BIG DATA GENERATED BY CONNECTED AND AUTOMATED VEHICLES FOR SAFETY MONITORING, ASSESSMENT AND IMPROVEMENT

FINAL REPORT, YEAR 2

SOUTHEASTERN TRANSPORTATION CENTER

Asad Khattak, Ph.D.
Jun Liu, Ph.D.
Behram Wali
Mohsen Kamrani
Meng Zhang

SEPTEMBER 2016

US DEPARTMENT OF TRANSPORTATION GRANT DTRT13-G-UTC34

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

ACKNOWLEDGEMENT

Data used for this project are from several sources, including the Research Data Exchange program maintained by Federal Highway Administration, National Highway Traffic Safety Administration, California and Georgia Departments of Transportation, Census, Google Earth, and the National Renewable Energy Laboratory. Software packages R, MATLAB and Google Earth were used for the data processing and visualization. Statistical software STATA was used for modeling. The research was supported through the Southeastern Transportation Center, sponsored by the United States Department of Transportation through grant number DTRT13-G-UTC34. Special thanks are extended to the following entities for their support: Transportation Engineering & Science Program, Initiative for Sustainable Mobility, and colleagues from University of Kentucky and University of Central Florida who also worked on the Big Data major research initiative of the Southeastern Transportation Center. The support from Dr. S. Nambisan, L. Han, and C. Cherry at University of Tennessee is gratefully acknowledged. The views expressed in the report are those of the authors, who are responsible for the facts and accuracy of information presented herein.

Form DOT F 1700.7 (8-72)

1. Report No. STC-2016-M4.UTK	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Big Data Generated by Connected and Automated Vehicles for Safety Monitoring, Assessment and Improvement		5. Report Date September 2016
		6. Source Organization Code N/A
7. Author(s) Khattak, Asad; Liu, Jun; Wali, Behram; Kamrani, Mohsen; Zhang, Meng		8. Source Organization Report No. STC-2016-M4.UTK
9. Performing Organization Name and Address Southeastern Transportation Center 309 Conference Center Building Knoxville, Tennessee 37996-4133 865.974.5255		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. DTRT12-G-UTC34
12. Sponsoring Agency Name and Address US Department of Transportation Office of the Secretary of Transportation Research 1200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered Final Report: August 2015 September 2016
		14. Sponsoring Agency Code USDOT/OST-R
15. Supplementary Notes: None		

16. Abstract

Increasing amounts of information generated by electronic sensors from various sources that include travelers, vehicles, infrastructure and the environment coupled with social, economic and spatial data, collectively referred to as "Big Data," represent an opportunity for innovation. The opportunities span across transportation system planning, design, operation and maintenance. The key objectives of this project are to: 1) Generate new frameworks for acquisition and use of Big Data to facilitate safety monitoring, assessment and improvement; 2) Visualize and analyze Big Data and develop tools/products that can be used (e.g., in transportation management centers) to improve safety; and 3) Take advantage of opportunities arising from Big Data to create safety products/tools and create new knowledge. In this report, we summarize efforts undertaken to extract, process and integrate data from multiple sources in order to and generate driver feedback based on vehicle-to-vehicle and vehicle-to-infrastructure communication data. The specific objectives pursued include examining instantaneous driving decisions and trip-level driving volatility at a microscopic level, generation of alerts and warnings from connected vehicle data, and analyzing location-specific volatility for developing and demonstrating a proactive safety methodology.

17. Key Words Big data, connected and autonomous vehicles, intelligent transportation systems, safety monitoring, driver feedback, V2V, V2I	18. Distribution Statement Unrestricted

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	20. No. of Pages 39	20. Price N/A

TABLE OF CONTENTS

LIST OF FIGURES iv
LIST OF TABLES iv
EXECUTIVE SUMMARY v
1. INTRODUCTION 1
1.1. ACTIVITIES 1
1.2. OBJECTIVES 2
1.3. SUMMARY OF RESULTS 2
2. CAN DATA GENERATED BY CONNECTED VEHICLES ENHANCE SAFETY? A PROACTIVE APPROACH TO INTERSECTION SAFETY MANAGEMENT 7
2.1. INTRODUCTION 8
2.2. LITERATURE REVIEW 9
2.2.1. Research Objective and Contribution 10
2.3. METHODOLOGY 10
2.3.1. Conceptual Framework 10
2.3.2. Location Based Volatility 12
2.3.3. Calculation of LBV 12
2.3.4. Modeling Approach 13
2.4. DATA 15
2.5. RESULTS 17
2.5.1. Descriptive Statistics 17
2.5.2. Modeling Results 19
2.6. DISCUSSION 23
2.7. LIMITATIONS 24
2.8. CONCLUSIONS 24
3. DELIVERING IMPROVED ALERTS, WARNINGS, AND CONTROL ASSISTANCE USING BASIC SAFETY MESSAGES 26
4. DYNAMICS OF DRIVING REGIMES EXTRACTED FROM BASIC SAFETY MESSAGES 27
5. IDENTIFYING AND ANALYZING EXTREME LANE CHANGE EVENTS USING BASIC SAFETY MESSAGES 28
6. REFERENCES 29

LIST OF FIGURES

Figure 1 Overall framework for study vi

Figure 2 Transition probabilities and durations of acceleration/deceleration. 3

Figure 3 Alerts/warnings generated during a trip. 3

Figure 4 Plots of extreme acceleration events (hallowed in side) 4

Figure 5 Generating warnings and control assists. 4

Figure 6 Methodology and key results for location-based volatility study. 5

Figure 7 Extreme Lane change events 6

Figure 8 Ann Arbor map created from BSM data 17

Figure 9 Mean-expected over actual number of crashes for fixed and random-parameter Poisson models (Green: fixed parameter models; Red: random parameter models) 22

LIST OF TABLES

Table 1 Description of key variables and Descriptive Statistics 16

Table 2 Description of Key Variables and Descriptive Statistics 18

Table 3 Modeling Results of Fixed- and Random-Parameter Poisson Regressions 21

EXECUTIVE SUMMARY

With driverless cars attracting widespread attention, and wireless exchange of safety critical and operational data between vehicles and the infrastructure, an important advance is the ability for vehicles to communicate their spatial coordinates, heading, speed, and acceleration data within a spatio-temporal proximal envelope. Recent developments can leverage a tremendous opportunity to utilize information becoming available from the nascent and emergent cyber-physical systems in the automobile-roadway operating realm. The activities undertaken in this project are transformative because they enable preemptive warnings and assists to drivers based on anticipated driving maneuvers that are potentially unsafe. The study is advancing knowledge of instantaneous driver decisions that lead to undesirable outcomes, and it is exploring innovative ways to avoid anomalous situations and behaviors.

A key idea behind pursuing activities undertaken is to understand (and where possible reduce) "driving volatility" in instantaneous driving decisions and increase driving and locational stability. Consequently, driving volatility was operationalized at both the driver level, when drivers undertake trips, and at the network level. Driving volatility helps us understand instantaneous driving decisions in a connected vehicle environment. The project team has expanded knowledge about instantaneous driving decisions by making conceptual, methodological, and empirical contributions. Methodologically, we introduce the applications of Dynamic Markov Switching models, and hierarchical models to large-scale CAV data. Such techniques are contributing to the scientific analysis of real-world connected vehicle data, and extracting actionable information embedded in such data.

Initially the concept of driving volatility was operationalized at the trip level as the percentage of time when a driver's acceleration or braking is more than the mean plus one or two standard deviations above the accelerations or braking of a large sample at that speed. In terms of developing new ideas, we have worked on driving volatility by understanding acceleration and braking regimes in instantaneous trip-level driving decisions. A database from a Connected Vehicles (CV) deployment initiative in Michigan was used to empirically explore regimes using Markov switching models and more broadly the microscopic decisions made by drivers. Questions answered include when these regimes change and how to quantify the volatility associated with each regime. Another new idea is extending volatility to specific locations in a network, termed "location-based volatility." For location-based volatility, a unique database was assembled by integrating CAV data with crash and inventory data at intersections. This enabled associating location-based volatility with actual crashes, fostering new insights.

The new ideas are to enhance development of tools and algorithms that address research needs in connected and automated vehicle space (Figure 1). In particular, by developing and applying new data analytic techniques, the study enables new discoveries and generation of actionable information for innovative real-life applications. Ultimately, the knowledge generated can provide a strong behavioral basis for CAV applications including hazard notification and real-time or predictive driver feedback. Overall, the foundational methods for extracting useful information from CAV and related data are developed that will enhance safety, fuel efficiency and reduce emissions.

In this study, the dynamics of driving regimes extracted from Basic Safety Messages (BSMs) transmitted between connected vehicles are analyzed at a microscopic level. We generate new knowledge for connected vehicle technologies by explicitly investigating time-series instantaneous driving decisions of connected vehicle drivers, and mapping such decisions to instantaneous driving contexts. This analysis is important because driving decisions primarily depend on surrounding traffic states, and a detailed understanding of driving decisions can significantly help us with better anticipating hazardous situations and providing warnings and alerts to drivers. By introducing application of Markov-Switching models to large-scale CAV data, we characterize typical driving cycle into different regimes. The exhaustive analysis generates meaningful information about the existence of different regimes, when do the regimes change and how long they last, and key correlates associated with each regime. Finally, the study provides a new foundational framework for making short-term instantaneous driving and regime predictions at specific instances in time—critical for developing connected vehicle hazard anticipation and notification systems.

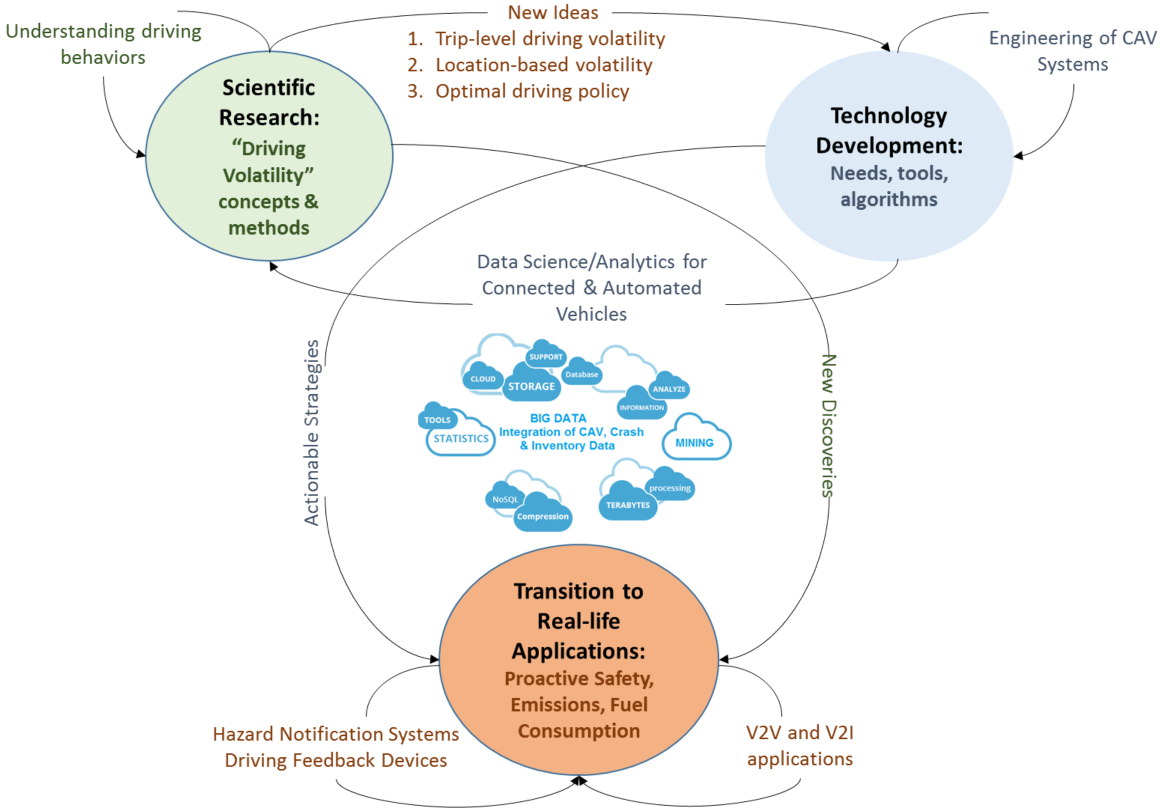

Understanding driving behaviors
New Ideas
1. Trip-level driving volatility
2. Location-based volatility
3. Optimal driving policy
Engineering of CAV Systems
Scientific Research:
"Driving Volatility" concepts & methods
Technology Development:
Needs, tools, algorithms
Data Science/Analytics for Connected & Automated Vehicles
Actionable Strategies
New Discoveries
SUPPORT
CLOUD
STORAGE
Database
ANALYZE
INFORMATION
TOOLS
STATISTICS
BIG DATA
Integration of CAV, Crash & Inventory Data
MINING
NoSQL
Compression
TERABYTES
processing
Transition to Real-life Applications:
Proactive Safety, Emissions, Fuel Consumption
Hazard Notification Systems
Driving Feedback Devices
V2V and V2I applications

Figure 1 Overall framework for study

A key idea explored in this study is to establish context-relevant alert, warning, and control assist thresholds based on extreme event information. A novel analytical methodology introduces speed varying thresholds to account for different driving situations and contexts. The study's innovative approach interprets high-frequency geo-referenced connected vehicle data, and extracts, through data analytic techniques, critical information about extreme events from connected vehicle data.

As a proactive safety measure and a leading indicator of safety, we expand the concept of volatility to specific locations by quantifying variability in instantaneous driving decisions at intersections. Traditionally, evaluation of intersection safety has been largely reactive, based on historical crash data. The conceptual framework developed uses emerging CAV data and historical crash data to proactively identify intersections with high levels of variability in instantaneous driving behaviors. The empirical contributions include assembling a unique database that integrates intersection crash and inventory data with more than 65 million real-world BSMs transmitted between connected vehicles and roadside units. This unique empirical database facilitates new and rigorous exploratory analyses. Additionally, the techniques used to extract useful information from integrated CAV, crash, and inventory data provide useful knowledge for enhancing proactive intersection safety.

It is important to note that a complementary project was funded by the US National Science Foundation, titled "Study of Driving Volatility in Connected and Cooperative Vehicle Systems" by the Division of Civil, Mechanical, & Manufacturing Innovation," Award Number: 1538139. The NSF project is highly complementary to this DOT research project and involves the development of analytic procedures for understanding driving volatility, and the use of data for generating information and driver feedback. The project is developing computationally efficient algorithms for predicting driver actions and volatility using information about their prior behaviors combined with positions and motions obtained via wireless communications. A Markov Decision Process framework is being developed to anticipate instantaneous driver maneuver decisions. Driver-specific estimates of rewards and penalties for available maneuver choices in driving situations are will be learned using Bayesian Inverse Reinforcement Learning and gossip algorithms. There is cross-fertilization between the NSF and DOT projects. For reporting purposes, we have assigned percentages to indicate the contribution of the project to the preparation and publication of the paper.

Overall, by studying driving volatility from different perspectives, the team is creating new analytical frameworks to extract useful information from raw CAV data with the purpose of enhancing safety and fuel economy. During the reporting period, the above mentioned activities led to the publication of refereed journal papers, and preparation several full-length research papers to be submitted to transportation research conferences and peer reviewed journals. Refereed journal publications include:

- Liu J. & A. Khattak. Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles, *Transportation Research*, Part C, Volume 68, pp. 83–100, 2016. (80%)
- Liu J. A. Khattak & X. Wang, A Comparative Study of Driving Performance in Metropolitan Regions Using Large-scale Vehicle Trajectory Data: Implications for Sustainable Cities, Forthcoming in *International Journal of Sustainable Transportation*, 2017. (50%)

Papers prepared for submission to transportation journals and/or conferences include:

- Liu, J., A. Khattak, & M. Zhang, Structuring and Integrating Data in Metropolitan Regions to Explore Multilevel Links Between Driving Volatility and Correlates. To be submitted to a conference for presentation review and a journal for publication review. (60%)
- Kamrani, M. A. Khattak, & B. Wali, Can Data Generated by Connected Vehicles Enhance Safety? A proactive approach to intersection safety management. To be submitted to a conference for presentation review and a journal for publication review. (50%)
- Khattak A., & B. Wali, Dynamics of Driving Regimes Extracted from Basic Safety Messages Transmitted Between Connected Vehicles. To be submitted to a conference for presentation review and a journal for publication review. (40%)
- Zhang, M. & A. Khattak, Identifying and Analyzing Extreme Lane Change Events Using Basic Safety Messages in a Connected Vehicle Environment. To be submitted to a conference for presentation review and a journal for publication review. (50%)

Presentations and Talks given during the reporting period include:

- Khattak A. Study of Micro-Driving Behaviors at Different Levels of Vehicle Automation and Connectivity, Plenary Session Invited Talk, 16th COTA International Conference of Transportation Professionals, CICTP, Shanghai, China, July 2016. Dr. Khattak was featured as a plenary session speaker in 16th International Conference of Transportation Professionals (CICTP) held during July 6-9, 2016, in Shanghai, China. This conference was jointly organized by Chinese Overseas Transportation Association (COTA) and Shanghai Maritime University (SMU). (50%)
- Khattak A. The Role of Connected and Automated Vehicles: How can urban areas use the data they create? Seminar presentation at National Center for Transportation Systems Productivity and Management, Civil Engineering Department, Georgia Institute of Technology, March 2016. (50%)
- Khattak A., & J. Liu, Improved Warning and Assistance Information from Connected Vehicle Basic Safety Messages, 2015 ITS World Congress, Bordeaux, France, 2015. (100%)
- Liu, J., & A. Khattak, Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles, Presented at the Transportation Research Board, National Academies, Washington, D.C., 2016. (100%)
- Liu, J., A. Khattak, & M. Zhang, Structuring and Integrating Data in Metropolitan Regions to Explore Multilevel Links Between Driving Volatility and Correlates, Presented at the Transportation Research Board, National Academies, Washington, D.C., 2016. (100%)
- Liu, J., A. Khattak & M. Zhang, Exploring Links between Naturalistic Driving Behaviors and Various Factors in Hierarchies: A Study Integrating Multiple Data Sources, 2015 Road Safety & Simulation International Conference, Orlando, FL, 2015. (100%)

Major research activities undertaken by the team and the result are summarized.

1. INTRODUCTION

Broadly speaking, the project is advancing the move toward somewhat disruptive driverless connected and automated vehicles. It is doing this by creating new knowledge about motor vehicle driving in transportation engineering and science. The techniques developed in this project provide the appropriate theoretical and application framework that leads to understanding instantaneous driving decisions, especially "driving volatility," and key correlates. Short-term driving decisions primarily depend on surrounding traffic states. An in-depth analysis of surrounding traffic states is important for understanding driver specific behavior and developing customized driver based safety applications. To this end, several methodological approaches are explored, including Markov Switching models, and hierarchical models. These models can deal with driver decisions made in real time (data velocity) and with data integrated from diverse sources (variety).

An idea that can substantially impact transportation system performance is "proactive" intersection safety management. Intersections are prevalent throughout the transportation system and represent points of conflict, delays, and safety problems. After integrating safety critical connected vehicle data with infrastructure data (crashes at intersections, road inventory), early findings suggest a positive association between location-based driving volatility and crash frequency at intersections. If many drivers behave in a volatile manner at a specific intersection, then such intersections can be identified before accidents happen. 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.

Overall, the findings from this study are advancing knowledge about instantaneous driver behavior that can form the basis for development of proactive early warnings and control assists to drivers. The results from this project are impacting the development of transportation engineering and science by leveraging the opportunity to utilize information becoming available from emergent connected and automated vehicle systems.

1.1. ACTIVITIES

During reporting period, the efforts of the team focused on following activities:

- Acquiring and analyzing real-world large-scale connected vehicle data to extract critical driving behavior information, especially driving volatility, embedded in raw BSMs.
- Applying modeling frameworks, e.g., Markov Switching models, to characterize driving regimes and embedded volatility during trips undertaken in connected vehicles.
- Assembling a unique database, which integrates intersection crash and inventory data with more

than 65 million real-world Basic Safety Messages transmitted between 3,000 connected vehicles and roadside units. The concept of volatility was expanded to specific locations and termed “location-based volatility” resulting in a proactive approach to intersection safety.

1.2. OBJECTIVES

The specific objectives are:

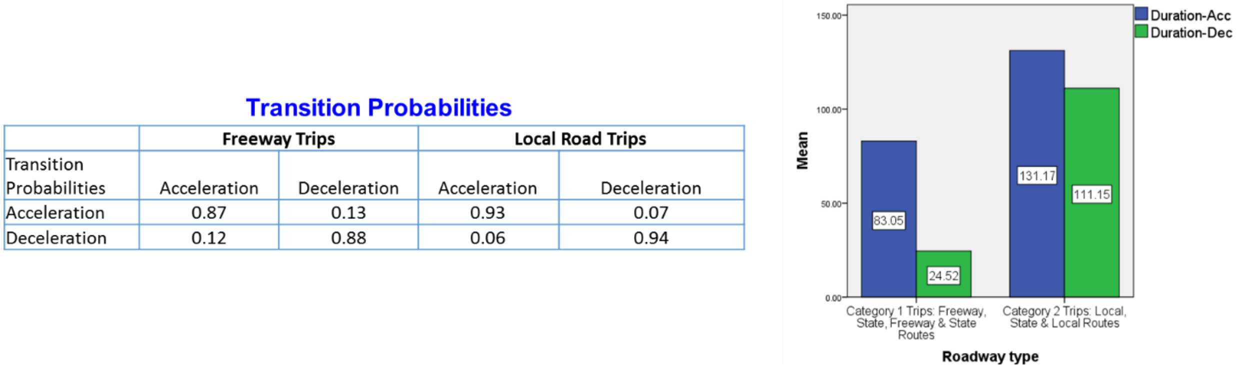
- Acquire experimental data related to Basic Safety Messages (BSM) for analysis.
- Integrate CAV databases with related databases (crash and road inventory) for analysis
- Find personally optimal policies that maximize the expected sum of rewards for individual drivers in terms of decisions related to accelerating, decelerating, and maintaining constant speed during trips.
- Investigate time-series instantaneous driving decisions of connected vehicle drivers at microscopic level and mapping such decisions to instantaneous driving contexts.
- Introduce rigorous Markov switching models for conceptualizing micro-level driving behavior into different regimes and mapping correlates to each regime.
- Establish a new methodology for proactive intersection safety management by using large-scale data generated by connected vehicles and infrastructure.
- Understand the relationship between intersection-specific volatility with crash frequencies, while controlling for other variables, using rigorous statistical tools.
- Use an MDP-induced metric and formalize the modeling method for estimating deviation from normal traffic pattern.
- Develop a hazard anticipation, notification and action-planning algorithm, optimized for a group of vehicles in close proximity and with multiple drivers in the loop (DIL).
- Identify extreme lane change behavior by investigating lateral driving behavior through extracting embedded information in CV data (i.e. lateral acceleration and distance of the host vehicle to the left and right line markings)

1.3. SUMMARY OF RESULTS

Research activities revolved around extracting useful information from big data for safety improvements and they focused on examining instantaneous driving decisions and trip-level driving volatility at a microscopic level, generation of alerts and warnings, and analyzing location-specific volatility. The results are summarized.

The results from Markov Switching models reveal that acceleration and braking are two distinct regimes in a typical driving cycle, with braking showing substantially greater volatility (Figure 2).

Compared to braking, the acceleration regime typically lasts longer. As expected, some drivers show greater volatility than others especially when driving on local and state routes. Importantly, when more objects surround a vehicle, the tendency is to accelerate even more if a driver is in acceleration regime, and to accelerate or lower the intensity of their braking if a driver is in the braking regime.



Transition Probabilities

	Freeway Trips		Local Road Trips	
Transition Probabilities	Acceleration	Deceleration	Acceleration	Deceleration
Acceleration	0.87	0.13	0.93	0.07
Deceleration	0.12	0.88	0.06	0.94

Figure 2 Transition probabilities and durations of acceleration/deceleration.

A novel framework is developed that helps in monitoring of acceleration and braking at a microscopic level, which can generate alerts and warnings, provided through advanced traveler information systems (Figure 3). The study shows that sequence of acceleration and braking events and vehicular jerk are key driving behavior factors to consider. Results revealed that the number of warnings/alerts varies significantly between driver groups and it is highly associated with young drivers, new vehicles, two-seat vehicles, AM and PM rush hours, and commuter trips.

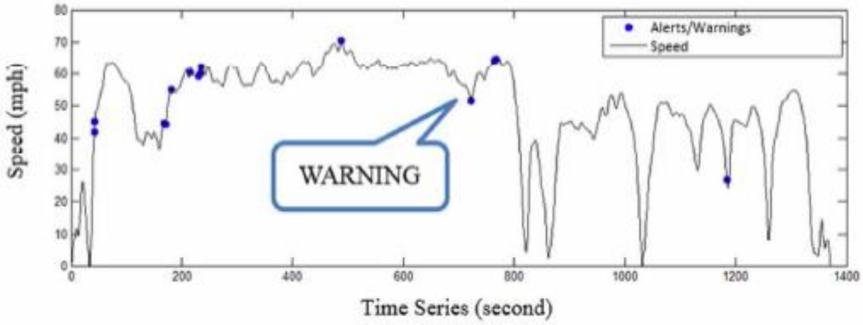


Figure 3 Alerts/warnings generated during a trip.

A fundamental understanding of instantaneous driving decisions is developed by examining two-dimensional instantaneous accelerations, i.e. longitudinal and lateral accelerations. Instantaneous driving volatility is captured by characterizing extreme driving events (Figure 4), and it clearly shows that driving behavior is strongly associated with driving contexts, e.g., whether driving on local roads or freeways. The results from empirical analysis revealed that extreme events identified in BSMs are strongly correlated with trip attributes, driver maneuvers, and driving contexts, which can help in real-time generation of warnings and alerts (Figure 5).

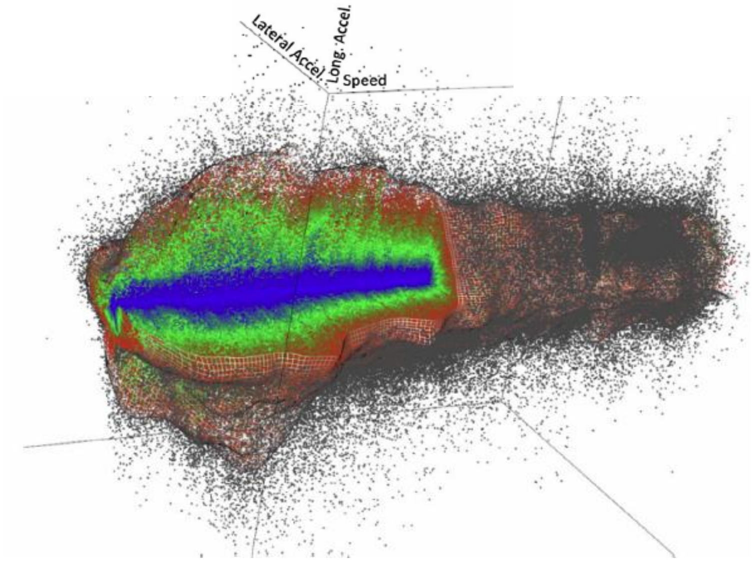


Figure 4 Plots of extreme acceleration events (hallowed in side)

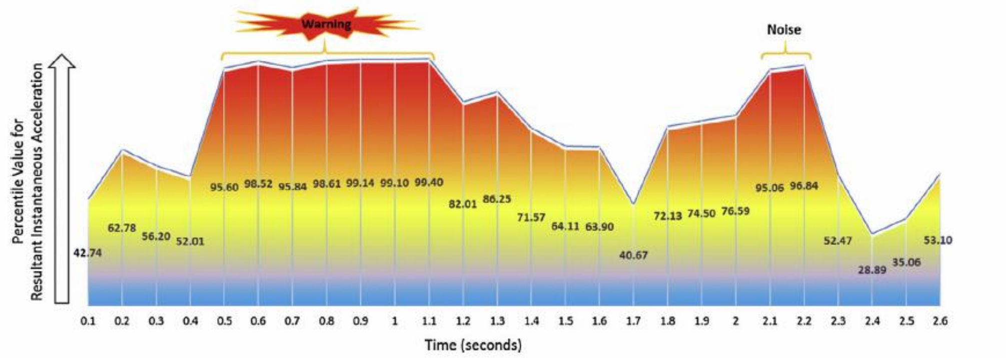


Figure 5 Generating warnings and control assists.

Using a unique database, we found that location-based volatility is positively associated with crash frequency (Figure 6). On average, a one-percent increase in the coefficient of variation in longitudinal accelerations/decelerations is associated with 0.25 increase in crash frequency for signalized intersections, while controlling for other factors. If many drivers behave in a volatile manner at a specific intersection, 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. The full text of this study is provided in the next section.

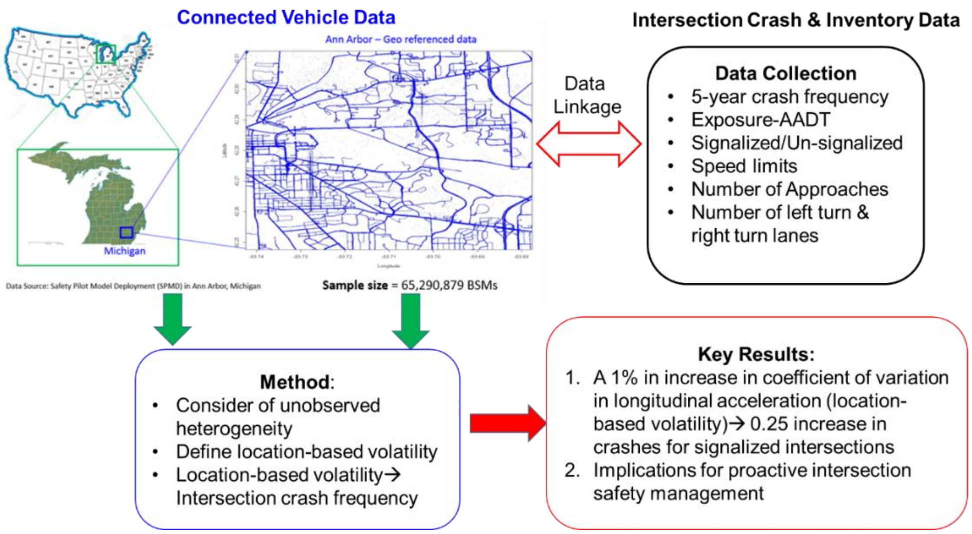
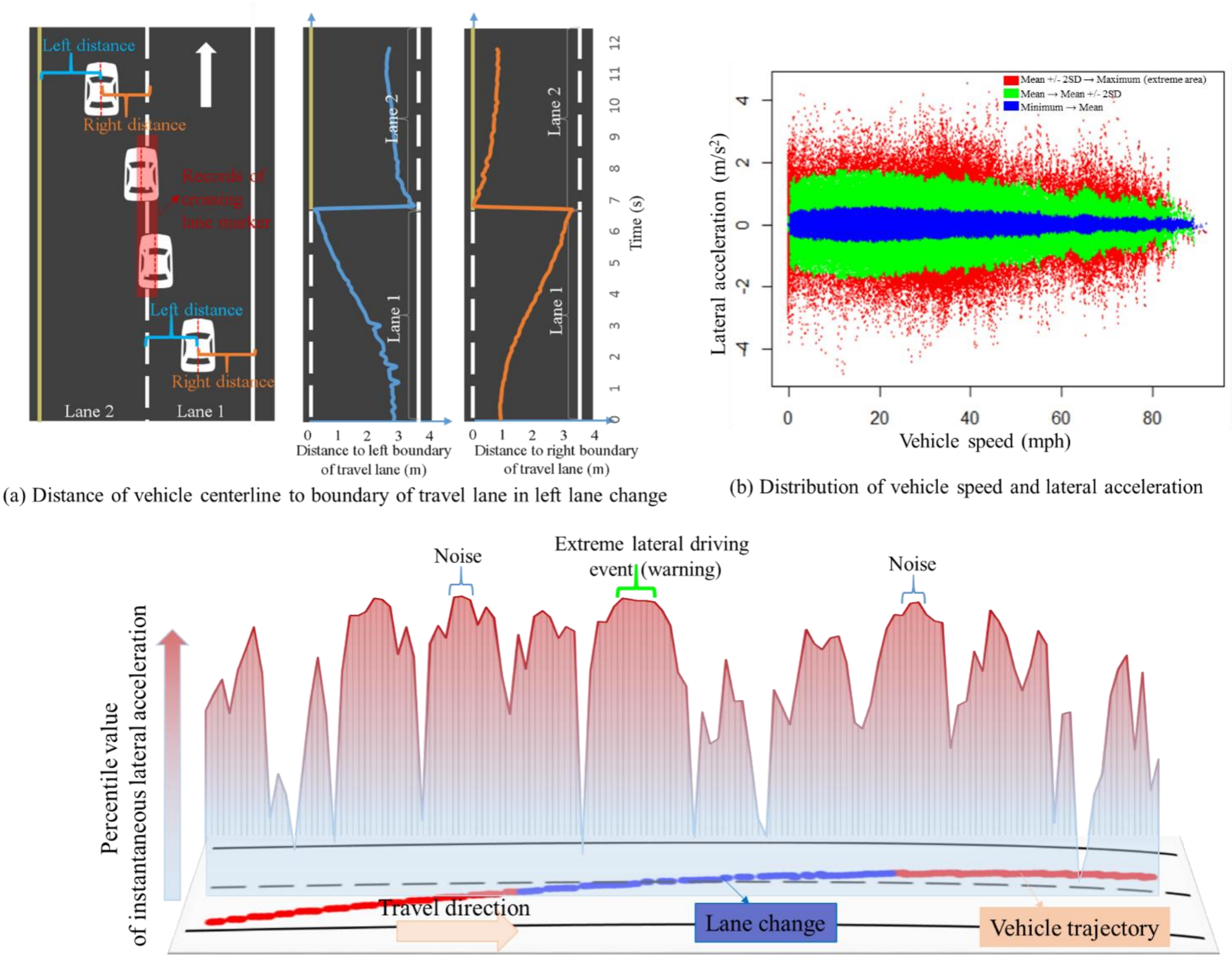

Connected Vehicle Data
Ann Arbor – Geo referenced data
Michigan
Data Source: Safety Pilot Model Deployment (SPMD) in Ann Arbor, Michigan
Sample size = 65,290,879 BSMs
Data Linkage
Intersection Crash & Inventory Data
Data Collection
• 5-year crash frequency
• Exposure-AADT
• Signalized/Un-signalized
• Speed limits
• Number of Approaches
• Number of left turn & right turn lanes
Method:
• Consider of unobserved heterogeneity
• Define location-based volatility
• Location-based volatility→ Intersection crash frequency
Key Results:
1. A 1% in increase in coefficient of variation in longitudinal acceleration (location-based volatility)→ 0.25 increase in crashes for signalized intersections
2. Implications for proactive intersection safety management

Figure 6 Methodology and key results for location-based volatility study.

Using geo-referenced data embedded in connected vehicles, a deeper understanding of normal and extreme lane change behaviors is developed by examining changes in vehicle positions and instantaneous lateral accelerations (Figure 7). Lane changes are identified based on multiple criteria, including vehicle position (i.e., sharp change in distance of a vehicle's centerline relative to lane boundary) and lane crossings recorded by on-board units (i.e., when a vehicle crosses over a lane marker). Extreme lane change events are then identified as those where lateral acceleration exceeds the 95th percentile threshold at the initiation and before the end of the lane change maneuver. The results show that, on average, the test vehicles generated 2.4 lane changes (0.5 extreme lane changes) with trip durations averaging 14 minutes. Based on the analysis of data, warnings can be generated to help drivers make more informed driving decisions about avoiding potential risks from extreme lane changes, through the application of connected vehicle technologies.



(a) Distance of vehicle centerline to boundary of travel lane in left lane change

(b) Distribution of vehicle speed and lateral acceleration

(c) Warnings of extreme lane change events (12 s)

Figure 7 Extreme Lane change events

The following section provides the full-length paper along with abstracts of relevant papers.

2. CAN DATA GENERATED BY CONNECTED VEHICLES ENHANCE SAFETY? A PROACTIVE APPROACH TO INTERSECTION SAFETY MANAGEMENT [1]

Abstract – Traditionally, evaluation of intersection safety has been largely reactive, based on historical crash frequencies 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 (SPMD) 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 transmitted between 3,000 connected vehicles and roadside units. 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; a one-percent increase in coefficient of variation in longitudinal acceleration/deceleration is associated with a 0.25 increase in crashes. The implications of the findings for proactive intersection safety management are discussed in the paper.

***Keywords*:** Proactive Safety, Driving Volatility, Connected Vehicles, Basic Safety Messages, Fixed and Random Parameters, Poisson Regression.

[1] Material in this section is based on: Kamrani, M. A. Khattak, & B. Wali, Can Data Generated by Connected Vehicles Enhance Safety? A proactive approach to intersection safety management. To be submitted for presentation and publication review.

2.1. INTRODUCTION

There is considerable evidence about vehicle conflicts at intersections resulting in crashes, making them among the most dangerous locations on roadways (5, 6). 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 (6). Safety treatments can then be applied to intersections based on historical crash data methodology. However, the emerging data from Connected and Automated (CAVs) are increasingly becoming available and they can supplement historical data to proactively identify intersections where crashes may happen in the future. Variability in instantaneous driving behaviors can be leading indicators of occurrence of unsafe outcomes such as crashes/incidents. The CAV data are in high-volume and high-resolution and are exchanged in real-time (1). In this study, we posit that expanding the concept of driving volatility (2-4) 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 (1). 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 and jerky maneuvers, and frequent switching between different driving regimes (3). Specifically, Wang et al. (4) and Liu and Khattak (2) examined the relationships between trip-based driving volatility and several factors such as demographics, trip purpose and duration, and detailed vehicle characteristics (2, 4). 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, 4).

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.

2.2. 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 (7) are still under investigation in order to establish a better vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) connection (7). Another is the platooning systems whose ultimate goal is the reduction of the gap between vehicles in order to increase roads capacity, as discussed in Bergenhem et al. (8). Optimization, analysis and improvement of traffic flow at intersection and roadways under connected vehicle environment (9) and the behavioral aspect of using CV features such as influence of warning or monitoring systems on drivers' behavior (10) are other study directions.

Also, there are a number of studies (not necessarily in CV area) trying to characterize aggressive, reckless or risky driving style (11). Among them, speed limits are usually the threshold that determines a driver's performance (12). 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 (13). 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, (14) investigated the effect of warning messages on drivers' ability to handle primary and secondary threats. The results showed an improved detection time of the primary threat while increased reaction time to the secondary threat which was placed after the primary threat. In another study (15), the impacts of dynamic route guidance on work zone safety under different market penetration of CV were explored. According to 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 (16). 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 in order to understand the drivers' behavior. For instance, a measure of driving performance in connected vehicles network has been defined as "driving volatility" (17). As such, trip-based driving volatility was introduced (17) to account for the variation of driving behaviors under different conditions using objective driving performance evaluation matrix i.e. vehicular jerk. More succinctly, (18) 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 (19) 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.

2.2.1. Research Objective and Contribution

The objectives of this study are:

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 (20), 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. 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.

2.3. METHODOLOGY

2.3.1. 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 (21). 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 other factors such as the road conditions rather than the drivers' decisions. Such information can be disseminated to connected vehicles 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 obtained data consisting of geo codes and instantaneous vehicles acceleration (longitudinal and lateral) were cleaned. In the next step, 116 signalized and un-signalized intersections were identified in Ann Arbor area (discussed later) and data were segregated based on 150 ft (as the boundary of intersections) from the center of each intersection. Then, location-based volatility (LBV) can be calculated at a specific intersection by using appropriate methods (discussed later). Given the hypothesis that LBV can be 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 appropriate modeling techniques can identify intersections where drivers, on average, have higher fluctuations in their instantaneous driving decisions (e.g. longitudinal acceleration), and where such fluctuations (LBV) are found to be correlated with historical crash data. By carefully analyzing high-resolution real-world data transmitted between connected vehicles and through application of appropriate statistical methods can ultimately help us in generating pro-active (rather than the traditional reactive safety approach) alerts and warnings to vehicles at a particular intersection. Such pro-active 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.

2.3.2. Location Based Volatility

Understanding instantaneous driving volatility at specific intersections is one of the most challenging aspects of the current study. For calculation of location-based volatility for sampled intersections in study area, different instantaneous driving measures can be used such as accelerations, steering angles or position of brakes, for details see Liu and Khattak (2). However, for calculation of LBV, the authors prefer using longitudinal and lateral acceleration as they are direct outcomes of vehicle maneuvering. However, Due to the considerable amount of erroneous lateral acceleration, only longitudinal acceleration data were used. The present study uses a standardized measure of dispersion called Coefficient of Variation (CV) (also known as the ratio of relative standard deviation) for quantifying the fluctuations in longitudinal acceleration and/or decelerations at a specific intersection (22). Note that different measures such as range, interquartile range, variance or standard deviation can be used for capturing variability in longitudinal accelerations at a specific intersection. However, standard deviation and variance are preferable as whole information embedded in data is used for calculation of variability. While standard deviation and its square (variance) can quantify variability in data, both of the 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 (for example) can be quantified as proportion of mean acceleration (or deceleration), at a specific intersection. This approach is used due to its simplicity and ability to 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.

2.3.3. Calculation of LBV

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 (2). 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. However, the authors could not use the available lateral acceleration data due to significant erroneous values for lateral accelerations i.e. out of total of 65,290,879 BSMs, lateral acceleration data is erroneous for 27,240,788 BSMs. However, 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.

In order to compute volatility for each intersection, speed bins with half mph width are considered. This is consistent with past literature on this topic (2). Longitudinal accelerations/decelerations were categorized into half-mile speed bins based on their related speed values. 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 values are separated, and the mean and the standard deviations are calculated for acceleration values for each speed bin within each intersection. Finally, in order to obtain CV as a measure of LBV, standard deviations of acceleration are divided by mean of accelerations within each bin. The final CV for longitudinal acceleration is then calculated by taking average of the CVs within each bin. This process is done separately for longitudinal accelerations and decelerations respectively, and CVs for longitudinal acceleration and deceleration are averaged to obtain CV for entire intersection. Finally, CVs are obtained for all 116 intersections by using instantaneous longitudinal accelerations collected from more than 30 million BSM packets. The calculated coefficient of variation (CV) for a specific intersection provides the relative measure of dispersion of longitudinal accelerations with respect to mean longitudinal accelerations, and thus different intersections can be compared on the basis of calculated CVs.

2.3.4. 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 (23, 24).

For a Poisson model, the probability of having a specific number of crashes "n" at intersection "i" can be written as (22):

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

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 (22):

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

Where X_i is a vector of explanatory variables and β is a vector of estimable parameter estimates.

The Poisson function defined in Equation 1 and 2 can be maximized by standard maximum likelihood procedure with the following likelihood function (22):

$$L(\beta)=\prod_i \frac{\exp[-\exp(\beta X_i)]\,[\exp(\beta X_i)]^n}{n_i!} \tag{3}$$

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 ∈, as follow:

$$\ln(\lambda_i)=\beta(X_i)+\in_i \tag{4}$$

Where $\exp(\in_i)$ in Equation 4 is a gamma-distributed error term with mean one and variance α (22). The conditional probability of crashes then becomes (23):

$$P(n_i|\in)=\frac{\exp[-\lambda_i \exp(\in_i)]\,[\lambda_i \exp(\in_i)]^{ni}}{n_i!} \tag{5}$$

Following (23), to obtain unconditional distribution of n_i, $\in_i$ can be integrated out of Equation 5, which results in the following maximum likelihood estimation problem:

$$P(n_i)=\frac{\Gamma(\theta+n_i)}{[\Gamma(\theta).n_i!]}.u_i^{\theta}(1-u_i)^{n_i} \tag{6}$$

Where: u_i is $\theta(\theta+\lambda_i)$ and $\theta=\frac{1}{\alpha}$, Γ is the gamma function. It can be seen in Equation 6 that Poisson is a limiting function of the Poisson-Gamma model as variance α approaches to zero. Following (22), if α is significantly different from zero, negative binomial regression should be favored and if not, Poisson model can be more appropriate (25). While presence of over-dispersion can be indicated by the mean and variance of crash data (22), formally a Lagrange multiplier can be performed to statistically test the existence of over- dispersion in Poisson model (26). The test statistic is defined as:

$$LM=\left[\frac{\sum_{i=1}^{n}[(y_i-\mu_i)^2-y_i]}{2\sum_{i=1}^{n}{\mu_i}^2}\right]^2 \tag{8}$$

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 8 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 (26).

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 (27). 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 (27). As such, random parameters can be included in the estimation framework as:

$$\beta_i = \beta + \varphi_i \tag{9}$$

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 9, the Poisson parameter in Equation 2 becomes:

$$\lambda_i | \varphi_i = EXP(\beta X) \tag{10}$$

And, the Poisson parameter in Equation 4 in Poisson-Gamma model becomes:

$$\lambda_i | \varphi_i = EXP(\beta X + \in_i) \tag{11}$$

Finally, the following likelihood function for random-parameter model can be maximized through maximum simulated likelihood technique (28):

$$LL = \sum_i ln \int_{\varphi_i}^{i} g(\varphi_i) P(n_i|\varphi_i) d\varphi_i \tag{12}$$

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 (22, 25).

2.4. DATA

The data used in this study are from BSMs sent and received by 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. The field test includes 75 miles of instrumented roadway with approximately 26 roadside units that are capable of communicating with appropriately equipped vehicles, and devices via Dedicated Short Range Communication (DSRC) technology. The data are stored in a transportation data sharing system, called Research Data Exchange (RDE, https://www.its-rde.net/home), maintained by the Federal Highway Administration under US DOT. This study used Basic Safety Message

(BSM) data set extracted from the SPMD. 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). Table 1 illustrates variables related to position and motion (for details see (1, 2)). One-month data (October 2012) were utilized in this study. Table 1 also summarizes descriptive statistics for key variables. Based on the distributions of key variables used for calculation of location-based volatility, the data seems to be of reasonable quality. Figure 8 shows Ann Arbor area map created from BSM data by plotting trajectory of connected vehicles participating in SPMD. The figure is a good indication of data precision and coverage of Ann Arbor area.

In order 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 Ann Arbor area, 115 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-year average number of crashes (2010-2014) along with other explanatory factors 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 115 intersections.

Table 1 Description of key variables and Descriptive Statistics

Variable	**Description**
Position	
Latitude	Current degree of latitude at which the vehicle is located
Longitude	Current degree of longitude at which the vehicle is located
Motion	
Speed	Current vehicle speed, as determined from the vehicle's transmission
Longitudinal Acceleration	Longitudinal acceleration measured by an Inertial Measurement Unit (IMU)
Heading	Vehicle heading/direction from North

Variable		**Mean**	**SD**	**Median**	**Min**	**Max**	**Range**
Position	Longitude	-83.72	0.02	-83.72	-83.77	-83.68	0.09
	Latitude	42.28	0.02	42.28	42.25	42.31	0.06
Motion	Speed (km/h)	61.03	38.73	59.98	0	177.26	177.26
	Longitudinal Acceleration (m/s^2)	-0.02	0.66	-0.01	-9.9	9.88	19.78

Sample size = 65,290,879 BSMs

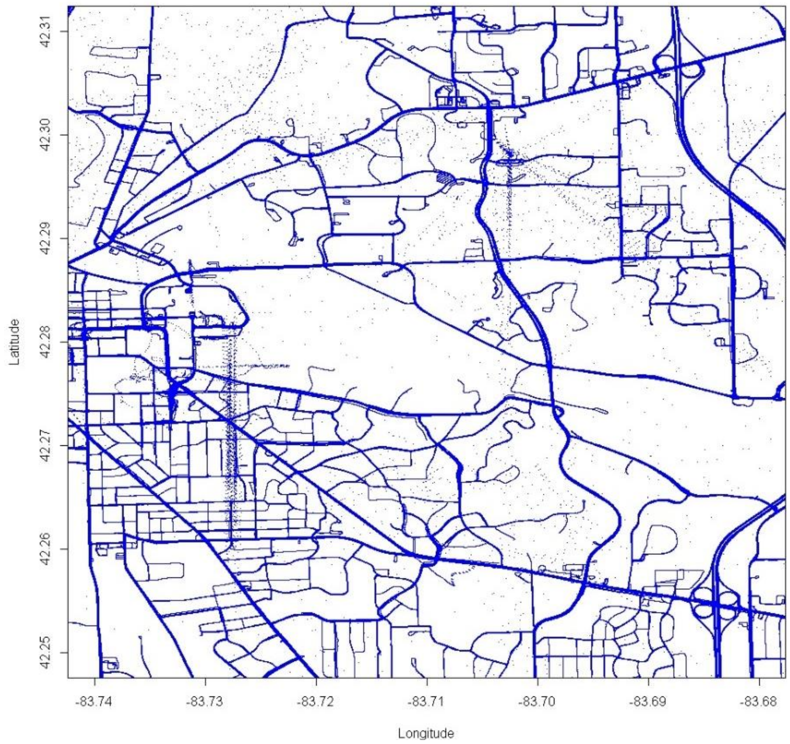


Figure 8 Ann Arbor map created from BSM data

2.5. RESULTS

2.5.1. **Descriptive Statistics**

Table 2 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 separately for signalized intersections (N=52) and unsignalized intersections (N=64). For all the intersections, signalized, and unsignalized intersections, the mean five-year crash frequency is 6.78, 11.73, and 2.765. As expected, signalized intersections have significantly higher crash frequency (on average) than unsignalized intersections. This finding is in agreement with Abdel Aty and Keller (29) who found approximately 9.6 crashes per year at signalized intersections as opposed to only 2 crashes per year on unsignalized intersections (29). 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 (29). 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” (24).

Regarding location-based volatility, it can be seen that the average coefficient of variations (CV) are 78.2%, 85.1%, and 78% for all intersections, signalized, and unsignalized intersections respectively.

Compared to unsignalized intersections, these statistics suggest that signalized intersections on average have higher variability in longitudinal accelerations/decelerations, and thus can be more volatile.

In order to avoid omitted variable bias in modeling (27), data on other variables such as five-year average AADT (major and minor approach), speed limits (major and minor approach), and number of approaches were collected. Regarding the number of approaches, 40% of all intersections, 63.4% of signalized intersections, and 21.8% of unsignalized intersections are four-legged intersections (Table 2). In terms of exposure on major and minor roads, signalized intersections have higher (on average) AADT than unsignalized intersections (24,550 vs 20,012 for major roads and 10009.62 vs 8479.68 for minor roads). Furthermore, it is likely that coefficient of variation (CV) may be correlated with other exposure related factors. Due to several identified and unidentified interactions among key factors in crash data, multicollinearity can arise and can affect model results significantly if not addressed carefully. Existence of multicollinearity among independent variables was checked by using the VIFs. As shown in Table 1, the VIF value of each variable is much smaller than 10, which indicates absence of significant multicollinearity (22).

Table 2 Description of Key Variables and Descriptive Statistics

All intersections (N=116)					
Variables	Mean	SD	Min	Max	VIF
Crash frequency	6.784	6.673	0	40	2.59
Speed limit major	35.344	7.244	25	45	1.46
Speed limit minor	30.474	3.955	25	45	1.64
CV%	78.273	6.555	55.071	91.657	1.07
Major road AADT	22046.55	8630.87	3100	52700	4.52
Minor road AADT	9165.51	4142.99	1100	27400	4.23
Major road AADT (log)	9.91	0.471	8.039	10.872	2.84
Minor road AADT (log)	9.028	0.466	7.003	10.218	3.36
Four-leg (1/0)	0.405	0.493	0	1	1.57
Signalized Intersections (N=52)					
Variables	Mean	SD	Min	Max	VIF
Crash frequency	11.73	6.851	1	40	2.13
Speed limit major	35.769	7.301	25	45	2.23
Speed limit minor	30.673	5.051	25	45	1.91
CV%	85.13	5.644	69.366	91.657	1.31
Major road AADT	24550	8317.48	13900	52700	3.12
Minor road AADT	10009.62	5722.03	3100	27400	3.23
Major road AADT (log)	10.059	0.306	9.539	10.872	5.12
Minor road AADT (log)	9.075	0.517	8.039	10.218	4.71
Four-leg (1/0)	0.634	0.486	0	1	1.37

Unsignalized Intersections (N=64)					
Variables	Mean	SD	Min	Max	VIF
Crash frequency	2.765	2.586	0	12	1.71
Speed limit major	35	7.237	25	45	2.87
Speed limit minor	30.312	2.799	25	40	2.91
CV%	78	7.256	55.07	89.454	1.08
Major road AADT	20012.5	8402.27	3100	39400	7.51
Minor road AADT	8479.68	1939.25	1100	13400	4.11
Major road AADT (log)	9.788	0.544	8.039	10.581	2.01
Minor road AADT (log)	8.989	0.421	7	9.503	1.14
Four-leg (1/0)	0.218	0.416	0	1	1.31

Notes: SD is standard deviation; Min is minimum value; Max is maximum value; VIF is variance inflation factor

2.5.2. Modeling Results

For examining the correlations between crash frequency and location-based volatility (as measured by CV), 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 unsignalized 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 number of approaches. Furthermore, the descriptive statistics for crash frequencies in Table 2 apparently reveal the existence of over-dispersion in data in which case Negative Binomial models should be preferred (22). Thus, three negative binomial regression models (fixed-parameters are also developed with explanatory variables shown in Table 2). Despite that the values of mean and variance for crash frequency in Table 2 reveal over-dispersion, statistical tests are conducted to confirm the existence of over-dispersion in data at hand (26). As explained in methodology section, Lagrange Multiplier tests were conducted for the three Poisson models. By using Equation 8, the Lagrange Multiplier (LM) values for the three models were 0.05, 0.031, and 0.15 for all intersections, signalized intersections, and unsignalized 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. Thus, the null hypothesis that Poisson regressions are more appropriate is failed to reject, and it would be more appropriate to use Poisson regressions (22). For brevity, the results of Negative Binomial model are not shown over here.

Due to the likely presence of unobserved heterogeneity in crash data (25) 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 (25). Regarding functional form of random-parameters, log-normal, Weibull, uniform, and triangular distributions are tested with normally distributed random parameters giving the best fit.

Table 3 Modeling Results of Fixed- and Random-Parameter Poisson Regressions

	All Intersections					**Signalized Intersections**					**Unsignalized Intersections**				
	Fixed Parameter		**Random Parameter**			**Fixed Parameter**		**Random Parameter**			**Fixed Parameter**		**Random Parameter**		
Variables	Coeff	t-stat	Coeff	t-stat	ME	Coeff	t-stat	Coeff	t-stat	ME	Coeff	t-stat	Coeff	t-stat	ME
Constant	-15.35	-11.01	-15.24	-11.07		-12.38	-7.33	-12.72	-7.52		-12.62	-3.77	-10.66	-2.74	
Coefficient of Variation	0.001	1.95	0.014	2.19	0.06	0.022	2.86	0.024	2.85	0.25	0.0004	0.95	-0.007	-0.72	-0.01
(Standard deviation)	---	---	*(0.008)*	*(15.21)*	---	---	---	*(0.001)*	*(3.51)*	---	---	---	---	---	
Major road AADT (log form)	1.614	12.21	1.493	11.32	6.82	1.218	6.72	1.238	6.05	13.22	1.257	5.56	1.124	5.36	2.46
Minor road AADT (log form)	0.153	1.61	0.128	1.98	0.58	0.081	0.74	0.088	1.05	0.94	0.1	0.37	0.061	0.18	0.13
(Standard deviation)	---	---	*(0.008)*	*(2.01)*	---	---	---	---	---		---	---	---	---	
Speed limit major	-0.02	-3.38	-0.026	-3.88	-0.11	-0.023	-2.81	-0.025	-2.85	-0.26	-0.015	-1.65	-0.024	-1.98	-0.05
(Standard deviation)	---	---	*(0.002)*	*(2.43)*	---	---	---	---	---		---	---	*(0.012)*	*(5.42)*	
Speed limit minor	0.01	1.1	0.019	1.93	0.08	0.021	1.92	0.021	1.93	0.22	0.023	0.91	0.042	1.65	0.09
4-legged intersection (1/0)	---	---	---	---	---	0.279	2.75	0.284	2.56	3.03	0.35	1.91	0.33	1.57	0.73
Log-likelihood at Zero	-522.81		-522.81			-199.42		-199.42			-148.2		-148.2		
Log-likelihood at Convergence	-394.08		-318.85			-142.8		-137.35			-125.71		-119.9		
McFadden ρ^2	0.2462		0.39			0.283		0.312			0.15		0.19		
Sample Size	116					52					64				

Notes: Coeff: parameter estimate; ME: Average Marginal Effects from Random Parameter Model.

The final results obtained from fixed and random parameter Poisson model are presented in Table 3. 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 ρ^2 (Table 3) (22). While this study does not focus on methodological approaches for modeling intersection crash data, the predicted vs actual values of crashes (Figure 9) are plotted and reveal statistical superiority of random parameter models in fitting the data at hand.

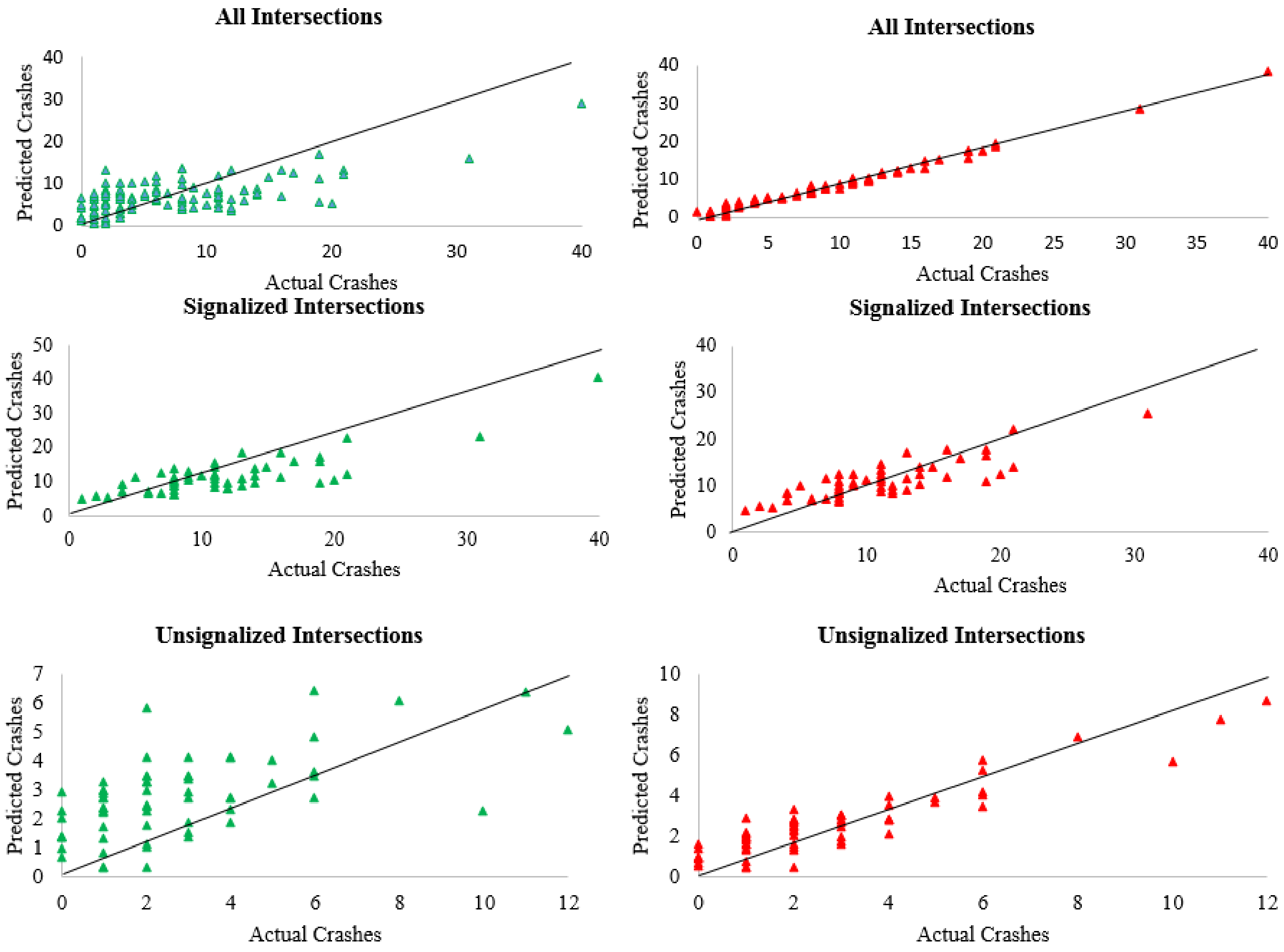


Figure 9 Mean-expected over actual number of crashes for fixed and random-parameter Poisson models (Green: fixed parameter models; Red: random parameter models)

2.6. DISCUSSION

Coming to the fixed-parameter estimation results for all intersections (Table 3), the results provide evidence that coefficient of variation (CV) is positively associated (statistically significant at 90% confidence level) with crash frequency. Likewise, the association between CV and crash frequency is also positive and statistically significant for signalized intersections. However, both for all intersections and for signalized intersections, CV is found to be normally distributed random parameter, suggesting heterogeneity in associations between coefficient of variation and crash frequency. Note that despite the presence of heterogeneity, the association is positive for 100% of the observations (see mean and standard deviations of coefficient of variation in random parameter models).

For example, referring to marginal effects for random parameter in Table 3, on average one-percent increase in CV is associated with 0.06 increase in crash frequency for all intersections and 0.25 increase in crash frequency for signalized intersections. These findings have implications for proactive intersection-related safety strategies. Also, it is interesting to note the significantly higher marginal effect of CV (0.25) 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 sense that if many drivers behave in a volatile (higher variability in longitudinal accelerations) manner at a specific intersection, the reasons of such volatile behaviors may be related to intersection conditions rather than driver's decisions. Given that signalized intersections are typically observed to have more crashes (29), proactive intersection-customized strategies can be designed. 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 unsignalized intersections, as can be seen in Table 3, the study did not find any statistical evidence for associations between CV and crash frequency either in fixed-parameter or random-parameter model. However, this requires further investigation in the future.

The estimation results also quantify associations between major and minor road AADT and crash frequency. Referring to marginal effects from random-parameter model, one-log unit increase in major road AADT is associated with 6.82, 13.22, and 2.46-unit increase in crash frequency for all intersections, signalized intersections, and unsignalized intersections respectively. While minor road AADT is significant in the random-parameter model for all intersections, the relationships are not statistically significant for signalized and unsignalized intersections (Table 3). Likewise, speed limit on major roads is negatively associated with crash frequency for in all three random-parameter models. These findings are consistent with past studies on this topic (5, 30). Finally, 4-legged signalized intersections are more likely to have higher crash frequencies. The marginal effect for this variable in signalized intersections model is 3.03;

suggesting 3.03-unit increase in crash frequency with each one-unit increase in 4-legged intersection. This finding is in agreement with Abdel-Aty and Haleem (5).

2.7. 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 methodology, 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. While the current sample size may not be enough to draw robust conclusions, the authors have used all available data for 116 intersections.

2.8. 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 was created by combining more than 65 million Basic Safety Messages transmitted between connected vehicles and roadside units at 115 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 (CV) (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 intersection-specific volatility is positively associated with crash frequency. On average, a one-percent increase in coefficient of variation in longitudinal accelerations/decelerations is associated with 0.25 increase in crash frequency for signalized intersections, while controlling for other factors. A statistically significant relationship was not found between this location-based volatility measure and crash frequencies for unsignalized intersections.

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 (29), 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.

3. DELIVERING IMPROVED ALERTS, WARNINGS, AND CONTROL ASSISTANCE USING BASIC SAFETY MESSAGES[2]

Abstract – When vehicles share their status information with other vehicles or the infrastructure, driving actions can be planned better, hazards can be identified sooner, and safer responses to hazards are possible. The Safety Pilot Model Deployment (SPMD) is underway in Ann Arbor, Michigan; the purpose is to demonstrate connected technologies in a real-world environment. The core data transmitted through Vehicle-to-Vehicle and Vehicle-to-Infrastructure (or V2V and V2I) applications are called Basic Safety Messages (BSMs), which are transmitted typically at a frequency of 10 Hz. BSMs describe a vehicle's position (latitude, longitude, and elevation) and motion (heading, speed, and acceleration). This study proposes a data analytic methodology to extract critical information from raw BSM data available from SPMD. A total of 968,522 records of basic safety messages, gathered from 155 trips made by 49 vehicles, was analyzed. The information extracted from BSM data captured extreme driving events such as hard accelerations and braking. This information can be provided to drivers, giving them instantaneous feedback about dangers in surrounding roadway environments; it can also provide control assistance. While extracting critical information from BSMs, this study offers a fundamental understanding of instantaneous driving decisions. Longitudinal and lateral accelerations included in BSMs were specifically investigated. Varying distributions of instantaneous longitudinal and lateral accelerations are quantified. Based on the distributions, the study created a framework for generating alerts/warnings, and control assistance from extreme events, transmittable through V2V and V2I applications. Models were estimated to untangle the correlates of extreme events. The implications of the findings and applications to connected vehicles are discussed in this paper.

***Keywords*:** Connected vehicles; Basic safety messages; Extreme events; Longitudinal and lateral accelerations

[2] Abstract is based on: Liu, Jun, and Asad J. Khattak. "Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles." *Transportation Research Part C: Emerging Technologies* 68 (2016): 83-100. A version this (full-length) paper was included in the previous report.

4. DYNAMICS OF DRIVING REGIMES EXTRACTED FROM BASIC SAFETY MESSAGES[3]

Abstract – Driving volatility captures the extent of speed variations when a vehicle is being driven. Extreme longitudinal variations signify hard acceleration or braking. Warnings and alerts given to drivers can reduce such volatility potentially improving safety, energy use, and emissions. This study develops a fundamental understanding of instantaneous driving decisions, needed for hazard anticipation and notification systems, and distinguishing normal from anomalous driving. The driving task is divided into two distinct yet unobserved regimes. The research issue is to characterize and quantify these regimes in typical driving cycles, explore when the two regimes change and the key correlates associated with each regime. Using Basic Safety Message (BSM) data from the Safety Pilot Model Deployment in Ann Arbor, Michigan, Markov switching models are estimated for various trip types. While thousands of instrumented vehicles with V2V and V2I communication systems are being tested, nearly 1.4 million records of BSMs, from 184 trips undertaken by 71 instrumented vehicles are analyzed in this study. Then even more detailed analysis of 43 randomly chosen trips that were undertaken on various roadway types is conducted. The results indicate that acceleration and deceleration are two distinct regimes, and as compared to acceleration, drivers decelerate at higher rates, and braking is significantly more volatile than acceleration. Different correlations of the two regimes with instantaneous driving contexts are explored. The study contributes to analyzing volatility in short-term driving decisions, and how changes in acceleration and braking can be mapped to a combination of local traffic states surrounding the vehicle.

***Keywords*:** Connected and Automated Vehicles, Driving Behavior, Markov Decision Processes, Inverse Reinforcement Learning, Basic Safety Messages

[3] Abstract is based on: Khattak A., & B. Wali, Dynamics of Driving Regimes Extracted from Basic Safety Messages Transmitted Between Connected Vehicles. To be submitted for presentation and publication review.

5. IDENTIFYING AND ANALYZING EXTREME LANE CHANGE EVENTS USING BASIC SAFETY MESSAGES[4]

Abstract – Traffic congestion and safety are challenging problems in the US, costing nearly a trillion dollars annually. A deeper understanding of driving behaviors, through emerging data from connected vehicles, can potentially reduce dangerous situations and unstable flows caused by aggressive or extreme behaviors. As lane change maneuvers are fundamental to traffic flow and safety, this study focuses on microscopic driver-level instantaneous decisions regarding situations where drivers make lane change maneuvers, especially extreme lane change events on various roadway types. A sub-sample of 534,509 Basic Safety Message records from 64 randomly-selected trips (5 minutes or longer) are analyzed from connected vehicles participating in Safety Pilot Model Deployment in Michigan. Since BSMs describe a vehicle's operation and performance, lane changes are identified based on multiple criteria, including vehicle position (i.e., sharp change in distance of a vehicle's centerline relative to lane boundary) and lane crossings recorded by on-board units (i.e., when a vehicle crosses a lane marker). Extreme lane change events are then identified as those where lateral acceleration exceeds the 95th percentile threshold at the initiation and before the end of the lane change maneuver. A total of 157 lane changes and 33 extreme lane changes were identified in the data. On average, the test vehicles generated 2.4 lane changes (0.5 extreme lane changes) with trip durations averaging 14 minutes. Modeling results show that high maximum speed during a trip is associated with more extreme lane changes. Based on the analysis of data, warnings can be generated to help drivers make more informed driving decisions about avoiding potential risks from extreme lane changes, through the application of connected vehicle technologies.

***Keywords*:** Lane Change Identification, Connected Vehicles, Basic Safety Messages, Extreme Lane Change Events, Lateral Acceleration

[4] Abstract is based on: Zhang, M. & A. Khattak, Identifying and Analyzing Extreme Lane Change Events Using Basic Safety Messages in a Connected Vehicle Environment. To be submitted for presentation and publication review.

6. REFERENCES

1. Henclewood, D., *Safety Pilot Model Deployment – One Day Sample Data Environment Data Handbook*, U.D.o.T. Research and Technology Innovation Administration. Research and Technology Innovation Administration, Editor. 2014: McLean, VA.
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, 2016. **68**: p. 83-100.
3. Khattak, A., S. Nambisan, and S. Chakraborty, *Study of Driving Volatility in Connected and Cooperative Vehicle Systems. National Science Foundation*. 2015.
4. Wang, X., et al., *What is the level of volatility in instantaneous driving decisions?* Transportation Research Part C: Emerging Technologies, 2015. **58**: p. 413-427.
5. Abdel-Aty, M. and K. Haleem, *Analyzing angle crashes at unsignalized intersections using machine learning techniques.* Accident Analysis & Prevention, 2011. **43**(1): p. 461-470.
6. 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, 1998(1635): p. 113-120.
7. Osman, O.A. and S. Ishak, *A network level connectivity robustness measure for connected vehicle environments.* Transportation Research Part C: Emerging Technologies, 2015. **53**: p. 48-58.
8. Bergenhem, C., et al. *Overview of platooning systems*. in *Proceedings of the 19th ITS World Congress, Oct 22-26, Vienna, Austria (2012)*. 2012.
9. Ngoduy, D., S. Hoogendoorn, and R. Liu, *Continuum modeling of cooperative traffic flow dynamics.* Physica A: Statistical Mechanics and its Applications, 2009. **388**(13): p. 2705-2716.
10. Farah, H. and H.N. Koutsopoulos, *Do cooperative systems make drivers' car-following behavior safer?* Transportation research part C: emerging technologies, 2014. **41**: p. 61-72.
11. NHTSA. *Resource Guide Describes Best Practices For Aggressive Driving Enforcement* 2000 [cited 2016 June 22nd]; Available from: http://www.nhtsa.gov/About+NHTSA/Traffic+Techs/current/Resource+Guide+Describes+Best+Practices+For+Aggressive+Driving+Enforcement.
12. 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, 2000. **3**(1): p. 39-51.
13. Paleti, R., N. Eluru, and C. Bhat, *Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes.* Accident Analysis & Prevention, 2010. **42**(6): p. 1839-1854.
14. 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.
15. 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, 2015. **30**(2): p. 04015020.
16. Du, L. and H. Dao, *Information Dissemination Delay in Vehicle-to-Vehicle Communication Networks in a Traffic Stream.* Intelligent Transportation Systems, IEEE Transactions on, 2015. **16**(1): p. 66-80.
17. Wang, X., et al., *What is the Level of Volatility in Instantaneous Driving Decisions?* Transportation Research Part C: Emerging Technologies, 2015.
18. Liu, J., X. Wang, and A. Khattak, *Generating Real-Time Driving Volatility Information*, in *2014 World Congress on Intelligent Transport Systems*. 2014: Detroit, MI.
19. 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.
20. Kludt, K., et al., *Human Factors Literature Reviews on Intersections, Speed Management, Pedestrians and Bicyclists, and Visibility*. 2006.
21. Henclewood, D., *Safety Pilot Model Deployment – One Day Sample Data Environment Data*

Handbook. 2014, Research and Technology Innovation Administration, US Department of Transportation: McLean, VA.

22. Washington, S.P., M.G. Karlaftis, and F. Mannering, *Statistical and econometric methods for transportation data analysis*. 2010: CRC press.
23. Poch, M. and F. Mannering, *Negative binomial analysis of intersection-accident frequencies.* Journal of Transportation Engineering, 1996. **122**(2): p. 105-113.
24. 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, 2004. **36**(2): p. 193-211.
25. Anastasopoulos, P.C. and F.L. Mannering, *A note on modeling vehicle accident frequencies with random-parameters count models.* Accident Analysis & Prevention, 2009. **41**(1): p. 153-159.
26. Greene, W.H., *Econometric analysis*. 2003: Pearson Education India.
27. Mannering, F.L. and C.R. Bhat, *Analytic methods in accident research: methodological frontier and future directions.* Analytic Methods in Accident Research, 2014. **1**: p. 1-22.
28. 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, 2016 (forthcoming).
29. Abdel-Aty, M. and J. Keller, *Exploring the overall and specific crash severity levels at signalized intersections.* Accident Analysis & Prevention, 2005. **37**(3): p. 417-425.
30. Ye, X., et al., *A simultaneous equations model of crash frequency by collision type for rural intersections.* Safety Science, 2009. **47**(3): p. 443-452.