BIG DATA FOR SAFETY MONITORING, ASSESSMENT AND IMPROVEMENT

FINAL REPORT

SOUTHEASTERN TRANSPORTATION CENTER

ASAD KHATTAK, JUN LIU & XIN WANG

AUGUST 2015

US DEPARTMENT OF TRANSPORTATION GRANT DTRT13-G-UTC34

DISCLAIMER

ACKNOWLEDGEMENTS

Data used for this project are from several sources, including the National Renewable Energy Laboratory, 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 Driving Simulator Laboratory of the University of Tennessee. 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 worked on the major research initiative. 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.

1. Report No. STC-2015-M4.UTK	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Big Data for Safety Monitoring, Assessment and Improvement		5. Report Date August 2015
		6. Source Organization Code N/A
7. Author(s) Khattak, Asad; Liu, Jun; Wang, Xin		8. Source Organization Report No. STC-2015-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: September 2013 – August 2015 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 our first year efforts undertaken to determine appropriateness of sensor data that can be used in transportation safety studies, extracting, processing and integrating data from multiple sources, characterizing volatile driving behaviors using sensor data and generating driver feedback from vehicle-to-vehicle and vehicle-to-infrastructure communication data. Sensor and behavioral data are structured to mine them using multi-level modeling, demonstrating how correlates of driving behaviors can be untangled.

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	21. No. of Pages #130	22. Price N/A

Form DOT F 1700.7 (8-72)

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Transportation safety is a priority of the US Department of Transportation, with substantial resources devoted to reducing costs of injuries, fatalities, and property damage. The tools developed for addressing safety problems largely rely on post-crash analysis of the data, i.e., police reports integrated with road inventory and traffic data. Recently, a limited amount of naturalistic driving data has become available, providing insights into associated pre-crash factors. This study explores whether safety problems can be ameliorated by extracting useful information from a wider range of large-scale behavioral and sensor data, which is increasingly available in digital form. Indeed, the task of extracting information from large-scale databases is challenging because of the enormous amounts of data involved, i.e., data from various sources that include travelers, vehicles, infrastructure and the environment coupled with social, economic and spatial data, collectively referred to as "Big Data." Nevertheless, data-rich environments represent an opportunity for innovation in transportation system planning, design, operation and maintenance and toward achieving safety goals. This project develops a framework for the use of Big Data to facilitate safety monitoring, assessment and improvement, reporting on the first year efforts of the work undertaken in this major research initiative of the Southeastern Transportation Center.

The research team assessed the quality of Big Data in terms of relevance to safety, variety of the databases, and their reliability/validity. While some of the Big Data identified and analyzed in this study are not routinely available in real-time operational safety monitoring and incident/accident management, such data are expected to become increasingly available. The data used in this study comes from a wide variety of sources that include:

- Global Positioning System data on movements of individuals through time and space.
- Regional surveys of travel behavior containing travel and socio-economic data.
- Traffic data, i.e., traffic counts.
- Socio-demographic information from Census (county and regional levels).

- Geographic information from various sources including Google Earth.
- Basic safety messages exchanged between vehicles and infrastructure.

The study contributes by demonstrating a way to integrate data from multiple sources to explore links between naturalistic driving behaviors and various factors that are structured in hierarchies.

The information needs of individual travelers and transportation system planners and managers were considered in developing the framework. Safety-enhancement strategies/solutions for individual users include delivering alerts, warnings, and control assists in crash-imminent situations. One way transportation system planners and managers can take advantage of Big Data is by accessing incoming data and examining measures of how people are driving overall in metropolitan areas. In line with their data needs, this report documents four major themes addressed by the project team:

1) Data quality to explore how much information is lost when using sensor data at different sampling rates, and identifying data needs at the data collection stage;
2) Characterizing driving behaviors as volatile or calm using large-scale sensor and behavioral data. Specifically, data about the volatility in driving (hard accelerations and braking) was explored. Over-aggressive driving situations were identified and GPS data from drivers was visualized.
3) Structuring and integrating safety data from multiple sources. Data analysis included estimation of statistical models, and testing of hypotheses to explore associations between various factors and safety-relevant driving decisions. Specifically, how driving volatility was correlated with driver, roadway, and vehicle factors was quantified using hierarchical models. The study further assessed and compared driving performance across geographic regions in the US.
4) To provide driver feedback (alerts, warnings, and control assists), the study provides applications for safety monitoring under connected vehicle

environments, extracting useful information from Vehicle-to-Vehicle and Vehicle-to-Infrastructure communications.

This project makes both theoretical and empirical contributions by: 1) developing measures to characterizing instantaneous driving behaviors using high resolution sensor data, which could be a key component of the future Big Data analytics in transportation safety; 2) providing examples of structuring and integrating safety data from multiple sources and delivering applications for safety assessment, monitoring and improvement. Figure 1 shows the overall organization of the report and highlights the efforts undertaken in the first year. Big data visualization was part of several efforts, e.g., driving trajectories, instantaneous driving time use, acceleration and vehicular jerk distributions are visualized in two- and three-dimensions.

The report first discusses the appropriateness of sensor data that can be used for transportation safety studies, in terms of information loss when sampling sensor data. "Undersampling" can cause loss of information and misinterpretations of the data, but "oversampling" can waste storage and processing resources. Data from a driving simulator study collected at 20 Hertz are analyzed (N=718,481 data points from 35,924 seconds of driving tests). The results show that marginally more information is lost as data are sampled down from 20 Hz to 0.5 Hz. However, the relationship between loss of information and sampling rates is non-linear. The study provides a sound basis to help scientists easily identify data needs at the experimental design stage, and it has implications for designing monitoring systems.

Second, an innovative way of characterizing driving behaviors using large-scale sensor data is provided. This effort contributes by leveraging a large-scale behavioral database to analyze short-term driving decisions and develop a new driver volatility index to measure the extent of variations in driving. The index captures variations in instantaneous driving behavior constrained by the performance of the vehicle from a decision-making perspective. Specifically, instantaneous driving decisions include maintaining speed, accelerating, decelerating, maintaining acceleration/deceleration, or jerks to vehicle, i.e., the decision to change marginal rate of acceleration or deceleration.

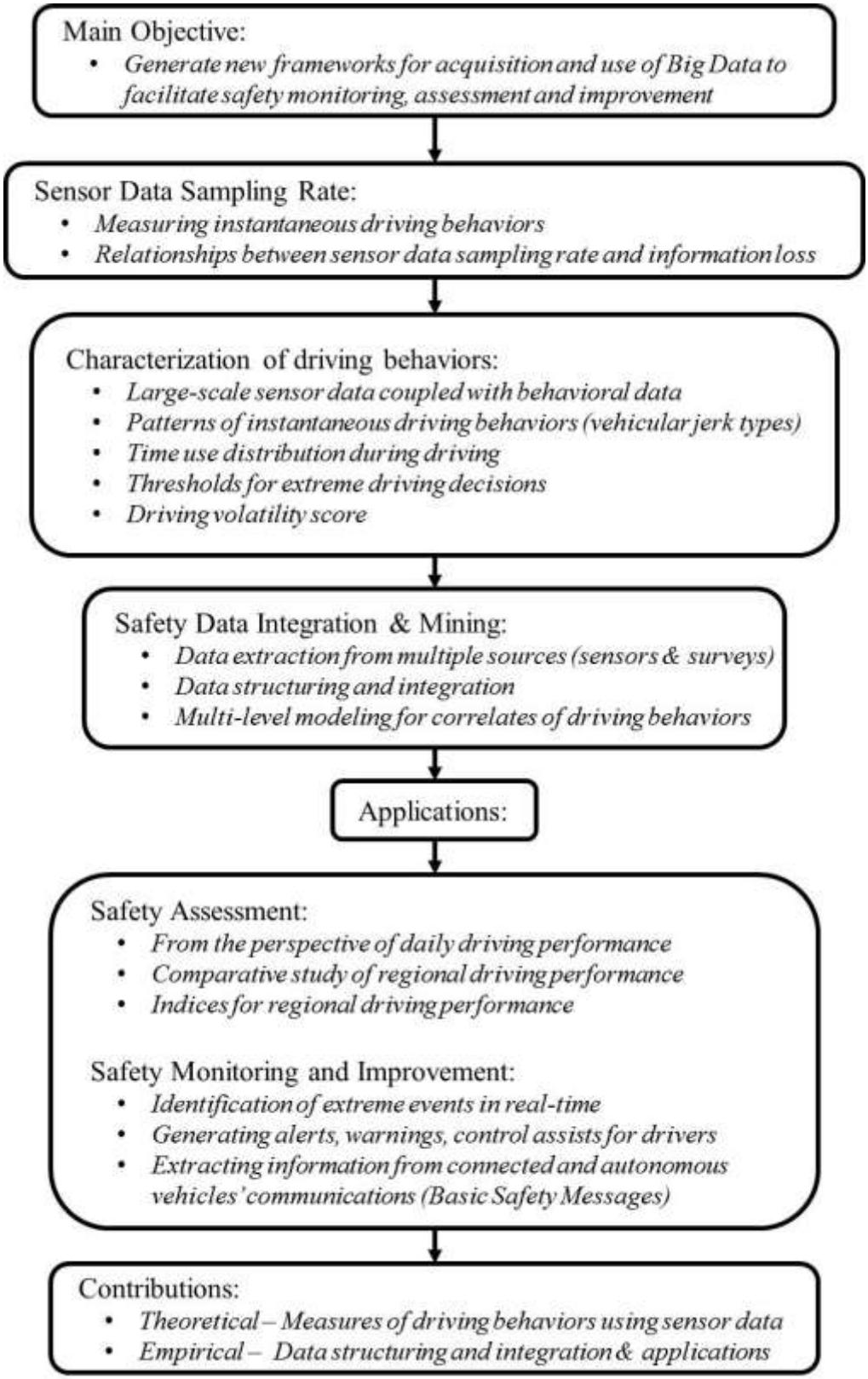


FIGURE 1 Report outline

Volatility in driving decisions, captured by jerky movements, is quantified using data collected in Atlanta, GA during 2011. The database contains 51,370 trips and their associated second-by-second speed data, totaling 36 million seconds. Rigorous statistical models explore correlates of volatility that include socioeconomic variables, travel context variables, and vehicle types.

Third, ways of extracting, structuring and integrating data from multiple sources are explored. Under this effort, we demonstrate creation of a unique database by integrating data from four seemingly disparate sources that include two large-scale travel surveys, historical traffic counts from California and Georgia Department of Transportation, socio-demographic information from Census, and geographic information from Google Earth. The database provides a rich resource to test hypothesis and model driving behaviors at the micro-level, i.e., second-by-second. The data include 117,022 trips made by 4,560 drivers residing in 78 counties of major metropolitan areas (Los Angeles, San Francisco, Sacramento, and Atlanta) across two states, representing various land use types and populations; all trips were recorded by in-vehicle GPS devices giving 90,759,197 second-by-second speed records. Appropriate multi-level models are estimated to extract valuable information from the data and study correlates of driving behaviors structured in hierarchies and compare driving performance across geographical regions.

Fourth, applications for safety monitoring under connected environments are developed. This effort is featured by exploring information embedded in basic safety messages (BSMs) transmitted between connected vehicles, and developing applications for delivering improved alerts, warnings, and control assistance using BSMs. A data analytic methodology extracts critical information from raw BSM data available from Safety Pilot Model Deployment (SPMD) underway in Ann Arbor, Michigan. The information extracted from BSM data captures 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. Thus, the study creates a framework for generating alerts, warnings, and control assistance from extreme events, transmittable through Vehicle-to-Vehicle and Vehicle-to-Infrastructure (or V2V and V2I) applications.

Research papers in refereed journals and conference presentations based on the results of efforts discussed above contributed directly to knowledge creation in the context of big data applications in safety. The presentations and publications related to this project include the following (with relevance of specific papers shown as percentage):

1. Liu J., A. Khattak & L. Han. How Much Information is Lost When Sampling Driving Behavior Data? TRB paper # 15-0968. Presented at the Transportation Research Board Annual Meeting, National Academies, Washington, D.C., 2015. (100%)
2. Wang X, A. Khattak, J. Liu, G. Masghati-Amoli & S. Son. What is the Level of Volatility in Instantaneous Driving Decisions? Forthcoming paper in Transportation Research Part C: Emerging Technologies, 2015. (50%)
3. Liu J., A. Khattak & X. Wang. Creating Indices for How People Drive in a Region: A Comparative Study of Driving Performance, TRB paper # 15-0966. Presented at the Transportation Research Board Annual Meeting, National Academies, Washington, D.C., 2015. (50%)
4. Khattak A., J. Liu & X. Wang. Supporting Instantaneous Driving Decisions through Vehicle Trajectory Data, TRB paper # 15-1345. Presented at the Transportation Research Board Annual Meeting, National Academies, Washington, D.C., 2015. (50%)
5. Liu J., X. Wang & A. Khattak. Generating Real-Time Volatility Information, Presented at 2014 Intelligent Transportation Systems World Congress, Detroit, MI, 2014. (50%)
6. Liu J. & A. Khattak. Improved Warning and Assistance Information from Connected Vehicle Basic Safety Messages, Accepted for presentation to 2015 Intelligent Transportation Systems World Congress, Bordeaux, France, 2015. (100%)
7. Liu J. A. Khattak, & M. Zhang, Exploring Links between Naturalistic Driving Behaviors and Various Factors in Hierarchies: A Study Integrating Multiple Data Sources, To be presented at 2015 Road Safety & Simulation International Conference, Orlando, FL. 2015. (100%)
8. Khattak A., & J. Liu, Transportation Data Needs for Making Transportation Decisions, Submitted to 2016 Transportation Research Board for review. (20%)

9. Liu J., A. Khattak & M. Zhang, Structuring and Integrating Data in Metropolitan Regions to Explore Multi-Level Links between Driving Volatility and Correlates, Submitted to 2016 Transportation Research Board for review. (100%)
10. Liu J. & A. Khattak, Delivering Improved Alerts, Warnings, and Control Assistance Using Basic Safety Messages Transmitted between Connected Vehicles, Submitted to 2016 Transportation Research Board for review. (100%)

This report synthesizes key papers and presentations that are relevant to this project.

HOW MUCH INFORMATION IS LOST WHEN SAMPLING DRIVING BEHAVIOR DATA COLLECTED FROM ELECTRONIC SENSORS?[1]

Abstract – Individuals' driving behavior data are becoming available widely through electronic sensors, such as Global Positioning System devices. These data can be used to make accurate estimates of vehicle fuel consumption, emissions, and safe driving. Storage and computing power have become readily available to the extent that scientists and engineers are presented with a wide range of options for balancing resource cost versus amount of data that needs to be stored. The incoming data can be sampled at rates ranging from one Hertz (or even lower) to hundreds of Hertz, i.e., one data point per second to hundreds of data points per second. Failing to capture substantial changes in vehicle movements over time by "undersampling" can cause loss of information and misinterpretations of the data, but "oversampling" can waste storage and processing resources. Empirical assessment of driving data is necessary because real-world vehicular movements are difficult to characterize mathematically and they vary substantially over time. A key objective of this study is to empirically explore how micro driving decisions to maintain speed, accelerate or decelerate, or change marginal rate of acceleration (known as vehicular jerk) can be best captured, without substantial loss of information. A framework for measuring information loss using several measures that are combined into an overall index is developed. Data from a driving simulator study collected at 20 Hertz are analyzed (N=718,481 data points from 35,924 seconds of driving tests). The results show that marginally more information is lost as data are sampled down from 20 Hz to 0.5 Hz. However, the relationship between loss of information and sampling rates is non-linear. The study provides a sound basis to help scientists easily identify data needs at the experimental design stage, and it has implications for designing monitoring systems.

Keywords: information loss, instantaneous driving decisions, sampling rate, undersampling

[1] Material in this section is based on: Liu J., A. Khattak & L. Han. How Much Information is Lost When Sampling Driving Behavior Data? TRB paper # 15-0968. Presented at the Transportation Research Board Annual Meeting, National Academies, Washington, D.C., 2015.

INTRODUCTION

Increasingly detailed driving data are being collected with well-developed data acquisition technologies, such as Global Positioning System (GPS), video, Bluetooth, and on-board diagnostics. With the increasing amount of data from sensors, digging through detailed transportation data helps explore micro-level driver behaviors that were not possible until fairly recently. Instantaneous driving decisions are of particular interest, because they are related to energy consumption, emissions and safety. They include accelerating, decelerating, maintaining speed, altering acceleration/deceleration, etc. Driving reflects a chain of instantaneous driving decisions made by drivers according to changes in surrounding circumstances, e.g., adjacent vehicles, roadway conditions, and geometric changes in the roadway, and weather conditions (*1*). The higher rate sampled data can capture more information about the instantaneous driving decisions. Current data collection in industry can go as high as 800 MHz (*2*) and it can contain valuable information (*3*). One question is that, whether driving data need to be sampled by such high rates in the transportation context. High sampling rates can be expensive in terms of requiring extra storage and processing time, which is called oversampling (*4*). Undersampling/inadequate sampling may cause loss of critical information (*3*). Next Generation Simulation Program (NGSIM) collected detailed vehicle trajectory data in 10 Hz to develop behavioral algorithms in support of traffic simulation on microscopic modeling (*5*). One problem for data sampled by high sampling rates is the data accuracy. The accuracy of NGSIM data is estimated at 2~4 ft. (*6*). For NGSIM data, in 0.1 second, the distance travelled by a 60 mph vehicle is about 8.8 ft. but with a 2~4 ft. error. Therefore, the accuracy of NGSIM data might be jeopardized with high sampling rates. Jackson et al., discussed the validity of using in-vehicle GPS second-by-second (1 Hz) velocity data to track the 1-second driving operation modes, including acceleration and deceleration. Their results imply that the 1-second operation modes can be successfully measured by using GPS data sampled by 1 Hz (*7*), while the driving operation modes within 1-second are unknown. For example, if a driving command –"acceleration →deceleration→acceleration" occurs within one second, the 1 Hz sampled data may lose the information about the deceleration, though the deceleration exists in a very short time. Thus,

another question is how much information we may lose if we only sampled data by 1 Hz or even lower rates. Current driving data are usually continuously sampled by rates from 0.2 to 10 Hz (*8-16*). Note that the continuous driving data are different from the traffic data collected by loop detectors (*17, 18*). The focus of this study is the continuous driving data used to explore micro-driving behavior. The key question to be answered is what sampling rates are appropriate to capture micro-driving behavior without losing much information (i.e., by undersampling).

In the field of signal processing, Nyquist–Shannon sampling theorem gives the appropriate sampling rates for continuous signal. The Nyquist criterion for sampling rates is twice the bandwidth of a bandlimited signal or a bandlimited channel. The key question is to find out the bandwidth of a signal (*19*). However, the driving behavior does not fulfill the features of bandlimited signal. Driving behavior varies according to the decisions a driver makes to respond the instantaneous driving circumstances. This study aims to find out the appropriate sampling rates for driving behavior data through exploring the nature of driver's micro-driving behavior.

DATA DESCRIPTION

Data used in this study comes from the University of Tennessee Driving Simulator Lab (DSL). This driving simulator, Drive Safety DS-600c, is fully integrated and immersive to driving test subjects with its visual and audio effects in the front half cab of a Ford Focus sedan and it provides 300° horizontal field-of-view via five projectors and back sight via three rear mirror liquid crystal display displays (*20*). The cab base is able to mimic pitch and 30 longitudinal motions. Since 2009, over 10 simulator studies have been conducted in DSL. The equipment has been recognized as a high-fidelity driving simulator and is qualified to be used to conduct driving behaviors associated research. The data of driver responses (e.g. speed) gathered from simulator driving tests can be used as surrogate measures of driving behavior (*21*, *22*). The driving data used in this study was collected from 24 subjects (13 males, 11 females, average licensed year – 17.6, standard deviation –7.87). Subjects were tested in a simulated driving scenario designed with various driving conditions, e.g., urban vs. rural environments. Each subject completed the driving test in 22 ~ 29 minutes,

depending on their travel speed and responses to traffic controls. The driving speed was sampled at 20 Hz. The final dataset used in this study includes 718,481 data points from 35,924 seconds (598 minutes) of driving tests.

METHODOLOGY

A fundamental question is how much information is lost in going to lower sampling rates? Driving can be volatile as drivers made driving decisions (e.g., accelerating and braking) according to the instantaneous changes of surrounding circumstances, e.g., adjacent vehicles, roadway conditions, geometric changes in the roadway, and weather conditions (*1*). Using the 20-Hz simulator driving data, this study creates a set of measures to quantify the magnitude of information loss (*MIL*):

a) MIL_1: Instantaneous driving decision loss (based on combined direct and indirect 'detectability' explained below) – Equations 1, 2, 3;

b) MIL_2: Percentage of out-of-range observations during driving– Equation 4;

c) MIL_3: Ratio of sampled to actual range in driving data– Equation 5;

d) MIL_4: Relative speed deviation from linear interpolation of under-sampled data (based on observed speed deviation over the under-sampled data) – Equations 6 and 7.

An index named *Extent of information loss (EIL)*, given a sampling rate is created and it is shown in Equation 8. The overall methodological framework for this study is shown in Figure 2 and explained in more detail below. Each measure is calculated as a percentage in order to index the extent of information loss in different situations. The *Extent of Information Loss (EIL)* is an overall measure of information loss that combines the above measures. The study quantifies the relationship between information loss measures and sampling rates. A user can then select thresholds, e.g., 5% or 1% of information loss may be acceptable and find the appropriate sampling rate.

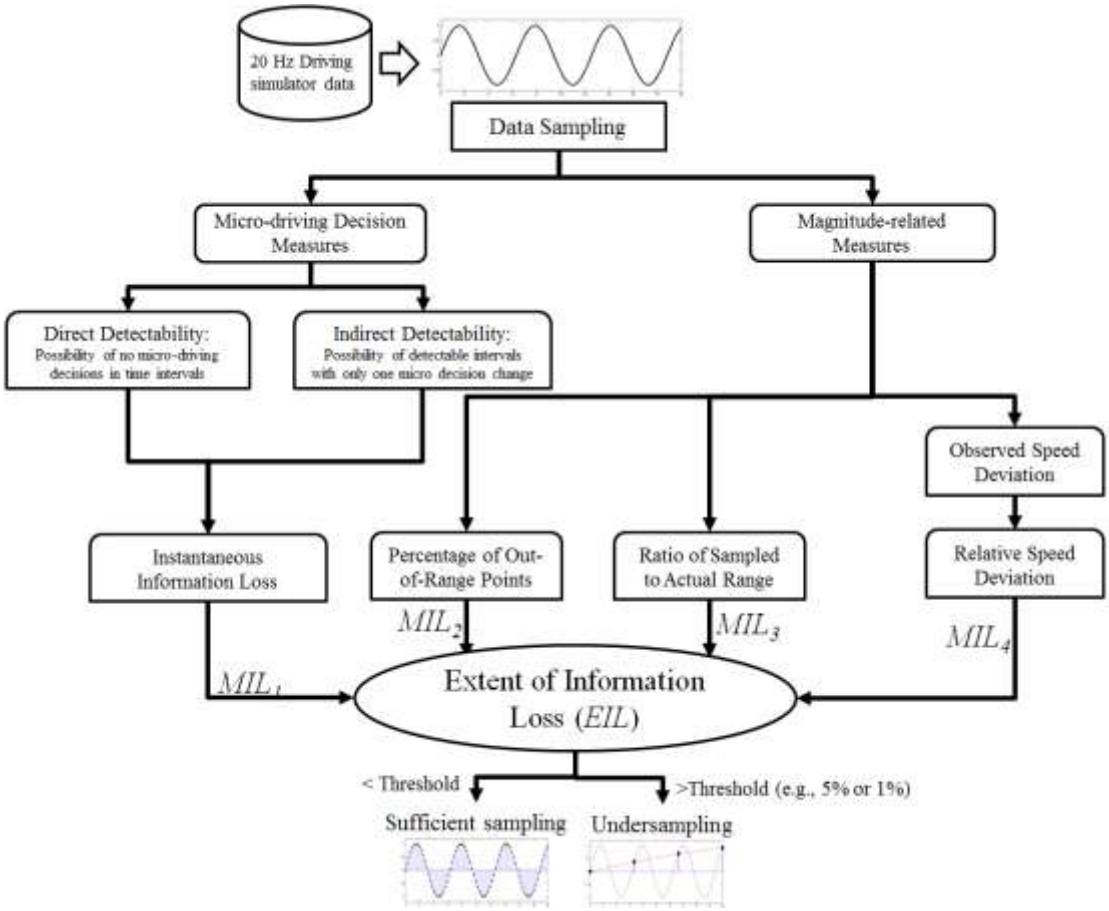


FIGURE 2 Study steps and measures

Direct Detectability of Driving Decisions

Driving decisions can be altered at any time and frequently when a vehicle is being operated. If the frequency of the driving decision alteration is considerably high and the data sampling rate is very low, then some driving decisions may be lost. As shown in Figure 3(i), the decision alteration– "acceleration to deceleration" between *n* and *n+1* second is missed by the 1-Hz sampled data (red points), as the speeds at *n* and *n+1* second are identical. In this case, undersampling causes information loss of micro driving decisions. The information about going from "acceleration to deceleration" between *n* and *n+1* second is lost, while the information –"deceleration" or "no decision alternation" between *n+1* and *n+2* second is detected directly by the sampled data.

This study uses the 20-Hz simulator driving data to count the number of decisions made given a specific time interval, and then computes the possibility of no decision made cases, termed *Direct Detectability of Driving Decisions*. The formula is as follows:

$$Direct\ Detectability = \frac{1}{N}\sum_{i=1}^{N} w_i^0 \qquad Equation\ (1)$$

Where,

$N = T \times f$, the number of time slices during total data duration T in second;

f = target sampling frequency/rates, e.g., 1 Hz;

$w_i^0 = \begin{cases} 1, if\ \max\{v_{ij} - v_{i(j-1)}\} \times \min\{v_{ij} - v_{i(j-1)}\} \geq 0, \\ 0, if\ \max\{v_{ij} - v_{i(j-1)}\} \times \min\{v_{ij} - v_{i(j-1)}\} < 0, \end{cases}$, indicator for micro-driving decision alternation during i^{th} time interval $t = \frac{1}{f}$, i= 1, 2, 3, …, N;

v_{ij} = Speed at j^{th} location in i^{th} time interval, j=1, 2, 3, …, n;

$n = \frac{T}{N} = \frac{F}{f}$, number of available data points in a given time interval;

F = sampling rate of original dataset, 20 Hz in this study.

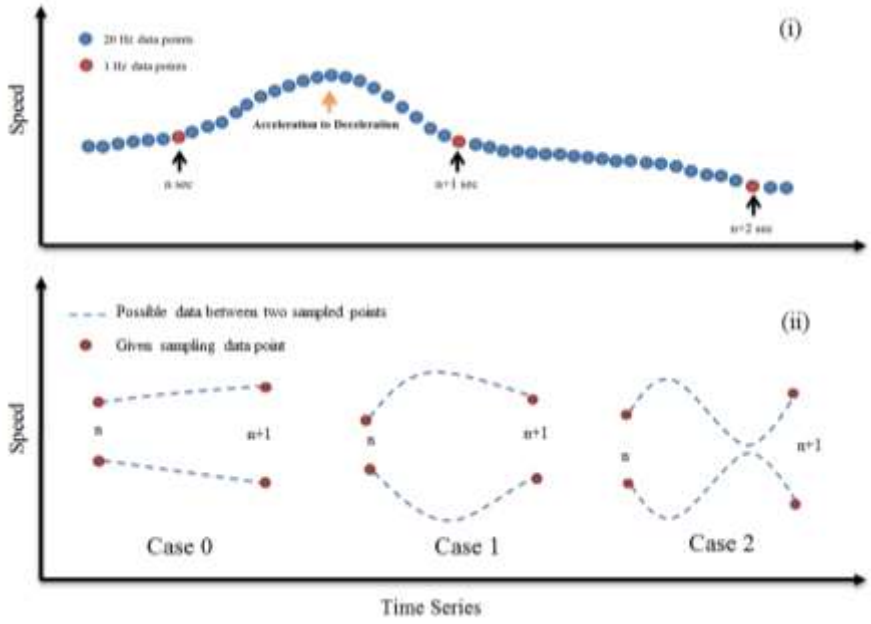


FIGURE 3 Example of information loss in instantaneous driving decisions

In this study, time intervals without decisions made belongs to Case 0 (this includes constant acceleration or deceleration), as shown in Figure 3(ii), with one micro-decision made are referred to as Case 1, and with two decision alternations are referred to as Case 2. Case 1 will be further discussed below.

Indirect Detectability of Driving Decisions

Direct detectability tells the chance of detecting micro driving decisions directly with the sampled data. Next, this study discusses the chance of detecting driving decisions in Case 1. An assumption is made before we discuss the indirect detectability. We assume that driving speed is a continuous changing measurement without sharp changes. A sine wave illustrates the example of continuous changing measures, while square wave and sawtooth wave are examples of sharp changes (*23*).

With this assumption, using 20-Hz data, this study takes one second interval (corresponding to 1-Hz sampling rate) as the example for illustrating detection of driving decision alternation. Figure 4(i) presents four possible types of micro driving behavior of Case 1 within one second. Types (a) and(c) show that there is a micro-decision made from accelerating to decelerating between *n* and *n+1* second. Types (b) and (d) show that there is a micro-decision made from decelerating to accelerating between *n* and *n+1* second.

For Type (a), there is a micro-decision made from accelerating to decelerating between *n* and *n+1* second, while the speed measurement at *n* and *n+1* second implies a deceleration during that second. Therefore, the missing micro-decision made within this second could be observed by using given sampling data points at *n* and *n+1* second, though the amount/intensity of the driving decision change is not necessarily accurate. In the same fashion, Type (b) illustrates information detection for the micro-decision made from decelerating to accelerating. Therefore, for Types (a) and (b), the micro-decision change can be detected but with an error.

Types (c) and (d) do not meet the situations in Types (a) and (b), since the sampled data do not show the correct micro-decision made between two sampled observations. Types (c) and (d) also include the cases that speed at n second is equal to n+1 second, shown in

Figure 4(i), since in these cases, the sampled observations can also not tell the micro-decision correctly.

Therefore, we move our sight to next second, as shown in Figure 4(ii). In Type (c_1), the sampled speeds at *n+1* and *n+2* second give a deceleration which uncovers the lost micro-decision made between *n* and *n+1* second, but with a temporal error. The time stamped for the micro-decision using sampled data is at n+1 second, but actually it occurred between *n* and *n+1* second. Type (d_1) is similar to Type (c_1), but for detecting a micro-decision from decelerating to accelerating.

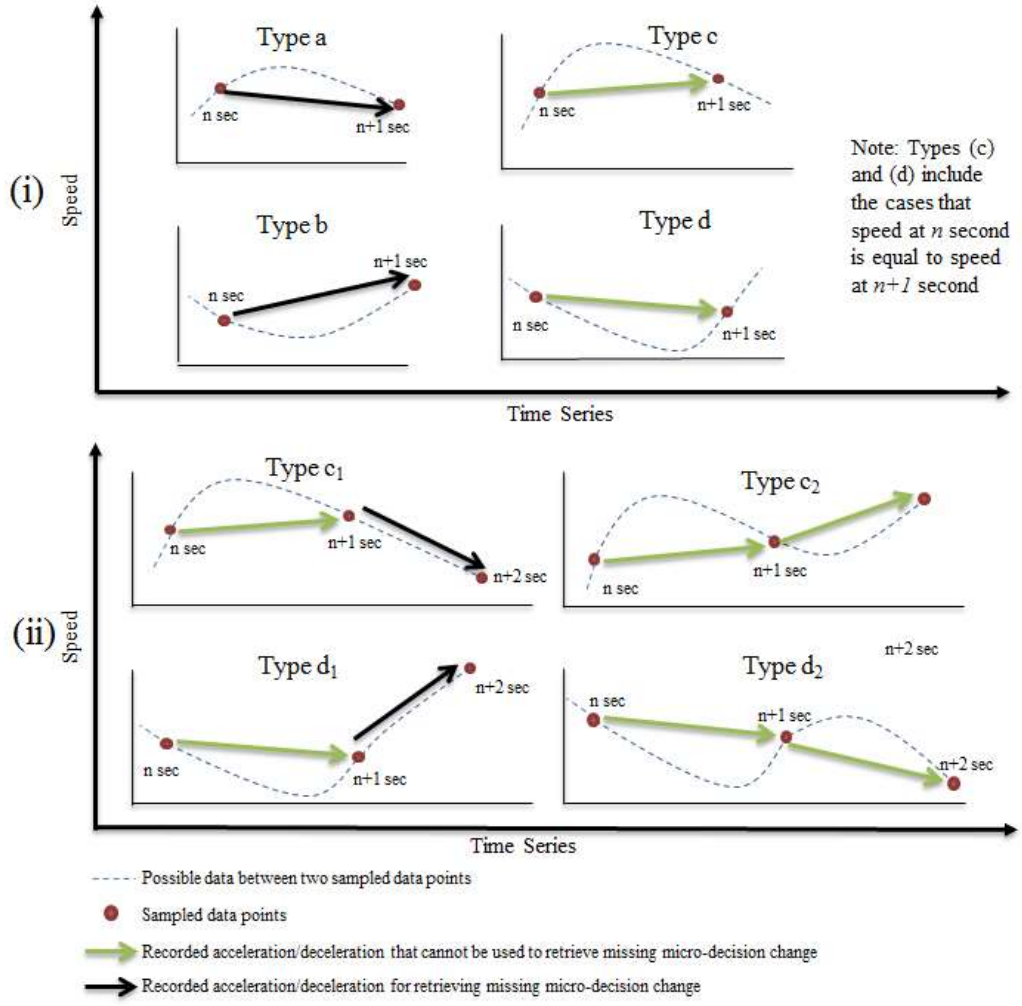


FIGURE 4 Examples of missing information when examining speed data over time

Types (c_2) and (d_2) illustrate these two types that the micro decision made between two sampled observations cannot be detected, since there are two micro-decisions made in two sequential time intervals. Besides, for cases with two or more micro-decisions made within one particular time interval, there is no way to detect them by above methods. This study mainly discusses Case1 with one micro-decision made and tries to find the possibilities of having Types (a), (b), (c_1) and (d_1) in Case 1 given a time interval. The measure, *Indirect Detectability of Driving Decisions*, is the sum of the possibilities of having Types (a), (b), (c_1) and (d_1).

The formula is as follows:

$$Indirect\ Detectability = \frac{1}{\sum_{i=1}^{N} d_i^1} \sum_{i=1}^{N} (w_i^a + w_i^b + w_i^{c_1} + w_i^{d_1}) \qquad Equation\ (2)$$

Where,

$N = T \times f$, the number of time slices during the total data duration T in second;

f = target sampling frequency/rates, e.g, 1 Hz;

$w_i^1 = \begin{cases} 1, if\ \sum_{j=1}^{n-1} z_j = 1 \\ 0, if\ \sum_{j=1}^{n-1} z_j \neq 1 \end{cases}$, indicator for whether there is only one decision change during i^{th} time interval $t = \frac{1}{f}$, i= 1, 2, 3, …, N;

$z_j = \begin{cases} 1, if\ (v_{ij} - v_{i(j-1)}) \times (v_{i(j+1)} - v_{ij}) < 0 \\ 0, if\ (v_{ij} - v_{i(j-1)}) \times (v_{i(j+1)} - v_{ij}) \geq 0 \end{cases}$, indicator for whether two consecutive driving statuses are both acceleration or deceleration;

v_{ij} = Speed at j^{th} location in i^{th} time interval, j=1, 2, 3, …, n;

$n = \frac{T}{N} = \frac{F}{f}$, the number of available data points in a given time interval;

F = sampling rate of original dataset, 20 Hz in this study.

$w_i^a = \begin{cases} 1, if\ d_i^1 = 1\ and\ (v_{ij} - v_{i(j-1)}) > 0\ and\ (v_{i(j+n)} - v_{i(j+n-1)}) < 0\ and\ v_{ij} > v_{i(j+n)} \\ 0 \end{cases}$,

indicator for Type (a) error;

$w_i^b = \begin{cases} 1, if\ d_i^1 = 1\ and\ (v_{ij} - v_{i(j-1)}) < 0\ and\ (v_{i(j+n)} - v_{i(j+n-1)}) > 0\ and\ v_{ij} < v_{i(j+n)} \\ 0 \end{cases}$,

indicator for Type (b) error.

$w_i^c =$

$$\begin{cases} 1, if\ d_i^1 = 1\ and\ d_{i+1}^0 = 1\ and\ (v_{ij} - v_{i(j-1)}) > 0\ and\ (v_{i(j+n)} - v_{i(j+n-1)}) < 0\ and, \\ \quad v_{ij} < v_{i(j+n)} \\ 0 \end{cases}$$

indicator for Type (c_1) error;

$w_i^d =$

$$\begin{cases} 1, if\ d_i^1 = 1\ and\ d_{i+1}^0 = 1\ and\ (v_{ij} - v_{i(j-1)}) < 0\ and\ (v_{i(j+n)} - v_{i(j+n-1)}) > 0\ and, \\ \quad v_{ij} > v_{i(j+n)} \\ 0 \end{cases}$$

indicator for Type (d_1) error;

Instantaneous Driving Decision Loss

With the direct and indirect detectability of driving decisions, we can detect micro-driving decision made given a particular sampling rate. The formula for instantaneous driving decision loss (MIL_1) is as follows:

$$Decision\ Loss = 1 - (Direct\ Detectability + \frac{1}{N}\sum_{i=1}^{N} d_i^1\ \times Indirect\ Dectetability)$$

$$Equation\ (3)$$

Empirical results are shown later. Theoretically, higher sampling rates lower the possibility of missing critical decisions, but they increase the possibility of "noise" in the data and the data storage and processing requirements. The challenge is to not lose decision information while reducing the noise in the data.

Measures Concerning Magnitudes

It is important to know whether sampled values represent the population and the magnitude of errors, if any. In other words, whether the one point (e.g., 1 Hz data) can represent the 20 data points (20 Hz data) during the same second? If the 20 data points provide only marginally more information (such as constant speed during one second), one data point might be sufficient for sampling this second.

Figure 5(i) shows an example using 20 Hz simulator data, along with two 1-Hz sampled points at the *n* and *n+1* second. The speed is 10 mph at *n* second and 12 mph at *n+1* second. The problem would be whether all speed values between *n* and *n+1* second are

within the micro speed range 10~12 mph. The example shows given one-second time interval, there are six data points, or 30% (6 out of 20) data points with speed values out of range 10~12 mph. In this case, two data points with records of 10 and 12 mph cannot fairly represent the driving behavior from *n* to *n +1* second. The *Percentage of Out-of-Range observation* (*MIL*$_2$) is a measure that captures how many data points are out of the sampled micro speed range.

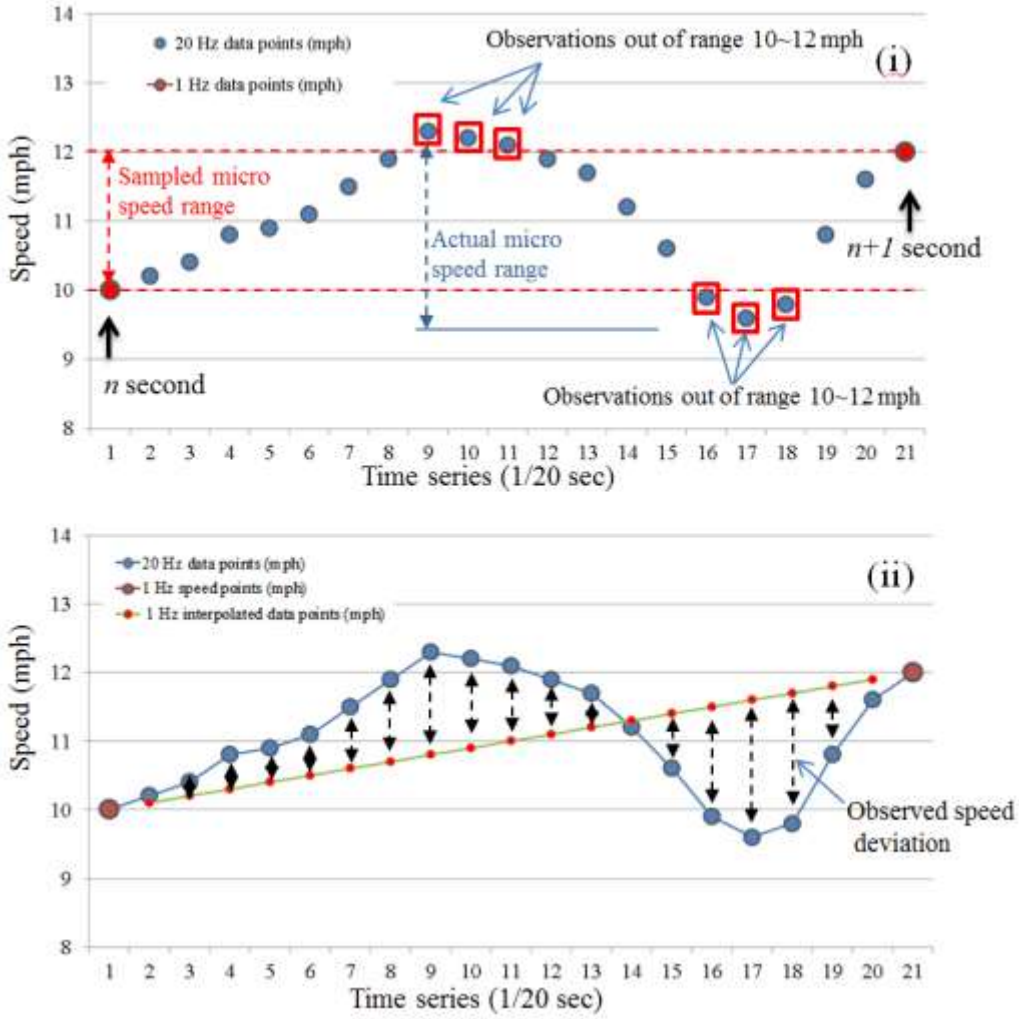


FIGURE 5 Quantifying magnitude errors in sampled data

The formula for *Percentage of Out-of-Range Observation* (*MIL*$_2$) is:

$$Percentage\ of\ Out\ Range\ Observations = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{n} OR_{ij}}{n} \qquad Equation\ (4)$$

Where,

$OR_{ij} = \begin{cases} 1, if\ v_{ij} > \max\{v_{i1,}v_{in}\}\ or\ v_{ij} < \min\{v_{i1,}v_{in}\} \\ 0 \end{cases}$, indicator for out-of-rang observation.

The ratio of sampled micro speed range over actual micro speed range during the same second is another measure of information loss and it is termed *Ratio of sampled to Actual Range* (*MIL*$_3$). In the example, the sampled micro speed range is 12-10=2 mph, while the actual micro speed range is 12.3-9.6=2.7 mph. The ratio is 2/2.7=0.74, or 74%. The formula is as follows:

$$Ratio\ of\ Sampled\ to\ Actual\ Range = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i^{Sampled}}{R_i^{Actual}} \qquad Equation\ (5)$$

Where,

$R_i^{Sampled} = |v_{i1} - v_{(i+1)1}|$, sampled speed range for i^{th} time slice;

$R_i^{Actual} = \max\{v_{ij}\} - \min\{v_{ij}\}$, actual speed range for i^{th} time slice.

A measure of information loss is through speed deviations. The deviations are measured based on the linear distance between observed speeds and sampled speeds. Sampled data can be used to linearly interpolate the data points in between two timestamps. This can be compared with observed data at higher frequency (20 Hz in this case). Figure 5(ii) uses 20 Hz driving simulator data and measures *Observed Speed Deviation*, which is the mean of absolute deviations within time intervals. Another measure is *Relative Speed Deviation* (*MIL*$_4$), which is the average deviations over interpolated speed values, providing the extent of deviations. The formulas are as follows:

$$Observed\ Speed\ Deviation = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n}\sum_{j=1}^{n}\left|v_{ij} - j \times \frac{v_{i1} - v_{i(n+1)}}{n}\right|\right) \qquad Equation\ (6)$$

$$Relative\ Speed\ Deviation = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\left|v_{ij} - j \times \frac{v_{i1} - v_{i(n+1)}}{n}\right|}{v_{ij}}\right) \qquad Equation\ (7)$$

An Index for Magnitude of Information Loss (MIL)

The *Instantaneous Driving Decision Loss*, *Percentage of Out-of-Range Observation*, *Ratio of Sampled to Actual Range*, and *Relative Speed Deviation* quantify the magnitude of information loss from different angles. All these measures are finally calculated in terms of percentage of information loss. Then, these measures can be combined (weighted equally) to create an index capturing the *Extent of Information Loss Index*, given a sampling rate. The formula is as follows:

$$Extent\ of\ Information\ Loss\ Index = \frac{MIL_1 + MIL_2 + (1 - MIL_3) + MIL_4}{4} \qquad Equation\ (8)$$

Where,

MIL_1= Instantaneous driving decision loss;

MIL_2= Percentage of out-of-range observations;

MIL_3= Ratio of sampled to actual range;

MIL_4= Relative speed deviation.

Users of data in the transportation context can either choose a threshold for information loss and find the appropriate sampling rate or vice versa.

RESULTS

Direct Detectability of Driving Decisions

To capture alternations between acceleration and deceleration within the given time interval (e.g., 1 second) corresponding to a sampling rate (e.g., 1 Hz), the number of alternations was counted by using 20 Hz data. All possible alternations within the data, given different time intervals and starting locations were counted. If all decisions made occur exactly at the sampled points, no information will be lost. For example in Figure 1, if the data was just sampled at *n+0.5* second and *n+1.5* second instead of *n* and *n+1* second, then the driving decisions from accelerating to decelerating can be detected accurately, even if the data are still sampled at 1 Hz. The example in Figure 1 shows that there are 20 possible locations to start sampling the 1 Hz data.

Figure 6(i) presents the direct detectability, possibility of no decision made, given a specific time interval, and 5(ii) presents the distribution of the possibilities of the three Cases (discussed above) in different time intervals. In Figure 6(i), the maximum and minimum detectability is also indicated, according to observations from the different sampling locations. For short time intervals, the location does not have a significant influence on the data sampling. Specifically, for time interval of 1 second, the direct detectability is around 89.90%, i.e., no micro decision made during one second intervals. The reason is probably related to the driver reaction time, which is usually more than 1 second (*24*). Thus, there is a large possibility that drivers do not make decisions during one second (N= 35,924 intervals out of 20-Hz sampled data).

In Figure 6(ii), the percentages of possibilities of the three Cases (i.e., no decision, one decision and two and more decisions made within the sample interval) are provided. Shorter time intervals (higher sampling rates) are related to the lower information loss in terms of instantaneous driving decisions, as expected.

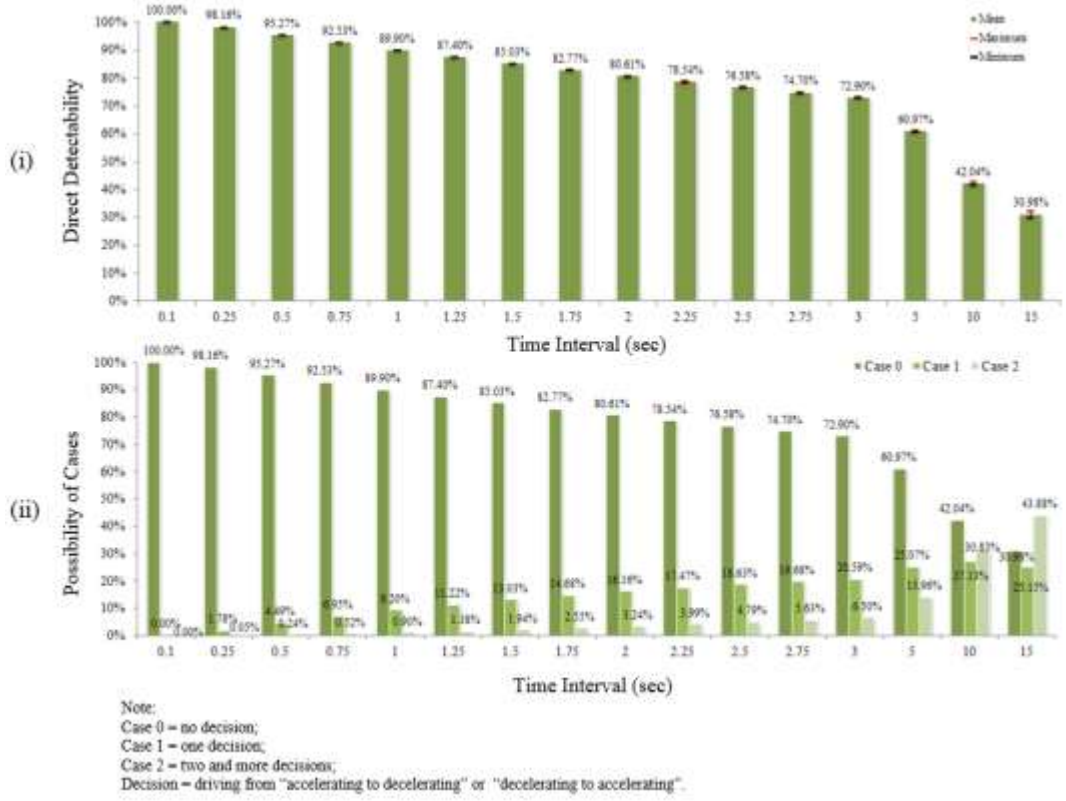


FIGURE 6 "Direct detectability" in different time intervals

Indirect Detectability of Driving Decisions

Figure 7(i) shows percentages of Types (a), (b), (c_1) and (d_1) in Case 1 (one decision change). Specifically, given a one second time interval (or 1-Hz sampling rate), Types (a), (b), (c_1) and (d_1) constitute 30.99%, 25.37%, 24.42% and 16.14% of the Case 1 where only one micro-decision made between two sampled data points. These four types of patterns contain detectable driving information. The indirect detectability is the sum of these possibilities, shown in Figure 7(b). For one second time interval (or 1-Hz sampling rate), the indirect detectability is around 30.99%+25.37%+24.42%+16.14%=93.92%. With the time interval getting longer, this indirect detectability decreases.

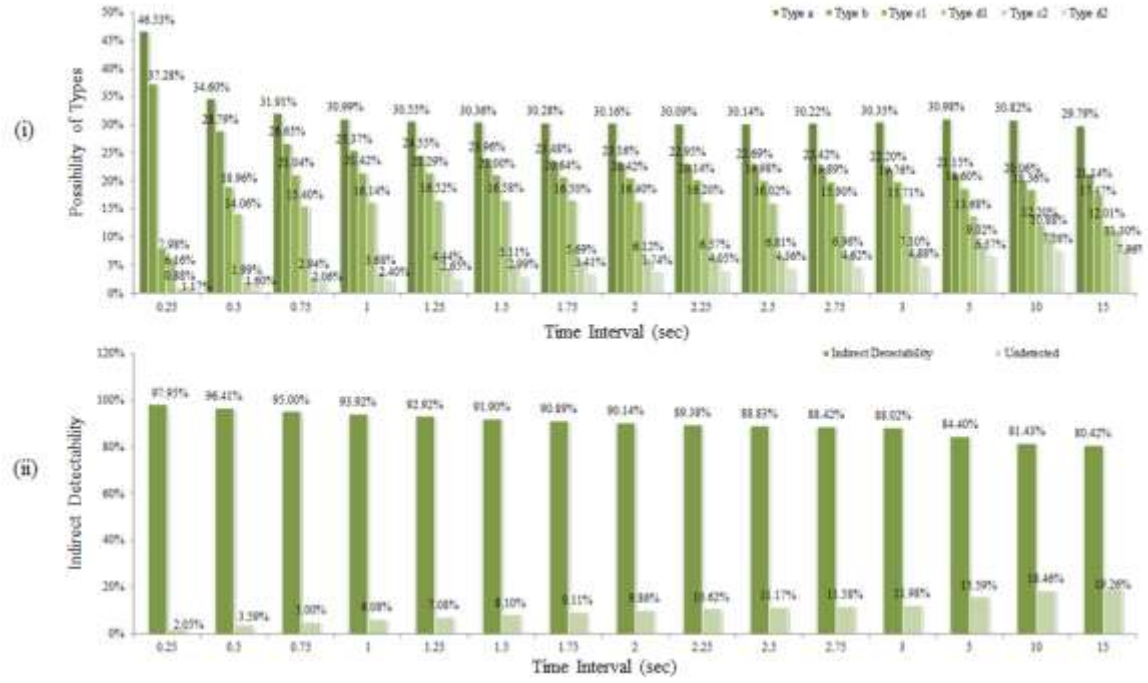


FIGURE 7 Indirect detectability in different time intervals

Instantaneous Driving Decision Information Loss

The combined results of instantaneous driving decision loss are shown in Table 1. There is an 89.90% chance that there is no micro-decision (Case 0) within one second (1-Hz sampling data, highlighted in Table 1) and 9.20% chance that there is one micro-decision (Case 1). For Case 1 with only one micro-decision, there is a 30.99% chance that the Type (a) decision pattern would occur, and 25.37%, 24.42% and 16.14% for Types (b), (c) and (d) respectively.

These four types include micro-decisions that can be detected. Therefore, in summary, the feasibility of detecting micro-driving decisions for 1 Hz sampling data are 89.90% + 9.20% × (30.99% + 25.37% + 24.42% + 16.14%) = 98.54%, and 1.46% of information about micro-decisions would be lost. Data sampled by rates higher than 0.5 Hz can reflect more than 95% of micro-decisions and the instantaneous driving decision loss is less than 5%.

TABLE 1 Instantaneous Driving Decisions Information Loss

Sampling Rate (Hz)	Time Interval (second)	Percentage of total sample			Percentage of Case 1						Feasibility of detecting micro-decisions	Instantaneous driving decision lost
		Case 0	Case 1	Case 2	Type a	Type b	Type c_1	Type d_1	Type c_2	Type d_2		
10	0.1	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00% ^
4	0.25	98.16%	1.78%	0.05%	46.53%	37.28%	7.98%	6.16%	0.88%	1.17%	99.91%	0.09%
2	0.5	95.27%	4.49%	0.24%	34.60%	28.79%	18.96%	14.06%	1.99%	1.60%	99.60%	0.40%
1.333	0.75	92.53%	6.95%	0.52%	31.91%	26.65%	21.04%	15.40%	2.94%	2.06%	99.13%	0.87%
1	1	89.90%	9.20%	0.90%	30.99%	25.37%	21.42%	16.14%	3.68%	2.40%	98.54%	1.46%
0.8	1.25	87.40%	11.22%	1.38%	30.55%	24.55%	21.29%	16.52%	4.44%	2.65%	97.83%	2.17%
0.667	1.5	85.03%	13.03%	1.94%	30.36%	23.96%	21.00%	16.58%	5.11%	2.99%	97.01%	2.99%
0.571	1.75	82.77%	14.68%	2.55%	30.28%	23.48%	20.64%	16.50%	5.69%	3.41%	96.11%	3.89%
0.5	2	80.61%	16.16%	3.24%	30.16%	23.16%	20.42%	16.40%	6.12%	3.74%	95.17%	4.83%
0.444	2.25	78.54%	17.47%	3.99%	30.09%	22.95%	20.14%	16.20%	6.57%	4.05%	94.16%	5.84%
0.4	2.5	76.58%	18.63%	4.79%	30.14%	22.69%	19.98%	16.02%	6.81%	4.36%	93.13%	6.87%
0.364	2.75	74.70%	19.68%	5.63%	30.22%	22.42%	19.89%	15.90%	6.96%	4.62%	92.10%	7.90%
0.333	3	72.90%	20.59%	6.50%	30.35%	22.20%	19.76%	15.71%	7.10%	4.88%	91.03%	8.97%
0.2	5	60.97%	25.07%	13.96%	30.98%	21.15%	18.60%	13.68%	9.02%	6.57%	82.13%	17.87%
0.1	10	42.04%	27.13%	30.83%	30.82%	20.06%	18.36%	12.20%	10.88%	7.58%	64.14%	35.86%
0.0667	15	30.98%	25.15%	43.88%	29.79%	21.14%	17.47%	12.01%	11.30%	7.96%	51.20%	48.80%

Note: ^Extremely close to 0%.

Measures Concerning Magnitudes

Results in Table 2 show that lower sampling rates (or longer time intervals) are associated with larger percentages of out-of-range points, smaller ratio of sampled to actual range, larger speed deviations and relative speed deviations, as expected. Percentage of out-of-range points concerns the sampled micro speed range within a time interval. The sampled micro speed range is determined by two sequential recorded data points, as shown in Figure 5. The results show that, on average, 1.75 points (or 8.75%) are out of the sampled micro speed range for 1-second time interval (or 1-Hz data), because there is a large possibility that

there is no micro-decision changes during one second. It is consistent with above finding that for the time interval of 1 second, the average possibility of no micro-decision change is 88.90%, see Figure 6. For 1-Hz data, the ratio of sampled to actual micro range is 0.957, which means the extent of representativeness of the 1-Hz data to 20-Hz data is about 95.7% in terms of magnitude. Though some data points are possibly out of the recorded micro ranges, these points do not deviate broadly. Further, 1-Hz data have an observed speed deviation of about 0.076 mph. Note that 1% percentile of 718,481 20-Hz speed records is 0.493 mph, thus the deviation of 0.076 mph is not substantial in the distribution of speed records. This is consistent with EPA drive cycle data, which is based on 10-Hz (*25*). Further, the relative speed deviation, ratio of deviation over interpolated speeds, shows that 1-Hz data has a relative speed deviation to 20-Hz speed records at 0.87%, substantially lower than the 5% threshold.

Extent of Information Loss

The overall extent of information loss is an equally weighted measure, calculated using Equation 8. The results are shown in Table 2. We know if the sampling rate is 1-Hz, the percentage of out-of-range points is 8.77%, ratio of sampled to actual range is 95.71%, relative speed deviation is about 0.87%, and the instantaneous driving decision loss is about 1.46%. So, the overall extent of information loss is (8.77% + (100%-95.71%) + 0.87% + 1.46%)/4 = 3.85%. Thus, overall about 3.85% of the driving information, including the micro-driving decisions and speed magnitude, might be lost if the sampling rate is 1-Hz instead of 20 Hz.

TABLE 2 Overall Magnitude of Information Loss

Sampling Rate (Hz)	Time Interval (second)	Count of out-of-range observations	MIL_2 Percentage of out-of-range observations	MIL_3 Ratio of sampled to actual range	Observed speed deviation (mph)	MIL_4 Relative Speed Deviation	MIL_1 Instantaneous driving decision loss (from Table 1)	*EIL* Extent of information loss
10	0.1	0.008	0.42%	100.00%	0.001	0.01%	0.00%	0.11%
4	0.25	0.100	2.00%	99.37%	0.005	0.05%	0.09%	0.69%
2	0.5	0.442	4.42%	98.11%	0.020	0.23%	0.40%	1.73%
1.3333333	0.75	1.010	6.73%	96.87%	0.045	0.52%	0.87%	2.81%
1	1	1.754	8.77%	95.71%	0.076	0.87%	1.46%	3.85%
0.8	1.25	2.677	10.71%	94.68%	0.115	1.24%	2.17%	4.86%
0.6666667	1.5	3.847	12.82%	93.38%	0.160	1.66%	2.99%	6.02%
0.5714286	1.75	5.050	14.43%	92.40%	0.208	2.00%	3.89%	6.98%
0.5	2	6.345	15.86%	91.66%	0.258	2.35%	4.83%	7.85%
0.4444444	2.25	7.848	17.44%	90.65%	0.316	2.78%	5.84%	8.85%
0.4	2.5	9.441	18.88%	89.53%	0.371	3.11%	6.87%	9.83%
0.3636364	2.75	11.216	20.39%	88.63%	0.426	3.45%	7.90%	10.78%
0.3333333	3	13.172	21.95%	87.70%	0.491	3.88%	8.97%	11.78%
0.2	5	30.058	30.06%	81.42%	0.974	6.15%	17.87%	18.17%
0.1	10	81.855	40.93%	71.10%	2.088	10.57%	35.86%	29.07%
0.0666667	15	139.545	46.51%	64.73%	3.131	14.52%	48.80%	36.28%

Figure 8 presents the final results quantifying various information loss measures and different sampling rates. The results show that different measures have different levels of information loss at a given sampling rate and the relationship is non-linear. As sampling rate drops, more information about the out-of-range observations (MIL_2) is lost. This measure may be critical for some purposes, e.g., crash reconstruction and reporting. Therefore, for studies dealing with crashes, especially crash reconstruction studies that are highly sensitive to speed magnitude, higher sampling rates can be beneficial. The curves, including the overall information loss measure show that information loss becomes rather high between at 1 to 2-Hz level.

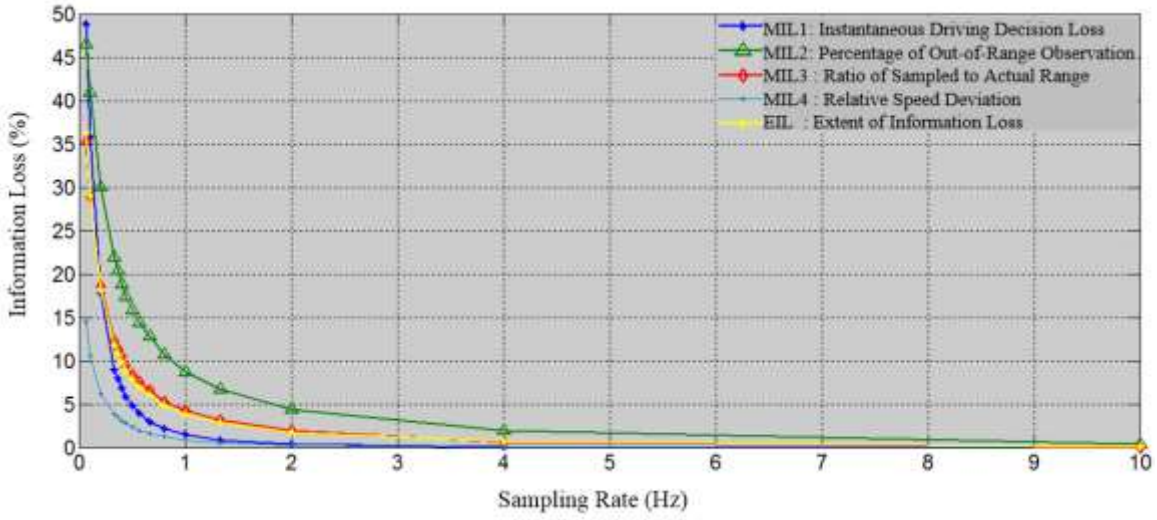


FIGURE 8 Extent of information loss with different sampling rates

LIMITATIONS

The data used in this study comes from a simulator driving test, i.e., they are from a hypothetical but controlled test environment. Having few test subjects is recognized as a limitation, though it is not very germane to this study. The data was sampled by 20 Hz. It is possible that micro driving decisions between the 20 Hz time-stamp data points were lost. This study assumes the chance of having micro decision changes within 0.05 second is very small, given a perception reaction times of about 1 second. In the future, driving data sampled at even higher sampling rates can be used to verify the results of this study. The proposed measures can be used for analysis of information loss with any range of sampling frequency.

CONCLUSIONS

The key question investigated in this study is: what sampling rates are appropriate to capture micro or short-term driving decisions? Oversampling can result in noisy data, and waste storage and processing resources. Undersampling can result in loss of information about important instantaneous driving decisions. This study developed measures of information loss and quantified their relationship with sampling rates. It discussed driving behavior information from two angles: instantaneous driving decisions and speed magnitudes. Four

main measures were created to quantify the magnitudes of driving behavior information loss: a) MIL_1 –Instantaneous driving decision loss (combined direct and indirect 'detectability'); b) MIL_2 – Percentage of out-of-range observations; c) MIL_3 – Ratio of sampled to actual range; and d) MIL_4 – Relative speed deviation from linear interpolation of sampled data (based on observed speed deviation over interpolated speed). These measures quantify the extent of information loss. With these four measures, the overall magnitude of information loss index was generated by equally weighting them. The index, termed by Extent of Information Loss (EIL), simply tells us how much information might be lost given a sampling rate.

The results show that shorter time intervals (i.e., higher sampling rates) are associated with larger direct detectability of instantaneous driving decisions. In other words, there is a smaller chance of having cases with micro-driving decisions between two sampled data points. Drivers typically keep constant acceleration/deceleration rates during a short time. Specifically, for the time interval 1 second (i.e., 1-Hz sampling rate) the direct detectability is 88.90%. The large possibility of no micro-decision in one second may be due to the driver reaction time. The reaction time includes the time for driver perception, identification, judgment and reaction (*26*). The whole process usually takes more than 1 second (*24*). This study further observed cases of one micro-driving decision made within a particular time interval and discussed the possibility of detecting such micro-driving decisions. Through defining the six possible micro driving decision patterns, the study found the four of six patterns include the micro-driving decisions that can be detected indirectly by using the sampled data points. These four patterns dominate the cases in short time intervals (less than 3 seconds). Specifically, the indirect detectability for one second time interval (or 1-Hz sampling rate) is around 93.92%. The feasibility of detecting micro-driving decisions combines direct detectability and indirect detectability. Thus, the feasibility of detecting micro-driving decisions by 1-Hz data are 89.90% + 9.20% × 93.92% = 98.54%, and 100%-98.54%= 1.46% of information about micro-decisions (MIL_1) will be lost by 1-Hz data.

The measures of information loss magnitude reveal that smaller sampling rates or longer time intervals are related to more missing data points because of their too large or too small values. Though there are some data points out of the micro speed ranges (about 8.77% of points out of the micro ranges for 1-Hz data, MIL_2), these points do not deviate broadly

when sampling rates are equal to or higher than 1 Hz. Specifically, the ratio of sampled to actual ranges (MIL_3) is 95.7% for 1-Hz data. And 1-Hz data has average speed deviation of about 0.076 mph. The small deviation supports the assumption that driving behavior within one second shows nearly constant acceleration (*25*). Further, the relative speed deviation (MIL_4) of 1-Hz data to 20-Hz is around 0.87%. With four measures of Magnitudes of Information Loss (MILs), the overall Extent of Information Loss (EIL) can be calculated. For 1-Hz sampling rate, the EIL is about 3.85%.

This study proposed measures to quantify the magnitude of information loss. The measures can be used individually or combined to create an index. The results show that lower sampling rates are associated with greater information loss, but the relationship is not linear. This study contributes by quantifying the relationship between sampling rates and information loss and depending on the objective of their study, researchers can choose the appropriate sampling rate necessary to get the right amount of accuracy. For some studies, e.g., quantifying energy consumption or emissions, 0.5 Hz sampling rate may be sufficient, whereas for safety studies, higher sampling rates may be required.

References

1. 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*, Vol. No. 2015: pp.DOI: 10.1016/j.trc.2014.12.014.
2. Linear. *LTC6412 - 800MHz, 31dB Range Analog-Controlled VGA*. 2014 [cited 2014 May 1st]; Available from: http://www.linear.com/product/LTC6412.
3. Meade, M.L., C.R. Dillon, and C.R. Dillon, *Signals and systems*. Vol. 8. 1991: Springer.
4. Chawla, N.V., *Data mining for imbalanced datasets: An overview*, in *Data Mining and Knowledge Discovery Handbook*. 2010, Springer. p. 875-886.
5. Punzo, V., M.T. Borzacchiello, and B. Ciuffo, On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data. *Transportation Research Part C: Emerging Technologies*, Vol. 19, No. 6, 2011: pp. 1243-1262.
6. *Open Source NGSIM community*. 2014.

7. Jackson, E., L. Aultman-Hall, B.A. Holmén, and J. Du, Evaluating the ability of global positioning system receivers to measure a real-world operating mode for emissions research. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1941, No. 1, 2005: pp. 43-50.
8. Int Panis, L., S. Broekx, and R. Liu, Modelling instantaneous traffic emission and the influence of traffic speed limits. *Science of the total environment*, Vol. 371, No. 1, 2006: pp. 270-285.
9. Ahn, K. and H. Rakha, The effects of route choice decisions on vehicle energy consumption and emissions. *Transportation Research Part D: Transport and Environment*, Vol. 13, No. 3, 2008: pp. 151-167.
10. Campbell, K.L., The SHRP 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety. *TR News*, Vol. No. 282, 2012: pp.
11. Wang, H., L. Fu, Y. Zhou, and H. Li, Modelling of the fuel consumption for passenger cars regarding driving characteristics. *Transportation Research Part D: Transport and Environment*, Vol. 13, No. 7, 2008: pp. 479-482.
12. Hung, W., H. Tong, C. Lee, K. Ha, and L. Pao, Development of a practical driving cycle construction methodology: A case study in Hong Kong. *Transportation Research Part D: Transport and Environment*, Vol. 12, No. 2, 2007: pp. 115-128.
13. Lyons, T., J. Kenworthy, P. Austin, and P. Newman, The development of a driving cycle for fuel consumption and emissions evaluation. *Transportation Research Part A: General*, Vol. 20, No. 6, 1986: pp. 447-462.
14. Boriboonsomsin, K., A. Vu, and M. Barth, Eco-driving: Pilot evaluation of driving behavior changes among US drivers. Vol. No. 2010: pp.
15. Simpson, M. and T. Markel. Plug-in Electric Vehicle Fast Charge Station Operational Analysis with Integrated Renewables. in *EVS26 (Electric Vehicle Symposium)*. 2012.
16. TSDC, *Secure Transportation Data Project*. 2014, Transportation Secure Data Center, National Renewable Energy Laboratory
17. Bikowitz, E.W. and S.P. Ross, *Evaluation and improvement of inductive loop traffic detectors*. 1985.
18. Oh, S., S.G. Ritchie, and C. Oh, Real-time traffic measurement from single loop inductive signatures. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1804, No. 1, 2002: pp. 98-106.

19. Landau, H., Sampling, data transmission, and the Nyquist rate. *Proceedings of the IEEE*, Vol. 55, No. 10, 1967: pp. 1701-1706.
20. Yang, Q., R. Overton, L.D. Han, X. Yan, and S.H. Richards, Driver behaviours on rural highways with and without curbs–a driving simulator based study. *International journal of injury control and safety promotion*, Vol. No. ahead-of-print, 2013: pp. 1-12.
21. Bédard, M., M. Parkkari, B. Weaver, J. Riendeau, and M. Dahlquist, Assessment of driving performance using a simulator protocol: Validity and reproducibility. *The American Journal of Occupational Therapy*, Vol. 64, No. 2, 2010: pp. 336-340.
22. Wang, Y., B. Mehler, B. Reimer, V. Lammers, L.A. D'Ambrosio, and J.F. Coughlin, The validity of driving simulation for assessing differences between in-vehicle informational interfaces: a comparison with field testing. *Ergonomics*, Vol. 53, No. 3, 2010: pp. 404-420.
23. Elmore, W.C. and M.A. Heald, *Physics of waves*. 2012: Courier Dover Publications.
24. AASHTO, A Policy on Geometric Design of Highways and Streets, 6th Edition, 2011. *American Association of State Highway and Transportation Officials, Washington, DC*, Vol. 1, No. 2011: pp. 990.
25. EPA. *Dynamometer Drive Schedules*. 2013 [cited 2014 March 3rd]; Available from: http://www.epa.gov/nvfel/testing/dynamometer.htm.
26. TRB, *Managing Speed: Review of current practice for setting and enforcing speed limits*. Transportation Research Board, National Research Council.No. 1998.

WHAT IS THE LEVEL OF VOLATILITY IN INSTANTANEOUS DRIVING BEHAVIORS?[2]

Abstract - Driving styles can be broadly characterized as calm or volatile, with significant implications for traffic safety, energy consumption and emissions. How to quantify the extent of calm or volatile driving and explore its correlates is a key research question investigated in the study. This study contributes by leveraging a large-scale behavioral database to analyze short-term driving decisions and develop a new driver volatility index to measure the extent of variations in driving. The index captures variation in driving behavior constrained by the performance of the vehicle from a decision-making perspective. Specifically, instantaneous driving decisions include maintaining speed, accelerating, decelerating, maintaining acceleration and deceleration, or jerks to vehicle, i.e., the decision to change marginal rate of acceleration or deceleration. A fundamental understanding of instantaneous driving behavior is developed by categorizing vehicular jerk reversals (acceleration followed by deceleration), jerk enhancements (increasing accelerations or decelerations), and jerk mitigations (decreasing accelerations or decelerations). Volatility in driving decisions, captured by jerky movements, is quantified using data collected in Atlanta, GA during 2011. The database contains 51,370 trips and their associated second-by-second speed data, totaling 36 million seconds. Rigorous statistical models explore correlates of volatility that include socioeconomic variables, travel context variables, and vehicle types. The study contributes by proposing a framework that is based on defining instantaneous driving decisions in a quantifiable way using big data generated by in-vehicle GPS devices and behavioral surveys.

[2] The idea of driving volatility originated in another project sponsored by the US DOT under the *TranLive* University Transportation Center. That project is titled "Reducing Energy Use and Emissions through Innovative Technologies and Community Designs." Driving volatility has implications for safety as well as energy and environment. Therefore, it was developed further in the safety context using large-scale trajectory data. Materials presented here are based on the following publication and presentations: 1) Wang X, A. Khattak, J. Liu, G. Masghati-Amoli & S. Son. What is the Level of Volatility in Instantaneous Driving Decisions? Forthcoming in Transportation Research Part C: Emerging Technologies, 2015. 2) Liu J., X. Wang & A. Khattak. Generating Real-Time Volatility Information, Presented at 2014 Intelligent Transportation Systems World Congress, Detroit, MI, 2014. 3) Khattak A., J. Liu & X. Wang. Supporting Instantaneous Driving Decisions through Vehicle Trajectory Data, TRB paper # 15-1345. Presented at the Transportation Research Board Annual Meeting, National Academies, Washington, D.C., 2015.

Keywords: instantaneous driving decision; big data; volatility; acceleration; speed

INTRODUCTION

As the most dominant transportation mode in USA, automobile driving has significant impacts on traffic safety, energy, and emissions. With widespread deployment of emerging information and communication technologies, massive amounts of driving data in high resolution are becoming available, allowing researchers to scrutinize driving behavior in far more detail than was possible before. Insights can be obtained by studying instantaneous decisions made during driving in nearly real-time. Also, such "Big data" provides opportunities that support visualization, analysis, and modeling in new ways that could not be imagined before. The combination of data and tools can help create new visions that can potentially transform the way we monitor and evaluate transportation system performance and potential improvement actions. This study takes advantage of the big data collected by in-vehicle Global Positioning System (GPS) devices and survey data to define instantaneous driving decisions as drivers' choices of a set of options during driving. Such choices include maintaining speed, accelerating, decelerating, maintaining acceleration/deceleration, and vehicular jerk, i.e., the decision to change marginal rate of acceleration and deceleration. The sequential chaining of these short-term driving decisions can be volatile because they are intended to respond to the instantaneous changes in surrounding circumstances, such as approach of adjacent vehicles, pavement conditions, geometric transitions in the roadway, and weather conditions. Fluctuations in traffic flow can create challenges for safety, as well as challenges for energy consumption, tailpipe emissions and public health (*1, 2*). Existing studies have shown that emissions and fuel usage vary significantly with different speed ranges (US EPA, *3*). Additionally larger deviations from mean speed can significantly increase crash risk (TRB, *4*). Accordingly it is important to understand and quantify variability in drivers' instantaneous decisions and explore the associations with socioeconomic, vehicular, and contextual variables.

Volatility in instantaneous driving decisions can be quantified by variability in vehicular movement, and the variability can be represented by speed and its derivative (acceleration/deceleration) as well as its second derivative (vehicular jerk). Micro level GPS

data along with behavioral survey data are used to answer the following fundamental questions:

1) How to develop measures of driving volatility?
2) What is the level of volatility in instantaneous driving decisions?
3) What are the key correlates of driving volatility?

LITERATURE REVIEW

Aggressive driving and its impacts on traffic safety has been a concern of the public and many other sectors, including public transportation agencies, policy agencies, insurance companies, various organizations such as American Automobile Association. No consensus exists regarding "aggressive driving" in the literature. Social psychology researchers define it from the perspective of intent (*5*); for instance, "road rage" refers to more criminal-oriented offenses (*6*), while NHTSA classify "aggressive driving" as "driving actions that markedly exceed the norms of safe driving behavior and that directly affect other road users by placing them in unnecessary danger"(NHTSA*7*). Other researchers had a list of "aggressive driving" (*8*) including "weaving in and out of traffic", "driving at speeds far in excess of the norm which results in frequent tailgating, frequent and abrupt lane changes", "passing one or more vehicles by driving on the shoulder and then cutting in", or through certain syndrome of frustration-driven behaviors or negative cognitions such as annoyance, hostility, sustained horn-honking, glaring at others, yelling, gesturing, etc. (*6, 9-11*). These studies in driving psychology largely depend on self-reported surveys of the driving public (*5, 12*), or video recording which requires manual identifications (*13*), with limitations on collecting data systematically and accurately. Critical research issues include: what are the so-called the norms of safe driving behavior; how to define a driver's extent of "aggressive driving" in a precise and quantifiable way.

While the research of "aggressive driving" in social psychology focuses more on peoples' intentions, the above driving behaviors and their cognitive processes in such driving situations are difficult to measure directly and continuously. Nevertheless, the speed profile as a common observable behavior is relatively easy to collect and has the potential of being utilized to characterize driving behavior. Measures used in the literature to identify

aggressive or calm driving styles include the ratio of the standard deviation and the average acceleration within a specified time window (*14*), ratio of standard deviation and vehicular jerk of the normal driving style (*15*).

Several critical cutoff points for aggressive behavior based on acceleration have been reported; 1.47 m/s^2 (4.82 ft/s^2) and 2.28 m/s^2 (7.47 ft/s^2) were reported as critical estimates of aggressive and extremely aggressive acceleration thresholds in urban driving environments (*16, 17*). However, there is no consensus threshold, for instance, other researchers reported 0.45-0.65 m/s^2 (1.48-2.13ft/s^2) as calm driving, 0.85-1.10m/s^2 (2.79-3.61$ft/^2$) as aggressive driving for urban journeys (*18*).

The percentage of time acceleration exceeds 1.5 m/s^2 (4.92 ft/s^2) was reported as one of the most important parameters (out of 16 parameters) contributing to increases in emissions and fuel consumption (*19*). However, researchers argued that using acceleration alone may not represent the driving style accurately; therefore the coefficient of variance were also used as a complementary measurement in order to identify aggressive driving. Accordingly, accelerating at a relative regular rate, along with driving with medium acceleration but high standard deviation of acceleration are both flagged as aggressive driving (*14*).

Connections between aggressive driving and safety were found in existing studies (*20, 21*). Paleti et al. (*21*) have explored aggressive pre-crash behaviors and defined aggressive driving to include "speeding, tailgating, changing lanes frequently, flashing lights, obstructing the path of others, making obscene gestures, ignoring traffic control devices, accelerating rapidly from stop, and stopping suddenly." Their results show a positive association between injury severity and aggressive driving (given a crash).

Regarding emissions and fuel consumption, studies have shown that emissions can vary according to the decisions including both strategic decisions (vehicle selection and maintenance tactical decisions (selection of routes, dealing with congestion, and operational decisions (idling, speed selection, and use of cruise control) (*22*). A large number of studies have linked microscopic "aggressive" driving with emissions. Research has shown that peak emissions are associated with aggressive driving behavior including high speeds and extreme speed-ups or brake-downs (*18, 23-26*). Factors describing speed, acceleration, power

demand, and gear changing behavior are significantly associated with emissions (HC, NOx, and CO2) as well as fuel consumption (*19*). An understanding of speed variation/ speed fluctuation/ driving dynamics, acceleration variation can further benefit research in energy and emissions.

While the literature provides insights, there is still a need to quantify the extent of volatile (aggressive) driving on routine urban journeys using continuous and reliable sources of data. This study is intended to close the gap between psychological studies and crash studies by applying appropriate empirical methods to quantify "volatile driving," and analyze the socio-demographic and travel correlates, which distinguishes this work from other driving behavior studies in social psychology, human factors, and safety fields. This study is also quite different from previous engineering-based aggressive driving studies because unique real-world GPS driving data along with reported behavior data from a survey are used to quantify the extent of variability in driving decisions. Considering that the word "aggressive" contains intent of the person, the use of term "volatility" in driving decisions is preferred in the paper, as it better suits the purpose of measuring the variability in instantaneous driving decisions.

DATA DESCRIPTION

Data used in this study come from the Atlanta Regional Commission — A Regional Travel Survey with GPS Sub-Sample conducted in 2011 (survey period covered Feb. 2011 through Oct. 2011). It was a well-executed regional survey using CATI (Computer-assisted telephone interviewing), with 6% final response rate and 34% participate rate. The sample is large-scale, covering about 20 counties in the region of Atlanta, representing various land use types and populations. Overall, the data quality was reasonable and efforts were made to make the sample representative of the region. More details about the survey are available in the report (*27*). Similar to a standard travel behavior survey, the instrument relies on the willingness of households to 1) provide demographic information about the household, its members and its vehicles; 2) have all household members recording all travel-related details for a specific 24-hour period on multiple travel days, including their trip purposes, travel modes and other standard trip diary questions; 3) in the GPS subsample, data were collected by in-vehicle

GPS devices for each trip. The device captured travel date, time, latitude and longitude (however this information was removed from the public released database), and the speed data. The GPS data points were collected at a sampling rate of at least 0.25 Hz and the raw GPS data was fed through a processing routine that removed outlying speed values, interpolated missing data and smoothed the speed profile (*28*).

The final database contains different levels of data-personal data; household data, trip data, and microscopic second-by-second data for each trip. In all, 51,370 trips made by 1,653 drivers from 850 households were included in the database, which contained a total of more than 36 million seconds of records, covering driving practices on different road types by different type of vehicles.

The data was collected professionally, using state-of-the-art methods and upon examination show that it is reasonable. Specifically, for driving data, the speed data has reasonable ranges, with highest speed of 80 mph, average speed of 37 mph; acceleration changes ranged between -5.2ft/s^2 and 7.64ft/s^2, which are consistent with the numbers reported in the literature, e.g., 7.47ft/s^2 as extremely aggressive driving (*29*). Vehicular jerk changes ranged between -5.53 ft/s^3 and 8.28 ft/s^3. For demographics, again the data are reasonable. Specifically, 47.24% of drivers were male; the average age of respondents was 47 years. This fairly represents the driving population in Atlanta. Comparing the sampled data with other data sources such as the census showed that 47.24% of male drivers in the sample is consistent with 47.4% in the Atlanta are population; average age of 47.18 years, this is consistent with Census (49% of population is between 25 to 54); and average vehicle age of 7.9 years is consistent with 33.8% of vehicles in Atlanta area that are between 6-10 years old.

METHODOLOGY

Measures of Instantaneous Driving Decisions

Distinct from strategic during decisions, instantaneous driving decisions refer to those micro-decisions to accommodate real-time situational changes during their journeys. These instantaneous driving decisions can include: accelerating, decelerating, maintaining constant speed (zero acceleration), jerking the vehicle (change in marginal rate of acceleration or

deceleration), or maintaining constant acceleration and deceleration (zero vehicular jerk). As shown in Equation 9, vehicular jerk is the derivative of acceleration or the second derivative of speed, representing abrupt movement of vehicles. Therefore, while an acceleration profile shows how fast a driver speeds up and slows down, a vehicular jerk profile shows how fast a driver accelerates and decelerates, which is more suited to capture drivers' abrupt adjustments in speeds. Figure 9 represents the speed, acceleration and vehicular jerk profile for a single sampled driving trip.

$$J = d(a)/d(t) = d^2(v)/d(t)^2 = d^3(d)/d(t)^3 \quad \textit{Equation (9)}$$

Where J is vehicular jerk; a is acceleration; v is velocity; d is distance

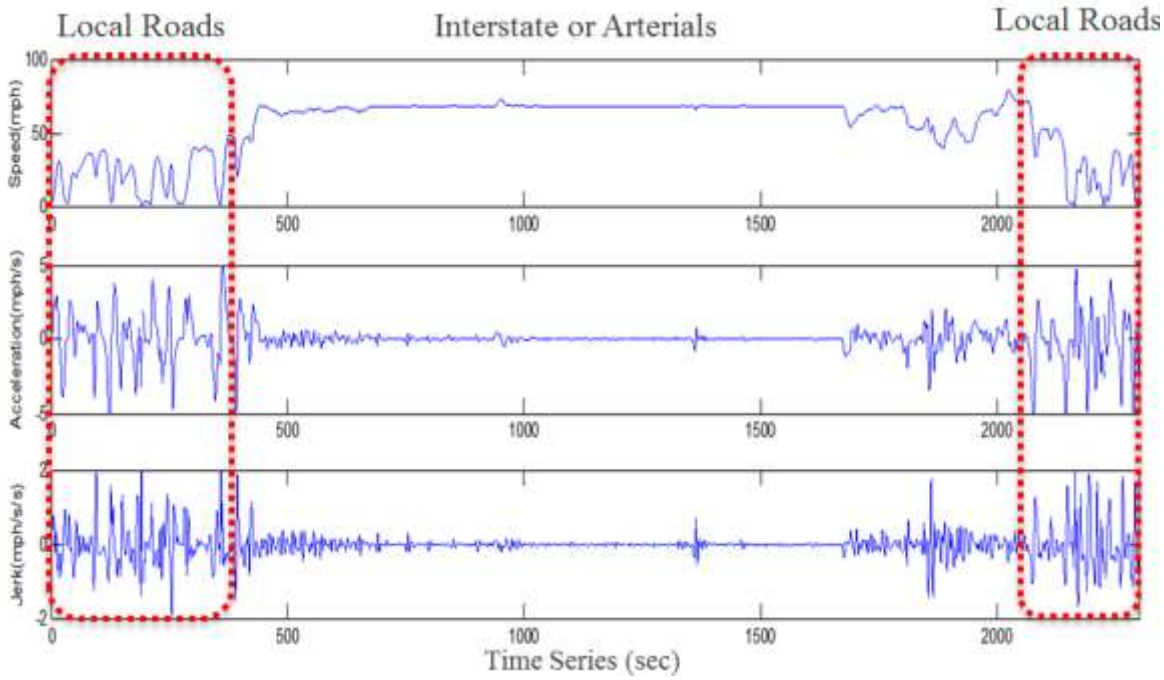
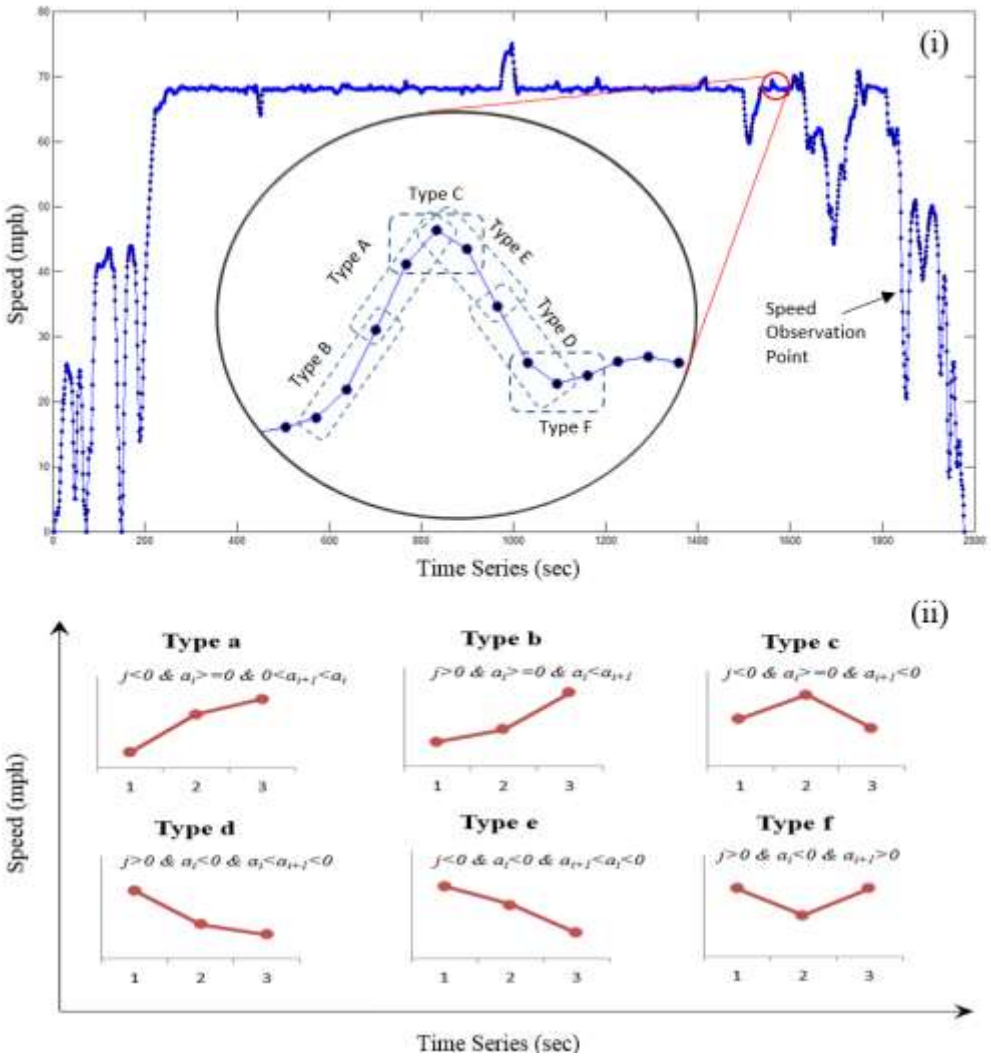


FIGURE 9 Comparison between speed, acceleration and vehicular jerk profiles on a trip

While these three profiles represent the same trip, they show significant differences, especially when speed fluctuates. The spikes in the vehicular jerk profile occur only when there are large changes in the accelerations, negatively or positively. The vehicular jerk profile acts as an amplification of speed changes since it is more sensitive to speed changes.

Patterns of Instantaneous Driving Decisions

Different patterns of instantaneous driving decisions can be observed based on how acceleration and deceleration are chained sequentially. Figure 10 shows six different vehicular jerk patterns during driving for illustrative purposes. The upper three graphs show vehicular jerks starting from acceleration and followed respectively by lower acceleration (a), higher acceleration (b), and deceleration (c). The lower graphs show vehicular jerks starting from a vehicle braking and followed respectively by a lower deceleration (d), higher deceleration (e), and acceleration (f). In these graphs, there is a decision point at second 10 when the driver has to decide whether he/she wants to change the current driving situation.



Notes: j=vehicular jerk; a_i=acceleration at time i; a_{i+1}=acceleration at time $i+1$

FIGURE 10 Different types of vehicular jerk during driving.

Since vehicular jerk is the second derivative of speed, it can be positive (b, d, f) or negative (a, c, e). Where vehicular jerk is zero, the driver operates the vehicle at a fixed acceleration/deceleration rate or simply maintains the speed. However, generally there can be a greater chance of collisions when negative vehicular jerk happens compared with positive vehicular jerk. In situations where vehicles are followed by other vehicles, negative vehicular jerks can result in abrupt shortening of distance between the vehicles and following vehicles, possibly creating a shockwave under condition c, e and a (a shockwave from strong to weak). Understanding the profiles of different vehicular jerk styles is important for safety and for energy and emissions.

Methodological Framework

Figure 11 shows the overall framework. The purpose of this study is to generate knowledge of short-term driving decisions by taking advantage of large-scale travel survey data that contain 36 million second-by-second trajectory records with travel behavioral data from 1,653 drivers. To do this, the research first defines different instantaneous driving decision patterns. Speed, acceleration, and vehicular jerk are extracted from the (large-scale) raw trajectory data, with decision patterns identified by chaining decisions with different sequences. Next, visualizing the data provides a complete picture of how drivers spend their time on these different driving decisions at different vehicular speeds. Then trip-based measures of short-term driving volatility are created based on acceleration and vehicular jerk profiles. Then, statistical models are estimated in order to explore the socio-demographic and travel correlates of driving volatility, generating new knowledge about volatility. Finally, potential applications for supporting calmer/smoother driving behavior and traffic management are proposed.

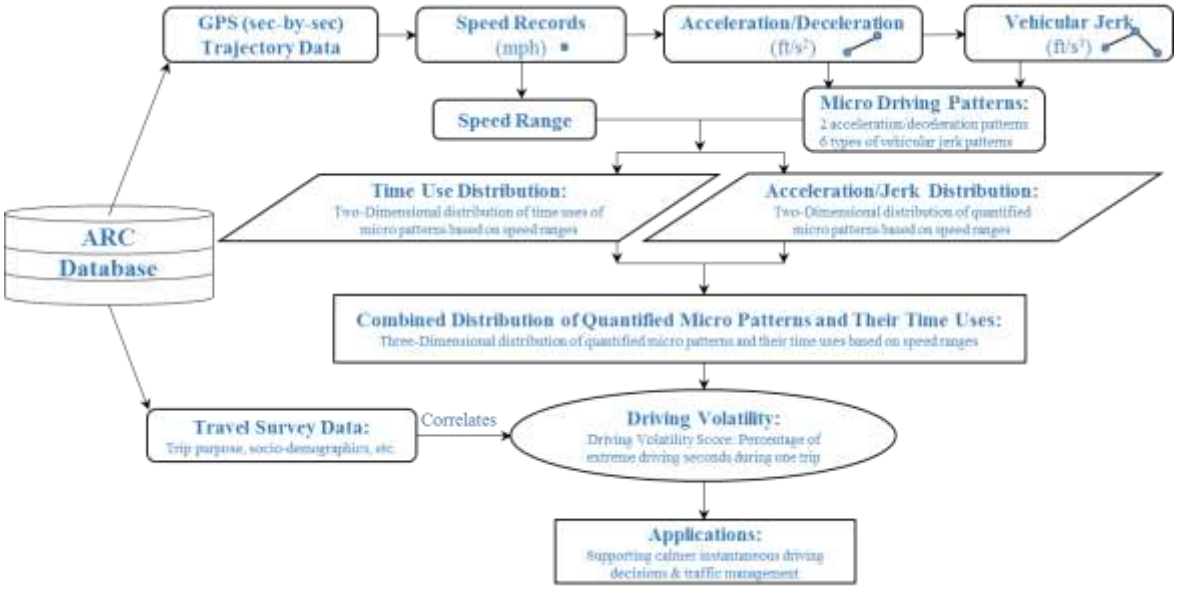


FIGURE 11 Methodological framework

RESULTS – EXTENT OF VOLATILITY IN DRIVING

Time Use Distribution

Acceleration/Deceleration

To understand driving time spent on different instantaneous decisions in a metropolitan environment, the frequency of acceleration, deceleration and zero acceleration by speed bin in 0.5 mph (mile per hour) increments were calculated based on 36 million driving seconds of total 51,370 trips (shown in Figure 12). On selection of speed bin, we have conducted sensitivity analysis and found that volatility can be somewhat sensitive to the selection of different speed bin widths. There is no ideal bin size, but we know that if the bin size is too large (e.g., 5 mph), then the data are overly aggregated and there is substantial loss of variability (note that there are only 16 bins for speeds ranging from 0 mph to 80 mph). If the bin size is too small (e.g., 0.1 mph), then data noise (random fluctuations) can become an issue, obscuring interpretation (for 0.1 mph speed bins there will be 800 bins for 0 to 80 mph range). The 0.5 mph (equivalent to 0.73 ft/s) speed bin is a reasonable compromise that gives a fairly accurate picture of the acceleration and jerk distributions with respect to driving speeds.

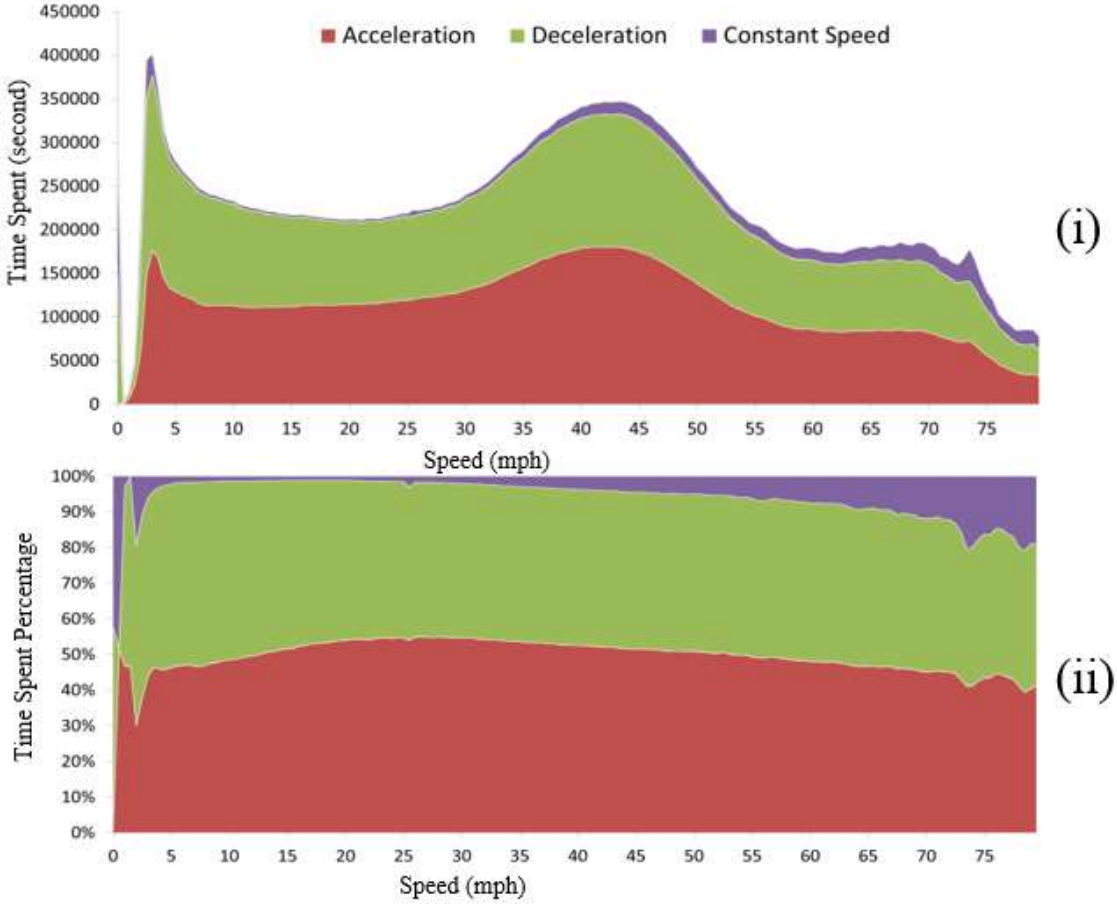


FIGURE 12 Time use in acceleration, deceleration and constant speed at different speeds (N= 36 Million)

Given that each sample represents one second of driving, the magnitude of frequency bars demonstrate the time used during trips on acceleration, deceleration and maintaining constant speed of the vehicle. Notably, very small accelerations or decelerations (0.03 mph, based on the 5th percentile of speed changes) were considered noise and coded as constant speed. Figure 12 (i) shows time use distribution and (ii) shows the percent of time spent on acceleration, deceleration and constant speed after standardization.

Overall 7% of driving time was spent driving at idling or low speeds (below 5 mph), 47% of driving time was spent on acceleration, 41% of driving time was spent on deceleration and 5% of driving time was spent maintaining constant speed, based on the massive amount of field data from GPS devices. The results can be compared with the

Federal Test Procedure (FTP) drive cycle test (known as FTP-75 for the city driving cycle), which involves a decelerating drive mode for 34.5% of the time, and idling mode for 17.9% of the time (*30, 31*). Table 3 shows major drive cycles designed to represent typical driving practices in order to certify vehicle fuel economy. The massive field driving data provides first-hand knowledge of real world driving practices, which can inform drive cycle design and provide insights.

TABLE 3 United States certification drive cycles compared with Atlanta drive cycle (*30*)

Drive Cycle	Description	Data Collection Method	Year of Data	Top Speed	Avg. Speed	Max. Acc.	Distance	Time (min)	Idling time
FTP	Urban/City	Instrumented Vehicles/Specific route	1969	56 mph	20 mph	1.48 m/s^2	17 miles	31 min	18%
C-FTP	city, cold ambient temp	Instrumented Vehicles/ Specific route	1969	56 mph	32 mph	1.48 m/s^2	18 miles	31min	18%
HWFET	Free-flow traffic on highway	Specific route Chase-car/ naturalistic driving	Early 1970s	60 mph	48 mph	1.43 m/s^2	16 miles	12.5 min	None
US06	Aggressive driving on highway	Instrumented Vehicles/ naturalistic driving	1992	80 mph	48 mph	3.78 m/s^2	13 miles	10min	7%
SC03	AC on, hot ambient temp	Instrumented Vehicles/ naturalistic driving	1992	54 mph	35 mph	2.28 m/s^2	5.8 miles	9.9 min	19%
Atlanta	Urban/City	In-veh. GPS devices, Travel survey	2011	80mph	37mph	5.10 m/s^2	7.1 mile^	12.7min^	7%*

Note:
1. FTP: Federal Test Procedure.
2. HWFET: The Highway Fuel Economy Test.
3. US06: The US06 Supplemental Federal Test Procedure (SFTP) for High Speed and High Acceleration Driving behavior.
4. SC03: A Supplemental Federal Test Procedure (SFTP) with Air Conditioning.
5. C- FTP: Federal Test Procedure under cold ambient temperature.
6. ^ mean values are used for Atlanta.
7. * idling & low speeds (below 5 mph)

Travel time spent at different speeds varies, depending on speed range, with 30-50 mph as the most common speed range. Less driving time was spent on driving at speeds higher than 50 mph. This result depends largely on regional road network structure. Overall

greater amounts of driving time were spent on acceleration than deceleration, especially when speed was between 10-50 mph. However, more time was spent on deceleration compared with acceleration in lower speed bins (less than 10 mph). When speed is higher than 50 mph the travel time spent on acceleration and deceleration was nearly equal.

Notably, time spent on maintaining constant speed is much less than time spent on speed alterations. Relatively higher proportion of time is spent on maintaining constant speed when speeds are higher; specifically, more than 10% in speed bins higher than 55 mph and more than 20% at speeds higher than 70 mph. This is reasonable since less stop-and-go traffic is expected on freeways with free flowing traffic, coupled with the use of cruise control on interstates. Notably, neither the data on the use of cruise control nor the road types and second-by-second geo-codes are available in the public use database. This makes it difficult to link the speed profile/bins with specific roadway types, especially when speed is less than 50 mph. For example, the roadway can be a congested interstate or signalized arterial with free flowing traffic. Nevertheless, the graphs reveal useful information that helps understand driving time use. Specifically, the driving time spent on idling (traveling below 5 mph) is below 10% in Atlanta; the time spent on accelerating and braking are roughly equal and substantially higher than time spent on maintaining speed during urban journeys.

Vehicular Jerk

To understand how much time drivers spent on different vehicular jerk decisions, the time spent for the speed bins was aggregated by different vehicular jerk types. Then the results were standardized by calculating the percent of time spent on each vehicular jerk style, shown in Figure 13. Similar to the time spent on acceleration, the percent of time spent on zero vehicular jerks remains a small portion, this is especially true when speed is more than 70 mph. Possible reasons are drivers seem to avoid jerks to vehicles at higher speeds, or the use of cruise control is more common at higher speeds. However, the cruise control usage information was not available in the database, otherwise it would have added valuable information to understand instantaneous driving decisions comprehensively.

Different vehicular jerk styles are observed within different speed bins. Specifically,

for the speeding up behaviors (a, b), Style (a) has a very small share when speed is less than 5 mph then reaches its peak (30%) when speed is around 30 mph, after that, it starts to shrink slightly but remains at least 20%. While style (b) has its largest share when speed is around 10 mph then remains at a 20% share constantly. As for slowing down behavior (d, e), style (d) has its largest share (30%) when speed is 5 mph, then remains relatively constant at 20% when speed increases; style (e) has its largest share when speed is close to zero, representing the hard braking behavior when coming to a stop. When speed increases, the percent of style (e) has peaks at 25% with moderate speeds (between 20 mph and 30 mph) and then remains constantly at 20% when speed is higher than 30 mph. As for the other two styles when acceleration and deceleration behavior are chained, both of style (c) and style (f) account for about 5% and this percentage remains relatively constant at various speeds.

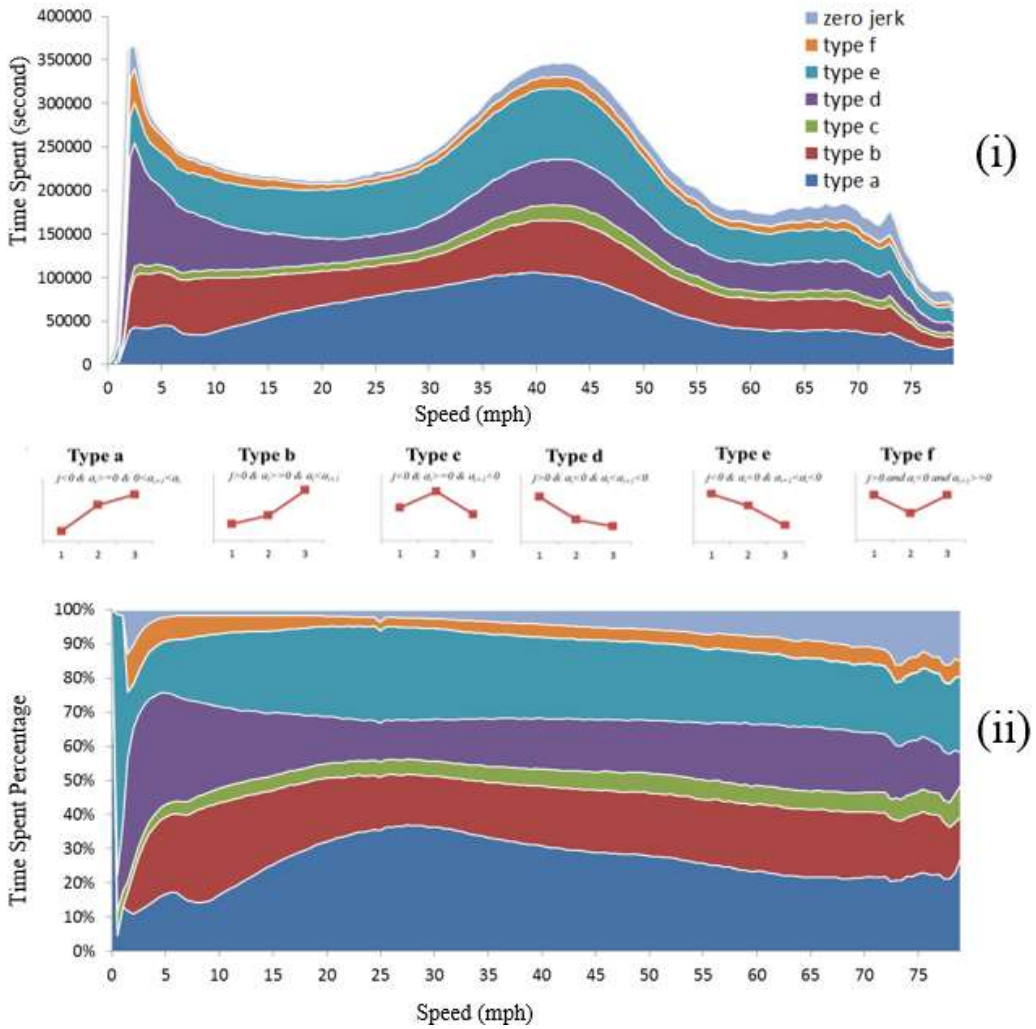


FIGURE 13 Time use in vehicular jerk patterns at different speeds (N= 36 Million)

Variation Distribution

Acceleration/Deceleration

Most existing studies have applied a single acceleration value as a threshold for identifying aggressive driving. Ahn et al. (*31*) have fitted a linear regression line showing that higher accelerations are associated with lower speeds. However, the nonlinear relationships between acceleration and speed in real-life driving situations are largely unexplored. Vehicle engines have to do more work in order to maintain the same acceleration at higher speeds to overcome the increasing air resistance. Therefore the ability to accelerate or decelerate a vehicle decreases naturally at higher speeds.

The speed vs. acceleration/deceleration profile (shown in Figure 14) is consistent with the above expectations. Upper and lower bands represent the means plus/minus one standard deviation bands for accelerations and they denote "typical driving practices." The (red) points that are out of the bands are the "volatile" driving seconds. In general, 15% of the 36 million seconds of driving are volatile (15.73% for acceleration and 14.50% for deceleration). This is reasonable since approximately 68% of the mass will be within one standard deviation for a bell-shaped normal speed distribution. Note that in order to separate the typical behaviors of drivers from moderately and highly risky behaviors, the use of 1 standard deviation threshold is reasonable. Using a 2 or 3 standard deviation threshold instead (i.e., capturing 95% and 99.7% of the observations for normally distributed data), will only leave extreme outliers, that are 5% or even lower (at 0.3%) portion of the data, i.e., high risk behaviors.

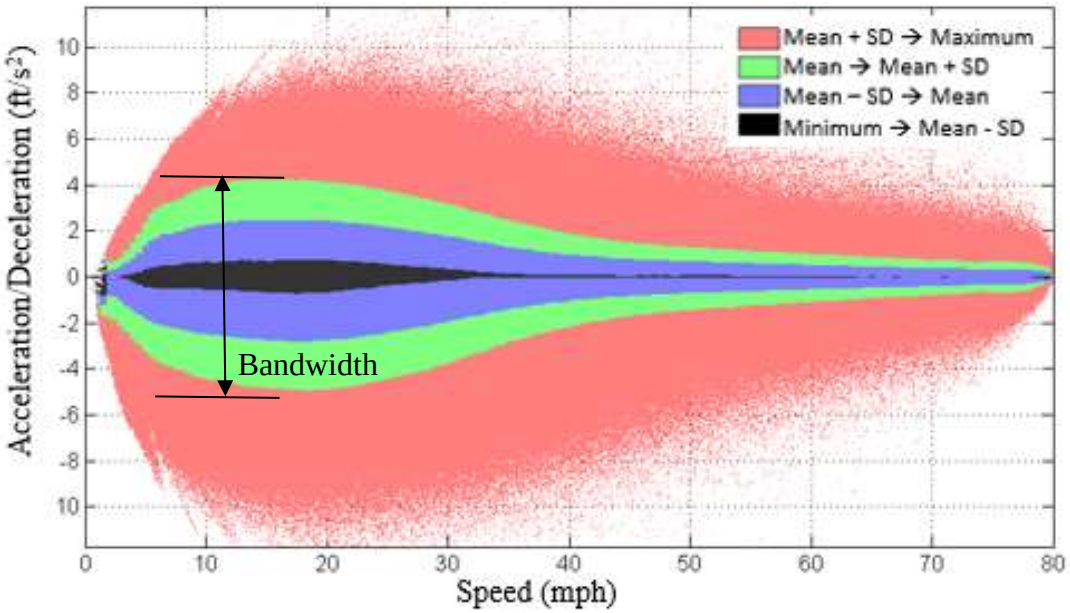


FIGURE 14 Average acceleration/deceleration at different speeds (N=36 Million)

Bandwidth is the difference between the upper band value and the lower band value. A falling bandwidth reflects decreasing variation and rising bandwidth reflects increasing variation in speed changes. The largest bandwidth is between 10 mph and 30 mph and it decreases substantially when speed is higher than 40 mph. This confirms that at higher

speeds (typically on freeways with a good level of service) drivers usually do not or simply cannot accelerate and decelerate abruptly. When speed is above 55 mph, accelerations scarcely exceed 1.5 feet/sec2, as reflected in the upper band.

A similar trend is observed in the deceleration profile with minor differences. Compared with acceleration, the magnitude of the maximum mean of deceleration is higher. It is -3.0 feet/sec2 for deceleration while the maximum mean value is less than 3.0 feet/sec2 for acceleration. This finding is interesting when combined with information contained in Figure 5. It revealed that in the Atlanta area, on average, drivers spend more time braking and they brake harder compared with accelerations.

Vehicular Jerk

Figure 15(i) shows the distribution of the average vehicular jerk by different types and Figure 15(ii) the mean and standard deviation of vehicular jerk at different speeds. The difference in absolute magnitude of vehicular jerk reveals their intensity. Types (c) and (f) show the highest absolute magnitudes which is reasonable since both of them represent drivers reversing vehicle acceleration, i.e., going from acceleration to deceleration or vice versa. Note that type (f) has a higher absolute magnitude than its negative counterpart, i.e., type (c). This means that on average drivers jerk their vehicles more forcefully to accelerate after braking compared with the opposite. This is especially true when speed is less than 40 mph. The other two positive and the two negative jerk types show similar trends and values.

The upper band and lower band (mean plus/minus one standard deviation) are created respectively for the aggregated positive and negative vehicular jerk. For speed bins higher than 40 mph, the lower band of positive vehicular jerk is below zero and the upper band of negative vehicular jerk is above zero; hence zero were used in calculating the bandwidth in those cases. The upper band of the positive vehicular jerk and lower band of negative jerk collectively create a profile of regular practice for vehicular jerk. In other words, it represents the most typical driving practice on roadways regardless of road type. The bands can also serve as a critical threshold for identifying volatile driving behaviors, which are the red points falling outside the bands in Figure 15(ii).

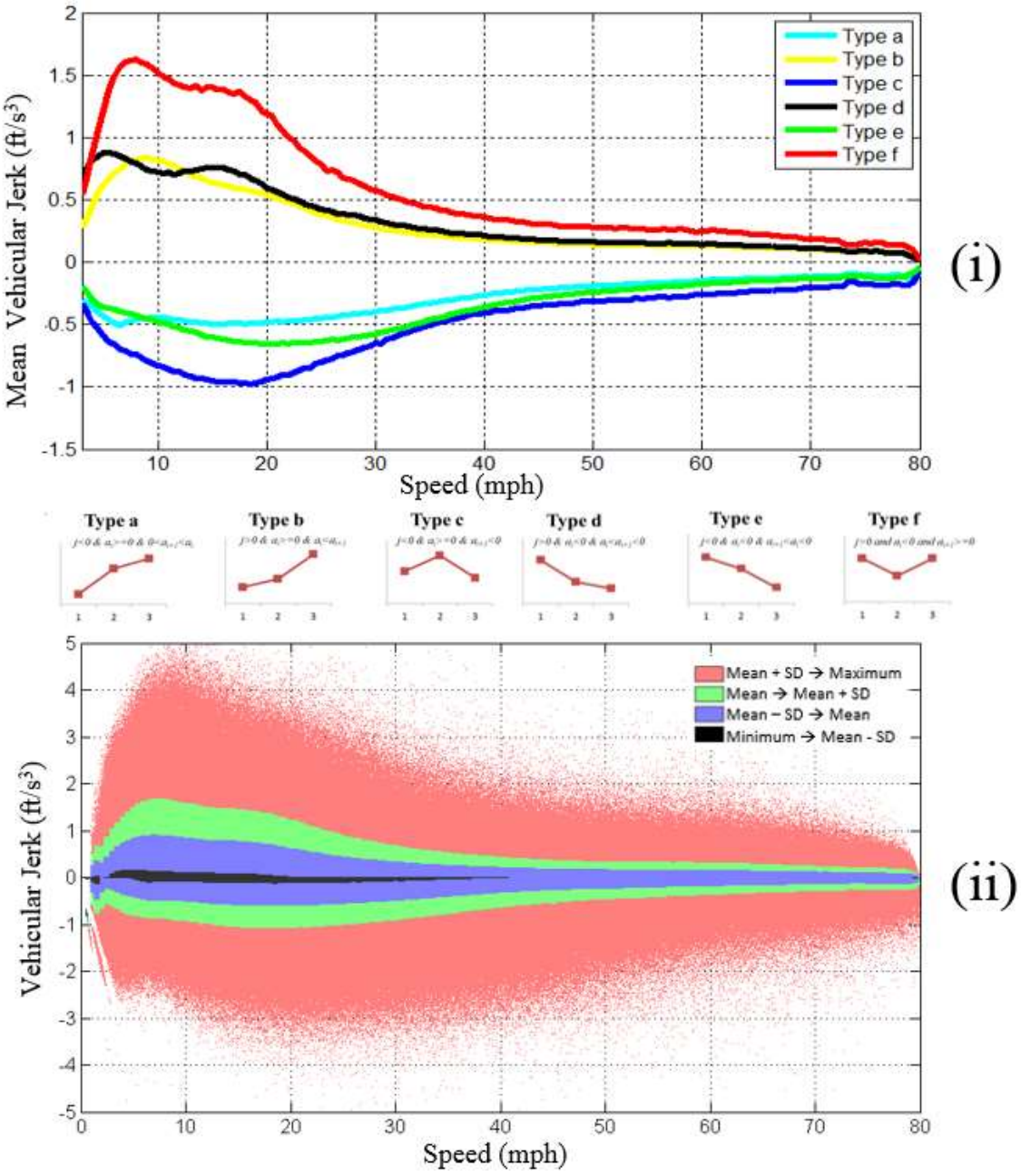


FIGURE 15 Vehicular jerk distribution by speed bins (N=36 Million)

Based on 36 million seconds of driving data, about 13.36% seconds are identified as volatile seconds when using the vehicular jerk profiles. This score represents the average volatility level for typical driving practices for the GPS subsample from the Atlanta Metropolitan Area. More volatile driving practices are found within at lower speeds, as

expected. Specifically, 16.4% of the total time drivers are volatile (above 1 standard deviation) when speed is lower than 20 mph, while 13.6% of the time they are volatile when speed is between 20-40 mph. This percentage drops to 12.00% for speed range between 40-60 mph and it is 11.9% for speeds larger than 60 mph.

The critical values of vehicular jerk associated with volatile driving behavior vary by speed. There is a peaking of this measure at speeds of 7.5 mph then it decreases gradually as vehicular speed goes up, until it reaches a steady line with minor fluctuations at speeds between 45-52 mph. In general, the bandwidth is larger at relatively low speeds (less than 20 mph) and it is relatively narrower at higher speeds. This is to say that lower speeds have a boarder range of volatile driving, but this is not the case for higher speeds.

Combined Distribution

Figure 16 shows three dimensional distribution of time use and variations of instantaneous driving decisions at different speeds. The height shows the number of driving records with corresponding driving status (i.e., speed and acceleration/deceleration or vehicular jerk). At speeds 10 ~30 mph there are fewer driving records with zero acceleration or deceleration (see the trough in Figure 6); for higher speeds (> 60 mph), a large portion of time is spent in maintaining speed with small acceleration or deceleration (see the ridge in Figure 16). Differing from acceleration distributions, vehicular jerk distributions are more concentrated at zero. This implies that any quantified jerk patterns that are different from zero can be easily identified as abnormal micro driving patterns, e.g., sudden braking or accelerating.

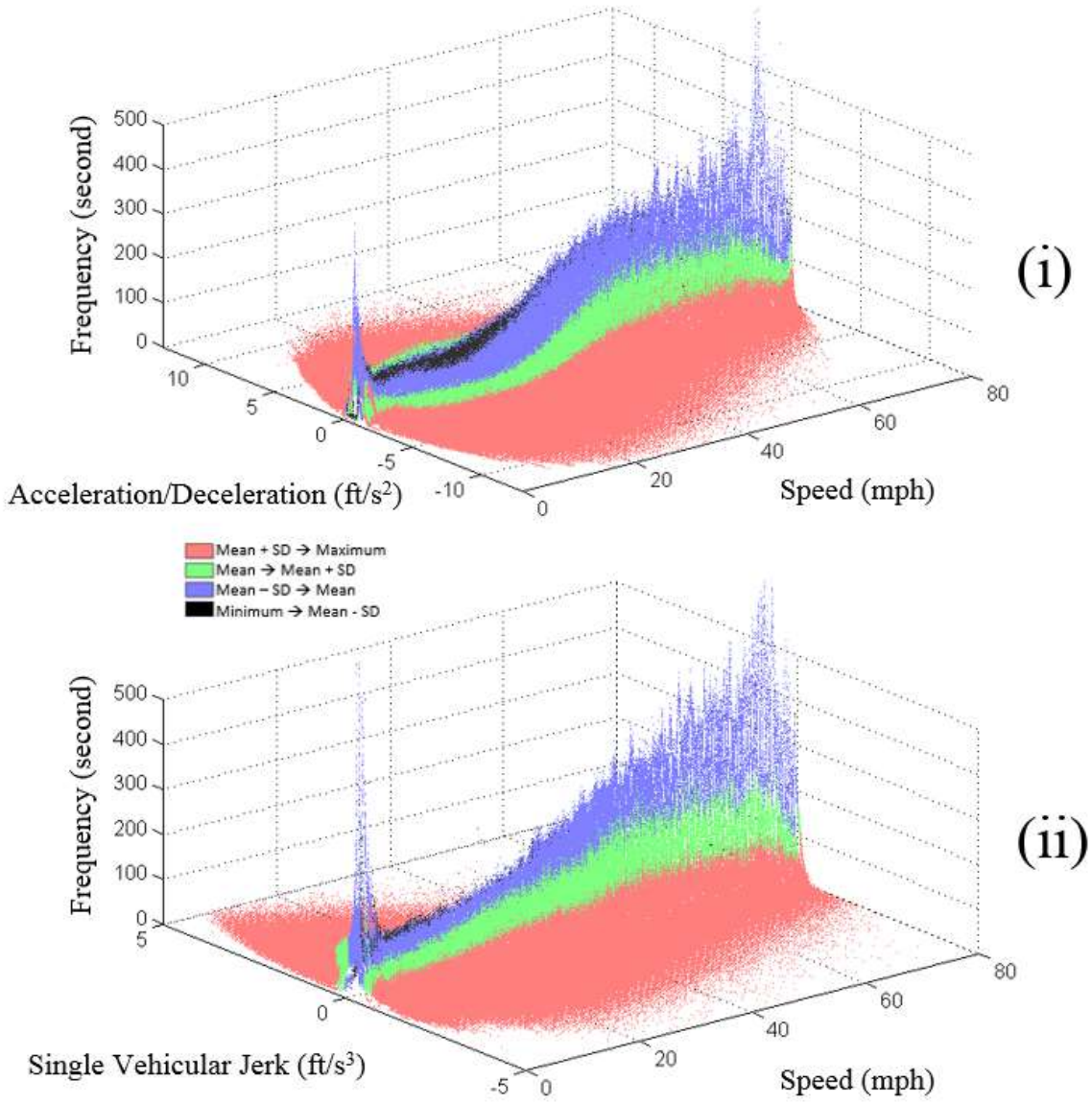


FIGURE 16 3D distribution of time use and variations of instantaneous driving decisions at different speeds (N=36 Million)

Driving Volatility Score

A new measure, termed driving volatility score was created after identify the volatile seconds. The idea is to measure individual volatility for each trip using the acceleration or vehicular jerk band. A driver's volatility score is defined as a percentage of time tagged as volatile seconds over the entire trip. In other words, volatility is measured as the percentage of time when the driver's acceleration or vehicular jerk goes beyond the typical driving

thresholds (acceleration or vehicular jerk bands). The driving volatility score can be calculated by following equation:

$$Volatility\ Score\ \% = \frac{Volatile\ Seconds}{Entire\ Trip\ Druation} \times 100 \qquad \textit{Equation}\ (10)$$

Figure 17 shows a comparison between the volatility scores generated using acceleration bands versus using vehicular jerk bands for a sampled trip. Less volatile seconds were identified using jerk bands compared with using acceleration bands; volatility score was 8.5% with jerk bands vs. 6.0% with acceleration bands for the trips analyzed. The jerk-based volatile seconds are not always in concordance with volatile acceleration-based volatile seconds. That is to say, sometimes the driver accelerated at a higher than the upper band level but he/she did not jerk the vehicle during this period.

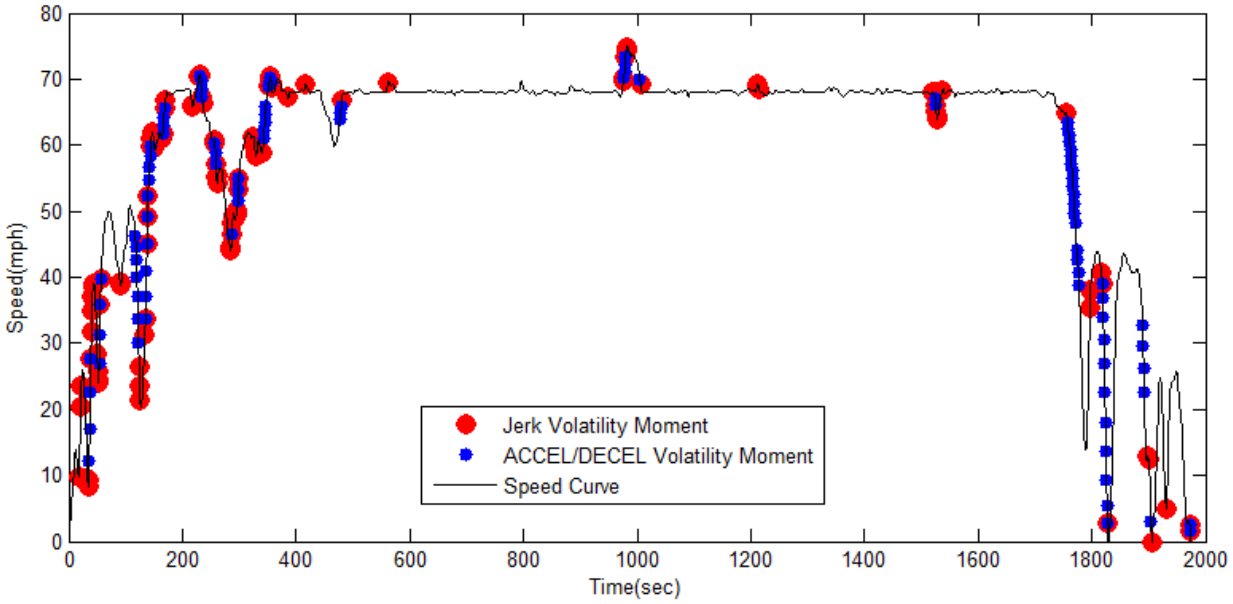


FIGURE 17 Volatile driving identified by different methods

Conceptually, it is important to understand and identify key decision points when the driver abruptly changes driving actions, e.g., goes from acceleration to deceleration. Based on the observations shown in Figure 17, jerk seems to capture critical decision points better

than acceleration while acceleration has more tolerance for volatility. Vehicular jerk can serve as an effective measurement to identify abrupt instantaneous decision changes. Since the volatility score is calculated for each trip, when data on multiple trips for a single driver are collected, average volatility score can be generated for each driver. This makes it is possible to compare both the intra-trip volatility and volatility between different drivers.

RESULTS – CORRELATES OF DRIVING VOLATILITY

After calculating the volatility scores (based on vehicular jerk bands) for each trip in the database, statistical models were estimated to investigate relationships between the volatility and driver demographics, vehicle characteristics and trip specifics. The database contained 51,370 trips made by 1,653 survey respondents in. After removing observations with missing information, the final database sample contained 40,240 trips by 1,486 respondents-—these are unique driver-vehicle pairs, labeled as driver-vehicle ID. Table 4 presents the descriptive statistics for the dependent and independent variables. The average volatility score is 13.84, which means that driving was volatile during 13.84% of the travel time (above or below mean vehicular jerk plus or minus one standard deviation). Some trips show calm driving (minimum score is 0.1%) while some were highly volatile when 55.46% of the time was spent on jerking vehicles at a higher level (outside of the bands).

In the final sample for modeling, 47.24% drivers were male; the mean age of respondent is 47.18, and a broad age range from 15 to 91. The mean vehicle age is 7.91 years and 43.88% of sampled vehicles were auto-sedans, 27.52% SUVs, and 13.59% pick-up trucks. As expected, 96.16% vehicles were gasoline-powered. 46.26% of trips were made during rush hours (6:00 am-10:00 am or 3:00 pm-7:00 pm); 24.37% were made on weekends; 19.49% were commute trips; the average trip duration was 14.17 minutes with an almost equal standard deviation–14.73. Overall, the data seems to be reasonable and in accordance with expectations.

TABLE 4 Descriptive statistics for dependent and independent variables

Variables			N	Frequency	Mean/Percent	Std. Dev	Min	Max
Dependent	Volatility Score		40240	-	13.840	6.701	0.1	55.46
Independent	Driver Variable	Gender [Male]	1486	702	47.24%	0.499	0	1
		Driver age (years)	1486	-	47.183	13.319	15	91
	Vehicle Age	Vehicle age (years)	1486	-	7.908	5.417	0	50
	Vehicle Type	Auto-sedan	1486	652	43.88%	0.496	0	1
		Two-seated	1486	58	3.90%	0.194	0	1
		Van	1486	131	8.82%	0.284	0	1
		RV	1486	3	0.20%	0.045	0	1
		SUV	1486	409	27.52%	0.447	0	1
		Station wagon	1486	31	2.09%	0.143	0	1
		Pickup	1486	202	13.59%	0.343	0	1
	Vehicle Fuel Type	Gasoline	1486	1429	96.16%	0.192	0	1
		Diesel	1486	29	1.95%	0.138	0	1
		Hybrid	1486	19	1.28%	0.112	0	1
		Flex fuel	1486	9	0.61%	0.078	0	1
	Trip Variable	Rush hour [Yes]	40240	18616	46.26%	0.499	0	1
		Weekend [Yes]	40240	9805	24.37%	0.429	0	1
		Trip duration (min)	40240	-	14.165	14.738	2.01	374.45
		Commute trip [Yes]	40240	7843	19.49%	0.396	0	1

Note: * Rush hours are AM (6:00 am-10:00 am) or PM (3:00 pm-7:00 pm)

The differences of volatility scores between trips can be result of the driving styles of different drivers (males vs. females, or young vs. older drivers), vehicle performance (new vehicles vs. older vehicles, body type, fuel type), or trip specifics (longer vs. shorter trips, commute vs. non-commute trips, and workday vs. weekend trips). Therefore simple Ordinary Least Squares (OLS) models were first estimated to test their associations. However, the traditional OLS models assume independence of observations and in this case multiple trips were made by the same drivers. Therefore, OLS will violate the independence assumption. One way to deal with correlated observations is to estimate a mixed-effect model, also called the mixed model. This model can capture correlated errors that arise from repeated observations in a group. In this study, the group variable is driver-vehicle pair; repeated variables are personal and vehicular characteristics; non-repeated variables are the measures for each specific trip. A "*Driver-Vehicle ID*" was created to represent different

driver-vehicle pairs in the sample and was used as the random term in the mixed-effects model. The random term quantifies the error due to repeated variables. The mixed-effects regression model can contain both fixed and random terms, as shown in following equations.

$$Y = \beta X + \gamma Z + \varepsilon \qquad \textit{Equation (11)}$$

$$\gamma \sim N(0, G)$$

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

Y is the response vector of volatility score for each trip in the data; *X* is a vector of fixed independent variables (age, gender, vehicle body type, fuel type, vehicle age, trip duration, commute or not, peak hour or off-peak, weekend or not); β is a vector of estimated fixed effects for matrix *X*; and *Z* is a vector of random independent variables (*Driver-Vehicle ID*); γ is a vector of estimated random effects for matrix *Z*; ε is a vector of unknown random errors; *G* is an diagonal matrix with identical entries for each fixed effect; I_n is an identity matrix; γ and ε are assumed to be independent.

Table 5 provides the modeling results for mixed models. Given that the distribution of vehicle jerk-based volatility scores is slightly right-skewed, square root transformed volatility score was tested as the dependent variable. However, the transformation improved the statistical properties of the model only marginally, e.g., significance of variables. Therefore, the original volatility score is used as the dependent variable, providing more intuitive parameter interpretation. Overall, the modeling results are reasonable, providing insights about a range of volatility correlates.

A key advantage of the mixed model over OLS model is that the random terms added into the mixed model structure can better model the effects of repeated observations within the group (driver-vehicle pair) by allowing various degrees of freedom for different variables according to their variations within groups. More specifically, all observations are treated equally in the OLS model regardless of their variations within or between groups. In this case, the overall sample size is 40,240 (the total number of trips). However, in the mixed model, only the sample size for generic variables (*32*), (i.e., trip characteristics) with variations within groups remains the same (40,240), while the sample size for alternative-

specific socioeconomic variables (i.e., driver and vehicle characteristics) become 1,486, which is the count of unique driver-vehicle pairs. As a result, larger standard errors are reported for alternative-specific socioeconomic variables in the mixed model. The estimated coefficients in the OLS and mixed models are nearly identical, but with different standard errors for driver and vehicle related terms, as expected. The following modeling interpretation is based on the mixed-effects model using the untransformed volatility score.

Full and final models are presented, with the final model containing only the statistically significant variables (10% level). The results of the final are discussed. The models have a reasonably good fit, explaining 40.3% of the variation in volatility score. As expected, younger drivers exhibit higher volatility in driving (5% level). A ten year increase in driver age is associated with a decrease of 0.57 in volatility scores. However, there is no statistical evidence for association between volatility score and drivers' gender. Driving volatility varies significantly with vehicle characteristics, including vehicle body type, vehicle age and fuel type. The results show that two-seat sports cars are associated with higher volatility, possibly due to their higher horse power. Trips made by two-seat sports cars drivers have 3.28 higher volatility scores, compared with trips made by drivers in the "base" category that includes sedans, RVs, station wagons, and SUVs. While van drivers show 1.82 lower volatility compared with drivers in the base category, perhaps due to their larger size and more sluggish performance. The use of hybrid vehicles shows lower volatility (-1.98) compared with gasoline and diesel vehicles. The volatility scores are lower for older vehicles, perhaps due to their engine performance. A year added to vehicle age is associated with a 0.10 units decline in the volatility score.

Volatility score also shows significant correlation with trip specific factors, including trip duration, time of day, day of the week, and trip purpose. Compared with non-rush hour trips, there is a 0.24 units increase in volatility score during rush hours; compared with workday trips, the decrease in volatility score for weekend trips is 0.30 units; for commute trips, the increase in volatility score is 0.36 units compared with non-commute trips; and a one-minute increase in trip duration is associated with a 0.04 units lower volatility score.

TABLE 5 Results of the mixed model using volatility score as the dependent variable

Dependent = Volatility Score		Full model		Final model	
Independent Variables		β	P-value	β	P-value
Constant		16.6983**	<.0001	17.6644**	<.0001
Driver Variables	Gender [Male]	-0.0018	0.9871	-	-
	Driver age (years)	-0.0573**	<.0001	-0.0574**	<.0001
Vehicle Age Variable	Vehicle age (years)	-0.1079**	<.0001	-0.1036**	<.0001
Vehicle Body Type Variable	Auto-sedan	Base		Base	
	Two-seated	3.8554**	<.0001	3.2830**	<.0001
	Van	-1.2621**	0.0084	-1.8231**	<.0001
	Recreational Vehicle-RV	-2.7353	0.1886	Base	-
	Sports Utility Veh.-SUV	0.3291	0.4249	Base	-
	Station wagon	-0.2914	0.6843	Base	-
	Pickup	-0.8836 *	0.0522	-1.5596**	<.0001
Vehicle Fuel Type Variable	Gasoline	Base		Base	
	Diesel	-0.9484	0.1760	Base	-
	Hybrid	-1.7512**	0.0295	-1.9825**	0.0101
	Flex fuel	1.8594*	0.0742	1.5765*	0.0947
Trip Variables	Rush hours [Yes]	0.2375**	<.0001	0.2376**	<.0001
	Weekend [Yes]	-0.3038**	<.0001	-0.3036**	<.0001
	Trip duration (min)	-0.0356**	<.0001	-0.0356**	<.0001
	Commute trip [Yes]	0.3627**	<.0001	0.3630**	<.0001
R^2		0.4028		0.4028	
R^2 Adjusted		0.4026		0.4027	
Root Mean Square Error-RMSE		5.2672		5.2672	
Mean of Response		13.8397		13.8397	
Observations (or Sum Weights)		40240		40240	
Bayesian Information Criterion-BIC		251937		251900	

Variance Component Estimates				
	Var. Comp.	Percent of Total	Var. Comp.	Percent of Total
Variance Between Driver-Vehicle Pairs	14.7136	34.66%	14.8319	34.84%
Remaining Variance	27.7429	65.34%	27.7430	65.16%
Total Variance	42.4564	100.00%	42.5749	100.00%

Note:
1. Rush hours: AM (6:00 am-10:00 am), PM (3:00 pm-7:00 pm);
2. ** = significant at a 95% confidence level;
3. * = significant at a 90% confidence level;
4. For mixed model, the random term is Driver-Vehicle ID (N=1486);
5. REML=Restricted Maximum Likelihood;
6. Statistically significant variables (90% level) are kept in the final model.

High levels of correlations among explanatory variables were checked and we did not

find them to be high. One example is that of commute trips which are typically made during peak hours. In the data, 46.28% of the trips were made during rush hours and 19.51% of the trips were for commute purposes. While these two variables capture different aspects of travel, i.e., time of day and trip purpose, the correlation between them was relatively low (0.156), justifying their joint inclusion in the model.

Examination of the random effects, reported as variance component estimates, shows a sizable variation (34.84%) in the volatility score across driver-vehicle pairs. This further justifies the use of the mixed model. Note that the models presented in this paper show an effort to test whether the measurement of volatility can be used to quantify the relationships between instantaneous driving decisions and other variables that include personal, vehicular, situational context factors. The random effects model confirmed that volatility score varies significantly between different driver-vehicle pairs. However, it does not fully disentangle volatility variations between different driving trips made by the same driver. A more sophisticated hierarchical modeling framework will be needed for answering such questions (*33*).

LIMITATIONS

This study depends heavily on GPS data collected by in-vehicle devices. To some extent the accuracy and availability of location data constrain the analysis. Compared with high industrial sampling rates (e.g. 96 kHz), these data are limited by relatively low sampling frequency which gives only second-by-second speeds. A reasonable question is whether second-by-second speed data are good enough for identifying instantaneous driving decisions. To address this issue, additional analyses were conducted by collecting driving data at 20 Hz using a driving simulator (*34*). This database includes 35,924 seconds speed data made by 24 drivers, generating 718,481 speed data points, which allows the investigation of micro-driving decision changes within one second. The results show that drivers made no change to their speed for 89.9% of the sampled seconds, i.e., drivers either kept accelerating, decelerating or just maintained speed during a second. Only 10.1% of the sampled seconds involve driver's decision change. Overall, the analysis found that at least 98.5% instantaneous driving decision changes can be detected using second-by-second data

compared with smaller intervals and that the second-by-second data are reasonably accurate for the purposes of this study.

Some other critical information remains unknown to the researchers due to privacy concerns. This includes the type of roads and the geo-codes for each second of driving. Missing geographically referenced information for trips prevents the researchers from extracting useful contextual factors. These include roadway segments used during trips and associated traffic counts, road geometry, traffic operations facilities, and surrounding land uses. Therefore, how the instantaneous decisions are associated with surrounding traffic, facility and land use can be analyzed adding interesting findings. This paper presents an attempt to enhance understanding of volatility in instantaneous driving decisions. More research is needed to investigate the impacts of network attributes, environmental attributes on instantaneous decisions, as shown in the conceptual framework. Expansion of the study can form the basis of future analysis of driver volatility and how it relates to energy, environment and safety.

CONCLUSIONS

In the context of using big data for traffic safety improvement, tailpipe emissions and energy use reduction in a driving dominant environment, it is essential to understand drivers' instantaneous driving decisions and their associated impacts. The research takes advantage of large-scale driving databases coupled by second-by-second GPS data to develop a framework for the research agenda in driving behavior studies addressing how to define the instantaneous driving decisions in a quantifiable way and how to quantify explicitly volatile driving in a defensible manner. The answer is to create a volatility indicator to measure the gap between an individual's driving practice and the typical driving practice in that region. Assuming the typical driving practice applied by most people represents the norm of driving culture in that region, the driving practices standing out of that norm could be defined as volatile driving. The paper demonstrates a methodology to measure the volatility, which is based on variance in vehicular jerk between individual drivers and regional sample profiles. The creation of a robust volatility score that is able to quantify the extent of volatility, instead of simply labeling a driver as aggressive or non-aggressive is a key contribution.

To create a typical driving profile for the study metropolitan area, acceleration or vehicular jerk distributions were analyzed using speed bins and enveloped by an upper and lower band (mean plus/minus one standard deviation). While typical driving practices are identified when the acceleration or vehicular jerk fall between the bands, volatile driving is defined as accelerations or vehicular jerks that fall out of the bands range. A volatility score for each trip or each driver can be calculated by the percent of travel time spent on volatile driving. In this sense, developing a regional driving profile is critical since this driving profile serves as a "standard" to define individual's driving volatility. Atlanta's driving profile was developed through an innovative visualization of data, the time spent on each driving behavior was calculated. Specifically, overall 14% of the travel time spent on high vehicular jerk; 7% of driving time was spent on idling or traveling at speeds below 5 mph, 47% of driving time was spent on acceleration, 41% of driving time was spent on deceleration and 5% of driving time was spent on maintaining constant speed. This information can be useful for designing driving cycle in a local context for better emissions estimations. The methodology has great potential to be expanded to measure driving volatility on road infrastructures as an indicator of roadway safety. Roads with higher risk (those experiencing more hard braking and negative jerks) can be identified and proactive strategies can be designed.

Individual level driving volatility also has a practical value. It can be potentially incorporated in advanced traveler information systems applications, e.g., driving behavior monitoring and feedback devices can use volatility information to provide alerts and warnings, network-based microscopic simulations and emission models can use volatility information for more accurate predictions (*35-37*). Drivers can check their volatility scores at the end of each day or even instantaneously, knowing where and when they exhibited high volatility. Moreover, statistical models confirmed that volatility scores significantly vary between drivers, which can support the early identification of risk-prone drivers. Volatile practices of risk-prone drivers can be potentially targeted through early warning systems. Applications can be embedded in navigation systems to send drivers warnings when they repeatedly show highly volatile driving during a trip.

References

1. Ji, S., C. Cherry, M. Bechle, Y. Wu, and J. Marshall, Electric Vehicles in China: Emissions and Health Impacts. *Environmental Science & Technology*, Vol. 46, No. 4, 2011: pp. 2018-2024.DOI: 10.1021/es202347q.
2. Wang, X., A. Khattak, and Y. Zhang, Is Smart Growth Associated with Reductions in Carbon Dioxide Emissions? *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2375, No. -1, 2013: pp. 62-70.DOI: 10.3141/2375-08.
3. U.S. Environmental Protection Agency. *Techinical Guidance on the Use of MOVES2010 for Emission Inventory Preparation in State Implementation Plans and Transportation Conformity*. 2010 [cited 2014 08-01]; Available from: http://www.epa.gov/otaq/models/moves/420b10023.pdf.
4. Transportation Research Board, *TRB Special Report 254: Managing Speed: Review of Current Practice for Setting and Enforcing Speed Limits*. Transportation Research Board, 1998.
5. Miles, D. and G. Johnson, Aggressive driving behaviors: are there psychological and attitudinal predictors? *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 6, No. 2, 2003: pp. 147-161.DOI: http://dx.doi.org/10.1016/S1369-8478(03)00022-6.
6. Shinar, D., Aggressive driving: the contribution of the drivers and the situation. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 1, No. 2, 1998: pp. 137-160.
7. National Highway Traffic Safety Administration, *Countermeasures that work: A highway safety countermeasure guide for state highway safety offices, Report No. DOT HS 811081*. National Highway Traffic Safety Administration, 2009.
8. James, L. and D. Nahl, *Road rage and aggressive driving: Steering clear of highway warfare*. 2000: Amherst, New York: Prometheus Books.
9. Nesbit, S. and J. Conger, Predicting aggressive driving behavior from anger and negative cognitions. *Transportation Research Part F: Traffic Psychology and*

Behaviour, Vol. 15, No. 6, 2012: pp. 710-718.DOI: http://dx.doi.org/10.1016/j.trf.2012.07.003.

10. Underwood, G., P. Chapman, S. Wright, and D. Crundall, Anger while driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 2, No. 1, 1999: pp. 55-68.DOI: http://dx.doi.org/10.1016/S1369-8478(99)00006-6.
11. Tasca, L. *A Review of the Literature on Aggressive Driving Research*. 2000 [cited 2014 11-03]; Available from: http://www.stopandgo.org/research/aggressive/tasca.pdf.
12. Lajunen, T. and H. Summala, Can we trust self-reports of driving? Effects of impression management on driver behaviour questionnaire responses. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 6, No. 2, 2003: pp. 97-107.DOI: http://dx.doi.org/10.1016/S1369-8478(03)00008-1.
13. Shinar, D. and R. Compton, Aggressive driving: an observational study of driver, vehicle, and situational variables. *Accident Analysis & Prevention*, Vol. 36, No. 3, 2004: pp. 429-437.DOI: http://dx.doi.org/10.1016/S0001-4575(03)00037-X.
14. Langari, R. and W. Jong, Intelligent energy management agent for a parallel hybrid vehicle-part I: system architecture and design of the driving situation identification process. *Vehicular Technology, IEEE Transactions on*, Vol. 54, No. 3, 2005: pp. 925-934.DOI: 10.1109/tvt.2005.844685.
15. Murphey, Y., *Intelligent Vehicle Power Management: An Overview*, in *Computational Intelligence in Automotive Applications*, D. Prokhorov, Editor. 2008, Springer Berlin Heidelberg. p. 169-190.
16. Kim, H., J. Oh, and R. Jayakrishnan, Application of Activity Chaining Model Incorporating a Time Use Problem to Network Demand Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1977, No. 2006: pp. 214-224.
17. Kim, E. and E. Choi, *Estimates of Critical Values of Aggressive Acceleration from a Viewpoint of Fuel Consumption and Emission*, in *TRB 2013 Annual Meeting*. 2013: Washington D.C.

18. 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.DOI: http://dx.doi.org/10.1016/S1352-2310(00)00217-X.
19. Ericsson, E., Independent driving pattern factors and their influence on fuel-use and exhaust emission factors. *Transportation Research Part D: Transport and Environment*, Vol. 6, No. 5, 2001: pp. 325-345.
20. Renski, H., A. Khattak, and F. Council, Effect of Speed Limit Increases on Crash Injury Severity: Analysis of Single-Vehicle Crashes on North Carolina Interstate Highways. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1665, No. -1, 1999: pp. 100-108.DOI: 10.3141/1665-14.
21. Paleti, R., N. Eluru, and C. 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.
22. Sivak, M. and B. Schoettle, Eco-driving: Strategic, tactical, and operational decisions of the driver that influence vehicle fuel economy. *Transport Policy*, Vol. 22, No. 0, 2012: pp. 96-99.DOI: http://dx.doi.org/10.1016/j.tranpol.2012.05.010.
23. Nam, E., C. Gierczak, and J. Butler, *A Comparison Of Real-World and Modeled Emissions Under Conditions of Variable Driver Aggressiveness*, in *Annual TRB Meeting*, T.A.M. CD-ROM, Editor. 2003: Washington DC.
24. Rouphail, N., C. Frey, J. Colyar, and A. Unal, *Vehicle Emissions and Traffic Measures: Exploratory Analysis of Field Observations at Signalized Intersections*, in *80th Annual Meeting of the Transportation Research Board*. 2001: Washington D.C.
25. Nam, E., *Proof of Concept Investigation for the Physical Emissions Estimator (PERE) for MOVES, EPA420-R-03-005*. Office of Transportation and Air Quality, EPA, 2003.
26. Holmén, B. and D. Niemeier, Characterizing the effects of driver variability on real-world vehicle emissions. *Transportation Research Part D: Transport and Environment*, Vol. 3, No. 2, 1998: pp. 117-128.DOI: http://dx.doi.org/10.1016/S1361-9209(97)00032-1.

27. PTV Nustats in Association with Geostats. *Atlanta Regional Commission Regional Travel Survey Final Report*. 2011 [cited 2014 08-01]; Available from: http://www.atlantaregional.com/File%20Library/Transportation/Travel%20Demand%20Model/tp_2011regionaltravelsurvey_030712.pdf.

28. Atlanta Regional Commission. *Atlanta Regional Commission (ARC) Regional Travel Survey*. 2011 [cited 2014 08-01]; Available from: http://www.atlantaregional.com/transportation/travel-demand-model/household-travel-survey.

29. 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. Washington DC.

30. Berry, I., *The Effects of Driving Style and Vehicle Performance on the Real-World Fuel Consumption of U.S. Light-Duty Vehicles*, in *Department of Mechanical Engineering and the Engineering Systems Division*. 2010, Massachusetts Institute of Technology.

31. Ahn, K., H. Rakha, A. Trani, and M. Van Aerde, Estimating Vehicle Fuel Consumption and Emissions based on Instantaneous Speed and Acceleration Levels. *Journal Of Transportation Engineering*, Vol. 128, No. 2, 2002: pp. 182-190.DOI: doi:10.1061/(ASCE)0733-947X(2002)128:2(182).

32. Ben-Akiva, M. and S. Lerman, *Discrete choice analysis: theory and application to travel demand*. 1985, Cambridge, Mass: MIT press.

33. Liu, J., A. Khattak, and X. Wang, The Role of Alternative Fuel Vehicles: Using Behavioral and Sensor Data to Model Hierarchies in Travel. *submitted to Transportation Research Part C: Emerging Technologies*, Vol. No. 2014: pp.

34. Liu, J., A. Khattak, and L. Han, *What is the Magnitude of Information Loss When Sampling Driving Behavior Data?*, in *94th Annual Meeting of the Transportation Research Board*. 2015: Washington D.C.

35. Khattak, A., J. Liu, and X. Wang, Supporting instantaneous driving decisions through vehicle trajectory data. *Submitted to Journal of Intelligent Transportation Systems*, Vol. No. 2015: pp.

36. Bandeira, J., T.G. Almeida, A.J. Khattak, N.M. Rouphail, and M.C. Coelho, Generating emissions information for route selection: Experimental monitoring and routes characterization. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, Vol. 13, No. 1, 2013: pp. 3-17.DOI: 10.1080/15472450.2012.706197.

37. Bandeira, J.M., D.O. Carvalho, A.J. Khattak, N.M. Rouphail, T. Fontes, P. Fernandes, S.R. Pereira, and M.C. Coelho, Empirical Assessment of Route Choice Impact on Emissions Over Different Road Types, Traffic Demands, and Driving Scenarios. *International Journal of Sustainable Transportation*, Vol. No. 2014: pp. null-null.DOI: 10.1080/15568318.2014.901447.

STRUCTURING AND INTEGRATING DATA IN METROPOLITAN REGIONS TO EXPLORE MULTI-LEVEL LINKS BETWEEN DRIVING VOLATILITY AND CORRELATES[3]

Abstract – This study demonstrates how large-scale data can be transformed into useful knowledge. This is done by creating a framework for combing data from multiple sources and comparing counties/regions in terms of volatility of resident drivers. Higher driving volatility (e.g., hard accelerations or braking) can imply unsafe outcomes, more energy use, and higher emissions. A unique database was integrated from four sources that include two large-scale travel surveys, historical traffic counts from California and Georgia Department of Transportation, socio-demographic information from Census, and geographic information from Google Earth. The database provides a rich resource to test hypothesis and model driving decisions at the micro-level, i.e., second-by-second. The large-scale travel survey data includes 117,022 trips made by 4,560 drivers residing in 78 counties of major metropolitan areas (Los Angeles, San Francisco, Sacramento, and Atlanta) across two states, representing various land use types and populations; all trips were recorded by in-vehicle GPS devices giving 90,759,197 second-by-second speed records. The study contributes by demonstrating a way to integrate data from multiple sources to explore links between naturalistic driving behaviors and various factors structured in hierarchies. That is, the data are structured at the levels of trips, drivers, counties, and regions. Appropriate hierarchical models are estimated to study correlates of driving performance and compare driving performance across regions.

Keywords: Data integration, naturalistic driving behavior, multi-level modeling

[3] Material based on: Liu, J., A. Khattak & M. Zhang, Exploring Links between Naturalistic Driving Behaviors and Various Factors in Hierarchies: A Study Integrating Multiple Data Sources, Accepted for presentation at 2015 Road Safety & Simulation International Conference, Orlando, FL, 2015. A revised version of this paper, titled "Structuring and Integrating Data in Metropolitan Regions to Explore Multi-Level Links between Driving Volatility and Correlates," was submitted to 2016 Transportation Research Board for review. This chapter also contains content from a paper: "Liu J., A. Khattak & X. Wang. Creating Indices for How People Drive in a Region: A Comparative Study of Driving Performance, TRB paper # 15-0966. Presented at the Transportation Research Board Annual Meeting, National Academies, Washington, D.C., 2015."

INTRODUCTION

Utilizing transportation-related data to perform driving-related studies is one compelling direction to uncover and ameliorate transportation-related problems, e.g., safety, energy consumption and emissions. With newly available information technologies, transportation data can be pulled from conventional and emerging data sources. Conventional data sources normally refer to those documenting roadway parameters, roadside elements, land use, safety facts and relevant transportation plans. Emerging data are typically generated by electronic sensors, such as Global Positioning Systems (GPS), video, Bluetooth, social media and weather reporting system (RWIS).

Understanding driving performance is key to implementation of transportation improvement strategies, and big data can help in this regard (*1*). Driving behaviors have been quantified in terms of driving volatility using large-scale naturalistic driving data (*2*). Driving volatility captures the instantaneous decisions about speed, acceleration, and vehicular jerk. It is expected that higher driving volatility (e.g., hard accelerations or braking) are associated with worst safety outcomes, higher energy consumption and emissions. With increasing amounts of information, generated by sensors from various sources that include travelers, vehicles, infrastructure and the environment coupled with social, economic and spatial data, the relationships between driving volatility and associated factors can be explored in a more comprehensive way.

A number of factors have been linked in the literature to explore associations of driving behaviors with various factors, such as law enforcement (*3, 4*), road network, road type and traffic conditions (*5-8*), terrain (*9, 10*), weather (*11-14*), driver education (*15-17*), vehicle type (*2, 18*), and driver demographics (*19-24*). Driving behavior is complex and it can be conceptualized as embedded in a hierarchical structure (*25*). Drivers in the same area face similar road network, terrain, and are influenced by the similar driving cultures (*26, 27*). However, drivers also have their own characteristics, such as gender, age, education, income, employment, etc. Further, while drivers are making different trips, their driving behaviors are influenced instantaneously by the trip features, such as time of day, trip length, trip purpose, etc. The increasing availability of a large number of factors structured in hierarchies increases the importance of applying appropriate statistical models. This study

explores how large-scale data can be structured and integrated in metropolitan regions to explore multi-level (trips, drivers, counties, and regions) links between driving volatility and its correlates.

DATA SOURCES

The data used in this study are extracted from four major sources: 1) Naturalistic driving data from regional travel surveys; 2) Geographic information about road networks from Google Earth; 3) Contextual data from Census; 4) Exposure data from the historical traffic data reported by the State Department of Transportation,

Naturalistic Driving Data

The naturalistic driving data used in the study were collected through regional travel surveys, including 2012-2013 California Household Travel Survey (CHTS) and 2011 Atlanta Regional Travel Survey (ARTS) (*28, 29*). The data is managed by Transportation Secure Data Center under the National Renewable Energy Laboratory (*30*). The data from CHTS cover 58 counties across the State of California and the data from ARTS cover 20 counties in the region of Atlanta Regional Commission. The data include 117,022 trips made by 4,560 drivers residing in 78 counties across two states, representing various land use types and populations; all trips were recorded by in-vehicle GPS devices giving 90,759,197 second-by-second speed records (*31*).

Geographic Information

Driving behaviors are highly correlated to the geographic features of road networks (*7, 8*). This study extracts the geographic information of road networks within a county and explores the county-level correlations between the road geographic and regular driving practices in a county. Data were pulled from the Google Earth (https://www.google.com/earth/). Geographic information explored in the study includes road network pattern (gridiron or non-gridiron), terrain (flat, rolling or hilly), elevation, coast (yes or no) and big city (yes or no).

Contextual Data

To account for the influence of contextual factors on driving behaviors, this study pulled county-level data from Census.com. The key factors that are considered in the investigation include population, population density, percent of residents 65 years or over, percent of female, percent of persons with a Bachelor's degree or higher, mean commute time, median household income and percent of persons in poverty. All these factors are hypothetically correlated to the driving behaviors in a county.

Exposure Data

The exposure data were pulled from the historical traffic data reported by the State Department of Transportation. Highway Performance Monitoring System (HPMS) under California Department of Transportation provides yearly traffic records for each county and city (*32*) and TravelSmart of Georgia Department of Transportation documents county-level traffic records as well (*33*). The exposure data used in the study include rural and urban road mileage within a county, and rural and urban daily vehicle miles traveled at the county-level. The ratio of numbers distinguishing the rural and urban areas can be useful in understanding the driving performances associated with the urbanized level of a county.

MODELING STRUCTURE

A key objective of this study is to explore the links between naturalistic driving behaviors and various factors structured in hierarchies. There are numerous measurements that have been utilized to characterize driving behaviors, including speed (*34, 35*), acceleration (*36-38*), acceleration noise/variations (*39, 40*), and vehicular jerk (*41*). Vehicular jerk is the derivative of acceleration (or the second derivative of speed), and is able to capture the instantaneous change of driving decisions (e.g., transitions from accelerating to decelerating). The authors' previous papers (*2, 42*) have developed a measure, termed driving volatility, to capture the instantaneous decisions about speed, acceleration, and vehicular jerk simultaneously. The driving volatility is defined as the percentage of "extreme" driving seconds (i.e., large vehicular jerk values) over the duration of an observation period (e.g., one

trip, or all trips made by one driver). Large values imply extreme variability of instantaneous driving decisions. The measure of driving volatility is proposed for qualifying driving behaviors relying on the large-scale trajectory data. More details about the driving volatility are available from our previous papers. Driving volatility is used as a key measure of the driving behaviors for the modeling.

The dataset structured using the data pulled from aforementioned sources contains a hierarchy, as shown in Figure 18. In the survey one driver could make multiple trips. Trips made by the same driver are not completely independent form each other, though trip-related factors (e.g., length, time and trip purpose) may vary across trips. Since factors driver's own attributes would not change within in the same driver making multiple trips, the assumption of independence of observations in traditional ordinary least square (OLS) models is violated. Trips are nested in drivers. The same nesting relationships can also be found between drivers and counties, and counties and regions. Note that, to see whether there is a significant differences in driving behaviors between major regions, this study partitioned the data into six regions: four metropolitan areas, delineated according to 2013 Census Metropolitan Statistical Areas (*43*), including Los Angeles Metropolitan Area (Los Angeles-Long Beach-Riverside), San Francisco Metropolitan Area (San Jose-San Francisco-Oakland) and Sacramento Metropolitan Area (Sacramento-Arden-Arcade-Yuba City) and the Atlanta metropolitan area, and California central valley area (Fresno-Stockton), and other areas that are not included in any specified regions.

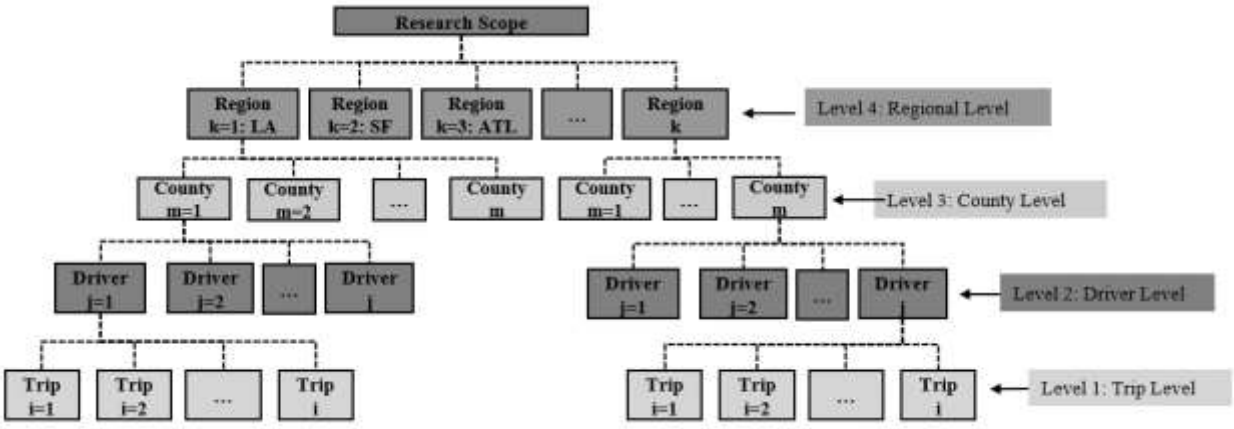


FIGURE 18 Hierarchical data structure

One method to statistically account for hierarchical structure of the data is to use multi-level or hierarchical modeling. Multi-level modeling can accommodate non-independence of observations. Given the four levels embedded in the data, a four-level model is applied in the study. The formulations of the four-level hierarchical modeling are shown in simplified forms as following.

The level-1 (trip level) model is a regression model with dependent variable Y (driving volatility) and trip-related factors.

$$Y = \beta^{(1)} X^{(1)} \qquad \textit{Equation } (11)$$

Where,

Y = a vector of responses—trip-level driving volatility;

$X^{(1)}$= a vector of level-1 factors (including the intercept term), e.g., trip length and trip purpose;

$\beta^{(1)}$= a vector of coefficients for level-1 factors, including the intercept;

The level-1 models treat trips made by the same driver as a whole observational set and generate one coefficient for each level-1 factor for this driver. These coefficients can be regarded as being fixed within one driver and used to predict response (e.g., driving volatility) given a set of trip specifics only for the same driver. For all drivers, there are thousands of coefficients for each level-1 factor. Thus, coefficients generated in level-1 models are random to thousands of drivers. The mean and variance of these random coefficients can be predicted by level-2 (driver level) predictors (e.g., driver age, gender, vehicle body type and fuel type), level-1 models discover the variance of driving performance within a driver, and level-2 models are used to explain the variance across drivers.

The level-2 (driver level) includes a series of regression models using coefficients $\beta^{(1)}$ from level-1 models as the dependent variable. Note that, β includes the intercept term. If only the intercepts from level-1 models are allowed to vary across the drivers, level-2 models are termed as random intercept models. The formulations shown below also allow the

coefficients from level-1 models vary across drivers, which is referred to as random intercept and slope model.

$$\beta^{(1)} = \beta^{(2)} X^{(2)} \qquad \textit{Equation } (12)$$

Where,

$\beta^{(1)}$ = a vector of coefficients from level-1 models;

$X^{(2)}$ = a vector of level-2 factors, e.g., driver age, gender, vehicle body type and fuel type;

$\beta^{(2)}$ = a vector of coefficients for level-2 factors.

Level-2 models explain the mean and variance of random coefficients of level-1 factors. Similarly, they treat all level-1 coefficients in one county as a whole observational set and generate one coefficient for each level-2 factor for the county. Through explaining coefficients for level-2 predictors, level-3 (county level) models can be estimated to uncover the variance across counties. Level-3 includes a series of regression models using level-2 coefficients, $\beta^{(2)}$ as the responses.

$$\beta^{(2)} = \beta^{(3)} X^{(3)} \qquad \textit{Equation } (13)$$

Where,

$\beta^{(2)}$ = a vector of coefficients from level-2 models;

$X^{(3)}$ = a vector of level-3 factors, e.g., terrain, road network pattern, population;

$\beta^{(3)}$ = a vector of coefficients for level-3 factors.

And so on, the level-4 model is estimated to examine the variations of level-3 coefficients within each county.

$$\beta^{(3)} = \beta^{(4)} X^{(4)} \qquad \textit{Equation } (14)$$

Where,

$\beta^{(3)}$ = a vector of coefficients from level-3 models;

$X^{(4)}$ = the level-4 factor, i.e., regional dummy variables;

$\beta^{(4)}$ = a vector of coefficients for level-4 factors.

Thus, only factor(s) at the highest level have stationary coefficients, and factors in three lower levels have coefficients varying across their nested groups. To report the variations in coefficients of factors at the three levels, hierarchical modeling estimation provides two components of coefficients for these factors: fixed effects and random effects. Fixed effects can be viewed as the general correlations between factors and responses. Random effects reflect the variations of the coefficients. Fixed effects are typically of primary interest and give a sense of how the factors are associated with the response variable. Random effects are estimated to accommodate variations across groups (i.e., drivers, counties, and regions). Finally, the four-level hierarchical modeling structure can be written in a simple way:

$$Y = \beta X + \gamma^{(2)}Z^{(2)} + \gamma^{(3)}Z^{(3)} + \gamma^{(4)}Z^{(4)} \qquad \textit{Equation } (15)$$

Where,

Y = a vector of responses, e.g., trip-level driving volatilities;

X = a vector of all factors;

β = a vector of the fixed components of coefficients for all factors;

$Z^{(2)}$ = a vector of the group factors at level-2, e.g., driver age, gender and vehicle body type;

$\gamma^{(2)}$ = a vector of random-effects at level-2;

$Z^{(3)}$ = a vector of the group factors at level-3, e.g., terrain and road network pattern;

$\gamma^{(3)}$ = a vector of random-effects at level-3;

$Z^{(3)}$ = a vector of the group factor at level-4, e.g., region dummy variable;

$\gamma^{(3)}$ = a vector of random-effects at level-4;

The output of hierarchical modeling generally has three components: the effect estimation of fixed-effects parameters, variance estimation of random-effects parameters and model goodness of fit.

RESULTS

Regional Driving Performance Comparisons

Time Use Distributions

Figure 19 presents the time spent in the four regions on acceleration, deceleration and constant speed by speed range in 0.5 mph increments, as well as standardized time shares, i.e., time shares for speed bins. Time spent accelerating or braking varies with speeds.

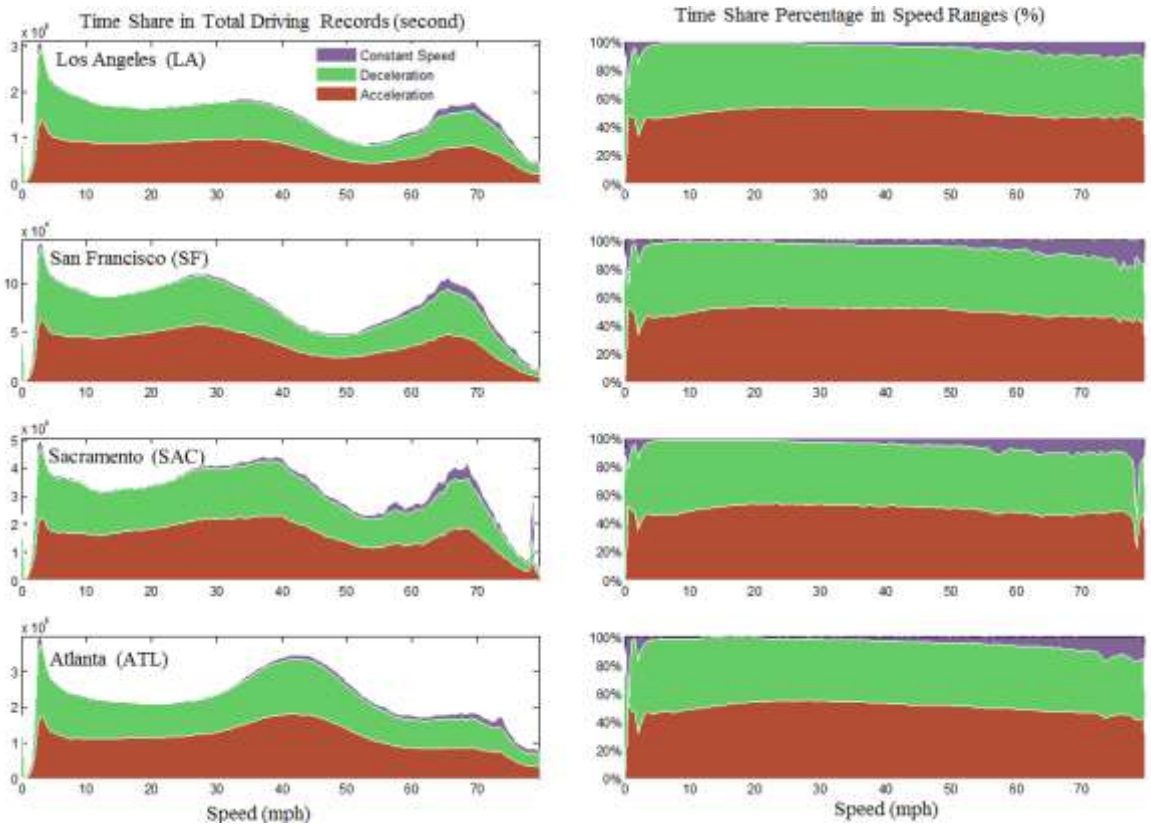


FIGURE 19 Time use distribution of acceleration, deceleration and constant speed in study regions

The findings regarding regional comparisons are:

- Regions in California show similarity in terms of larger time shares in high speeds ranges (60 ~70mph), which is different from Atlanta. This likely depends on a host of different factors that include road networks and trip lengths. Trips in regions of California are relatively longer (average-14 minutes) in the sampled datasets than those in Atlanta (average- 12 minutes). Longer urban trips are typically associated with a greater possibility of driving

on interstates or expressways. Highway and arterial density of a region might be another reason for regional differences. Compared with California regions, Atlanta has the least density of highways and arterials (*44*, *45*).

- For normal local driving practices (10~50 mph), there is no clear peak time in Los Angeles. Traffic congestion is a likely contributing reason. According to the 2012 Urban Mobility Plan (*46*), Los Angeles region is ranked highest in terms of congestion index among these four regions.
- Atlanta has a clear peaking of time spent at around 40 mph, and this region has the smallest congestion index (*46*).
- Driving in San Francisco Bay Area seems shows more time spent in lower speed ranges, partly because of the hilly terrain and strong grades in the region.

Note that, very small acceleration or deceleration rates (0.04 ft/s^2, based on the 5^{th} percentile of speed changes in one second) were considered noise and coded as constant speed. Standardized time shares show that driving time is mostly devoted to accelerating or decelerating rather than maintaining speed. Acceleration and deceleration have about equal time share. Increasing time is spent on maintaining constant speed when speeds are higher (>60 mph). Notably, Sacramento region has a large time share percentage of constant speed at speeds around 76 mph. The large shares of constant speed possibly associate with traffic congestion on freeways, and to some extent the use of cruise control. Unfortunately, the information about the use of cruise control was not recorded or released to the public in the database. Comparisons between regions reveal that Atlanta shows lower times spent on driving at freeway speeds of 70 mph or above.

Acceleration and Vehicular Jerk Distributions

The distributions of micro driving patterns reveal the magnitude/intensity of micro driving decisions (accelerating, decelerating, changing acceleration/deceleration). Figure 20 shows the distribution of quantified acceleration/deceleration and vehicular jerk patterns along with the speeds. Owing to the large number of observations (N= 78.7 million records), the four regions have similar distributions. The overlapping distributions are presented but with the

means and two-standard-deviations separately indicated for each region in the Figure. The gradual change of color shows the concentration extent of magnitude to zero (i.e. constant speed or zero acceleration/jerk). The mean and two-standard-deviations are for each speed range (0.5 mph increments) is plotted. Major findings are:

- San Francisco Bay Area has smaller magnitudes of accelerations (closer to zero) than other areas. This may be associated with the hilly terrain and strong grades in the region.
- Los Angeles and Sacramento regions have closer means and standard deviations across speed bins. Both regions have grid road patterns with similar terrains and road densities, so similar driving performance in terms of acceleration is expected and observed.
- Atlanta region seems to have larger means and standard deviations.
- In general, large accelerations and decelerations occur at lower speeds (10 ~ 30 mph) and after speeds reach higher ranges (> 40 mph) the magnitude tends to zero, as shown in Figure 20(i).
- Acceleration/deceleration rates are not distributed homogeneously along with the speeds.
- There are no apparent regional differences at speeds less than 15 mph or above 60 mph, in terms of mean and standard deviation shown in Figure 20(ii).

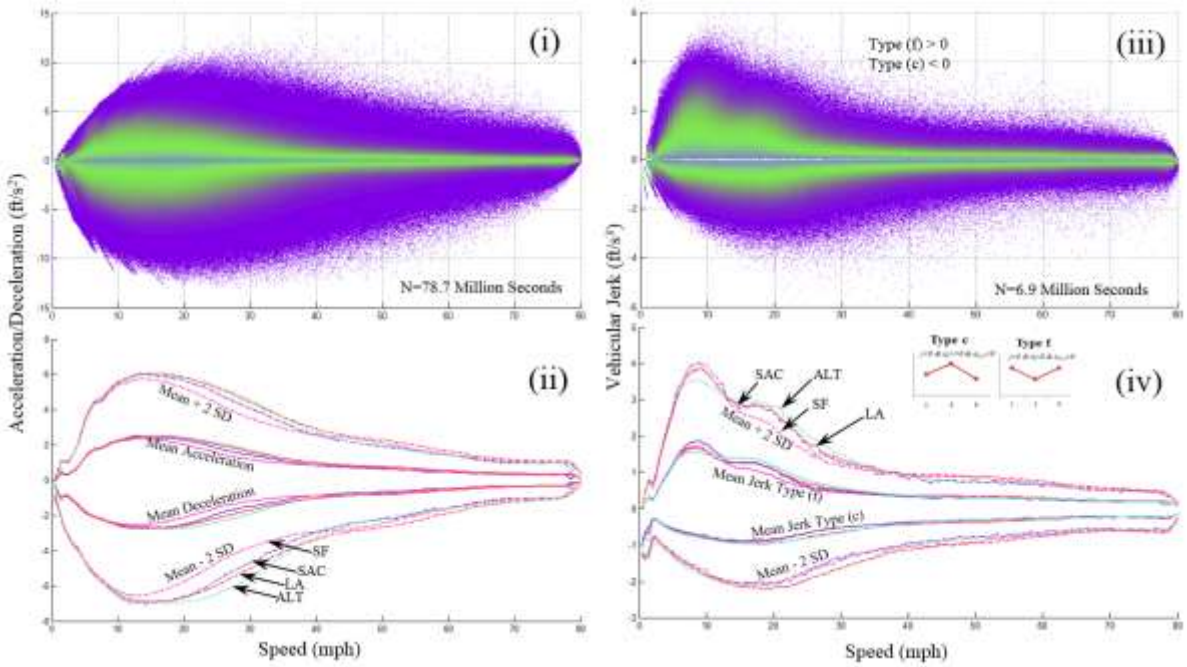


FIGURE 20 Distributions of acceleration and jerk at various speeds

Six types of vehicular jerk patterns show different distributions. Owing to the limited space, only Types (c) and (f) are presented in this paper, as shown in Figure 20 (iii) and (iv). Types (c) and (f) may represent the most volatile micro-driving patterns with acceleration chained with deceleration in a very short time frame. Unlike the acceleration distribution, vehicular jerk does not have a symmetric distribution. The positive jerk patterns have larger magnitudes than negative ones. The large magnitudes of jerk are mainly observed at speeds of 5~20 mph. After speeds reach 30 mph, the vehicular jerk magnitudes are relatively constant around zero.

Type (f) has positive values and SF has smaller magnitudes than other three regions at speeds 15~30 mph. The small magnitudes are associated with smooth or careful driving. The hilly terrain in SF may encourage drivers to change their micro-driving decisions (i.e. Type (f), from decelerating to accelerating) more smoothly. LA and SAC have close magnitudes in Type (f) and ATL has the largest magnitudes. There are no significant differences between regions in terms of the Type (c) driving pattern. SF and SAC have close magnitudes that are smaller than regions LA and ATL. Overall, this study observed and quantified heterogeneity of micro driving patterns at different speeds.

Combined Distribution of Time Use and Accelerations/Vehicular Jerk

Due to space limitations, only the overall distributions of speeds, accelerations and time use can be presented (N=78.7 million records). Figure 21 shows a wealth of information about vehicle performance; distributions for each region are available from the authors. The height shows the number of driving records with corresponding driving status (i.e., speed and acceleration/deceleration or vehicular jerk).

At speeds 10 ~30 mph there are fewer driving records with zero acceleration or deceleration (see the trough in Figure 21); for higher speeds (> 60 mph), a large portion of time is spent in maintaining speed with small acceleration or deceleration (see the ridge in Figure 21). Differing from acceleration distributions, vehicular jerk distributions are more concentrated at zero. This implies that any quantified jerk patterns that are different from zero can be easily identified as abnormal micro driving patterns, e.g., sudden braking or accelerating.

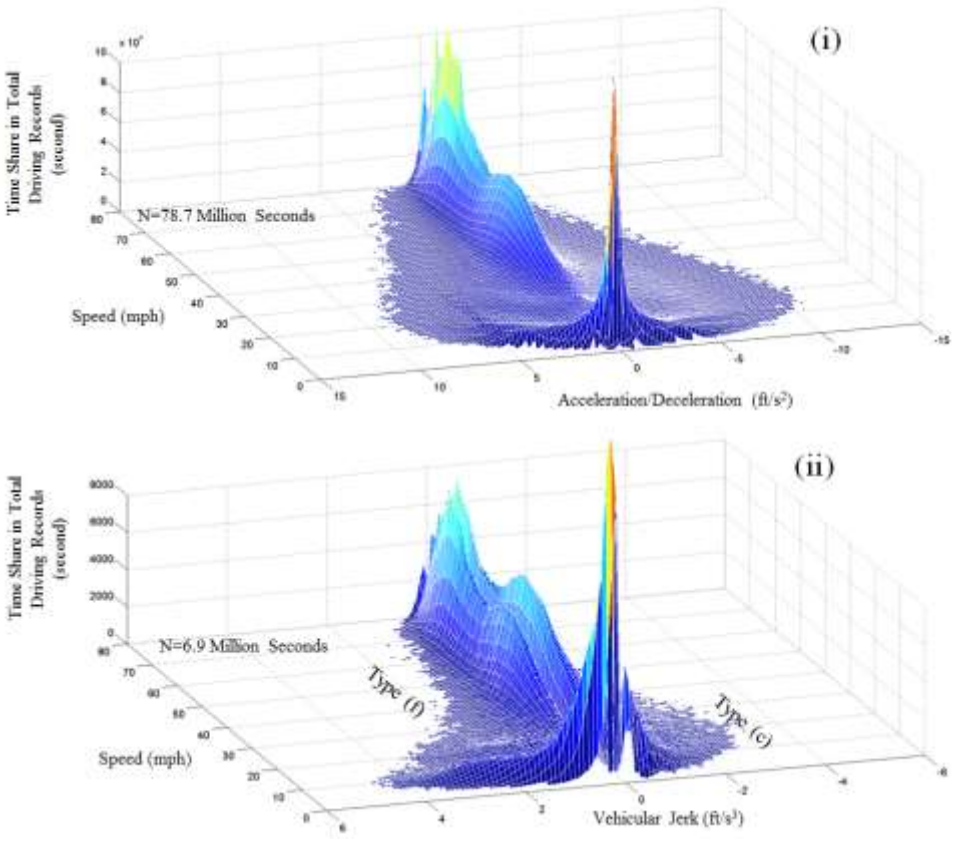


FIGURE 21 Combined distributions of speed, acceleration (or vehicular jerk), and time use

Regional Driving Index

To generate a new measure of driving performance at the regional level, similar to Congestion Index in the Urban Mobility Report (*30*), this study created the following two indices:

$$Regional\ Accleration\ Index = \sum_{i=1}^{n}\left(\frac{v_i * t_i}{\sum_{i=1}^{n}(v_i * t_i)} * |a_i|\right) \qquad \textit{Equation (16)}$$

$$Regional\ Vehicular\ Jerk\ Index = \sum_{i=1}^{n}\left(\frac{v_i * t_i}{\sum_{i=1}^{n}(v_i * t_i)} * |j_i|\right) \qquad \textit{Equation (17)}$$

Where,

v_i= speed record of sampled vehicle during i^{th}time slice in selected region, *i*=1, 2, 3, …, *n*, *n* is the total driving records for one region (e.g., for LA: 24,185,380 seconds);

t_i = duration of i^{th}time slice, i.e., one second if using the second-by-second data;

a_i= acceleration during i^{th}time slice;

j_i= vehicular jerk during i^{th}time slice;

$\sum_{i=1}^{n}(v_i * t_i)$ = total distance traveled in the sample in one region.

These two indices represent the intensity and variability of instantaneous driving decisions respectively. They can be used to compare the driving patterns across metropolitan areas. If Time is sliced equally, the formals can be simplified to speed-weighted mean. Using the data, LA has 24,185,380 time slices (*n*=24,185,380); SF has *n*=12,579,345; SAC has *n*=5,229,874 and ATL has *n*=36,715,308 slices. Each time slice consists of one second and the speed and acceleration for each second is known. The results are shown below:

TABLE 6 Regional Driving Index

Region	Acceleration	Vehicular Jerk
LA:	0.951	0.268
SF:	0.826	0.250
SAC:	0.847	0.241
ATL:	0.909	0.254

Hypothetically, the acceleration index can range from 0 up to 15 or even 20 ft/s^2, and similarly for vehicular jerk index. Acceleration index captures the intensity of micro-driving decisions and jerk index captures the variability in micro-driving decisions in a region (using a sample). The results show that overall LA has the largest values for both indices and ATL is ranked second. SAC has a larger acceleration index than SF but the vehicular jerk index is smaller.

Comparisons of County-level Characteristics

Table 7 summarizes the basic information for five selected counties, pulled from multiple sources for illustrative purposes. Particularly, the travel patterns and driving behaviors are extracted using naturalistic driving data from millions of driving records. Comparisons of selected counties (related to large cities) are made in terms of travel patterns and driving behaviors quantified using large-scale trajectory data. T-tests were done with the county of Los Angeles. Key results from the comparisons are:

- On average, drivers in Fulton County (Atlanta) reported making more daily trips than four California counties, i.e., the county of Los Angeles, San Francisco, Sacramento and San Diego. The difference is statistically significant ($p<0.05$). No significant differences were observed among four California states.
- The daily distances travelled for drivers in San Francisco were longer than Los Angeles drivers, as well as the lengths and durations of trips.
- Drivers in San Francisco spent significantly longer times traveling daily than those in Los Angeles, as well as Fulton drivers. Sacramento and San Diego drivers spent less time traveling daily than Los Angeles drivers.
- In terms of the trip length and duration, drivers in Fulton County made significantly shorter trips than Los Angeles drivers. While slightly longer trips were made by San Diego drivers, their trip durations were shorter than Los Angeles drivers'.
- Drivers in Los Angeles spent significantly more time on idling and less time on extended stable driving, than other four counties.
- Drivers in Fulton spent more time on accelerating and less time on decelerating than those in Los Angeles than other four counties.

- In terms of the trip mean speed, drivers in San Diego and Fulton had a higher mean speed than drivers in other counties.
- In terms of the mean acceleration/deceleration, drivers in Los Angeles made greater accelerations/decelerations than drivers in other counties.
- Drivers in Los Angeles show greater vehicular jerk values (including positive and negative values) than those in other counties.
- Compared with trips made in Los Angeles, trips made by drivers in San Francisco were associated with a higher level of driving volatilities (in terms of the mean driving volatility), and trips made in other three counties were less volatile.

Differences in the travel patterns and driving performances can be easily drawn by the simple comparisons shown in Table 7. The basic information about these counties can be used to explain why these differences exist. For example, perhaps because of the hilly terrain and strong grades in the county of San Francisco, drivers there had a higher level of driving volatilities than those in other four counties where the terrain is flat or rolling. However, the relationships between driving contexts and driving performance quantified by various measurements are very complex. Simple comparisons can only reveal a simple picture about the driving contexts and driving performance. Driver behaviors are potentially influenced by multiple factors synergistically. Thus, this study establishes a sophisticated hierarchical model to untangle those complex relationships between driving performance and driving contexts. As introduced, driving volatility is a measure being able to capture the speed, acceleration and vehicular jerk simultaneously. Driving volatility is used as a key measure of the driving practices for the further exploration.

TABLE 7 Comparisons of Selected Counties

Data sources	Attributes	Los Angeles	San Francisco	Sacramento	San Diego	Fulton (Atlanta)
Sample size	Number of drivers in dataset	528	39	146	184	258
	Total trips in dataset	12329	989	3074	4355	8226
Naturalistic driving data	*Mean daily trips*	*4.47*	*4.50*	*4.27**	*4.39*	*5.95***
	Total VMT (mile)	97242.12	9273.40	25125.90	36034.22	48919.07
	Mean daily VMT (mile)	*35.28*	*42.15***	*34.95**	*36.36**	*35.40*
	Mean trip length (mile)	*7.89*	*9.38***	*8.17*	*8.27*	*6.08***
	Total duration (min)	170365.30	15963.42	41462.88	57666.77	96491.60
	Mean daily duration (min)	*61.82*	*72.56***	*57.67***	*58.19***	*69.82***
	Mean trip duration (min)	*13.82*	*16.14***	*13.49*	*13.24***	*12.00***
	% of short trips (<3 miles)	45.23%	50.25%	40.11%	41.38%	50.52%
	% of long trips (> 25 miles)	6.70%	12.44%	5.69%	5.21%	3.11%
	Mean % time on idling per trip	*11.14%*	*9.68%***	*10.15%***	*9.73%***	*9.80%***
	Mean % time on ext. stable driving per trip[#]	*4.59%*	*6.69%***	*5.01%***	*5.31%***	*4.63%*
	Mean % time on acceleration per trip	*44.25%*	*43.22%***	*44.46%**	*44.57%***	*45.11%***
	Mean % time on declaration per trip	*39.04%*	*39.62%***	*39.49%***	*39.54%***	*39.53%***
	Trip mean speed (mph)	*26.884*	*29.181***	*27.859***	*28.595***	*28.358***
	Mean maximum speed (mph)	*49.727*	*51.078***	*50.742***	*51.832***	*49.924*
	Trip mean acceleration (ft/s^2)	*1.028*	*0.929***	*0.973***	*0.988***	*0.977***
	Mean maximum acceleration (ft/s^2)	*4.064*	*4.138***	*4.005***	*4.111***	*4.016***
	Trip mean deceleration (ft/s^2)	*-1.112*	*-0.977***	*-1.053***	*-1.069***	*-1.068***
	Mean maximum deceleration (ft/s^2)	*-4.722*	*-4.642***	*-4.652***	*-4.825***	*-4.713*
	Trip mean positive vehicular Jerk (ft/s^3)	*0.377*	*0.337***	*0.358***	*0.364***	*0.321***
	Mean maximum positive vehicular Jerk	*3.171*	*3.028***	*3.097***	*3.188*	*1.806***
	Trip mean negative vehicular Jerk (ft/s^3)	*-0.293*	*-0.268***	*-0.278***	*-0.286***	*-0.288***
	Mean maximum negative vehicular Jerk	*-1.403*	*-1.385*	*-1.362***	*-1.419***	*-1.382***
	Mean driving volatility (%)	*15.79*	*18.10***	*14.36***	*14.85***	*15.08***
	Maximum driving volatility (%)	64.68	60.20	55.25	55.22	56.25
Contextual data	Population	10017068	837442	1462131	3211252	984293
	Population density (per square mile)	2419.60	17179.10	1470.80	735.80	1748.00
	Percent of persons 65 years and over	11.90	14.20	12.40	12.30	10.10
	Percent of females	50.70	49.10	51.10	49.70	51.30
	Percent of persons with college degrees	29.50	52.00	27.90	34.40	48.40
	Mean commute time for work (minutes)	29.10	29.90	25.70	24.20	26.90
	Median household income (dollars)	27900.00	47278.00	26856.00	30683.00	37238.00
	Percent of persons in poverty	17.10	13.20	16.50	13.90	16.80
	Land size (square miles)	4057.88	46.87	964.64	4206.63	526.64
Geographic information	Road pattern (Gridiron/Non-Gridiron)	Gridiron	Non-G.	Gridiron	Gridiron	Non-
	Road network density (m. per square	5.35	20.53	5.35	2.67	7.96
	General terrain (Flat/Rolling/Hilly)	Flat	Hilly	Flat	Flat	Rolling
	Coast area (Yes/No)	Yes	Yes	No	Yes	No
Exposure data	Maintained highway miles, Rural	2074.09	0.72	871.21	4048.89	198.85
	Maintained highway miles, Urban	19620.00	961.71	4285.30	7164.90	3990.98
	Maintained highway miles, Total	21694.09	962.43	5156.51	11213.79	4189.83
	Daily VMT, Rural (1000 miles) ^	7709.93	17.56	3305.59	8051.40	394.00
	Daily VMT, Urban (1000 miles)	206772.51	9082.22	29631.74	67600.61	31443.00
	Daily VMT, Total (1000 miles)	214482.44	9099.78	32937.33	75652.01	31837.00
Safety facts	Annual average fatal crashes (2008-2012)	643.55	36.91	117.91	242.91	92.73

Notes: 1. [#]: Extended stable driving was defined by speed is above 30 mph and acceleration less than 0.088 (ft/s2). Thresholds were calibrated using test driving data; 2. Variables in *Italics* show results of t-tests, for comparisons with county of Los Angeles; 3. ** = t-test significant at a 95%confidence level; * = t-test significant at a 90% confidence level. 4. ^: VMT = Vehicle Miles Traveled, calculated by multiplying the AADT × the Section Length.

Descriptive Statistics

The raw dataset includes 117,022 trips made by 4,560 drivers residing in 78 counties across two states. Observations with missing information (e.g., no driver age or gender) were removed from the final data set. The final dataset contains 90, 511 trips made by 3,842 drivers from 78 counties. Counties were partitioned into six regions: Los Angeles (LA), San Francisco (SF), Sacramento (SAC), Atlanta (ATL), California central valley (CCV), and other areas in California. These regions were delineated according to 2013 Census Metropolitan Statistical Areas (*43*). Table 8 presents the distributions of observations at each hierarchy. Specifically, for level-1, the trip level, all observations are independent to each other and there are no clusters/groups. For level-2, the driver level, there are 3,842 groups (or drivers); on average each driver made 23.6 trips (min - 1, max - 154) during their travel survey periods. For level-3, the county level, one average, there are 953 trips (or observations) made within each county (min - 3, max - 9,615). For level-4, the regional level, the distributions of observations is shown in Table 9. 22,706 trips are from LA and were made by 1,027 drivers; 10,708 trips were made by 499 drivers in SF, 5,103 trips were made by 255 drivers in SAC, 5,658 trips were made by 245 drivers from CCV, 40,322 trips were made by 1,493 drivers in ATL, and 6,044 trips were made by 323 drivers from other areas in California.

TABLE 8 Observation Distributions at Each Level

Level	No. of Groups	Trips per Group		
		Minimum	Average	Maximum
Level 1	90,551 trips	1	-	1
Level 2	3,842 drivers	1	23.6	154
Level 3	78 counties	3	953	9,615
Level 4	6 regions	5103	15091.8	40332

TABLE 9 Distributions of Observations in Each Region (Level 4)

Region (not county)	Trips in Region	Percentage	Drivers in Region	Percentage
LA	22,706	25.08%	1,027	26.73%
SF	10,708	11.83%	499	12.99%
SAC	5,103	5.64%	255	6.64%
CCV	5,658	6.25%	245	6.38%
Other CA	6,044	6.67%	323	8.41%
ATL	40,332	44.54%	1,493	38.86%
Total	90,511	100.00%	3,842	100.00%

Table 10 presents the descriptive statistics of key variables. The number of observations at each level is different, which is unlike ordinary models that all variables have the same number of counts. The numbers shown in the table seem reasonable and were error checked. The driving volatility is the dependent variable and it was measured at the trip-level, which has 90,511 observations.

The average trip distance was 8.53 miles (min – 0.1, max - 431.52). 46.29% trips were made during rush hours, 23.56% were made during weekends and 17.76% were commute trips (from home to work or school). At Level 2, there are 3,842 observations (or drivers). 48.23% of them were males. Driver ages ranged from 15 to 91 years old. The mean vehicle age was 7.38 years old, ranging from 0 (i.e., new car) to 52 years old. In the dataset, 43.02% vehicles were regular auto-sedans, 22.46% were SUVs and 13.25% were Pickups. These vehicles consumed a variety of types of fuels—the majority is gasoline vehicles. At Level 3, descriptive statistics of 78 counties are summarized by contextual information, geographic information and exposure data. Among these counties, 49 generally have non-gridiron road networks, 46 are collectively flat areas and 16 counties are next to the coast. The mean road network density is 2.926 miles per square miles (min – 0.199, max – 20.534). Mean daily vehicle miles traveled is 13410.51 miles (min – 167, max – 214482).

TABLE 10 Descriptive Statistics of Key Variables

Variables			N	Frequency	Mean /Percent	Std. Dev.	Min	Max
Dependent	Driving Volatility (%)		90,551		13.934	7.314	0	67.188
Level-1	Trip distance (Mile)		90,551		8.529	14.319	0.1	431.52
	*Rush hour [Yes]		90,551	41,920	46.29%	0.499	0	1
	Weekend [Yes]		90,551	21,332	23.56%	0.424	0	1
	Commute trip [Yes]		90,551	16,083	17.76%	0.382	0	1
Level-2	Gender [Male]		3,842	1,853	48.23%	0.5	0	1
	Driver age (years)		3,842		48.257	13.406	15	91
	Vehicle age (years)		3,842		7.384	5.024	0	52
	Body Type	Auto-sedan	3,842	1,653	43.02%	0.495	0	1
		Two seated	3,842	197	5.13%	0.221	0	1
		Van	3,842	265	6.90%	0.253	0	1
		Hatchback	3,842	3	0.08%	0.028	0	1
		SUV	3,842	863	22.46%	0.417	0	1
		Station wagon	3,842	124	3.23%	0.177	0	1
		Pickup	3,842	509	13.25%	0.339	0	1
		Convertible	3,842	33	0.86%	0.092	0	1
		Unknown body type	3,842	195	5.08%	0.22	0	1
	Fuel Type	Hybrid electric vehicles	3,842	326	8.49%	0.279	0	1
		Gasoline vehicles	3,842	3,242	84.38%	0.363	0	1
		Diesel vehicles	3,842	115	2.99%	0.17	0	1
		Plug-in hybrid electric vehicle	3,842	19	0.49%	0.07	0	1
		CNG (compressed natural gas)	3,842	23	0.60%	0.077	0	1
		BEV (battery electric) vehicle	3,842	86	2.24%	0.148	0	1
		Flex fuel vehicle	3,842	9	0.23%	0.048	0	1
		Unknown fuel type	3,842	22	0.57%	0.075	0	1
Level-3	Contextual data	Percent of persons 65 years and over	78		14.36	4.461	7.8	25.1
		Percent of females	78		49.924	2.13	36.6	52.6
		Percent of college degrees	78		25.632	10.492	12.4	54.6
		Mean commute time (minutes)	78		26.044	5.142	13.9	37.4
		Median household income (dollars)	78		26917.79	7203.295	16667	55695
		Percent of persons in poverty	78		15.297	4.68	6.7	24.8
	Geographic information	Road pattern [Non-Gridiron]	78	49	62.82%	0.486	0	1
		Road density (m. per square mile)	78		2.926	3.561	0.199	20.534
		General terrain [Flat]	78	46	58.97%	0.495	0	1
		General terrain [Rolling]	78	19	24.36%	0.432	0	1
		General terrain [Hilly]	78	13	16.67%	0.375	0	1
		Coast area [Yes]	78	16	20.51%	0.406	0	1
	Exposure data	Maintained highway miles	78		2737.615	3129.114	270	21694
		Percent of urban maintained miles	78		44.265	32.897	0	100
		Daily VMT (1000 miles)	78		13410.51	28021.57	167	214482
		Percent of urban daily miles traveled	78		56.091	33.178	0	100
Level-4[#]	Region Indicator	LA	6	-	-	-	0	1
		SF	6	-	-	-	0	1
		SAC	6	-	-	-	0	1
		CCV	6	-	-	-	0	1
		Other CA areas	6	-	-	-	0	1
		ATL	6	-	-	-	0	1

Note:

*: Rush hours are AM (6:30 am-10:00 am) or PM (3:30 pm-7:00 pm);

[#]: Level-4 predictors are regional indicators that are indicator variables (0 or 1).

Model Selection

Table 4 presents all plausible variables (even more variables, e.g., population and density) that can be examined in the hierarchical modeling. However, not all variables are ensured to have a significant correlation with the response variable. Thus, the model selection is performed before the giving the final model. Standard information criteria, i.e., Akaike information criterion (AIC) and Schwarz's Bayesian information criterion (BIC), are used to compare models with different sets of variables. Given the same data, a smaller value of AIC and BIC indicates a better goodness-of-fit (*47-49*).

For any statistic models, the AIC and BIC can be calculated by

$$AIC = -2\mathrm{Ln}L + 2k \qquad \textit{Equation} (18)$$

$$BIC = -2\mathrm{Ln}L + k\mathrm{Ln}N \qquad \textit{Equation} (19)$$

Where,

$\mathrm{Ln}L$ = maximum log-likelihood of a model;

k = the number of variables in a model;

N = the sample size.

In the Equation 19, the second part ($k\mathrm{Ln}N$) implies that BIC would favor a simpler model (i.e., smaller number of variables included in the model) if the number of observations (k) is large. Thus, given such a large dataset (N=90,511 at trip-level), this study takes AIC as the major information criterion when AIC and BIG disagrees on which model is the best, in terms of the goodness-of-fit. Considering the massive computation of multi-level model with a large number of observation as well as the fact that most variables show significant correlations with driving volatility, the backward elimination method is applied for the variable selection (*50*). Table 11 shows the model selection results.

TABLE 11 Model Selection

Y = Driving Volatility			Model #1	Model #2	Model #3	Model #4
Level-1	Trip Distance (Mile)		√	√	√	√
	Rush Hour [Yes]		√	√	√	√
	Weekend [Yes]		√	√	√	√
	Commute Trip [Yes]		√	√	√	√
Level-2	Gender [Male]		√	×	×	×
	Driver Age (years)		√	√	√	√
	Vehicle Age (years)		√	√	√	√
	Body Type	Auto-Sedan	Base	Base	Base	Base
		Two Seated	√	√	√	√
		Van	√	√	√	√
		Hatchback	√	√	√	√
		SUV	√	√	√	√
		Station Wagon	√	√	√	√
		Pickup	√	√	√	√
		Convertible	√	√	√	√
		Unknown body type	√	√	√	√
	Fuel Type	Hybrid electric Vehicles	√	√	√	√
		Gasoline vehicles	Base	Base	Base	Base
		Diesel vehicles	√	√	√	√
		Plug-in hybrid electric vehicle	√	√	√	√
		CNG (Compressed Natural Gas)	√	√	√	√
		BEV (Battery Electric) vehicle	√	√	√	√
		Flex fuel vehicle	√	√	√	√
		Unknown fuel type	√	√	√	√
Level-3	Contextual data	Percent of persons 65 years and over	√	√	√	×
		Percent of females	√	√	√	√
		Percent of persons with college	√	√	×	×
		Mean commute time for work	√	√	×	×
		Median household income (dollars)	√	√	√	√
		Percent of persons in poverty	√	√	×	×
	Geographic information	Road pattern [Non-Gridiron]	√	√	√	√
		Road network density (miles per	√	×	×	×
		General terrain [Flat]	Base	Base	×	×
		General terrain [Rolling]	√	√	×	×
		General terrain [Hilly]	√	√	×	×
		Coast area [Yes]	√	√	√	×
	Exposure Data	Maintained highway miles, Total	√	×	×	×
		Percent of urban maintained highway	√	√	√	√
		Percent of urban daily vehicle miles	√	×	×	×
Level-4	Region Indicator	LA	√	√	√	√
		SF	√	√	√	√
		SAC	√	√	√	√
		CCV	√	√	√	√
		Other CA areas	Base	Base	Base	Base
		ATL	√	√	√	√
Log Likelihood (Model)			-287186.1	-287186.4	-287188.1	-287190.0
Degree of Freedom			47	43	38	36
AIC			574466.3	574458.8	574452.2	574452.0
BIC			574908.7	574863.6	574809.9	574790.9

Model 1 is the full model with all plausible variables. Most variables are significantly correlated with the response variable – driving volatility. Variables that are not significant at 50% level (i.e., p-value > 0.5) are removed from Model 1. These variables include driver gender, road network density, maintained highway miles and percent of urban daily vehicle miles traveled. In general, a three-point reduce in AIC indicates a significant improvement of the model's goodness-of-fit (*51*). The significantly reduced AIC and BIC indicate Model 2 has a better goodness-of-fit than Model 1. From Model 2, variables that are not significant at 70% level (i.e., p-value > 0.3) are removed. These variables include percent of persons with college degrees, percent of persons in poverty, and terrain. Model #3 has significantly smaller AIC and BIC estimates than Model 2, indicating a significant improvement of goodness-of-fit. From Model 3, variables that are not significant at 90% level (i.e., p-value > 0.1) are removed to get Model 4. However, Model 4 does not show a significantly reduced AIC estimate, which implies that variables eliminated from Model 3 should be kept for a better goodness-of-fit. Thus, the final model used for modeling the correlates of driving volatility with associate factors embedded in a hierarchy.

Modeling Results

Tables 12 and 13 present the outputs of the full model and the final mode after model selection: fixed effects, random effects and summary statistics. The model summary statistics seem reasonable. Note that, the likelihood ratio test of hierarchical modeling versus OLS modeling shows that the multi-level hierarchical model is significantly better than an OLS model, in terms of explaining the variances of driving volatility embedded in hierarchies. The signs of coefficients of variables (i.e., fixed effects) are expected. The random effects in Table 13 show that the total variances of driving volatility at each level, the unexplained variance and the percent of explained variable by variables. Variances of driving volatility between regions were 100% explained. Variances of driving volatility across counties are also explained very well. However, the Level 2 and Level 1 still have a sizable unexplained variables, which means more variables that describe these two levels are needed.

TABLE 12 Modeling Results (Fixed Effects)

Model			Full Model (Model #1)		Final Model (Model #3)	
Y = Driving Volatility			*β*	P-value	*β*	P-value
Constant			5.373	0.253	5.238	0.248
Level-1	Trip Distance (Mile)		-0.042**	0.000	-0.042**	0.000
	Rush Hour [Yes]		0.465**	0.000	0.465**	0.000
	Weekend [Yes]		-0.585**	0.000	-0.586**	0.000
	Commute Trip [Yes]		0.656**	0.000	0.656**	0.000
Level-2	Gender [Male]		-0.017	0.915		
	Driver Age (years)		-0.058**	0.000	-0.058**	0.000
	Vehicle Age (years)		-0.125**	0.000	-0.126**	0.000
	Body type	Auto-Sedan	Base			
		Two Seated	1.809**	0.000	1.814**	0.000
		Van	-2.394**	0.000	-2.405**	0.000
		Hatchback	-3.134	0.258	-2.708	0.323
		SUV	-0.885**	0.000	-0.886**	0.000
		Station Wagon	-1.002**	0.022	-1.009**	0.021
		Pickup	-2.186**	0.000	-2.195**	0.000
		Convertible	2.441**	0.003	2.450**	0.003
		Unknown body type	-0.804**	0.034	-0.808**	0.033
	Fuel type	Hybrid electric Vehicles	-1.490**	0.000	-1.479**	0.000
		Gasoline vehicles	Base			
		Diesel vehicles	-1.110**	0.013	-1.100**	0.014
		Plug-in hybrid electric vehicle	-0.520	0.634	-0.535	0.623
		CNG (Compressed Natural Gas)	-0.726	0.453	-0.715	0.459
		BEV (Battery Electric) vehicle	-2.572**	0.000	-2.579**	0.000
		Flex fuel vehicle	0.928	0.537	0.911	0.545
		Unknown fuel type	-0.777	0.438	-0.803	0.422
Level-3	Contextual data	Percent of persons 65 years and over	-0.099	0.138	-0.084	0.175
		Percent of females	0.127	0.223	0.178*	0.054
		Percent of persons with college	-0.039	0.299		
		Mean commute time for work	0.042	0.340		
		Median household income (dollars)	^0.000**	0.016	^0.000**	0.001
		Percent of persons in poverty	0.036	0.429		
	Geographic information	Road pattern [Non-Gridiron]	-0.539*	0.083	-0.561*	0.052
		Road network density (miles per	0.014	0.751		
		General terrain [Flat]	Base			
		General terrain [Rolling]	-0.261	0.402		
		General terrain [Hilly]	-0.175	0.643		
		Coast area [Yes]	0.363	0.290	0.364	0.224
	Exposure data	Maintained highway miles, Total	0.000	0.917		
		Percent of urban maintained highway	2.084*	0.066	2.859**	0.000
		Percent of urban daily vehicle miles	0.945	0.518		
Level-4	Region Indicator	LA	0.942	0.108	1.370**	0.003
		SF	-0.492	0.411	0.001	0.998
		SAC	0.458	0.390	0.669	0.179
		CCV	0.307	0.631	0.671	0.248
		Other CA areas	Base			
		ATL	0.019	0.979	0.159	0.774

Notes:

** = significant at a 95%confidence level; * = significant at a 90% confidence level, ^: the coefficient of household income is positive.

TABLE 13 Modeling Results for Random Effects and Summary Statistics

Model	Full Model (Model #1)			Final Model (Model #3)		
Random effects: Variance across groups						
Hierarchies	Total variance	Unexplained variance	Percent of explained	Total variance	Unexplained variance	Percent of explained
Level 4: Region level	1.236	0.000	100.00%	1.236	0.000	100.00%
Level 3: County level	2.138	0.103	95.20%	2.138	0.101	95.27%
Level 2: Driver level	21.361	18.947	11.30%	21.361	18.972	11.18%
Level 1: Trip level	30.308	29.825	1.59%	30.308	29.825	1.59%
Summary Statistics						
Log Likelihood (Model)	-287186.1			-287188.1		
Degree of Freedom	47			38		
AIC	574466.3			574452.2		
BIC	574908.7			574809.9		
Wald Chi-square	2176.020			2167.82		
Prob. > Chi-square	0.000			0.000		
Likelihood Ratio Test: Hierarchical vs. OLS	0.000			0.000		

All Level 1 variables for trip features have a statistically significant correlation (at 95% level) with the driving volatility, however, the correlation varies across drivers. Driver age, vehicle age, vehicle body type and fuel type at Level 2 also have a statistically significant coefficient, which may vary across counties. The variable at Level 3, household income, is significantly related to the driving volatility at county level. Regional dummy variables at Level 4 show the differences of driving volatility across regions. Results show drivers in Los Angeles have significantly greater driving volatilities than drivers in other California areas.

Discussion of Key Variables

Vehicle Features

As expected, vehicle features including vehicle age, body types and fuel types are significantly correlated to the trip-level driving volatilities. Older vehicles are associated with smaller driving volatilities. One unit increase in vehicle age corresponds to 0.126 units decrease in the driving volatility. Compared with auto-sedan, two-seated vehicles and convertibles are associated with higher level of driving volatilities, by a 1.814 and 2.45 units increase respectively. Note that one unit increase in trip-level driving volatility indicates one point increase in the percent of extreme driving seconds (i.e., large vehicular jerk values)

over the duration of one trip. Vehicles in other body types are associated with smaller driving volatilities. The modeling also reveals that vehicles consuming different type of fuels or energies perform differently in terms of the driving volatility. Compared with the gasoline vehicles, hybrid vehicles are associated with 1.479 units lower driving volatilities, battery electric vehicles are with 2.579 units lower driving volatilities and diesel vehicles are seen to be 1.1 units lower driving volatilities.

Driver Socio-Demographics

Senior drivers seem to be less volatile in driving than young drivers. Specifically, one year increase in driver age is related to 0.058 units decrease in driving volatilities. The modeling did not reveal a significant different driving volatility between male and female drivers.

Trip factors

All trip factors included in the model show a significant correlation with the driving volatility. Long trips are seen to be less volatile, perhaps because long trips are often made on freeways or interstates. Compared with trips made in non-rush hours, rush hour trips are associated with 0.456 units increase in driving volatilities. Weekend travel trips are less volatile than weekday trips. Trips made for work or school (commute trips) are more volatile than trips made for other purposes.

Geographic information

The geographic features of the road networks in a county was hypothesized to be associated with the driving volatility. However, as these are county-level geographic features, this modeling did not reveal strong correlations with the driving volatility, except the road pattern. Non-gridiron road patterns are likely associated with smaller driving volatilities. The insignificant estimates imply that there are not strong direct links between the county-level geographic features of road networks and the driving volatility, perhaps because these county-level features can hardly represent the environment of a specific road segment a driver is instantaneously perceiving during driving. Thus, understanding the influence of road characteristics on driving volatility needs detailed road information.

Contextual Information

The contextual information in this study may represent some aspects of driving contexts in a county. The modeling results show that counties with higher household incomes are associated with higher driving volatilities. The percent of females in a county is negatively correlated with the driving volatility, which is a marginally significant estimate. A county with more seniors is associated with smaller driving volatilities, as expected. However, the association is not statistically significant.

LIMITATIONS

The data used in this study are from multiple sources. Driving volatility is generated from large-scale trajectory data. These data were collected using in-vehicle GPS and OBD devices during two regional travel surveys. The accuracy of the trajectory data should be considered carefully.

Given the privacy issues for the data released online for public research, critical information, such as locations, are not available from the trajectory database. Thus, it is difficult for this study to link the driving behaviors with the instantaneous driving contexts.

CONCLUSIONS

This study demonstrates integration and use of large-scale data. It extends the understanding of naturalistic driving performance measured by driving volatility (*2*, *42*) and answers an important research question about whether the driving volatility varies across spatial contexts. Large-scale behavioral and vehicle trajectory data are coupled with data from various sources including FARS, historic traffic counts, google earth and census. Statistical analysis extracts useful information from hierarchically structured data, exploring links between driving volatility and correlates.

County-level factors were investigated in terms of their associations with the driving volatility. Counties with higher percentages of urban roadway mileage are associated with a higher level of driving volatilities for drivers in this county, indicating that driving in urban areas are more volatile driving. However, other county-level factors, including contextual and geographical information, were not statistically significantly associated with driving

volatility. In addition, vehicle features, driver socio-demographics and trip attributes are found to be significantly correlated, given their more direct relationship with the trip-level driving volatility.

This study is useful for both practitioners and researchers who are aware of and accessible to various transportation data sources; it offers insights in integrating data from multiple sources and exploring links between driving behaviors and casual factors. With the increasing popularity of GPS and other information technology for collecting travel information, transportation-related "Big Data" will increase. Extracting useful information from the data will increase in its importance. This study has shown that hierarchical modeling provides an opportunity to extract useful information from big data with complex hierarchical structure. This study extends the understanding of driving performance and it will benefit the future work of mining large-scale transportation data. In terms of future research, similar models can be applied to study correlates of other regional transportation performance measures, such as vehicle miles or hours traveled.

References

1. Lohr, S., *The age of big data*, in *New York Times*. 2012: New York, NY.
2. 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*, Vol. No. 2015: pp.DOI: 10.1016/j.trc.2014.12.014.
3. Young, T., J. Blustein, L. Finn, and M. Palta, Sleepiness, driving and accidents: sleep-disordered breathing and motor vehicle accidents in a population-based sample of employed adults. *Sleep*, Vol. 20, No. 8, 1997: pp. 608-613.
4. Ramaekers, J.G., H. Robbe, and J. O'Hanlon, Marijuana, alcohol and actual driving performance. *Human Psychopharmacology Clinical and Experimental*, Vol. 15, No. 7, 2000: pp. 551-558.
5. Wang, Q., H. Huo, K. He, Z. Yao, and Q. Zhang, Characterization of vehicle driving patterns and development of driving cycles in Chinese cities. *Transportation research part D: transport and environment*, Vol. 13, No. 5, 2008: pp. 289-297.

6. Brundell-Freij, K. and E. Ericsson, Influence of street characteristics, driver category and car performance on urban driving patterns. *Transportation Research Part D: Transport and Environment*, Vol. 10, No. 3, 2005: pp. 213-229.
7. Aarts, L. and I. Van Schagen, Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, Vol. 38, No. 2, 2006: pp. 215-224.
8. Ericsson, E., Variability in urban driving patterns. *Transportation Research Part D: Transport and Environment*, Vol. 5, No. 5, 2000: pp. 337-354.
9. Bang, K.L. and A. Carlsson, Development of speed-flow relationships for Indonesian rural roads using empirical data and simulation. *Transportation research record*, Vol. No. 1484, 1995: pp. 24-32.
10. Akamatsu, M., N. Imachou, Y. Sasaki, H. Ushiro-Oka, T. Hamanaka, and M. Onuki. Simulator study on driver's behavior while driving through a tunnel in a rolling area. in *Proceedings of the Driving Simulation Conference. Michigan, USA*. 2003.
11. Kilpeläinen, M. and H. Summala, Effects of weather and weather forecasts on driver behaviour. *Transportation research part F: traffic psychology and behaviour*, Vol. 10, No. 4, 2007: pp. 288-299.
12. Maze, T.H., M. Agarwai, and G. Burchett, Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. *Transportation research record: Journal of the transportation research board*, Vol. 1948, No. 1, 2006: pp. 170-176.
13. Golob, T.F. and W.W. Recker, Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of Transportation Engineering*, Vol. 129, No. 4, 2003: pp. 342-353.
14. Keay, K. and I. Simmonds, The association of rainfall and other weather variables with road traffic volume in Melbourne, Australia. *Accident analysis & prevention*, Vol. 37, No. 1, 2005: pp. 109-124.
15. Savage, I., Does public education improve rail–highway crossing safety? *Accident Analysis & Prevention*, Vol. 38, No. 2, 2006: pp. 310-316.
16. Mayhew, D.R. and H.M. Simpson, The safety value of driver education an training. *Injury Prevention*, Vol. 8, No. suppl 2, 2002: pp. ii3-ii8.

17. Liu, J., B. Bartnik, S.H. Richards, and A.J. Khattak, *How are driver characteristics related to safety at railroad-crossings? The case of passive railroad grade crossings*, in *93rd Annual Meeting of the Transportation Research Board* 2014: Washington D.C.
18. Khattak, A. and M. Rocha, Are SUVs" Supremely Unsafe Vehicles"?: Analysis of Rollovers and Injuries with Sport Utility Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1840, No. 1, 2003: pp. 167-177.
19. Cestac, J., F. Paran, and P. Delhomme, Young drivers' sensation seeking, subjective norms, and perceived behavioral control and their roles in predicting speeding intention: How risk-taking motivations evolve with gender and driving experience. *Safety science*, Vol. 49, No. 3, 2011: pp. 424-432.
20. Hung, W.-T., K.-M. Tam, C.-P. Lee, L.-Y. Chan, and C.-S. Cheung, Comparison of driving characteristics in cities of Pearl River Delta, China. *Atmospheric Environment*, Vol. 39, No. 4, 2005: pp. 615-625.
21. Elander, J., R. West, and D. French, Behavioral correlates of individual differences in road-traffic crash risk: An examination of methods and findings. *Psychological Bulletin*, Vol. 113, No. 2, 1993: pp. 279-294.
22. Garrity, R.D. and J. Demick, Relations Among Personality Traits, Mood States, and Driving Behaviors. *Journal of Adult Development* Vol. 8, No. 2, 2001: pp. 109-118.
23. Ulleberg, P. and T. Rundmo, Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. *Safety Science*, Vol. 41, No. 5, 2003: pp. 427-443.DOI: 10.1016/s0925-7535(01)00077-7.
24. Oltedal, S. and T. Rundmo, The effects of personality and gender on risky driving behaviour and accident involvement. *Safety Science*, Vol. 44, No. 7, 2006: pp. 621-628.DOI: 10.1016/j.ssci.2005.12.003.
25. Liu, J., A. Khattak, and X. Wang, The role of alternative fuel vehicles: Using behavioral and sensor data to model hierarchies in travel. *Transportation Research Part C: Emerging Technologies*, Vol. 55, No. 2015: pp. 379-392.DOI: 10.1016/j.trc.2015.01.028.

26. Moeckli, J. and J. Lee, The making of driving cultures. *Improving Traffic Safety Culture in the United States*, Vol. 38, No. 2, 2007: pp. 185-192.
27. Zaidel, D.M., A modeling perspective on the culture of driving. *Accident Analysis & Prevention*, Vol. 24, No. 6, 1992: pp. 585-597.
28. Caltrans. *California Household Travel Survey*. 2013 [cited 2014 May 2nd]; Available from: http://www.dot.ca.gov/hq/tsip/otfa/tab/chts_travelsurvey.html.
29. ARC, *Regional Travel Survey - Final Report*. Atlanta Regional Commission, 2011.
30. TSDC, *Secure Transportation Data Project*. 2014, Transportation Secure Data Center, National Renewable Energy Laboratory
31. Liu, J., A. Khattak, and X. Wang, *Creating Indices for How People Drive in a Region: A Comparative Study of Driving Performance*, in *94th Annual Meeting of the Transportation Research Board* 2015: Washington D.C.
32. Caltrans. *Highway Performance Monitoring System (HPMS)*. 2015 [cited 2015 March 25]; Available from: http://www.dot.ca.gov/hq/tsip/hpms/.
33. GDOT. *TravelSmart*. 2015 [cited 2015 March 25]; Available from: http://www.dot.ga.gov/AboutGeorgia/Pages/TravelSmart.aspx.
34. Várhelyi, A., M. Hjälmdahl, C. Hydén, and M. Draskóczy, Effects of an active accelerator pedal on driver behaviour and traffic safety after long-term use in urban areas. *Accident Analysis & Prevention*, Vol. 36, No. 5, 2004: pp. 729-737.
35. 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.
36. Ericsson, E., Independent driving pattern factors and their influence on fuel-use and exhaust emission factors. *Transportation Research Part D: Transport and Environment*, Vol. 6, No. 5, 2001: pp. 325-345.
37. Kim, E. and E. Choi, *Estimates of Critical Values of Aggressive Acceleration from a Viewpoint of Fuel Consumption and Emission*, in *TRB 2013 Annual Meeting*. 2013: Washington D.C.

38. 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.
39. Drew, D.R., *Traffic flow theory and control*. 1968.
40. Langari, R. and W. Jong-Seob, Intelligent energy management agent for a parallel hybrid vehicle-part I: system architecture and design of the driving situation identification process. *Vehicular Technology, IEEE Transactions on*, Vol. 54, No. 3, 2005: pp. 925-934.DOI: 10.1109/tvt.2005.844685.
41. Murphey, Y., 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*. 2009. IEEE.
42. Liu, J., X. Wang, and A. Khattak, *Generating Real-Time Driving Volatility Information*, in *2014 World Congress on Intelligent Transport Systems*. 2014: Detroit, MI.
43. Census. *Metropolitan and Micropolitan Statistical Areas Main*. 2013 [cited 2014 May 3rd]; Available from: http://www.census.gov/population/metro/.
44. Caltrans, *2010 California Public Road Data*. California Department of Transportation, 2011.
45. Sorensen, P., Reducing Traffic Congestion and Improving Travel Options in Los Angeles. Vol. No. 2010: pp.
46. Schrank, D., B. Eisele, and T. Lomax, TTI's 2012 urban mobility report. *Proceedings of the 2012 annual urban mobility report. Texas A&M Transportation Institute, Texas, USA*, Vol. No. 2012: pp.
47. Akaike, H., A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, Vol. 19, No. 6, 1974: pp. 716-723.
48. 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.
49. Schwarz, G., Estimating the dimension of a model. *The annals of statistics*, Vol. 6, No. 2, 1978: pp. 461-464.

50. Sutter, J.M. and J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchemical journal*, Vol. 47, No. 1, 1993: pp. 60-66.

51. Fotheringham, A.S., C. Brunsdon, and M. Charlton, *Geographically Weighted Regression, the analysis of spatially varying relationships*, ed. U. University of Newcastle. 2002: John Wiley & Sons, Ltd.

DELIVERING IMPROVED ALERTS, WARNINGS, AND CONTROL ASSISTANCE USING BASIC SAFETY MESSAGES TRANSMITTED BETWEEN CONNECTED VEHICLES[4]

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 (BSM), which are sampled 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 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.

[4] Material based on: Liu J. & A. Khattak. Improved Warning and Assistance Information from Connected Vehicle Basic Safety Messages, Accepted for presentation to 2015 Intelligent Transportation Systems World Congress, Bordeaux, France, 2015. A revised version of this paper, titled "Delivering Improved Alerts, Warnings, and Control Assistance Using Basic Safety Messages Transmitted between Connected Vehicles" was submitted to 2016 Transportation Research Board for review.

Keywords: Connected vehicle, basic safety messages, extreme events, speed and acceleration

INTRODUCTION

The United States has one of the largest highway transportation systems in the world. According to the highway statistics from US Department of Transportation (US DOT), since 2010, the highway system in the US has exceeded 8.5 million lane miles (*1*) and vehicle miles traveled is around three trillion (*2*). Over 5.5 million police-reported traffic crashes occur annually. According to 2012 traffic safety facts (*2*), about 34,000 people were killed and 2.3 million people were injured. In recent years (2007-2011), the number of fatalities has declined, but the death toll is still too high. We still need critical improvements to make highway transportation systems safer. Connected vehicles can improve safety through exchange of critical information between vehicles and infrastructures.

Currently, there is no universally agreed-upon definition for connected vehicles. According to the Intelligent Transportation Systems Joint Program Office, connected vehicle technology is "the creation of a safe, interoperable wireless communications network that includes cars, buses, trucks, trains, traffic signals, cell phones, and other devices (*3*)." Like the Internet that connects computers, smart phones, servers and other terminals, connected vehicle communication networks connect vehicles, facilities, operation centers, and other utilities.

Connected vehicles are spawning new applications. Most previous safety applications such as air bags and seat belts help occupants survive crashes while connected vehicle applications are expected to help people avoid crashes altogether. Researchers have proposed applications of vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) to inform drivers of roadway hazards and situations that they cannot immediately perceive using driver assist systems. This should help drivers make decisions to avoid dangers (*4-8*). V2V application enables vehicles to transmit data to and from surrounding vehicles, and V2I allows vehicles to communicate with infrastructures such as traffic signals. According to a US DOT report, V2V and V2I applications potentially address about 80 percent of traffic crashes (*3*), possibly because over 90 percent of traffic crashes are attributed to driver errors, including recognition

errors, decision errors, and performance or nonperformance errors (*9*). V2V and V2I applications are expected to help drivers perceive dangers and hazards and make better decisions.

A connected vehicle safety pilot program, Safety Pilot Model Deployment (SPMD), underway in Ann Arbor, Michigan intends to demonstrate V2V and V2I technologies in a real-world environment (*10*). Approximately 3,000 vehicles are equipped with V2V communication devices, and 75 miles of roadway are instrumented with roadside equipment, mainly placed at signalized intersections. Data acquisition systems (DAS) are installed in vehicles participating in the program to facilitate V2V and V2I communications. Data transmitted through V2V and V2I are called Basic Safety Message (BSM), sampled at a frequency of 10 Hz. The core contents of BSM are data elements that describe a vehicle's position (latitude, longitude, and elevation) and motion (heading, speed, and acceleration) (*10*). BSM also contains data pertaining to the vehicle's component status (lights, brakes, wipers) and vehicle safety information (path history, events) (*10*). Given the availability of advanced communication and sensor technologies such as Global Positioning Systems (GPS), Radar, and Bluetooth, there is no doubt that BSMs can be successfully sent and received by vehicles and roadside equipment. SPMD is a demonstration of such communication technologies.

Using these technologies, what kind of critical information can be extracted and provided to drivers to alert them to present dangers or assist in vehicle control to help drivers make safe decisions? Driver-oriented information is supposed to be simple and informative, such as head-on collision warnings. This study proposes an original methodology, based on data analytics, to extract critical information from basic safety messages transmitted between connected vehicles and infrastructure.

LITERATURE REVIEW

The literature reflects substantial activities in connected vehicles (CVs), covering a wide range of topics from how CVs will be adopted and used, to their applications and implications for safety, energy, and the environment (*11-15*). A report by Hill et al. discusses CV infrastructure deployment approaches and strategies, including the time horizon, key

issues related to drivers and vehicles, operations, benefits for state and local agencies, and early CV adopters (*16*). Substantial work has been done to establish connected vehicle networks, such as Vehicular Ad-hoc Networks (*5, 17*). Both the public and private sectors are interested in applications and implications of CVs. Applications include intersection signals (*6, 18-22*), pavement assessments (*23*), traffic queue estimation (*24*), vehicle routing and travel time estimation (*25-28*), driving behavior monitoring and warnings (*29-33*), and fuel efficiency (*34*).

The key to the success of connected vehicles lies in how well connectivity of vehicles and infrastructure works in real life. Recent innovations that enable connectivity include applications of V2V and V2I. supported by wireless communication technologies such as Dedicated Short Range Communication (DSRC)(*35, 36*), Wi-Fi (*37, 38*), Bluetooth (*39, 40*), and cellular networks (*41, 42*).

Safety Pilot Model Deployment (SPMD) uses Basic Safety Messages (BSMs) to describe a vehicle's position, motion, its component status, and other relevant travel information (*10*). However, the 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. However, abnormal and extreme driver behaviors determine the safety of driving the short-term. Thus, it is critical to identify abnormal or extreme behaviors from BSMs, and warn drivers through the V2V, V2I, or other connected vehicle applications.

A number of studies have focused on investigating driving behaviors. Vehicle motion (speed and acceleration) has been regarded as the core information describing driving behaviors. Fast driving is normally characterized as an aggressive or reckless driving style, and speed limits are usually the threshold that determines a driver's performance (*43-46*). However, speed choice depends mainly on the conditions of speed limits (or road conditions) and the traffic. A driver is supposed to comply speed limits, but he or she is also affected by the traffic (*47, 48*). Researchers give several acceleration cut-off points as thresholds to identify abnormal, extreme, and aggressive driving behaviors. Kim and Choi report thresholds for aggressive and extremely aggressive accelerations in urban driving environments (*49*), while De Vlieger et al. did similar work for calm driving, normal driving,

and aggressive driving (*50*).

Driving occurs in various conditions, and driver behaviors may vary in different contexts. To account for the variation of driving behaviors under different conditions, Liu et al. and Wang et al. introduced a varying acceleration threshold to identify extreme driving behaviors (*51, 52*). Most previous studies investigating driver behaviors overlook the directions (longitudinal and lateral) of driving decisions. Driving decisions in two directions (longitudinal and lateral accelerations) are under-explored. This study proposes an innovative way to identify extreme driving behaviors embedded in BSMs that may provide warning messages to drivers through V2V and V2I applications.

While previous studies propose ideas for warnings or alerts to drivers using the CV applications (*29-33*), they have not fully assimilated the value of information transmitted between connected vehicles. For example, Noble et al. (*53*) used only naturalistic driving data collected through the Strategic Highway Research Program 2 (but not BSMs) for analysis, and Osman et al. (*33*) used driving simulator based data. This study fully mines the geo-referenced data transmitted between vehicles and infrastructure in a real-life CV deployment. Specifically, it extracts useful information about extreme events from new data sources made possible by communication between connected vehicles.

METHODOLOGY

Recently, NHTSA and SAE have come out with levels of automation that range from no automation to full automation (*54*). Figure 22 shows the how drivers can transition from controlling all aspects of the dynamic driving task to relinquishing control of these tasks. Based on the taxonomy, this study focuses on Level 0 in the SAE taxonomy. Within Level 0, the study transitions from driving without alerts, warnings, or intervention systems to using these for enhanced driving safety. Increasingly, vehicles are incorporating driver decision support systems, while drivers retain control of steering and braking controls, except in crash imminent situations. Alerts, warnings, and control assists can be divided into two broad categories:

1) Internal to the functioning and performance of the driver or vehicle. Examples of these include warnings about hard accelerations or braking, or frequent lane changes, sharp turns, or functioning of wipers, head- and tail-lights, turn signals, etc.

2) External to the vehicle. These are warnings that relate to proximity of other vehicles, objects, infrastructure, and the environment, and include forward collision warning or lane departure warning.

This paper considers both types of alerts, warnings, and control assists, which are still part of Level "0." Such warnings are based on BSM data, and can eventually lead to higher levels of automation.

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
***Human driver* monitors the driving environment**						
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	**Driver Assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	**Partial Automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/ deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
***Automated driving system* ("system") monitors the driving environment**						
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	**High Automation**	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

FIGURE 22 Six levels of driving automation. Source: SAE, J3016, 2014 (*54*).

Data Description - Basic Safety Message

The data used in this study are from BSMs sent and received by vehicles and roadside equipment participating the SPMD in Ann Arbor, Michigan (*10*). SPMD provides a rich database for research on connected vehicles. 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 uses datasets collected from participating vehicles equipped with Data Acquisition Systems (DAS). Datasets contain vehicles' instantaneous driving statuses (sampled at 10 Hz) of position (altitude, latitude and longitude), motion (speed and acceleration), the status of major components (accelerator, brakes, lights, cruise control, and wipers), and instantaneous driving contexts (surrounding objects, and distance to closest objects). Table 14 presents the detailed descriptions of key data variables. More variable descriptions are available in SPMD Data Handbook (*10*).

TABLE 14 Variable Descriptions from Safety Pilot Model Deployment, Ann Arbor, Michigan

Variable		Description
Position	Altitude	A GPS-based estimate of height above sea level (height above the reference ellipsoid that approximates mean sea level)
	Latitude	Current degree of latitude at which the vehicle is located
	Longitude	Current degree of longitude at which the vehicle is located
Motion	Speed (host vehicle)	Current vehicle speed, as determined from the vehicle's transmission
	Longitudinal Acceleration	Longitudinal acceleration measured by an Inertial Measurement Unit (IMU)
	Lateral Acceleration	Lateral acceleration measured by an IMU
Vehicle Maneuvering	Accelerator Pedal	Reflects the amount the accelerator pedal is displaced with respect to its neutral position
	Brake Pedal	Indicates whether the brake light is on or off
	Cruise Control	Indicates whether cruise control is active/engaged
	Turn Signal	Provides information regarding the state of the vehicle turn signals
Driving Context	Number of objects	Number of identified objects, as determined by the Mobileye sensor
	Distance to the closest object	Position of the closest object, relative to a reference point on the host vehicle, according to the Mobileye sensor
	Relative speed of the closest object	Longitudinal velocity of the closest object, relative to the host vehicle according to the Mobileye sensor

Source: SPMD Data Handbook (*10*).

One-day sample data were used in this study. Observations with errors (e.g., speeds > 200 mph and altitude > 30,000 ft) were removed from the sample. The final one-day sample contains 968,522 records of basic safety messages, from 155 trips made by 49 vehicles. The sum of trip durations is about 26.9 hours, and the average duration per trip is about 10.4 minutes. Most of the trips were made within the road networks of Ann Arbor, Michigan, and some long trips reached the neighboring towns of Dexter, Chelsea, and Livonia in Michigan. Table 15 shows the descriptive statistics of selected BSM variables in the final datasets. Based on the error-checked descriptive statistics and the distributions, the data seemed to be of reasonably good quality. Figure 23 presents the spatial distribution of sampled data. Distributions of variables seemed reasonable in terms of magnitude and spatial characteristics.

TABLE 15 Descriptive Statistics of Selected BSM Variables

Variable		Mean/ Percentage	Std. Dev.	Min	Max
Position	Altitude (ft)	724.603	81.076	496.388	1345.492
	Latitude (degree)	42.307	0.135	42.044	42.977
	Longitude (degree)	-83.745	0.291	-85.635	-83.280
Motion	Host Vehicle Speed (mph)	38.507	23.249	0.000	83.346
	Longitudinal Acceleration (ft/s^2)	-0.367	2.107	-21.818	22.420
	Lateral Acceleration (ft/s^2)	-0.090	2.246	-22.310	22.330
Vehicle Maneuvering	Accelerator Pedal (%)	13.299%	-	0.000	1.000
	Brake Pedal (engaged)	20.186%	-	0.000	1.000
	Cruise Control (engaged)	38.822%	-	0.000	1.000
	Turn Signal (None)	93.847%	-	0.000	1.000
	Turn Signal (Left)	3.927%	-	0.000	1.000
	Turn Signal (Right)	2.226%	-	0.000	1.000
Driving Context	Number of objects	1.618	1.469	0.000	9.000
	Distance to the closest object (ft)	143.301	133.912	0.000	839.690
	Relative Speed of the closest object (mph)	-3.893	21.829	-157.984	159.941

Note: Sample size = 968,522 records.

FIGURE 23 Spatial Distribution of Trajectory Data in the Final Datasets

Conceptual Framework

The main objective of this study is to use data analytics to extract critical information embedded in BSMs sent and received by vehicles and infrastructure. Figure 3 shows the proposed framework to explore the BSMs and compile the raw BSMs into messages that can be communicated to drivers. These messages can inform the host vehicle drivers (i.e., raw BSMs are from the same vehicle) to adjust their driving behaviors, and also to give warnings and control assistance to remote vehicle drivers (i.e., raw BSMs are from other vehicles) to avoid potential dangers. In real-world environments, real-time BSMs are compiled into real-time advisory or warning messages, directed to local drivers through V2V and V2I applications.

Understanding instantaneous driving volatility was one of the most challenging aspects of this study; this understanding can be accomplished by the BSM Compiler designed in Figure 24. Data sampled at a high frequency, 10 Hz, yielded deeper insights into instantaneous driving behaviors. This study used various data visualization tools to show the extent of instantaneous driving volatility, including distributions of longitudinal and lateral acceleration, speed-based distributions of instantaneous yaw rate, three-dimensional distributions of longitudinal acceleration-lateral acceleration-speed, and driving volatility on different road types. This paper provides data visualization details. Then, extreme driving events will be identified in accordance with special rules. One of the rules this study used is speed-based thresholds (*51*), since driving behaviors vary at different speeds, implying different driving contexts. These rules are discussed along with data visualization and

analysis. Further, these extreme events were linked to instantaneous vehicle control statuses and driving contexts to understand why they occur. Finally, advisory or warning messages and vehicle control assistance to drivers can be generated.

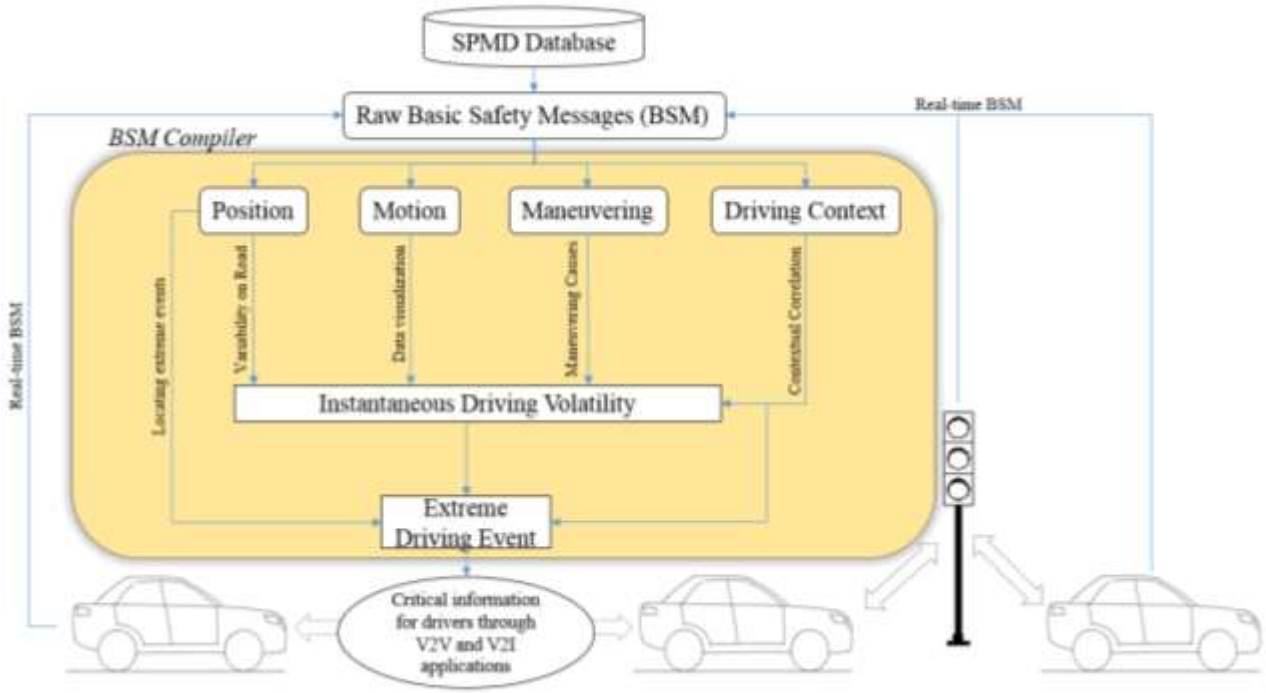


FIGURE 24 Conceptual Framework

EXTENT OF INSTANTANEOUS DRIVING VARIABILITY

Driving on Road

To observe how a vehicle moves on a road, 10-Hz motion data (speed, longitudinal, and lateral acceleration) from BSMs were visualized on maps according to position data (longitude and latitude). Figure 25 shows two sample trips made on different types of roads. Figure 25(i) presents one trip going through downtown Ann Arbor and Figure 25(ii) shows a trip that includes freeway driving. To illustrate the volatility of instantaneous driving decisions (i.e., variability of instantaneous accelerations), Figure 25 also shows the variance of longitudinal and lateral accelerations within one second, calculated by Equations (20) and (21) below.

$$VAR_i^{Longitudinal} = \frac{1}{n}\sum_{j=i-n+1}^{i}\left(a_j^{Longitudinal} - \frac{1}{n}\sum_{j=i-n+1}^{i} a_j^{Longitudinal}\right)^2 \quad \textit{Equation (20)}$$

$$VAR_i^{Lateral} = \frac{1}{n}\sum_{j=i-n+1}^{i}\left(a_j^{Lateral} - \frac{1}{n}\sum_{j=i-n+1}^{i} a_j^{Lateral}\right)^2 \quad \textit{Equation (21)}$$

Where,

VAR = Variance of accelerations

n = number of observations within one second, *n* = 10 for 10 Hz data;

i = time series (0.1 second), $i \geq n$;

$a^{Longitudinal}$ = Record of instantaneous longitudinal acceleration;

$a^{Lateral}$ = Record of instantaneous lateral acceleration.

As expected, driving on local roads is more volatile, in terms of variance of accelerations, than driving on a freeway because distractions such as pedestrians and roadside attractions are more frequent on local roads. For this reason, critical information extraction from BSMs should consider the difference of driving performance or behavior under different driving contexts. An easy way to distinguish the driving context would be travel speed. The travel speed on freeways would often be higher than on local roads, as indicated by speed limits. Figure 5 shows that higher speeds are associated with smaller variations of acceleration in two directions. Ahn et al. (*55*) report that higher accelerations are associated with lower speeds. Consequently, driving volatility might also be associated with speeds.

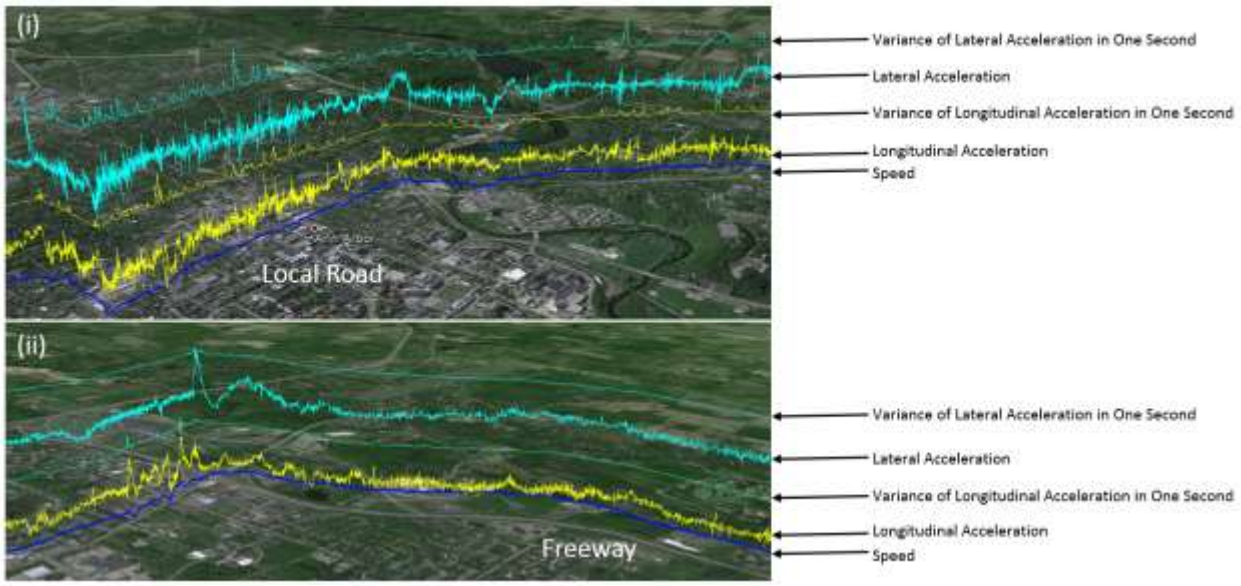


FIGURE 25 Instantaneous Driving Decisions Visualized on Road

Distributions of Acceleration

To clarify the relationship between speeds and acceleration, distributions of acceleration were visualized in two directions: longitudinal and lateral. Figure 26 presents the magnitude distribution of acceleration along speeds and the density of possible acceleration values along speeds. It shows that generally higher speeds (>50 mph) were associated with smaller acceleration magnitudes, which is partially consistent with Ahn et al. Vehicle engines have to do more work in order to maintain the same acceleration at higher speeds to overcome increasing air resistance. Therefore, the ability to accelerate or decelerate a vehicle decreases naturally at higher speeds (*52*). Ahn et al. (*55*) points out a linear relationship between acceleration and speeds, and this study reveals a nonlinear relationship between acceleration and speed in real-life driving situations. The varying distributions of instantaneous accelerations along speeds confirm the above findings that driving behavior varies in different driving situations, as reflected by driving speeds. In addition, Figure 26 also presents the distribution of longitudinal vs. lateral accelerations. The lozenge shaped distribution implies that longitudinal and lateral accelerations do not have large magnitudes simultaneously. In terms of their magnitude, longitudinal and lateral accelerations seem to be inversely correlated (correlation = -0.8343, p-value < 0.01).

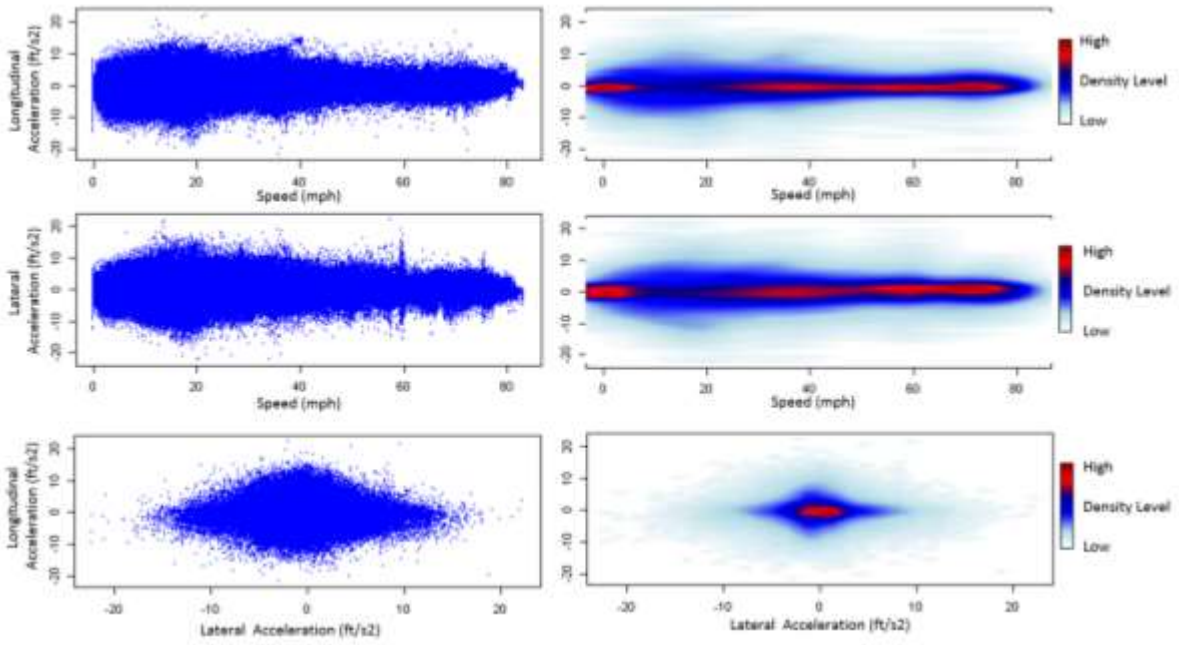


FIGURE 26 Distributions of Longitudinal and Lateral Acceleration

The distributions of longitudinal and lateral accelerations were visualized in three-dimensional space, according to their relative magnitude and direction. The resultant instantaneous acceleration of a vehicle is the sum of motion vectors of longitudinal and lateral acceleration, as shown in Equation (22).

$$\vec{A} = \vec{a}_{Longitudinal} + \vec{a}_{Lateral} \quad \text{Equation (22)}$$

The magnitude of resultant instantaneous acceleration $\vec{A}$ is

$$|\vec{A}| = \sqrt[2]{(\vec{a}_{Longitudinal}{}^{2} + \vec{a}_{Lateral}{}^{2})} \quad \text{Equation (23)}$$

The direction of resultant instantaneous acceleration $\vec{A}$ is

$$\Delta = 180 \frac{\tan^{-1}\left(\frac{\vec{a}_{Longitudinal}}{\vec{a}_{Lateral}}\right)}{\pi} \quad \text{Equation (24)}$$

Δ is the counter-clockwise angle between vehicle heading direction and the direction of resultant instantaneous acceleration.

As shown in Figure 26, varying distributions implied variations in driving behaviors at given speeds and motion in different directions. To show the magnitude of distribution at

various speeds and directions, this study used different colors to indicate the relative magnitudes of accelerations by speed × sector bin in 0.5 mph × 2°. An example bin is shown in Figure 27. All magnitudes of accelerations were compared within a bin.

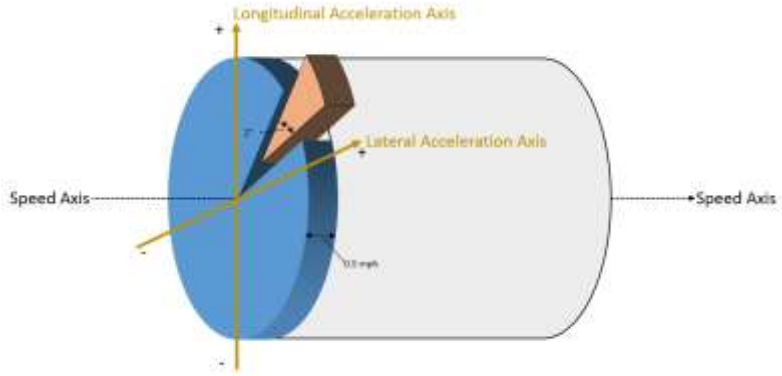


FIGURE 27 Speed × Sector Bin

Figure 28 shows the three-dimensional distribution of accelerations at different speeds and directions. The view of sectional drawings reveals the magnitude of instantaneous accelerations. Blue implies that the magnitudes of accelerations are close to zero compared to other accelerations within the same bin, and red indicates greater magnitudes. Same color indicates that magnitudes of acceleration within different bins were at the same percentiles, forming percentile bands. Magnitudes of longitudinal and lateral accelerations varied with different speeds. Magnitudes at lower speeds were relatively larger than at higher speeds, illustrated by the percentile bands, as shown in Figure 28(iii) and (iv). The cross orthogonal shape, shown by the blue and yellow area in Figure 28 (ii), implies that magnitudes of accelerations that were parallel or perpendicular to the heading directions were relatively greater than those diagonal to heading directions. This confirmed that instantaneous driving decisions varied in different directions, which is useful for identifying extreme driving events such as sudden lane change behaviors.

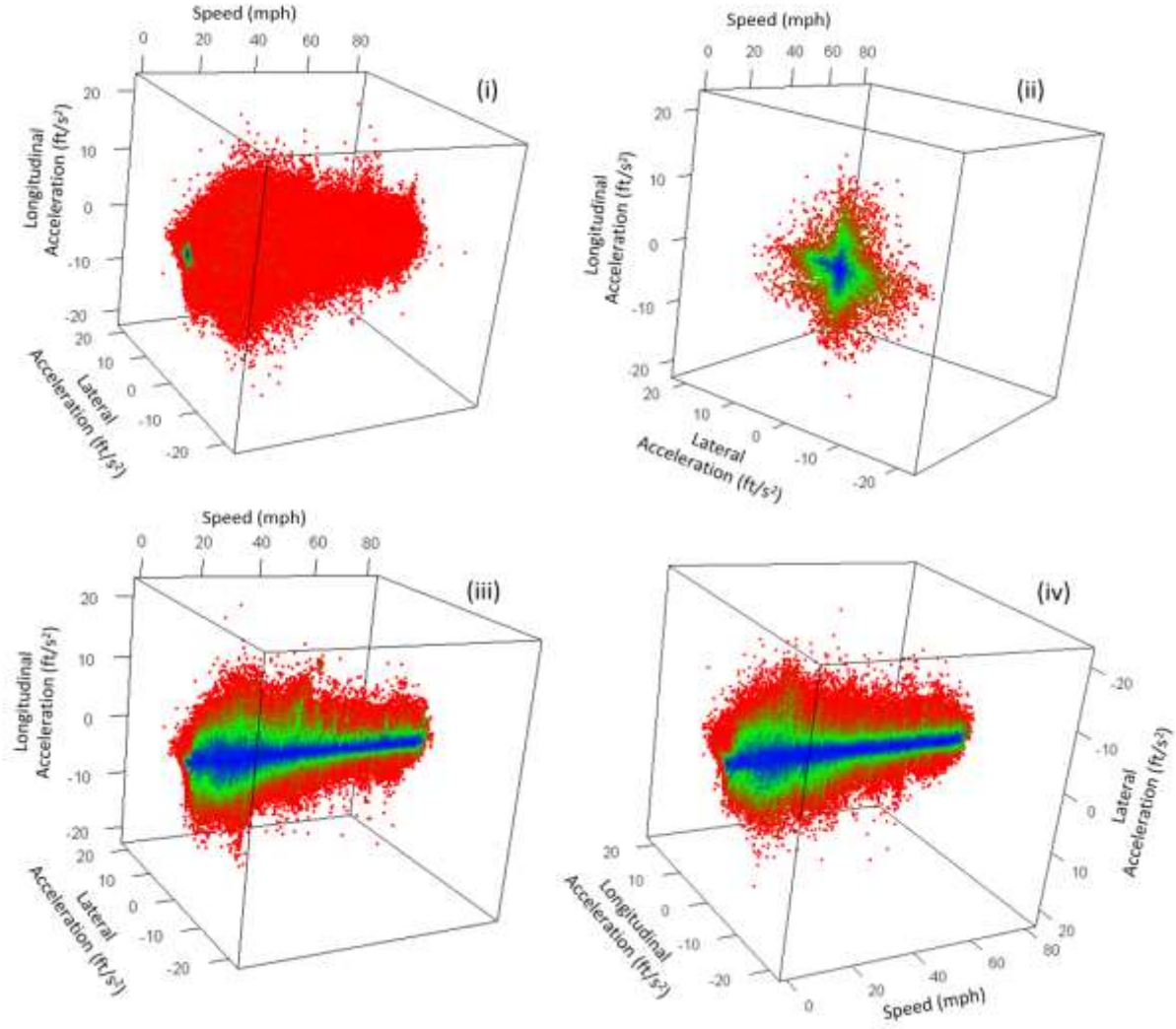


FIGURE 28 Distributions of Longitudinal and Lateral Acceleration in 3-D Space

IDENTIFICATION OF EXTREME EVENTS

Previous studies indicate that extreme driving events (e.g., hard braking or acceleration) are associated with aggressive driving behaviors. However, aggressive driving might be one reason for extreme events. Road situations and vehicle conditions such as obstacles on roads, poor pavements, slippery road surface, sharp curves, and sensitive accelerating or brake systems can also be reasons for extreme driving events. Researchers have given cut-off acceleration values as a threshold for defining extreme (aggressive) driving and calm (normal) driving (*49*, *50*, *56*, *57*). In light of the varying distributions of instantaneous

accelerations along speeds and at various directions, this study proposes an innovative way to identify extreme events, which are the core information generated for drivers from BSMs and transmitted thru V2V and V2I applications. Our previous studies revealed the extreme acceleration events based on driving speeds (*51, 52*), caused by the limited dimension of vehicle motion data. This study used data that contains both longitudinal and lateral accelerations. Extreme events of acceleration in different directions potentially correspond to different warning and control assist messages. For example, if an extreme event is acceleration going straight ahead, head-on warnings might be generated, and if it is an acceleration going to the right side, vehicles on the right lane might be warned through V2V applications.

Unlike previous studies that give cut-off values as a threshold regardless of driving situations, this study proposes thresholds that change with speeds that account for different driving situations to a greater extent. In addition, the directional variation of acceleration distributions was considered when defining the thresholds. This study used 95th percentile values in each bin as the thresholds. The thresholds could be customized according to the acceptable levels of driving volatility. Since the thresholds were generated based on possible values in a specific bin, they varied at different speeds and directions. Figure 29 presents: (i) an aggregated threshold surface that is enclosed like a cylinder with varying radii at different speeds and directions; (ii) identified extreme acceleration events (i.e., gray dots out of the surface); and (iii) an enlarged image of figure (ii). Note that the threshold surface was fitted using the 95^{th} percentile magnitudes of longitudinal and lateral accelerations within the one speed×sector bin. The radii of the enclosed surface were greater for lower speeds and narrower for high speeds, as shown in Figure 28 (iii) and (iv). Accelerations parallel or perpendicular to the heading directions had greater radii than did those diagonal to heading directions, as shown in Figure 28 (ii).

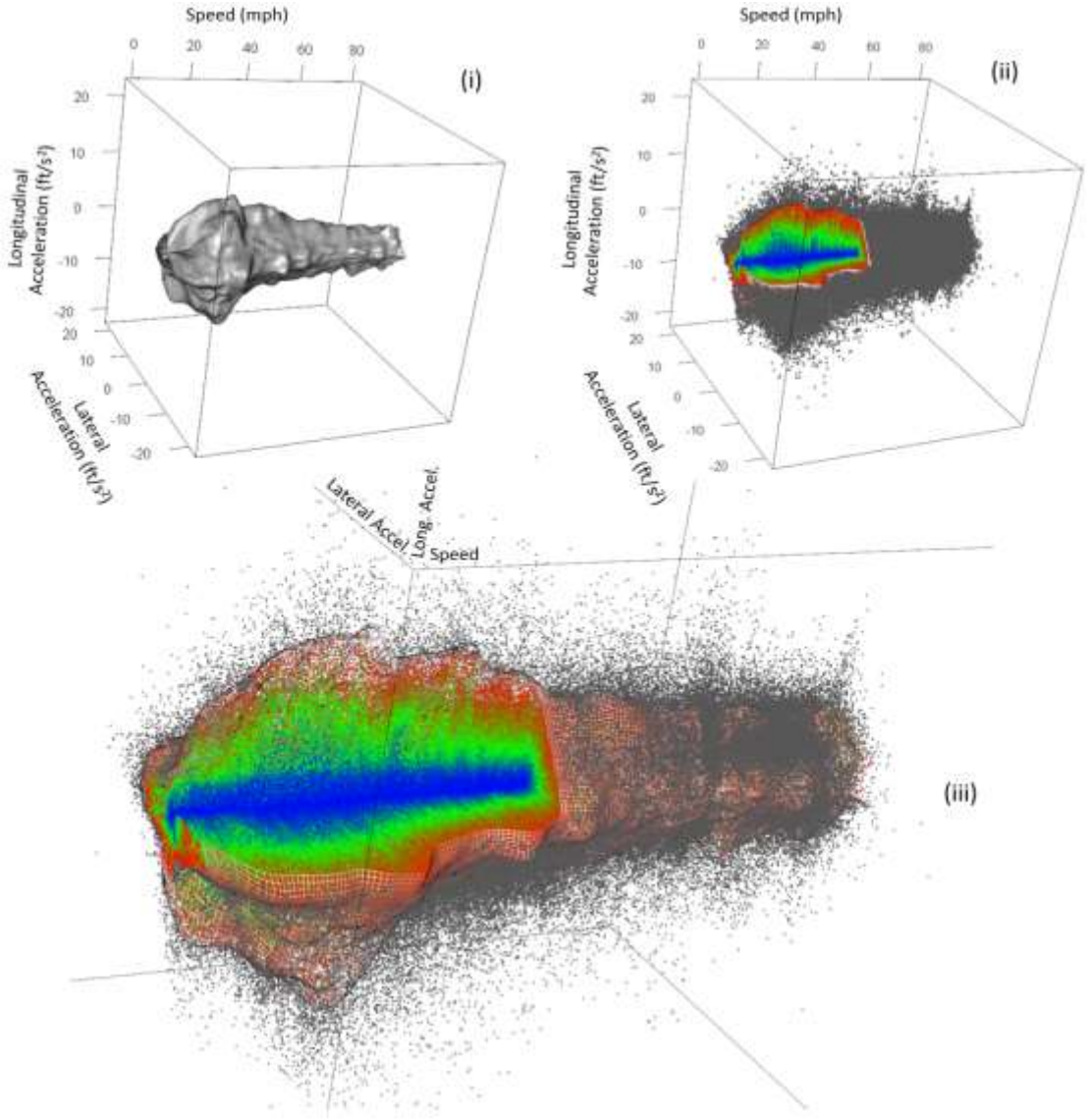


FIGURE 29 Plots of Extreme Acceleration Events (hollowed in side)

Figure 31 presents a sample trip with identified extreme events (see Warnings and Control Assist) and their locations on a map. Note that the warnings and control assists were generated based on the 95th percentile thresholds introduced above, and if more than five successive BSMs (> 0.5 seconds) beyond the 95th percentile thresholds were identified together one warning or control assist can be generated as shown in Figure 30.

Extreme events seemed to be located at critical driving conditions, such as sharp turns and complex intersections. A zoomed-in view of the warning and control assist locations shows that this is a six-way intersection consisting of three two-way roads with pedestrian

sidewalks, as shown in Figure 31(ii). Three locations generated three separate warnings and control assists. Warning and control assist 1 indicated the potential for poor sight distance caused by roadside plantings on East Madison Street. Warning and control assist 2 pointed out driver behaviors influenced by the intersecting traffic from Packard Street. Warning and control assist 3 was possibly associated with the pedestrians crossing South Division Street. The methodology proposed in this study identifies extreme events and locates them accurately in space.

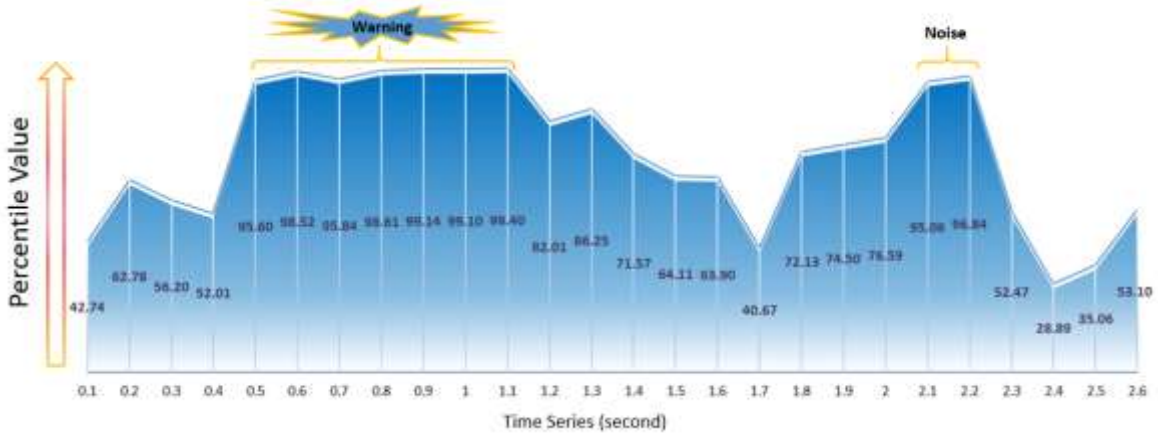


FIGURE 30 Generating Warnings and Control Assists

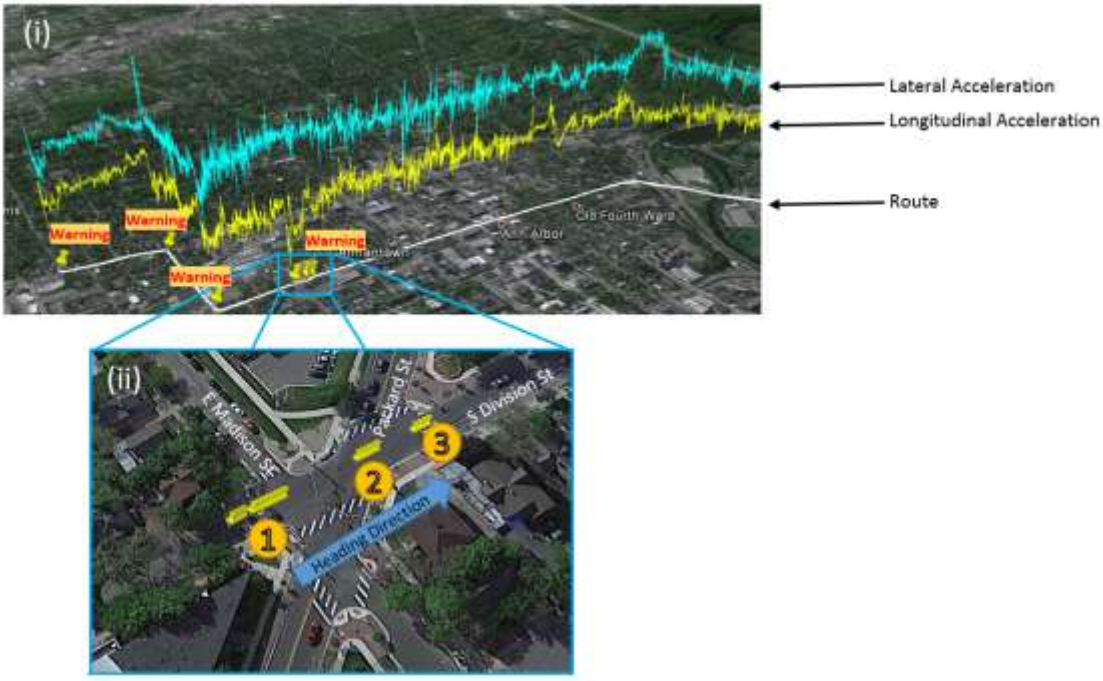


FIGURE 31 Extreme Events Identified in Space

UNDERSTANDING EXTREME EVENTS

One question to be answered after the identification of extreme events was whether these extreme events made sense in terms of vehicle maneuvering and driving. To understand the correlates of extreme events, simple regression models were estimated. Factors included vehicle maneuvering (accelerator, brakes, and cruise control), and instantaneous driving contexts (surrounding objects, distance to closest objects). This study applied Negative Binomial (NB) regression model to predict the number of warning messages generated from BSMs during one trip. The NB regression model is given by:

$$\mu_i = \exp(X_i\beta + \varepsilon_i) \qquad \text{Equation (25)}$$

Where, μ_i = number of warnings and control assists during a trip i, i = 1, 2, 3,…, n; X_i = explanatory variables; β = a vector of estimated coefficients; $\exp(\varepsilon_i)$ = a gamma-distributed error term. The coefficients in NB model are estimated under the NB probability function,

$$\Pr(y_i = k) = \frac{\Gamma[(1/\alpha) + k]}{\Gamma(1/\alpha)\Gamma(k+1)}\left[\frac{1/\alpha}{(1/\alpha) + \mu_i}\right]^{1/\alpha}\left[\frac{\mu_i}{(1/\alpha) + \mu_i}\right]^{k} \qquad \text{Equation (26)}$$

Where, y_i = observed number of warning and control assists during trip i; k = possible number of warning and control assists during a trip; α= over-dispersion parameter. Note that $\mu_i \sim Gamma\left(\frac{1}{\alpha}, \alpha\mu\right)$ and μ are both the mean and variance of observed number of warning and control assists fitted in Poisson distribution. α is the over-dispersion parameter. For Poisson regression model, α = 0. For Negative binomial regress model, α is significantly larger than zero.

Table 16 shows the descriptive statistics of selected variables at trip level. Number of warnings and control assists is the dependent variable in NB model, and other variables related to warning and control assists are also shown in the table for future research purposes. On average, there are 7.23 warnings and control assists per trip. The trip with the greatest number had 39 warning and control assists while some trips had no warning and control assists. The longest warning and control assist lasted 19.1 seconds, as shown in Table 16.

Noticeably, there were some extremely short trips (less than 1 minute) because drivers did not always initiate data collection at the beginning of their trips. For those trips, the data relates to one segment of a trip. Distributions of other variables seemed reasonable, based on error checking. Also, correlations between explanatory variables were checked, and were found to be reasonable for modeling purposes.

TABLE 16 Descriptive Statistics of Variables at Trip-Level

Variable		N	Mean	Std. Dev.	Min	Max
Warning	Number of warning and control assists	155	7.23	8.22	0.00	39.00
	Total warning and control assist duration (second)	155	7.96	10.14	0.00	58.00
	Average duration per warning and control assist (second)	155	0.80	0.62	0.00	3.35
	Longest warning and control assist duration (second)	155	2.11	2.67	0.00	19.10
Trip Attribute	Average speed (mph)	155	33.55	15.01	2.74	71.22
	Speed variance	155	228.96	158.60	0.06	728.58
	Maximum speed (mph)	155	53.06	14.62	12.66	83.35
	Trip length (mile) *	155	6.68	9.46	0.03	72.14
	Trip duration (minute) *	155	10.41	10.92	0.07	60.78
Vehicle Maneuvering	Average accelerator pedal displace (%)	155	13.97%	11.79%	0.00%	64.30%
	Maximum accelerator pedal displace (%)	155	38.28%	28.79%	0.00%	100.00%
	Brake pedal (engaged) (%)	155	23.08%	16.39%	0.00%	86.02%
	Cruise control (engaged) (%)	155	30.16%	37.78%	0.00%	100.00%
	Turn signal (left) (%)	155	4.71%	10.82%	0.00%	97.54%
	Turn signal (right) (%)	155	2.38%	5.70%	0.00%	60.03%
	Number of left turns	155	1.53	2.03	0.00	10.00
	Number of right turns	155	1.53	2.42	0.00	13.00
	Number of turns	155	3.06	3.99	0.00	21.00
Driving Context	Average number of objects	155	1.61	0.95	0.00	3.91
	Maximum number of objects	155	5.02	2.30	0.00	9.00
	Average distance to the closest object (ft)	155	163.76	149.98	21.33	839.69
	Min distance to the closest object (ft)	155	50.33	160.87	0.41	839.69
	Average relative speed of the closest object (mph)	155	-4.38	10.80	-80.86	12.03
	Max relative speed of the closest object (mph)	155	48.52	62.35	-66.27	159.94
	Min relative speed of the closest object (mph)	155	-68.81	45.89	-157.98	3.64

*: There are some extreme short trips because drivers did not always initiate the data collection at the beginning of the trips. For those trips, the data refer to one segment of a trip.

Table 17 presents the NB modeling results, including the full and final models. The likelihood ratio test showed that the over-dispersion parameter α is significantly greater than zero in both models, which validates the use of NB regression instead of Poisson regression. Overall, the modeling results were reasonable, providing insights about the correlates of extreme events embedded in BSMs. Since some variables were highly correlated, such as trip length and trip durations, the final model was estimated by stepwise selection technique (*58*). As expected, higher trip average speeds were correlated with less warning and control assists while longer trips were associated with more warning and control assists. Interestingly, if the maximum speed during a trip is higher, then more warnings and control assists were generated. One possible reason could be that drivers who reach higher speeds may be pressing the accelerator harder and thus are more likely to get warnings and control assists.

For vehicle maneuvering, more time spent on braking was associated with more warnings and control assists, as expected. Driving context was highly correlated to the amount of warnings and control assists. Average number of objects encountered during a trip was negatively correlated to the number of warnings and control assists. The large number of objects indicated a higher level of traffic density and driving complexity, and drivers may have compensated by being more cautious. However, the maximum number of objects encountered was positively associated with the number of warnings and control assists generated. The distance to the closest object was positively associated with the number of warnings and control assists, which is expected. There was a positive correlation between the average relative speed of the closest object and the number of warnings and control assists, implying that the oncoming objects influenced driver behavior significantly. Note that the modeling results were limited to a relatively small sample size (N=155). The estimated coefficients may change as more data become available.

TABLE 17 Negative Binomial Models for Frequency of Extreme Events
(Y=number of warnings/control assists during one trip)

Variable		β	P-value	β	P-value
Constant		-1.030	0.190	-0.361	0.396
Trip Attributes	Average speed (mph)	-0.014	0.582	-0.042**	0.000
	Speed variance	0.000	0.675		
	Maximum speed (mph)	0.043**	0.027	0.049**	0.000
	Trip length (mile)	-0.042	0.212		
	Trip duration (minute)	0.052*	0.083	0.024**	0.041
Vehicle Maneuvering	Average accelerator pedal displacement (%)	-0.004	0.774		
	Maximum accelerator pedal displacement (%)	-0.001	0.858		
	Brake pedal (engaged) (%)	1.546	0.226	2.000**	0.006
	Cruise control (engaged) (%)	-0.455	0.151		
	Turn signal (left) (%)	-0.297	0.804		
	Turn signal (right) (%)	-0.298	0.846		
	Number of turns	0.035	0.214		
Driving Context	Average number of objects	-0.476**	0.014	-0.449**	0.007
	Maximum number of objects	0.212**	0.018	0.205**	0.007
	Average distance to the closest object (ft)	0.002	0.453	0.002**	0.001
	Min distance to the closest object (ft)	0.001	0.669		
	Average relative speed of the closest object (mph)	0.045**	0.012	0.036**	0.008
	Max relative speed of the closest object (mph)	-0.002	0.284		
	Min relative speed of the closest object (mph)	-0.001	0.690		
SUMMARY STATISTICS					
α		0.797		0.854	
Likelihood-ratio test of α=0		354.020**		385.020**	
Number of observations		155		155	
Log Likelihood		-425.012		-428.885	
Log Likelihood $\chi 2$		80.960		73.210	
Prob. > $\chi 2$		0.000		0.000	
Pseudo R^2		0.087		0.079	

Note:
1. ** = significant at a 95% confidence level;
2. * = significant at a 90% confidence level;

LIMITATIONS

Data used in this study were BSMs sent and received by connected vehicles through V2V and V2I applications. Participating vehicles equipped with Data Acquisition Systems (DAS)

collected the BSMs. Thus, the extent of measurement errors in the data was unknown, although the results from data visualization and modeling were reasonable.

A threat to the validity of this study is the limited sample size. Only one-day sample data from the SPMD is publicly available in Research Data Exchange (RDE, https://www.its-rde.net/home). Consequently, this study can be regarded as an exploration of BSMs transmitted by real-world connected vehicle technologies. As more data become available, the methodology of this study can be applied directly for a broader exploration, and the results of this study can be validated by expanded sample data.

CONCLUSION

The connected vehicle technologies discussed in this paper are capable of transmitting high-frequency data between vehicles and infrastructure, which has the potential to improve mobility, safety, energy consumption, and the environment. In this context, it is important to maximize the value embedded in data generated by the entire ecosystem of vehicles and infrastructure to support driver decision-making. SPMD provides data on basic safety messages, which are the core data sent and received by connected vehicles and infrastructure through V2V and V2I applications. The content of BSMs describe a vehicle's position (latitude, longitude, and elevation) and motion (heading, speed, and acceleration) (*10*). BSMs also contains data pertaining to the vehicle's component status (lights, brakes, wipers) and vehicle safety information (path history, events) (*10*). The raw BSMs are complex and not informative to drivers. This study proposes a data analytic methodology to extract critical information from raw BSMs. The information can be provided to drivers and inform them about their driving behaviors or about dangers in surrounding roadway environments. The information is simple and informative, and helps drivers make informed decisions. The research is timely and has long-term value, as connected and automated vehicles are likely to have substantial impacts throughout the world, and have seen substantial research activity.

Our previous studies have explored the critical information embedded in vehicle trajectory data, which were collected through GPS devices (*51, 52*). This study extended the methodology of extracting critical information from BSMs transmitted by V2V and V2I applications. This study established a fundamental understanding of instantaneous driving

decisions by investigating two-dimensional instantaneous accelerations, i.e., longitudinal and lateral accelerations. Instantaneous driving volatility was visualized, and it clearly showed that driving behavior is strongly associated with driving contexts, whether driving on local roads or freeways. This study untangled the relationship between speeds and acceleration through the distribution of instantaneous accelerations at different speeds. Higher speeds (>50 mph) are associated with smaller acceleration magnitudes, which is consistent with Ahn et al. This study further revealed a nonlinear relationship between acceleration and speed in real-life driving situations. The lozenge shaped joint distribution implies that longitudinal and lateral acceleration hardly reached a large magnitude simultaneously. In terms of magnitude, longitudinal and lateral accelerations seemed inversely correlated.

This research presents an original idea, which is to establish context-relevant alert, warning, and control assist thresholds based on extreme event information embedded in BSMs. Most previous studies give fixed cut-off values for thresholds regardless of driving situations (*49, 50, 56, 57*). However, some of the thresholds for warnings and control assists may be flexible and can change with speeds to account for different driving situations and contexts. In addition, the directional variation of acceleration distributions is considered in establishing the thresholds. Information about extreme driving decisions can be used for control assists and provided as feedback to drivers in real time to help them shift to calmer driving, and transmitted to other drivers surrounding through V2V applications to warn them about potential dangers.

Results from rigorous statistical modeling revealed that the extreme events identified from BSMs are highly associated with trip attributes, driver maneuvering, and driving contexts. The results from modeling provide correlates of extreme events.

This study contributes by making sense of high-frequency geo-referenced connected vehicle data, and extracts critical information about extreme events from new data sources created by communications between connected vehicles. Connected vehicles are a relatively new and emerging area of research activity in intelligent transportation systems, with strong interest from a wide audience that includes government agencies, auto makers, practitioners and researchers who are interested in implementing connected vehicles.

The findings of this study are relevant to promoting transportation system objectives

by incorporating alerts, warnings, and control assists in V2V and V2I applications of connected vehicles. This will help drivers identify extreme events surrounding them quickly so they may avoid dangers by taking evasive actions. For example, the warnings can identify dangers of side-collisions (if there are extreme lateral accelerations), forward or rear collisions (if there are extreme longitudinal accelerations), and so on. In addition, vehicle information can be provided to analysts in a traffic operations center to better manage traffic through eco-routing and route diversion decisions.

References

1. BTS. *Estimated U.S. Roadway Lane-Miles by Functional System*. 2014; Available from: http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_06.html.
2. NHTSA, *Overview Traffic Safety Facts 2012*. 2014, National Highway Traffic Safety Administration, US Department of Transportation.
3. Office, T.I.T.S.J.P. *Connected Vehicle Research in the United States*. 2014; Available from: http://www.its.dot.gov/connected_vehicle/connected_vehicle_research.htm.
4. Qin, W.B. and G. Orosz. Digital effects and delays in connected vehicles: linear stability and simulations. in *ASME 2013 Dynamic Systems and Control Conference*. 2013. American Society of Mechanical Engineers.
5. Li, Y., M. Zhao, and W. Wang. Intermittently connected vehicle-to-vehicle networks: detection and analysis. in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*. 2011. IEEE.
6. Lee, J. and B. Park, Development and evaluation of a cooperative vehicle intersection control algorithm under the connected vehicles environment. *Intelligent Transportation Systems, IEEE Transactions on*, Vol. 13, No. 1, 2012: pp. 81-90.
7. Jin, Q., G. Wu, K. Boriboonsomsin, and M. Barth. Advanced intersection management for connected vehicles using a multi-agent systems approach. in *Intelligent Vehicles Symposium (IV), 2012 IEEE*. 2012. IEEE.

8. Goodall, N.J., B.L. Smith, and B. Park, Traffic Signal Control with Connected Vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2381, No. 1, 2013: pp. 65-72.
9. NHTSA, National Motor Vehicle Crash Causation Survey: Report to Congress. *National Highway Traffic Safety Administration Technical Report DOT HS*, Vol. 811, No. 2008: pp. 059.
10. 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.
11. Fagnant, D.J. and K. Kockelman, PREPARING A NATION FOR AUTONOMOUS VEHICLES: 1 OPPORTUNITIES, BARRIERS AND POLICY RECOMMENDATIONS FOR 2 CAPITALIZING ON SELF-DRIVEN VEHICLES 3. *Transportation Research*, Vol. 20, No. 2014: pp.
12. Zhang, W.-B., J.A. Misener, C.-Y. Chan, K. Zhou, and J.-Q. Li. Feasibility Assessment of A Truck Automation Deployment Framework. in *Transportation Research Board 93rd Annual Meeting*. 2014.
13. Koulakezian, A. and A. Leon-Garcia. CVI: Connected vehicle infrastructure for ITS. in *Personal Indoor and Mobile Radio Communications (PIMRC), 2011 IEEE 22nd International Symposium on*. 2011. IEEE.
14. Zeng, X., K.N. Balke, and P. Songchitruksa, *Potential Connected Vehicle Applications to Enhance Mobility, Safety, and Environmental Security*. Southwest Region University Transportation Center, Texas Transportation Institute, Texas A&M University System, 2012.
15. Olia, A., H. Abdelgawad, B. Abdulhai, and S.N. Razavi. Assessing the Potential Impacts of Connected Vehicles: Mobility, Environmental, and Safety Perspectives. in *Transportation Research Board 93rd Annual Meeting*. 2014.
16. Hill, C.J. and J.K. Garrett, *AASHTO connected vehicle infrastructure deployment analysis*. 2011.

17. Zhang, L. and G. Orosz. Designing network motifs in connected vehicle systems: delay effects and stability. in *ASME 2013 Dynamic Systems and Control Conference*. 2013. American Society of Mechanical Engineers.
18. Christofa, E., J. Argote, and A. Skabardonis, Arterial queue spillback detection and signal control based on connected vehicle technology. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2356, 2013: pp. 61-70.
19. Wu, G., K. Boriboonsomsin, H. Xia, and M. Barth, Supplementary Benefits from Partial Vehicle Automation in an Ecoapproach and Departure Application at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. No. 2424, 2014: pp. 66-75.
20. Guler, S.I., M. Menendez, and L. Meier, Using connected vehicle technology to improve the efficiency of intersections. *Transportation Research Part C: Emerging Technologies*, Vol. 46, No. 2014: pp. 121-131.
21. Feng, Y., S. Khoshmagham, M. Zamanipour, and K.L. Head. A Real-time Adaptive Signal Phase Allocation Algorithm in a Connected Vehicle Environment. in *Transportation Research Board 94th Annual Meeting*. 2015.
22. Bagheri, E., B. Mehran, and B. Hellinga. Real-time Estimation Of Saturation Flow Rates For Dynamic Traffic Signal Control Using Connected Vehicle Data. in *Transportation Research Board 94th Annual Meeting*. 2015.
23. Bridgelall, R., Connected vehicle approach for pavement roughness evaluation. *Journal of Infrastructure Systems*, Vol. 20, No. 1, 2013: pp. 04013001.
24. Li, J.-Q., K. Zhou, S. Shladover, and A. Skabardonis, Estimating Queue Length Under Connected Vehicle Technology: Using Probe Vehicle, Loop Detector, and Fused Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. No. 2356, 2013: pp. 17-22.
25. 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. No. 2015: pp. 04015020.
26. Tian, D., Y. Yuan, J. Zhou, Y. Wang, G. Lu, and H. Xia. Real-time vehicle route guidance based on connected vehicles. in *Green Computing and Communications*

(GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing. 2013. IEEE.

27. Kianfar, J. and P. Edara, Placement of Roadside Equipment in Connected Vehicle Environment for Travel Time Estimation. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. No. 2381, 2013: pp. 20-27.
28. Moylan, E. and A. Skabardonis. Reliability-and Median-Based Identification of Toll Locations in a Connected Vehicle Context. in *Transportation Research Board 94th Annual Meeting*. 2015.
29. 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.
30. Doecke, S., A. Grant, and R.W. Anderson, The real-world safety potential of connected vehicle technology. *Traffic injury prevention*, Vol. 16, No. sup1, 2015: pp. S31-S35.
31. Goodall, N.J., B.L. Smith, and B.B. Park, Microscopic estimation of freeway vehicle positions from the behavior of connected vehicles. *Journal of Intelligent Transportation Systems*, Vol. No. ahead-of-print, 2014: pp. 1-10.
32. 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.
33. Osman, O.A., J. Codjoe, and S. Ishak, Impact of Time-to-Collision Information on Driving Behavior in Connected Vehicle Environments Using A Driving Simulator Test Bed. *Journal of Traffic and Logistics Engineering Vol*, Vol. 3, No. 1, 2015: pp.
34. Kishore Kamalanathsharma, R. and H.A. Rakha, Leveraging connected vehicle technology and telematics to enhance vehicle fuel efficiency in the vicinity of signalized intersections. *Journal of Intelligent Transportation Systems*, Vol. No. ahead-of-print, 2014: pp. 1-12.
35. Cheng, L., B.E. Henty, D.D. Stancil, F. Bai, and P. Mudalige, Mobile vehicle-to-vehicle narrow-band channel measurement and characterization of the 5.9 GHz

dedicated short range communication (DSRC) frequency band. *Selected Areas in Communications, IEEE Journal on*, Vol. 25, No. 8, 2007: pp. 1501-1516.

36. Chan, C.-Y. Connected vehicles in a connected world. in *VLSI Design, Automation and Test (VLSI-DAT), 2011 International Symposium on*. 2011. IEEE.
37. Goel, S., T. Imielinski, and K. Ozbay. Ascertaining viability of WiFi based vehicle-to-vehicle network for traffic information dissemination. in *Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on*. 2004. IEEE.
38. Chou, C.-M., C.-Y. Li, W.-M. Chien, and K.-c. Lan. A feasibility study on vehicle-to-infrastructure communication: WiFi vs. WiMAX. in *Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on*. 2009. IEEE.
39. Nusser, R. and R.M. Pelz. Bluetooth-based wireless connectivity in an automotive environment. in *Vehicular Technology Conference, 2000. IEEE-VTS Fall VTC 2000. 52nd*. 2000. IEEE.
40. Sugiura, A. and C. Dermawan, In traffic jam IVC-RVC system for ITS using Bluetooth. *Intelligent Transportation Systems, IEEE Transactions on*, Vol. 6, No. 3, 2005: pp. 302-313.
41. Mosyagin, J. Using 4G wireless technology in the car. in *Transparent Optical Networks (ICTON), 2010 12th International Conference on*. 2010. IEEE.
42. Abid, H., T.C. Chung, S. Lee, and S. Qaisar. Performance analysis of lte smartphones-based vehicle-to-infrastrcuture communication. in *Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th International Conference on*. 2012. IEEE.
43. NHTSA. *Resource Guide Describes Best Practices For Aggressive Driving Enforcement* 2000 [cited 2014 May 15th]; Available from: http://www.nhtsa.gov/About+NHTSA/Traffic+Techs/current/Resource+Guide+Describes+Best+Practices+For+Aggressive+Driving+Enforcement.
44. Lajunen, T., A. Corry, H. Summala, and L. Hartley, Impression management and self-deception in traffic behaviour inventories. *Personality and individual differences*, Vol. 22, No. 3, 1997: pp. 341-353.

45. Lajunen, T., A. Corry, H. Summala, and L. Hartley, Cross-cultural differences in drivers' self-assessments of their perceptual-motor and safety skills: Australians and Finns. *Personality and Individual Differences*, Vol. 24, No. 4, 1998: pp. 539-550.
46. Hoedemaeker, M. and K.A. Brookhuis, Behavioural adaptation to driving with an adaptive cruise control (ACC). *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 1, No. 2, 1998: pp. 95-106.
47. Åberg, L., L. Larsen, A. Glad, and L. Beilinsson, Observed vehicle speed and drivers' perceived speed of others. *Applied Psychology*, Vol. 46, No. 3, 1997: pp. 287-302.
48. 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.
49. 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. Washington DC.
50. 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.
51. Liu, J., X. Wang, and A. Khattak, *Generating Real-Time Driving Volatility Information*, in *2014 World Congress on Intelligent Transport Systems*. 2014: Detroit, MI.
52. 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*, Vol. No. 2015: pp.DOI: 10.1016/j.trc.2014.12.014.
53. Noble, A.M., S.B. Mclaughlin, Z.R. Doerzaph, and T.A. Dingus, Crowd-sourced Connected-vehicle Warning Algorithm using Naturalistic Driving Data. No. 2014.
54. SAE, *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. SAE International, 2014.
55. Ahn, K., H. Rakha, A. Trani, and M. Van Aerde, Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. *Journal of Transportation Engineering*, Vol. 128, No. 2, 2002: pp. 182-190.

56. Lajunen, T., J. Karola, and H. Summala, Speed and acceleration as measures of driving style in young male drivers. *Perceptual and motor skills*, Vol. 85, No. 1, 1997: pp. 3-16.
57. Langari, R. and J.-S. Won, Intelligent energy management agent for a parallel hybrid vehicle-part I: system architecture and design of the driving situation identification process. *Vehicular Technology, IEEE Transactions on*, Vol. 54, No. 3, 2005: pp. 925-934.
58. StataCorp, *Stepwise Estimation*. 2013.