_v = \frac{S_{dev}}{|\bar{x}|} * 100 \quad (2)$$

where $\bar{x}$ is the standard deviation ad the mean.

Quartile Coefficient of Variation

Using quantiles to capture the amount of dispersion in the data especially when the distribution is not normal is another measure of driving volatility which is called quartile coefficient of variation (*50*):

$$Q_{CV} = \frac{Q_3 - Q_1}{Q_3 + Q_1} * 100 \quad (3)$$

where Q_1 and Q_3 are the sample 25th and 75th percentiles respectively.

Time-varying stochastic volatility

Given the time series property of vehicles speed, use of time-varying stochastic volatility migh capture the amount of fluctuation and dispersion in the speed time-series data. This measure is defined as (*52; 53*):

$$V_f = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(\mathrm{r_t} - \bar{r})^2} \quad \text{from } i = 1 \text{ to } n \quad (4)$$

where

$$\mathrm{r_i} = \ln\left(\frac{x_t}{x_{t-1}}\right) * 100, \qquad x_t \; and \; x_{t-1} \geq 0 \quad (5)$$

and x_t and x_{t-1} are the observation at times *t* and *t-1* and *ln* is the natural logarithm. Given the constraint in Equation 5, this measure is applicable to the speed only.

Exclusion of Evasive Maneuver Duration

To calculate driving volatilities that can be used as pre-crash predictors, it is crucial to exclude the evasive maneuver duration when drivers were trying to avoid crashes. Figure 37 shows two speed and acceleration profiles; a baseline and a crash from the NDS data. For baseline cases (e.g., Figure 37, left), all seconds of speed and acceleration values (around 20 sec.) are used to obtain the driving volatilities. However, for a crash or near-crash case, only seconds up to the moment of driver reaction to the incident are used to calculate driving volatilities. In Figure 37 (right) for instance, 21 seconds of the data are used. There are several cases where the drivers did not react or reacted after the incident. Therefore, in such cases, seconds up to the moment of impact are used for calculation of driving volatilities. In other words, we used either the impact moment or the driver reaction moment whichever came first, to obtain the number of seconds to be included in calculating pre-crash driving volatility.

Modeling Approach

Structural equation modeling (hereafter path analysis) is used in this study to investigate direct and indirect factors contributing to crash and near-crash risks. Figure 38 depicts the simplified and overall structure of the model. Therefore, it can be said:

$$Y_1 = \beta_{11}X_1 + \beta_{21}X_2 + \epsilon_1 \tag{6}$$

where the model coefficients ($\beta_{11}$and β_{21}) are estimated using ordinary least square (OLS) method. Given the the categorical nature of event type as the dependent variable (Y_2) multinomial logistic regression (MNL) is used where the utility of each outcome can be written as:

$$U_{outcome\ i} = \beta_{12_{outcome\ i}}X_1 + \beta_{22_{outcome\ i}}X_2 + \gamma_{outcome\ i}Y_1 + \epsilon_{2_{outcome\ i}} \tag{7}$$

and the probability of outcome *i* is obtained by:

$$P_{outcome\ i} = \frac{Exp(U_{outcome\ i})}{\sum_{\forall I} Exp(U_{outcome\ i})} \tag{8}$$

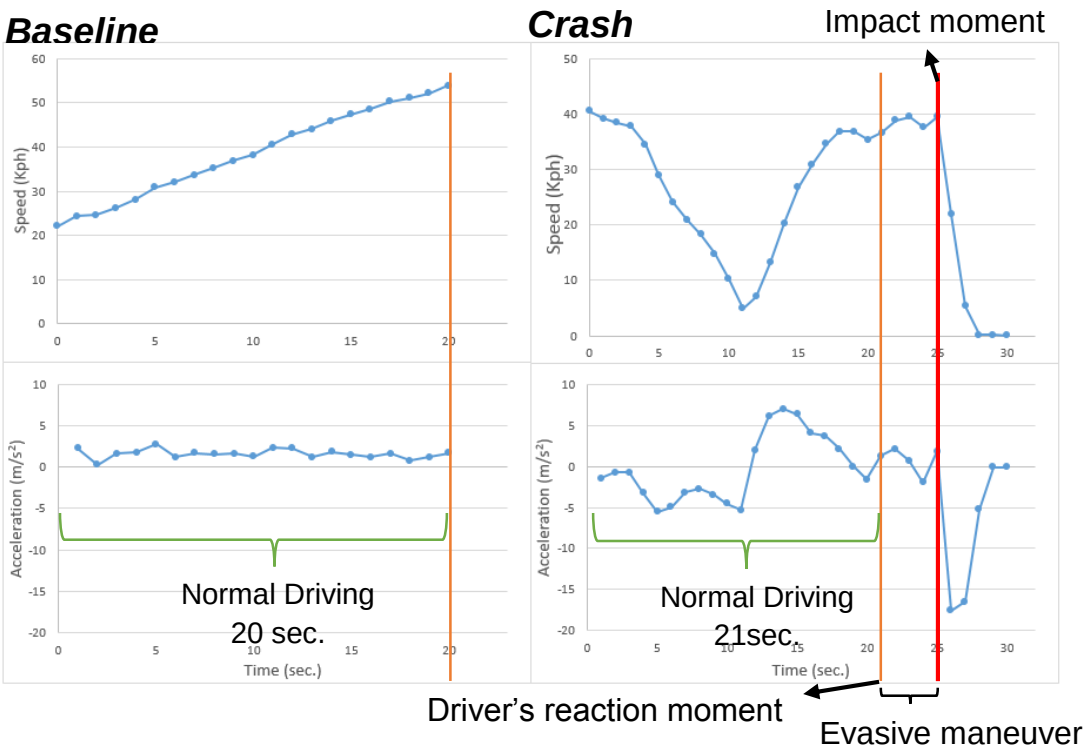


Figure 37. Comparison of baseline with crash/near crash driving profiles

In MNL regression, the coefficients (here $\beta_{12}, \beta_{22}, \gamma$ for each outcome) are estimated using maximum likelihood method.

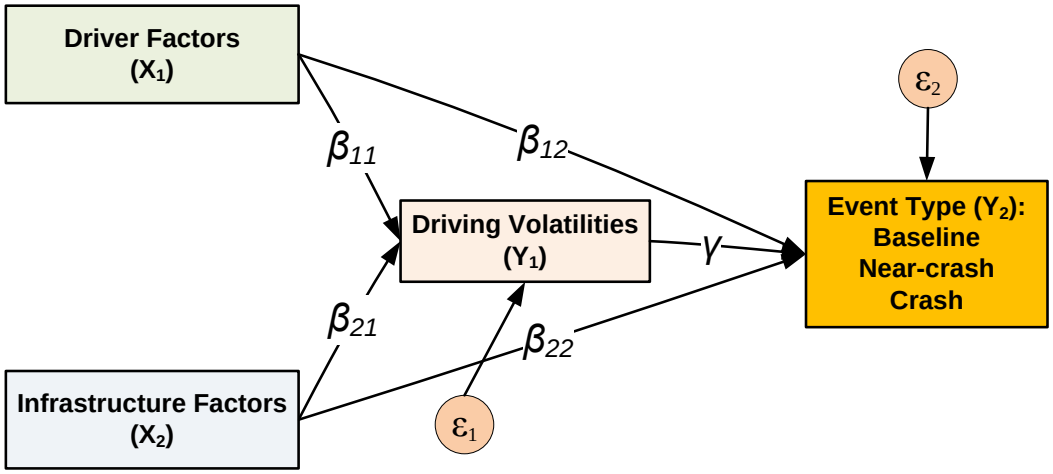


Figure 38. Conceptual framework of the model

Marginal Effects

Providing the marginal effects (ME) makes the estimated models easier to understand. Marginal effect is defined as the effect of one unit change in a given independent variable on the dependent variable. In case of linear regression, the marginal effects are indeed, the estimated coefficients of the variables. For multinomial logistic regression, they need to be calculated. As such, the impact of variable *z* with estimated coefficient of β_z is obtained by (*177*):

$$ME_{direct} = \frac{\partial P_{outcome\ i}}{\partial z_{outcome\ i}} = \beta_z P_{outcome\ i}\,(1 - P_{outcome\ i}) \tag{9}$$

Having obtained the direct marginal effects, indirect marginal effects of the variables on the outcome can be calculated. For example, referring to Figure 38, the indirect marginal effects of drivers factors on outcome *i* are calculated by:

$$ME_{indirect} = ME_{X_1 on\ Y_1} * \gamma \tag{10}$$

Therefore, total marginal effects of driver factors on outcomes will be the summation of direct and indirect outcomes:

$$ME_{total} = ME_{direct} + ME_{idirect} \tag{11}$$

Data

During Strategic Highway Research Program 2 (SHRP2) around four petabytes of naturalistic driving data (NDS) from around 3400 vehicles (35 million vehicle miles) from 2010 to 2013 in selected cities from states of Florida, Indiana, New York, North Carolina, Pennsylvania, and Washington were collected (*160; 161*). Data were collected using eight sensors and cameras capturing different views from the vehicles and drivers. The data also includes kinematics (e.g., speed, acceleration), and information about the drivers' behavior from the videos coded by data reductionists. They were following appropriate protocols for such reductions making it a high quality and rich dataset.

This study had access to a subset of the data. Having done the cleaning and error-checking and removing erroneous cases, the final dataset comprised of 9,268 trips taken by 1,546 drivers. For each case, the measures of driving volatility as explained earlier in the methodology section, were calculated from the speed and acceleration profiles and linked with trip summary to form the final dataset. The final dataset should have gone through some pre-processing, which is discussed next, to become ready for the final analysis.

Data Pre-processing and Recoding

Given the categorical nature of majority of variables related to driving behavior and secondary task, some recoding were necessary. For instance, variable "Secondary Task 1" had 62 different categories, most of which were similar and could be merged to form a cleaner and more intuitive variable. It should be noted in some cases, the reason of having so many categories for one variable was the different ways of coding it by different data reductionists. Overall, similarity, number of observations, intuition, judgment and authors' discretions contributed to determination of categories and merging process. Table 15 lists the recoded variables along with their original values and the final variable categories after recoding process. For the full list of variables including the ones used without recoding, refer to the descriptive statistics shown in Table 16.

Results

Descriptive Statistic

Table 16 provides the variables descriptive statistics. The table contains three types of variables in terms of their sources of information: driver, infrastructure and the vehicle. Driving behavior, secondary task, maneuver judgments are driver related variables. Variables such as weather, intersection influence, and traffic density are infrastructure related variables while seatbelt and airbag deployment data are vehicle related variables. Also, calculated driving volatilities shown separately in Table 3 are also vehicle related

Table 15. List of recoded variables and their final Categories

Variable	Affected Original Values in the Variable*	Coded as	Final Categories
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	• Aggressive • Sleepy, Fatigued • Improper Action • Low Speed • Speeding • Violation • Other • None
	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 Task	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	• Object Related • Body Related • Cell Phone • Interaction • External • Internal • Drink/Eat/Smoke • Talking/Singing • Other • None
	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	
	Other non-specific internal eye glance	Internal	
	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	
Weather	Fog/Raining/Snowing/Rain and Fog/Sleeting/Snow, sleet and Fog	Adverse Conditions	• No Adverse Conditions • Mist/Light Rain • Adverse Conditions
Surface Condition	Icy/Snowy	Ice/Snow	• Dry • Ice/Snow • Wet • Other
	Gravel, Dirt Road/ Gravel over Asphalt/ Unknown	Other	
Road Alignment	Curve left/Curve Right/Other	Other	• Straight • Curve

** Only modified variables have been provided in this table along with their final categories. For the full list of variables, see descriptive statistics Table 2.*

variables because they are obtained from vehicle kinematics. Table 16 columns also provide the descriptive statistics of variable according to the events type i.e. baseline, near-crash and crash. Such breaking provides beneficial insights about the key variables variation across event types which are discussed in the following before presenting the modeling results.

Driving Behavior

The percentage column of Table 16 (under total) for driving behavior shows that improper action is the most prevalent with 5.00% of occurrence followed by speeding (4.23%) and traffic violation (1.92%). Looking at the table row-wise, reveals some interesting observations. For example, the percentage of cases with speeding behavior in baseline is 3% while it is 9.28% in crashes. The change in the percentage of improper action category across the event type is the most among all other categories (baseline: 2.7%, near-crash: 7.26%, crash: 28.01%). The results also show that the percentage of aggressive behavior in near-crashes (4.39%) is considerably higher than baselines (0.11%).

Secondary Task

Although the overall variation of percentages across the event type for secondary task is less than driving behavior, in some of its categories, the variation is considerable. For instance, while cell phone usage is reported in 7.84% of the baselines, the percentage is more than two times in near-crashes (16.35%) and is 12.38% in crash profiles. Object distraction category also exhibits meaningful changes from baseline to near-crash and crash profiles (3.86%, 7.42% and 8.31% respectively). There are other categories whose percentages either do not vary significantly or if they do, they are not intuitive. For example, drink/eat/smoke percentages are 2.88%, 3.03% and 2.77% for baseline, near-crash and crash respectively. Similar trend are observed in internal, talking/singing, interaction and body related interaction. The statistical modeling results will shed more lights on the significance of these variables and their level of association with near-crash and crash risks.

Seatbelt Use

The overall seatbelt use percentage for the whole dataset is 94.57%. However, while the percentage of seatbelt use is 95% for baseline, it declines to 93.94% and 90.55% for near-crash and crashes respectively. Given that the seatbelt use does not intuitively affect the event types (being a baseline, near-crash or crash), the trend itself is interesting in a sense that probably drivers with non-use seatbelt attitude are more likely to be involved in crash and near-crash incidents. Although this variable was included in the statistical model, its association with crash and near-crash probability turned out to be not statistically significant and therefore, was omitted from the final analysis.

Table 16. Descriptive statistics of the categorical variables in NDS data

Variable	Category	Total (N = 9,268)		Baseline (N = 7,400)		Near-Crash (N = 1,254)		Crash (N = 614)	
		Perc.	Freq.	Perc.	Freq.	Perc.	Freq.	Perc.	Freq.
Driving Behavior	Aggressive	0.71%	66	0.11%	8	4.39%	55	0.49%	3
	Drowsy, sleepy	1.28%	119	1.26%	93	1.44%	18	1.3%	8
	Improper Action	5%	463	2.7%	200	7.26%	91	28.01%	172
	Low Speed	0.95%	88	1%	84	0.24%	3	0.16%	1
	None	85.27%	7,903	89.99%	6,659	72.41%	908	54.72%	336
	Other	0.64%	59	0.39%	29	1.2%	15	2.44%	15
	Speeding	4.23%	392	3.04%	225	8.77%	110	9.28%	57
	Violation	1.92%	178	1.38%	102	4.31%	54	3.58%	22
Distraction	Cell	9.29%	861	7.84%	580	16.35%	205	12.38%	76
	Drink/Eat/Smoke	2.89%	268	2.88%	213	3.03%	38	2.77%	17
	External	9.41%	872	9.43%	698	8.37%	105	11.24%	69
	Interaction	12.87%	1,193	13.22%	978	10.05%	126	14.5%	89
	Internal	3.24%	300	3.18%	235	3.91%	49	2.61%	16
	None	46.75%	4,333	49.27%	3,646	37.48%	470	35.34%	217
	Object Distraction	4.64%	430	3.86%	286	7.42%	93	8.31%	51
	Other	1.79%	166	1.39%	103	3.51%	44	3.09%	19
	Talking/singing	5.9%	547	5.89%	436	6.06%	76	5.7%	35
	Body Related	3.22%	298	3.04%	225	3.83%	48	4.07%	25
Maneuver Judgment	Safe and legal	89.35%	8,281	93.16%	6,894	75.52%	947	71.66%	440
	Safe but illegal	2.11%	196	2.31%	171	0.64%	8	2.77%	17
	Unsafe and illegal	5.89%	546	3.45%	255	16.11%	202	14.5%	89
	Unsafe but legal	2.64%	245	1.08%	80	7.74%	97	11.07%	68
Airbag Deployed	No	6.87%	637	0%	0	2.63%	33	98.37%	604
	Not Applicable	93.02%	8,621	100%	7,400	97.37%	1,221	0%	0
	Yes	0.11%	10	0%	0	0%	0	1.63%	10
Seatbelt	None used	5.43%	503	4.99%	369	6.06%	76	9.45%	58
	Used	94.57%	8,765	95.01%	7,031	93.94%	1,178	90.55%	556
Light	Darkness, lighted	13.78%	1,277	13%	962	15.39%	193	19.87%	122
	Darkness, not lighted	5.56%	515	6.09%	451	2.63%	33	5.05%	31
	Dawn/Dusk	4.52%	419	4.76%	352	3.67%	46	3.42%	21
	Daylight	76.14%	7,057	76.15%	5,635	78.31%	982	71.66%	440
Weather	Adverse Conditions	6.15%	570	5.91%	437	6.46%	81	8.47%	52
	Mist/Light Rain	4.09%	379	3.85%	285	4.63%	58	5.86%	36
	No Adverse Conditions	90%	8,319	90.24%	6,678	88.92%	1,115	86%	526
Surface	Dry	84.4%	7,822	85.65%	6,338	81.9%	1,027	74.43%	457
	Ice/snow	0.86%	80	0.58%	43	1.36%	17	3.26%	20
	Other	0.23%	21	0.2%	15	0.32%	4	0.33%	2
	Wet	14.51%	1,345	13.57%	1,004	16.43%	206	21.99%	135
Density (Level-of-service)	A1	40.19%	3,725	42.5%	3,145	19.38%	243	54.89%	337
	A2	30.13%	2,792	32.34%	2,393	22.73%	285	18.57%	114
	B	20.19%	1,871	18.47%	1,367	31.26%	392	18.24%	112
	C	6.11%	566	4.55%	337	15.71%	197	5.21%	32
	D	2.11%	196	1.27%	94	7.26%	91	1.79%	11
	E	1.01%	94	0.72%	53	2.95%	37	0.65%	4
	F	0.25%	23	0.14%	10	0.72%	9	0.65%	4
	Unknown	0.01%	1	0.01%	1	0%	0	0%	0

Table 16. (Continued)

Variable	Category	Total (N = 9,268)		Baseline (N = 7,400)		Near-Crash (N = 1,254)		Crash (N = 614)	
		Perc.	Freq.	Perc.	Freq.	Perc.	Freq.	Perc.	Freq.
Road Alignment	Curve	13.58%	1,259	13.97%	1,034	10.9%3	137	14.33%	88
	Straight	86.42%	8,009	86.03%	6,366	89.07%	1,117	85.67%	526
Road Grade	Grade Down	4.01%	372	3.3%	244	7.58%	95	5.37%	33
	Grade Up	7.62%	706	7.03%	520	10.29%	129	9.28%	57
	Level	87.88%	8,145	89.23%	6,603	81.5%	1,022	84.69%	520
	other	0.49%	45	0.45%	33	0.64%	8	0.65%	4
Construction Zone	No	96.21%	8,917	96.66%	7,153	93.78%	1,176	95.77%	588
	Yes	3.79%	351	3.34%	247	6.22%	78	4.23%	26
Traffic Flow	Divided	40.06%	3,713	41.39%	3,063	40.83%	512	22.48%	138
	No lanes	3.99%	370	3.03%	224	3.19%	40	17.26%	106
	Not divided - center	8.79%	815	9.03%	668	9.01%	113	5.54%	34
	Not divided - simple	43.25%	4,008	43.22%	3,198	40.91%	513	48.37%	297
	One-way traffic	3.91%	362	3.34%	247	6.06%	76	6.35%	39
Intersection Influence	No	71.9%	6,664	81.22%	6,010	31.98%	401	41.21%	253
	Unknown	0.08%	7	0.03%	2	0.24%	3	0.33%	2
	Yes, Interchange	3.18%	295	2.23%	165	8.77%	110	3.26%	20
	Yes, Other	1.12%	104	0.47%	35	4.39%	55	2.28%	14
	Yes, Parking lot, Driveway	4.53%	420	2.01%	149	11.96%	150	19.71%	121
	Yes, Stop Sign	3.34%	310	3.15%	233	3.19%	40	6.03%	37
	Yes, Traffic Signal	12.52%	1,160	9%	676	29.51%	370	19%	114
	Yes, Uncontrolled	3.32%	308	1.76%	130	9.97%	125	8.63%	53

Driving Volatilities

Table 17 mainly shows the descriptive statistics of the variables obtained from vehicle kinematics which quantify the fluctuations and the dispersion in vehicles instantaneous speed and acceleration (negative values of acceleration are named as deceleration). The table also provides the descriptive statistics of the moments of event occurrence in the profiles, evasive maneuver duration, whether the driver reacted to the event (not applicable to baselines), and the number of seconds used to calculate driving volatilities as per our discussion earlier. According to our previous study that developed several measures of driving volatility and comprehensively investigated their contribution to crash frequency (*162*), we were expecting that the volatilities included in this study, show intuitive variations across the event types. In Table 17, speed standard deviation as one of the simplest yet important correlates of crash risk shows an ascending trend from baseline to near-crash and crashes (1.64 m/s, 4.3 m/s and 3.9 m/s respectively). Similar trends can be seen for other driving volatilities of the speed and acceleration. Overall, near-crash and crash cases show higher driving volatilities than baselines.

Table 17 shows that the mean of pre-crash and pre-near-crash durations of observations are both 23.3 seconds. Furthermore, 99% of the drivers in near-crashes reacted before the impact while only 51% of driver did so in crashes. The mean of evasive maneuver durations are 0.96 and 1.58 seconds respectively which can also be observed from distributions of evasive maneuver durations in Figure 39, left. On average, drivers in near-crash situations reacted 0.62 seconds earlier than those in crashes.

Modeling Results

The results presented in this sections consist of three models of speed volatility, deceleration volatility and the event type. In volatility models, the driving behavior factors contributing to driving volatility are studied. Among the presented driving volatilities in Table 17, Speed V_f and Deceleration C_v resulted more intuitive and better models in terms of goodness of fit. These driving volatilities in turn, along with other variables, play roles as independent variables for the event type model. Therefore, in our path analysis framework, driving volatilities are both endogenous and exogenous, event type is endogenous (caused) and all other variables are exogenous (not caused) variables. Figure 40 depicts the final model structure which is obtained through forward stepwise variable selection process based on variables statistical significance and intuition as well as model fit and parsimony. Given the probable correlation between speed and deceleration volatilities, their covariance is also included in the model (curved arrow). According to the established structure in Figure 40, pre-crash driving behavior and secondary task contribute to near-crash and crash events directly. However, they also indirectly affect the probability of such events through contributing to speed and deceleration driving volatilities. Among the infrastructure variables data which are included in the model as control variables, traffic flow and intersection influence contribute to driving volatilities while traffic density (level of service) and light are directly correlated with event type. In the following, the models are discussed in more details.

Table 17. Descriptive statistics of calculated variables from NDS data

Variable	Total (N = 9,268)		Baseline (N = 7,400)		Near-crash (N = 1,254)		Crash (N = 614)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Speed S_{Dev} (m/s)	2.15	1.96	1.64	1.5	4.3	2.16	3.9	2.35
Speed V_f	0.17	0.35	0.08	0.22	0.54	0.46	0.6	0.43
Speed V_f (log)	-2.98	1.53	-3.46	1.2	-1.16	1.19	-0.91	1.04
Speed C_V	0.23	0.3	0.14	0.16	0.55	0.42	0.65	0.4
Speed Q_{CV}	0.18	0.23	0.11	0.14	0.41	0.31	0.51	0.3
Acceleration S_{Dev}	0.65	0.54	0.51	0.45	1.26	0.51	1.01	0.5
Acceleration C_V	0.76	0.26	0.73	0.24	0.87	0.3	0.91	0.33
Deceleration C_V	0.84	0.32	0.76	0.23	1.24	0.4	1.04	0.37
Acceleration Q_{CV}	0.58	0.19	0.57	0.19	0.64	0.19	0.67	0.19
Deceleration Q_{CV}	0.62	0.18	0.59	0.18	0.72	0.17	0.71	0.18
Impact Moment (sec.)	--	--	--	--	23.27	1.75	23.29	2.00
Reacted to the Event (yes = 1)	--	--	--	--	0.99	--	0.51	--
Evasive Maneuver Duration (sec.)*	--	--	--	--	1.58	1.15	0.96	1.07
Seconds Used for Calculation of Driving Volatilities (sec.)	21.18	0.93	20.95	0.03	21.71	1.54	22.80	2.06

** Only for cases where drivers reacted ($N_{nearcrash}$ = 1241, N_{crash} = 314)*

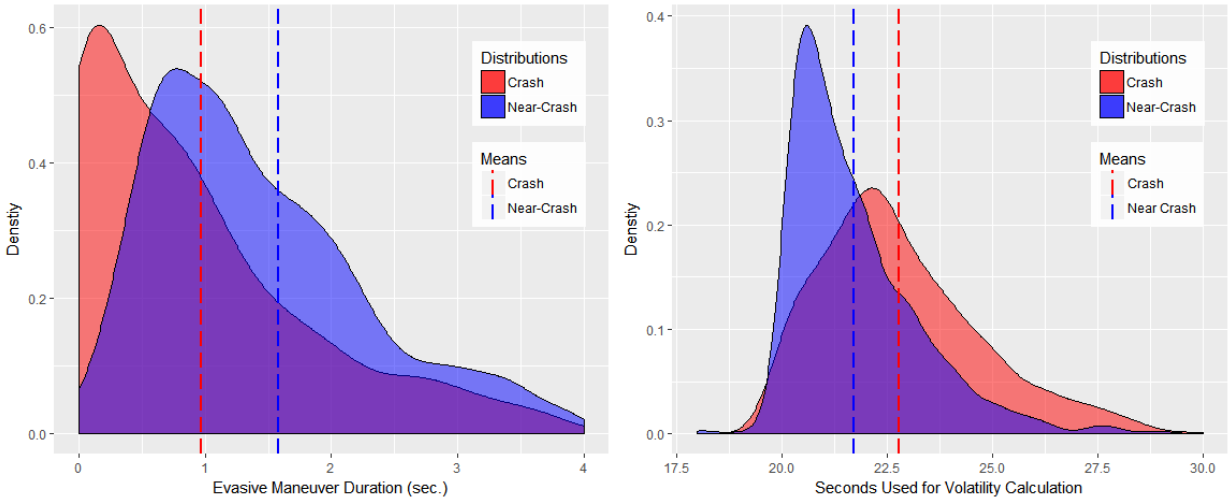


Figure 39. Distribution of evasive maneuver (left) and second used for volatilities calculation

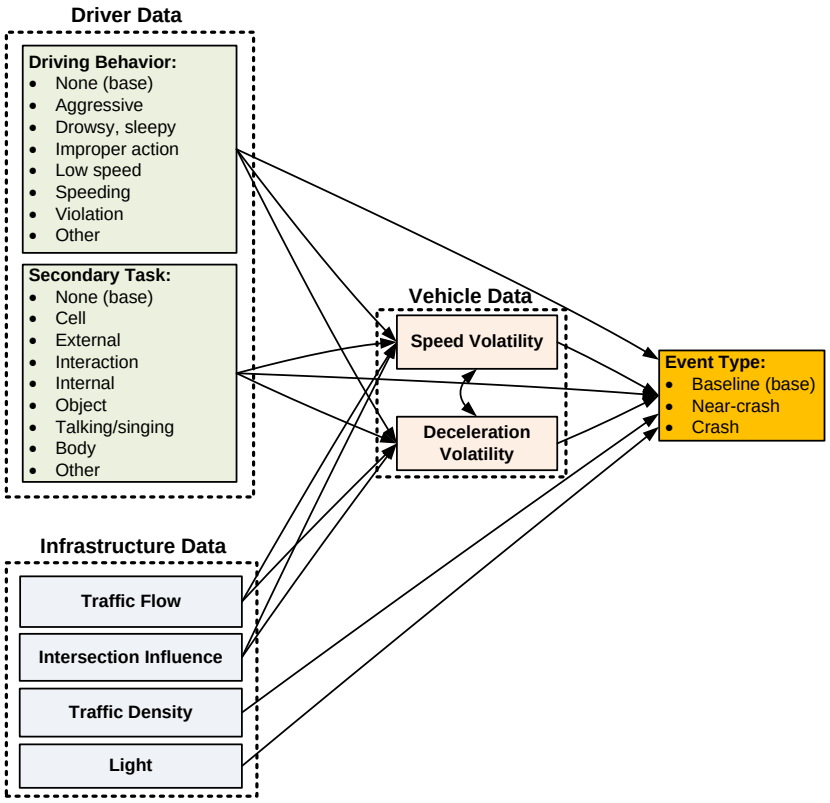


Figure 40. Final model structure

Speed Volatility Model

This model evaluates the correlates of speed volatility (Speed V_f) in terms of driving behavior. Table 18 presents the modeling results and fit summary statistics. Given the continuous nature of Speed V_f as the dependent variable, a log transformation was performed to make the distribution closer to normal distribution. The descriptive statistics of the transformed variable of Speed V_f was provided earlier in Table 17. According to Table 18, the main contributing factor to speed volatility is aggressive driving behavior. In the data, behaviors such as directed menacing actions, cutting in, driving too close behind or front of other vehicles were coded as aggressive driving behavior. Improper action, speeding and traffic violation behaviors are found to be associated with increased speed volatility. Improper action variable includes several actions listed in Table 15 that are not illegal per se but are considered risky, such as driving in other vehicle's blind spot, fail to signal (it is illegal in some states), sudden braking, passing on right etc. When it comes to secondary task, almost all of its categories are statistically associated with increased speed volatility among which cell phone is the factor with most contributory effect. Although the coefficients of the "other" category is a bit higher than cell, it should be noted that the former is the product of merging less uniform secondary tasks accumulating their effects in one estimated coefficient.

Deceleration Volatility Model

Table 19 shows the results for the model with deceleration volatility as the dependent variable. As compared to the previous model, speed volatility model was a better fit in terms of adjusted R^2. However, still aggressive driving, speeding and almost all of secondary tasks were found to be associated with increased deceleration volatility. While being sleepy was not correlated with speed volatility, it is associated with increased deceleration volatility in this model.

Event Type Model

Table 20 shows the results of event type model where driving volatilities along with other key variables were included in a logistic regression model. According to the results, both speed and acceleration driving volatilities are associated with increased probability of being involved with near-crash and crash incidents. The marginal effects of the variables are provided in the last column of the table. For continuous variables, the marginal effect is defined as the amount of increase in the risk of near-crashes or crashes with one unit increase in the continuous independent variables which are speed and deceleration volatilities. Therefore, one unit increase in speed and deceleration volatilities is correlated with 4.34% and 18.12% higher chances of involving in near-crashes and 3.65% and 3.49% higher chance of crashes respectively. For categorical variables, the marginal effect of each *category* (of the variable) indicates the amount of risk increase or decrease of near-crash and crashes compared with the base category.

Table 18. The speed volatility model

Variable	Category	Estimate [a]	Std. Err.	Z value
Intercept	--	-4.059***	0.026	-158.49
Driving Behavior	None	--	--	--
	Aggressive	1.387***	0.150	9.23
	Drowsy, sleepy	0.172	0.112	1.53
	Improper Action	0.611***	0.060	10.2
	Low Speed	-0.156	0.130	-1.19
	Speeding	0.144**	0.063	2.29
	Violation	0.627***	0.096	6.51
	Other	0.764***	0.159	4.81
Secondary Task	None	--	--	--
	Cell	0.370***	0.045	8.14
	Drink/Eat/Smoke	0.177**	0.076	2.32
	External	0.242***	0.045	5.36
	Interaction	0.118***	0.040	2.97
	Internal	0.214***	0.072	2.96
	Object	0.365***	0.061	5.96
	Other	0.378***	0.096	3.94
	Talking/singing	0.085***	0.055	1.55
	Body Related	0.244***	0.073	3.36
Traffic Flow	Divided	--	--	--
	No lanes	1.808***	0.069	26.36
	Not divided – center	0.585***	0.047	12.39
	Not divided – simple	0.761***	0.029	26.67
	One-way traffic	0.742***	0.069	10.78
Intersection Influence	No	--	--	--
	Yes, Interchange	0.737***	0.075	9.78
	Yes, Parking lot, Driveway	1.500***	0.064	23.45
	Yes, Stop Sign	1.212***	0.074	16.27
	Yes, Traffic Signal	1.649***	0.039	42.2
	Yes, Uncontrolled	1.516***	0.071	21.22
	Yes, Other	1.979***	0.121	16.42
	Unknown	2.369***	0.459	5.16
Summary Statistics	Number of observations	9268		
	R^2	0.3726		
	Adjusted R^2	0.3707		
	Root mean square error	1.2132		
	AIC	29,905		
	BIC	30,105		

[a] Significance codes: ***0.1%, **1%, *5%

Table 19. Deceleration volatility model

Variable	Category	Estimate [a]	Std. Err.	Z value
Intercept	--	0.770***	0.007	116.91
Driving Behavior	None	--	--	--
	Aggressive	0.231***	0.039	5.97
	Drowsy, sleepy	0.047*	0.029	1.63
	Improper Action	-0.016	0.015	-1.02
	Low Speed	-0.023	0.033	-0.68
	Speeding	0.102***	0.016	6.28
	Violation	0.032	0.025	1.3
	Other	0.119***	0.041	2.91
Secondary Task	None	--	--	--
	Cell	0.088***	0.012	7.52
	Drink/Eat/Smoke	0.034**	0.020	1.73
	External	0.023*	0.012	1.94
	Interaction	0.016	0.010	1.6
	Internal	0.054***	0.019	2.92
	Object	0.055***	0.016	3.51
	Other	0.124***	0.025	5.03
	Talking/singing	0.040***	0.014	2.82
	Body Related	0.024	0.019	1.28
Traffic Flow	Divided	--	--	--
	No lanes	0.011	0.018	0.6
	Not divided - center	0.005	0.012	0.45
	Not divided - simple	-0.008	0.007	-1.11
	One-way traffic	-0.036**	0.018	-2.05
Intersection Influence	No	--	--	--
	Yes, Interchange	0.153***	0.019	7.91
	Yes, Parking lot, Driveway	0.189***	0.016	11.48
	Yes, Stop Sign	0.043**	0.019	2.22
	Yes, Traffic Signal	0.149***	0.010	14.81
	Yes, Uncontrolled	0.227***	0.018	12.33
	Yes, Other	0.268***	0.031	8.65
	Unknown	0.516***	0.118	4.37
Summary Statistics	Number of observations	9268		
	R^2	0.0763		
	Adjusted R^2	0.0736		
	Root mean square error	0.3121		
	AIC	4746.0		
	BIC	4945.8		
[a] Significance codes: ***0.1%, **1%, *5%				

Table 20. Crash and near-crash model

	Variable	Category	Estimate [a]	Std. Err.	Z value	M.E.
Baseline		--	--	--	--	--
Near-crash	Intercept	--	-4.361***	0.211	-20.72	--
	Speed V_f (log)	--	1.354***	0.044	30.71	4.34%
	Deceleration C_V	--	4.289***	0.165	25.98	18.12%
	Driving Behavior	None	--	--	--	--
		Aggressive	3.248***	0.502	6.47	24.43%
		Drowsy, sleepy	0.594	0.428	1.39	1.58%
		Improper Action	1.443***	0.174	8.28	0.98%
		Low Speed	-0.082	0.765	-0.11	0.43%
		Speeding	2.677***	0.202	13.28	13.50%
		Violation	1.024***	0.248	4.13	5.21%
		Other	1.386***	0.458	3.03	2.92%
	Distraction	No Distraction	--	--	--	--
		Cell	0.850***	0.157	5.43	2.78%
		Drink/Eat/Smoke	0.124	0.291	0.43	-0.18%
		External	-0.125	0.181	-0.69	-1.85%
		Interaction	0.171	0.159	1.08	-0.54%
		Internal	0.319	0.269	1.18	1.38%
		Object	0.930***	0.212	4.39	2.28%
		Other	0.827***	0.317	2.61	2.92%
		Talking/singing	0.025	0.208	0.12	-0.29%
		Body Related	0.548**	0.267	2.06	1.70%
	Light	Daylight	--	--	--	--
		Darkness, lighted	0.222	0.139	1.6	0.32%
		Darkness, not lighted	0.584**	0.254	2.29	1.87%
		Dawn/Dusk	-0.380	0.257	-1.48	-1.40%
	Density	Level of service A1	--	--	--	--
		Level of service A2	1.079***	0.134	8.08	6.26%
		Level of service B	1.742***	0.137	12.72	10.03%
		Level of service C	2.114***	0.179	11.79	13.31%
		Level of service D	2.555***	0.259	9.87	17.04%
		Level of service E	2.015***	0.332	6.06	13.88%
		Level of service F	1.253*	0.715	1.75	6.37%
		Unknown	-8.951	500.407	-0.02	-7.90%
Crash	Intercept	--	-3.086***	0.233	-13.27	--
	Speed V_f (log)	--	1.567***	0.053	29.49	3.65%
	Deceleration C_V	--	3.117***	0.190	16.42	3.49%
	Driving Behavior	None	--	--	--	--
		Aggressive	1.824**	0.760	2.4	0.33%
		Drowsy, sleepy	0.895**	0.489	1.83	2.61%
		Improper Action	2.799***	0.163	17.17	14.25%
		Low Speed	-0.513	1.137	-0.45	-1.50%
		Speeding	2.815***	0.230	12.26	8.20%
		Violation	0.667**	0.283	2.35	0.44%
		Other	2.087***	0.452	4.62	2.28%
	Distraction	No Distraction	--	--	--	--
		Cell	1.088***	0.188	5.8	2.77%
		Drink/Eat/Smoke	0.412	0.344	1.2	1.40%
		External	0.519***	0.197	2.63	2.48%
		Interaction	0.673***	0.178	3.77	2.50%
		Internal	0.239	0.345	0.69	0.24%
		Object	1.442***	0.234	6.16	4.55%
		Other	0.981***	0.371	2.64	2.28%
		Talking/singing	0.206	0.251	0.82	0.75%

Table 20. (Continued)

	Variable	Category	Estimate [a]	Std. Err.	Z value	M.E.
	Distraction	Body Related	0.721**	0.303	2.38	1.77%
	Light	Daylight	--	--	--	--
		Darkness, lighted	0.419***	0.151	2.77	1.37%
		Darkness, not lighted	0.716***	0.274	2.61	1.89%
		Dawn/Dusk	-0.332	0.310	-1.07	-0.52%
	Density	Level of service A1	--	--	--	--
		Level of service A2	-0.014	0.150	-0.09	-2.87%
		Level of service B	0.379**	0.157	2.42	-2.93%
		Level of service C	0.296	0.241	1.23	-4.23%
		Level of service D	0.375	0.382	0.98	-4.99%
		Level of service E	-0.398	0.578	-0.69	-6.09%
		Level of service F	0.376	0.822	0.46	-1.53%
		Unknown	-10.228	715.134	-0.01	-8.77%
	Var (Speed V_f (log))	--	1.467	0.021	1.425	--
	Var (Deceleration C_V)	--	0.097	0.001	0.094	--
	Covariance	--	0.125	0.004	30.13	--
Summary Statistics		Number of observations	9268			
		Pseudo R^2	0.5350			
		Likelihood at intercept	-5840.1			
		Likelihood at convergence	-2715.9			
		AIC	5547.7			
		BIC	5691.5			
[a] Significance codes: ***0.1%, **1%, *5%						

Therefore, aggressive driving behavior is associated with increased risk of near-crash by 24.43% compared with the base which means no specific driving behavior. In the same vein, drivers with speeding behavior are 13.50% more likely to be involved with near-crash events compared with the base. When it comes to crash events, speeding is correlated with increased risk of crashes by 8.20%. Behaviors attributed to improper actions show association with increased risk of being involved in crash events by 14.25%. Among the categories of the secondary task variable, “cell phone”, “object” and “other” are associated with increased risk of near-crashes for 2.78%, 2.28% and 2.92% respectively. In crash cases, those tasks are associated with 2.77%, 4.25% and 2.28% increased crash risks. It should be noted, the presented marginal effects of aggressive, speeding and secondary tasks were indeed the direct effects. In the next section, their overall marginal effects which includes both direct and indirect effects (through impacting driving volatilities) will be discussed.

Path Analysis and Overall Marginal Effects

Table 21 provides the marginal effects of the three discussed models where the effect of independent variables on near-crash and crash events collectively are shown in the last two columns. In the table, the values under “**Direct Marginal Effects**” are the marginal effects of the model discussed in Table 6. The “**Indirect Marginal Effects**” of each

variable are indeed their estimated coefficients from Table 4 (for speed V_f) and Table 5 (for deceleration C_v) multiplied by the direct marginal effects of speed V_f and deceleration C_v shown in first two rows of Table 7. For example: indirect effect of speed V_f in case of aggressive driving on near-crash is calculated in this way: 1.387 * 4.344% = 6.03%. Therefore, the total marginal effect of aggressive driving for near-crash is obtained by summation of the direct and indirect marginal effects: 24.43% + 6.03% + 4.18% = 34.64%. In Table 21, top three marginal effects in terms of magnitude for driving behavior and secondary task are in bold face.

Aggressive driving, speeding and violation are found to be associated with 36.64%, 15.97% and 8.52% increased chance of being involved in near-crashes respectively. In addition, improper action, speeding and aggressive driving are correlated with increased risk of crashes by 16.43%, 9.08% and 6.20% respectively. Therefore, aggressive driving and speeding are major contributing factors in both near-crash and crash events.

In case of secondary tasks, “cell”, “object” and “other” are top three contributing factors to near-crash and crashes. As such, use of cell phone (and other similar devices such as tablets) while driving are associated with 5.98% and 4.43% increased chance of near-crash and crashes respectively. “Other” tasks (refer to Table 15 for the categories) also found to be correlated with 6.82% and 4.10% increase in near-crash and crash risks respectively. The total marginal effects of object related tasks on near-crash and crash chances are 4.87% and 6.07% respectively.

Table 21. Marginal effects

Variable	Category	Direct Marginal Effect		Indirect M.E. via speed V_f		Indirect M.E. via deceleration C_v		Total Marginal Effect	
		Near-crash	Crash	Near-crash	Crash	Near-crash	Crash	Near-crash	Crash
Speed V_f	--	4.34%	3.65%	--	--	--	--	4.34%	3.65%
Deceleration C_v	--	18.12%	3.49%	--	--	--	--	18.12%	3.49%
Behavior	None	--	--	--	--	--	--	--	--
	Aggressive	24.43%	0.33%	6.03%	5.06%	4.18%	0.80%	**34.64%**	**6.20%**
	Drowsy, sleepy	1.58%	2.61%	0.75%	0.63%	0.85%	0.16%	3.18%	3.40%
	Improper Action	0.98%	14.25%	2.65%	2.23%	-0.29%	-0.05%	3.35%	**16.43%**
	Low Speed	0.43%	-1.50%	-0.68%	-0.57%	-0.41%	-0.08%	-0.65%	-2.15%
	Speeding	13.50%	8.20%	0.63%	0.53%	1.84%	0.35%	**15.97%**	**9.08%**
	Violation	5.21%	0.44%	2.72%	2.29%	0.58%	0.11%	**8.52%**	2.83%
	Other	2.92%	2.28%	1.64%	1.38%	2.25%	0.43%	6.82%	4.10%
Secondary Task	None	--	--	--	--	--	--	--	--
	Cell	2.78%	2.77%	1.61%	1.35%	1.59%	0.31%	**5.98%**	**4.43%**
	Drink/Eat/Smoke	-0.18%	1.40%	0.77%	0.65%	0.62%	0.12%	1.21%	2.16%
	External	-1.85%	2.48%	1.05%	0.88%	0.41%	0.08%	-0.39%	3.45%
	Interaction	-0.54%	2.50%	0.51%	0.43%	0.30%	0.06%	0.27%	2.99%
	Internal	1.38%	0.24%	0.93%	0.78%	0.98%	0.19%	3.30%	1.21%
	Object	2.28%	4.55%	1.59%	1.33%	1.00%	0.19%	**4.87%**	**6.07%**
	Other	2.92%	2.28%	1.64%	1.38%	2.25%	0.43%	**6.82%**	**4.10%**
	Talking/singing	-0.29%	0.75%	0.37%	0.31%	0.72%	0.14%	0.81%	1.20%
	Body	1.70%	1.77%	1.06%	0.89%	0.43%	0.08%	3.19%	2.74%
Intersection Influence	No	--	--	--	--	--	--	--	--
	Yes, interchange	--	--	3.20%	2.69%	2.78%	0.54%	5.98%	3.23%
	Yes, parking	--	--	8.60%	7.22%	4.86%	0.94%	13.46%	8.16%
	Yes, stop sign	--	--	6.52%	5.47%	3.42%	0.66%	9.94%	6.13%
	Yes, traffic light	--	--	5.26%	4.42%	0.77%	0.15%	6.03%	4.57%
	Yes, uncontrolled	--	--	7.16%	6.02%	2.70%	0.52%	9.86%	6.54%
	Yes, other	--	--	6.59%	5.53%	4.11%	0.79%	10.69%	6.32%
	Unknown	--	--	10.29%	8.64%	9.36%	1.80%	19.65%	10.44%
Traffic Flow	Divided	--	--	--	--	--	--	--	--
	No lanes	--	--	7.85%	6.59%	0.19%	0.04%	8.04%	6.63%
	Not divided - center	--	--	2.54%	2.13%	0.10%	0.02%	2.64%	2.15%
	Not divided - simple	--	--	3.31%	2.78%	-0.15%	-0.03%	3.16%	2.75%
	One-way traffic	--	--	3.22%	2.71%	-0.66%	-0.13%	2.56%	2.58%
Density (Level of service)	A1	--	--	--	--	--	--	--	--
	A2	6.26%	-2.87%	--	--	--	--	6.26%	-2.87%
	B	10.03%	-2.93%	--	--	--	--	10.03%	-2.93%
	C	13.31%	-4.23%	--	--	--	--	13.31%	-4.23%
	D	17.04%	-4.99%	--	--	--	--	17.04%	-4.99%
	E	13.88%	-6.09%	--	--	--	--	13.88%	-6.09%
	F	6.37%	-1.53%	--	--	--	--	6.37%	-1.53%
	Unknown	-7.90%	-8.77%	--	--	--	--	-7.90%	-8.77%
Light	Daylight	--	--	--	--	--	--	--	--
	Darkness, lighted	0.32%	1.37%	--	--	--	--	0.32%	1.37%
	Darkness, not lighted	1.87%	1.89%	--	--	--	--	1.87%	1.89%
	Dawn/Dusk	-1.40%	-0.52%	--	--	--	--	-1.40%	-0.52%

Limitations

From the methodological standpoint, as with any statistical method, path analysis has its own limitations. A detailed discussion of the limitations is provided by Tomarken and Waller (*178*). For instance, the omitted variables might result in biased estimation of coefficients and the standard errors. It should be noted that even inclusion of residual terms in the model does not address the limitation because their variances are estimated freely (*178*). Note that obtained estimates and standard errors are unbiased with assumption of the correctly specified model. To elaborate, the maximum likelihood estimation process utilizes all the information existed in the covariance matrix while ignoring the effects of misspecified parameters that were not included in the equations. In path analysis, the convergence can be also a problem sometimes. To address the heterogeneity from drivers taken several trips, the participants unique ID were used in a hierarchical structure. However the model did not converge. Random parameter multinomial logistic and random parameter linear regression models can be used instead to account for heterogeneity.

The NDS data also has limitations. Given that drivers are voluntarily participated in the program (with monetary incentives), the data is not representative. Also, the difference between the proportion of events in real-world and in the data might have affected the results. In reality, crashes are rare events while they form 6.6% of the NDS data. In addition, there might be some variations among the judgments of data reductionists who created and coded the variables used in this study from the recorded videos. Apart from person to person judgment differences, a specific data reductionist can be biased in coding driving behaviors if they know that video is a crash event in the sense that they might have tried to find a behavior for the justification of the event.

Conclusions

The available information in naturalistic driving study data coupled with the capability of the developed measures of driving volatility, made it possible to analyze near-crash and crash events by incorporating the information from the famous triangle of driver, infrastructure and vehicle. Using path analysis, the direct and indirect effects of different pre-crash driving behaviors on near-crash and crash risks were studied. There are two new aspects of this paper that contribute to transportation safety. First, is adding a new dimension to the study by incorporating measures of driving volatility capturing microscopic driving behavior in terms of speed and acceleration as both endogenous and exogenous variables. Second is the study of near-crashes and its correlates in addition to traditional study of crash risks.

The results revealed that aggressive driving, speeding, cell phone and object related tasks are the top contributing factors to near-crash and crashes. As such, aggressive driving is associated with increased risk of near-crash and crashes by 35% and 6%

respectively. Speeding also was found to be associated with increased risk of near-crash and crash events by 16% and 9% respectively. The omnipresence of cellphones has made them the main contributing factors among secondary tasks. According to the results, cell phone is correlated with 6% and 4.4% increased risk of being involved with near-crash and crashes respectively. Tasks related to other objects in the vehicle, are also associated with 4.9% and 6% increased near-crash and crash risks.

The paper also showed how different types of secondary tasks are correlated with driving volatility. As expected, among pre-crash driving behaviors, speeding and aggression were also associated with higher driving volatilities. Almost all types of secondary tasks were found to be correlated with increased driving volatility among which cell phone was the main factor. Future research can explore how technologies such as connected and automated vehicles, advance traveler information systems, driving assist systems, etc. can be leveraged in order to detect and reduce identified dangerous pre-crash driving behavior.

Acknowledgement

This paper is based upon work supported by National Science Foundation (Award number: 1538139). Additional support was provided by the US Department of Transportation through the Collaborative Sciences Center for Road Safety, a consortium led by The University of North Carolina at Chapel Hill in partnership with The University of Tennessee. The data were provided by Dr. Tom Karnowski, Oak Ridge National Laboratory, for which the authors are very grateful. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect views of the sponsors.

CONCLUSION

This dissertation conceptualized measures of driver, vehicle, and road infrastructure volatility and proposed a framework for monitoring them by fusing and processing streams of data. Specifically, real-time driving event processing was suggested for learning from the streams and taking proactive measures at critical moments to avoid crashes. The proposed event processing framework provides a way to bring the data to life and gives us a greater situational awareness.

In the second chapter, we created new and unique database (BSM data integrated with crash and inventory data) and mined critical information from large-scale BSM data. More than 2,500,000,000 BSMs were processed along with crash and inventory data from 116 intersections in the city of Ann Arbor, Michigan. To explore relationships between measures of driving volatility and crash frequency at intersections, rigorous statistical models were estimated. Three measures of driving volatilities show positive and statistically significant association with crash frequencies at the intersections. In practice, depending on the detected reasons, proactive countermeasures can be taken to reduce drivers' volatility. The study showed how to extract useful information in the form of driving volatility from newly available BSM data. Such data are increasingly becoming available, providing a valuable resource for studying vehicle kinematics and microscopic behaviors of drivers, e.g., instantaneous vehicle speed, acceleration, and jerk.

Chapter three proposed the concept of location-based volatility that captures the extent of variations in instantaneous driving decisions. Five-year crash frequencies, AADT, speed limits, and number of approaches for all intersections are extracted and linked with location-based volatilities. Significant efforts went into data processing, collection, and linkage. The modeling results revealed that most of computed C_V s (coefficient of variation as measures of volatilities) are positively associated with crash frequency. The study has implications for proactive intersection safety management. The findings are important in the sense that if many drivers behave in a volatile manner at a specific intersection (exhibit higher variability in longitudinal accelerations), then such intersections can be identified before accidents happen. Proactive warnings and alerts can be generated about potential hazards at specific intersections and transmitted to drivers via connected vehicle technologies such as road-side equipment. These can in turn increase drivers' situational and safety awareness, and help them pursue safer driving at dangerous intersections.

Chapter four contributes by using high velocity and diverse driving data to learn driving decisions. Specifically, the reward and states were defined theoretically and real-life driving decisions are analyzed to explore correlations with contextual factors i.e., proximity to surrounding objects. Markov Decision Process (MDP) is applied to learn personally revealed choices of drivers in terms of acceleration, deceleration and

maintaining a constant speed. The results show that with increasing objects around a host vehicle, drivers would rather accelerate to avoid crowdedness around them. One potential application of this study is to generate short-term predictive information about driver decisions, which can be used to warn the driver when they deviate substantially from their own Personally Revealed Choices (PRC's) based on their own historical driving performance. The other is to use policies from expert drivers for training novice drivers or drivers who are more prone to near-misses and crashes. Yet another is to disseminate preferred actions of a driver to surrounding vehicles, enabling surrounding vehicles to foresee the states and actions of other drivers' potentially avoiding collisions.

Chapter five used data from the Naturalistic Driving Study (NDS) to estimate several crash risk models. Several statistical and machine learning methods were used including decision trees, bagging, boosting, random forest, logistic regression, lasso logistic regression, genetic algorithm logistic regression and artificial neural networks. The predictions of these models were assessed in terms of misclassification rate. Twenty three measures of driving volatility were applied to vehicles speed, acceleration and jerk and used as predictors in the models. The results show that neural networks and random forest outperformed the other methods with overall (crash and non-crash) training and test errors of 4.74% and 5.19%, respectively which means that these methods are able to correctly classify crash and non-crash events roughly 95% of the time. The importance of these results lies in the fact that the features used to classify normal (baseline) and crash events are solely extracted from speed profiles. It means such features can be quantified constantly in real-time through the use of sliding windows as inputs to the models for crash prediction. Proper warnings can be issued to the driver for crash avoidance purposes.

Finally, chapter six analyzed near-crash and crash events by incorporating the information from the famous triangle of driver, infrastructure and vehicle. There are two new aspects of this chapter that contribute to transportation safety. First, is adding a new dimension to the study by incorporating measures of driving volatility capturing microscopic driving behavior in terms of speed and acceleration as both endogenous and exogenous variables. Second is the study of near-crashes and its correlates in addition to traditional study of crash risks. The results revealed that aggressive driving, speeding, cell phone and object related tasks are the top contributing factors to near-crash and crashes. As such, aggressive driving is associated with increased risk of near-crash and crashes by 35% and 6% respectively. Speeding also was found to be associated with increased risk of near-crash and crash events by 16% and 9% respectively. The omnipresence of cellphones has made them the main contributing factors among secondary tasks. According to the results, cell phone is correlated with 6% and 4.4% increased risk of being involved with near-crash and crashes respectively. Tasks related

to other objects in the vehicle, are also associated with 4.9% and 6% increased near-crash and crash risks.

REFERENCES

1 . Kamrani, M., B. Wali, and A. J. Khattak. Can Data Generated by Connected Vehicles Enhance Safety? Proactive Approach to Intersection Safety Management. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2659, 2017, pp. 80-90. http://dx.doi.org/10.3141/2659-09.
2 . Liu, J., and A. J. Khattak. Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles. *Transportation Research Part C: Emerging Technologies,* Vol. 68, 2016, pp. 83-100.
3 . Wang, X., A. J. Khattak, J. Liu, G. Masghati-Amoli, and S. Son. What is the level of volatility in instantaneous driving decisions? *Transportation Research Part C: Emerging Technologies,* Vol. 58, 2015, pp. 413-427.
4 . Rigas, G., Y. Goletsis, and D. I. Fotiadis. Real-time driver's stress event detection. *IEEE Transactions on intelligent transportation systems,* Vol. 13, No. 1, 2012, pp. 221-234.
5 . Healey, J. A., and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems,* Vol. 6, No. 2, 2005, pp. 156-166.
6 . Singh, R. R., S. Conjeti, and R. Banerjee. An approach for real-time stress-trend detection using physiological signals in wearable computing systems for automotive drivers.In *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, IEEE, 2011. pp. 1477-1482.
7 . Baek, H. J., H. B. Lee, J. S. Kim, J. M. Choi, K. K. Kim, and K. S. Park. Nonintrusive biological signal monitoring in a car to evaluate a driver's stress and health state. *Telemedicine and e-Health,* Vol. 15, No. 2, 2009, pp. 182-189.
8 . Sun, Y., and X. B. Yu. An innovative nonintrusive driver assistance system for vital signal monitoring. *IEEE journal of biomedical and health informatics,* Vol. 18, No. 6, 2014, pp. 1932-1939.
9 . Ji, Q., Z. Zhu, and P. Lan. Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE transactions on vehicular technology,* Vol. 53, No. 4, 2004, pp. 1052-1068.
10 . Lee, B.-G., and W.-Y. Chung. Driver alertness monitoring using fusion of facial features and bio-signals. *IEEE Sensors Journal,* Vol. 12, No. 7, 2012, pp. 2416-2422.
11 . Rigas, G., Y. Goletsis, P. Bougia, and D. I. Fotiadis. Towards driver's state recognition on real driving conditions. *International Journal of Vehicular Technology,* Vol. 2011, 2011.
12 . Tawari, A., and M. Trivedi. Speech based emotion classification framework for driver assistance system.In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, IEEE, 2010. pp. 174-178.
13 . Wang, J., and Y. Gong. Recognition of multiple drivers' emotional state.In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, IEEE, 2008. pp. 1-4.
14 . Jabon, M., J. Bailenson, E. Pontikakis, L. Takayama, and C. Nass. Facial expression analysis for predicting unsafe driving behavior. *IEEE Pervasive Computing,* Vol. 10, No. 4, 2011, pp. 84-95.

15 . Wu, M., S. Zhang, and Y. Dong. A novel model-based driving behavior recognition system using motion sensors. *Sensors,* Vol. 16, No. 10, 2016, p. 1746.
16 . Lee, B.-G., and W.-Y. Chung. A smartphone-based driver safety monitoring system using data fusion. *Sensors,* Vol. 12, No. 12, 2012, pp. 17536-17552.
17 . Lee, J.-D., J.-D. Li, L.-C. Liu, and C.-M. Chen. A novel driving pattern recognition and status monitoring system. *Advances in Image and Video Technology*, 2006, pp. 504-512.
18 . Wang, J., W. Xu, and Y. Gong. Real-time driving danger-level prediction. *Engineering Applications of Artificial Intelligence,* Vol. 23, No. 8, 2010, pp. 1247-1254.
19 . Figlewski, S. Forecasting volatility using historical data. 1994.
20 . Kamrani, M., R. Arvin, and A. J. Khattak. Extracting useful information from connected vehicle data: An empirical study of driving volatility measures and crash frequency at intersections. *Transportation Research Record: Journal of the Transportation Research Board (accepted forthcoming)*, 2018.
21 . DailyCare. *Introductory guide to identifying ECG irregularities*. http://www.dcbiomed.com/proimages/materials/EKG_Introduction/portableEKG_weitere_Informationen.pdf. Accessed 25/07, 2017.
22 . Medicine, S. S. o. *EKG Interpretive skills*. http://www.meddean.luc.edu/lumen/meded/medicine/skills/ekg/les1prnt.htm. Accessed 07/25, 2017.
23 . BIOPAC. *FNIR FUNCTIONAL NEAR INFRARED OPTICAL BRAIN IMAGING*. https://www.biopac.com/application/fnir-functional-near-infrared-optical-brain-imaging/#tabs. Accessed 07/25, 2017.
24 . Appelrath, H., D. Geesen, M. Grawunder, T. Michelsen, and D. Nicklas. Odysseus: a highly customizable framework for creating efficient event stream management systems.In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*, ACM, 2012. pp. 367-368.
25 . Oldenburg, U. o. *Odysseus - The Event Processing System*. http://odysseus.informatik.uni-oldenburg.de/. Accessed 07/24, 2017.
26 . Team, T. O. *Aggregation and Window*. https://wiki.odysseus.informatik.uni-oldenburg.de/display/ODYSSEUS/Aggregation+and+Window. Accessed 07/24, 2017.
27 . “Brian” Park, B., I. Yun, and K. Ahn. Stochastic optimization for sustainable traffic signal control. *International Journal of Sustainable Transportation,* Vol. 3, No. 4, 2009, pp. 263-284.
28 . Arvin, R., M. Kamrani, A. J. Khattak, and J. Rios-Torres. Safety Impacts of Automated Vehicles in Mixed Traffic. Presented at Transportation Research Board 97th Annual Meeting, Washington DC, 2018.
29 . Ghiasi, A., O. Hussain, Z. S. Qian, and X. Li. A mixed traffic capacity analysis and lane management model for connected automated vehicles: A Markov chain method. *Transportation Research Part B: Methodological,* Vol. 106, 2017, pp. 266-292.

30 . Henclewood, D. Safety Pilot Model Deployment–One Day Sample Data Environment Data Handbook. *Research and Technology Innovation Administration. Research and Technology Innovation Administration, US Department of Transportation, McLean, VA*, 2014.
31 . Zhang, M., and A. J. Khattak. Identifying and Analyzing Extreme Lane Change Events Using Basic Safety Messages in a Connected Vehicle Environment. Presented at Transportation Research Board 97th Annual Meeting, Washington DC, 2018.
32 . Kludt, K., J. Brown, J. Richman, and J. Campbell. Human factors literature reviews on intersections, speed management, pedestrians and bicyclists, and visibility.In, 2006.
33 . Arvin, R., M. Khademi, and H. Razi-Ardakani. Study on mobile phone use while driving in a sample of Iranian drivers. *International journal of injury control and safety promotion,* Vol. 24, No. 2, 2017, pp. 256-262.
34 . Wali, B., A. J. Khattak, and J. Xu. Contributory fault and level of personal injury to drivers involved in head-on collisions: application of copula-based bivariate ordinal models. *Accident Analysis & Prevention,* Vol. 110, 2018, pp. 101-114.
35 . Lajunen, T., D. Parker, and S. G. Stradling. Dimensions of driver anger, aggressive and highway code violations and their mediation by safety orientation in UK drivers. *Transportation Research Part F: Traffic Psychology and Behaviour,* Vol. 1, No. 2, 1998, pp. 107-121.
36 . Özkan, T., T. Lajunen, D. Parker, N. Sümer, and H. Summala. Symmetric relationship between self and others in aggressive driving across gender and countries. *Traffic injury prevention,* Vol. 11, No. 3, 2010, pp. 228-239.
37 . Papazikou, E., M. A. Quddus, and P. Thomas. Detecting deviation from normal driving using SHRP2 NDS data. 2017.
38 . Ghasemzadeh, A., and M. M. Ahmed. 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*, No. 2663, 2017, pp. 99-108.
39 . Ahmed, M. M., and A. Ghasemzadeh. Exploring the Impacts of Adverse Weather Conditions on Speed and Headway Behaviors Using the SHRP2 Naturalistic Driving Study Data. Presented at Transportation Research Board 96th Annual Meeting, Washington DC, 2017.
40 . Kim, E., and E. Choi. Estimates of critical values of aggressive acceleration from a viewpoint of fuel consumption and emissions.In *2013 Transportation Research Board Annual Meeting*, 2013.
41 . De Vlieger, I., D. De Keukeleere, and J. Kretzschmar. Environmental effects of driving behaviour and congestion related to passenger cars. *Atmospheric Environment,* Vol. 34, No. 27, 2000, pp. 4649-4655.
42 . Han, I., and K. Yang. Characteristic analysis for cognition of dangerous driving using automobile black boxes. *International journal of automotive technology,* Vol. 10, No. 5, 2009, pp. 597-605.

43 . Murphey, Y. L., R. Milton, and L. Kiliaris. Driver's style classification using jerk analysis.In *Computational Intelligence in Vehicles and Vehicular Systems, 2009. CIVVS'09. IEEE Workshop on*, IEEE, 2009. pp. 23-28.
44 . Bagdadi, O., and A. Várhelyi. Development of a method for detecting jerks in safety critical events. *Accident Analysis & Prevention,* Vol. 50, 2013, pp. 83-91.
45 . Feng, F., S. Bao, J. R. Sayer, C. Flannagan, M. Manser, and R. Wunderlich. Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data. *Accident Analysis & Prevention,* Vol. 104, 2017, pp. 125-136.
46 . Stewart, A. E., and J. H. Lord. Motor vehicle crash versus accident: a change in terminology is necessary. *Journal of traumatic stress,* Vol. 15, No. 4, 2002, pp. 333-335.
47 . Khattak, A. J., and B. Wali. Analysis of volatility in driving regimes extracted from basic safety messages transmitted between connected vehicles. *Transportation Research Part C: Emerging Technologies,* Vol. 84, 2017, pp. 48-73.
48 . Wali, B., A. J. Khattak, H. Bozdogan, and M. Kamrani. How is Driving Volatility Related to Intersection Safety? A Bayesian Heterogeneity-Based Analysis of Instrumented Vehicles Data. *Transportation Research Part C: Emerging Technologies*, 2018. Revised version submitted.
49 . Huber, P. J. *Robust statistics*. John Wiley & Sons, 2005.
50 . Bonett, D. G. Confidence interval for a coefficient of quartile variation. *Computational Statistics & Data Analysis,* Vol. 50, No. 11, 2006, pp. 2953-2957.
51 . Kamrani, M., A. J. Khattak, and T. Li. A Framework to Process and Analyze Driver, Vehicle and Road infrastructure Volatilities in Real-time. Presented at Transportation Research Board 97th Annual Meeting, Washington DC, 2018.
52 . Figlewski, S. Forecasting volatility using historical data. *Working Paper, New York University Stern School of Buisiness*, 1994.
53 . Kamrani, M., R. Arvin, and A. J. Khattak. Analyzing Highly Volatile Driving Trips Taken by Alternative Fuel Vehicles. Presented at Transporation Research Board 97th Annual Meeting, Washington DC, 2018.
54 . Anastasopoulos, P. C., and F. L. Mannering. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention,* Vol. 41, No. 1, 2009, pp. 153-159.
55 . Zhang, M., A. J. Khattak, and E. Shay. Analysis of Crashes Involving Pedestrians Across the United States: Implications for Connected and Automated Vehicles. Presented at Transportation Research Board 97th Annual Meeting, Washingtom DC, 2018.
56 . Azizi, L., and A. Sheikholeslami. Safety Effect of U-Turn Conversions in Tehran: Empirical Bayes Observational Before-and-After Study and Crash Prediction Models. *Journal of Transportation Engineering,* Vol. 139, No. 1, 2012, pp. 101-108.
57 . Greene, W. H. *Econometric analysis*. Pearson Education India, 2003.
58 . Washington, S. P., M. G. Karlaftis, and F. Mannering. *Statistical and econometric methods for transportation data analysis*. CRC press, 2010.

59 . Wali, B., A. Ahmed, and N. Ahmad. An ordered-probit analysis of enforcement of road speed limits.In *Proceedings of the Institution of Civil Engineers-Transport*, Thomas Telford Ltd, 2017. pp. 1-10.
60 . Li, X., A. J. Khattak, and B. Wali. Role of Multiagency Response and On-Scene Times in Large-Scale Traffic Incidents. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2616, 2017, pp. 39-48.
61 . Jamali, A., and Y. Wang. Estimating pedestrian exposure for small urban and rural areas. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2661, 2017, pp. 84-94.
62 . Wang, Y., K. Kockelman, and A. Jamali. A Synthesis of Spatial Models for Multivariate Count Responses.In *Regional Research Frontiers-Vol. 2*, Springer, 2017. pp. 221-237.
63 . Wali, B., A. J. Khattak, J. Waters, D. Chimba, and X. Li. Development of Safety Performance Functions: Incorporating Unobserved Heterogeneity and Functional Form Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. Accepted.
64 . Bhat, C. R. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B: Methodological,* Vol. 37, No. 9, 2003, pp. 837-855.
65 . Train, K. Halton sequences for mixed logit. *Department of Economics, UCB*, 2000.
66 . Poirier, J., F. Ludwig, and S. Motamedi. Traffic Light Assistant Simulation: Foggy Weather.In *Advances in Human Aspects of Transportation*, Springer, 2017. pp. 1129-1138.
67 . Abdel-Aty, M., and K. Haleem. Analyzing angle crashes at unsignalized intersections using machine learning techniques. *Accident Analysis & Prevention,* Vol. 43, No. 1, 2011, pp. 461-470.
68 . Persaud, B., and T. Nguyen. Disaggregate safety performance models for signalized intersections on Ontario provincial roads. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1635, 1998, pp. 113-120.
69 . Kamrani, M., S. M. H. E. Abadi, and S. R. Golroudbary. Traffic simulation of two adjacent unsignalized T-junctions during rush hours using Arena software. *Simulation Modelling Practice and Theory,* Vol. 49, 2014, pp. 167-179. https://doi.org/10.1016/j.simpat.2014.09.006.
70 . Khattak, A., S. Nambisan, and S. Chakraborty. Study of Driving VOlatility in Connected and Cooperative Vehicle Systems. National Science Foundation.In, 2015.
71 . Henclewood, D. *Safety Pilot Model Deployment – One Day Sample Data Environment Data Handbook*, 2014.
72 . Osman, O. A., and S. Ishak. A network level connectivity robustness measure for connected vehicle environments. *Transportation Research Part C: Emerging Technologies,* Vol. 53, 2015, pp. 48-58.

73 . Ghiasi, A., J. Ma, F. Zhou, and X. Li. Speed Harmonization Algorithm Using Connected Autonomous Vehicles. Presented at Transportation Research Board, 2017.
74 . Li, X., A. Ghiasi, and Z. Xu. Exact Method for a Simplified Trajectory Smoothing Problem with Connected Automated Vehicles. Presented at Transportation Research Board, 2017.
75 . Bergenhem, C., S. Shladover, E. Coelingh, C. Englund, and S. Tsugawa. Overview of platooning systems.In *Proceedings of the 19th ITS World Congress, Oct 22-26, Vienna, Austria (2012)*, 2012.
76 . NHTSA. *Resource Guide Describes Best Practices For Aggressive Driving Enforcement* National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, DC. http://www.nhtsa.gov/About+NHTSA/Traffic+Techs/current/Resource+Guide+Describes+Best+Practices+For+Aggressive+Driving+Enforcement. Accessed June 22nd, 2016.
77 . Haglund, M., and L. Åberg. Speed choice in relation to speed limit and influences from other drivers. *Transportation Research Part F: Traffic Psychology and Behaviour,* Vol. 3, No. 1, 2000, pp. 39-51.
78 . Wali, B., A. Ahmed, S. Iqbal, and A. Hussain. Effectiveness of enforcement level of speed limit and drink driving laws and associated factors – Exploratory empirical analysis using a bivariate ordered probit model. *Journal of Traffic and Transportation Engineering (English Edition)*, 2017 (forthcoming).
79 . Paleti, R., N. Eluru, and C. R. Bhat. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accident Analysis & Prevention,* Vol. 42, No. 6, 2010, pp. 1839-1854.
80 . Chrysler, S. T., J. M. Cooper, and D. Marshall. The Cost of Warning of Unseen Threats: Unintended Consequences of Connected Vehicle Alerts.In *Transportation Research Board 94th Annual Meeting*, 2015.
81 . Genders, W., and S. N. Razavi. Impact of connected vehicle on work zone network safety through dynamic route guidance. *Journal of Computing in Civil Engineering,* Vol. 30, No. 2, 2015, p. 04015020.
82 . Du, L., and H. Dao. Information Dissemination Delay in Vehicle-to-Vehicle Communication Networks in a Traffic Stream. *Intelligent Transportation Systems, IEEE Transactions on,* Vol. 16, No. 1, 2015, pp. 66-80.
83 . Wang, X., A. Khattak, J. Liu, G. Masghati-Amoli, and S. Son. What is the Level of Volatility in Instantaneous Driving Decisions? *Transportation Research Part C: Emerging Technologies*, 2015. 10.1016/j.trc.2014.12.014.
84 . Liu, J., X. Wang, and A. Khattak. Generating Real-Time Driving Volatility Information. Presented at 2014 World Congress on Intelligent Transport Systems, Detroit, MI, 2014.
85 . Liu, J., and A. Khattak. Delivering Improved Alerts, Warnings, and Control Assistance Using Basic Safety Messages Transmited between Connected Vehicles. . *Transportation Research Part C: Emerging Technologies*, 2016.

86 . Lord, D., and F. Mannering. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice,* Vol. 44, No. 5, 2010, pp. 291-305.
87 . Li, X., A. J. Khattak, and B. Wali. Large-Scale Traffic Incident Duration Analysis: The Role of Multi-agency Response and On-Scene Times. *Transportation Research Record: Journal of the Transportation Research Board*, 2017 (forthcoming).
88 . Khattak, A. J., J. Liu, B. Wali, X. Li, and M. Ng. Modeling Traffic Incident Duration Using Quantile Regression. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2554, 2016, pp. 139-148.
89 . Henclewood, D. Safety Pilot Model Deployment – One Day Sample Data Environment Data Handbook.In, Research and Technology Innovation Administration, US Department of Transportation, McLean, VA, 2014.
90 . Abdel-Aty, M., and J. Keller. Exploring the overall and specific crash severity levels at signalized intersections. *Accident Analysis & Prevention,* Vol. 37, No. 3, 2005, pp. 417-425.
91 . Schneider, R. J., R. M. Ryznar, and A. J. Khattak. An accident waiting to happen: a spatial approach to proactive pedestrian planning. *Accident Analysis & Prevention,* Vol. 36, No. 2, 2004, pp. 193-211.
92 . Mannering, F. L., and C. R. Bhat. Analytic methods in accident research: methodological frontier and future directions. *Analytic Methods in Accident Research,* Vol. 1, 2014, pp. 1-22.
93 . Ye, X., R. M. Pendyala, S. P. Washington, K. Konduri, and J. Oh. A simultaneous equations model of crash frequency by collision type for rural intersections. *Safety Science,* Vol. 47, No. 3, 2009, pp. 443-452.
94 . Guzzella, L., and A. Sciarretta. *Vehicle propulsion systems*. Springer, 2007.
95 . Kiencke, U., and L. Nielsen. Vehicle modelling. *Automotive Control Systems: For Engine, Driveline, and Vehicle*, 2005, pp. 301-349.
96 . Burnham, G., J. Seo, and G. Bekey. Identification of human driver models in car following. *IEEE transactions on Automatic Control,* Vol. 19, No. 6, 1974, pp. 911-915.
97 . Prokop, G. Modeling human vehicle driving by model predictive online optimization. *Vehicle System Dynamics,* Vol. 35, No. 1, 2001, pp. 19-53.
98 . Macadam, C. C. Understanding and modeling the human driver. *Vehicle System Dynamics,* Vol. 40, No. 1-3, 2003, pp. 101-134.
99 . Liu, A., and A. Pentland. Towards real-time recognition of driver intentions.In *Intelligent Transportation System, 1997. ITSC'97., IEEE Conference on*, IEEE, 1997. pp. 236-241.
100 . Cooper, R. A. System identification of human performance models. *IEEE transactions on systems, man, and cybernetics,* Vol. 21, No. 1, 1991, pp. 244-252.
101 . Kiencke, U., R. Majjad, and S. Kramer. Modeling and performance analysis of a hybrid driver model. *Control Engineering Practice,* Vol. 7, No. 8, 1999, pp. 985-991.

102 . Toledo, T., H. N. Koutsopoulos, and M. Ben-Akiva. Integrated driving behavior modeling. *Transportation Research Part C: Emerging Technologies,* Vol. 15, No. 2, 2007, pp. 96-112.
103 . Brackstone, M., and M. McDonald. Car-following: a historical review. *Transportation Research Part F: Traffic Psychology and Behaviour,* Vol. 2, No. 4, 1999, pp. 181-196.
104 . Rothery, R. W. Car following models. *Trac Flow Theory*, 1992.
105 . Gipps, P. G. A behavioural car-following model for computer simulation. *Transportation Research Part B: Methodological,* Vol. 15, No. 2, 1981, pp. 105-111.
106 . Yang, Q., and H. N. Koutsopoulos. A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research Part C: Emerging Technologies,* Vol. 4, No. 3, 1996, pp. 113-129.
107 . Gipps, P. G. A model for the structure of lane-changing decisions. *Transportation Research Part B: Methodological,* Vol. 20, No. 5, 1986, pp. 403-414.
108 . Salvucci, D., E. Boer, and A. Liu. Toward an integrated model of driver behavior in cognitive architecture. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1779, 2001, pp. 9-16.
109 . Li, L., X. M. Chen, and L. Zhang. A global optimization algorithm for trajectory data based car-following model calibration. *Transportation Research Part C: Emerging Technologies,* Vol. 68, 2016, pp. 311-332.
110 . Ossen, S., and S. P. Hoogendoorn. Car-following behavior analysis from microscopic trajectory data. *Transportation Research Record,* Vol. 1934, No. 1, 2005, pp. 13-21.
111 . ---. Heterogeneity in car-following behavior: Theory and empirics. *Transportation Research Part C: Emerging Technologies,* Vol. 19, No. 2, 2011, pp. 182-195.
112 . Farah, H., and H. N. Koutsopoulos. Do cooperative systems make drivers' car-following behavior safer? *Transportation Research Part C: Emerging Technologies,* Vol. 41, 2014, pp. 61-72.
113 . Hoogendoorn, R., B. van Arem, and S. Hoogendoorn. Incorporating driver distraction in car-following models: Applying the TCI to the IDM.In *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*, IEEE, 2013. pp. 2274-2279.
114 . Koutsopoulos, H. N., and H. Farah. Latent class model for car following behavior. *Transportation Research Part B: Methodological,* Vol. 46, No. 5, 2012, pp. 563-578.
115 . Papathanasopoulou, V., and C. Antoniou. Towards data-driven car-following models. *Transportation Research Part C: Emerging Technologies,* Vol. 55, 2015, pp. 496-509.
116 . Hoogendoorn, S., R. G. Hoogendoorn, and W. Daamen. Wiedemann revisited: new trajectory filtering technique and its implications for car-following modeling. *Transportation Research Record,* Vol. 2260, No. 1, 2011, pp. 152-162.
117 . Ossen, S., S. P. Hoogendoorn, and B. G. Gorte. Interdriver Differences in Car-Following: A Vehicle Trajectory–Based Study. *Transportation Research Record,* Vol. 1965, No. 1, 2006, pp. 121-129.

118 . Choudhury, C., V. Ramanujam, and M. Ben-Akiva. Modeling acceleration decisions for freeway merges. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2124, 2009, pp. 45-57.
119 . Shimosaka, M., K. Nishi, J. Sato, and H. Kataoka. Predicting driving behavior using inverse reinforcement learning with multiple reward functions towards environmental diversity.In *Intelligent Vehicles Symposium (IV), 2015 IEEE*, IEEE, 2015. pp. 567-572.
120 . Kuderer, M., S. Gulati, and W. Burgard. Learning driving styles for autonomous vehicles from demonstration.In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, IEEE, 2015. pp. 2641-2646.
121 . Hayeri, Y. M., K.-E. Kim, and D. Lee. An Inverse Reinforcement Learning Approach to Car Following Behaviors.In, 2016.
122 . Bellman, R. The theory of dynamic programming.In, DTIC Document, 1954.
123 . Sutton, R. S., and A. G. Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.
124 . Kamrani, M., R. Arvin, and A. J. Khattak. Extracting useful information from Basic Safety Message Data: an empirical study of driving volatility measures and crash frequency at intersections. *Transportation Research Record*, 2018. http://dx.doi.org/10.1177/0361198118773869.
125 . Train, K. *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*. MIT press, 1986.
126 . Train, K. E., and C. Winston. Vehicle choice behavior and the declining market share of US automakers. *International economic review,* Vol. 48, No. 4, 2007, pp. 1469-1496.
127 . Train, K. A validation test of a disaggregate mode choice model. *Transportation Research,* Vol. 12, No. 3, 1978, pp. 167-174.
128 . Ramming, M. S. Network knowledge and route choice. *Unpublished Ph. D. Thesis, Massachusetts Institute of Technology*, 2001.
129 . Henclewood, D. Safety Pilot Model Deployment – One Day Sample Data Environment Data Handbook. . *Research and Technology Innovation Administration, US Department of Transportation: McLean, VA*, 2014.
130 . Hossain, M., and Y. Muromachi. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention,* Vol. 45, 2012, pp. 373-381.
131 . Sun, J., and J. Sun. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies,* Vol. 54, 2015, pp. 176-186.
132 . Wang, L., M. Abdel-Aty, Q. Shi, and J. Park. Real-time crash prediction for expressway weaving segments. *Transportation Research Part C: Emerging Technologies,* Vol. 61, 2015, pp. 1-10.
133 . Xu, C., A. P. Tarko, W. Wang, and P. Liu. Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention,* Vol. 57, 2013, pp. 30-39.
134 . Yu, R., and M. Abdel-Aty. Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention,* Vol. 51, 2013, pp. 252-259.

135 . Ahmed, M. M., and M. A. Abdel-Aty. The viability of using automatic vehicle identification data for real-time crash prediction. *IEEE Transactions on intelligent transportation systems,* Vol. 13, No. 2, 2012, pp. 459-468.
136 . Abdel-Aty, M. A., and R. Pemmanaboina. Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. *IEEE Transactions on intelligent transportation systems,* Vol. 7, No. 2, 2006, pp. 167-174.
137 . Xu, C., W. Wang, and P. Liu. A genetic programming model for real-time crash prediction on freeways. *IEEE Transactions on intelligent transportation systems,* Vol. 14, No. 2, 2013, pp. 574-586.
138 . Everitt, B., and A. Skrondal. *The Cambridge dictionary of statistics*. Cambridge University Press Cambridge, 2002.
139 . Zwillinger, D., and S. Kokoska. Standard Probability and Statistical Tables and Formula.In, Chapman & Hall, Boca Raton, 2000.
140 . James, G., D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Springer, 2013.
141 . Breiman, L. Bagging predictors. *Machine learning,* Vol. 24, No. 2, 1996, pp. 123-140.
142 . Ho, T. K. Random decision forests.In *Document analysis and recognition, 1995., proceedings of the third international conference on, No. 1*, IEEE, 1995. pp. 278-282.
143 . Schapire, R. E. The strength of weak learnability. *Machine learning,* Vol. 5, No. 2, 1990, pp. 197-227.
144 . Washington, S. P., M. G. Karlaftis, and F. L. Mannering. Statistical and econometric methods for transportation data analysis. *Chapman Hall/CRC, Boca Raton, FL*, 2003.
145 . Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, pp. 267-288.
146 . Holland, J. H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.
147 . Mitchell, M. *An introduction to genetic algorithms*. MIT press, 1998.
148 . Bozdogan, H. Akaike's information criterion and recent developments in information complexity. *Journal of mathematical psychology,* Vol. 44, No. 1, 2000, pp. 62-91.
149 . ---. Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix.In *Information and Classification: Concepts, Methods and Applications Proceedings of the 16th Annual Conference of the "Gesellschaft für Klassifikation e.V." University of Dortmund, April 1–3, 1992*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1993. pp. 40-54.
150 . Akaike, H. Information theory and an extension of the maximum likelihood principle.In *Selected papers of hirotugu akaike*, Springer, 1998. pp. 199-213.
151 . Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika,* Vol. 52, No. 3, 1987, pp. 345-370.

152 . ---. ICOMP: A new model-selection criterion.In *1. Conference of the International Federation of Classification Societies*, 1987. pp. 599-608.
153 . Bozdogan, H. The theory and applications of information-theoretic measure of complexity (ICOMP) as a new model selection criterion. *Unpublished research report, the Institute of Statistical mathematics, Tokyo, Japan, and the Department of mathematics, University of Virginia, Charlottesville, VA*, 1988.
154 . Bozdogan, H. On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics-Theory and Methods,* Vol. 19, No. 1, 1990, pp. 221-278.
155 . ---. Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity.In *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach*, Springer, 1994. pp. 69-113.
156 . Bozdogan, H. A new informational complexity criterion for model selection: the general theory and its applications. *Invited paper presented in the Session on Information Theoretic Models,* Vol. 6, 1996.
157 . ---. Informational complexity and multivariate statistical modeling.In, manuscript, 1998.
158 . Werbos, P. Beyond regression: new fools for prediction and analysis in the behavioral sciences. *PhD thesis, Harvard University*, 1974.
159 . ---. Beyond regression: New tools for prediction and analysis in the behavior science. *Unpublished Doctoral Dissertation, Harvard University*, 1974.
160 . TRB. *The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset*. https://insight.shrp2nds.us/.
161 . Hankey, J. M., M. A. Perez, and J. A. McClafferty. Description of the SHRP 2 Naturalistic Database and the Crash, Near-Crash, and Baseline Data Sets.In, Virginia Tech Transportation Institute, 2016.
162 . Kamrani, M., R. Arvin, and A. J. Khattak. Extracting Useful Information from Basic Safety Message Data: An Empirical Study of Driving Volatility Measures and Crash Frequency at Intersections. *Transportation Research Record,* Vol. 0, No. 0, 2018. http://dx.doi.org/10.1177/0361198118773869.
163 . Shaheed, M. S., and K. Gkritza. A latent class analysis of single-vehicle motorcycle crash severity outcomes. *Analytic Methods in Accident Research,* Vol. 2, 2014, pp. 30-38.
164 . Chang, H.-L., and T.-H. Yeh. Risk factors to driver fatalities in single-vehicle crashes: comparisons between non-motorcycle drivers and motorcyclists. *Journal of Transportation Engineering,* Vol. 132, No. 3, 2006, pp. 227-236.
165 . Zhang, T., and A. H. Chan. The association between driving anger and driving outcomes: A meta-analysis of evidence from the past twenty years. *Accident Analysis & Prevention,* Vol. 90, 2016, pp. 50-62.
166 . Choudhary, P., and N. R. Velaga. Mobile phone use during driving: Effects on speed and effectiveness of driver compensatory behaviour. *Accident Analysis & Prevention,* Vol. 106, 2017, pp. 370-378.

167 . Márquez, L., V. Cantillo, and J. Arellana. Mobile phone use while driving: A hybrid modeling approach. *Accident Analysis & Prevention,* Vol. 78, 2015, pp. 73-80.
168 . Dozza, M., C. A. Flannagan, and J. R. Sayer. Real-world effects of using a phone while driving on lateral and longitudinal control of vehicles. *Journal of safety research,* Vol. 55, 2015, pp. 81-87.
169 . Choudhary, P., and N. R. Velaga. Modelling driver distraction effects due to mobile phone use on reaction time. *Transportation Research Part C: Emerging Technologies,* Vol. 77, 2017, pp. 351-365.
170 . Dingus, T. A., F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences,* Vol. 113, No. 10, 2016, pp. 2636-2641.
171 . Dingus, T. A. Estimates of prevalence and risk associated with inattention and distraction based upon in situ naturalistic data. *Annals of advances in automotive medicine,* Vol. 58, 2014, p. 60.
172 . Neyens, D. M., and L. N. Boyle. The effect of distractions on the crash types of teenage drivers. *Accident Analysis & Prevention,* Vol. 39, No. 1, 2007, pp. 206-212.
173 . Klauer, S. G., T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. 2006.
174 . Schneidereit, T., T. Petzoldt, A. Keinath, and J. F. Krems. Using SHRP 2 naturalistic driving data to assess drivers' speed choice while being engaged in different secondary tasks. *Journal of safety research,* Vol. 62, 2017, pp. 33-42.
175 . Ye, M., O. A. Osman, S. Ishak, and B. Hashemi. Detection of driver engagement in secondary tasks from observed naturalistic driving behavior. *Accident Analysis & Prevention,* Vol. 106, 2017, pp. 385-391.
176 . Wali, B., A. J. Khattak, H. Bozdogan, and M. Kamrani. How is driving volatility related to intersection safety? A Bayesian heterogeneity-based analysis of instrumented vehicles data. *Transportation Research Part C: Emerging Technologies,* Vol. 92, 2018, pp. 504-524.
177 . Train, K. E. *Discrete choice methods with simulation*. Cambridge university press, 2009.
178 . Tomarken, A. J., and N. G. Waller. Structural equation modeling: Strengths, limitations, and misconceptions. *Annu. Rev. Clin. Psychol.,* Vol. 1, 2005, pp. 31-65.

VITA

Mohsen Kamrani earned his bachelor's degree in Industrial Engineering from Mazandaran University of Science & Technology, Iran in 2006. After obtaining his bachelor's, he worked for approximately three years in the industry and then decided to continue his education. He moved to Malaysia and received his master's degree in Industrial Engineering from Universti Teknologi Malaysia where he was awarded the 2014 Best Industrial Engineering Student Award. In 2015, Mohsen was admitted to the Civil Engineering doctorate program with the concentration of Transportation Engineering at the University of Tennessee, Knoxville under Professor Asad Khattak's mentorship. During his Ph.D., he worked on a National Science Foundation (NSF) granted project titled "Driving Volatility in a Connected and Cooperative Vehicle Environment: Algorithms for Driver Warnings and Control Assists," which is very competitive grant awarded nationally by the Civil and Infrastructure Systems division of NSF. He also served as a graduate teaching assistant in surveying, transportation safety, and intelligent transportation safety courses. He published multiple journal and peer-reviewed conference papers in the area of transportation safety, connected and automated vehicles and driving behavior and contributed to multiple research proposals. He actively participated conferences; in particular, the Transportation Research Board (TRB) where he received the 2017 Outstanding Paper Award from TRB Data, Analysis and Evaluation Committee for the paper titled "Can Data Generated by Connected Vehicles Enhance Safety? A Proactive Approach to Intersection Safety Management." He also managed to earn his Master's degree in statistics from the University of Tennessee concurrently with his Ph.D. in December 2018.