Characterizing Human Driving Behavior Through an Analysis of Naturalistic Driving Data

Gibran Ali

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Mechanical Engineering

Mehdi Ahmadian, Chair
Christopher L North
John B Ferris
Azim Eskandarian
Steve C Southward

December 12, 2022
Blacksburg, Virginia

Keywords: Driving style, vehicle acceleration, crash risk, interactive analytics

Copyright 2023, Gibran Ali

Characterizing Human Driving Behavior Through an Analysis of Naturalistic Driving Data

Gibran Ali

ABSTRACT

Reducing the number of motor vehicle crashes is one of the major challenges of our times. Current strategies to reduce crash rates can be divided into two groups: identifying risky driving behavior prior to crashes to proactively reduce risk and automating some or all human driving tasks using intelligent vehicle systems such as Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS). For successful implementation of either strategy, a deeper understanding of human driving behavior is essential.

This dissertation characterizes human driving behavior through an analysis of a large naturalistic driving study and offers four major contributions to the field. First, it describes the creation of the Surface Accelerations Reference, a catalog of all longitudinal and lateral surface accelerations found in the Second Strategic Highway Research Program Naturalistic Driving Study (SHRP 2 NDS). SHRP 2 NDS is the largest naturalistic driving study in the world with 34.5 million miles of data collected from over 3,500 participants driving in six separate locations across the United States. An algorithm was developed to detect each acceleration epoch and summarize key parameters, such as the mean and maxima of the magnitude, roadway properties, and driver inputs. A statistical profile was then created for each participant describing their acceleration behavior in terms of rates, percentiles, and the magnitude of the strongest event in a distance threshold.

The second major contribution is quantifying the effect of several factors that influence acceleration behavior. The rate of mild to harsh acceleration epochs was modeled using negative binomial distribution-based generalized linear mixed effect models. Roadway speed category, driver age, driver gender, vehicle class, and location were used as fixed effects, and a unique participant identifier was as the random effect. Subcategories of each fixed effect were compared using incident rate ratios. Roadway speed category was found to have the largest effect on acceleration behavior, followed by driver age, vehicle class, and location. This methodology accounts for the major influences while simultaneously ensuring that the comparisons are meaningful and not driven by coincidences of data collection.

The third major contribution is the extraction of acceleration-based long-term driving styles and determining their relationship to crash risk. Rates of acceleration epochs experienced on ≤ 30 mph roadways were used to cluster the participants into four groups. The metrics to cluster the participants were chosen so that they represent long-term driving style and not short-term driving behavior being influenced by transient traffic and environmental conditions. The driving style was also correlated to driving risk by comparing the crash rates, near-crash rates, and speeding behavior of the participants.

Finally, the fourth major contribution is the creation of a set of interactive analytics tools that facilitate quick characterization of human driving during regular as well as safety-critical driving events. These tools enable users to answer a large and open-ended set of research questions that aid in the development of ADAS and ADS components. These analytics tools facilitate the exploration of queries such as how often do certain scenarios occur in naturalistic driving, what is the distribution of key metrics during a particular scenario, or what is the relative composition of various crash datasets? Novel visual analytics principles such as video on demand have been implemented to accelerate the sense-making loop for the user.

Characterizing Human Driving Behavior Through an Analysis of Naturalistic Driving Data

Gibran Ali

GENERAL AUDIENCE ABSTRACT

Naturalistic driving studies collect data from participants driving their own vehicles over an extended period. These studies offer unique perspectives in understanding driving behavior by capturing routine and rare events. Two important aspects of understanding driving behavior are longitudinal acceleration, which indicates how people speed up or slow down, and lateral acceleration, which shows how people take turns. In this dissertation, millions of miles of driving data were analyzed to create an open access acceleration database representing the driving profiles of thousands of drivers. These profiles are useful to understand and model human driving behavior, which is essential for developing advanced vehicle systems and smart roadway infrastructure. The acceleration database was used to quantify the effect of various roadway properties, driver demographics, vehicle classification, and environmental factors on acceleration driving behavior. The acceleration database was also used to define distinct driving styles and their relationship to driving risk.

A set of interactive analytics tools was developed that leverage naturalistic driving data by enabling users to ask a large set of questions and facilitate open-ended analysis. Novel visualization and data presentation techniques were developed to help users extract deeper insight about driving behavior faster than previously exiting tools. These tools will aid in the development and testing of automated driving systems and advanced driver assistance systems.

Dedication

To Ali & Rukhsana, for making it all possible.

To Hafsah, for her companionship.

To Amal, for the joy she brings in our lives.

Acknowledgments

My PhD journey was full of challenges that would have been impossible to overcome without the help of my mentors, colleagues, family, and friends. I am deeply indebted to all of them. First, I would like to thank my advisor, Dr. Mehdi Ahmadian, for his consistent support and guidance throughout this journey. He guided me on the technical aspects of my research and provided great support and mentorship through various professional and personal challenges. I would also like to thank the committee members, Dr. Chris North, Dr. Steve Southward, Dr. John Ferris, and Dr. Azim Eskandarian for their guidance on various aspects of my research. The courses taught by Dr. North and Dr. Southward were highly influential and have positively impacted my research. The discussions with Dr. Ferris and Dr. Eskandarian provided important perspectives that helped improve this dissertation.

My colleagues and friends at Virginia Tech Transportation Institute (VTTI) played a crucial role in the development of ideas presented in this dissertation. I have learned a lot from Dr. Shane McLaughlin through our numerous discussions on disparate topics ranging from research questions presented in this thesis to philosophical queries about living a good life. I will always be grateful for his mentorship. For my PhD research, I have used tools developed by colleagues like Cameron Rainey, Calvin Winkowski, Neal Feierabend, Dr. Zeb Bowden, and Brian Daily. I am deeply thankful for their help and generosity. Numerous discussions with other VTTI researchers such as Vicki Williams, Andy Schaudt, Mac McCall, Dr. Miguel Perez, and Jon Atwood were instrumental in refining the ideas presented in this dissertation. I am also thankful to Carl Cospel, Andy Petersen, and the rest of the VTTI hardware team for happily answering my questions about how something was done over ten years ago. I would also like to thank the VTTI editing team for their help in reviewing this document.

I would like to acknowledge leadership provided by Dr. Jon Hankey through The National Surface Transportation Safety Center for Excellence (NSTSCE) which was instrumental in exploring several ideas during the initial stages of development. I am also deeply thankful to the leadership of Automated Mobility Partnership (AMP), which provided me with the opportunity to bring ideas about interactive analytics to life. Dr. Kevin Kefauver, Dr. Zac Doerzaph, Dr. Tom Dingus, Dr. Shane McLaughlin, Michelle Chaka, and the rest of the AMP team at VTTI played a crucial role in my success.

I am grateful to the members of the mechanical engineering department and the larger Virginia Tech community which created a nurturing environment for exploration of ideas. I have thoroughly enjoyed my time here. A special thanks to Paola Jaramillo and Cameron Rainey for their help in getting me started at VTTI. I would also like to thank my friends and colleagues at Clemson University and National Institute of Technology Srinagar.

I am extremely fortunate to have had the support of my loving family throughout my life. My parents nurtured me and my sister from an early age to value education and to be curious about the world. I vividly remember, when I was still in elementary school, how my father first explained to me what a PhD thesis was. Despite growing up in a region rife with uncertainty and turmoil, my parents prioritized our education over personal comfort. I am forever grateful to them.

Completing my PhD would not have been possible without the constant support of my wife, Hafsah. I am grateful to her for all the love, support, and understanding that she has shown throughout this journey with me. From moving our lives across continents to taking the small things off my plate so that I could focus on my research, she has done it all.

Finally, I am grateful for my daughter, Amal. She brings me immense joy just by her presence and provides perspective about what is truly important in life.

Contents

List of Figures

List of Tables

List of Abbreviations

β	Vector of regression parameters
λ_i	Rate of events per mile
λ_z	The rate of epochs corresponding to the Z^{th} percentile value
μ_i	Number of epochs experienced by the i$^{\text{th}}$ driver
a_x	Longitudinal acceleration
D_T	Total distance traveled by a participant
k	Number of clusters
LCL	The lower limit of the 95% confidence interval
m_i	Number of miles driven by the i$^{\text{th}}$ driver
N_T	Total number of epochs experienced by a driver
N_X	Number of epochs stronger than $X(g)$ experienced by a driver
R_X	Rate of epochs stronger than $X(g)$ per mile
$RR_{category\ A/category\ B}$	The ratio of rate of events observed for subcategory A versus subcategory B
UCL	The upper limit of the 95% confidence interval
X_i	Matrix of covariates for the i$^{\text{th}}$ driver
AD	Automated Driving

ADAS Advanced Driver-Assistance System

ADS Automated Driving System

AMP Automated Mobility Partnership

CAN Controller Area Network

CNC Crashes and Near-Crashes

DAS Data Acquisition System

FARS Fatality Analysis Reporting System

GES General Estimates System

GPS Global Positioning System

IMU Inertial Measurement Unit

IRB Institutional Review Board

MVMT Million Vehicle Miles Travelled

NAS National Academy of Sciences

NCSA National Center for Statistics and Analysis

NHTSA National Highway Traffic Safety Administration

NMVCCS National Motor Vehicle Crash Causation Survey

NRC National Research Council

PII Personally Identifiable Information

SAFETEA-LU The Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users

SHRP 2 NDS The Second Strategic Highway Research Program Naturalistic Driving Study

SUV Sports Utility Vehicle

SVM Automated Driving System

TRB Transportation Research Board

VTTI Virginia Tech Transportation Institute

Chapter 1

Introduction

1.1 Motivation

In 2010, motor vehicle crashes caused 32,999 fatalities, 3.9 million non-fatal injuries, and 24 million damaged vehicles just in the United States of America. The resulting economic cost from loss of life and productivity was $242 billion [20]. Since then, the economic cost of crashes has increased as the total number of crashes has grown. Motor vehicle crashes are one of the top five leading causes of death for ages 1 to 44 and are the most common cause of death for ages 8 to 24 [139]. Therefore, preventing motor vehicle crashes and mitigating their severity is one of the major challenges of our times.

Considerable research has shown that the human driver is the critical factor in over 90% of automotive crashes [34, 117]. Therefore, most crash mitigation strategies focus on preventing driver error. Two general themes have emerged in crash mitigation methods. The first one focuses on identifying drivers having elevated crash risk and providing training to encourage better driving habits. These include department of transportation training programs for vulnerable drivers or even the use of smartphone apps that nudge drivers towards safer driving habits such as avoiding harsh braking events. However, to effectively target vulnerable drivers, it is important to identify them before they cause crashes.

The second theme focuses on the development of technologies that replace the human driver

for some or all driving tasks through advanced driver assistance systems (ADAS) or automated driving systems (ADS). However, for effective implementation of such systems, characteristics of human driving need to be understood and modeled. This is essential information for intelligent vehicle systems to predict the behavior of other road users as well as to meet the comfort expectations of human passengers.

Several parameters such as vehicle speed, acceleration, and following distance are used to characterize human driving behavior. Acceleration is an important indicator of transience between various driving states. For example, a longitudinal acceleration epoch represents a period of continuous speed gain, and a longitudinal deceleration epoch represents a period of continuous slowing down. Similarly, lateral acceleration epochs represent a continuous change in direction. The intensity, duration, and frequency of longitudinal and lateral acceleration epochs contain valuable information that needs to be characterized for understanding the norms, as well as the extremes, of human driving behavior. This will be beneficial to not only identify vulnerable driving behavior before crashes happen, but also for better emulation of human driving style by intelligent transportation systems.

There are many data sources for characterizing driving behavior, such as national crash datasets, test track studies, field operation tests, and naturalistic driving studies. Each of these sources offers unique perspectives for characterizing human driving behavior. However, large naturalistic driving studies such as the 2nd Strategic Highway Research Program Naturalistic Driving Study (SHPR 2 NDS) offer deeper insights into driving behavior as they capture routine driving as well as rare safety-critical events for the same driving population. For effective utilization of SHR P2 NDS data to characterize driving behavior, efficient data mining algorithms and interactive analytics tools are necessary.

To summarize, the work presented in this dissertation is motivated by the need to:

- characterize human driving behavior from an acceleration perspective
- extract insights about human driving that will help identify vulnerable drivers before crashes occur
- model human comfort preferences embedded in driving style and present them in a form that can be emulated by intelligent driving systems
- build tools that analyze naturalistic driving data and allow for quick extraction of insights about human driving characteristics

1.2 Objectives

The objective of this dissertation is to characterize human driving behavior through an analysis of naturalistic driving data for improving driving safety and to advance the development of intelligent vehicle systems. This will be accomplished by achieving the following objectives.

1. Developing a standardized acceleration catalog that is based on a large and diverse naturalistic driving study.
2. Quantifying the effect of roadway characteristics, driver factors, vehicle class, and location on acceleration behavior.
3. Extracting acceleration-based driving style and assessing its relationship to crash risk for implementation in intelligent vehicle systems.
4. Creating a set of interactive analytics tools that enable users to quickly characterize human driving through an exploration of naturalistic driving data.

1.3 Approach

The data used in this research has been sourced from the SHRP 2 NDS. It is the world's largest naturalistic driving dataset, with over 34 million miles of driving data collected from 3,500 participants driving their own vehicles in naturalistic conditions for several months. The participants were chosen to represent a diverse dataset on the basis of numerous factors such as driver age, gender, location, and vehicle class. Data collected in the SHRP 2 NDS is ideal for this research due to its large scale, diversity, and rich context.

The Surface Accelerations Reference, a catalog of longitudinal and lateral accelerations, was created by analyzing all 34 million miles of driving data in the SHRP 2 NDS. An algorithm was designed to detect and summarize epochs of longitudinal and lateral accelerations. The summary measures included values describing mean and maximum magnitude within an epoch, roadway properties, and driver inputs. Statistical profiles were then created for each participant describing their driving history in the study. These included rates of acceleration at various thresholds per mile, percentiles, and the magnitude of the strongest acceleration in a certain unit of distance. An interactive analytics tool was also created to visualize and query the dataset, facilitating users interested in asking specific questions about acceleration behavior.

To understand the simultaneous effect of various factors on the rates of mild to harsh acceleration epochs, a generalized linear mixed effect modeling approach was used. Roadway speed category, driver age, driver gender, location, and vehicle class were modeled as fixed effects, and a unique driver identifier was modeled as the random effect. Incident rate ratios were then calculated to compare subcategories of each fixed effect. Smaller rate ratios indicated less difference between two subcategories, and larger value indicated a larger effect.

To differentiate long-term driving style from driving behavior, a novel approach was used.

Data from the Surface Accelerations Reference was specifically selected so that long-term driving style is the main influence and the effect of short-term driving behavior in response to traffic and environmental conditions is largely eliminated. Rates of longitudinal and lateral acceleration epochs on roads with speed category of $\leq$ 30 mph at thresholds of $\geq$ 0.3 g, $\geq$ 0.4 g, $\geq$ 0.5 g, and $\geq$ 0.6 g were chosen. Based on this data, an unsupervised clustering algorithm was used to characterize the drivers into four groups named "*low*", "*mid A*", "*mid B*", and "*high*". The safety record of each group was compared using crash rate, crash plus near-crash rate, and speeding behavior.

A set of interactive analytics tools have been developed to characterize human driving through an analysis of naturalistic driving data. These tools are a part of the Automated Mobility Partnership (AMP) which brings together 13 industry leaders to work on their immediate ADS deployment needs. These tools enable users to:

- compare the composition of crashes between naturalistic driving and national datasets such as Fatality Analysis Reporting System (FARS) and General Estimates System (GES)
- compare rates of various epochs of interest in naturalistic driving datasets
- visualize distribution of parameters such as speed and accelerations during various epochs of interest
- develop a multidimensional understanding of epochs of interest and filter down to the appropriate subset needed for their analysis

The interactive analytics tools have been optimized for providing deeper insights by accelerating the user's sense-making loop through novel details-on-demand features.

1.4 Outline

- Chapter 1 discusses the motivation, objectives, approach, future work, and outline of the dissertation.

- Chapter 2 gives a background of driving safety and the use of naturalistic driving data to answer to answer safety related questions.

- Chapter 3 gives an overview of the data used in this dissertation by describing the SHRP 2 NDS, the various variables recorded, the standardization and augmentation process, and the composition of the data.

- Chapter 4 discusses the creation of the Surface Accelerations Reference. It describes the various steps performed to detect and summarize all acceleration epochs in the SHRP 2 NDS. It also describes the creation of the statistical profiles representing each driver's acceleration behavior. A brief overview of the interactive analytics tool is also provided. Finally, it illustrates the utility of the Surface Accelerations Reference by comparing the rates of longitudinal and lateral acceleration epoch in different roadway speed categories.

- Chapter 5 quantifies the effect of roadway, driver, vehicle, and location characteristics on the frequency of longitudinal and lateral accelerations.

- Chapter 6 describes the process of extracting long-term acceleration-based driving styles and their relationship with driving risk.

- Chapter 7 discusses the interactive analytics tools developed to answer a large set of research questions from naturalistic driving data. These tools are used by a technically diverse audience to develop ADAS and ADS.

- Chapter 8 lists the major contributions and possible future work.
- Appendix A analyzes the longitudinal and lateral acceleration signal processing that takes place on the data acquisition system before being stored for long-term use.

Chapter 2

Background

2.1 Historical Perspective of Roadway Transportation Safety Research

Reducing motor vehicle crashes has been a major challenge since the advent of the automobile. Figure 2.1 shows the historical trends of mileage in million vehicle miles traveled (MVMT), motor vehicle crash fatalities, and the corresponding fatality rate per 100 MVMT for the last hundred years in the United States of America. The total number of miles driven per year has been consistently increasing, and now more than 3 trillion miles are driven every year. The number of net fatalities peaked in the 1970s and there has since been a slow overall decline despite the considerable increase in total miles driven. The rate of fatalities per 100 MVMT has had a downward trend that has fluctuated around 1.1 fatalities per 100 MVMT over the last decade.

Figure 2.2 illustrates the yearly numbers for fatal crashes, non-fatal injury-causing crashes, property damage causing crashes, yearly MVMT, total crash rates, and fatality rates. It should be noted that the number of non-fatal crashes is about two orders of magnitude higher than fatal crashes. Since these data are based on police reports, the actual number of non-fatal crashes would be even higher as is shown by other crash data sources such as naturalistic driving studies [98][54].

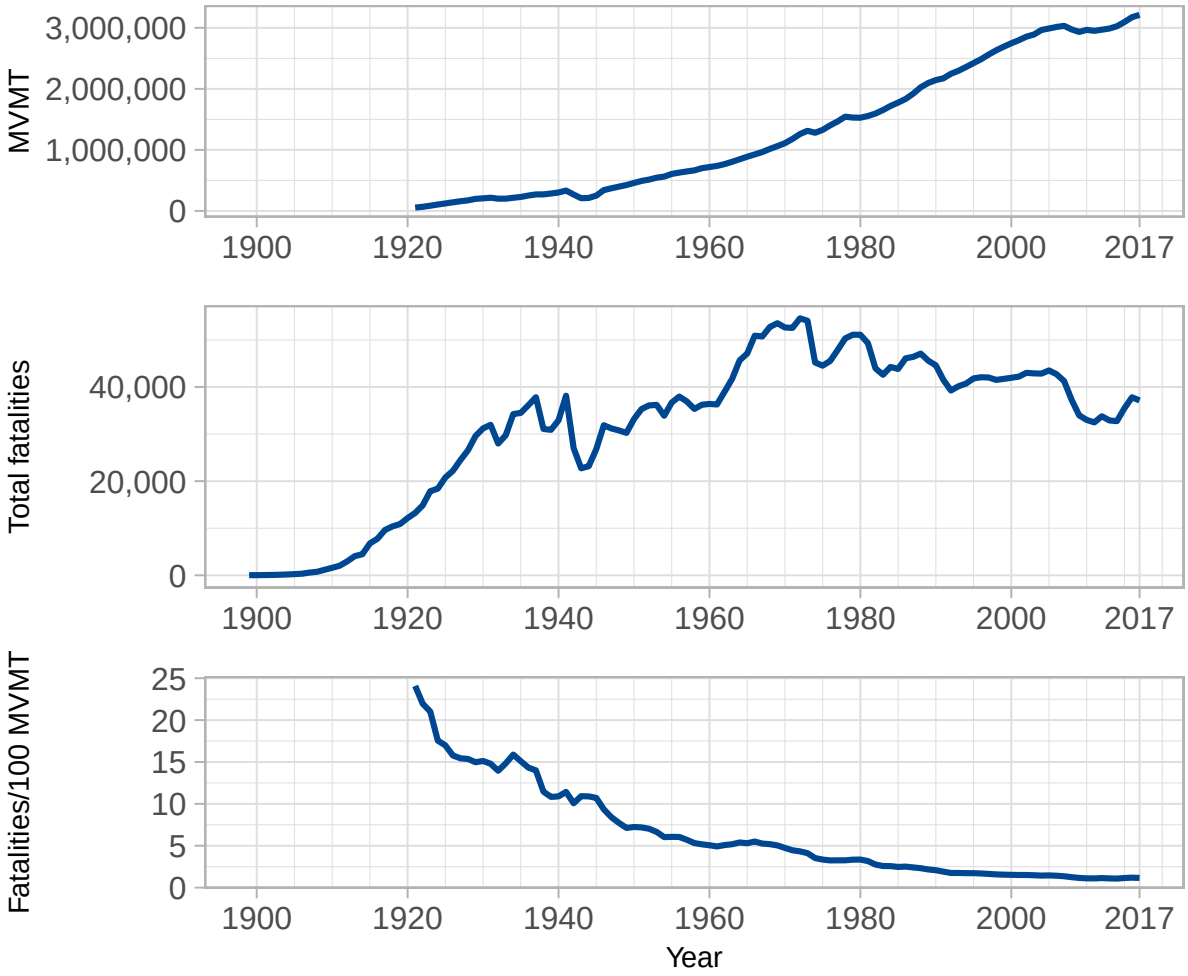


Figure 2.1: Yearly data for million vehicle miles travelled (MVMT), motor vehicle crash fatalities, and fatalities per 100 MVMT in USA from 1899 to 2017 [97] [79].

Despite the considerable reduction in fatalities over the last 50 years, progress has plateaued because of the corresponding increase in vehicle miles traveled. To bring the vision of "Towards Zero Deaths" closer to reality, newer strategies and approaches are needed [123]. Risky drivers need to be targeted before they cause safety-critical events. Newer technologies such as driver assistance and self-driving systems also offer possible solutions that could drastically reduce motor vehicle crashes. However, development of such systems would require deeper understanding of driver behavior for better emulation as well as prediction on a shared road.

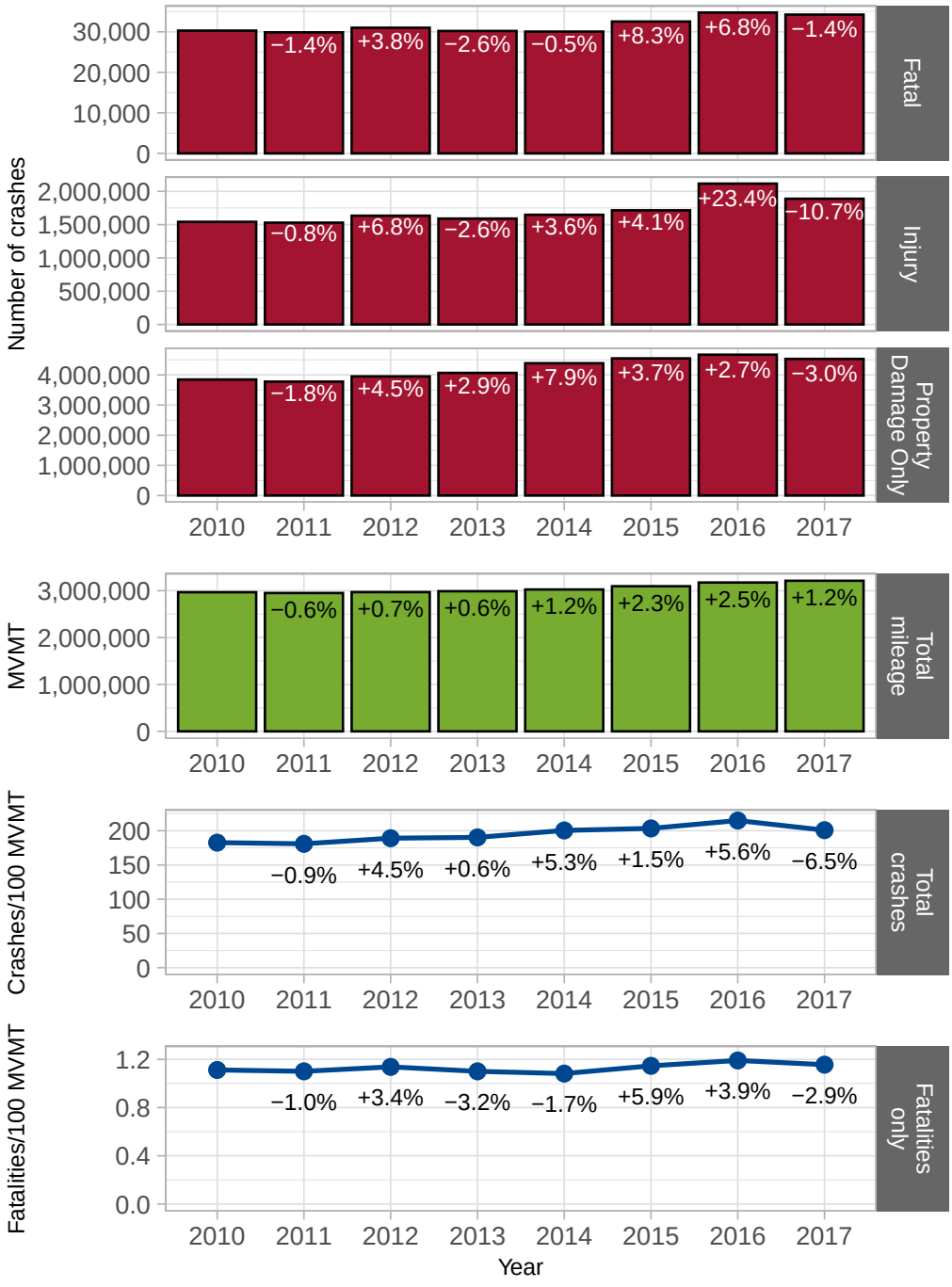


Figure 2.2: NHTSA crash counts, mileage in MVMT, and crash rates with yearly percentage change in USA from 2010 to 2017 [97].

2.2 Predominant Causes of Motor Vehicle Crashes

To understand the factors leading up to crashes, NHTSA's National Center for Statistics and Analysis (NCSA) conducted the National Motor Vehicle Crash Causation Survey (NMVCCS) from 2005 to 2007 and determined the critical reason for the pre-crash event [117] [96]. Singh describes the critical reason as "often the last failure in the causal chain of events leading up to the crash"[117]. Figure 2.3 clearly shows that the driver is the predominant "critical reason" for most crashes (93.5%) with environment (2.4%), vehicles (2%), and other reasons (2.1%) constituting a much smaller proportion.

Figure 2.4 illustrates the composition of the the 93.5% driver-related crashes by type of driver error. Recognition error, such as inattention, internal and external distractions, and inadequate surveillance, was the critical reason for 41% of the driver-related crashes. Decision error, such as driving too fast for conditions, too fast for curve, false assumptions of other's actions, misjudgment of gap or other vehicle's speed, and illegal maneuvers, was the critical

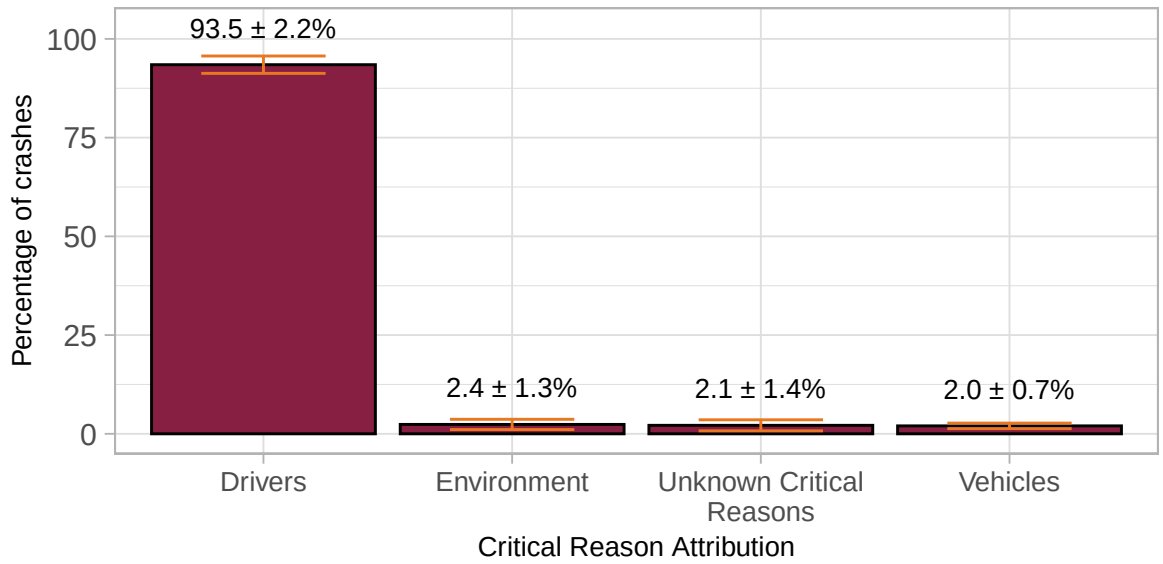


Figure 2.3: Critical reasons for crashes investigated in NMVCCS [117].

reason for 33% of the driver-related crashes. Performance errors, such as overcompensation and poor direction control, constituted another 10.3% of the driver-related crashes. Sleep was the most common type of non-performance error, which constituted 7.1% of driver-related crashes. The remaining 7.1% driver-related crashes had other driver-related critical errors that could not be grouped together into a single category. However, it should be noted that this analysis only includes crashes where emergency medical services were dispatched, police were present when the NMVCCS researcher arrived, at least one of the vehicles was towed, and a completed police accident report was available [117]. These limitations could bias the crash sample and make it different from that of other sources.

A similar analysis was conducted by Dingus et al. using naturalisic driving data that also found the driver to be the predominant factor in most crashes [34]. Figure 2.5 summarizes the findings from the analysis and illustrates that in only 12.3% of the crashes, the driver was not distracted, did not commit an error, and was not impaired. In all the other 87.7% crashes, some combination of a driver factor was present. Even though Singh used epidemiological

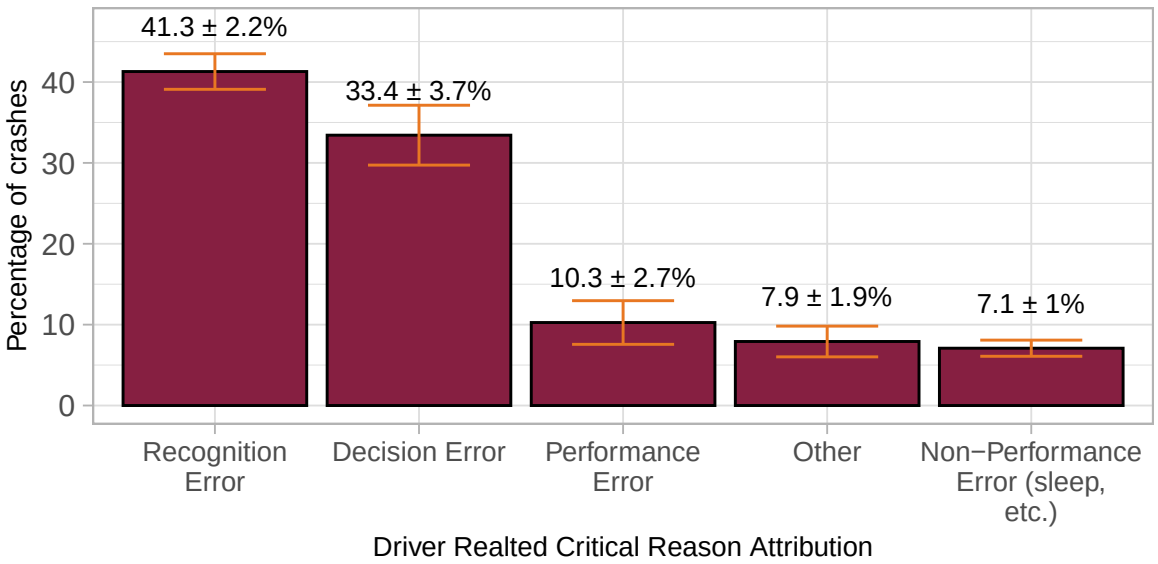


Figure 2.4: Driver related critical reasons for crashes investigated in NMVCCS [117].

data [117], and Dingus et al. used naturalistic data [34], the proportion of crashes with driver factors are similar.

It is clear that driver error, distraction, or impairment are, by far, the leading cause of unsafe driving on roadways. Therefore, newer technologies automating driving tasks have great potential for improving driving safety. However, for such technologies to be adopted widely, it is important to develop a deep understanding of driving behavior. As automation technologies need to meet driver and passenger expectations, there is need for a database of normative driver behavior in various driving conditions and scenarios. Similarly, even fully autonomous vehicles would need to share the road with human drivers and therefore conform to their expectations.

A deeper understanding of driver behavior can also help identify at-risk drivers before they cause crashes and proactively improve roadway safety. For example, demographic markers, such as age group, combined with certain kinematic driving signatures, can be used to identify vulnerable groups and tailor specific training or tests to ensure that they are fully prepared for the road.

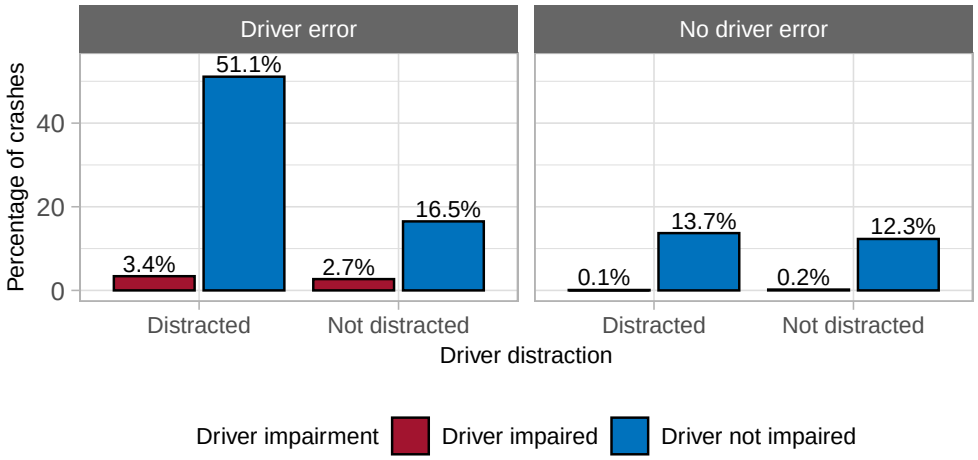


Figure 2.5: Prevalence of driver factors before crashes in SHRP 2 NDS. [34].

2.3 Traditional Approaches to Understanding Driver Behavior

Figure 2.6 shows the pyramid of traffic events based on Hayden [62] and Tarko [125]. It illustrates the relative rarity of crashes and near-crashes with respect to normal driving events when no hazard is present or driving events when a hazard is present but the scenario does not escalate to a near-crash. This relative rarity is a major challenge in understanding driver behavior during crashes and near-crashes. Traditionally, researchers have relied on either simulator and test track-based empirical methods or epidemiological datasets to understand driver behavior. Both these methods have certain advantages but lack the holistic context needed for truly comprehending the problem space [33].

2.3.1 Simulator or test track-based empirical data collection

In a simulator or test track-based empirical data collection study, test subjects are usually given a task while a researcher observes them. In such a study, the dependent variables being studied are predetermined, and the independent variables can be proactively set. This

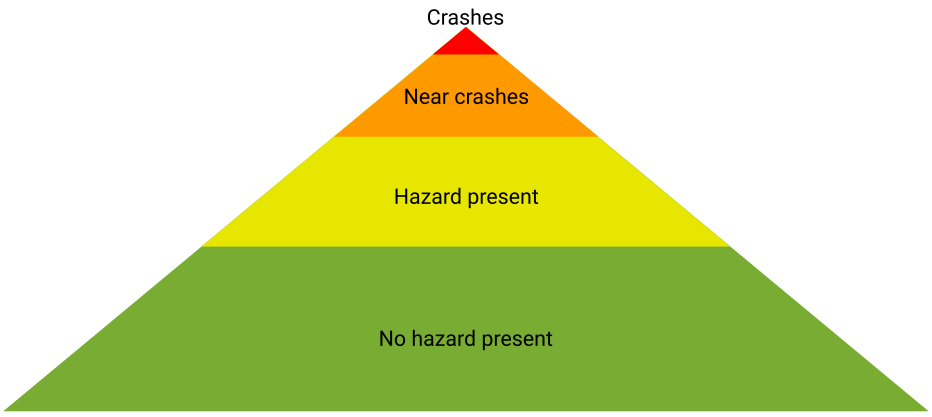


Figure 2.6: Pyramid of traffic events. [62][125].

allows the experimenter to get orthogonal crash risk information. However, in such studies, the test subject behavior is often altered by the presence of the experimenter and the natural behavior may not be captured. Another major drawback of such studies is that they rely on unproven safety surrogates. Often, the simplifications necessary to make such studies feasible prevent the researchers from capturing real-world complexity [33].

2.3.2 Epidemiological data collection

Epidemiological crash datasets, such as the Fatality Analysis Reporting System (FARS) maintained by NHTSA, are invaluable sources of crash risk information as they capture crashes from all over the country and provide important information about the circumstances that led to the crash. However, all the pre-crash information is recreated and is not captured while the crash was taking place. Also, these datasets are reactive, i.e., they only capture events once they have happened and do not capture near-crashes or other conflicts. These datasets also lack driver behavior information during normal driving scenarios, which prevents identification of at-risk markers that could be used to proactively target training programs [33].

Therefore, to get a comprehensive understanding of driver behavior, a data collection method was needed that overcame the disadvantages of the above mentioned data collection methods. Over the last two decades, researchers at VTTI have pioneered naturalistic driving studies that not only overcome the disadvantages of simulator/test track and epidemiological datasets, but also facilitate new types of driving behavior analyses that were hitherto not possible.

2.4 Naturalistic Driving Studies

Naturalistic driving studies observe drivers in their natural driving environment. This is usually achieved by unobtrusively instrumenting a driver's own vehicle and collecting data over a long period of time ranging from a few months to even years. The usual instrumentation consists of a data acquisition system (DAS) collecting data from video cameras, a GPS module, an IMU module, a radar module, and various vehicle sensors communicating on the CAN bus [35, 36, 54, 56, 98]. The drivers are not given any special instructions and no experimenters are present, which ensures that the Hawthorne effect is not a major factor.

The naturalistic driving study data collection method was applied in 1999 by Hanowski et al.[56], to study fatigue and driver performance in long-/short-haul drivers. This study enrolled 42 drivers and accumulated about 1,000 hours of driving. In 2001, Dingus et al. [36] studied the impact of sleeper berth usage on driver fatigue using similar methods. The 100-Car Naturalistic Driving Study [33] was the first such large-scale study with 241 participants and 100 instrumented vehicles accumulating about 2 million miles and 43,000 hours of driving. The SHRP 2 naturalistic driving study was an even larger scaled study conducted between 2010 and 2013 by Dingus et al. [35]. This study accumulated about 34.4 million miles and 1.19 million hours of driving from 3,546 participants driving 3,353 vehicles in six locations across the United States of America. In the last few years, similar but smaller naturalistic studies have been conducted in Canada and the European Union [25][68]. Table 2.1 lists the major known naturalistic driving studies.

So far, over 1,900 crashes and over 8,000 near-crashes have been discovered from the SHRP 2 NDS. Therefore, such large naturalistic driving studies are ideal for driving behavior research as they offer insights into all the regions shown in Figure 2.6. Not only do these studies provide examples of normal driving, hazardous driving without safety-critical events, near-

Table 2.1: Major naturalistic driving studies.

Naturalistic driving study	**Participants**	**Vehicles**	**Mileage (million miles)**	**Time (hours)**
The Impact of Local/Short Haul Operations on Driver Fatigue (1999) [55]	42	4	-	1,000
Impact of Sleeper Berth Usage on Driver Fatigue (2002) [36]	56	2	-	250
100-Car NDS (2004 - 2005) [33]	241	100	2	43,000
SHRP2 NDS (2010 - 2013) [35]	3,546	3,353	34.4	1,191,615
Canada NDS (2013 - 2015) [68]	140	146	1.59	49,000
UDRIVE: the European NDS (2015 - 2017) [25]	281	281	1.42	87,871

crashes, and crashes, they provide examples of each for the same driving population.

2.5 Using Naturalistic Driving Studies for Driver Behavior Research

Traditionally, naturalistic driving studies such as the SHRP 2 NDS have been used to understand safety-critical events (SCEs) and their correlation with certain behaviors. This has been achieved through an analysis of SCEs and baselines. Each trip in the SHRP 2 NDS has been analyzed through an SCE detector algorithm that flags trip IDs and timestamps on the basis of certain kinematic signatures such as crossing a deceleration threshold [54, 104]. These flagged epochs were then viewed by video reductionists, who determined whether the epoch was a crash, a near-crash, or a false positive. If the epoch was an SCE, further reduction was performed to describe various environmental, traffic, and driver factors. Baselines are selected randomly from the whole dataset and therefore can be regarded as representative. The same variables are annotated as those for SCEs to facilitate comparisons between the two.

In the past, using a complex and computationally expensive algorithm on all of SHRP 2 NDS has been prohibitive due to the time required to process all the data. However, with advancements in algorithm efficiency, as well as cheaper parallel computing resources, it is now possible to analyze such large datasets more extensively. This has opened several avenues of utilizing large naturalistic driving studies for analyzing routine human driving behavior. For example, instead of analyzing regular acceleration behavior through baselines which form less than 1 percent of driving, what if each and every acceleration epoch in the 34.5 million miles of the SHRP 2 NDS could be searched, summarized, and logged? It would be a paradigm shift in how driving behavior could be analyzed.

Chapter 3

Data Overview

3.1 Introduction

The driving data analyzed in this dissertation was collected during the Second Strategic Highway Research Program naturalistic driving study (SHR P2 NDS), which is the largest collection of naturalistic driving data in the world. The study collected 34.4 million miles and 1.2 million hours of driving from 3,546 participants driving 3,353 instrumented vehicles in six locations across the United States of America. The data was collected over 2 years between November 2010 and December 2013 and has a diverse set of driver demographics and driving scenarios. Participants' own vehicles were instrumented unobtrusively, and participants were not given any special instructions, resulting in minimal Hawthorne effect.

The drivers' ages ranged from 16 to 95 years with younger and older drivers being deliberately over sampled. The SHRP 2 NDS vehicle fleet comprises cars, pickup trucks, SUVs, and vans, thereby representing all light vehicle types in the U.S. national fleet. The aggregated dataset includes a wide variety of geographic features, roadway types, and climates. The SHRP 2 NDS data is inclusive of the national data in many respects, but depending on the research question appropriate weighting can be applied for direct comparisons [12]. All these features together make the SHRP 2 NDS the most comprehensive resource for studying and understanding driver behavior.

This chapter first gives a brief background of the SHRP 2 program and how it came into being. It lists all the time series variables available as a part of the dataset. Then, the various data standardization and augmentation processes are described. Finally, the composition of the dataset is illustrated in terms of driver demographics, vehicle classification, driving type, and various roadway characteristics. This is also done to highlight the inherent biases of the dataset so that appropriate conclusions can be drawn from the data.

3.2 SHRP 2 NDS Overview

The U.S. Congress authorized SHRP 2 in August 2005 to research roadway safety and congestion as part of the Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU). The program was managed by the Transportation Research Board (TRB) on behalf of the National Research Council(NRC). A major portion of the SHRP 2 initiative was to conduct a massive naturalistic driving study that would give insight into drivers' behavior, their secondary tasks, and the major factors leading up to a crash. The goal of the NDS was to be a rich source of driving data for researchers, regulators, law makers, advocates, and other interested parties for a least a generation [35] [11].

SHRP 2 focused on the four research areas of safety, renewal, reliability, and capacity. The safety portion of SHRP 2 consisted of 12 projects (S01 - S12, and the following were directly involved in the naturalistic driving study design or collection:[2][11]

- S01: Develop NDS Analysis Methods (Q2 2007 to Q1 2009)
- S02: Develop NDS Analysis Plan (Q2 2008 to Q3 2009)
- S05: Naturalistic Driving Study Design (Q2 2007 to Q2 2009)

- S06: Technical Coordination and Quality Control (Q2 2009 to Q2 2013)
- S12A: DAS Procurement (Q3 2009 to Q4 2010)
- S07: NDS Sites - actual data collection (Q1 2010 to Q2 2013)

Protecting participant privacy and data was one of the most critical tasks of the SHRP 2 NDS. The data collection procedures during the study, as well as the storage and use processes after completion, complied with the Institutional Review Board (IRB) policies of Virginia Tech and the National Academy of Sciences (NAS). The policies were directly drawn from the Code of Federal Regulations, Title 45 Public Welfare, Department of Health and Human Services, Part 46, Protection of Human Subjects (45 CFR 46) [35]. For this dissertation, a version of the overall dataset was used that did not contain any personally identifiable information (PII) and therefore, none of the data published or included with this dissertation has any PII.

3.3 Data Recorded

The data acquisition system was developed by VTTI and every vehicle in the study used the same type of DAS. Table 3.1 lists the various data variables that are available as a part of the SHRP 2 NDS dataset. Variables sourced from the VTTI-installed modules, such as the inertial measurement unit (IMU), cameras, radars, etc., are standard across the collection, and variables that are sourced from the vehicle Controller Area Network (CAN) may not be available in all vehicles as data format on CAN buses differs between the manufacturers. The non-standardized CAN variables may also differ in frequency and units and therefore need standardization to be used in a manufacturer agnostic way.

Table 3.1: List of time series signals available as a part of SHRP 2 NDS dataset.

Variable name	**Description**	**Units**	**Frequency**	**Source**
abs	Identifies when antilock braking system is activated	binary	State change	Vehicle CAN
accel_x	Longitudinal acceleration of the vehicle	g	10 Hz	IMU installed by VTTI
accel_y	Lateral acceleration of the vehicle	g	10 Hz	IMU installed by VTTI
accel_z	Vertical acceleration of the vehicle	g	10 Hz	IMU installed by VTTI
cruise_state	Identifies when cruise control is activated	Binary	State change	Vehicle CAN
engine_rpm_instant	Engine crankshaft rotations per minute	rpm	Variable	Vehicle CAN
esc	Identifies when electronic stability control is activated	binary	State change	Vehicle CAN
gyro_x	Roll rate (rate of rotation along x axis) of the vehicle	degree/sec	10 Hz	IMU installed by VTTI
gyro_y	Pitch rate (rate of rotation along y axis) of the vehicle	degree/sec	10 Hz	IMU installed by VTTI

Table 3.1 continued from previous page

Variable name	Description	Units	Frequency	Source
gyro_z	Yaw rate (rate of rotation along z axis) of the vehicle	degree/sec	10 Hz	IMU installed by VTTI
heading_gps	Vehicle heading	degrees from north	1 Hz	IMU installed by VTTI
headlight	Identifies when vehicle headlights are on	binary	State change	Vehicle CAN
lane_distance_off_center	Distance between center of vehicle to center of lane	cm	15 Hz	Machine vision algorithms on video feeds
lane_width	Lane width of current lane	cm	15 Hz	Machine vision algorithms on video feeds
latitude	Latitude of the vehicle	degrees	1 Hz	GPS sensor installed by VTTI
left_line_right_distance	Distance of the center of the vehicle from the left lane line markings	cm	15 Hz	Machine vision algorithms on video feeds
left_marker_probability	Probability that the left lane line markings are correctly identified	0-1000	15 Hz	Machine vision algorithms on video feeds

Table 3.1 continued from previous page

Variable name	Description	Units	Frequency	Source
left_marker_type	Type of left lane line marker	Categorical	15 Hz	Machine vision algorithms on video feeds
light_level	Illuminance in front of the vehicle	lux	10 Hz	Light level sensor installed by VTTI
longitude	Longitude of the vehicle	degrees	1 Hz	GPS sensor installed by VTTI
number_of_satellites	The number of satellites tracked by the GPS sensor	N	1 Hz	GPS sensor installed by VTTI
object_id_t1	Unique identifier of object tracked by radar. 7 objects are tracked simultanously	Number	10 Hz	SMS Radar
pdop	Position of dilution of precision for the GPS system	N.A.	1 Hz	GPS sensor installed by VTTI
pedal_brake_state	Status of the brake pedal	binary	State change	Vehicle CAN
pedal_gas_position	Position of gas pedal	Percent pressed	Variable	Vehicle CAN

Table 3.1 continued from previous page

Variable name	Description	Units	Frequency	Source
prndl	Transmission gear selector	Categorical	Variable	Vehicle CAN
range_rate_x_t1	Relative longitudinal speed of the object tracked by radar with respect to the subject vehicle	meter/sec	10 Hz	SMS Radar
range_rate_y_t1	Relative lateral speed of the object tracked by radar with respect to the subject vehicle	meter	10 Hz	SMS Radar
range_x_t1	Relative longitudinal position of the object tracked by radar with respect to the subject vehicle	meter/sec	10 Hz	SMS Radar
range_y_t1	Relative lateral position of the object tracked by radar with respect to the subject vehicle	meter	10 Hz	SMS Radar
right_line_left_distance	Distance of the center of the vehicle from the right lane line markings	cm	15 Hz	Machine vision algorithms on video feeds

Table 3.1 continued from previous page

Variable name	Description	Units	Frequency	Source
right_marker_probability	Probability that the right lane line markings are correctly identified	0-1000	15 Hz	Machine vision algorithms on video feeds
right_marker_type	Type of right lane line marker	Categorical	15 Hz	Machine vision algorithms on video feeds
seatbelt_driver	Driver seatbelt status	binary	State change	Vehicle CAN
speed_gps	Vehicle speed	kph	1 Hz	GPS sensor installed by VTTI
speed_network	Vehicle speed	kph	1-100 Hz Variable	Vehicle CAN
steering_wheel_position	Steering angle	degrees	Variable	Vehicle CAN
temperature_interior	Temperature of the head unit installed by VTTI	degree C	1 Hz	Headunit installed by VTTI
timestamp	Timestamp from the start of the trip	millisecond	1000 Hz	Head unit installed by VTTI
traction_control_state	Identifies when the traction control system is active	binary	State change	Vehicle CAN
turn_signal	Status of the turn signal	binary	State change	Vehicle CAN

Table 3.1 continued from previous page

Variable name	Description	Units	Frequency	Source
wiper	Status of the wiper	binary	State change	Vehicle CAN

3.4 Data Standardization

Over 3,500 participants were enrolled in SHRP 2 NDS driving 3,353 different vehicles made by 36 manufacturers shown in Figure 3.1. This presented a major standardization challenge as the variables collected from different manufacturers would often have different units, frequencies, and data structures. To overcome this problem, researchers at VTTI created a standardization process that would transform a variable from different vehicles into a standard variable that would have the same units irrespective of the original variable. This was essential step that significantly reduced the complexity of processing the data.

3.5 Data Augmentation through Map Matching

Another major advantage of using SHRP 2 NDS is the data augmentation that had already been performed. Map matching is a process by which roadway properties can be assigned to specific epochs of driving by using GPS location data. It can vary from simple algorithms that find the the nearest roadway for each GPS trace to more complex algorithms that take into account various assumptions about what is physically and legally possible while driving. The map matching algorithm used on SHRP 2 NDS was created in house by VTTI researchers and used sophisticated search functions to find the most appropriate roadways

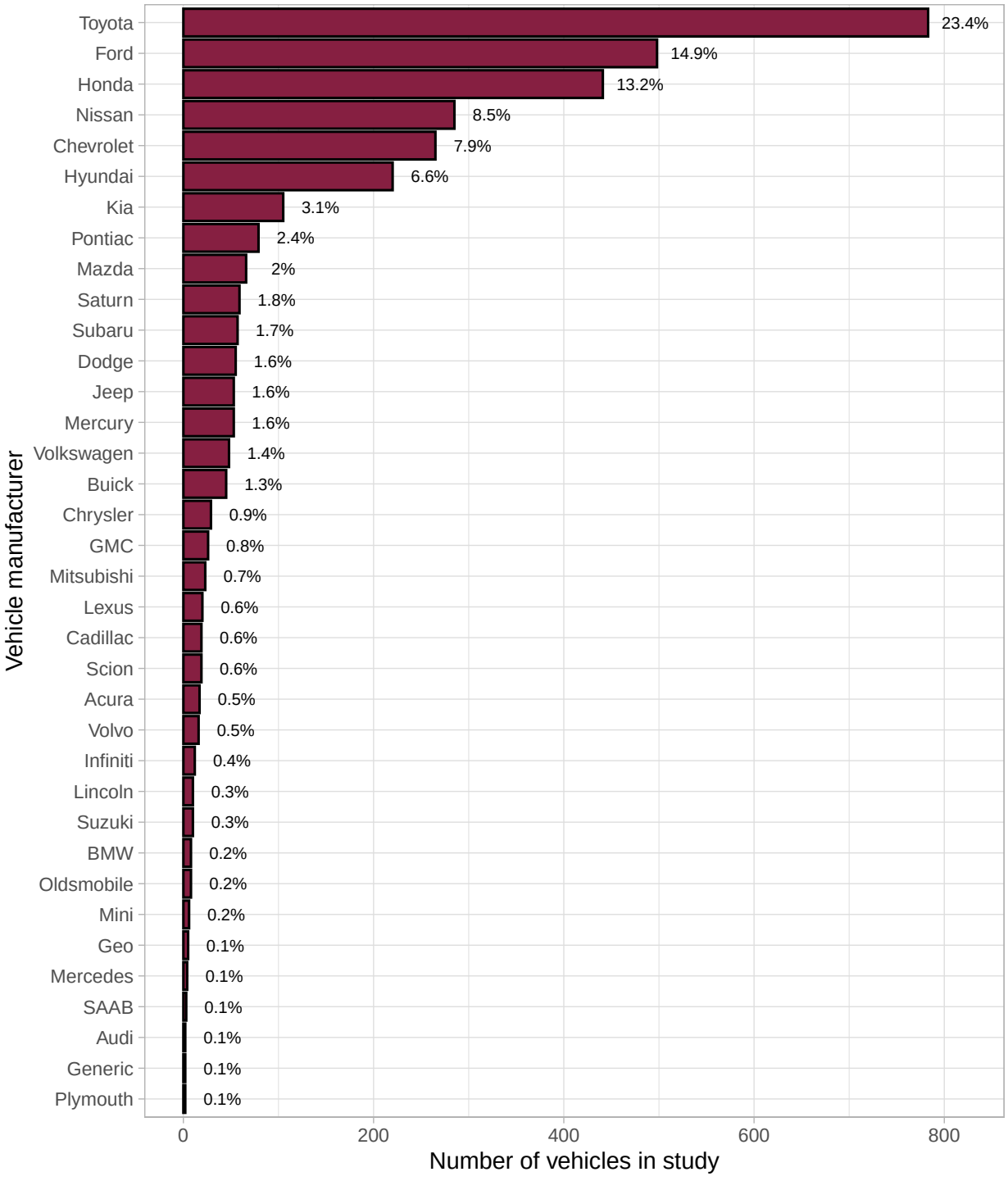
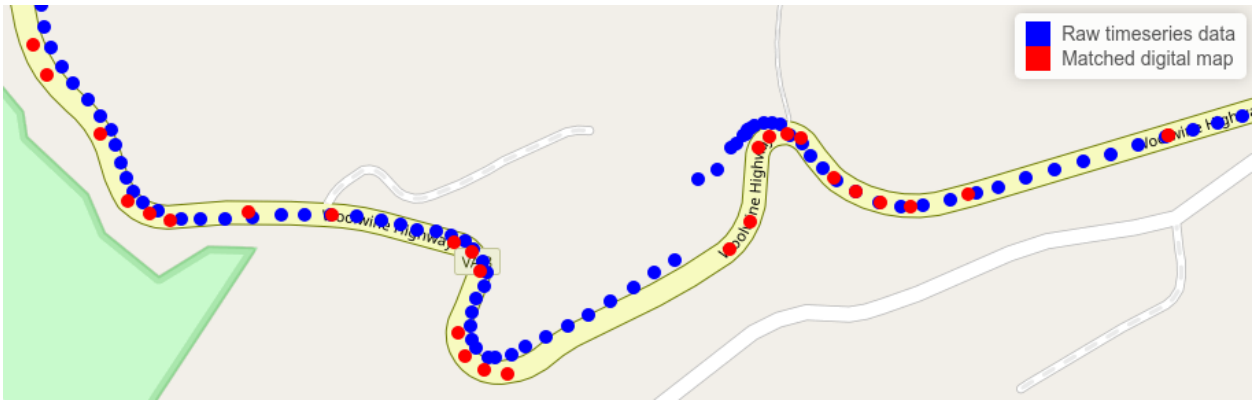
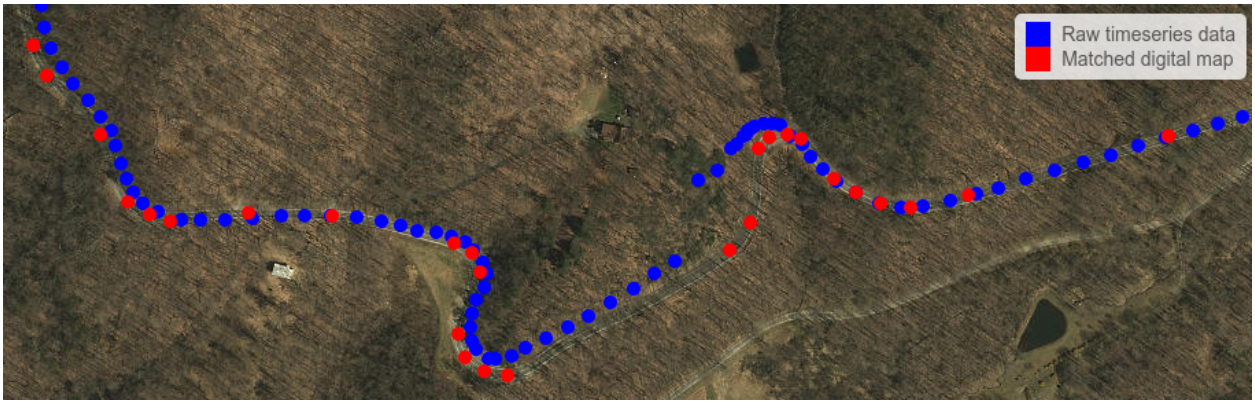


Figure 3.1: Number of vehicles by manufacturer belonging to participants enrolled in SHRP 2 NDS .



(a) Data points overlaid on open street map tiles.



(b) Data points overlaid on Esri satellite map tiles.

Figure 3.2: Comparison of raw time series location data and the corresponding matched digital map data.

from a HERE Technologies digital map database [87]. Figure 3.2 illustrates that even when GPS data quality was suboptimal, the algorithm was able to appropriately assign the correct roadway properties.

3.5.1 Roadway Properties

Several roadway properties are available through the HERE Technologies digital map database. Of these, the ones used in this research are functional class, speed category, speed limit, number of lanes, controlled access status, and ramp status. Table 3.2 lists the various HERE

Table 3.2: Roadway functional class definitions from Here Technologies and their possible FHWA classes[43] [58].

Functional Class	**Here Technologies definition**	**Probable FHWA classification**
FC 1	Allowing for high volume, maximum speed traffic movement	Principal Arterial
FC 2	Allowing for high volume, high-speed traffic movement	Principal Arterial
FC 3	Providing a high volume of traffic movement	Minor Arterial
FC 4	Providing for a high volume of traffic movement at moderate speeds between neighbourhoods	Collector
FC 5	Roads whose volume and traffic movement are below the level of any functional clas	Local

Table 3.3: Roadway speed category definitions from Here Technologies and their possible FHWA classes [43] [59].

Speed Category	**Speed range in km/h**	**Speed range in mph**	**Probable FHWA classification**
SC 1	>130 km/h	>80 mph	Principal Arterial
SC 2	101-130 km/h	65-80 mph	Principal Arterial
SC 3	91-100 km/h	55-64 mph	Principal Arterial / Minor Arterial
SC 4	71-90 km/h	41-54 mph	Minor Arterial
SC 5	50-70 km/h	30-40 mph	Collector
SC 6*	<50 km/h	$\leq$ 30 mph	Local

Technologies functional classes available, the definitions provided on their website, and a probable Federal Highway Administration (FHWA) classification. It should be noted that this translation between Here Technologies subcategories and FHWA classification is not perfect for all roadways. However, since the Here Technologies classification is hard to understand for people unfamiliar with their dataset, this translation has been provided.

Table 3.3 lists the various Here Technologies speed categories, their ranges in km/h and mph, and probable FHWA classification. The speed category is better populated than posted speed limit and is more representative of actual driving speeds on a particular roadway.

3.6 Data Composition

It is essential to understand the composition of the dataset as it helps determine whether the biases in this dataset are consistent with our assumptions of how driving mileage is accumulated across the country. This will help determine when to directly apply the inferences extracted from this data and when to apply appropriate corrections in the form of weighting factors. It will also help us choose appropriate subsets of the data to explore certain phenomenon where confounding factors need to be eliminated.

3.6.1 Driving Composition by Temporal and Demographic Characteristics

Figure 3.3 illustrates the distributions of trips by trip length in the SHRP 2 NDS through a box plot, histogram, and a cumulative distribution. The median trip length was 3.3 miles and 75 percent of the trips were shorter than 7.7 miles. Even though the majority of the trips were short, they contributed to a relatively smaller proportion of the total mileage.

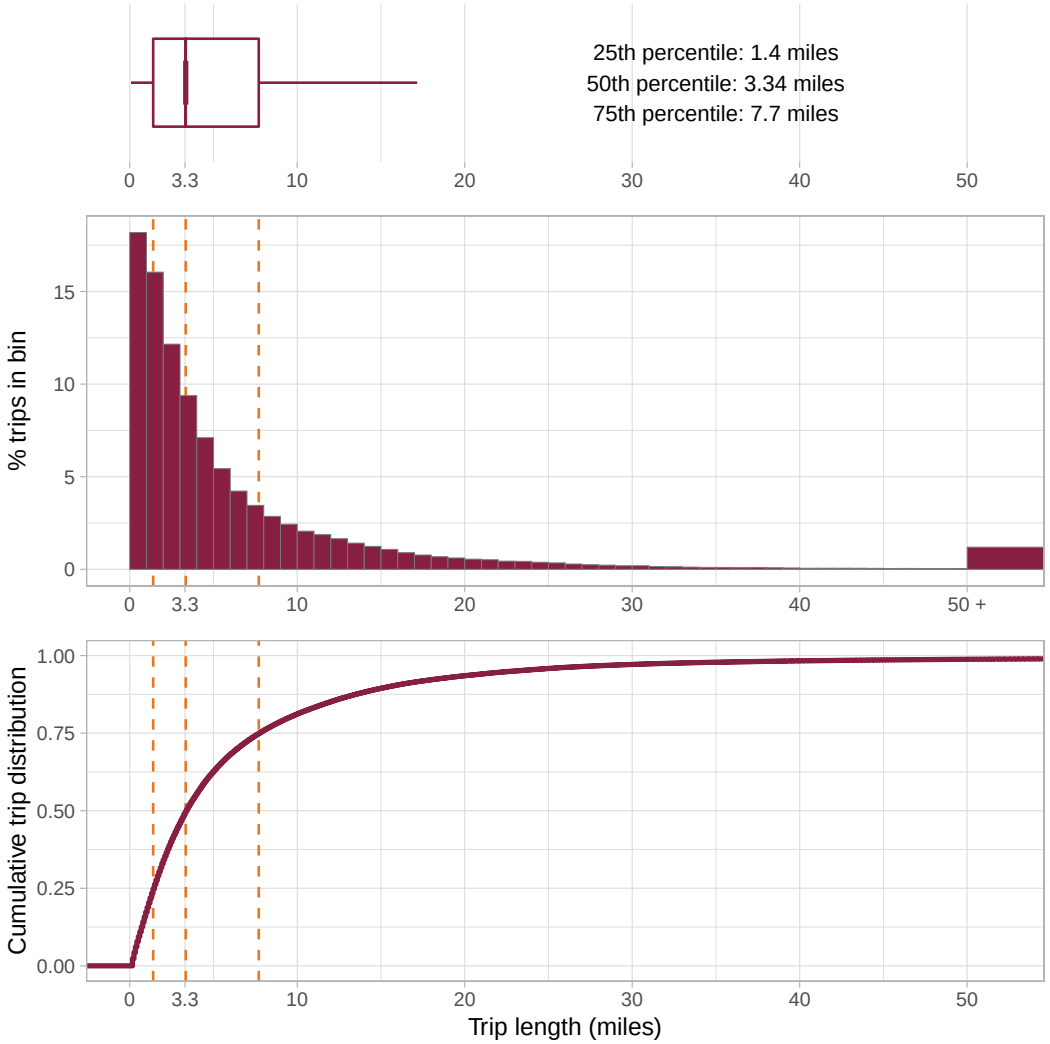


Figure 3.3: Distribution of trips by trip length in SHRP 2 NDS.

Figure 3.4 illustrates the distribution of mileage accumulated by trip length in SHRP 2 through a histogram and a cumulative distribution plot. The histogram has bins of 5 miles each, and all trips longer than 50 miles are grouped in the same bin. As can be clearly seen in the Figures 3.3 and 3.4, longer trips form a higher proportion of the mileage and a smaller proportion of the trip numbers.

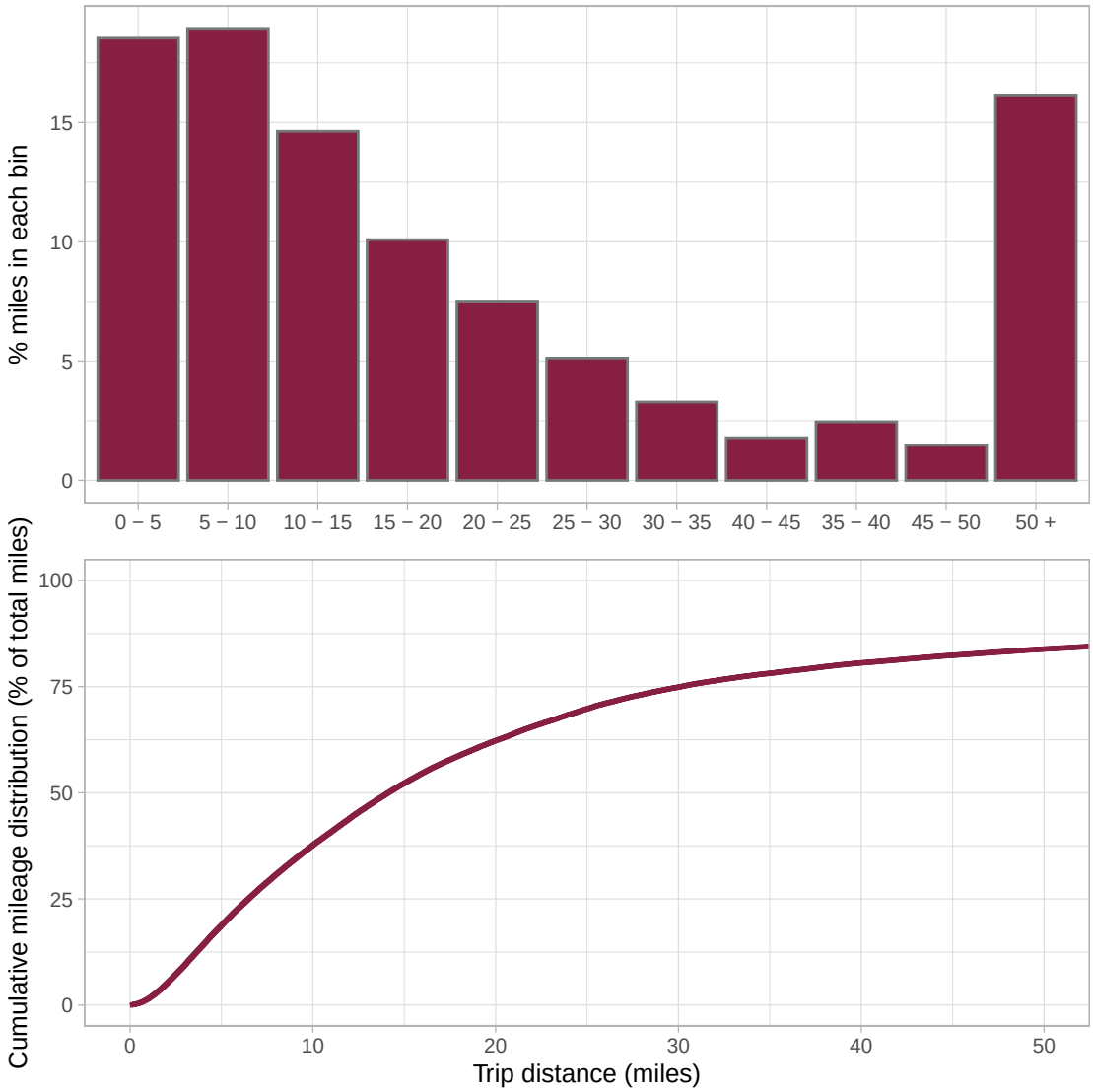


Figure 3.4: Distribution of mileage accumulated by trip length bin in SHRP 2 NDS.

Figure 3.5 illustrates distribution of mileage driven by month and year through a histogram and a cumulative distribution plot. The data collection started in November 2010 and culminated in December 2013 with about 50% of the data having been accumulated by October 2012. It is important to note the time period of the data collection because as the technological landscape of the U.S. light vehicle fleet changes over time, the inferences of this study may need to be adjusted.

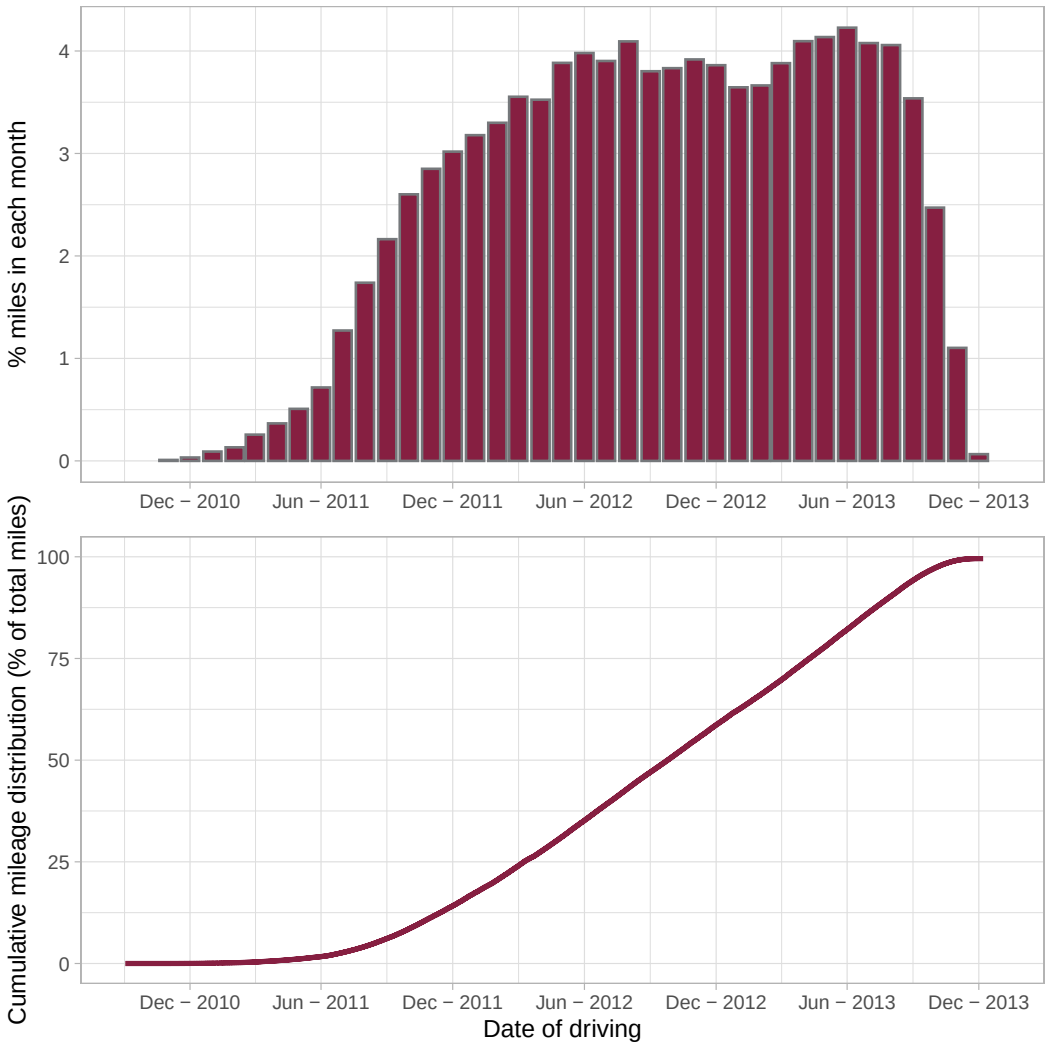


Figure 3.5: Distribution of mileage accumulated during month of data collection in SHRP 2 NDS.

Figure 3.6 describes the distribution of mileage traveled by each participant in the study through a box plot and a histogram. The box plot is overlaid on a "jittered" plot of the individual data points with the y-axis representing random noise to improve readability. As the figure illustrates, the 25th percentile participant drove 4,242 miles, the median participant drove 7,992 miles, and the 75th percentile participant drove 13,133 miles. This figure clearly shows that the mileage accumulation was well distributed across the study and did not come form a small group of participants. The most extreme outliers on the right tail of the distribution drove about 60,000 miles, which is less than 0.2% of the total distribution. On the left side of the distribution, some participants accrued very small mileage but that is mostly because they were secondary drivers in a vehicle that had a different primary driver.

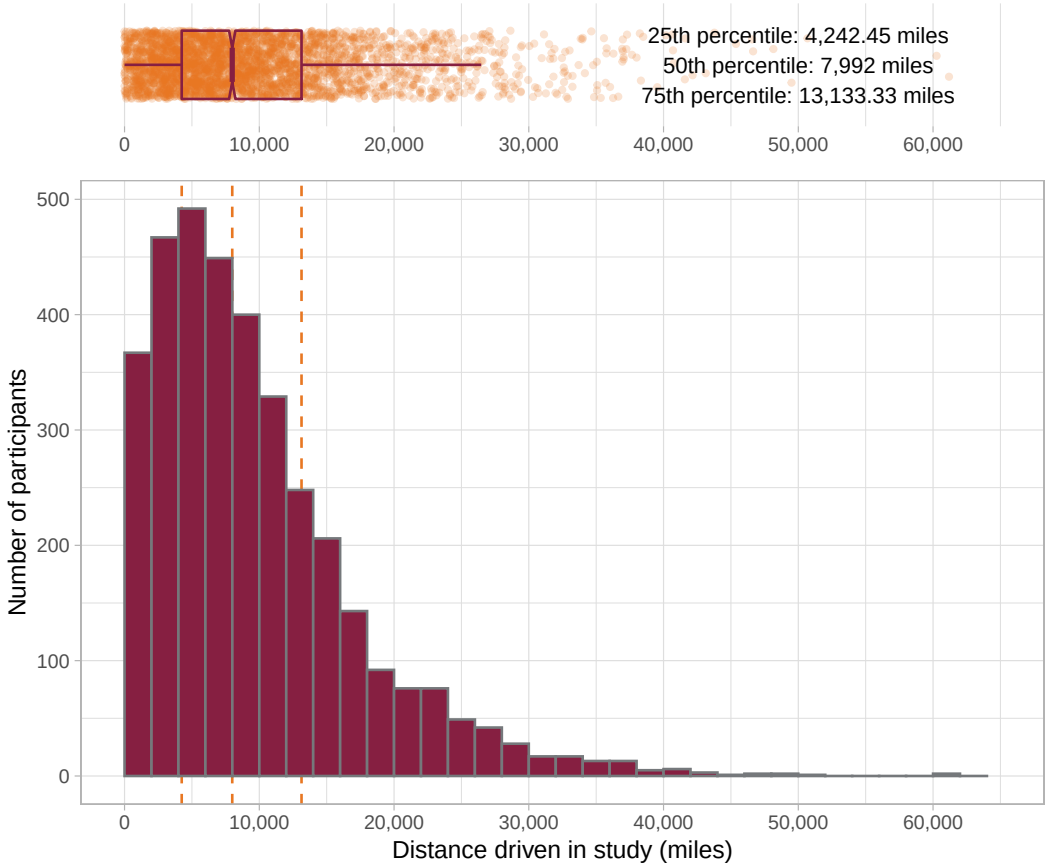


Figure 3.6: Distribution of participants by the total number of miles driven in SHRP 2 NDS.

Figure 3.7 illustrates the distribution of mileage accumulated and participants by age group and gender through a bar plot and a histogram. Both the participants, as well as the mileage, are almost equally distributed by gender between male and female drivers. Younger and older drivers are deliberately over sampled when compared to the national statistics because the study was aimed to examine questions about inexperienced drivers as well as driver impairment due to age. This can easily be corrected by weighting the sample or combining age groups in the middle, as age-related effects are mostly seen at the two ends of the distribution.

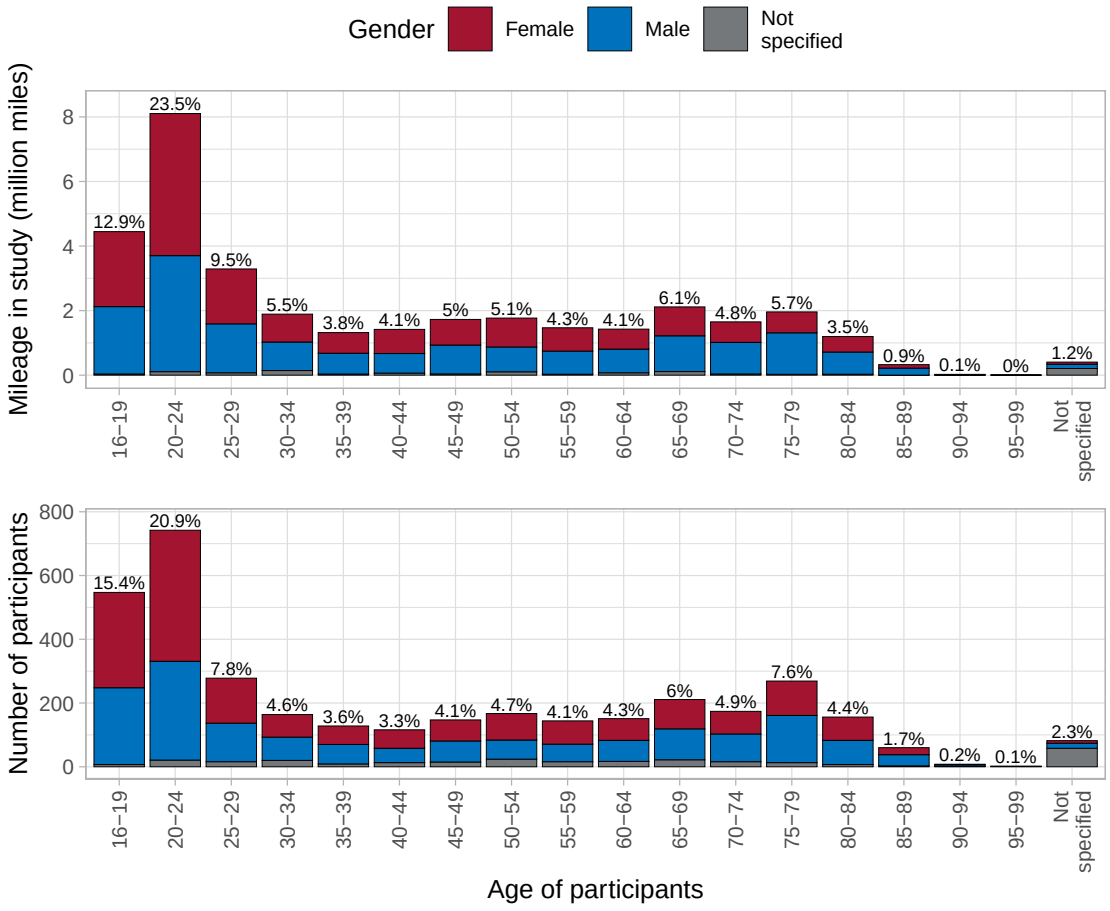


Figure 3.7: Number of participants and mileage driven by age and gender group in SHRP 2 NDS.

Figure 3.8 illustrates the distribution of mileage and participants by location in the study through two bar plots. As the figure shows, Tampa, Florida, Buffalo, New York, and Seattle, Washington, had the largest share of driving, as well as participants, with each location having more than 20% of the data in the study. Raleigh, North Carolina, saw about 18% of the driving, with Bloomington, Indiana, and State College, Pennsylvania, being the smallest collection sites with only 6-8% of driving at each location. Tampa, Buffalo, and Seattle also tended to be more urban compared to the other locations.

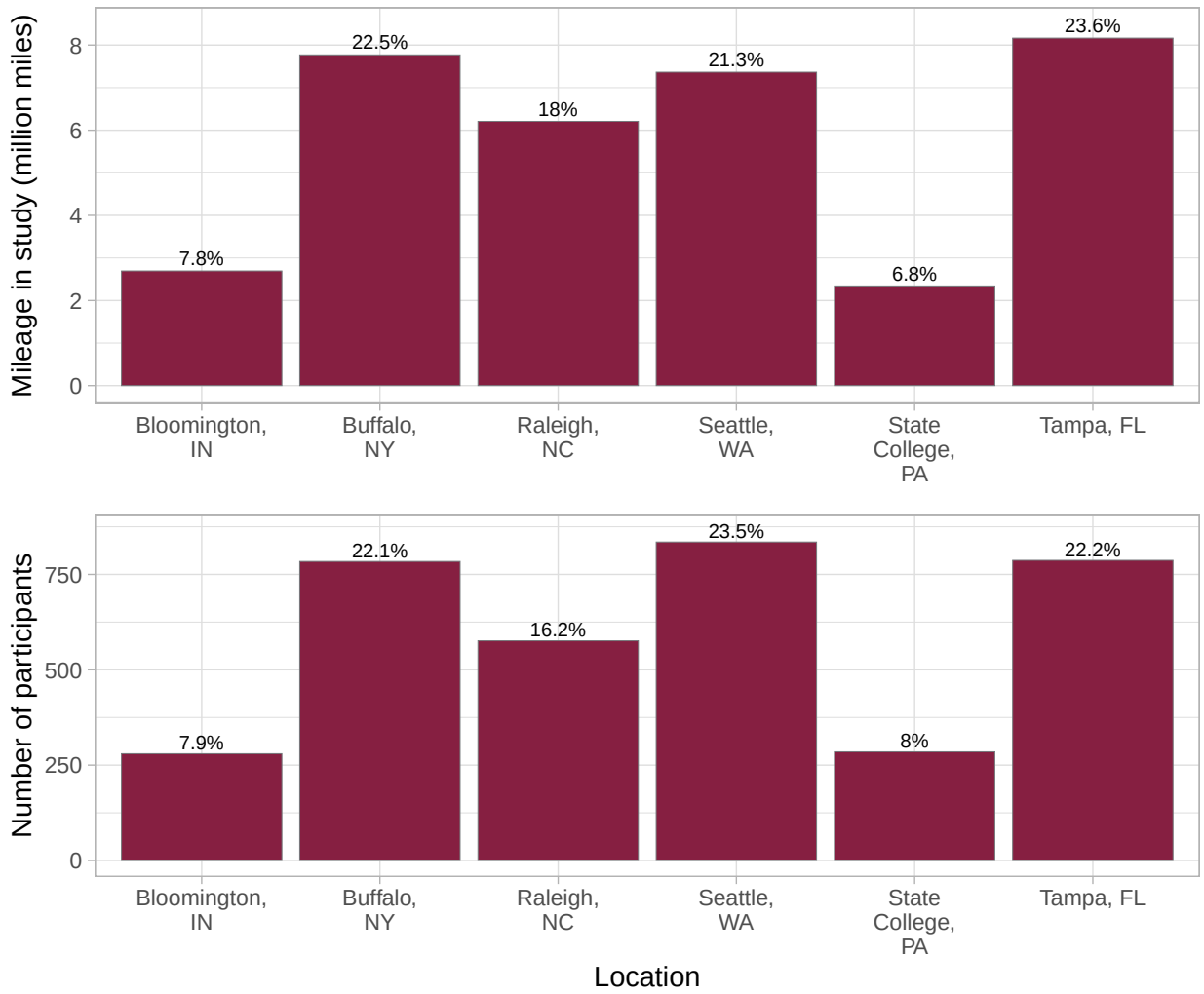


Figure 3.8: Number of participants and mileage driven by location in SHRP 2 NDS.

Figure 3.9 illustrates the distribution of mileage and participants by vehicle class through two bar plots. Cars accounted for about 71% of the mileage as well as participants, followed by SUVs (20%), pickup trucks (5%), and minivans (4%).

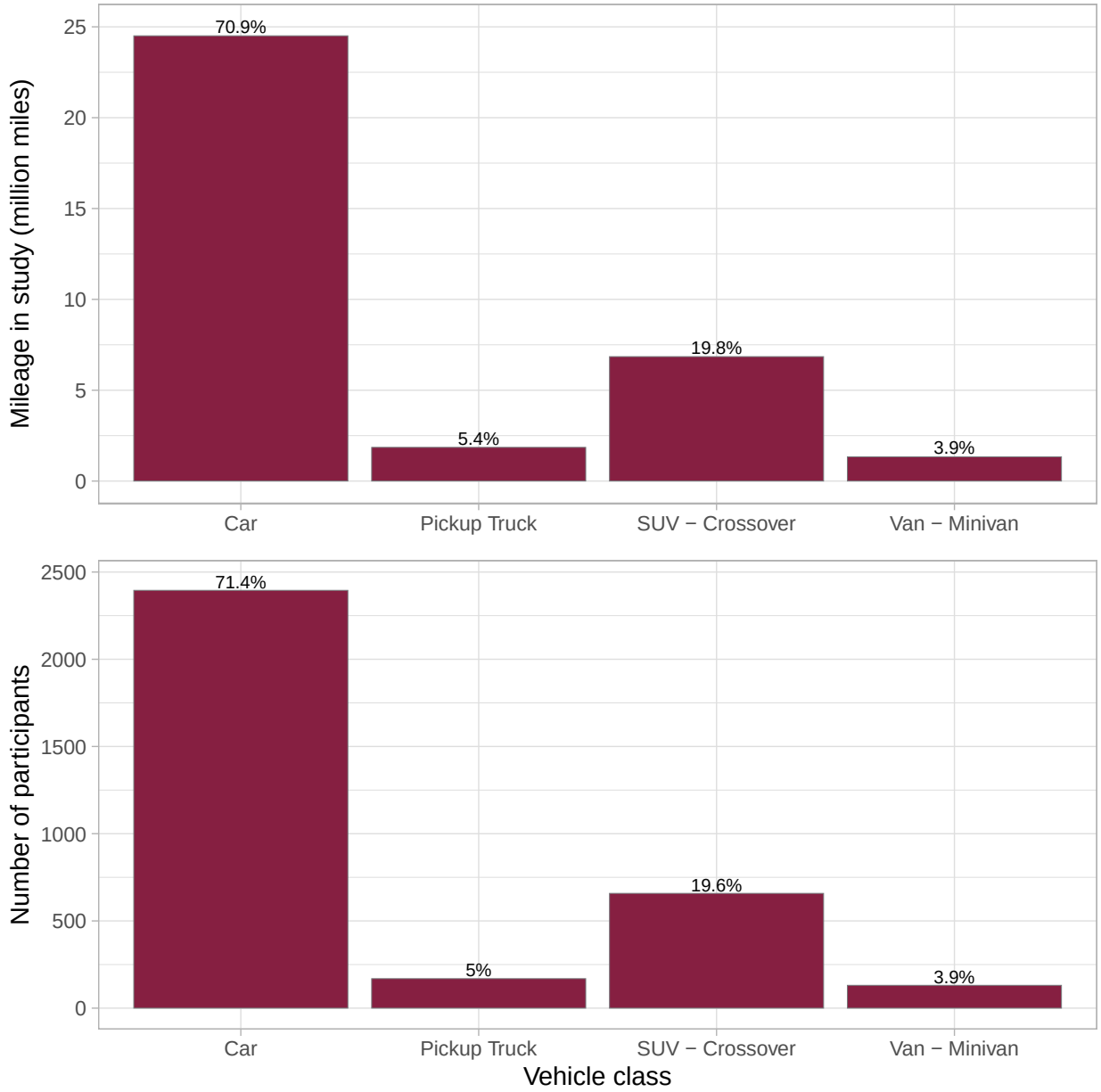


Figure 3.9: Number of participants and mileage driven by vehicle class in SHRP 2 NDS.

Figures 3.10 and 3.11 illustrate the distributions of mileage accumulation by day of week and hour of day. The driving is distributed relatively evenly between the days 3.12 shows the distribution of mileage accumulation by hour of day and day of week. As expected, on week days, there is a slightly higher percentage of mileage accumulated around 8 A.M. and 6 P.M. that is not seen on the weekends. It is important to retain this context during analysis as it may determine how the driving took place.

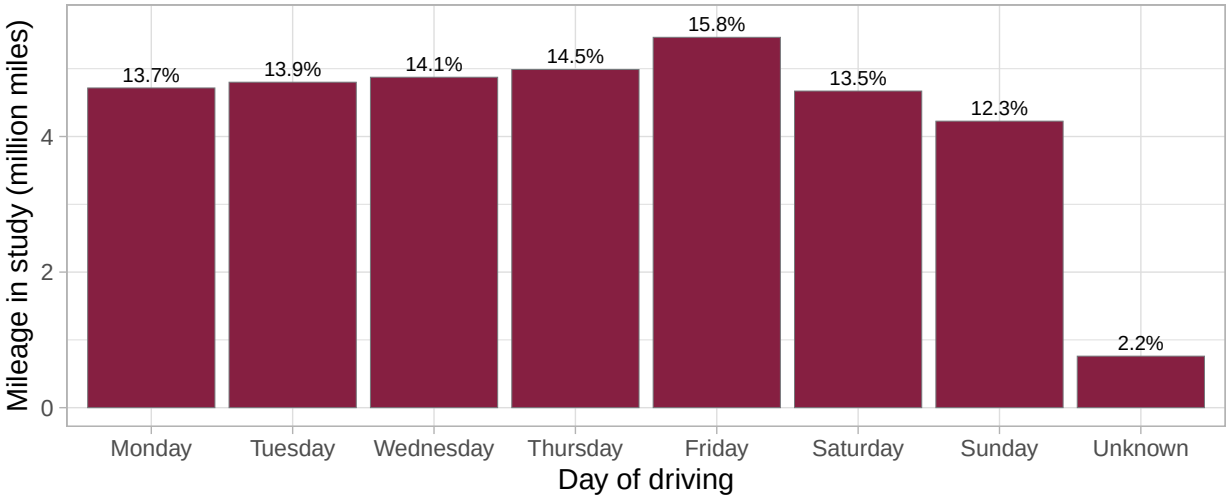


Figure 3.10: Mileage driven by day of week in SHRP 2 NDS.

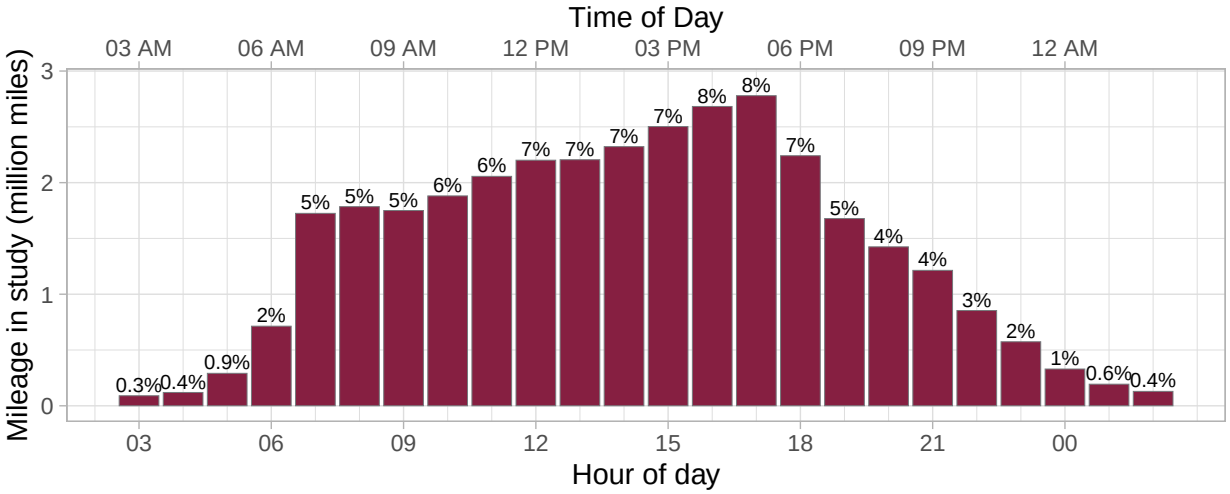


Figure 3.11: Mileage driven by hour of day in SHRP 2 NDS.

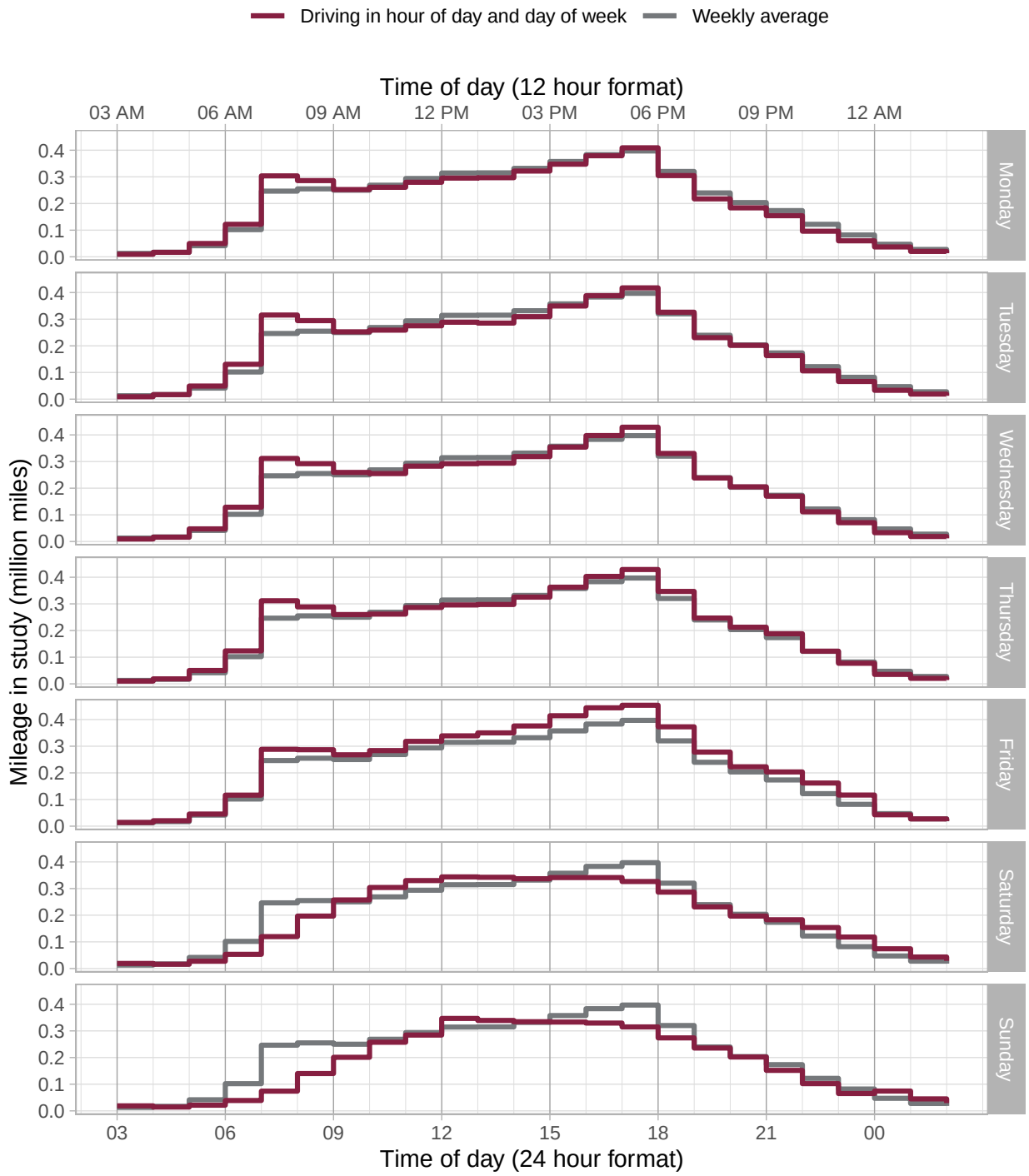


Figure 3.12: Mileage driven by hour of day and day of week in SHRP 2 NDS.

3.6.2 Driving Composition by Roadway Characteristics

The most important environmental factors that affect driver behavior are roadway properties such as functional class, control access, speed category, and speed limit. Figure 3.13 shows the distribution of mileage by functional class. The percentage marked as unknown represents the driving proportion where the map matching algorithm was unable to find roadways due to poor GPS data. This is often due to the GPS module taking up to a few minutes after ignition has been turned on to start and triangulate before it can provide useful information. Figures 3.14 and 3.15 illustrate the distribution of SHPR 2 NDS mileage by roadway speed categoy and roadway speed limit, respectively. The difference between the size of the unknown categories between the speed category and speed limit is due to the speed limit being available for fewer roadways. These figures also show that the distribution of mileage between high-speed roadways and low-speed roadways is fairly even.

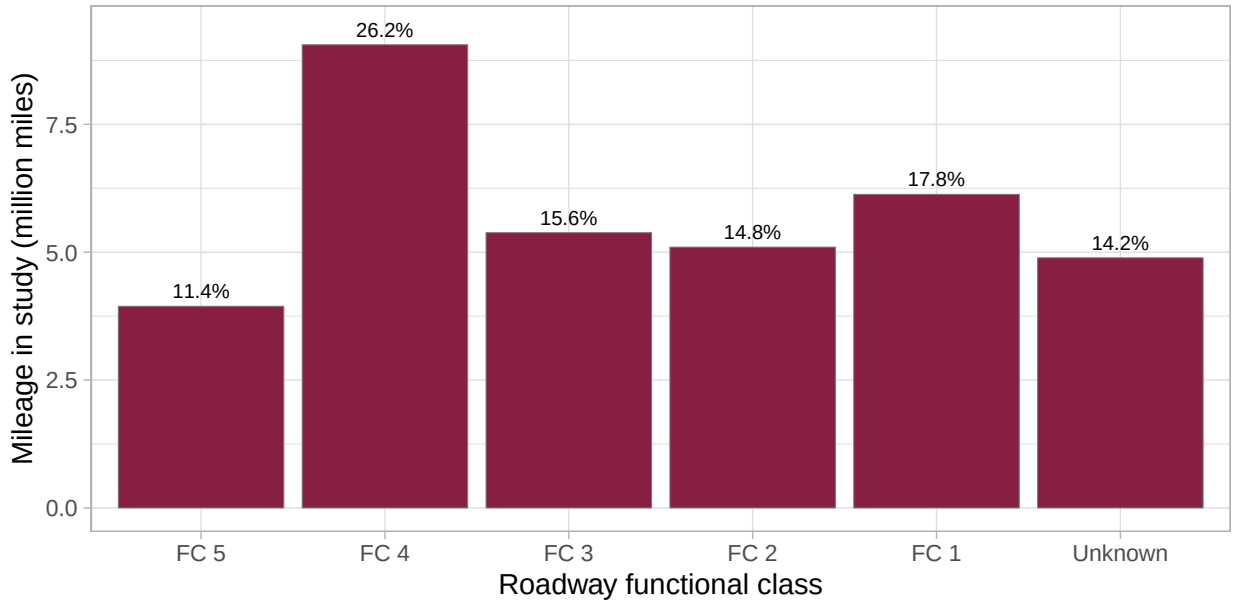


Figure 3.13: Mileage driven by roadway functional class in SHPR 2 NDS.

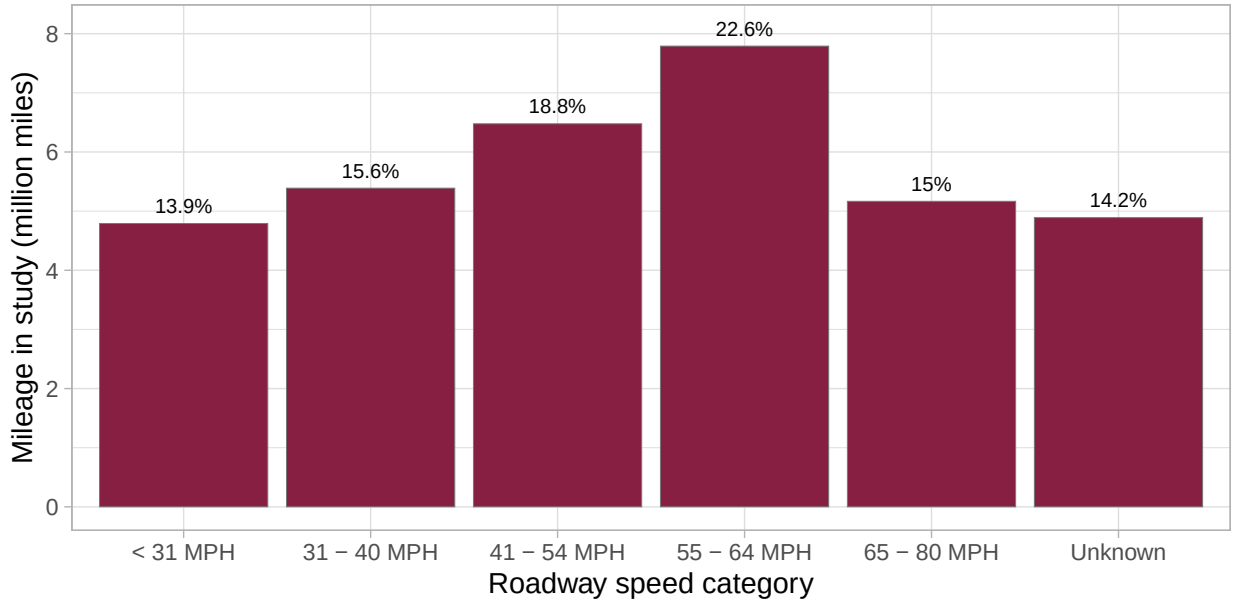


Figure 3.14: Mileage driven by roadway speed category in SHRP 2 NDS.

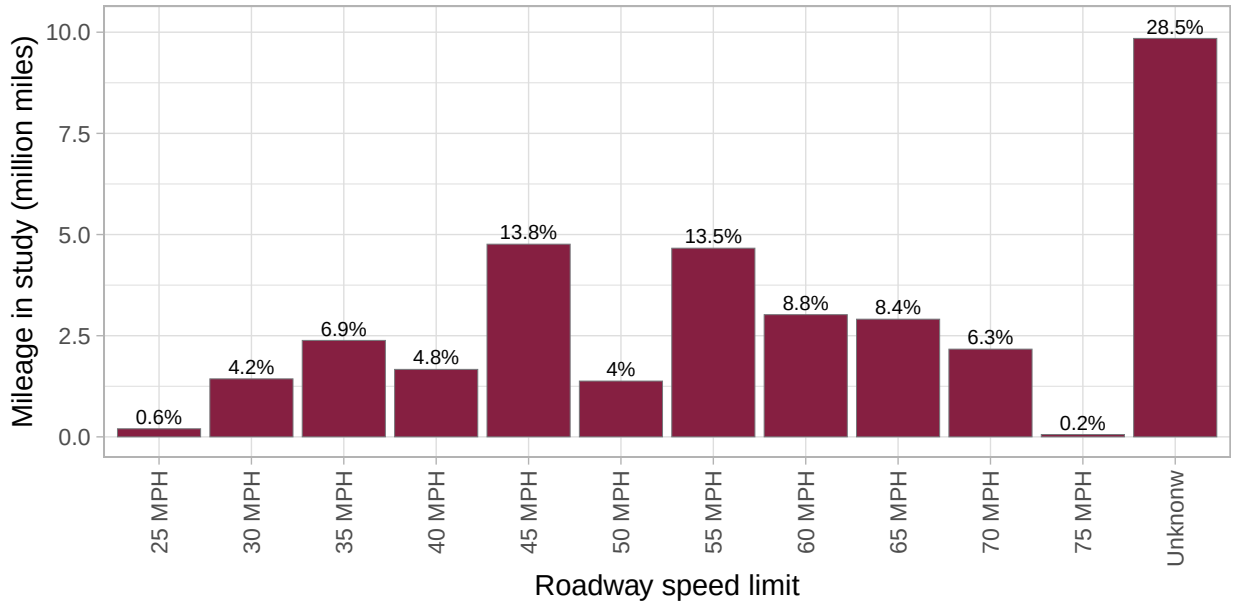


Figure 3.15: Mileage driven by roadway speed limit in SHRP 2 NDS.

Chapter 4

The Surface Accelerations Reference — A Large-scale, Interactive Catalog of Passenger Vehicle Accelerations

Ali, G., McLaughlin, S., & Ahmadian, M. (2022). The Surface Accelerations Reference — A Large-scale, Interactive Catalog of Passenger Vehicle Accelerations. Manuscript under review.

Abstract

There is a need for a large-scale, real world, diverse, and context rich vehicle acceleration catalog that can be used to design, analyze, and compare various intelligent transportation systems. This paper fulfills three primary objectives. First, it provides such a catalog through the Surface Accelerations Reference, which is openly available as an interactive analytics tool as well as an open and downloadable dataset. The Surface Accelerations Reference statistically describes the driving profiles of about 3,500 individuals contributing 34 million miles of continuous driving data collected in the Second Strategic Highway Research Program Naturalistic Driving Study (SHRP 2 NDS). These profiles were created by summarizing billions of longitudinal and lateral acceleration epochs experienced by the participants. Second, this

paper introduces a standardized methodology for creating such a catalog so that similar acceleration profiles can be produced for other human cohorts or automated driving systems. Finally, the data are used to analyze the effect of roadway speed category on the rates of lateral and longitudinal acceleration epochs at various thresholds. It is observed that, for the median driver, the rates of epochs are upto 3 orders of magnitude higher on low-speed roads as compared to high-speed roads. This catalog will facilitate intelligent vehicle system designers to compare and tune their systems for safer driving experiences. It will also allow agencies with similar data to create comparable catalogs facilitating safety and behavioral comparisons between populations.

4.1 Introduction

The purpose of this project was to catalog longitudinal and lateral accelerations experienced in passenger vehicles and create an easily accessible dataset for users from a variety of fields. The Second Strategic Highway Research Program Naturalistic Driving Study (SHRP 2 NDS), the largest collection of naturalistic driving data to date, is an ideal data set to create such a catalog, with more than 34 million miles of data collected from a diverse set of participants [35]. The SHRP 2 NDS dataset was analyzed to create the Surface Accelerations Reference, which describes the acceleration norms and distributions of the 3,500 participants in the study.

This catalog of accelerations experienced in passenger vehicles will be valuable for users in a number of fields [21, 72]. For example, considerable work has been done to model behavior, determine driving style and detect anomalous driving events using vehicle accelerations and other driving parameters [23, 24, 30, 31, 49, 60, 73, 81, 83, 90, 107, 110, 119, 134]. However, such research has mostly relied on small studies, nonrepresentative experiments, or test track

data. The surface accelerations reference provides a dataset that is sufficient to describe thousands of individuals, support exploration of numerous groupings of individual styles, and quantify the effect of a broad range of roadway environments, traffic conditions, and vehicle classes.

Advanced driver assistance systems and automated driving systems will need to operate within the expectations of passengers as well as surrounding road users. Automated systems that assist drivers or fully autonomous vehicles will likely benefit from control that feels comfortable and natural [148]. These systems also rely on predicting the behavior of other road users to keep occupants safe. Both developers and policy makers will benefit from objective understanding of the probability of acceleration behavior in a given context and the driving behavior of different road users.

In addition, roadway engineers consider vehicle kinematic parameters, such as longitudinal and lateral accelerations, speed, etc. to design roadways [1]. Often, such parameters come from models that were developed decades ago, in constrained experimental conditions, and which do not reflect current vehicle specifications [42]. Sometimes, such models are purely based on vehicle characteristics and do not reflect driver preferences or usage patterns [78]. The Surface Acceleration Reference, which represents the driving patterns of a diverse driver demographics and vehicle classes in a broad range of conditions, will therefore also be beneficial to roadway planners and departments of transportation.

Driver behavior research, which has traditionally relied on using adverse driving events to identify and evaluate risky driving behavior [39, 86, 93], can also benefit from establishing kinematics-based surrogate measures to classify risky driving[26, 27, 49]. Doing so would allow the prevention of safety critical events by identifying at-risk drivers and providing adequate driver training [94, 114]. Driver behavior researchers are also interested in categorizing various driving styles, such as sporty, aggressive, conservative, etc., and determining their

relationship with risky driving. The Surface Acceleration Reference, based on SHRP 2 data, which also includes driver crash involvement, would greatly benefit such researchers, as a driver's kinematic profile could be easily linked to their crash rates.

Many insurance companies have started pay-as-you-drive programs that incentivize good driving through better insurance rates [52, 76, 133, 146]. The Surface Acceleration Reference will also be useful to companies behind such programs, as it would let them compare the kinematic profiles of their customers with a dataset that includes data from a broader sensor suite as well as a range of demographic measures and risk analyses. This catalog also has the potential to serve as a standard within the industry, making the process transparent and comprehensible for customers. Practitioners from other fields, such as traffic modeling, crash reconstruction, ISO standard development, etc., can also greatly benefit from this comprehensive acceleration catalog [4, 130, 138].

A number of studies have been conducted on understanding the acceleration preferences of drivers and passengers [3, 45, 61, 77, 118, 132]. These studies provide valuable insights but have not been adopted as a cross-discipline reference on accelerations for various reasons, such as subjective interpretation of surveys; datasets being small, non-representative, or outdated; or lack of standard methods. These shortcomings had to be taken into consideration when developing the Surface Acceleration Reference.

To create a comprehensive acceleration catalog that can serve as an industry standard across various research fields, five major challenges need to be overcome. These challenges can be characterized into data scale, algorithm standardization, data correction, result interpretability, and output accessibility. The Surface Accelerations Reference addresses all these challenges and therefore is an ideal candidate to be used as a standard of comparison across various fields.

First, the underlying dataset needed to generate such a catalog must be diverse, large, and contain high quality data. The SHRP 2 dataset was designed to be such a dataset. More than 3,500 participants were chosen from six locations across the country and were representative of all ages, genders, ethnicities, races, levels of education, and socioeconomic statuses [35]. Though there are differences between the SHRP 2 sample population and U.S. national statistics, SHRP 2 data are fairly representative of the nation's population [12]. The SHRP 2 dataset is also the largest collection of naturalistic driving data, with more than 34 million miles of recorded data, including recordings from four video feeds (front, rear, and two in cabin views); various vehicle network variables such as speed, gas pedal input, brake state, etc.; location data in the form of GPS coordinates; and kinematic measures from an inertial measurement unit (IMU). A major advantage of such a dataset is that video feeds often provide additional context that helps validate outliers and detect malfunctioning sensors. The SHRP 2 dataset also has the advantage of containing recordings of more natural behaviors than datasets from studies where participants self-reported, drove with a researcher, or were enrolled for a shorter period of time. The lack of a standard acceleration catalog to date was in part due to the lack of a dataset that fulfilled all the requirements.

Second, for an acceleration catalog to be used across disciplines, there is a need for standardization of event definitions as well as the definition of rates. As various fields use different measures to compare event rarity, it is important to develop standard measures that can be used against other existing measures. For this purpose, an epoch definition, explained in the methodology section, was created and documented for a standardized acceleration epoch.

Third, collection over a long period of time is necessary to answer questions of interest related to accelerations, but IMU data are prone to noise and sources of error such as bias and drift. The process used to generate an acceleration catalog should recognize such errors

and incorporate processes to isolate problematic data. On a related note, the scale of data processing requires considerable computing resources as well as efficient algorithms.

Fourth, the output of such a catalog should be usable by practitioners from various fields. This poses multiple challenges as various disciplines are inclined to using different tools and rating scales in their analyses. To further ensure the Surface Acceleration Reference's broad utility, the acceleration epoch identification, summarization algorithms, and methods for aggregating accelerations and generating participant driving profiles are well documented in this paper to ensure reproducibility. Finally, a consolidated acceleration profile table is available to practitioners from various fields through a downloadable CSV file as well as an interactive visualization. The interactive visualization can be accessed through any web browser and allows users to compare various participant groups based on age, location, gender, and vehicle class.

4.2 Methodology

The Surface Accelerations Reference consists of an acceleration profile for each driver (or driver-vehicle combination if a participant drove more than one vehicle) created by analyzing their entire driving history in the SHRP 2 dataset. Figure 4.1 shows the data flow for creation of the Surface Accelerations Reference. The original sensor-based time series data was read from a database and analyzed by an algorithm that identified acceleration epochs. Each acceleration epoch was summarized as a multi feature data point and written to a database table. After the acceleration analysis algorithm was run on the entire SHRP 2 dataset, a profile generation algorithm aggregated all summarized accelerations in a participant's history and created an acceleration profile. The acceleration profile contains about 400 data points for each of the 3,670 driver-vehicle combinations and can be stored in either a

database or a downloadable CSV file. These acceleration profiles are made available as a downloadable CSV file as well as an interactive visualization through a web interface that lets users from a wide variety of fields parse data to answer questions relevant to their area of interest.

The creation of the Surface Acceleration Reference consisted of different processes that can be categorized into three stages: (1) identifying and summarizing acceleration epochs from time series data, (2) creating acceleration profiles from summarized epochs, and (3) generating an interactive visualization comparing these acceleration profiles.

4.2.1 Identifying and Summarizing Acceleration Epochs

An algorithm was developed to read time series data from the database, identify epochs of longitudinal and lateral acceleration, summarize them, and write the summary to a table. The algorithm analyzed 10 variables, described in Table 4.1, for every trip in the SHRP 2 dataset. The summarized epochs were linked to the original file and its associated metadata using a unique identifier and timestamps. Metadata associated with each trip included in the Surface Accelerations Reference is listed in Table 4.2.

To efficiently process the vast amount of data, a Db2 database cluster and a computing cluster were used to run 384 instances of the algorithm in parallel. The algorithm first ingested all the variables from the Db2 database cluster. The signals were then cleaned and preprocessed by removing invalid values, deleting repeating timestamps, and clearing artifacts. As Table 1 shows, the original variables had different sampling rates. To make the algorithm more efficient, all variables were interpolated to align on a common timestamp series. After the variables were standardized as described above, longitudinal acceleration, lateral acceleration, and yaw rate were analyzed to find epochs. Having a standardized

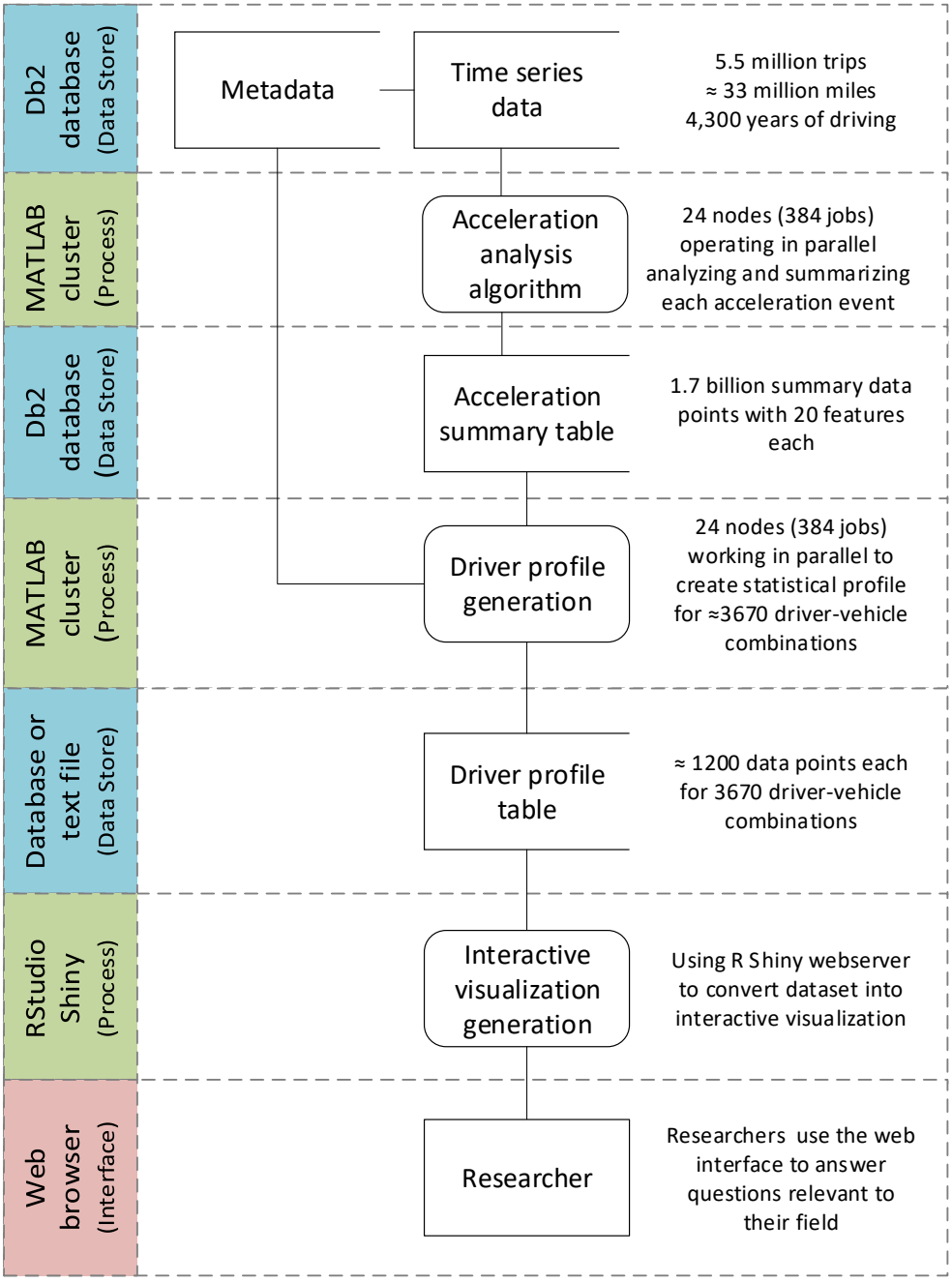


Figure 4.1: Data flow for creation of the Surface Accelerations Reference.

method to identify epochs that could be replicated on other studies was vital to the success of this project. An acceleration event or epoch is the duration for which the absolute value

of a signal is continuously greater than a threshold value. For longitudinal and lateral accelerations, this value is ± 0.01(g) and for yaw rate the threshold value is ± 1 deg/sec. The threshold values were chosen to ensure that fluctuations around the noise floor were not considered events, while still including cases that were quite low in amplitude [95]. Figure 4.2 shows the longitudinal acceleration and speed signals and illustrate the selection of acceleration and deceleration epochs and their peaks.

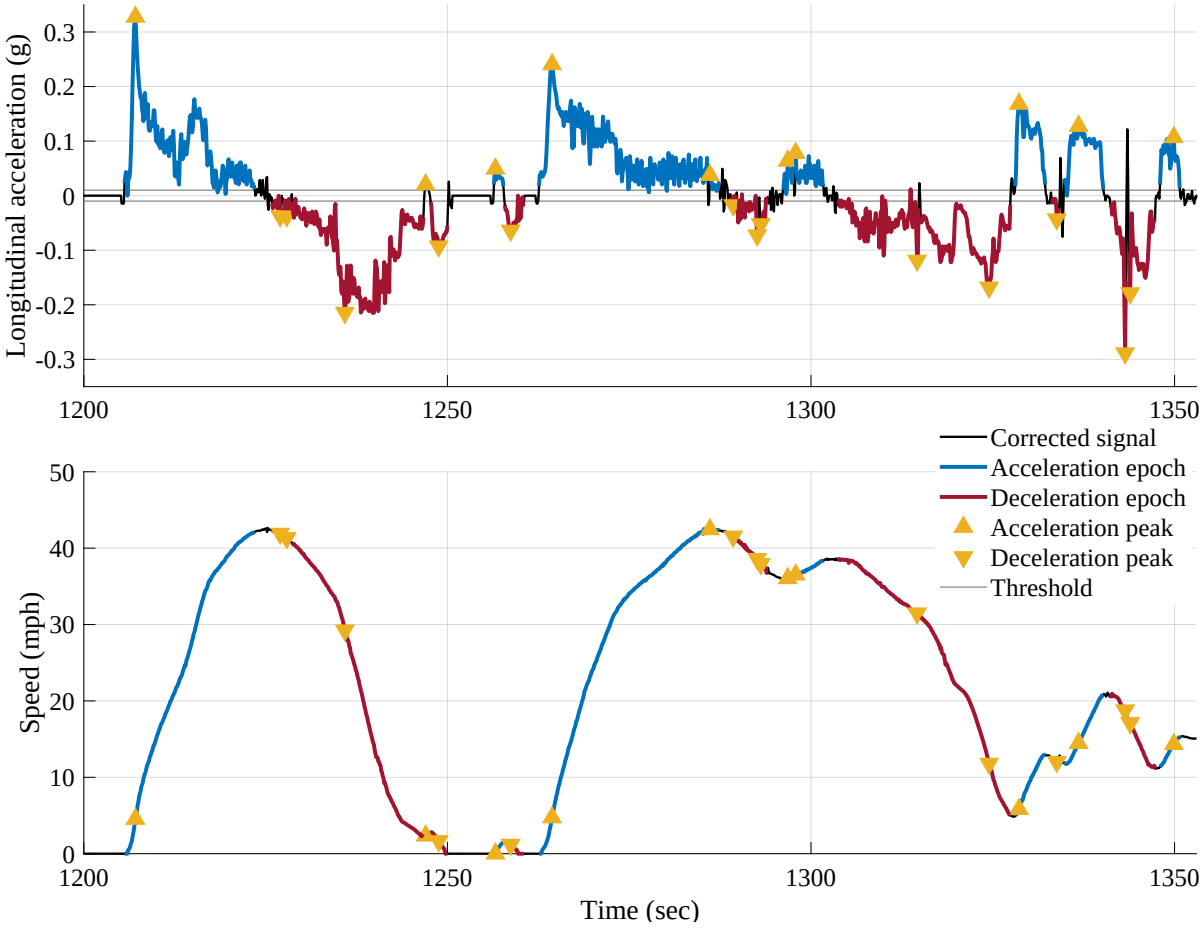


Figure 4.2: Identifying acceleration and deceleration epochs in longitudinal acceleration and speed time series data.

One of the challenges of analyzing acceleration signals is that when a vehicle is stopped on

a slope, the accelerometer output shows a constant non zero value. According to the epoch defintion, this would constitute an acceleration or decelration epoch. To avoid these false positives, a method was developed where the acceleration signal was made equal to zero if the speed of the vehicle was zero and if the moving standard deviation of the acceleration signal was below 0.01. This ensured that these potential false positives would not be identified as epochs and at the same time that the start and end of the acceleration epochs would be set appropriately given the speed signal lag relative to the acceleration signal. A comparison between the orignal and the modified acceleration signals is shown in Figure 4. The effect of road grade on the absolute value of acceleration was not taken into account as it was assumed the the vehicle would be going up and down grades an equal number of times, causing the change in absolute acceleration values to cancel each other out.

A similar approach was used to identify epochs of lateral acceleration and yaw rate. Once the epochs were identified, the algorithm summarized each acceleration epoch and generated statistical measures described in Table 3. These measures fall into three categories,

1. **Trip properties**, such as unique trip ID and timestamp within trip. These variables are useful to link each epoch to other metadata, such as driver, vehicle, road type on which the epoch happened, etc., associated with the trip.

2. **Epoch properties**, such as maximum value of signal in epoch, average value, distance traveled, etc. These summary variables provide addition details about the data as well as a means of qualifying epochs.

3. **Variables describing driver inputs**, such as the duration of brake pedal being pressed. These are useful for filtering and qualifying epochs based on driver inputs.

4. **Variables describing roadway characteristics**, such as road functional class and speed category [58, 59, 87]. These are valuable for contextualizing each event based

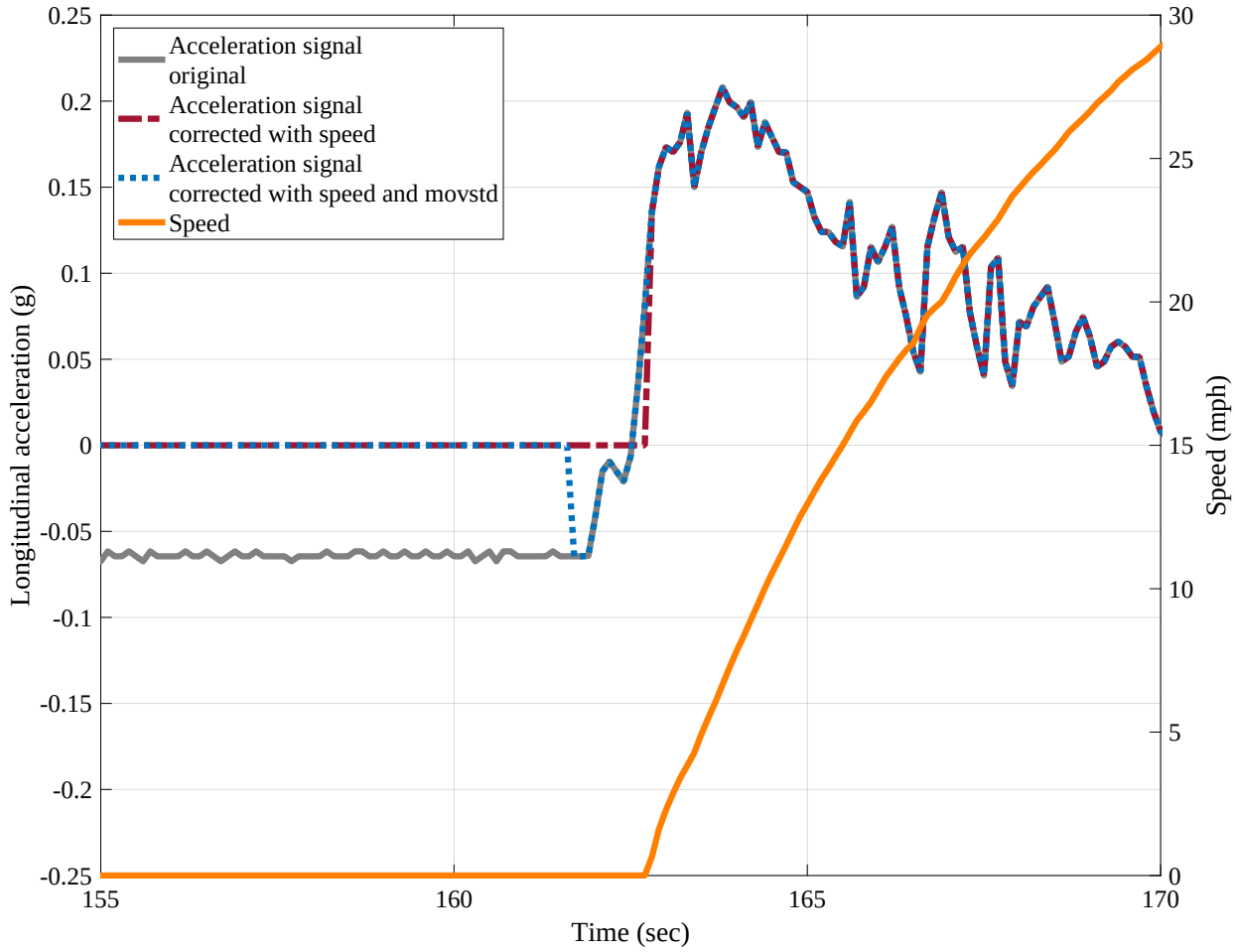


Figure 4.3: Comparing acceleration signal correction using speed and moving standard deviation.

on the road type.

Similar tables were created for lateral acceleration and yaw rates. A total of 1.7 billion summarized points were created and stored in a relational database. The table structure not only facilitates the creation of driver acceleration profiles but can also be used for contextualizing driving data in the SHRP 2 database using machine learning algorithms.

4.2.2 Creating Driver Acceleration Profiles from Summarized Epochs

Once all the acceleration epochs were identified, summarized, and recorded in the acceleration summary tables, a profile generation algorithm aggregated all epochs belonging to a driver-vehicle combination, creating a statistical profile. Such a profile was created for each of the following categories:

- Acceleration epochs
- Deceleration epochs
- Positive lateral acceleration epochs (vehicle moving towards right)
- Negative lateral acceleration epochs (vehicle moving towards left)

Table 4.1: A description of the variables used to create the Surface Accelerations Reference.

Variable	**Variable description**	**Source**	**Frequency**
accel_x	Longitudinal acceleration	DAS IMU	10 Hz
accel_y	Lateral acceleration	DAS IMU	10 Hz
accel_z	Vertical acceleration	DAS IMU	10 Hz
gyro_x	Roll rate	DAS IMU	10 Hz
gyro_y	Pitch rate	DAS IMU	10 Hz
gyro_z	Yaw rate	DAS IMU	10 Hz
speed_network	Vehicle speed	Vehicle CAN	Variable (1-100 Hz)
speed_gps	Vehicle speed	DAS GPS	1 Hz
pedal_brake_state	Brake pedal state	Vehicle CAN	State change timestamps
pedal_gas_position	Gas pedal position	Vehicle CAN	Variable (1-100 Hz)

Table 4.2: Metadata associated with each trip in the SHRP 2 dataset used in Surface Accelerations Reference.

Metadata type	**Metadata description**	**Categories**
Location	The state in which the vehicle was instrumented	New York Florida Washington North Carolina Pennsylvania Indiana
Age group	The age group of the participant	16-19 20-24 25-29 ⋮ 95-99
Gender	The gender of the participant	Male Female Did not specify
Vehicle class	The type of the vehicle used by the participant	Car Pickup SUV/Crossover Minivan

Table 4.3: List of summary variables calculated for each acceleration epoch identified within a trip.

Summary Variable	**Description**
FILE_ID	Trip ID that uniquely identifies each trip in SHRP 2 dataset
TIME_START	Within trip start time of acceleration epoch
TIME_END	Within trip end time of acceleration epoch
ACCELERATION_X_MAX	Maximum acceleration value in epoch
ACCELERATION_X_MEAN	Mean acceleration value in epoch
ACCELERATION_X_MEDIAN	Median acceleration value in epoch
SPEED_START	Speed at start of epoch
SPEED_END	Speed at end of epoch
DISTANCE	Distance travelled in epoch
DURATION	Duration of epoch
SPEED_CHANGE	Difference in speed between end and start of epoch

ACCELERATION_SPEED_MEAN	Mean acceleration calculated from change in speed and duration
TIME_PEAK_ACCELERATION	Timestamp of peak acceleration during epoch
ACCELERATION_TO_PEAK_MEAN	Mean acceleration from start to peak of epoch
ACCELERATION_FROM_PEAK_MEAN	Mean acceleration from peak to end of epoch
SPEED_AT_PEAK_ACCEL	Speed at peak acceleration
SPEED_MAX	Maximum speed during the epoch
GAS_PEDAL_MEAN	Mean of gas pedal signal during the epoch
GAS_PEDAL_MEDIAN	Median of gas pedal signal during the epoch
GAS_PEDAL_MAX	Maximum of gas pedal signal during the epoch
GAS_PEDAL_MIN	Minimum of gas pedal signal during the epoch
GAS_PEDAL_STDEV	Standard deviation of gas pedal signal during the epoch

BRAKE	Whether brake was used during epoch
BRAKE_DURATION	Duration for which brake was used in epoch
FUNCTIONAL_CLASS	Functional class of the roadway at the time of peak acceleration
SPEED_CAT	Speed category of the roadway at the time of peak acceleration

Three methods were used to compute statistical measures for each category to allow comparisons of various participants. Even though all three methods used the ACCELERATION_X_MAX (or equivalent measures for lateral acceleration), other measures mentioned in Table 4.3, such as ACCELERATION_X_MEAN, could also be used. These methods are described in further detail in the following sections.

4.2.2.1 Comparing Percentiles

Using the percentile method, measures were computed at the 50^{th}, 60^{th}, 70^{th}, 75^{th}, 80^{th}, 90^{th}, 95^{th}, 99^{th}, and the 99.9^{th} percentile. All epochs linked to a driver-vehicle combination were queried and the percentile values listed above were calculated for the maximum acceleration within the epoch. The percentiles allow users to compare frequent as well as rare epochs. For example, frequent epochs, such as 50^{th} percentile (the stronger of two random epochs) to 90^{th} percentile (the strongest of 10 random epochs), may indicate driving style, whereas rare epochs, such as 99^{th} or 99.9^{th} percentile, would indicate an atypical situation, either for the involved driver or nearby roadway users. This latter category may be indicative of

potential safety critical events. In other words, the percentile measure relates the magnitude of the epoch and its rarity for each participant. Therefore, comparing drivers on the same percentile shows the difference in magnitude for equally rare epochs. An example finding would be that for one driver or group of drivers, an 80^{th} percentile epoch would be 0.2 g deceleration whereas for another driver or group of drivers, the 80^{th} percentile epoch would be 0.3 g deceleration. The latter driver/group of drivers would execute hard decelerations much more often, though this metric cannot reveal how often per mile the behavior occurred.

4.2.2.2 Comparing Rate of Epochs Stronger Than a Threshold Magnitude

This method measured the rate of epochs stronger than a certain threshold per mile. The following rates were calculated:

- Number of epochs stronger than 0.1g per mile/kilometer
- Number of epochs stronger than 0.2g per mile/kilometer
- Number of epochs stronger than 0.3g per mile/kilometer
- Number of epochs stronger than 0.4g per mile/kilometer
- Number of epochs stronger than 0.5g per mile/kilometer
- Number of epochs stronger than 0.6g per mile/kilometer
- Number of epochs stronger than 0.7g per mile/kilometer
- Number of epochs stronger than 0.8g per mile/kilometer
- Number of epochs stronger than 0.9g per mile/kilometer

These metrics were calculated using Equation 4.1

$$R_X = \frac{N_X}{D_T} \tag{4.1}$$

where R_x is the rate of epochs stronger than X(g) per mile, N_X is the number of epochs stronger than X(g), and D_T is the total distance traveled by the participant. The different thresholds used to calculate these metrics differentiate epochs based on their severity. This measure compares the frequency of epochs stronger than a particular threshold among various drivers. The rates for different thresholds can have different implications. For example, for lower thresholds, the rates would imply a higher number of epochs per mile that can be caused by numerous factors, such as traffic, road type, and driving style. Higher thresholds could imply a stronger correlation to safety critical events. Through this metric it is possible to understand the range of amplitudes and frequency of acceleration epochs per unit distance traveled.

4.2.2.3 Comparing the Strongest Epoch in a Threshold Distance

This method compared participants based on the strongest epoch they experienced within a certain driving distance. The following measures were calculated using this method:

- Strongest epoch experienced in 1 mile/kilometer
- Strongest epoch experienced in 5 miles/kilometers
- Strongest epoch experienced in 10 miles/kilometers
- Strongest epoch experienced in 50 miles/kilometers
- Strongest epoch experienced in 100 miles/kilometers

- Strongest epoch experienced in 1000 miles/kilometers.

To calculate the strongest epoch experienced in a certain distance, the percentile equivalent of that distance was first calculated. Equation 4.2,

$$\lambda_z = \frac{\left(1 - \frac{z}{100}\right) \times N_T}{D_T} \tag{4.2}$$

shows the rate of epochs, λ_Z, above the Z^{th} percentile, with N_T being the total number of epochs experienced and D_T being the total distance travelled. The distance travelled per Z^{th} percentile epoch is therefore given by inverting the equation as follows in Equation 4.3:

$$\zeta = \frac{1}{\lambda_Z} = \frac{D_T}{\left(1 - \frac{Z}{100}\right) \times N} \tag{4.3}$$

Therefore, on average, the Z^{th} percentile epoch is the strongest epoch in ζ miles. To calculate the strongest epoch in a certain distance, say $\zeta = Y$ miles, the corresponding percentile can be found using Equation 4.4:

$$Z = 100 \times \left(1 - \frac{D_T}{Y \times N_T}\right) \tag{4.4}$$

Since the epoch rate, N_T/D_T , is different for different drivers, equation 4 allows a comparison of drivers based on unequal percentiles that correspond to an equal amount of driving.

Like the other statistical measures used to create the profile, these measures also allow driver comparison for frequent (strongest epoch experienced in 1 mile) as well as rare epochs (strongest epoch in 100 miles). This approach enables users of the Surface Acceleration Reference to identify what extreme epoch might be expected within a certain distance of travel, and is important for understanding the differences in epochs that are due to driving

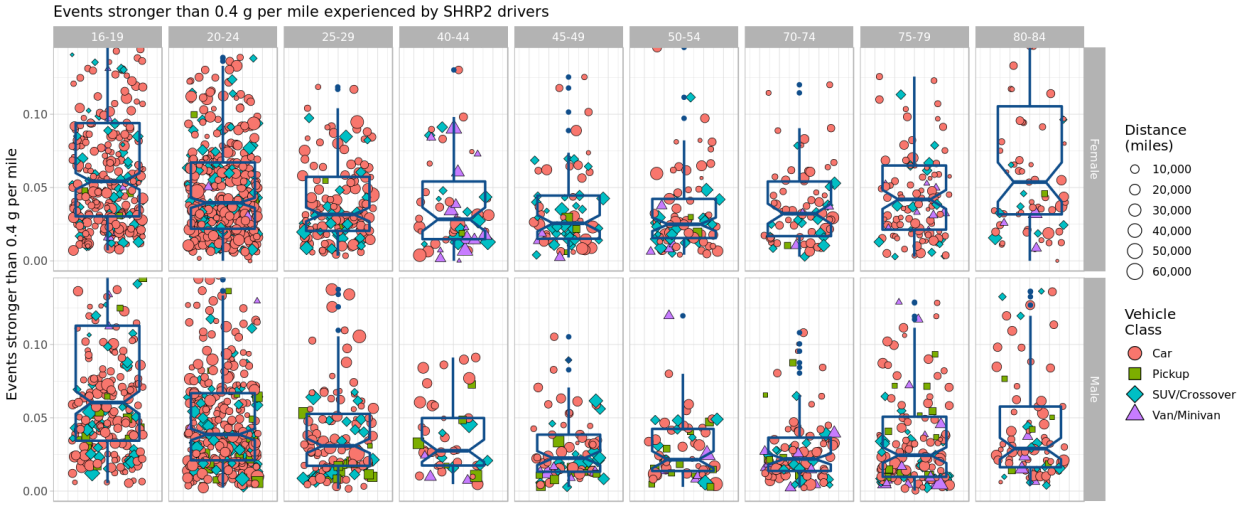


Figure 4.4: Interactive visualization available via the Surface Acceleration Reference web app comparing data based on age range and gender.

style versus those that are due to adverse events. The three categories of measures together provide a comprehensive driving profile for each participant in the SHRP 2 dataset and return units of analysis that are useful for users from multiple domains. All three types of measures were contextualized by roadway functional class and speed category on which they occurred. These measures have been made available for download in the form of a CSV file here.

4.2.3 Creating Interactive Visualization from Driver Acceleration Profiles

To achieve the project's purpose of creating an easily accessible dataset for a wide variety of users from various fields and backgrounds, it was important to have a low-effort threshold for data analysis. Even though the statistical driving profile for each participant is available via a downloadable CSV file, considerable effort is still required to ingest the data, understand its structure, and make meaningful inferences. To lower this threshold, a web app based on the

same dataset was developed using R, ggplot2, and Shiny [28, 140, 141]. This web app can be accessed through any web browser at www.vtti.vt.edu/surface-accelerations-reference.html. This interface allows researchers to compare SHRP 2 participants on any and all of the aforementioned measures with the ability to interrogate data by age, gender, location, and vehicle type. Figure 4.4 shows an example comparison for rate of epochs greater than 0.3 g where the data has been parsed by gender (category shown on right side of charts) along the vertical axis and age along the horizontal axis (category shown on top of charts).

Figure 4.4 provides an example of one combination of conditions that can be created using the visualization tool. Across all plots that might be created using the visualization tool, each mark in a plot represents one participant. The mark's position along the y-axis represents the measures described previously (i.e., percentile value, epoch rate, etc.) for a specific driver. The position of a point along the x-axis is randomly generated noise that aids visualization by reducing overlapping points. As illustrated in the legend in Figure 5, across plots, the size of a mark represents the amount of driving in the study and the shape and color represent the vehicle class. The various plot tiles or cells represent the categories to which each driver belongs making it easy to compare different classes. A box plot is overlaid on the scatter plot to show the median, quantiles and the outliers of the distribution. Hovering over a data point reveals additional information about the driver, including the exact value of the measure, gender, age, vehicle class, and total distance in the study. In addition, users are provided the ability to choose the units of distance, clip outliers for better resolution, display summary statistics, and plot the kernel distribution of the group in the form of a violin plot.

4.3 Findings

The main purpose of this project was to create an accessible catalog of accelerations experienced by road users in passenger vehicles that would enable users to answer questions pertinent to their fields. Therefore, most of the findings will come from the exploration of the Surface Accelerations Reference dataset and visualization tool. As an example of its application, differences in rates of accelerations on roads of various road speed categories are explored below.

Figure 4.5a shows the variation of epoch rates at the various maximum deceleration thresholds for all participants in SHRP 2 NDS grouped by the road speed category. The box plots show the distribution of rates at key thresholds and the dashed line represents the median calculated at a finer resolution. The threshold magnitudes are scaled linearly and the epoch rates are scaled logarithmically. A zero value for the median or the 25^{th} percentile lies at $-\infty$ on a log scale and therefore, is shown by keeping the box plot open. Figures 4.5b, 4.6a, and 4.6b describe similar information for longitudinal accelerations, lateral neagtive accelerations (left leaning), and lateral positive accelerations (right leaning) respectively.

The relationship between the epoch rates and the deceleration thresholds is log piecewise linear, i.e., the rate decreases exponentially as the threshold is increased linearly. Within the epoch rate distribution at each threshold, the rates for individual road speed categories differ by 1 to 2 orders of magnitude. For example, the deceleration epoch rate for the maximum deceleration threshold of 0.2 g is 2.8 epochs per mile for < 31 mph roads and 2.8 epoch per 100 miles for 65-80 mph roads. Tables 4.4, 4.5, 4.7, and 4.6 in the Appendix list the values of the 25^{th}, 50^{th}, and 75^{th} percentile epoch rates at key maximum acceleration thresholds.

Within each distribution shown in Figures 4.5a, 4.5b, 4.6a, and 4.6b the interquartile range is significant and increases with road speed as well as g-force magnitude. This implies that

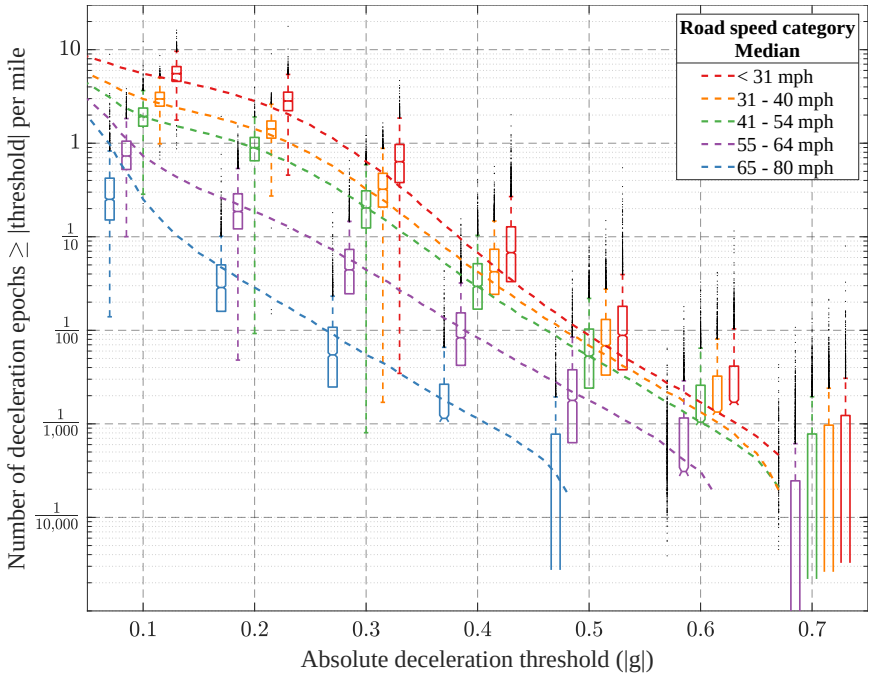
kinematic behaviour of the various participants has a higher variance for stronger thresholds. In other words, the proportional difference in rate between participants increases for stronger threshold epoch as well as on higher speed roads. These differences between various drivers can be explored using the accelerations reference and valuable insights can be drawn about driver kinematic behaviour.

These inferences will be useful across a number of fields. For example, designers of advanced driver assistance and automated driving systems can use Figures 4.5a to 4.6b and Tables 4.4 to 4.7 to compare the accelerations produced by their systems to a diverse driving population while controlling for roadway speed. This will enable system tuning to match certain driving populations such as the median driver or a driver in the interquartile range.

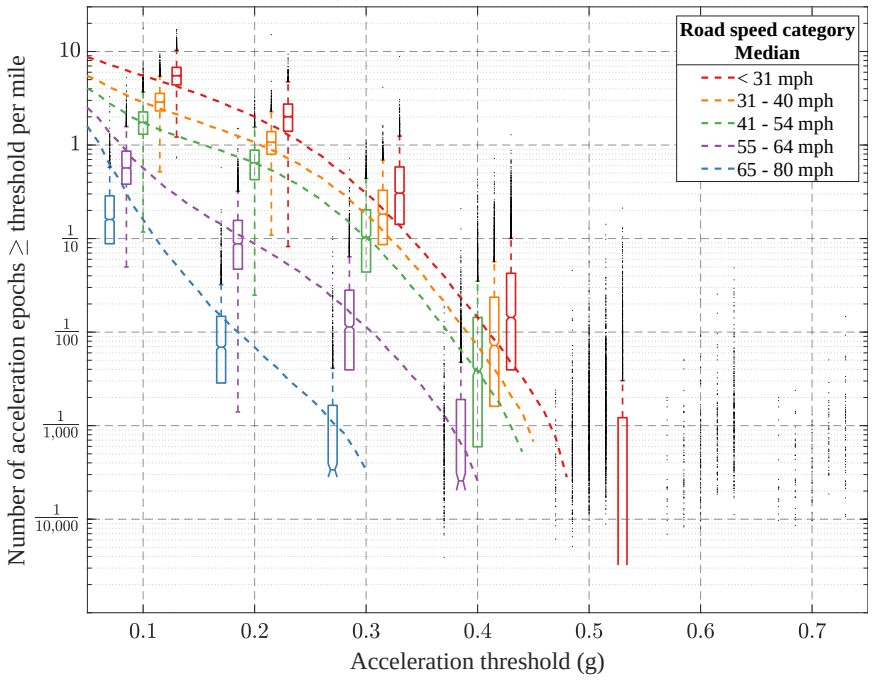
Organizations analyzing cohort data such as ride share service providers and taxi companies will be able to use the Surface Accelerations Reference for comparisons of their drivers with a national driving population. Similarly, "pay as you drive" and "pay how you drive" insurance programs will be able to utilize this data to compare their customers with a standardized national distribution allowing more transparency and, therefore, better adoption.

4.4 Conclusions

This study fulfills three major unmet needs in the study of accelerations based driving characteristics. First, the Surface Acceleration Reference provides a representative catalog of lateral and longitudinal accelerations in the form of a downloadable dataset and interactive analytics tool available at dataviz.vtti.vt.edu/Surface_Accelerations_Reference/. This catalog summarizes the acceleration behavior of over 3,500 participants in the SHRP2 NDS and is the largest openly available dataset of its kind. It can benefit a number of fields. For example, it will facilitate intelligent vehicle system engineers to design systems that safer

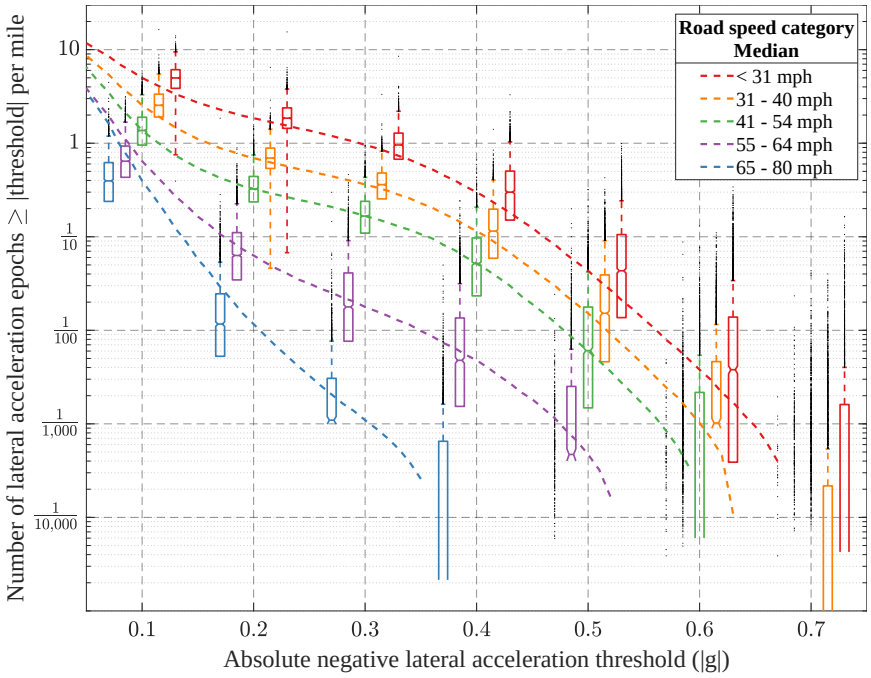


(a) Deceleration

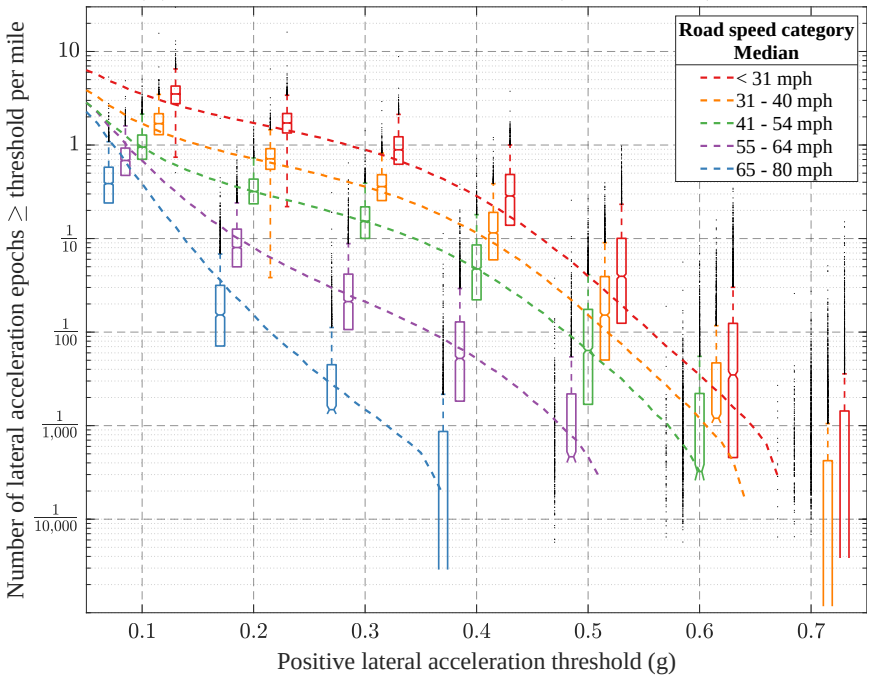


(b) Acceleration

Figure 4.5: Comparison of longitudinal driver epoch rates at various maximum thresholds.



(a) Negative lateral acceleration (left leaning).



(b) Positive lateral acceleration (right leaning).

Figure 4.6: Comparison of lateral driver epoch rates at various maximum thresholds.

and better representative of driver preferences. It will also enable behavioral researchers to understand the effect of various roadway, driver, and vehicle characteristics on the frequency and magnitude of accelerations.

Second, this paper introduces a methodology to create a standardized catalog of accelerations that can be replicated on similar datasets. A standardized approach has been missing in the field of acceleration behavior as different studies have used varying methods of comparing accelerations across populations. An algorithm was developed to summarize all the acceleration epochs in a driver's history to create profiles describing each participants acceleration norms and frequencies at different levels. This methodology includes signal processing, epoch definitions, summary extraction and standardization techniques. The statistical profiles created for every driver had three types of measures based on percentiles, rates, and distance thresholds, which makes this methodology applicable for a wide array of fields. This methodology can be used by agencies having similar datasets such as vehicle fleet operators, ride share services, government departments of transportation, and research institutes around the world to compare driving populations.

Finally, an analysis was conducted to measure the effect of road speed category on the distribution of longitudinal and lateral acceleration rates. It was observed that lower speed roads ($<$ 31 mph) have 2 to 3 orders of magnitude higher rates of strong g-force events as compared to high speed roads (65-80 mph). It was also observed that the relationship between the rate of epochs (i.e., accelerations) and the g-force magnitude was log-piecewise linear. I.e., for a linear increase in g-force magnitude, the rate of occurrence decreases exponentially. Even when controlling for the road speed category, the variance for high g-force events was significant and could indicate differences in driving style and effects of traffic conditions, driver demographics, etc.

Intelligent driving systems such as ADAS and automated vehicles interact with drivers, pas-

sengers, and other road users in safety critical ways with very little margin for error. Therefore, it is essential that such systems are developed based on data accurately representing the breadth of human behaviors and driving environments. The Surface Accelerations Reference provides a large-scale, real-world, diverse, and context rich catalog of accelerations. In addition to vehicle system development, these data will also provide value to many transportation related fields such as roadway, vehicle and safety systems design.

Table 4.4: The 25th, 50th, and 75th percentile deceleration epoch rates for all drivers at various thresholds grouped by road speed category.

Road speed category (mph)	**Measure (%tiles)**	**Rate of epochs per mile greater than threshold**					
		0.1 (g)	**0.2 (g)**	**0.3 (g)**	**0.4 (g)**	**0.5 (g)**	**0.6 (g)**
All	25	2.2E+00	9.3E-01	1.7E-01	1.9E-02	2.7E-03	4.7E-04
All	50	2.9E+00	1.3E+00	2.8E-01	3.4E-02	5.3E-03	1.1E-03
All	75	3.7E+00	1.8E+00	4.6E-01	6.5E-02	1.0E-02	2.4E-03
<31	25	4.6E+00	2.2E+00	3.8E-01	3.3E-02	3.8E-03	0
<31	50	5.5E+00	2.8E+00	6.3E-01	6.7E-02	8.8E-03	1.7E-03
<31	75	6.6E+00	3.5E+00	9.7E-01	1.3E-01	1.8E-02	4.1E-03
31 - 40	25	2.5E+00	1.1E+00	2.1E-01	2.4E-02	3.3E-03	0
31 - 40	50	3.0E+00	1.4E+00	3.2E-01	4.2E-02	6.8E-03	1.3E-03
31 - 40	75	3.5E+00	1.7E+00	4.7E-01	7.3E-02	1.3E-02	3.2E-03
41 - 54	25	1.5E+00	6.3E-01	1.2E-01	1.6E-02	2.3E-03	0
41 - 54	50	1.9E+00	8.8E-01	2.0E-01	2.9E-02	5.3E-03	9.7E-04
41 - 54	75	2.4E+00	1.1E+00	3.1E-01	5.1E-02	1.0E-02	2.6E-03
55 - 64	25	5.2E-01	1.2E-01	2.5E-02	4.2E-03	5.2E-04	0
55 - 64	50	7.3E-01	1.9E-01	4.5E-02	8.3E-03	1.7E-03	2.1E-04
55 - 64	75	1.0E+00	2.9E-01	7.4E-02	1.5E-02	3.8E-03	1.1E-03
65 - 80	25	1.5E-01	1.5E-02	2.3E-03	0	0	0
65 - 80	50	2.5E-01	2.8E-02	5.2E-03	1.0E-03	0	0
65 - 80	75	4.1E-01	4.9E-02	1.1E-02	2.6E-03	6.8E-04	0

Table 4.5: The 25th, 50th, and 75th percentile acceleration epoch rates for all drivers at various thresholds grouped by road speed category.

Road speed category (mph)	**Measure (%tiles)**	**Rate of epochs per mile greater than threshold**					
		0.1 (g)	**0.2 (g)**	**0.3 (g)**	**0.4 (g)**	**0.5 (g)**	**0.6 (g)**
All	25	1.9E+00	6.0E-01	6.7E-02	2.0E-03	0	0
All	50	2.6E+00	8.8E-01	1.4E-01	6.8E-03	1.0E-04	0
All	75	3.5E+00	1.3E+00	2.7E-01	2.1E-02	5.6E-04	0
<30 mph	25	4.4E+00	1.4E+00	1.4E-01	4.0E-03	0	0
<30 mph	50	5.5E+00	2.0E+00	3.1E-01	1.4E-02	0	0
<30 mph	75	6.8E+00	2.7E+00	5.9E-01	4.3E-02	1.2E-03	0
31 - 40	25	2.3E+00	7.9E-01	8.5E-02	1.6E-03	0	0
31 - 40	50	2.9E+00	1.1E+00	1.8E-01	7.3E-03	0	0
31 - 40	75	3.6E+00	1.4E+00	3.3E-01	2.4E-02	0	0
41 - 54	25	1.3E+00	4.1E-01	4.1E-02	4.3E-04	0	0
41 - 54	50	1.7E+00	6.3E-01	1.0E-01	3.7E-03	0	0
41 - 54	75	2.2E+00	8.7E-01	2.0E-01	1.4E-02	0	0
55- 64	25	3.8E-01	4.7E-02	3.9E-03	0	0	0
55- 64	50	5.7E-01	9.0E-02	1.2E-02	1.6E-04	0	0
55- 64	75	8.6E-01	1.6E-01	2.9E-02	1.9E-03	0	0
65 - 80	25	8.7E-02	2.6E-03	0	0	0	0
65 - 80	50	1.6E-01	6.6E-03	1.2E-04	0	0	0
65 - 80	75	2.8E-01	1.5E-02	1.6E-03	0	0	0

Table 4.6: The 25th, 50th, and 75th percentile lateral positive acceleration epoch rates for all drivers at various thresholds grouped by road speed category.

Road speed category (mph)	**Measure (%tiles)**	**Rate of epochs per mile greater than threshold**					
		0.1 (g)	**0.2 (g)**	**0.3 (g)**	**0.4 (g)**	**0.5 (g)**	**0.6 (g)**
All	25	1.4E+00	5.7E-01	2.5E-01	5.8E-02	6.4E-03	4.7E-04
All	50	1.8E+00	7.7E-01	3.7E-01	1.2E-01	1.8E-02	2.0E-03
All	75	2.3E+00	1.0E+00	5.2E-01	2.0E-01	4.2E-02	6.2E-03
<30 mph	25	2.7E+00	1.3E+00	6.2E-01	1.4E-01	1.2E-02	4.0E-04
<30 mph	50	3.5E+00	1.7E+00	8.9E-01	2.9E-01	4.0E-02	3.5E-03
<30 mph	75	4.3E+00	2.2E+00	1.2E+00	4.9E-01	1.0E-01	1.2E-02
31 - 40	25	1.3E+00	5.5E-01	2.5E-01	5.8E-02	4.9E-03	0
31 - 40	50	1.7E+00	7.1E-01	3.6E-01	1.1E-01	1.5E-02	1.1E-03
31 - 40	75	2.1E+00	9.1E-01	4.8E-01	1.9E-01	3.9E-02	4.7E-03
41 - 54	25	6.9E-01	2.3E-01	9.7E-02	2.1E-02	1.5E-03	0
41 - 54	50	9.4E-01	3.2E-01	1.5E-01	4.6E-02	6.2E-03	1.9E-04
41 - 54	75	1.3E+00	4.3E-01	2.2E-01	8.4E-02	1.7E-02	2.2E-03
55- 64	25	4.5E-01	4.8E-02	1.1E-02	1.7E-03	0	0
55- 64	50	6.7E-01	7.9E-02	2.1E-02	5.3E-03	4.0E-04	0
55- 64	75	9.2E-01	1.3E-01	4.2E-02	1.3E-02	2.2E-03	0
65 - 80	25	2.4E-01	6.7E-03	0	0	0	0
65 - 80	50	3.8E-01	1.5E-02	1.3E-03	0	0	0
65 - 80	75	5.8E-01	3.2E-02	4.4E-03	7.6E-04	0	0

Table 4.7: The 25th, 50th, and 75th percentile lateral negative acceleration epoch rates for all drivers at various thresholds grouped by road speed category.

Road speed category (mph)	**Measure (%tiles)**	**Rate of epochs per mile greater than threshold**					
		0.1 (g)	**0.2 (g)**	**0.3 (g)**	**0.4 (g)**	**0.5 (g)**	**0.6 (g)**
All	25	1.8E+00	5.8E-01	2.6E-01	6.1E-02	6.0E-03	4.3E-04
All	50	2.4E+00	7.9E-01	3.8E-01	1.2E-01	1.8E-02	1.7E-03
All	75	3.2E+00	1.1E+00	5.5E-01	2.0E-01	4.2E-02	5.9E-03
<30 mph	25	3.9E+00	1.4E+00	6.8E-01	1.5E-01	1.4E-02	4.5E-04
<30 mph	50	5.0E+00	1.9E+00	9.6E-01	3.0E-01	4.4E-02	3.8E-03
<30 mph	75	6.1E+00	2.4E+00	1.3E+00	5.1E-01	1.1E-01	1.4E-02
31 - 40	25	1.9E+00	5.3E-01	2.5E-01	5.9E-02	4.6E-03	0
31 - 40	50	2.5E+00	6.9E-01	3.6E-01	1.2E-01	1.5E-02	1.0E-03
31 - 40	75	3.3E+00	8.9E-01	4.8E-01	2.0E-01	3.9E-02	4.6E-03
41 - 54	25	9.4E-01	2.3E-01	1.1E-01	2.3E-02	1.4E-03	0
41 - 54	50	1.3E+00	3.2E-01	1.6E-01	5.1E-02	6.1E-03	0
41 - 54	75	1.9E+00	4.4E-01	2.4E-01	9.7E-02	1.8E-02	2.1E-03
55- 64	25	4.2E-01	3.4E-02	7.4E-03	1.4E-03	0	0
55- 64	50	6.4E-01	6.2E-02	1.8E-02	4.7E-03	4.1E-04	0
55- 64	75	9.2E-01	1.1E-01	4.1E-02	1.3E-02	2.5E-03	0
65 - 80	25	2.3E-01	5.1E-03	0	0	0	0
65 - 80	50	3.9E-01	1.1E-02	9.9E-04	0	0	0
65 - 80	75	6.2E-01	2.5E-02	3.1E-03	6.0E-04	0	0

Chapter 5

Quantifying the Effect of Roadway, Driver, Vehicle, and Location Characteristics on the Frequency of Longitudinal and Lateral Accelerations

Ali, G., McLaughlin, S., & Ahmadian, M. (2021). Quantifying the effect of roadway, driver, vehicle, and location characteristics on the frequency of longitudinal and lateral accelerations. Accident Analysis & Prevention, 161, 106356.

Abstract

The purpose of this study is to understand and quantify the simultaneous effects of roadway speed category, driver age, driver gender, vehicle class, and location on the rates of longitudinal and lateral acceleration epochs. The rate of usual as well as harsh acceleration epochs are used to extract insights on driving risk and driver comfort preferences. However, an analysis of acceleration rates at multiple thresholds incorporating various effects while

using a large scale and diverse dataset is missing. This analysis will fill this research gap. Data from the 2nd Strategic Highway Research Program Naturalistic Driving Study (SHRP2 NDS) was used for this analysis. The rate of occurrence of acceleration epochs was modeled using negative binomial distribution based generalized linear mixed effect models. Roadway speed category, driver age, driver gender, vehicle class, and location were used as the fixed effects and the driver identifier was used as the random effect. Incidence rate ratios were then calculated to compare subcategories of each fixed effect. Roadway speed category has the strongest effect on longitudinal and lateral accelerations of all magnitudes. Acceleration epoch rates consistently decreases as the roadway speed category increases. The difference in the rates depends on the threshold and is up to three orders of magnitude. Driver age is another significant factor with clear trends for longitudinal and lateral acceleration epochs. Younger and older drivers experienced higher rates of longitudinal accelerations and decelerations. However, the rate of lateral accelerations consistently decreased with age. Vehicle class also has a significant effect on the rate of harsh accelerations with minivans consistently experiencing lower rates.

5.1 Introduction

Every year in the United States, motor vehicle crashes cause thousands of fatalities, millions of injuries, and economic losses nearing a quarter trillion dollars [20, 139]. The critical factor in a majority of crashes is the driver (93.5%), followed by environment (2.4%), vehicle (2%), and other reasons (2.1%) [34, 117]. Therefore, significant research has focused on understanding driver behavior and its relationship to safety critical events. Driver behavior has been studied using various observable measures such as acceleration, speeding, following distance, and driver attentiveness. All of these are important and shed light on different

aspects of driving safety.

A number of studies have shown that a positive correlation exits between the crash rates of drivers and the frequency of harsh acceleration epochs [13, 14, 40, 70, 114, 120, 133, 143, 144]. Most of these studies are based on naturalistic driving data as it provides insight into normal driving as well as safety critical events for the same set of drivers. The positive correlation between crash and harsh acceleration rates has been shown for both longitudinal as well as lateral acceleration events [114]. However, when trying to explore these correlations, varied acceleration thresholds have been used to define harsh acceleration events and a consistent methodology does not exist [65]. Moreover, most studies analyzing harsh acceleration events do not simultaneously control for the various factors that could be affecting driver acceleration behavior such as roadway properties, driver demographics, vehicle characteristics, and environmental features.

Age and driving experience are an important factor in crash risk. Considerable research has shown that young drivers under the age of 25 years and older drivers above the age of 74 years experience higher driving risk than the drivers between the ages of 25 to 74 years [8, 29, 32, 51, 84, 85, 97, 99, 108, 115, 127, 136, 139, 142]. Studies have also shown that driver age, experience, and other driver demographic factors affect the rate of harsh acceleration epochs [8, 82, 102, 116]. However, there is a need for further investigation to quantify the effect of age on the rates of acceleration events at multiple thresholds based on real-world naturalistic driving data.

A number of studies have shown that roadway properties affect driving risk as well the rate of harsh acceleration epochs [102, 103, 105]. Roadway properties such as speed limit, type of access (e.g, controlled or not controlled), and functional class are major predictors for traffic speed and influence the frequency of stops and turns. Similarly, vehicle properties such as age and weight have been shown to affect driving risk and harsh acceleration frequency [22, 63].

Therefore, it is important that these factors also be simultaneously taken into account with the driver demographics when making comparisons about driving behavior and risk.

The aggregated driving complexity of a location can be influenced by several factors such as population density, traffic conditions, urban versus rural proportions, weather, intersection density, etc. Since these factors vary by location, so does driving complexity. Several studies have shown that location of driving can also affect the driving risk and acceleration behavior [71, 102, 103]. This, again, shows that analyses comparing acceleration based driving behavior should control for the the location in which the driving occurs.

Even though many studies explore the different factors affecting driving behavior, some major gaps still exist. Most of the studies simultaneously account for only one or two factors affecting driver behavior. Also, when identifying harsh acceleration behavior, most studies use a simple threshold in various driving conditions. To fill these gaps in our understanding of acceleration behavior, a comprehensive analysis based on a large scale and representative dataset and accounting all the major factors was needed. This study aims to fulfill this need. In this paper, the longitudinal and lateral acceleration rates of drivers at multiple thresholds were modeled using generalized linear mixed effect models. The various factors such as roadway speed, driver demographics, vehicle properties and location were used as fixed effects. These models were then used to calculate the rate ratios for acceleration frequency at different thresholds between subcategories of roadway speed, vehicle type, location, driver age, and driver gender. The value of the rate ratio quantified how each factor affects acceleration behavior.

This research will be beneficial in a number of fields. With better understanding of age effects on acceleration behavior, policy makers can devise programs to better detect at risk younger and older drivers and devise training programs to help them. Vehicle system designers can use the rate ratio based quantification to better specify or tune their systems for different

user demographics as well as roadway conditions. Various organizations that collect and process cohort based driving data such as insurance companies, transportation providers, and national traffic safety agencies can use this analysis to compare the trends in their user populations with a diverse and representative dataset.

5.2 Literature Review

Table 5.1 summarizes some of the major studies analyzing the relationship between acceleration based driving behavior, crash risk, driver demographics, vehicle characteristics, and roadway properties. Even though these studies have many major insights, Table 5.1 only summarizes the insights relevant to our discussion. The data source used by the researchers is also of importance and therefore has been briefly described as well.

5.2.1 Data Sources Used in Literature

The data sources described in Table 5.1 belong to one of the following categories:

1. Crash datasets such as the Fatal Accident Reporting System. These datasets are often maintained by federal agencies or police departments and offer details about hard to find crashes. However, they often need to be augmented with other datasets to estimate mileage based rates.

2. Naturalistic driving studies such as the 100 car NDS and the SHRP2 NDS. Large naturalistic driving studies are a great resource for studying driver behavior. However, finding rare and specific scenarios requires advanced data mining methods. Extremely rare driving events such as fatal crashes are often not captured in such studies.

Table 5.1: A summary of existing literature about acceleration behavior and its relationship to various factors of interest.

Study	Data source	Driving behavior insights relevant to this study
Relationship between acceleration behavior and safety critical events		
[70]	100 Car Naturalistic Driving Study with 109 participants and data collected for over a year	Drivers with higher rates of crashes and near crashes showed significantly higher rates of decelrations $\geq$ 0.3g and lateral accelerations $\geq$ 0.3g.
[13, 14]	Naturalistic driving study with 166 participants. 100 Car Naturalistic Driving Study with 109 participants and data collected for over a year.	A relationship exits between self reported accidents and jerky driving. For each additional critical jerk that the driver causes, a regression model showed that the number of accidents increases by about 1.13.
[114]	Teenage Naturalistic Driving Study with 42 newly licensed participants and data collected for 18 months.	Elevated g-force event rates can predict the crash and near crash likelihood in the near future. The correlation between the crashes/near crashes and elevated g-force rates was 0.60.
[120]	Smartphone GPS data from 4,000 drivers for 21,000 trips map matched to roadways	Hard braking and hard acceleration events were positively correlated to historical crash frequency at the roadway level.
Relationship between driver behavior and driver demographics, roadway properties, and vehicle characteristics.		

Table 5.1: A summary of existing literature about acceleration behavior and its relationship to various factors of interest.

Study	Data source	Driving behavior insights relevant to this study
[84]	Passenger vehicle travel data from the U.S. 1990 Nationwide Personal Transportation Survey, fatal crash data from 1990 Fatal Accident Reporting System (FARS), and police accident reports from the 1990 General Estimates System to produce crash involvement rates.	Rates of police reported crashes as well as fatalities is higher for younger and older drivers in the U.S. Drivers between the ages of 30 and 64 years are the safest.
[108]	Rates of crash involvement from the Road Injury Database of the Road Accident Prevention Research Unit at the University of Western Australia. The database contains all crashes that resulted in police reports or hospital admissions.	Younger and older drivers in Western Australia had the highest rates of crashes. Women had higher rates of crashes than men for all age groups.

Table 5.1: A summary of existing literature about acceleration behavior and its relationship to various factors of interest.

Study	Data source	Driving behavior insights relevant to this study
[32]	A linked dataset that contains licensing and crash data for drivers in New Jersey, U.S.A. 410,230 drivers were selected who got their intermediate license at 17-20 years of age from 2006-2009.	Crash rates are influenced by a combined effect of age and driving experience. For young drivers, crash rates reduce with increased driving experience.
[139]	A dataset cataloging the leading causes of death in the US available through the Web-based Injury Statistics Query and Reporting System (WISQARS) maintained by the Center of Disease Control and Prevention (CDC) USA.	Establishes that motor vehicle crashes are the leading cause of death for persons 16-24 years of age.
[51]	The SHRP2 naturalistic driving study with 3,542 participants accumulating 34.5 million miles of driving data over 3 years.	Younger and older drivers are more adversely affected by secondary-task engagement during driving than middle-aged drivers.
[29]	Dataset containing 1,614 police reported crashes that happened on 51 intersections in Seoul, South Korea.	Ageing drivers experience more crashes at intersections than younger drivers. This difference is greater when turning movements are involved.

Table 5.1: A summary of existing literature about acceleration behavior and its relationship to various factors of interest.

Study	Data source	Driving behavior insights relevant to this study
[116]	Teenage Naturalistic Driving Study with 42 newly licensed participants and data collected for 18 months in the US.	Young drivers experienced higher rates of hard braking events during first month of licensure. The rates were also significantly higher when no passengers were present as compared to with adult passengers. Male drivers experienced higher rates of hard braking events as compared to felame drivers.
[82]	Field operation test with 46 participants based in Wuhan, China. Each participant drove the same vehicle once over the same loop with a 14 km warm up.	This study concludes that there are gender based differences in drivers. Male drivers recognize more driving risk but also show more aggressive driving tendencies.
[8]	A driving simulator study in the CARRS-Q Advanced Driving Simulator with 78 participants who drove in four driving conditions. [9]	Age and gender-related effects are observed with young and male drivers having a higher probablity of engaging in a hard-braking event when driving without driving aids.

Table 5.1: A summary of existing literature about acceleration behavior and its relationship to various factors of interest.

Study	Data source	Driving behavior insights relevant to this study
[112]	A driving simulator study in the CARRS-Q Advanced Driving Simulator with 78 participants who drove in two driving environments.	Acceleration noise reduces in a connected environment. Young drivers take more advantage of connected environment relative to the middle-aged or old drivers.
[102]	Data from in vehicle data recorders collected by a Pay as You Drive insurance provider. 600 vehicles were sampled from a larger set based on involvement in crashes. All the selected vehicles were from Italy. The data recorders aggregated GPS signals every 2,000 meters before transmitting the data for storage.	Crash risk varies by time of day, day of week, road type, and vehicle speed. Driving is riskier at night versus during the day, during weekdays versus the weekends, on low speed roads versus high speed roads. This study also shows that driving is the safest in the mid-range velocities of 60-90 km/h as compared to speeds above or below.
[105]	Driving behavior data collected via smartphone sensors for 303 drivers while driving on two urban expressways in Athens, Greece. Traffic characteristics were obtained by 26 inductive loops on the two expressways.	Traffic characteristics such as traffic flow and speed have the most statistically significant impact on the rates of harsh acceleration events.

Table 5.1: A summary of existing literature about acceleration behavior and its relationship to various factors of interest.

Study	Data source	Driving behavior insights relevant to this study
[22]	The British STATS19 national road accident reporting system in combination with the Vehicle Registration Mark to find the data of first registration, make, model, and weight of vehicles involved in an accident.	The driver casualty rate falls significantly with the size of the car. The rate of casualty increases with the time since first registration.
[63]	Norwegian crash data from 2000 to 2016.	There are fewer drivers killed or seriously injured in newer and heavier cars.

3. Driving studies conducted on simulators such as the CARRS-Q Advanced Driving Simulator. Such studies are ideal when prior real world data does not exist or when behavior during rare and complex scenarios needs to be studied.

4. GPS traces from smartphones or in-vehicle data recorders. These studies offer a cheaper way to acquire datasets with many participants and large mileages. However, these datasets often lack the context around specific incidents offered by naturalistic driving studies. These studies can be augmented by other data sources such as insurance records to fill the missing context around crashes.

5. Field operation tests. In such studies, usually one vehicle is instrumented for data acquisition and multiple participants drive the same vehicle on the same roadways. These studies have been used to study specific predetermined scenarios in real world

traffic conditions. However, the effect of the participant driving a unfamiliar vehicle and being observed by an experimenter may influence their behavior.

It is important to consider the data source along with the insights from each study as it affects the application of the results and the confidence in the conclusions.

5.2.2 Relationship Between Acceleration Behavior and Driving Risk

There is ample evidence to show that drivers with high crash rates have different acceleration behavior than drivers with low crash rates. Using the 100 Car Naturalistic Driving Study data, [70] show that drivers with higher rates of crashes and near crashes also have significantly higher rates of decelerations and lateral accelerations $\geq$ 0.3g. While analyzing the effect of jerky driving on crash rates using two different naturalistic driving studies, [14] and [13], show that there is a relationship between critical jerk and self reported crash rates. Using GPS traces from smartphone data, [120] show that a positive correlation exists between historical crash frequency and hard braking/acceleration events by analyzing the location of the events. Therefore, understanding the factors that affect the frequency of harsh longitudinal and lateral acceleration events can help understand their role in crashes and near crashes as well.

5.2.3 Relationship Between Driving Behavior, Driver Demographics, Vehicle Characteristics, and Roadway Properties

A multitude of studies have examined the factors affecting driver behavior, especially when it concerns driving risk. [29, 32, 84, 108] and [139] use aggregated crash datasets to show

that age is an important factor affecting crash risk in the US, Australia, and South Korea. In all the studies younger and older drivers had higher crash rates than middle aged drivers. [51] use the SHRP2 NDS data to show that younger and older drivers are more adversely affected by secondary-task engagement during driving than middle-aged drivers.

Using the Teenage Naturalistic Driving Study, [116] show that younger drivers experienced higher rates of hard braking events during the first month of licensure. The same study also shows that male driver experienced higher rates of hard braking events as compared to female drivers. [82] used data from a field operation test to show that male drivers recognize more driving risk but also show more aggressive driving tendencies. Data from a driving simulator based study is used by [8] to show that young and male drivers have a higher probability of engaging in hard-braking events when driving without driving aids. Therefore, substantial evidence exists to show that driver age and gender can affect acceleration based driving behavior.

Using summarized GPS traces collected form vehicle data recorders [102] show that crash risk varies by time of day, day of week, road type, and vehicle speed. [105] uses GPS traces collected via smartphone sensors and traffic data collected via inductive loops to show that traffic flow and speed have statistically significant impact on the rates of harsh acceleration events. [22] and [63] use national crash data from Britain and Norway to show that vehicle size has an effect on driver casualty rates. Even though all these studies do not directly link to acceleration behavior to vehicle characteristics or roadway properties, sufficient evidence exists to show that such factors should be taken into account when studying acceleration behavior.

5.2.4 Gaps in Previous Studies

There are four major gaps in our current understanding of acceleration behavior. First, to be best of the author's knowledge, none of the existing studies examining acceleartion behavior simultaneously account for the multiple factors that affect behavior such as driver age, driver gender, roadway properties, vehicle class, and location. Second, most of the current studies use a simple constant threshold for analyzing harsh acceleration events which cannot be accurate in all driving conditions. For example, the median driver in SHRP2 NDS experiences a 0.3 g braking event once every mile or so while driving on low speed neighborhood roads but once every 200 miles while driving on an interstate [7]. Third, none of the existing studies have relied on large scale and representative driving studies such as the SHRP2 NDS. Even when larger datasets were available, only a small sample was analyzed. Finally, to the best of the author's knowledge, no existing studies analyze how the effect of earlier mentioned factors changes as the acceleration requirements for inclusion are increased from mild to harsh. This study aims to fill these gaps in our understanding of acceleration behavior.

5.3 Data

SHRP2 NDS is the ideal data set to effectively study the influence of roadway characteristics, driver demographics, and vehicle class on acceleration behavior. It is the single largest naturalistic driving study, consisting of over 34 million miles of driving conducted by about 3,500 participants recruited in six locations across the United States of America [11, 35, 53]. In addition to it's scale, the SHRP2 NDS was designed to capture diverse driver populations and driving conditions [12]. Finally, the data richness created through high frequency vehicle kinematics, driver inputs, video feeds, and GPS signals make the study an unparalleled data source for this analysis.

Table 5.2 summarizes the number of participants and corresponding mileage driven for each factor and its subcategories under consideration in this analysis. Almost all subcategories have more than a million miles of driving which illustrates the scale, diversity, and richness of the data.

Most of the participants have driven on every roadway speed category subtype and the mileages are uniformly distributed. The roadway speed category represents the usual driving speeds on a road segment and was obtained from HERE.com digital map data through map matching [59][87]. The map matching process uses latitude, longitude, and other vehicle timeseries data and finds the most appropriate matching segments from HERE.com digital maps. Since there are driving segments when the GPS system did not acquire meaningful data due to start up delay or low signal availability, such driving is marked as "Unknown" and has been excluded from the final analysis.

Even though there is a slightly higher number of female participants, SHPR2 NDS mileage is almost evenly split by gender. However, when it comes to age, this study was designed to over represent younger and older drivers in comparison to the national driving population. The difference is significant but will not affect the inferences drawn in this analysis as both age range and gender and examined as fixed effects in the analysis. The date of birth and gender were declared by the participants at the time of enrollment. Each driver was assigned an age group based on their age at the time of enrollment in the study. This was an additional measure to further obfuscate any personally identifiable information (PII) about the participants. It should be noted that the minimum age for issuance of driving license in the United States can vary by state. However, all collection sites for SHPR2 NDS were in states that allowed 16 year old drivers to apply for a restricted license.

All the vehicles were assigned one of four vehicle classes with cars being the most common subcategory, followed by SUVs/crossover, pickup trucks, and vans/minivans. These

Table 5.2: SHRP2 NDS participant and mileage summary by roadway speed category, driver gender, driver age, vehicle class, and collection site.

Factor	Sub categories	Number of participants	Number of million miles
Roadway speed category	0-30 MPH	3,545	4.79
	31-40 MPH	3,540	5.39
	41-54 MPH	3,535	6.48
	55-64 MPH	3,527	7.79
	65-80 MPH	3,159	5.17
	Unknown	3,546	4.89
Driver gender	Male	1,563	16.85
	Female	1,705	16.52
	Did not specify	278	1.12
Driver age range	16-19	547	4.45
	20-24	742	8.10
	25-29	278	3.29
	30-34	164	1.89
	35-39	128	1.32
	40-44	116	1.42
	45-49	147	1.72
	50-54	167	1.77
	55-59	144	1.47
	60-64	151	1.43

Table 5.2 continued from previous page

Factor	Sub categories	Number of participants	Number of million miles
	65-69	211	2.11
	70-74	174	1.64
	75-79	269	1.96
	80-84	156	1.19
	85+	70	0.351
	Did not specify	82	0.40
Vehicle class	Car	2,557	24.48
	Pickup Truck	170	1.84
	SUV - Crossover	721	6.84
	Van - Minivan	152	1.34
Collection site	Bloomington, IN	280	2.69
	Buffalo, NY	784	7.77
	Raleigh, NC	576	6.19
	Seattle, WA	835	7.36
	State College, PA	285	2.34
	Tampa, FL	787	8.15

categories are informative but should not be considered monolithic as seemingly different vehicles can have the same classification. There were 6 collection sites setup to recruit participants and swap filled hard drives. It should be noted that the actual location of driving strongly correlates with these collection sites but may, in some cases, be different. Given

that there is enough data in each subcategory and that vehicle class and location are examined as fixed effects, the over representation of one category should not affect the inferences drawn during analysis.

To analyze the effect of the above mentioned factors on acceleration behavior, a comprehensive dataset of acceleration epochs in SHRP2 NDS was needed. This requirement was fulfilled by The Surface Accelerations Reference which was created to catalog all longitudinal and lateral acceleration epochs experienced by drivers in SHRP2 NDS [7]. Using an algorithm, researchers discovered over 1.2 billion epochs and summarized them into multi-dimensional data points stored in an SQL queryable DB2 database. Each datapoint has multiple measures for acceleration, vehicle speed, driver inputs, and roadway properties. These summarized data points were then used to calculate three types of measures for each driver separated by roadway properties:

- Rate of events stronger than a threshold per mile. For example, rate of deceleration events stronger than 0.5 g per mile experienced on roadways with <30 mph speed category.
- Event magnitude at key percentiles. For example, magnitude of 95th percentile acceleration experienced on roadways with 31-40 mph speed category.
- Strongest event magnitude experienced at key measures of distance. For example, strongest right leaning lateral acceleration experienced in 100 miles of driving.

An online interactive tool was created to let users analyze and download the data [7]. Through this tool, user can visualize the above mentioned measures and also visually compare the effect of roadway properties, driver demographics, and vehicle class.

Using the same data, Figure 5.1 compares the distribution of longitudinal and lateral accel-

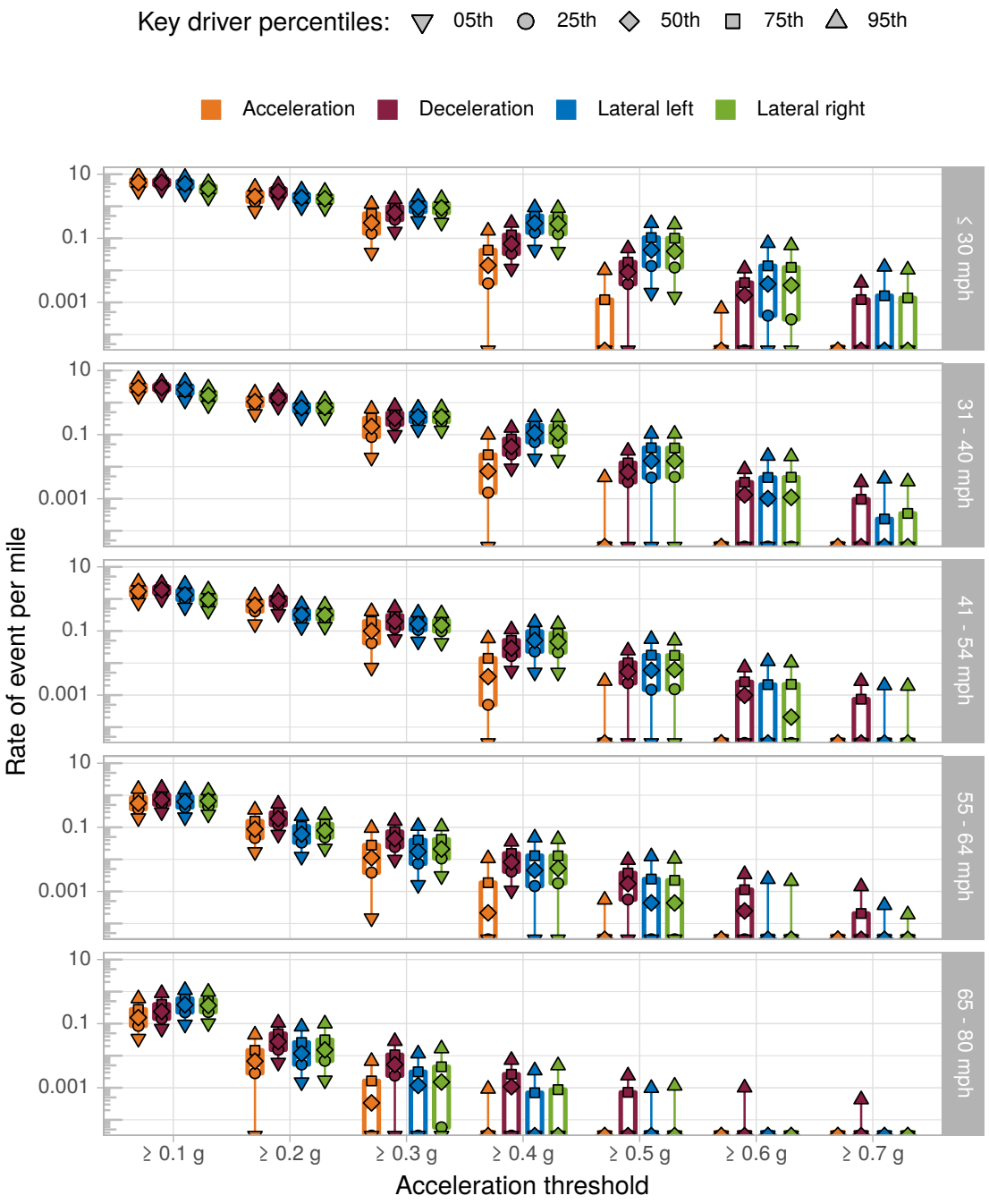


Figure 5.1: Distribution of acceleration rates for SHRP2 NDS drivers at various thresholds separated by roadway speed category.

eration rates. This is done by plotting the rates for 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentile drivers along the y-axis. The x-axis represents the acceleration threshold at which the rates are calculated and the data is separately plotted for each speed category. Since the y-axis spans over five orders of magnitude, it shows that even for the median drivers, some acceleration events are experienced about 10 times per mile whereas other can be as rare as once every 10,000 or more miles. The purpose of this figure is to set the general context of rarity by threshold and roadway type. This context is important when used with rate ratio plots in the results section.

For example, consider the deceleration rates for the median driver on ≤ 30 mph roadways, which mostly represent driving on neighborhood and minor collector roads. Deceleration ≥ 0.1 g and ≥ 0.2 g occur multiple times per mile and are more representative of the traffic conditions than driver preferences. ≥ 0.3 g deceleration occur about once every two miles and may be influenced by driving style and traffic conditions equally. ≥ 0.4 g, ≥ 0.5 g and ≥ 0.6 g deceleration epochs occur every 20, 100, and 1,000 miles respectively. These are likely to be created by the driver's response to unusual or rare incidents. Drivers with lower rates for such thresholds are either driving in safer environments, better at observation and prediction, or have superior vehicle control skills. The median driver in SHRP2-NDS never experienced a ≥ 0.7 g deceleration.

Figure 5.1 also illustrates that irrespective of acceleration type, as the threshold is increased, the distribution changes towards rates becoming lower. Also, the difference between the 25^{th} and the 75^{th} percentile driver grows with higher thresholds. Therefore, the rate distribution for each of the four acceleration types at every threshold should be modeled separately.

Even though the rates data provided in the accelerations reference informs on the differences by roadway, it does not quantify the individual contributions of various other factors in the rates of events. For example, if drivers in a particular age group have elevated rates of de-

celeration, is it because they primarily drove on lower speed roads or because that age group has a higher likelihood of producing such rates? Therefore, there is a need for quantifying the independent contributions of the roadway, vehicle, location, and driver attributes. This has been achieved using generalized linear mixed effect models in the analysis described in the next few sections.

5.4 Methods

The purpose of this analysis was to quantify the difference in rates of acceleration epochs based on roadway speed category, driver age, driver gender, vehicle class, and collection site location. To illustrate the methodology, consider the simpler analysis of comparing rate of strong deceleration on two types of roadways with speed categories of "$\leq$ 30 mph" and "65 - 80 mph".

Let $\lambda_{\leq 30\ mph}$ be the rate of deceleration epochs $\geq$ 0.5 g on roadways with speed category "$\leq$ 30 mph" and let $\lambda_{65\ -\ 80\ mph}$ be the rate of deceleration epochs $\geq$ 0.5 g on roadways with speed category "65 - 80 mph". Therefore, the rate ratio between the two is given by,

$$RR_{\frac{65\ -\ 80\ mph}{\leq 30\ mph}} = \frac{\lambda_{65\ -\ 80\ mph}}{\lambda_{\leq 30\ mph}} \tag{5.1}$$

To show that "65 - 80 mph" roadways produce lower rates of deceleration epochs than "$\leq$ 30 mph" roadways, the value of $RR_{65\ -\ 80\ mph/\leq 30\ mph}$ needs to be less than 1. However, to show that this would be true for 95% of such studies conducted, the upper limit of a 95% confidence interval for the rate ratio also needs to less than 1 as well. Therefore, for conclusive evidence of difference in behavior, both the rate ratio as well as the 95% confidence interval need to be calculated. For this analysis, these measures have been calculated using negative

binomial distribution based generalized linear mixed effect models which are often used for over dispersed data [50, 80].

There are four acceleration types under consideration and six thresholds ranging from ≥0.1 g to ≥0.6 g. For each combination of acceleration type and threshold, one mixed effect model was fitted to the data. The fixed effects were:

- roadway speed category
- vehicle class
- collection site location
- driver age
- driver gender

Since each driver drove over multiple roadways, the unique driver and vehicle identifier was used as the random effect. If λ_i is the rate of a particular acceleration type for a given threshold, the mixed effect model can be described as

$$\log \lambda_i = X_i\beta + \mathcal{Z}u_i \tag{5.2}$$

where X_i is the matrix of predictor variables for the driver, β is the column vector of fixed-effects regression coefficients for all drivers, $\mathcal{Z}$ is the design matrix for random effects and u_i is the random effect for the driver. The rate λ_i can also be expressed as

$$\lambda_i = \frac{\mu_i}{m_i} \tag{5.3}$$

where μ_i is the number of of relevant epochs experienced and m_i is the number of miles driven. Substituting equation 5.3 in equation 5.2:

$$\log \mu_i = X_i\beta + \mathcal{Z}u_i + \log m_i \tag{5.4}$$

This is a mixed effect count regression problem and the outcome variable can be represented by Poisson or Negative Binomial distributions. Since high threshold counts can be over dispersed, the negative binomial family was better suited and hence chosen for this analysis. Equation 5.5 symbolically represents the model for `lme4` package based `glmer.nb` function in R [15].

$$\log(\text{number of acceleration events}) \sim \text{roadway speed category} + \text{vehicle class} + \text{location} + \text{gender} + \text{age group} + (1|\text{driver vehicle id}) + \text{offset}(\log(\text{distance traveled})) \quad (5.5)$$

To facilitate incident rate ratio comparison, the model in equation 5.2 can also be represented as:

$$\lambda_i = e^{X_i\beta + \mathcal{Z}u_i} \quad (5.6)$$

To carry the simpler example in equation 5.1 forward, the two roadway speed subcategories can be compared using the above form of the predicted rate equation. To study incident rate ratios in different subcategories of a fixed effect, we can assume identical random effects. Also, since all fixed effects are categorical, the matrix of predictor variables is made up of 1's and 0's further simplifying the equation as:

$$\begin{aligned} RR_{\frac{65\ -\ 80\ mph}{\leq\ 30\ mph}} &= \frac{\lambda_{65\ -\ 80\ mph}}{\lambda_{\leq\ 30\ mph}} \\ &= e^{\beta\left(X_{65\ -\ 80\ mph} - X_{\leq\ 30\ mph}\right)} \\ &= e^{\left(\beta_{65\ -\ 80\ mph}\right)} \end{aligned} \quad (5.7)$$

There is a baseline level for each fixed effect that corresponds to 0, in this case $X_{\leq\ 30\ mph}$, and an indicator variable that corresponds to 1. The rate ratio only needs $\beta_{65\ -\ 80\ mph}$ as $\beta_{\leq\ 30\ mph}$ is already included in the intercept being the first subcategory of the fixed effect.

To calculate the 95% confidence interval, we can calculate the upper confidence limit (UCL) and lower confidence limit (LCL) as

$$RR_{\frac{65\ -\ 80\ mph}{\leq\ 30\ mph}\ UCL} = e^{\left(\beta_{65\ -\ 80\ mph} + 1.96 \times S.E\beta_{65\ -\ 80\ mph}\right)} \tag{5.8}$$

$$RR_{\frac{65\ -\ 80\ mph}{\leq\ 30\ mph}\ LCL} = e^{\left(\beta_{65\ -\ 80\ mph} - 1.96 \times S.E\beta_{65\ -\ 80\ mph}\right)} \tag{5.9}$$

where $S.E\beta_{65\ -\ 80\ mph}$ is the standard error associated with the $\beta_{65\ -\ 80\ mph}$ estimate. All other incident rate ratio comparisons discussed in Section 6.5 are calculated using similar methodology. For each of combination of the four acceleration types and six thresholds, the data was fitted with a separate model making a total of 24 models used in the results section. A good fit was ensured by trying multiple optimizers and comparing results to ensure similar convergence.

5.5 Results

The purpose of this analysis is to understand the effect of roadway speed category, driver age, vehicle class, location, and driver gender on the rates of common as well as rare acceleration epochs. To achieve this, the rate of the four acceleration types at thresholds of "$\geq$ 0.1g" to "$\geq$ 0.6g" was estimated using generalized linear mixed effect models with the the above mentioned factors as fixed effects and the driver-vehicle identifier as a random effect. Then the regression coefficients of the models were used to compare the subcategories of each fixed effect. Significant influence of roadway speed category and driver age is observed with consistent trends across all acceleration types and thresholds. Certain vehicle types and location also showed consistently different behavior. Driver gender had the smallest

influence among all the fixed effects. These results are discussed in greater detail in the following subsections.

Figures 5.2 to 5.6 have 24 subplots each with every subplot derived from a separately fitted generalized mixed effect model. For each plot, the y-axis represents the incidence rate ratio along with the 95% confidence interval shown as a circle with error bars and the x-axis represents the subcategories of the fixed effect being examined. For each subplot, the first subcategory on the left is the basis of comparison and hence has a rate ratio of 1. These plots are visualizing the relative frequency among the various subcategories and not the actual frequency of epochs. Therefore, Figure 5.1 should be also be used in conjunction for contextualizing the inferences. The underlying data produced by the analysis is also available for download with this paper. Subsections 5.5.1 to 5.5.5 discuss the important results.

5.5.1 Effect of Roadway Speed Category

Roadway speed category shows the most significant impact on the rates of acceleration epochs for all acceleration types and thresholds. Figure 5.2 shows the effect of roadway speed category on rate ratios with respect to the "$\leq$ 30 *mph*" speed category. The main takeaways are:

- For norms as well as extreme acceleration thresholds, as the roadway speed increases, the frequency of acceleration epochs decreases. For example, drivers on "$\leq$ 30 *mph*" roads experience about 100 times more $\geq$0.2 g deceleration than on "65 $-$ 80 *mph*" roadways. The much lower rates for high speed roadways can be explained in two ways. First, the driving is much more stable with speed or direction changing less frequently and the mixing of traffic traveling at different speeds is rare. Secondly, these rates are mileage based, and higher speed roadways accumulate more mileage in a lesser time.

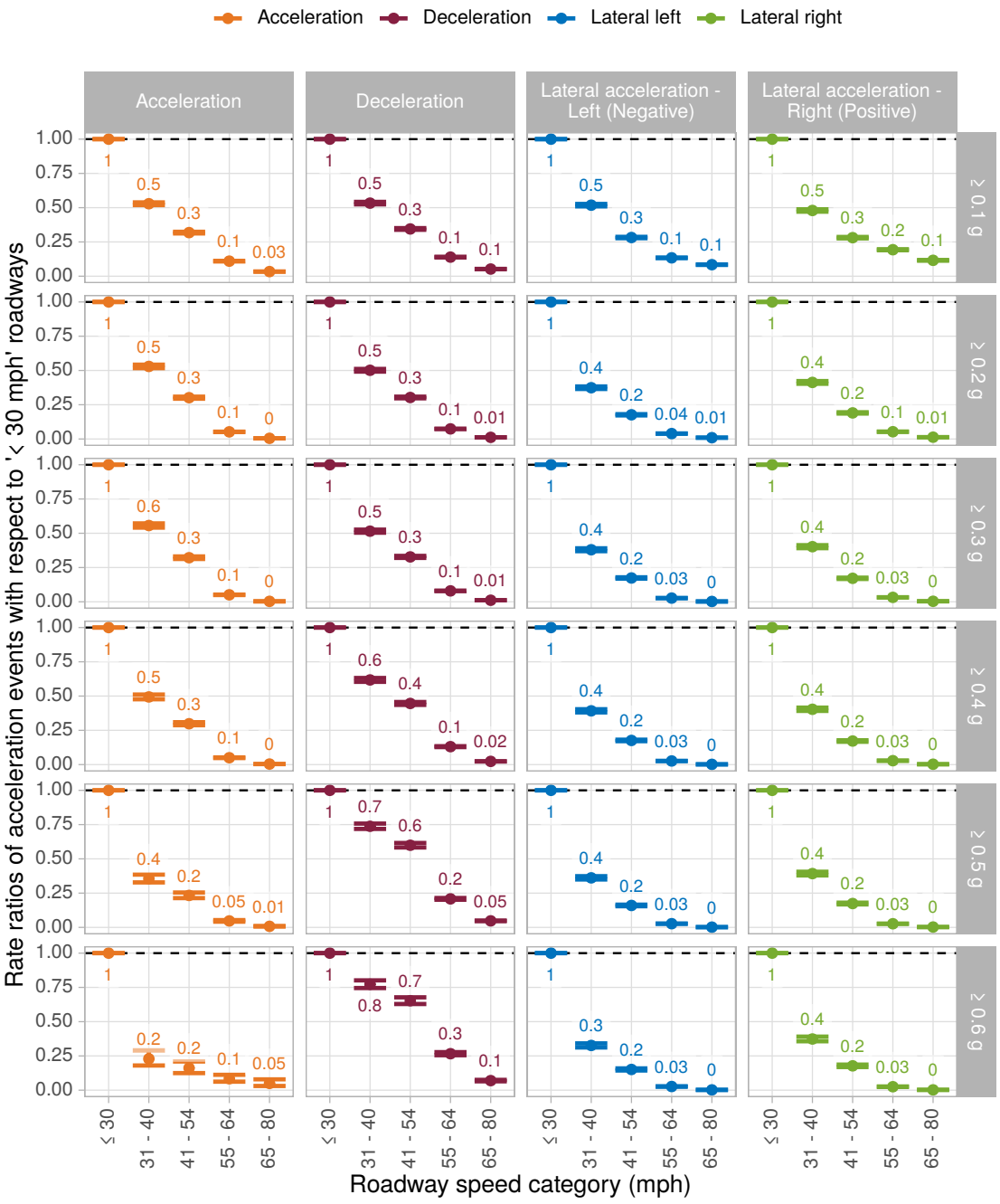


Figure 5.2: Comparing the effect of roadway speed category on incident rate ratios of the 4 acceleration types at the 6 thresholds with respect to speed category "≤ 30 *mph*".

- For higher thresholds, the drop in the relative rate is sharpest in longitudinal accelerations followed by lateral left and right accelerations. Deceleration epochs have a

slightly different trend with higher speed roads also producing relatively higher rates than other acceleration types. This is due to the vehicle's ability to produce higher deceleration than acceleration and the more frequent need to do so because of traffic lights and slowdowns.

- For all roadway speed comparisons, the 95% confidence intervals are relatively small when compared to other fixed effects. This indicates a strong characteristic influence of each subcategory. The tight confidence intervals are also due to most drivers having driven on all roadway speed categories, and hence a larger sample size as compared to other fixed effect subcategories.

5.5.2 Effect of Driver Age Range

Driver age group has significant effect on the relative rates of all acceleration types, especially at higher thresholds. Figure 5.3 shows the effect of driver age range on the rate ratios with respect to the "16 - 19" years age range. Some important observations are:

- For longitudinal acceleration at higher thresholds, as the age range increases from "16-19", the rate ratio keeps decreasing until the "70 - 74" group, after which it starts increasing again. "16-19" year old drivers experience the highest rates as compared to any other group. Other younger and older drivers also experience more frequent high-g longitudinal acceleration epochs when compared with ages 30 to 70. However, the 95% confidence intervals are wider than other acceleration types suggesting more variance within age groups.
- For longitudinal deceleration at higher thresholds, a similar trend is observed. Hard braking is experienced more frequently by younger and older drivers. For example, a $\geq$0.5 g deceleration is experienced more than twice as often by teenagers than by "50

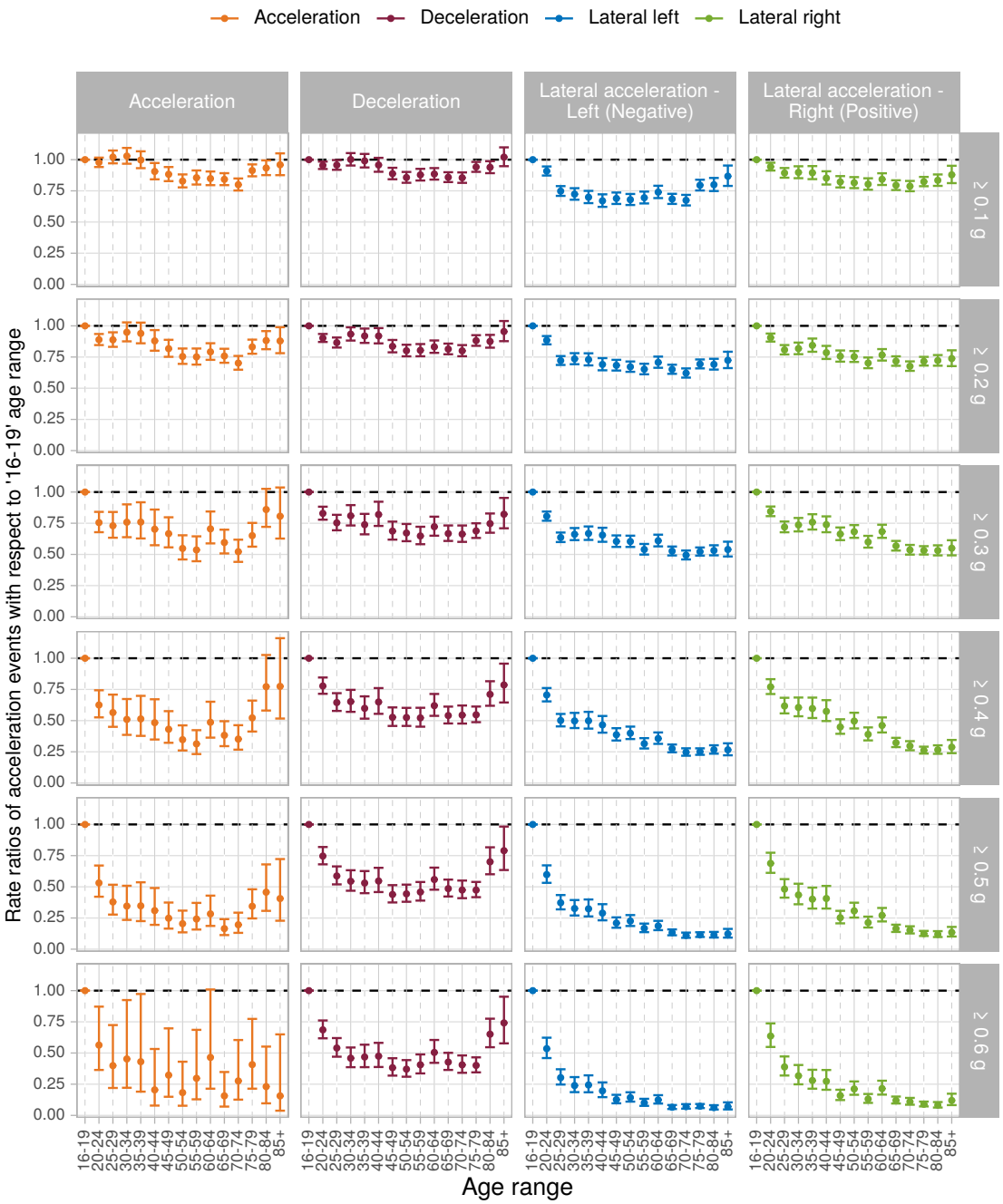


Figure 5.3: Comparing the effect of driver age on incident rate ratios of the four acceleration types at the six thresholds with respect to age range "16 - 19" years.

- 54" year old drivers.

In case of high-g deceleration rate ratios, there are certain bands with similar behavior

within the general age based trends. “16 - 19” age range has the highest rates followed by the “20 - 24” age group. Then groups within ages 25 to 44 seem to have similar values. Ages 45 to 79 also show similar but slightly lower values than groups within ages 25 to 44. 80 and above drivers show increased rates of hard deceleration epochs. Exploration of the exact mechanism of the age based differences is outside the scope of this paper. However, it should be noted that the mechanism for higher rates in younger and older drivers is probably different. For example, younger drivers could be produce higher rates due to inexperience whereas older drivers may produce it due to reduced perceptual capabilities and slower reaction times.

- The rates of high-g lateral acceleration epochs keeps decreasing as the driver age increases. The rate is highest for the teenage group followed by the “20 - 24” driver age group. Drivers between 25 to 44, 45 to 64, and 65 to 85 + form three bands of age groups that have progressively decreasing high-g lateral acceleration rates. For example, teenage drivers experience about 10 times as many ≥ 0.5 g left leaning lateral accelerations as “70 - 74” year old drivers. This is a stark difference in driving behavior purely based on the driver age.

  The driver age based trend is true for both left as well as right leaning lateral accelerations. This is in contrast to longitudinal acceleration and deceleration rates which showed higher rates for younger and older drivers. The difference in age based trends for high-g longitudinal versus lateral accelerations are due to the differences in the mechanism by which they occur. High-g longitudinal acceleration and deceleration epochs occur due to rapidly changing vehicle speed. However, high-g lateral acceleration epochs occur during rapidly changing vehicle direction such as turns at high speeds, tightly curved roads, and swerving maneuvers. Therefore, if certain groups are predisposed to driving at slower speeds, they may naturally experience lower rates of

high-g lateral acceleration epochs.

- For longitudinal acceleration and deceleration rates at lower thresholds of ≥ 0.1 g and ≥ 0.2 g, the differences are significant but less severe. Drivers from 16 to 44 years show similar rates whereas drivers from ages 45 to 74 show slightly lower rates.
- For lateral acceleration rates at lower thresholds of ≥ 0.1 g and ≥ 0.2 g, the differences are significant but less severe. Also, the trends are slightly different than longitudinal accelerations. Teenage drivers have the highest rates followed by the "20 - 24" age group. Drivers between the ages of 25 to 75 have similar rates which a slightly lower than the two youngest groups.

Exploring age effects on the rate ratios of high g-force acceleration rates could provide insight about change of driving expertise with age. This could be especially useful for designing age based training and certification programs for drivers.

5.5.3 Effect of Vehicle Class

Vehicle class has significant effect on the relative rates of all acceleration types at higher thresholds. Figure 5.4 shows the effect of vehicle class on the rate ratios with respect to the "Car" vehicle class. Some important observations are:

- At the lower thresholds of $\geq 0.1g$ and $\geq 0.2g$, all vehicle classes have similar rate ratios with minor differences.
- For moderate to strong acceleration thresholds, minivans have consistently lower rates of acceleration, deceleration, and both types of lateral accelerations. For example, on average cars experience 2.5 times the ≥ 0.5 g lateral accelerations that minivans do.

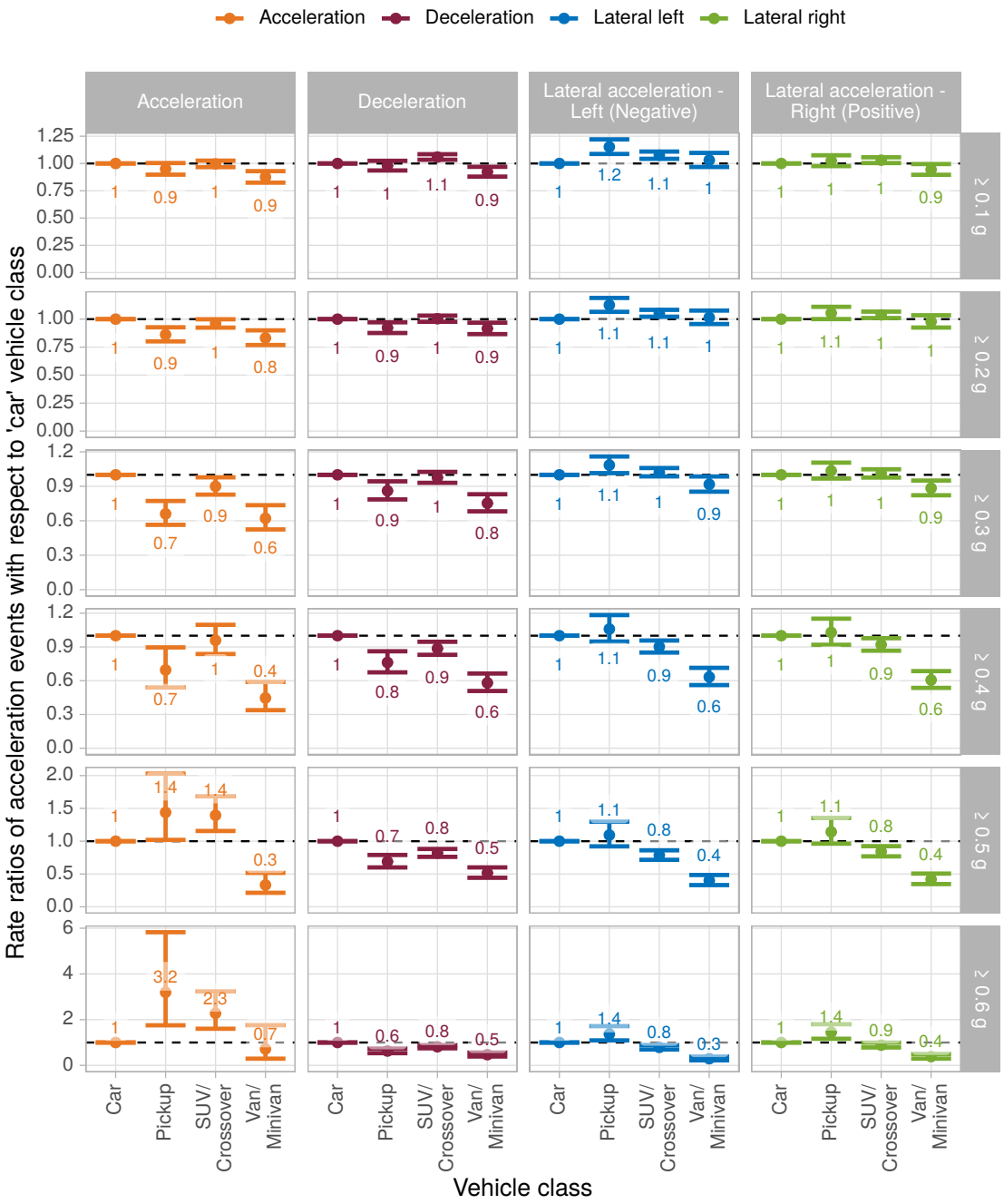


Figure 5.4: Comparing the effect of vehicle class on incident rate ratios of the 4 acceleration types at the six thresholds with respect to vehicle class "Car".

In general, vehicle class does not have as big an effect as the roadway category or the driver age. However, many comparisons show significant differences in various vehicle classes. Since

there are many factors that distinguish vehicles, such as body or engine specifications, some of the acceleration rate trends may be unclear due to the high dimensional space being represented by just four subcategories. Also, since some type of drivers may be more likely to drive a particular vehicle classes, there may be confounding factors that are outside the scope of this study.

5.5.4 Effect of Collection Site Location

Collection site location has significant effect on the relative rates for some comparisons. Figure 5.5 shows the effect of collection site on the rate ratios with respect to the "Tampa, FL" site. The collection site location can be seen to represent quite a few building blocks that constitute driving complexity in and around a location. This is in addition to the roadway speed category which is already a fixed effect in this analysis. For example, the layout of the location, whether it has curvy roads or perpendicular grids, proportion of rural versus urban areas, and average traffic conditions. Therefore, even when there are significant differences between locations, the mechanism causing these differences would be better understood by breaking out the driving into such constituent elements. However, as that is outside the scope of this analysis, only some significant observations are reported below:

- Bloomington, Indiana and State College, Pennsylvania have consistently lower rates of longitudinal acceleration and deceleration when compared to the other locations.

- Raleigh, North Carolina has the highest lateral acceleration rates for mild to moderate thresholds.

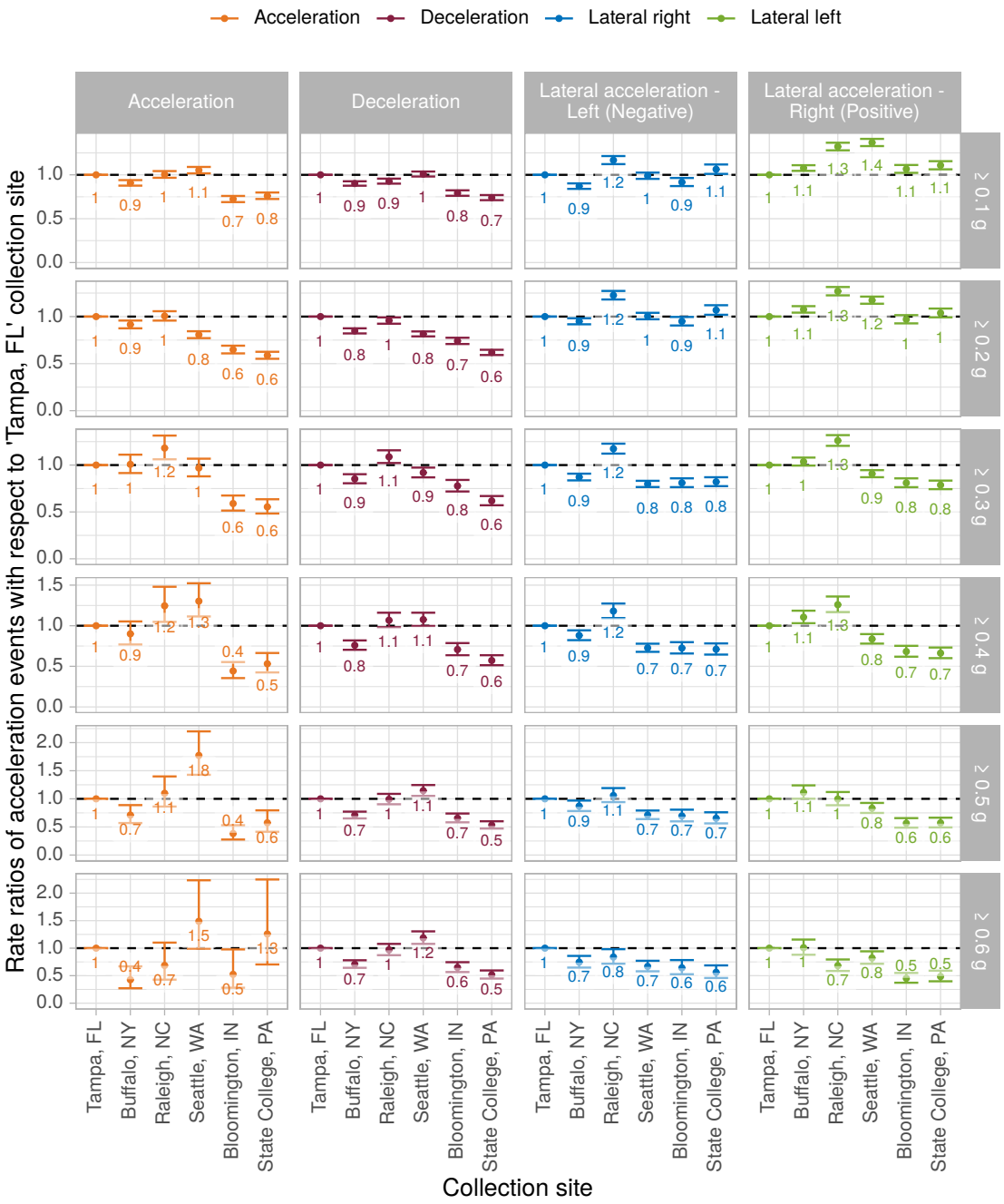


Figure 5.5: Comparing the effect of location on incident rate ratios of the four acceleration types at the six thresholds with respect to location "Tampa, FL".

5.5.5 Effect of Driver Gender

For most comparisons, driver gender does not have significant effect on the relative acceleration rates. However, for some comparisons, there is a difference between male and female drivers. Figure 5.6 shows the effect of driver gender on the rate ratios with respect to female drivers. Some important observations are:

- For low thresholds, which represent most of driving, the rates are very similar across driver genders with slightly lower deceleration rates for male drivers.
- For higher thresholds, male drivers experienced higher longitudinal acceleration rates, equal deceleration rates, and slightly higher lateral acceleration rates.

5.5.6 Comparing the Influence Various Fixed Effects

Figure 5.7 compares the influence of various fixed effects on the rate of different acceleration types with a threshold of $\geq$0.5 g. The x-axis represents log of the ratio between the maximum rate and the minimum rate within the subcategories of each fixed effect shown on the y-axis. For example, in case of decelerations $\geq$0.5 g, roadway speed category has the largest range with the rate experienced on $\leq$ 30 mph roadways being 21 times the rate experienced on $60 - 80$ mph roadways. Similarly, the ratio of maximum to minimum rate within each category is calculated for all acceleration types. It should be noted that the subcategories with maximum or minimum rates could be different across acceleration types, as is shown in the figure.

The primary purpose of this figure is to illustrate the comparative magnitude of the effect size across various fixed effects. Even though this figure compares rates for a threshold of 0.5g, similar comparisons can be made at other thresholds using the supplementary data provided

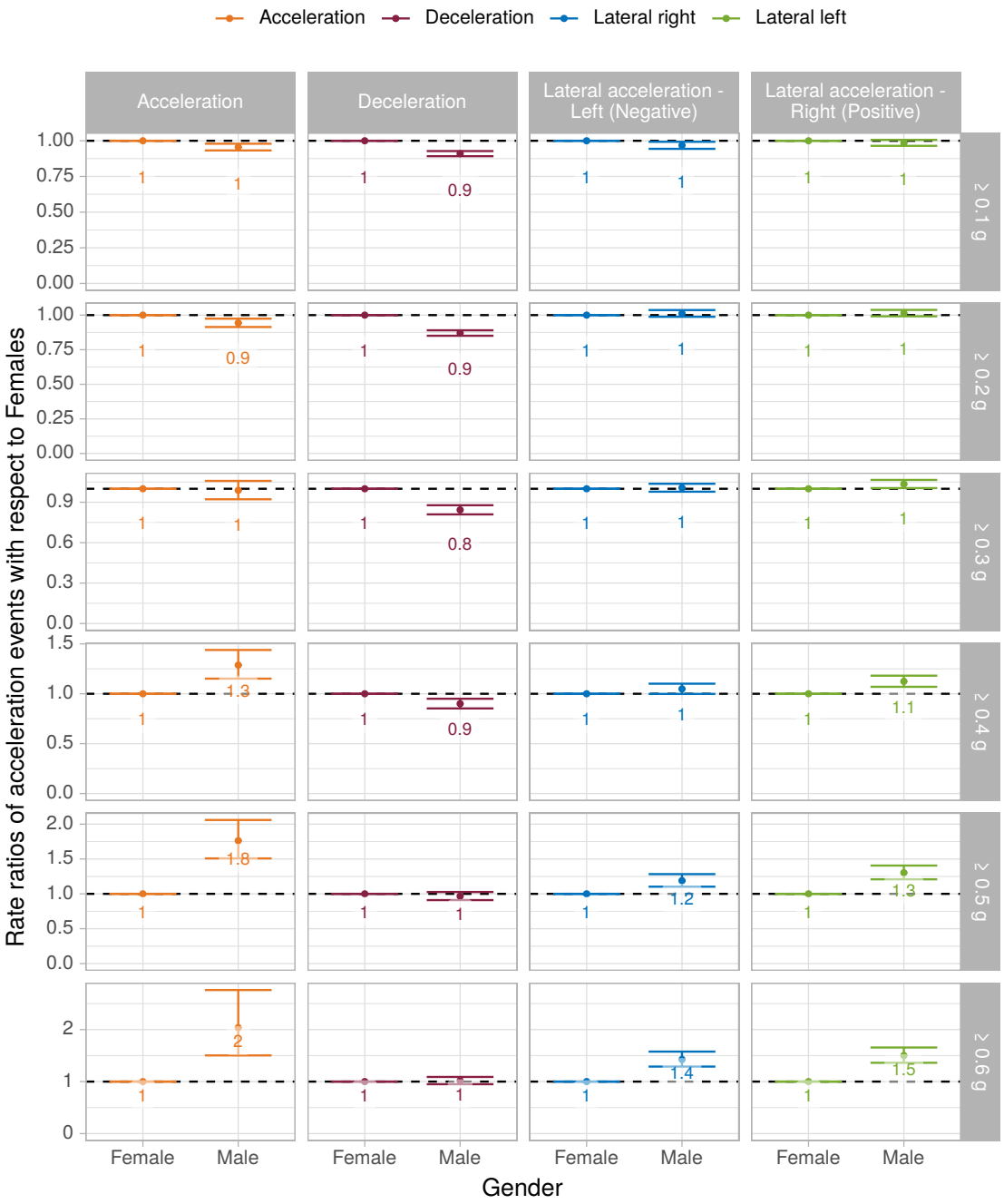


Figure 5.6: Comparing the effect of driver gender on incident rate ratios of the four acceleration types at the six thresholds with respect to driver gender "Female".

with this paper. There is consensus that driving within the acceleration values of ±0.05g is considered steady state driving [10, 100]. However, various studies have used different

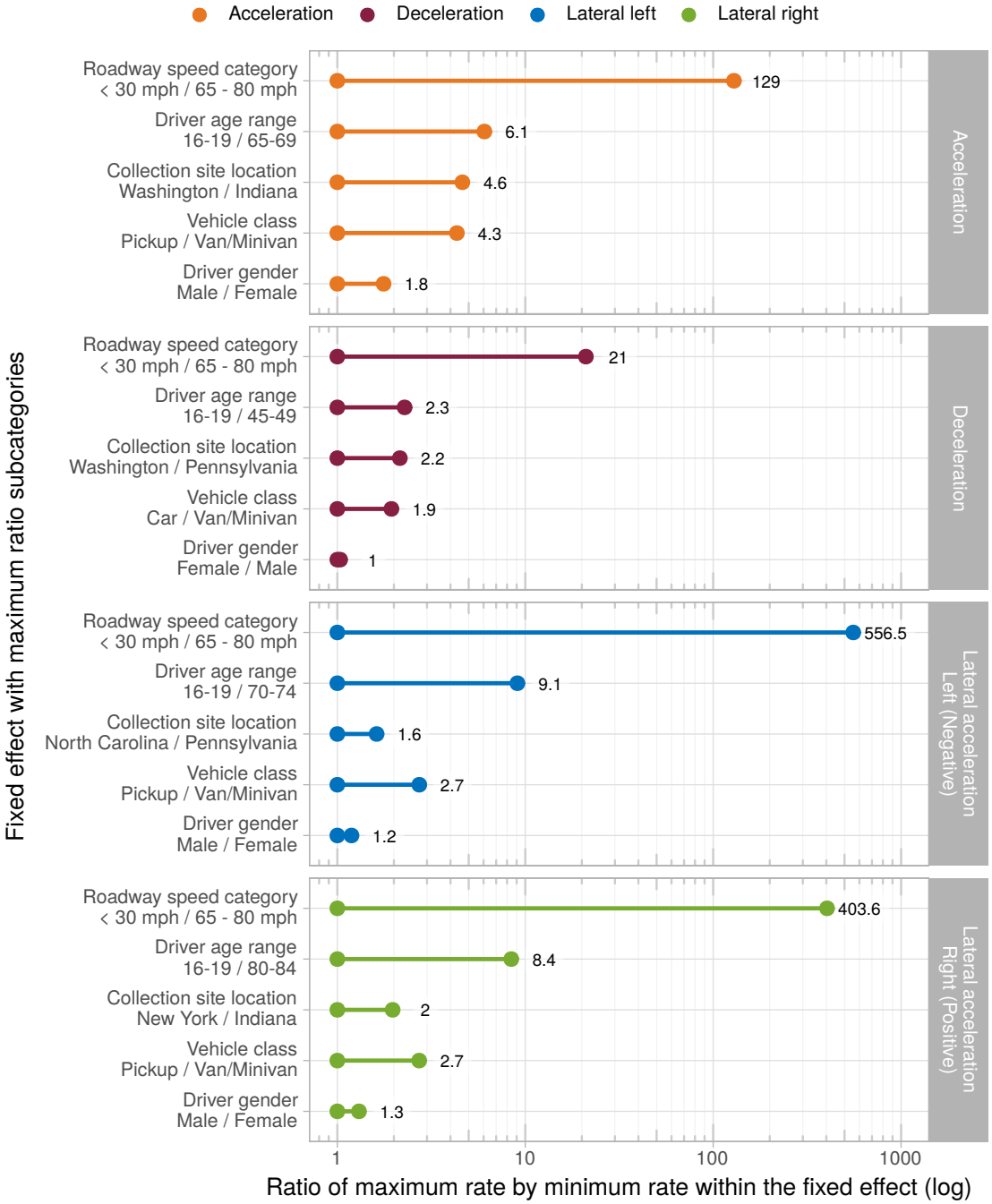


Figure 5.7: Comparing the ratio of maximum to minimum rate within each fixed effect for a threshold of 0.5 g.

thresholds for classifying harsh acceleration events [65]. A threshold of around 0.5g has often been used in transportation studies to classify harsh braking and lateral accelerations

events. For example, the following studies used a threshold between 0.4 to 0.6 g to classify some form of a harsh acceleration event. [5, 18, 33, 46, 47, 57, 69, 114, 116]. Therefore, Figure 5.7 was created for this particular threshold.

For each of the four acceleration types, roadway speed category has the most significant effect with the same subcategories having the highest and lowest rates. The ratio is 129 for longitudinal acceleration, 21 for longitudinal deceleration, 557 for left leaning lateral acceleration, and 403 for right leaning lateral acceleration. Age range is consistently the second most important factor determining the rate of all four acceleration types ≥ 0.5 g. Collection site location and vehicle class are the next two important factors but their order differs between the longitudinal and lateral acceleration types. Finally, driver gender has the smallest effect across all four acceleration types ranging between 1 and 1.8.

This study advances the understanding of acceleration behavior in three new ways. First, it simultaneously accounts for important factors such as roadway properties, driver characteristics, and vehicle classification. This ensures that the resulting differences are appropriately assigned to various factors and are not due to coincidences of data collection. Second, it simultaneously analyzes all four acceleration types. This ensures that the same methodology is applied to longitudinal as well as lateral accelerations and therefore the differences can be meaningfully compared. Finally, it analyzes all acceleration behavior at multiple thresholds ranging from ≥ 0.1 g to ≥ 0.7 g. This ensures that the behavior is investigated for mild, medium, and harsh events. Having all three regions enables researchers to examine traffic behavior, driving style, and safety critical events.

5.6 Conclusions

Acceleration epochs represent a change of state in driving. Be it the change of speed during longitudinal accelerations or the change of direction in case of lateral accelerations. The study of accelerations, their magnitudes, and their frequencies can reveal important insights into factors affecting driving. A comprehensive examination based on a large scale study and analyzing multiple factors has been missing from the literature. This analysis has quantified the simultaneous effects of various roadway, driver, and vehicle factors on the rate of acceleration epochs at various thresholds. It is based on SHRP2 NDS and the surface accelerations reference which are large scale, highly diverse, and context rich datasets. Generalized mixed effect models were used to predict the rates of acceleration epochs at various thresholds with the above mentioned factors as fixed effects and the vehicle-driver combination as a random effect. The regression coefficients from these models were then used to calculate incident rate ratios between subcategories of each fixed effect.

The analysis showed that roadway speed category had the strongest effect on rate of occurrence for mild as well as strong accelerations, with the maximum difference being as large as two orders of magnitude between 65 - 80 mph and $\leq$30 mph roads. Driver age also plays a major role in how often drivers experience high-g acceleration, deceleration, and lateral acceleration epochs. For some comparisons such as strong lateral accelerations, teenage drivers could experience the same type of epoch 10 times as often as older drivers in their 70's. Vehicle class also shows significant effect with minivans providing consistently lower rates than other vehicle types. The collection site location showed significant effect for some comparisons but to better understand the underlying mechanism, the driving would need to be broken out by its constituent elements. Driver gender had the smallest effect with most comparisons showing nearly identical rates.

This study will be beneficial to a number of fields. Vehicle system engineers focused on ADAS and AD systems will be able to use the results presented in Figures 5.1 - 5.7 to inform the design of their systems to better represent or respond to real world driving. Safety researchers and transportation regulators will be able to use these results to develop targeted programs for at risk drivers using anomalous driver detection techniques. Roadway designers could use these data to recognize accelerations that are atypical for a given category of roadway. Insurance companies and vehicle cohort operators will also be able to use these results to compare their populations with a representative sample and draw conclusions about driver behavior and driving risk.

This study opens up several possibilities for future work. To better understand the role of vehicle characteristics in acceleration behavior, the vehicle classification could be broken into several constituent factors such as weight, power, and length. Similarly, to better understand the role of location, the current categories could be expanded to include better indicators of driving complexity such as road curvature, average traffic density, intersection density, and road width. Including such factors in the model will help researchers further understand what causes certain acceleration behaviors.

Chapter 6

Extracting Acceleration-Based Driving Styles and Determining Their Relationship to Crash Risk

Ali, G., McLaughlin, S., & Ahmadian, M. (2022). Acceleration-Based Driving Styles and Their Relationship to Crash Risk. Manuscript in preparation.

Abstract

This study offers a novel data-driven approach to understanding real world acceleration-based driving styles. An unsupervised clustering algorithm was used to find four different driving styles in a cohort of 3,489 drivers. First, rates of lateral and longitudinal acceleration epochs experienced on $\leq$ 30 mph roadways were calculated at thresholds of $\geq$ 0.3 g, $\geq$ 0.4 g, $\geq$ 0.5 g, and $\geq$ 0.6 g. These rates are more likely to represent long term driving style instead of short term driving behavior influenced by transient factors such as traffic or environmental conditions. These metrics were analyzed through a clustering algorithm to distribute the cohort into four clusters, named "*low*", "*mid A*", "*mid B*", and "*high*" based on the relative frequency of acceleration epochs. The relationship between driving style and safety was established by comparing the crash rates and the crash plus near-crash rates of each group.

Statistically significant differences in crash risk were observed between the clusters using Tukey's honest significance test. *Low* was the safest driving cluster followed by *mid A*. *High*, followed by *mid B*, were the riskiest driving styles. The differences in driving styles were also reflected in speeding behavior, with the *high* cluster driving a larger proportion of their mileage above speed limit as compared to the *low* cluster. The centroids representing the different driving clusters have been provided, making the results of this study implementable in various intelligent vehicle systems, such as Automated Driving Systems (ADS) or advanced driver assistance systems (ADAS). This study fills important gaps in our understanding of acceleration-based driving styles and their correlation with crash risk. It introduces a new bottom-up approach of extracting long term driving style without assigning subjective meaning to acceleration epochs of various magnitudes. The results from this study will help ADS system designers to emulate human driving styles and therefore facilitate ADS adoption. Additionally, this study will be beneficial for naturalistic study-based research as it enables researchers to select certain driving styles based on safety and comfort requirements.

6.1 Introduction

Preventable motor vehicle crashes cause immense harm to life and property. Every year, thousands of people die, millions are injured, and nearly a quarter of a trillion dollars are lost just in the United States [20, 139]. In over 90% of these crashes, the driver is a primary contributing factor [34, 117]. Over many years, efforts have been made to reduce the risk associated with driving. On the human side, identifying risky drivers and designing intervention programs to better train them has shown some success [48, 101, 131]. However, one of the leading frontiers of reducing driving risk is the development of intelligent vehicle systems such as Automated Driver-Assistance Systems (ADAS) and Automated Driving Systems

(ADS). These systems have the potential to considerably reduce driving risk by taking over many or all driving tasks from the human driver and achieving a higher level of performance. To drastically reduce crashes, there is a need for safer drivers, better transportation safety systems, and wider adoption of such tools.

The design of intelligent transportation systems, however, also introduces many challenges. For wide scale adoption, these systems need to emulate human driving styles that are safe as well as comfortable. This is to not only ensure that such systems meet the driver and passenger expectations but also that these systems are made safer by better estimating the behavior of other road users [17, 19, 37, 75, 83, 106, 137, 147]. To achieve these goals, transportation researchers have continued to try to quantify driving styles from an engineering as well as human factors and psychology perspectives.

Even though differences in driving styles have been studied since the 1940s, a standard definition applicable across fields does not exist [83, 109, 129]. However, there is some agreement on what constitutes driving style. First, style is different from driving behavior, which can depend on several temporal factors, such as roadway speed limit, road type, traffic, weather, time of day, and various driver factors. Driving style is habitual, less likely to change over time, and therefore, a stable aspect of driving behavior [109]. Second, driving style differs systematically between drivers or groups of drivers [109]. Finally, an individual driver can exhibit multiple driving styles, depending on conditions, such as road type, speed limit, or trip purpose [109]. Therefore, driving style, as understood in this study, represents certain aspects of driving behavior that remain stable over a long time and differ systematically across groups of drivers. In terms of driving tasks, driving style can be broken into three levels: operational, such as steering and acceleration habits; tactical, such as choice of speed and headway; and strategic, such as habitual route choice [89, 109]. This study focuses on the operational driving style by analyzing longitudinal and lateral acceleration

behavior.

Various aspects of driving style have been examined through data collected from observational driving studies. These include speeding, acceleration, jerky driving, headway preferences, and curve negotiation [14, 41, 44, 81, 83, 88, 109]. However, to the best of the author's knowledge, acceleration-based driving styles have not been thoroughly examined. This is especially true from an implementational perspective, as none of the existing studies provides actionable output that intelligent vehicle systems could use to emulate human driving styles. For example, what are the magnitudes and frequencies of accelerations produced by various types of driving styles? This study aims to fill this gap by providing acceleration-based driving styles that can be directly used by vehicle systems.

6.2 Literature Review

Three recent survey papers summarize advances in the study of driving styles. Sagberg et al. primarily focus on the driver and discuss the major findings from a human factors and psychology perspective [109]. Martinez et al. discuss recent results from an engineering perspective, focusing on the application of driving styles to intelligent vehicle control [83]. Meiring and Myburgh focus on the algorithmic perspective, considering the literature for its potential to expose methods of uniquely identifying drivers [88]. There are interesting differences in how various fields analyze the problem. For example, the relationship between driving style and fuel efficiency is restricted to the engineering side of the discussion whereas driver aggressiveness and its underlying motivation is mostly restricted to the psychology side. However, there is also considerable overlap in the data and methods being discussed in these papers. The next few paragraphs discuss how this study fits into existing literature based on the data collection methods, the signals being used, and the methods employed to

analyze driving styles.

6.2.1 Data Sources Used in Literature

Most existing investigations of driving style use data from one of three types of sources: self reported survey results, simulator studies, and observational data acquired from instrumented vehicles [83, 88, 109]. Self reported surveys are mostly in the form of driving style questionnaires designed with the explicit aim of measuring driving style [64, 126, 149]. However, these measures mostly focus on the driver's mental state and usually do not contain actionable output that could be used to design vehicle systems to emulate desirable characteristics of human driving. Driving style studies that use simulator data have the advantage of being able to analyze specific scenarios, with control for a variety of factors that affect driving behavior. Some driving simulators such as the National Advanced Driving Simulator at the University of Iowa have also incorporated kinematic feedback mechanisms that make the driving more realistic [111]. However, most simulator driving differs considerably from real world experiences in a number of important aspects. First, the accelerations experienced in a simulator may not be fully realistic. Second, in most simulator-based driving studies, the participants are in a novel vehicle for a relatively short amount of time. Therefore, using simulator-based driving studies to understand long term acceleration-based driving style is not ideal.

There have been two types of observational studies based on instrumented vehicles. The first type are the controlled field studies that usually have a single vehicle shared by multiple participants driving on a test track or a predetermined public roadway loop. The second type are naturalistic driving studies that observe a participant in their natural setting by instrumenting their car and letting them drive it for a long period of time. Large scale naturalistic

driving studies are best suited for driving style determination because long term habitual style can be differentiated from short term behavior. Also, the relatively high number of participants makes the determination of group differences much more meaningful. Data from the Second Strategic Highway Research Program Naturalistic Driving Study (SHRP 2 NDS) were used for this study. The SHRP 2 NDS enrolled over 3,500 participants and collected 34.5 million miles of driving data over three years by instrumenting the participants' own vehicles.

6.2.2 Methods Used in Literature

From a methodological perspective, driving style recognition is usually achieved through one of three types of techniques: rule-based, model-based, and machine learning-based [83]. Rule-based methods usually assign some subjective meaning to a particular type of driving that could be defined through a rule. For example, decelerations ≥ 0.4 g are often considered harsh braking events and their prevalence has been used to define driving styles. [122] and [92] have used such rule-based methods to label maneuvers as aggressive by comparing acceleration and jerk signals to specific thresholds. To include a higher number of variables, simple rules based on predefined thresholds are often combined with fuzzy logic maps [38, 48, 113]. Rule-based methods have the advantage of being easy to interpret but the driving style discovery can be heavily influenced by subjectively defined thresholds.

Model-based approaches define driving style through a set of equations of pre-defined characteristics [83]. The parameters of the models are tuned using real world or simulation data and usually require extensive data collection and labeling. One of the major drawbacks of such methods is that validation across all the parameter ranges can be prohibitively expensive.

Machine learning-based approaches can be subdivided into two categories: supervised and

unsupervised learning algorithms. Supervised learning algorithms such as support vector machines (SVM), k-nearest neighbor (kNN), and neural networks have been used to classify driving styles in multiple studies [16, 66, 74, 124, 134, 145, 148]. However, such algorithms require labeled data as ground truth. Since driving style is not well understood, getting labeled data can be prohibitive because of the subjective biases being introduced and the cost of generating a large enough expert annotated dataset. On the other hand, unsupervised algorithms such as k-means do not require labeled data and the classification is achieved by statistical analysis of the inputs [91].

In this study, a combination of rule-based metrics and machine learning is used to recognize driving styles. An unsupervised machine learning algorithm is employed on measures that were created using domain knowledge-based rules. Frequencies of lateral and longitudinal acceleration at increasing thresholds were calculated for low speed roads and analyzed using a k-means clustering algorithm. This was to ensure that domain knowledge-based rules guided the discovery of driving styles but did not assign subjective meaning without first assessing the safety impact and relative distribution of such measures in the population.

6.2.3 Significane of Work

This study fills a number of gaps in the determination of acceleration-based driving styles and their relationship to driving risk. First, it uses a large scale, diverse, and context rich dataset collected over multiple years. This is essential in recognizing driving styles in a population instead of a small number of drivers. This condition is also important to differentiate long term driving style from general driving behavior. Second, it ensures that the various drivers are compared in similar driving domains, assuring that the effect of roadway properties on driving styles is considerably diminished. Third, it combines domain knowledge-based

rules with the unsupervised machine learning algorithms to make meaningful comparisons without levying subjective judgment on a particular threshold. Fourth, this driving style analysis presents actionable output that would help intelligent vehicle systems to emulate different human driving styles. Finally, the work correlates driving styles with indicators of driving risk, such as crash rates, near crash rates, and speeding behavior. Inclusion of this correlation is arguably a required consideration prior to implementation of any driving style related systems.

6.3 Data

6.3.1 Data Source

The SHRP 2 NDS is an ideal dataset for this study because it meets four important criteria essential to effectively recognize and quantify driving styles. First, it is large scale and collected over an extended period of time, meaning that long-term driving style can be differentiated from short-term driving behavior. The SHRP 2 NDS is the single largest collection of naturalistic driving and represents 34.5 million miles collected from 3,500 participants over three years [35].

Second, the driving sample selected has a large number of diverse participants so that various ages, genders, vehicle classes, and locations are represented [11, 12, 35]. This prevents a small group of individuals or factors from predominating the quantification of driving style.

Third, the data collected is rich in context, allowing various contributing factors to be accounted for or eliminated. SHRP 2 NDS collected data from an extensive suite of sensors, such as accelerometers, GPS, gyroscopes, vehicle CANs, and multiple camera feeds. This allows association of roadway properties to driving epochs and facilitates orthogonal validation

of various driving events.

Finally, the SHRP 2 NDS dataset has been thoroughly searched for crashes, near-crashes, and other safety critical events during previous studies [34, 35, 53, 104]. An algorithm analyzed signals of acceleration, yaw rate, vehicle traction and stability control, etc. looking for and flagging signatures that may represent safety critical events. Every epoch that was flagged was then reviewed by a team of naturalistic driving experts who determined whether the epoch was a crash, near-crash, other safety critical event or a false positive flag. This adds another important facet to the data that is used in this study to compare various driving styles on the basis of driving risk.

6.3.2 Data Preprocessing

To study acceleration-based driving style from SHRP 2 NDS, the raw data first needed to be processed to extract and summarize all longitudinal and lateral acceleration epochs. This was done through the creation of the Surface Accelerations Reference, which cataloged over 1.2 billion acceleration epochs in the SHRP 2 NDS [6, 7]. An algorithm analyzed longitudinal and lateral acceleration, yaw rate, speed, and other vehicle signals from each trip to find all detectable lateral and longitudinal acceleration epochs. An acceleration epoch was defined as the region in time when the acceleration signal had a continuous value above 0.01 g or below -0.01 g. Various signal processing techniques were used to clean signals, correct sensor artifacts, and augment multiple data sources, such as roadway properties, using map-matching [7, 87]. Each acceleration epoch was summarized into a data point with over 20 metrics representing statistical measures of acceleration, vehicle speed, driver input, and roadway properties.

All the epochs in a driver's history were then used to create a statistical driving profile

consisting of rate based measures at various magnitude thresholds for lateral and longitudinal acceleration. These measures represent the rate at which a driver experiences a certain type of acceleration epoch with a maximum magnitude above a certain threshold on a given road type. For example, the rate of "decelerations $\geq$ 0.5 g" on "$\leq$ 30 mph roadways." These measures were calculated for all four types of acceleration from 0.1 g to 1 g in intervals of 0.1 g.

Even though other types of measures were also calculated, for the purposes of this study, only these rate-based measures were chosen. The physical significance of these measures is easier to understand, which is important for a driving style analysis. Since the driving styles identified during this study will be represented in terms of the centroids of clusters, it is essential to choose metrics that have a physical significance, are easily measurable, can be interpreted by drivers, and are implementable in intelligent driving systems.

6.4 Methods

Accelerations are one of the most easily perceivable factors affecting ride quality. Therefore, emulation of desired acceleration characteristics is a primary concern for intelligent transportation system designers. Longitudinal and lateral acceleration epochs represent a change in the state of the vehicle and the magnitude of the acceleration represents the rapidity of the change. Longitudinal acceleration and deceleration represent a change in driving speed whereas lateral accelerations represent a change in vehicle direction. Therefore, the frequency of epochs at various thresholds contains information about the driver's driving style, traffic conditions, roadway properties, and other external factors. However, to extract and compare driving styles in a population of drivers, it is important to separate the driving style from other externally influencing factors. This is achieved using a two fold strategy:

first, choosing data from similar roadway conditions, and second, choosing measures that are not dependent on short term external effects.

Figure 6.1 shows the distribution of SHRP 2 NDS participants' longitudinal and lateral acceleration rates. The x-axis represents the threshold magnitude from "≥ 0.1 g" to "≥ 0.7 g" and the y-axis represents the rate—i.e., the number of acceleration epochs with the maximum magnitude above that threshold per mile of driving on roadways with a particular speed category. The data are separated into longitudinal acceleration, longitudinal deceleration, lateral left leaning acceleration, and lateral right leaning acceleration, which are plotted adjacent to each other for each threshold. The data are also separated by roadway speed category, which are shown as the five facets in the plot. The roadway speed category is a Here Technologies digital map attribute that represents usual driving speeds on a particular roadway and is derived from speed limit and other contributing factors, such as curvature and traffic levels [59]. The distribution is visualized by plotting the rate of drivers representing the 5^{th} percentile, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, and 95^{th} percentile. The range between the 75^{th} percentile and the 25^{th} percentile represents the middle 50% of the driving population.

6.4.1 Choosing Appropriate Subset of Preprocessed Data

The primary objective of the data selection strategy is to separate drivers' driving style from extraneous factors such as roadway properties and traffic conditions. For example, are two drivers driving differently because their driving style is different or because they predominantly drove on different types of roadways and in different traffic conditions? The effect of these factors are considerably eliminated from the analysis by separately calculating rates for roads with different speed categories. This ensures that when drivers are compared for a

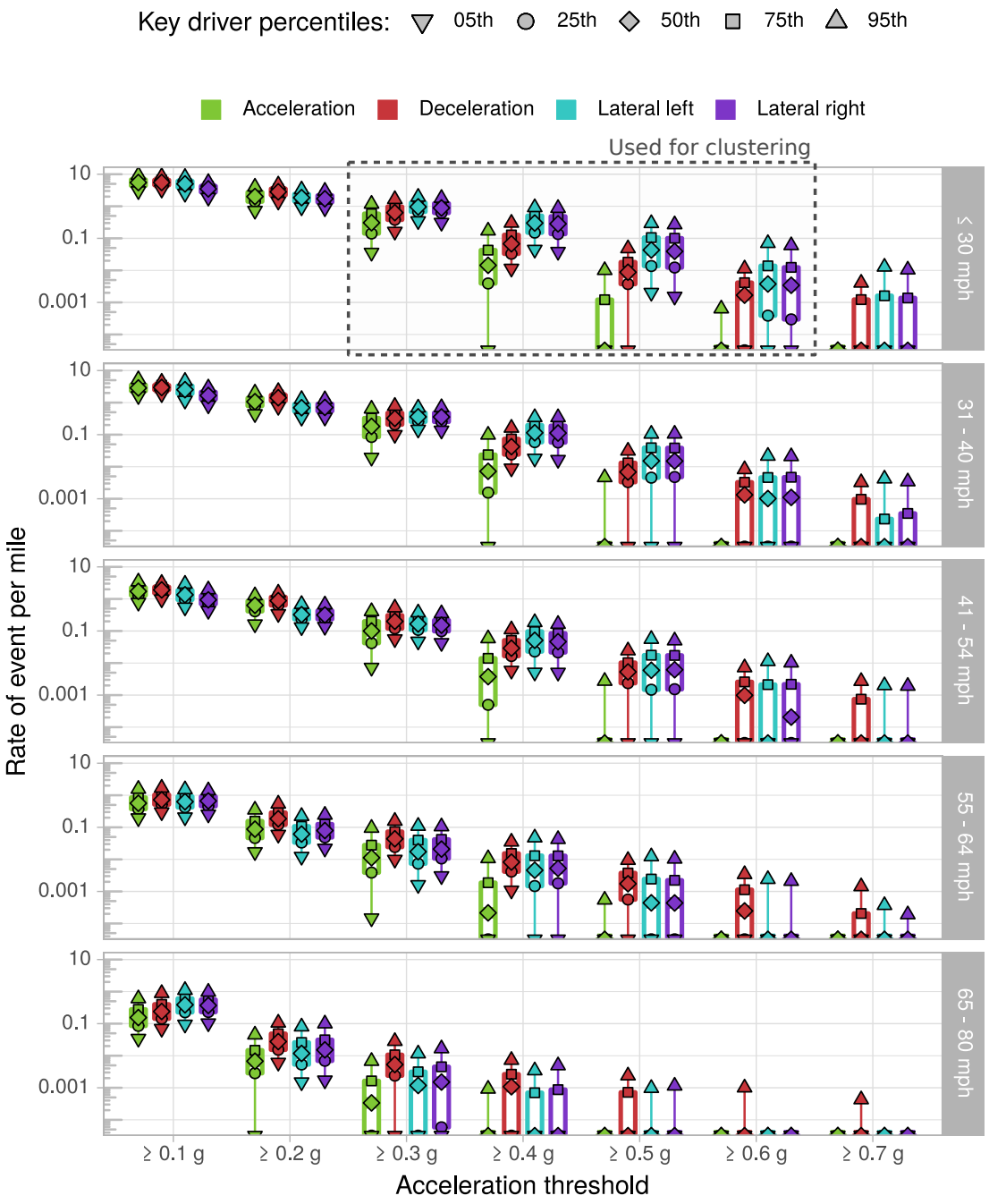


Figure 6.1: Distribution of acceleration rates for SHRP 2 NDS drivers at various thresholds separated by roadway speed category.

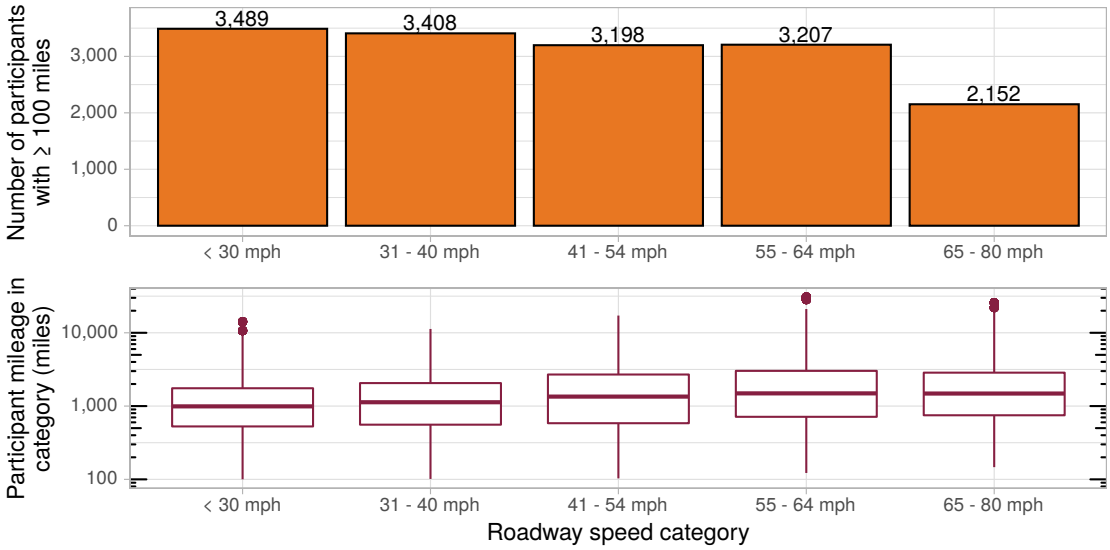


Figure 6.2: Distribution of mileage and the number of participants for each roadway category.

particular metric, the properties of the roadways on which those metrics were accumulated are taken into account. To ensure that the driving style determination is not heavily influenced by the differences in traffic around the participant's vehicle, rates for only certain magnitudes are included in the analysis.

6.4.1.1 Selecting Driving Domain

This study included only rates aggregated for $\leq$ 30 mph roadways from participants who drove more than 100 miles on those roadways. This deliberate selection serves several purposes. First, it ensures that the driving is compared on similar roadways. Second, an acceleration-based driving style is most perceivable on low-speed roads since the vehicle is frequently stopping, starting from stop, and taking sharp turns. Even though there is variance in acceleration values on high-speed roads, that variance is not as high as it is on low speed roads. This is clearly shown in Figure 6.1 for 65–80 mph roads where the median driver doesn't experience any lateral accelerations $\geq$ 0.4 g.

Third, $\leq$ 30 mph roadways had the highest number of participants with more than 100 miles of driving. Choosing this roadway speed category maximizes the number of participants for the analysis. Figure 6.2 illustrates the mileage distribution and the number of participants with at least 100 miles of driving in a particular roadway category. The 100-miles threshold in a roadway category ensures that the data are collected over several trips since it is unlikely that a participant drove more than a hundred miles on low speed roads in one trip. The median driver included in this study drove about a thousand miles on $\leq$ 30 mph roadways. This ensures that the driving style determined is a reflection of the driver's long term habits and not of short term externally influenced driving behavior. Finally, choosing data from only one roadway category significantly simplifies the interpretation of driving clusters determined using the unsupervised learning approach. It would be a notably harder problem to simultaneously account for driving styles across all five roadway categories.

6.4.1.2 Selecting Acceleration Thresholds

The rates at various magnitude thresholds for the four acceleration types are affected by different factors, which lets us separate driving style from other effects. The rates of acceleration epochs between $\geq$ 0.1 g and $\geq$ 0.2 g are mostly influenced by traffic conditions and roadway properties. Even though such acceleration epochs occur several times per mile for low speed roads, the rate of the 95$^{\text{th}}$ percentile driver is just 3–5 times that of the 5$^{\text{th}}$ percentile driver, as shown in Table 6.1. This shows that the range of these rates is well within one order of magnitude for 90% of the drivers in the SHRP 2 NDS. Therefore, these measures are excluded from this analysis to significantly eliminate effect of traffic on the determination of driving style.

Only the rates of longitudinal and lateral accelerations $\geq$ 0.3 g, $\geq$ 0.4 g, $\geq$ 0.5 g, and $\geq$ 0.6 g on $\leq$ 30 mph roadways are included in this analysis. Acceleration epochs with these mag-

Table 6.1: The distribution of longitudinal and lateral acceleration rates on ≤ 30 mph speed category roadways.

Roadway speed category	Threshold	Acceleration type	Driver percentile					Ratios	
			5th	25th	50th	75th	95th	75th:25th	95th:5th
≤ 30 mph	≥ 0.1 g	Acceleration	3.0E+00	4.4E+00	5.5E+00	6.8E+00	9.3E+00	1.5E+00	3.1E+00
≤ 30 mph	≥ 0.1 g	Deceleration	3.3E+00	4.6E+00	5.5E+00	6.6E+00	8.6E+00	1.4E+00	2.6E+00
≤ 30 mph	≥ 0.1 g	Lateral left	2.5E+00	3.9E+00	5.0E+00	6.1E+00	8.2E+00	1.6E+00	3.3E+00
≤ 30 mph	≥ 0.1 g	Lateral right	1.9E+00	2.8E+00	3.5E+00	4.3E+00	5.6E+00	1.5E+00	3.0E+00
≤ 30 mph	≥ 0.2 g	Acceleration	7.3E-01	1.4E+00	2.0E+00	2.7E+00	4.0E+00	2.0E+00	5.5E+00
≤ 30 mph	≥ 0.2 g	Deceleration	1.4E+00	2.2E+00	2.8E+00	3.5E+00	4.7E+00	1.6E+00	3.3E+00
≤ 30 mph	≥ 0.2 g	Lateral left	9.3E-01	1.4E+00	1.9E+00	2.4E+00	3.3E+00	1.7E+00	3.5E+00
≤ 30 mph	≥ 0.2 g	Lateral right	8.7E-01	1.3E+00	1.7E+00	2.2E+00	3.0E+00	1.6E+00	3.4E+00
≤ 30 mph	≥ 0.3 g	Acceleration	3.6E-02	1.4E-01	3.1E-01	5.8E-01	1.2E+00	4.1E+00	3.3E+01
≤ 30 mph	≥ 0.3 g	Deceleration	1.6E-01	3.8E-01	6.3E-01	9.7E-01	1.7E+00	2.6E+00	1.0E+01
≤ 30 mph	≥ 0.3 g	Lateral left	3.5E-01	6.7E-01	9.6E-01	1.3E+00	2.0E+00	1.9E+00	5.6E+00
≤ 30 mph	≥ 0.3 g	Lateral right	3.1E-01	6.2E-01	8.9E-01	1.2E+00	1.8E+00	2.0E+00	5.8E+00
≤ 30 mph	≥ 0.4 g	Acceleration	0.0E+00	3.9E-03	1.4E-02	4.3E-02	1.7E-01	1.1E+01	-
≤ 30 mph	≥ 0.4 g	Deceleration	1.1E-02	3.3E-02	6.7E-02	1.3E-01	3.0E-01	3.9E+00	2.6E+01
≤ 30 mph	≥ 0.4 g	Lateral left	4.5E-02	1.5E-01	3.0E-01	5.0E-01	9.2E-01	3.3E+00	2.1E+01
≤ 30 mph	≥ 0.4 g	Lateral right	3.8E-02	1.4E-01	2.8E-01	4.8E-01	8.7E-01	3.5E+00	2.3E+01
≤ 30 mph	≥ 0.5 g	Acceleration	0.0E+00	0.0E+00	0.0E+00	1.2E-03	1.0E-02	-	-
≤ 30 mph	≥ 0.5 g	Deceleration	0.0E+00	3.7E-03	8.8E-03	1.8E-02	4.8E-02	4.8E+00	-
≤ 30 mph	≥ 0.5 g	Lateral left	2.0E-03	1.4E-02	4.3E-02	1.1E-01	2.9E-01	7.7E+00	1.5E+02
≤ 30 mph	≥ 0.5 g	Lateral right	1.5E-03	1.2E-02	3.9E-02	1.0E-01	2.7E-01	8.2E+00	1.8E+02
≤ 30 mph	≥ 0.6 g	Acceleration	0.0E+00	0.0E+00	0.0E+00	0.0E+00	6.4E-04	-	-
≤ 30 mph	≥ 0.6 g	Deceleration	0.0E+00	0.0E+00	1.7E-03	4.1E-03	1.1E-02	-	-
≤ 30 mph	≥ 0.6 g	Lateral left	0.0E+00	3.9E-04	3.8E-03	1.4E-02	7.0E-02	3.6E+01	-
≤ 30 mph	≥ 0.6 g	Lateral right	0.0E+00	3.0E-04	3.4E-03	1.2E-02	6.0E-02	4.2E+01	-
≤ 30 mph	≥ 0.7 g	Acceleration	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	-	-
≤ 30 mph	≥ 0.7 g	Deceleration	0.0E+00	0.0E+00	0.0E+00	1.2E-03	4.1E-03	-	-
≤ 30 mph	≥ 0.7 g	Lateral left	0.0E+00	0.0E+00	0.0E+00	1.6E-03	1.3E-02	-	-
≤ 30 mph	≥ 0.7 g	Lateral right	0.0E+00	0.0E+00	0.0E+00	1.4E-03	1.0E-02	-	-

nitudes are more likely to represent driving style and less likely to be indicators of traffic or other external factors. The variance in driving style for longitudinal accelerations is mostly captured in rates for ≥ 0.3 g and ≥ 0.4 g. The median driver in SHRP 2 NDS never experienced a longitudinal acceleration of ≥ 0.5 g, as that is often beyond the capabilities of most vehicles. However, the longitudinal deceleration and both lateral acceleration components of driving styles are represented by all four thresholds. Even though an epoch with ≥ 0.5 g lateral acceleration or longitudinal deceleration is sometimes considered a hard braking or hard turning event, these occur frequently enough to be not included as indicators of driving style [114, 116]. For example, the 75^{th} percentile driver experiences a longitudinal deceleration of magnitude ≥ 0.5 g about once every 55 miles, and a lateral left or right acceleration once every 5 miles. This implies that 25% of the drivers in the SHRP 2 NDS experienced such events at similar or higher rates. Acceleration epochs ≥ 0.7 g were excluded, as the median driver never experienced them in any of the four acceleration types and therefore these epochs are not representative of driving style.

Therefore, for the purposes of this study, the following 16 variables were extracted for every participant and used in the determination of driving style.

- Rates of longitudinal acceleration epochs per mile at thresholds of ≥ 0.3, ≥ 0.4, ≥ 0.5, and ≥ 0.6 g on ≤ 30 mph roadways.
- Rates of longitudinal deceleration epochs per mile at thresholds of ≥ 0.3, ≥ 0.4, ≥ 0.5, and ≥ 0.6 g on ≤ 30 mph roadways.
- Rates of right leaning lateral acceleration epochs per mile at thresholds of ≥ 0.3, ≥ 0.4, ≥ 0.5, and ≥ 0.6 g on ≤ 30 mph roadways.
- Rates of left leaning lateral acceleration epochs per mile at thresholds of ≥ 0.3, ≥ 0.4, ≥ 0.5, and ≥ 0.6 g on ≤ 30 mph roadways.

6.4.1.3 Significance of Data Selection Strategy

This data selection approach offers a novel methodology of using domain knowledge to choose frequency-based metrics at varying thresholds without assigning qualitative judgment to them. This approach achieves four major objectives. First, it ensures that the drivers are compared in similar but generalizable driving domains. Second, it significantly reduces the influence of various extraneous factors such as traffic levels that may act as confounding factors in the driving style determination. Third, requiring greater than 100 miles of driving ensures that long term habits are represented instead of short term driving behavior that could change from one trip to another. Finally, it establishes the significance of the acceleration thresholds without levying a qualitative judgment on the metric. This allows the relative rate of accelerations for a participant compared to the driver population to determine the driver's classification. Since the underlying dataset is a collection of large scale, diverse, and real world driving, the clusters extracted by this methodology should be useful for many different applications.

6.4.2 *k*-means Clustering

In this study, k-means clustering is used to determine different types of acceleration-based driving styles within the SHRP 2 NDS population. k-means clustering is a unsupervised machine learning technique that aims to partition n observations into k clusters by choosing centroids that minimize within-cluster sum of squares. The number of clusters, k, is a chosen input to the process along with the the data in the form of n observations. The clustering process is completed in three steps. In the first step, the algorithm determines the cluster centroids by choosing k random observations within the total n observations. This step is the initialization step and can be affected by the choice of the random number generator

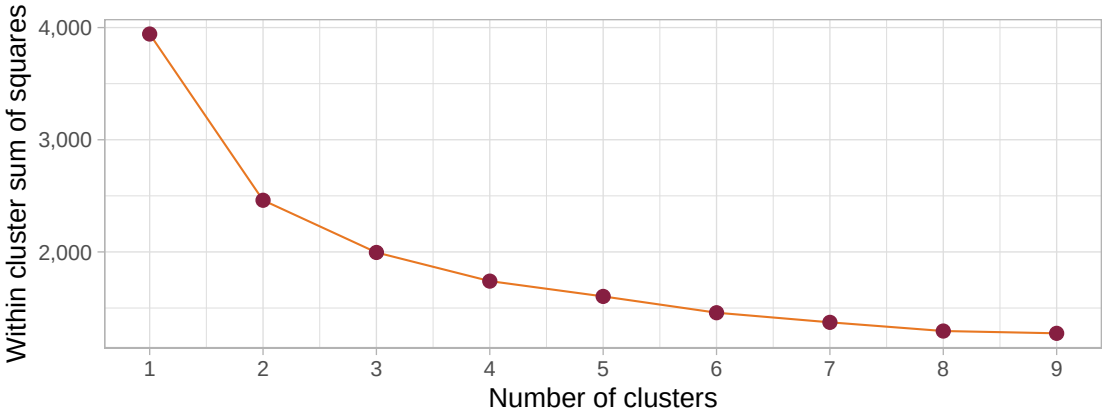


Figure 6.3: Within cluster sum of squares versus number of clusters.

used by the algorithm. In the second step, each observation is assigned a cluster based on the nearest cluster centroid. In the third step, the cluster centroids are calculated as the mean of all the observations within a cluster. This causes the centroids of the clusters to move in the Euclidean space formed of the same number of dimensions as the input data. The algorithm loops over steps two and three until the displacement of the centroids is below a certain threshold.

To effectively use k-means clustering, three considerations need to be examined. The first is k, number of clusters. Since k is an input to the algorithm and therefore a design decision of the analysis, its choice needs to be justified. The within-cluster sum of squares is a useful metric in deciding the appropriate number of clusters. Figure 6.3 shows the change in within cluster sum of squares for a varying number of clusters ranging from 1 to 9. The clearest knee is in the transition from two to three clusters. While additional mathematical improvement might be gained by retaining consideration of more clusters, to support the goal of human interpretable guidance on driving styles, four clusters will be used. This offers sufficient reduction in within cluster sum of squares while retaining enough coarseness that the driving styles can be differentiated and have physical significance.

It should also be noted that in this analysis, the number of clusters serves as an aid in

understanding different types of driving styles and not as a ground truth to be discovered. In other words, the driving styles in the population are not as cleanly delineated as the outcome of clustering algorithms may suggest. However, the centroids of the clusters will provide useful guidance in finding representational regions in the chosen metrics that could be emulated by intelligent driving systems.

The second consideration is how to address the rapidly expanding Euclidean space with the increasing number of dimensions, a phenomenon that is often known as "the curse of dimensionality." For higher dimension spaces, such as the metrics chosen for clustering, if the data are uniformly distributed, the measure of Euclidean distances can become meaningless. Since the cluster centers are based on Euclidean distances, it is essential to ensure that the analysis is not suffering from this problem. Figure 6.4 shows the variance explained by the principal components of the selected dataset. The explained variance is the ratio between the variance of the principal component and the total variance of the original dataset. The first six principal components explain 92% of the original dataset's variance.

Figure 6.5 compares the results of the clustering algorithm on two types of datasets. The top subplot shows the result of the k-means clustering algorithm on original data that was

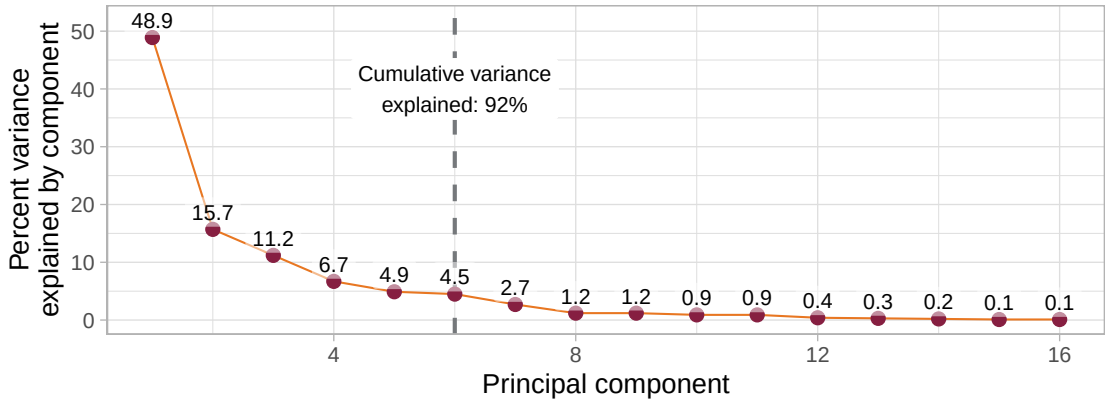


Figure 6.4: Variance explained by the principal components as a percentage of the total variance of the data.

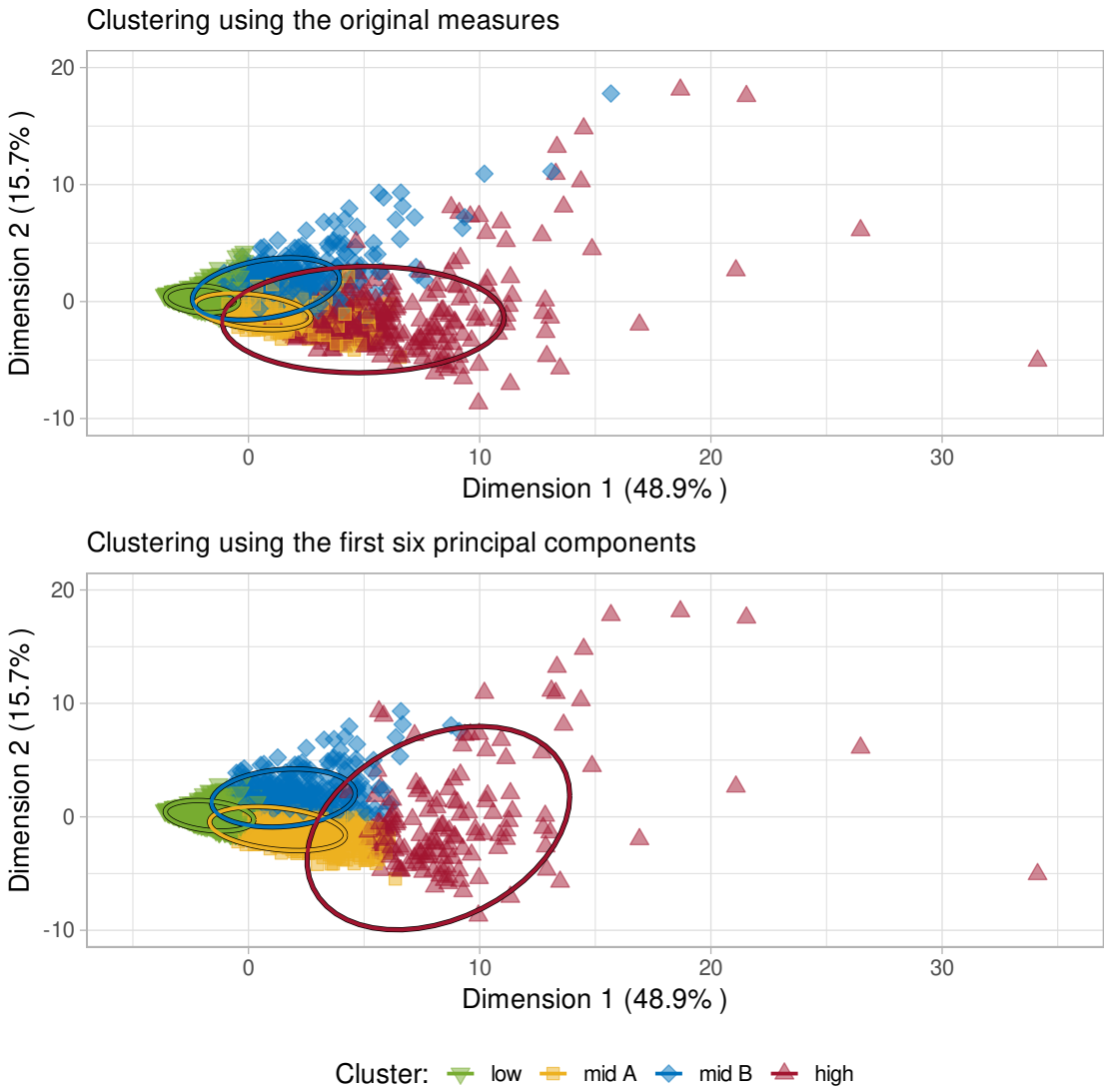


Figure 6.5: Comparison of clusters based on whether the k-means algorithm was operated on the 16 original measures or the first six principal components.

converted to principal components after clustering. The bottom subplot illustrates the result of the k-means clustering algorithm on the first six principal components of the data. Even though there are slight differences in the clustering results, the overall structure of the clustering seems similar. There are four clusters identified around similar centroids. This shows that the higher number of dimensions is not adversely affecting the clustering output. The final results of this analysis were based on the clustering algorithm that used the original

data and not the principal components because the centroids represent driving styles and, for interpretability, are best defined through the original measures and using the complete dataset.

Finally, the last consideration is the effect of the random seed used by the algorithm. Since the algorithm randomly assigns cluster centers during the initialization step, it is important to determine whether the cluster centers are robust to different initialization steps. This can be confirmed by changing the random number generator seed a few times and seeing how it affects the cluster centers. The centroids discovered in this analysis were robust to such changes.

In summary, the methodology of this study is a combination of data selection strategy and clustering approaches that identify acceleration-based driving styles. The data selection strategy ensures that drivers are compared on similar domains using metrics that reflect long term driving style instead of short term driving behavior. The chosen metrics also ensure that an appropriate spectrum of acceleration magnitude thresholds for all four acceleration types is included with minimal influence of subjective interpretation or even requiring imposing limits based on previous work. This sets up the unsupervised k-means clustering process to effectively identify similar groups of drivers. This process has been shown to choose the appropriate number of clusters, be resistant to the perils of high dimension Euclidean spaces, and be robust against the random components of the algorithm. This combined approach lets the relative frequencies of each magnitude for various acceleration types within the population determine the driving styles. Therefore, the cluster centroids representing the driving styles are data points located in the 16 dimensional "acceleration type – magnitude – frequency" space.

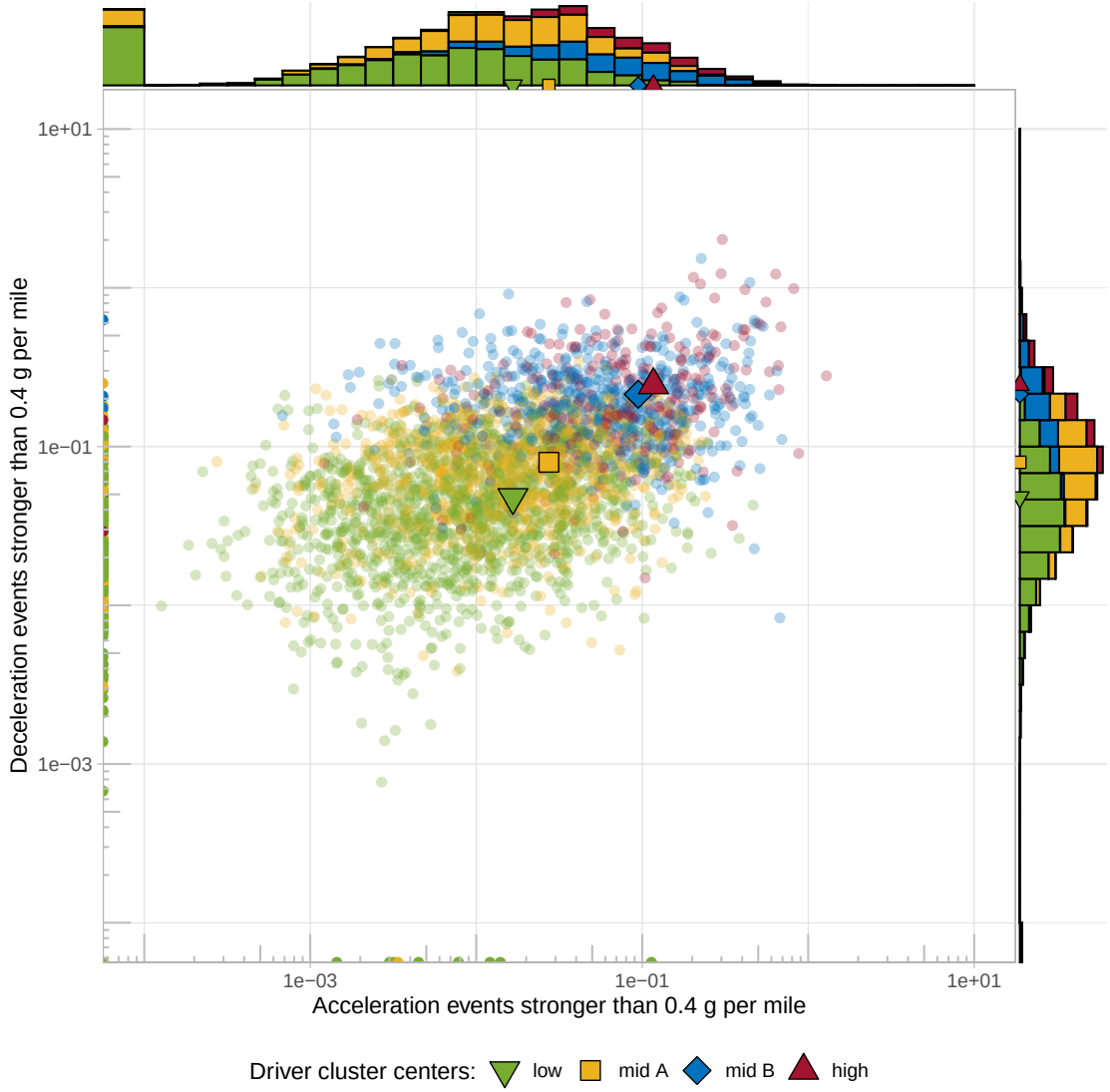


Figure 6.6: Scatter plot of $\geq 0.4g$ longitudinal acceleration and deceleration rates for SHRP 2 NDS drivers broken out into 4 clusters.

6.5 Results

In this analysis, a k-means clustering algorithm assigned 3,489 drivers to four clusters by analyzing 16 measures representing rates of four acceleration type epochs at magnitude thresholds of $\geq$ 0.3 g, $\geq$ 0.4 g, $\geq$ 0.5 g, and $\geq$ 0.6 g. These measures were extracted from driving on $\leq$ 30 mph roadways and only those participants who drove at least 100 miles on

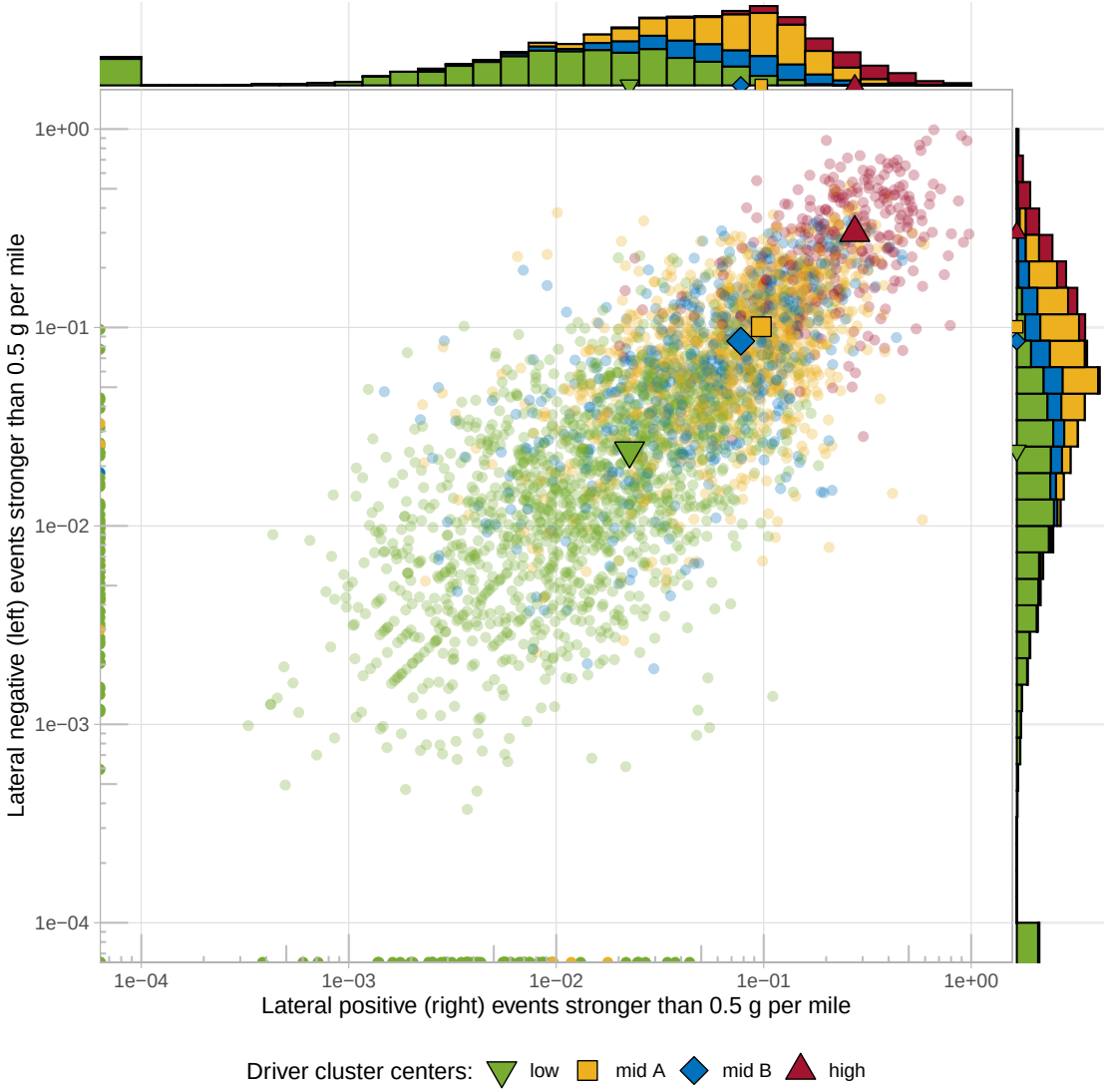


Figure 6.7: Scatter plot of $\geq 0.5g$ right and left leaning lateral acceleration rates for SHRP 2 NDS drivers broken out into 4 clusters.

such roads were included in the analysis. The data selection strategy ensured that long term driving style is the predominant information contained in these measures. The results of the analysis can be understood by observing the relative location of the various clusters along the 16 measures.

Figure 6.6 plots the driving cluster centroids and the associated participants in each cluster

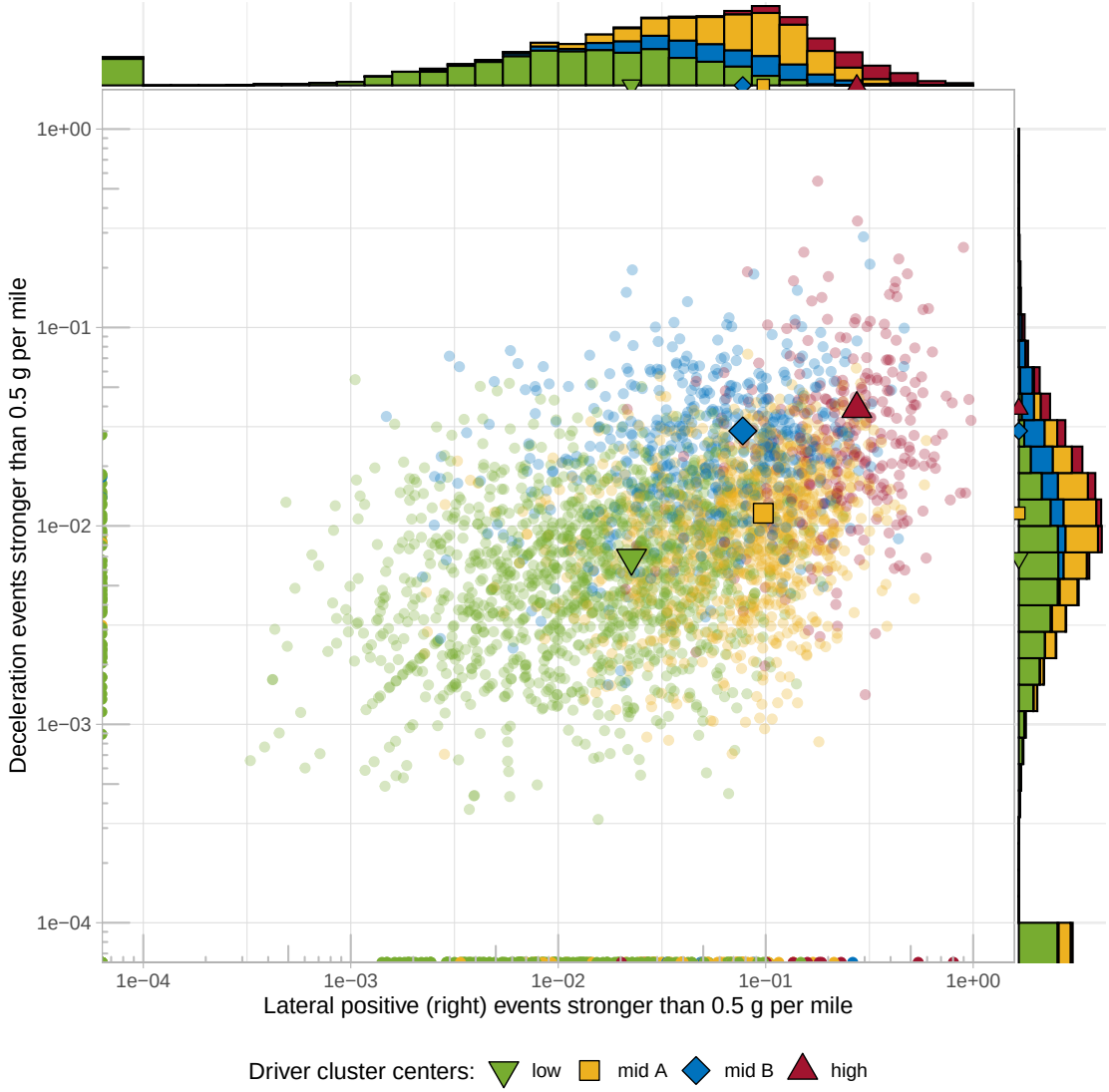


Figure 6.8: Scatter plot of $\geq 0.5g$ right leaning lateral acceleration versus longitudinal deceleration rates for SHRP 2 NDS drivers broken out into 4 clusters.

with rates of longitudinal acceleration epochs ≥ 0.4 g along the x-axis and longitudinal decelration epochs ≥ 0.4 g along the y-axis. It should be noted that this plot only represents 2 of the 16 total measures used in clustering. The four clusters identified have been named *low*, *mid A*, *mid B*, and *high* for reasons that will become clear in this section. The *low* and *high* clusters have the lowest and highest values for all measures respectively. The *mid A* and *mid B* clusters have different trends between lateral and longitudinal acceleration rates.

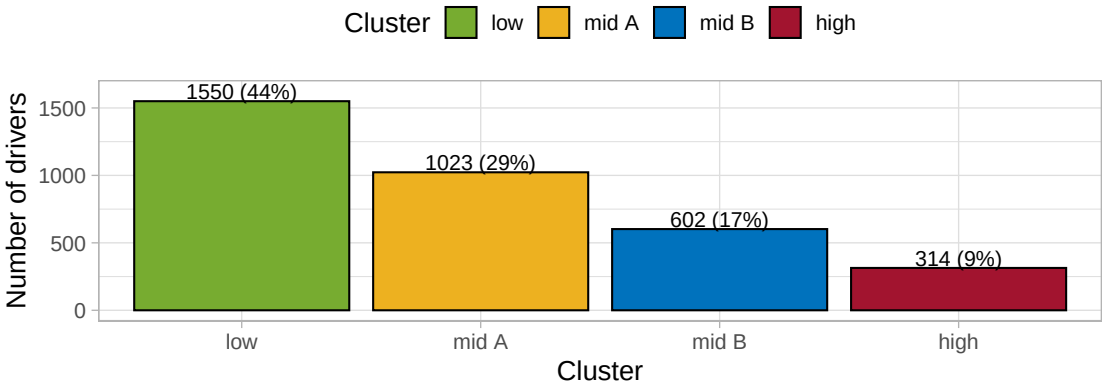


Figure 6.9: Proportion of the SHRP 2 NDS population in each of the driving style clusters.

Figure 6.6 is a combination of a scatter plot representing the values of the two variables and two histograms representing the distributions of the two variables. Both the axes have been transformed to the logarithmic scale with base 10. Since the rates of some of the measures are zero epochs per mile, the values of these measures are at -∞ and therefore have been counted in the extreme open bin in the histograms. This adjustment is essential to show the relative position of the various driving style clusters on the scatter plot as well as the histogram. As can be clearly seen in Figure 6.6, the value of longitudinal acceleration and deceleration increases log-linearly in the comparison of *low* to *mid A*, *mid B*, and *high* cluster centroids.

Figure 6.7 is similar to the previous figure, as it plots the driving cluster centroids and the associated participants in each cluster with rates of lateral right epochs $\geq$ 0.5 g along the x-axis and lateral left epochs $\geq$ 0.5 g along the y-axis. In this figure, all four cluster centroids have almost identical values for the two measures plotted (i.e., left and right accelerations). This symmetry between left and right lateral accelerations is missing from longitudinal acceleration and deceleration. Similar to the previous figure, *low* and *high* have the lowest and highest rates. However, the values of *mid A* and *mid B* have different trends for lateral acceleration than they do for longitudinal acceleration. In Figure 6.6, *mid A* was

significantly lesser than *mid B* but in Figure 6.7, *mid A* is slightly greater than *mid B*. This implies that when comparing *mid A* and *mid B* clusters, drivers belonging to *mid A* experience significantly lower longitudinal acceleration and deceleration rates but similar lateral left and right acceleration rates. This is further illustrated by Figure 6.8, which directly compares ≥ 0.5 g epoch rates of lateral right acceleration along the x-axis with longitudinal deceleration along the y-axis. Together Figures 6.6, 6.7, and 6.8 show the relative location of the cluster centroids along 4 of the 16 variables used for clustering. The names of these clusters have deliberately been chosen to only represent the relative frequencies and not be overly subjective in describing these participants' driving styles.

Figure 6.9 shows the number of drivers in each cluster. *Low* has the highest proportion (44%), followed by *mid A* (29%), *mid B* (17%), and *high* (9%). Figures 6.6, 6.7, and 6.8 also help in visualizing the distributions behind the cluster centroids and show that often for a particular measure, a participant may be closer to the centroid of a different cluster but the cluster assignment is based on the distance along all 16 measures. Therefore, to simplify the discussion, the acceleration-based driving styles derived through this analysis will mostly be discussed in terms of the location of the centroids and not the distributions around them. This will also be useful in determining the physical significance of each driving style and provide intelligent transportation systems clear data points for emulation. However, the actual driving population is a continuum in the "acceleration type – magnitude – frequency" space and it should not be assumed that it perfectly fits itself into these styles.

6.5.1 Driving Style Clusters and Their Physical Interpretation

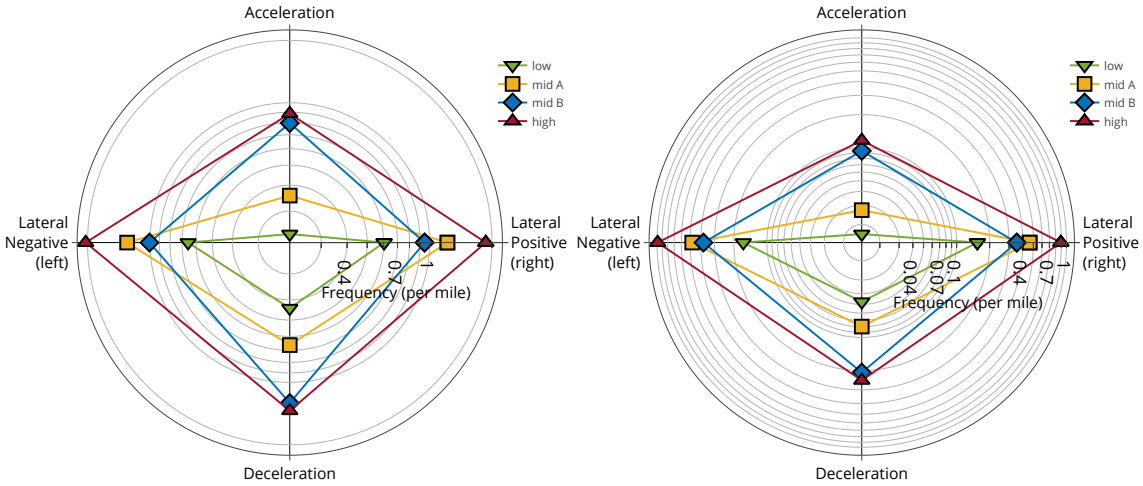
Figure 6.10a simultaneously displays the cluster centroids in terms of rates of epochs ≥ 0.3 g for all four acceleration types using a polar plot. Similarly, Figures 6.10b, 6.10c, and 6.10d

illustrate the cluster centroids for rates of epochs ≥ 0.4 g, ≥ 0.5 g, and ≥ 0.6 g respectively. These plots are especially useful in understanding the cluster centroids, as each axis of the polar plot coincides with the physical direction of the acceleration type, making the shape of the cluster physically meaningful. Table 6.2 displays the values of the cluster as well as the ratio of each value with respect to the *low* cluster. Therefore, the ratio column indicates the relative frequency of a particular type of acceleration epoch with respect to the *low* cluster.

To understand the driving style through cluster centroids and quantify the physical significance of their differences, the cluster centroids need to be evaluated from three perspectives. First, how do the centroids differ from each other at each threshold? Second, how do the centroids change as the magnitude of the thresholds increases? And finally, what is the difference in the values of each acceleration type within a cluster?

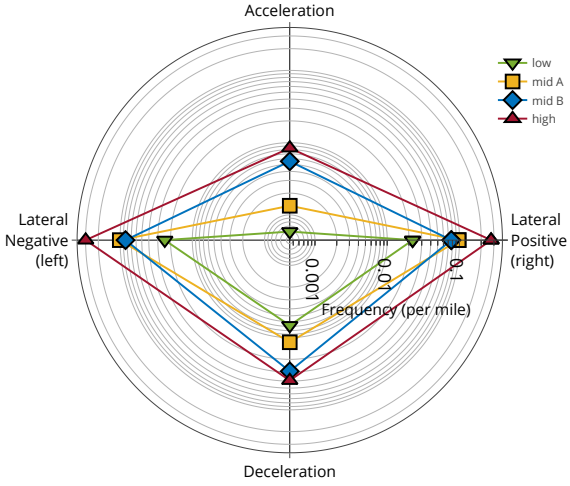
Figure 6.10a as well as Table 6.2 show that for longitudinal acceleration and deceleration epochs ≥ 0.3 g, *low* has the smallest value followed by *mid A*, *mid B*, and *high*. The relative values are such that *mid A* experiences 1.5 times as many longitudinal accelerations and decelerations as the *low* cluster. This value increases to about 3–4 times for *mid B*:*low* and *high*:*low* comparisons. The value of the centroid shows that a driver near the center of the *high* cluster has an acceleration epoch ≥ 0.3 g once every mile while that of a driver in the low cluster is only once every four miles. Similar differences can be seen when comparing deceleration values of the clusters.

Looking at all four thresholds for longitudinal acceleration and deceleration shown in Figures 6.10a - 6.10d and Table 6.2, some interesting trends emerge. As the threshold increases, the frequency of epochs decreases, which is expected. However, for longitudinal accelerations, as threshold increases, the ratio to *low* cluster also increases. For example, the ratio of *high*/*low* acceleration rates is 3.8 for ≥ 0.3 g, 7.1 for ≥ 0.4 g, 14.3 for ≥ 0.5 g, and 21.8 for ≥ 0.6 g thresholds. This implies that the longitudinal acceleration-based driving style becomes

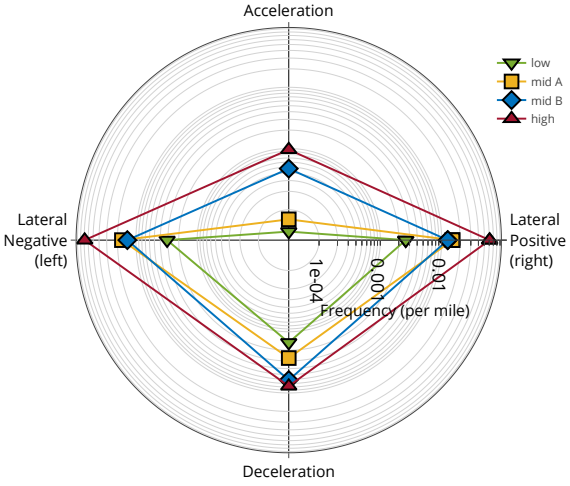


(a) 0.3 g threshold.

(b) 0.4 g threshold.



(c) 0.5 g threshold.



(d) 0.6 g threshold.

Figure 6.10: Cluster centroids plotted for various acceleration types at different threshold.

more distinct as the threshold is increased. However, this trend is not seen for longitudinal deceleration epochs. The ratio of *high*/*low* A for deceleration is 3.1 for ≥ 0.3 g and around 5.5 for the rest of the thresholds. A possible explanation is that the magnitude of the longitudinal acceleration is predominantly driven by user preference whereas the magnitude of the longitudinal deceleration or braking is also driven by the need for rapid evasive action,

Table 6.2: The value of the centroids for the various clusters in terms of rates of epochs greater than a threshold per mile.

Acceleration Type	Cluster Name	Centroid Value (epochs ≥ threshold per mile)				Ratio with respect to *low* cluster			
		0.3 g	0.4 g	0.5 g	0.6 g	0.3 g	0.4 g	0.5 g	0.6 g
Acceleration	*low*	2.3E-01	1.6E-02	5.9E-04	4.4E-05	1.0	1.0	1.0	1.0
	mid A	3.6E-01	2.7E-02	1.3E-03	7.0E-05	1.5	1.6	2.3	1.6
	mid B	8.0E-01	9.4E-02	5.5E-03	4.7E-04	3.4	5.7	9.4	10.7
	high	8.9E-01	1.2E-01	8.4E-03	9.6E-04	3.8	7.1	14.3	21.8
Deceleration	*low*	4.4E-01	4.7E-02	6.8E-03	1.5E-03	1.0	1.0	1.0	1.0
	mid A	6.6E-01	8.0E-02	1.2E-02	2.7E-03	1.5	1.7	1.7	1.8
	mid B	1.3E+00	2.1E-01	2.9E-02	6.2E-03	2.9	4.4	4.3	4.1
	high	1.4E+00	2.5E-01	3.9E-02	7.7E-03	3.1	5.2	5.8	5.2
Lateral Negative (left)	*low*	6.5E-01	1.6E-01	2.4E-02	3.1E-03	1.0	1.0	1.0	1.0
	mid A	1.3E+00	4.8E-01	1.0E-01	1.7E-02	2.0	2.9	4.2	5.5
	mid B	1.0E+00	3.8E-01	8.4E-02	1.4E-02	1.5	2.3	3.5	4.5
	high	2.0E+00	9.9E-01	3.0E-01	6.9E-02	3.1	6.0	12.5	22.2
Lateral Positive (right)	*low*	6.0E-01	1.5E-01	2.2E-02	2.6E-03	1.0	1.0	1.0	1.0
	mid A	1.2E+00	4.6E-01	9.7E-02	1.5E-02	2.0	3.0	4.3	5.8
	mid B	9.5E-01	3.5E-01	7.7E-02	1.3E-02	1.6	2.3	3.4	4.8
	high	1.9E+00	8.9E-01	2.7E-01	6.1E-02	3.1	5.8	12.2	23.4

which may limit how different driving-style-based groups could be.

The lateral acceleration components of the various cluster centroids also show interesting trends. First, the values of lateral left and lateral right components of the cluster centroids are nearly identical. This is despite the roadways' infrastructure being asymmetrical. For example, on average, right turns are sharper than left turns but also taken at lower speeds.

The opposite effect of the two factors causes the centroid components to be symmetrical. Second, the order of the clusters is consistently *low* < *mid B* < *mid A* < *high* for all four thresholds. This is different than longitudinal accelerations in two ways. The order of *mid B* and *mid A* are reversed and their values are much closer to each other than for longitudinal acceleration. Another interesting trend, similar to longitudinal acceleration, is that as the thresholds increase, the cluster centers move farther apart. The ratio of *high*/*low* is 3.1, 6, 12.5, and 22.2 for lateral left with similar values for lateral right components of the cluster centroids.

In terms of real world driving, these cluster centroid-based driving styles have distinct physical meaning. For example, consider the *low* cluster versus the *high* cluster. A *low* cluster driver experiences $\geq$ 0.4 g accelerations once every 63 miles whereas a *high* cluster driver experiences the same once every 8 miles. For deceleration or braking of magnitude $\geq$ 0.5 g, a *low* cluster driver experiences this event every 147 miles as compared to once every 25 miles for a *high* cluster driver. For lateral left or right accelerations with $\geq$ 0.5 g magnitude, the rates are once every 20 miles for *low* and once every one to two miles for *high* cluster drivers. These are significant differences and a passenger sitting in a vehicle being driven based on the *high* cluster would have a completely different ride quality experience than one being driven based on the *low* cluster.

It should be noted that even though these clusters are strictly based on acceleration values, when an intelligent transportation system is designed to emulate them, it will affect most other perceivable aspects of driving. For an ADS to operate within a certain region of the "acceleration type – magnitude – frequency" space, it will have to adjust its driving speed, headway, following distance, and other driving parameters as well. This is because the constraint of meeting a certain acceleration frequency will automatically create a window in which the system has to operate to be able to meet that frequency.

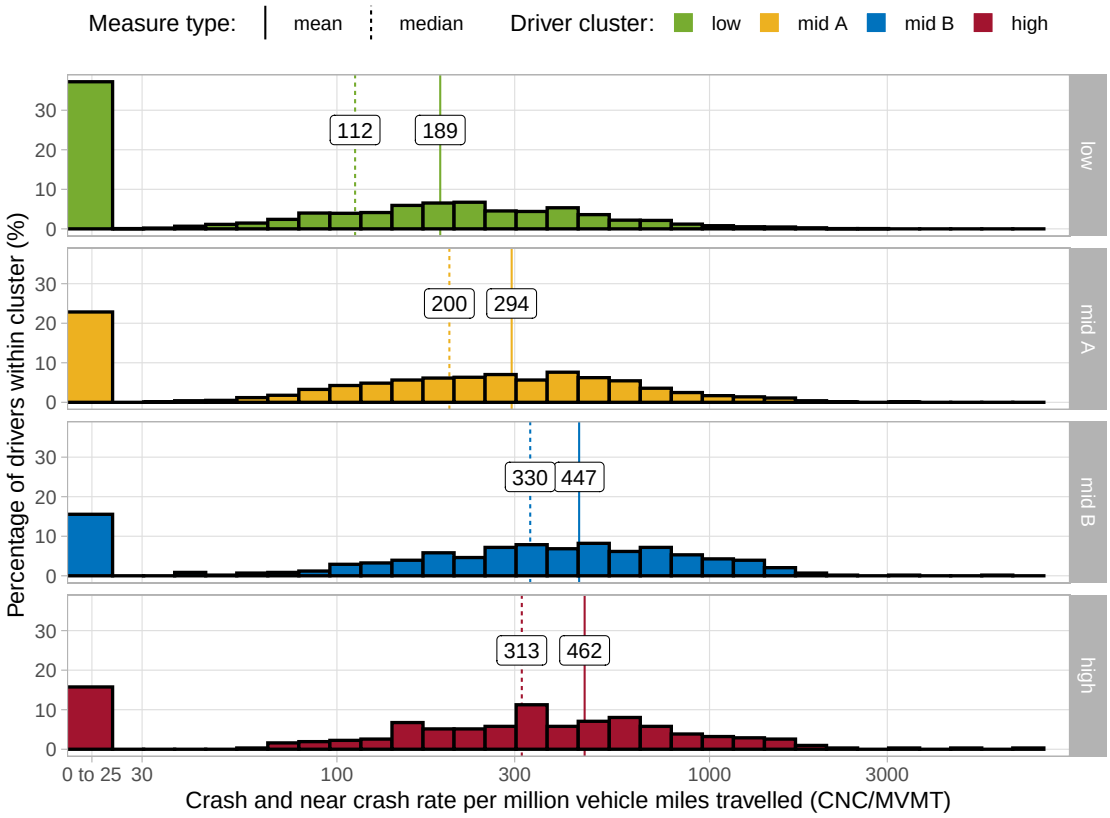


Figure 6.11: Histogram of crash and near crash rates for the four driver clusters.

6.5.2 Relationship to Safety Metrics

For a holistic driving style study, it is essential to evaluate the safety implications of each individual driving style determined. To achieve this, crashes and near-crashes discovered during previous analyses of the SHRP 2 NDS were used to calculate the crash rate and crash plus near-crash rate for each participant included in this analysis. The process set up for the discovery of crashes and near-crashes is quite robust and therefore most, if not all, safety critical incidents in a participant's driving history are identified. This is done in a two step process. In the first step, each timestamp of a trip is evaluated using an algorithm that flags possible safety critical events. This algorithm uses multiple recorded variables, such as the vehicle speed, acceleration signals, safety system statuses, etc., to determine

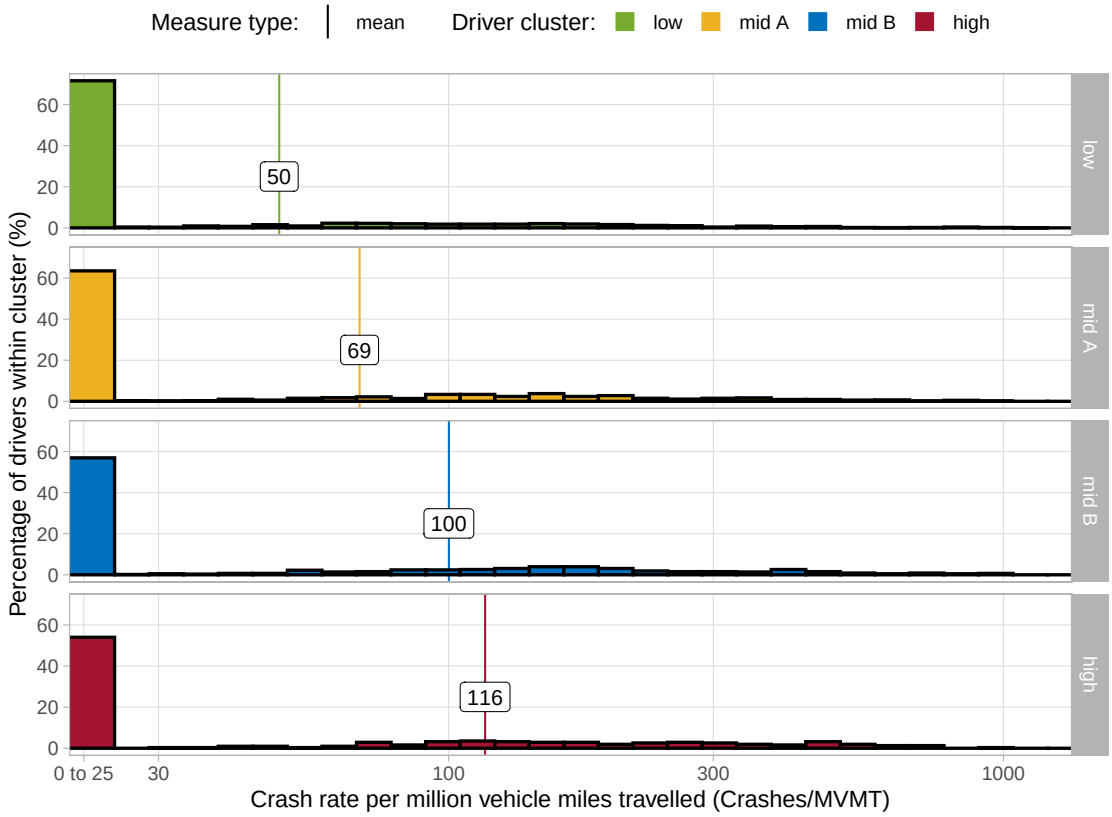


Figure 6.12: Histogram of crash rates for the four driver clusters.

the flags that would need review. In the second step, multiple reduction specialists analyze the corresponding video footage and sensor data for each case to determine if the flagged timestamp was actually a crash, a near-crash, or a false positive.

Figures 6.11 and 6.12 show the distribution of the crash plus near-crash and crash only rates per million vehicle miles traveled (MVMT) for the participants included in this study. The distributions are grouped by the driving style cluster and the solid/dashed vertical lines represent the calculated mean and median rates for each group. Since crashes are quite rare, near-crashes are often used as surrogate measures for estimating driving risk. In this analysis, both crash rates as well as crash and near-crash rates have been used. As can be clearly seen for both types of comparison, the *low* cluster has lowest collective rates and therefore the

Table 6.3: Comparison of the crash and "crash plus near-crash" rates for drivers from four clusters using Tukey honest significant differences test.

Rate Type	Comparison	Difference	Lower interval	Upper interval	Adjusted p-value
Crashes per MVMT	low - high	-66.8	-90.6	-43.0	6.0E-10
	low - mid B	-50.6	-69.2	-32.0	6.2E-10
	mid A - high	-47.2	-72.0	-22.5	5.8E-06
	mid B - mid A	31.0	11.2	50.8	3.5E-04
	low - mid A	-19.6	-35.1	-4.1	6.6E-03
	mid B - high	-16.2	-43.0	10.5	4.0E-01
CNC per MVMT	low - high	-272.8	-334.0	-211.5	6.0E-10
	low - mid B	-257.4	-305.3	-209.5	6.0E-10
	mid A - high	-167.7	-231.5	-104.0	6.9E-10
	mid B - mid A	152.3	101.3	203.4	6.0E-10
	low - mid A	-105.1	-145.1	-65.0	7.0E-10
	mid B - high	-15.4	-84.4	53.6	9.4E-01

lowest associated driving risk. *Low* is followed by *mid A*, *mid B*, and finally, *high*, which has the greatest crash rate of any group. The comparisons show that a significant percentage of the population has zero cases for both crashes as well as near-crashes. Since the x-axis is transformed to a log base 10 scale, these points occur at $-\infty$ and therefore are represented by the left most open bin in the histogram.

To conclusively show that these groups differ in terms of driving risk, a 95% confidence interval was calculated using Tukey's honest significant difference method. Table 6.3 shows the difference in crash and crash plus near-crash rate along with the 95% confidence intervals and the p-value adjusted for multiple comparisons. All comparisons, except between *mid B* and *high*, indicated a statistically significant difference. This shows that not only are the various driving style clusters markedly different in terms of ride quality and feel, but they

are also have statistically significant differences in driving risk. The *low* driving style is the safest of all the driving styles detected followed by *mid A*, *mid B*, and *high*. Even though *mid B* has a lower driving risk associated with it than *high*, the difference is not statistically significant for either of the two rates. Also, it is interesting that *mid A* has a much lower driving risk than *mid B* even though it has higher values of lateral acceleration components.

Besides crash rates, another orthogonal measure to evaluate these driving clusters is to analyze drivers' speeding behavior. Figure 6.13 compares the speeding behavior of the four driving styles on low speed roadways. The top plot in Figure 6.13 shows the absolute percentage and the bottom plot shows the percentage difference relative to the *low* driving style cluster. This metric is calculated by measuring the relative speed of the participant's vehicle with respect to the speed limit of the roadway, which was acquired through map-matching. The main bins of interest in this plot are from -10 to +15, which may indicate a group's preferred driving speed given the speed limit. For each of the bins in this region, the *low* cluster drivers have a relatively higher proportion of their driving below or at the speed limit. For the *high* cluster drivers, a relatively higher proportion of their mileage was accumulated well above the speed limit when compared to the other drivers. Interestingly, *mid A* also has a relatively higher proportion of driving at or above the speed limit.

The four cluster centroids can be summarized as:

- ***low***: This cluster has the lowest rates for all four types of acceleration at all thresholds. This cluster also has the lowest associated average crash rate of 50 crashes per MVMT. The drivers of this cluster drive a higher proportion of their mileage at or below speed limits when compared to the other driving styles. Therefore, these drivers represent a safe and cautious driving population.

- ***mid A***: This cluster has the second lowest longitudinal acceleration and deceleration

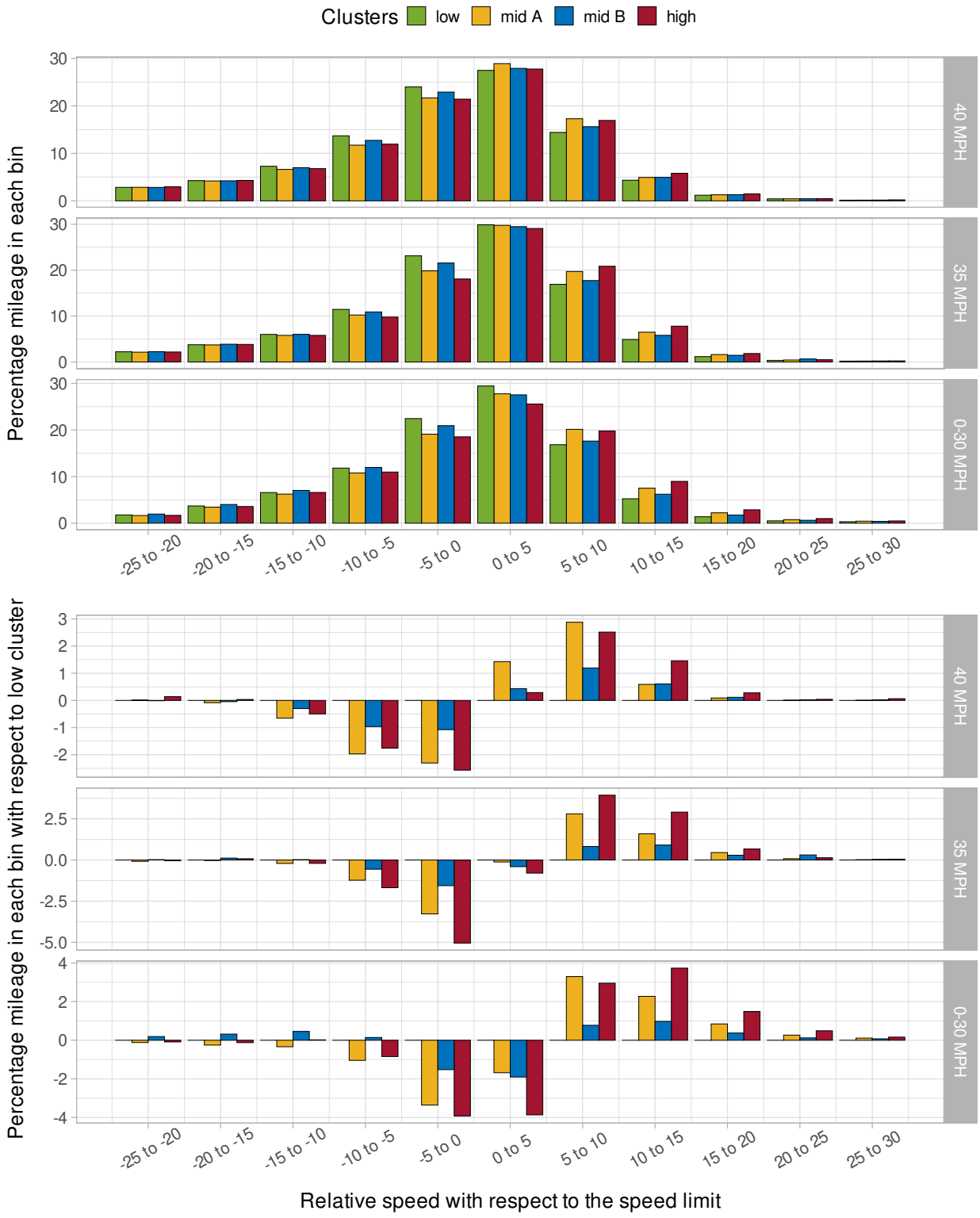


Figure 6.13: Percentage of mileage driven in each speeding bin by the four driving style clusters.

rates that are closer to the *low* cluster. However, this cluster has the second highest lateral acceleration rates. This cluster has the second lowest average crash rate of 69 crashes per MVMT, which is 38% higher than *low* cluster. The drivers in this cluster drive a higher proportion of their mileage above the speed limit as compared to the other clusters. The relatively different behavior in longitudinal versus lateral acceleration rates as well as the crash rate and speeding behavior make this cluster quite interesting. These drivers are still relatively low risk but may have a sportier driving style, as they take turns at higher speeds.

- ***mid B***: This cluster has the second highest longitudinal acceleration and deceleration rates, which are quite far from the *mid A* and *low* driving style clusters. However, the lateral accelerations are slightly lower than *mid A* and generally fall between *high* and *low*. This group has the second highest average crash rate of 100 crashes per MVMT which is 100% higher than than that of the *low* cluster and 45% higher than that of *mid A*. As far as speeding behavior is concerned, the drivers from this group drive a higher proportion of their mileage at or below speed limit than *mid A* or *high* cluster drivers. But when compared to the *low* cluster drivers, *mid B* drivers drive a higher proportion of their mileage above the speed limit. Overall, this group can be classified as having high longitudinal acceleration rates, intermediate lateral acceleration rates, and relatively high driving risk.

- ***high***: This driving style cluster has the highest longitudinal and lateral acceleration rates. The average crash rates associated with this cluster are also the highest, with 116 crashes per MVMT which is 132% higher than that of the *low* cluster, 68% higher than that of *mid A*, and 16% higher than that of *mid B*. However, as shown in Table 6.3, the difference to *mid B* is not statistically significant, whereas the other the comparisons are statistically significant with a very high degree of certainty, as shown by the low

p-values. This cluster also has the highest relative proportion of driving above the speed limit and the lowest relative proportion of driving at or below the speed limit. All these indicators show that this driving style cluster represents the riskiest driving as well as the highest acceleration rates.

6.6 Conclusions

The purpose of this study is to identify acceleration-based driving styles, differentiate them on the basis of their physical significance, and evaluate their corresponding driving risk. This has successfully been accomplished by using a combination of rule-based metrics and unsupervised machine learning algorithms. The acceleration rate-based metrics were chosen so that short term effects, such as traffic, were less relevant and long term driving style was the dominant factor. A consistent driving domain of low-speed roadways was chosen for the analysis. A k-means clustering algorithm was used to identify four clusters in a 16 dimensional "acceleration type – magnitude – and frequency" space. These clusters were named *low*, *mid A*, *mid B*, and *high* based on the relative values of longitudinal and lateral acceleration components of the cluster centroids.

The cluster centroids represent fairly distinct driving styles and safety implications. The *low* cluster based driving style represents the lowest rates of acceleration for all thresholds and has the least associated driving risk. These drivers also exhibit a cautious driving style from a speeding perspective. The *high* cluster based driving style represents the highest rates of acceleration for all thresholds and has the highest associated driving risk. These drivers also exhibit the most drastic speeding behavior of the four clusters. There are two more clusters between *low* and *high* identified as *mid A* and *mid B* which have similar later acceleration rates but fairly different longitudinal rates.

This study fills an important research gap in the understanding of acceleration-based driving styles and could benefit a number of fields. Designers of intelligent transportation systems, such as Advanced Driver-Assistance Systems (ADAS) and Automated Driving Systems (ADSs), can compare their driving styles with human driving styles determined in this study. System designers could also use the driving style clusters to emulate human driving so that passengers feel more comfortable in vehicles using such systems. The clusters can also be used to better predict the behavior of other vehicles on the road. Therefore, this study can be significantly beneficial to intelligent transportation systems by improving their safety and making them more comfortable for a wider range of passengers.

This study will also improve data mining and safety research based on naturalistic driving studies such as the SHRP 2 NDS. When using such studies for modeling human driving behavior, researchers often want to exclude unsafe drivers so that the resulting models can be implemented in various transportation systems. Traditionally this was done purely on the basis of crash rate. As Figure 6.12 shows, a considerable portion of drivers from each cluster have zero crashes, as crashes are fairly rare. Using a denser measure such as the driving style classification instead of a purely crash rate criteria for driver inclusion would offer advantages in driver selection for emulation. This would ensure that the selected drivers not only have a better safety record but also other desirable driving behaviors.

Considerable scope for future work has opened through this analysis. First, instead of purely counting the four types of acceleration epochs at various thresholds, the acceleration epochs could first be classified as maneuvers. This would create a maneuver-based clustering space that could group drivers based on the similarities or differences within a driving maneuver. Second, other important driving measures representing headway, driving speed, and speeding behavior could be included in addition to lateral and longitudinal accelerations for a holistic driving style determination. Finally, other domains not included in this analysis, such as

high speed roadways, could also be included to see how the driving style determination changes. However, it should be noted that adding more dimensions to a clustering algorithm will increase the Euclidean space and could make the data quite sparse. Therefore, the scale of future driving style recognition studies needs to be carefully considered by weighing the size of the data versus the possible feature space.

Chapter 7

Interactive Analytics Tools to Characterize Human Driving Through Naturalistic Driving Data

7.1 Introduction

Traditionally, data analyses are conducted around a few handpicked research questions. However, recent developments in interactive analytics have enabled building tools that explore a research area from multiple perspectives without being confined to a single research question. This can be especially useful when building tools for a broad audience that is approaching a problem from different perspectives. An opportunity to build such tools was presented through the AMP program that brought 13 industry leaders together to leverage naturalistic driving data for the development and testing of ADS and ADAS. The work presented in this chapter discusses the major ideas around building these tools and how they facilitate quicker and deeper insight extraction.

7.1.1 Unique Perspective Provided by Naturalistic Driving Studies

Several data sources are used to answer various questions about driving behavior and other transportation phenomenon. They vary in scale, specificity, and comprehensiveness. Data from test track studies or field operation tests is highly specific to a research question but lack scale and comprehensiveness. These types of studies are also affected by the observer effect. On the other end of the spectrum, datasets generated through large-scale epidemiological studies such as FARS and GES provide considerable statistical power for a wide variety of research questions. However, data from such studies focuses on extreme crashes that cause fatalities or generate police accident reports. This data also lacks comprehensiveness as it does not contain any routine driving behavior.

Naturalistic driving studies such as the SHRP 2 NDS offer a unique perspective due to their large scale and comprehensiveness. These studies record driving behavior of selected participants driving their own vehicles from key on to key off for several months at a time. The participants are specifically sampled to give statistical power for a wide variety of research questions. The DAS collects kinematic signals such as speed, acceleration, and yaw rates, driver inputs such as steering, brake and accelerator pedal, traffic sensors such as radar, and various video feeds. Such a setup offers comprehensiveness in two fundamental ways. First, it records all the needed information that will give invaluable context for each driving behavior. This is often missing from the other types of datasets. Second, it records all the driving executed by the participants for an extended period, therefore capturing routine as well as rare events.

VTTI houses 70 million miles of naturalistic driving data collected from motorcycles, passenger vehicles, and heavy trucks. Data from these studies is a valuable resource for developing

a deeper understanding of crash risk, frequency and composition of driving scenarios, real-world edge cases, and routine driving behavior. This information is useful for a wide variety of use cases. For example, vehicle system engineers can use this data to develop ADS and ADAS. Safety researchers can also use such data for evaluating safety performance of various vehicle systems. Behavioral researchers can use such data for better understanding human driving behavior and embedding human preferences into models. Therefore, considerable insight that can be extracted from this data to benefit multiple vehicle manufacturers and vehicle system suppliers. The AMP was set up to perform this role in a partner-driven and high impact format.

7.2 Capitalizing on Naturalistic Driving Data through the Automated Mobility Partnership

The AMP program brings together 13 industry leaders in a pre-competitive setting to promote the development of tools, techniques, and data resources that support ADS and ADAS deployment. The steering committee members, consisting of vehicle system manufacturers, Tier 1 suppliers, and tech companies, decide the priorities and the areas of focus. AMP provides the members access to a variety of real-world driving data from the naturalistic driving studies and a suite of support tools developed to extract insight from the data. A library of cases consisting of crashes, near-crashes, and driving epochs facilitate the exploration of routine and SCEs.

7.2.1 Defining the Tasks for the Analytics Tools

To appropriately leverage the data, there was a requirement for creating a set of analytics tools that lets the end user extract deeper insights from the naturalistic driving data. After considerable thought by the AMP team, four fundamental analytics tasks were identified. These are:

1. What is the relative composition of crashes across national, naturalistic, and AMP datasets? As national crash databases such as FARS and GES primarily rely on police accident reports, they are skewed towards higher severity when compared with crashes from naturalistic driving studies. Since complete trips are recorded from start to end, it is possible to detect nearly all crashes and near-crashes in naturalistic driving, which may be less severe but equally valuable for ADS and ADAS developers. Also, since the AMP case library is a subset of the overall NDS data, further differences can be introduced between NDS and the AMP dataset. Therefore, to make meaningful assessments of safety and risk using such datasets, it is essential to determine the differences in composition between them.

2. What is the relative frequency of crashes, near-crashes, and various driving epochs in naturalistic driving data? Naturalistic driving studies are uniquely positioned to inform about rates of occurrences as they capture full trip data of large cohorts over an extended period of time. This is in addition to the rich context captured by an extensive sensor suite consisting of kinematic, driver input, roadway, video, and environmental data. Therefore, the same set of trips can be used to calculate the numerator, i.e., the number of occurrences of a certain scenario, and the denominator, which is the total number of miles traveled. Given the rich context available, rates can also be calculated in a certain pre-specified operational design domain. This information will help AMP

users to better design vehicle systems knowing how often certain scenarios are expected over a component lifetime. Crash and near-crash rates can be used in comparing risk and safety metrics of vehicle systems with human drivers as the baseline.

3. What is the distribution of various metrics of interest during crashes, near-crashes, and different driving epochs? A distinguishing quality of naturalistic driving studies is the availability of data representing kinematic, roadway, and environmental factors. Knowing the expected ranges of these factors in various operational design domain segments is essential for vehicle system design and operation.

4. How can the user understand the multidimensional nature of data and choose the appropriate subset of cases for their analysis? Describing driving scenarios is fairly complex as there are many different factors that need to be taken into consideration. For example, a passenger vehicle driving straight at 50 mph on a controlled access highway in clear weather during day time with no lead vehicle present can be considered a driving scenario. Being able to narrow down to such a scenario will require a multidimensional approach using data variables representing vehicle class, speed, and maneuver, roadway properties, and environmental factors. Therefore, it is essential for users to be able to narrow down and understand a set of cases from a multidimensional perspective and be able to filter based on multiple parameters at the same time.

7.2.2 General Challenges

To create effective analytics tools for the tasks laid on in Section 7.2.1, some general challenges need to be overcome. These challenges can be summed up as:

1. A technically diverse user base will be using these tools to answer questions relevant for them. For example, even within the same organization, some users could use the

data for system design while others could be using these for regulatory purposes. To further complicate matters, some of the users may not be familiar with naturalistic driving data at all and, therefore, would not know the general biases of the dataset and various issues to watch out for. Therefore, it is important to provide sufficient information for all users to be able to utilize these tools properly.

2. How to maximize insight extraction for each query? Considerable thought will need to be given when designing the visualizations and the user interactions so that each new visualization quickly answers the query of the user or at least guides them in the right direction. This will be accomplished by incorporating the standard information visualization principles applied in the field.

3. How to accelerate the sense-making loop? Visual analytics or interactive analytics need the human and the machine to work together to derive insight. The analytics process usually works best when the user first visualizes the data at an overview level, then filters the data according to their needs, and is provided details on demand. To accelerate the sense-making loop, it is essential to make these steps intuitive and provide the most appropriate details on demand.

4. How to quickly on-board new users by speeding up the learning curve? Since these tools will have many features, it will be important to help new users to quickly get an understanding of how to use the tools. This can be achieved by a two-pronged strategy: first, creating tools that are intuitive so that users can learn the functions of the tool as they use them; second, helping the user through demonstrations, tutorials, walkthroughs, and workshops.

5. How to enable the users to determine if the data has correctly been processed or if some errors have propagated? Any complex dataset that has gone through several

layers of processing will have some errors introduced through faulty sensors or imperfect algorithms. It will be essential for the end user to be able to determine if a particular value is real or if the value has been corrupted by sensor/algorithm errors. This is especially true for outliers. Such challenges can be overcome by incorporating video into the interactive analytics tools through details on demand features.

7.3 Interactive Analytics for Insight Extraction

Large-scale NDS such as SHRP 2 present a unique opportunity for understanding human driving behavior. However, the scale and complexity of such datasets also present considerable challenges for analysis. Quite often, it takes researchers a substantial amount of time to understand the data structures and be able to analyze the data. Interactive analytics tools offer possible solutions as they enable a researcher working on a particular research problem to condense the complexity and large scale of the overall dataset into a more manageable size so that multiple users can work on that problem without having to fully wade into the larger dataset.

Interactive analytics tools are visual interfaces that facilitate analytical reasoning by human users through interaction [67, 128]. Figure 7.1 illustrates the sense-making loop in an interactive analytics tool. The designer of the tool performs an initial analysis of the data and creates a visualization that serves as a starting point in the analysis. The perception of the visualization by the user results in new insights and therefore knowledge acquisition. This can lead the user to form new hypotheses, which are then translated into new specifications for the visualization. This iterative loop is the crux of the sense-making process for most interactive analytics tools. Accelerating this process is key to enabling the user to gain more insight from the data.

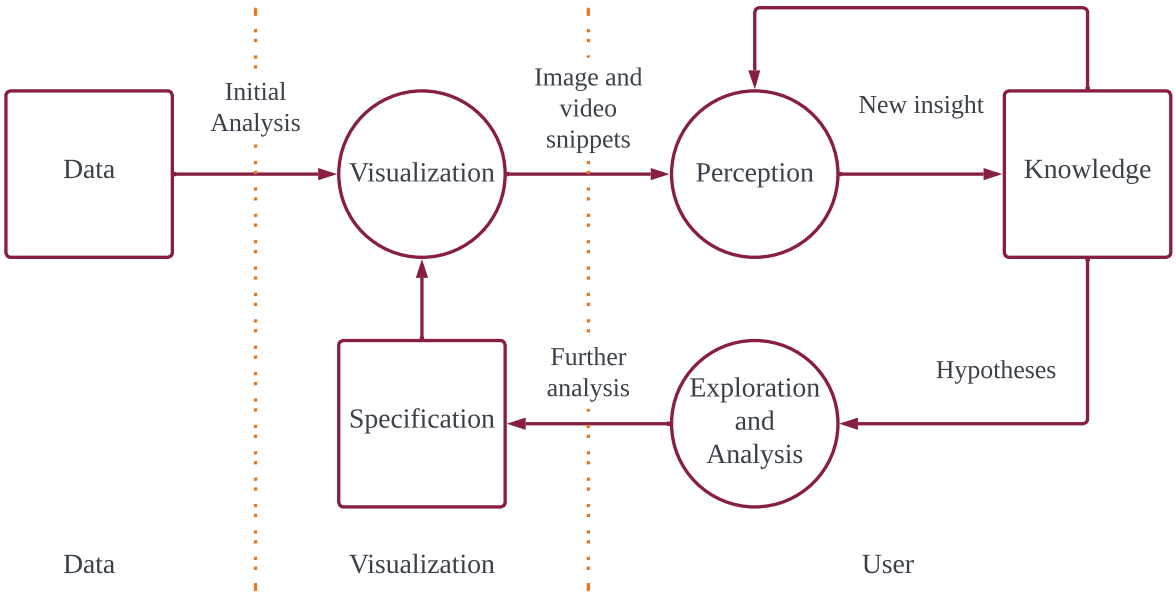


Figure 7.1: The sense-making loop in interactive analytics tools [67, 135].

For accelerating the sense-making loop in AMP analytics tools, it is important to leverage two characteristics of NDS data through a "details on demand" approach. The NDSs used in AMP have continuous front video feed from the ego vehicle perspective. This is an invaluable resource since all users of the tools will have experienced such a perspective and would be able to glean a very high quantity of information in a short amount of time. All the tools discussed in this chapter link every data point to a video snippet corresponding to the same time as when the data was collected. These snippets are available for the user by simply clicking on the glyph representing a data point on the graph. This not only helps the user quickly correlate complex metrics with actual driving scenarios but also enables them to differentiate between real outliers and outliers caused by sensor or algorithm errors.

The rich context of the NDS data available through the numerous variables collected can also be leveraged by a "details on demand" approach. To help the user gain a multivariate perspective of data points, additional information about other important metrics can be

provided through hover displays. Further, simultaneous plots representing different variables can be provided in the tools and interactive features, like "brush to select," can be used to link different data points across the plots.

Further, to remove friction from the sense-making process, it is essential that the visualization or video loading time after the user's input is minimal. This has been achieved by optimizing data structures, creating fast loading defaults, and avoiding higher video resolution when unnecessary. The interactive features, drop-down menus, and data filter options have been homogenized across the tools to create a faster learning curve for the user.

7.4 Design and Iteration Methodology

These tools were designed through an iterative approach. After the initial conception of the main tasks, whiteboard wireframing was performed to test out rough ideas. This included the design of the main visualization as well as the user interface for data selection and interactivity. Once a basic structure was agreed upon, a basic functioning product was created using R Shiny [28]. There were three levels of user testing performed for each version of the tool. The basic testing was done by the author himself to make sure the features were working as intended. After passing these basic tests, the beta versions of these tools were tested by other researchers also working on AMP at VTTI. These users provided feedback about bugs, user experience, and features that would improve the tools. The tools were then released to the intended users at VTTI and the other organization of the AMP steering committee. Feedback as well as feature requests were consolidated from all the users and planned into future releases.

7.5 Tools

7.5.1 Database Comparison Tool

7.5.1.1 Objective

The objective of the database comparison tool is to illustrate the differences between the national, naturalistic, and AMP crash datasets. The user chooses the comparison factor, the type of mapping, and other constraints for the desired comparison. The tool visualizes the three datasets to help the user understand the relative composition of each for the chosen comparison factor. This information will enable the user to draw appropriate inferences from the AMP dataset as they now know how it differs from the overall naturalistic and national datasets such as FARS and GES. Figures 7.2 and 7.3 show the various elements of the tool. A brief explanation for each element has been provided in the following sections.

7.5.1.2 Major Elements

1–4. **Primary navigation tabs**: The primary navigation tabs enable the user to navigate between the database comparison tool, the variable mapping, the methods section, and the demonstration section. The various elements of the first tab have been explained below. The variable mapping tab explains how various categories in the nation datasets and the naturalistic datasets were harmonized and mapped. The methods tab contains information about the methods used, the definition of the various metrics, and information about how to use the tool. The Demonstration tab contains a video explaining the various features of the tool. This high level navigation allows function-based compartmentalization of various aspects of the data analytics tools.

5. **Variable mapping selector**: Both the national and naturalistic crash datasets have

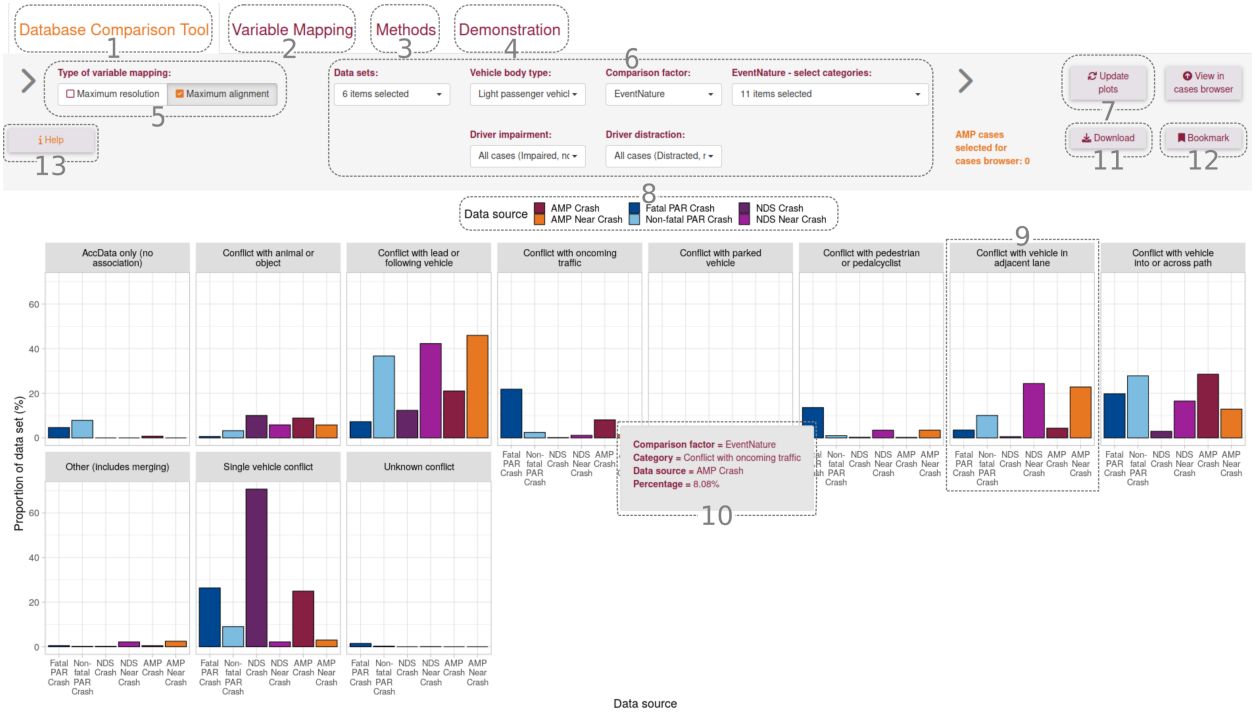


Figure 7.2: The major elements of the database comparison tool.

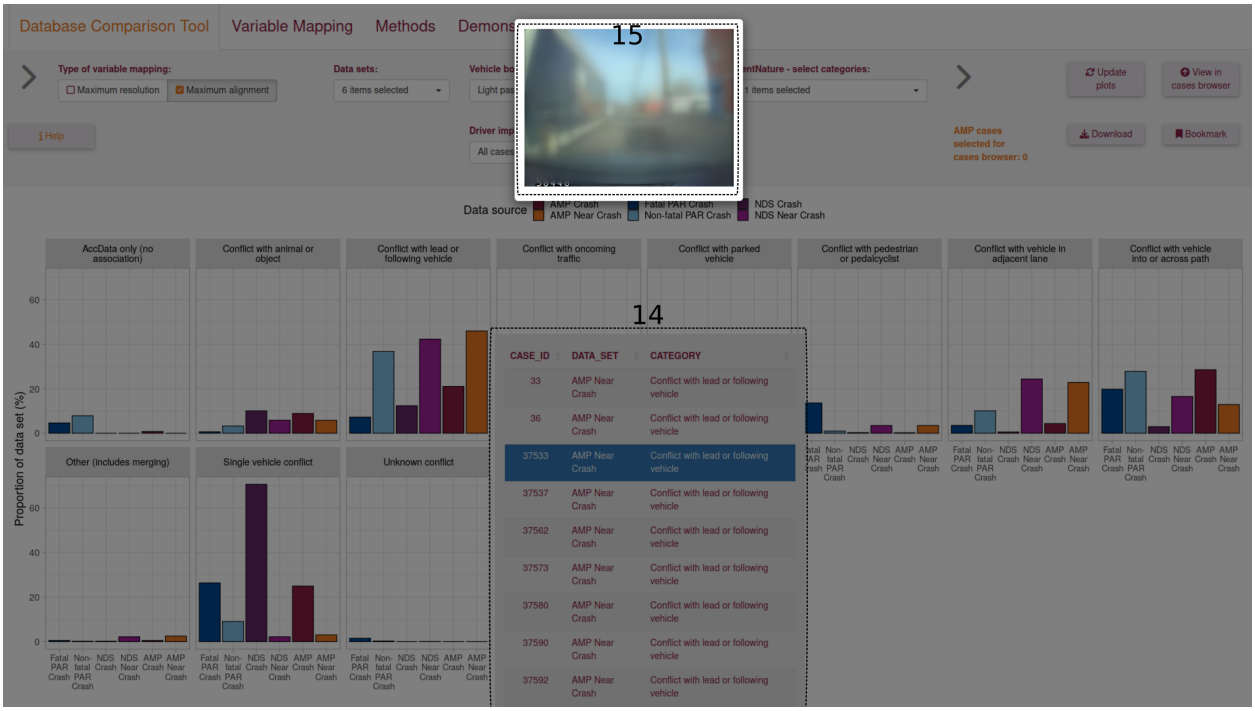


Figure 7.3: The interactive video elements of the database comparison tool.

a large number of categories and subcategories that are not perfectly aligned. To make the comparisons presented in this tool possible, VTTI researchers performed extensive variable mapping between the national and naturalistic datasets. Two types of mappings were created to facilitate maximum resolution and maximum alignment. This toggle switch enables the user to choose between the two.

6. **Data selection drop-down menus**: These menus allow the user to choose the datasets, vehicle classification, comparison factor, and driver behavioral metrics. These choices enable the user to make the most appropriate comparisons for their use case.

7. **Plot update button**: The plot update button controls when the visualization updates after changes have been made to the data selection drop-down menus. This is a deliberate design choice to add an additional step for the user after changing the data selections. The alternative was to spontaneously update the plot as the user changed the data selection options. However, when making multiple changes, the plot would keep updating and would result in a suboptimal experience.

8. **Legend**: This element shows the legend. The dataset variable is mapped to two channels, which are fill color and the position on the x-axis. This redundancy helps the user to efficiently perform the main task, which is comparing the relative percentages between the various datasets.

9. **Plot facet**: One of the primary requirements of this task was to compare the dataset composition using the comparison factor. Each comparison factor had a different number of subcategories. Therefore, the plot format needed to be flexible to incorporate such variation. After testing out different variations of the tool, providing a facet for each subcategory was the optimal solution.

10. **Hover information**: The hover information is activated by hovering the cursor over

the bar glyph. The hover display gives information about the comparison factor, category, data source, and the percentage.

11. **Download button**: The download button packages all the essential data, figures, and state of the tool into an Excel sheet ready to be saved on the user's computing device. The state of the tool is saved using a bookmark URL, which can be used by the same or a different user to generate the same figures and populate the same data selection again. The download feature has three fundamental benefits for the end user. First, it enables convenient saving of the analysis. Second, it facilitates users to share their analysis with other team members easily. And finally, it provides a simple way for the users to ingest the data and insights from the tool into the user's own data science tools such as MATLAB, Python, and R.

12. **Bookmark button**: The bookmark button saves the state of the analytics tool and provides the user with a URL that will open the tool in the exact same state.

13. **Help button**: The help button guides the user with the various features of the tool and helps them quickly get up to speed on how to use it.

14. **Case table**: Each bar glyph is an aggregate value representing several cases. A table containing all the cases represented by each glyph is activated by clicking on it as shown in Figure 7.3. This allows for the user to see each underlying case and update their mental models.

15. **Video viewer**: The video viewer is activated by clicking on the case row in the table and automatically starts playing a 6-second video of the driving segment from the ego vehicle front view perspective. The video snippet contains a significant amount of information about the scenario, roadway properties, and environmental conditions. This enables the user to quickly correlate the driving conditions to the visualized case

type, conditions, etc., and update their mental models. The integration of video into traditional visualization methods accelerates the sense-making loop that is at the core of any interactive visualization task.

7.5.2 Rates Tool

7.5.2.1 Objective

The objective of the rates tool is to illustrate the frequency of crashes, near-crashes, and various driving scenarios in NDS. This information is helpful for AMP users to understand the relative rates of occurrences of various types of driving events. Knowing the expected frequency of scenarios helps vehicle system engineers to better prioritize and design components for the product's expected life. Such data is also valuable for safety researchers to compare performance with human drivers. Figures 7.4, 7.5, and 7.6 show the various elements of the tool. A brief explanation of each element has been provided in the following sections.

7.5.2.2 Major Elements

1–3. **Primary navigation tabs**: The primary navigation tabs enable the user to navigate between the Rates tool, the methods section, and the demonstration section. The various elements of the analytics tab have been explained below. The methods tab contains information about the methods used, the definition of the various metrics, and information about how to use the tool. The Demonstration tab contains a video explaining the various features of the tool. This high level navigation allows function-based compartmentalization of various aspects of the data analytics tools.

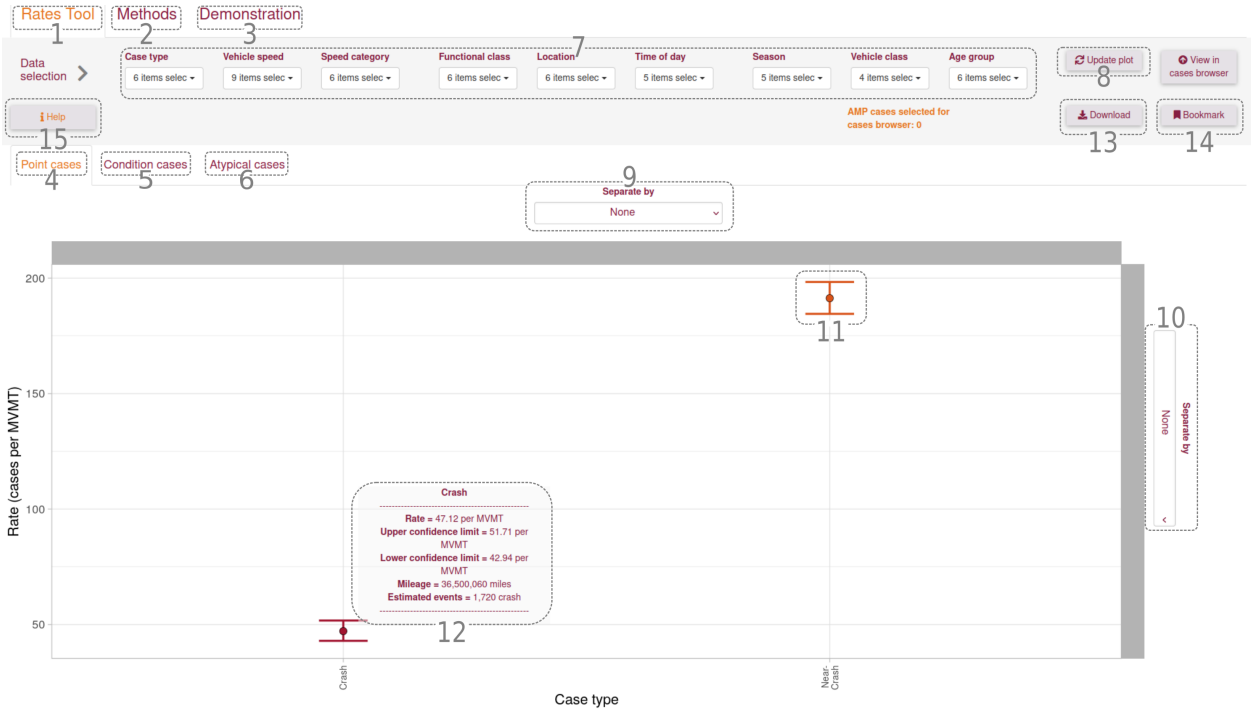


Figure 7.4: The major elements of the rates tool.

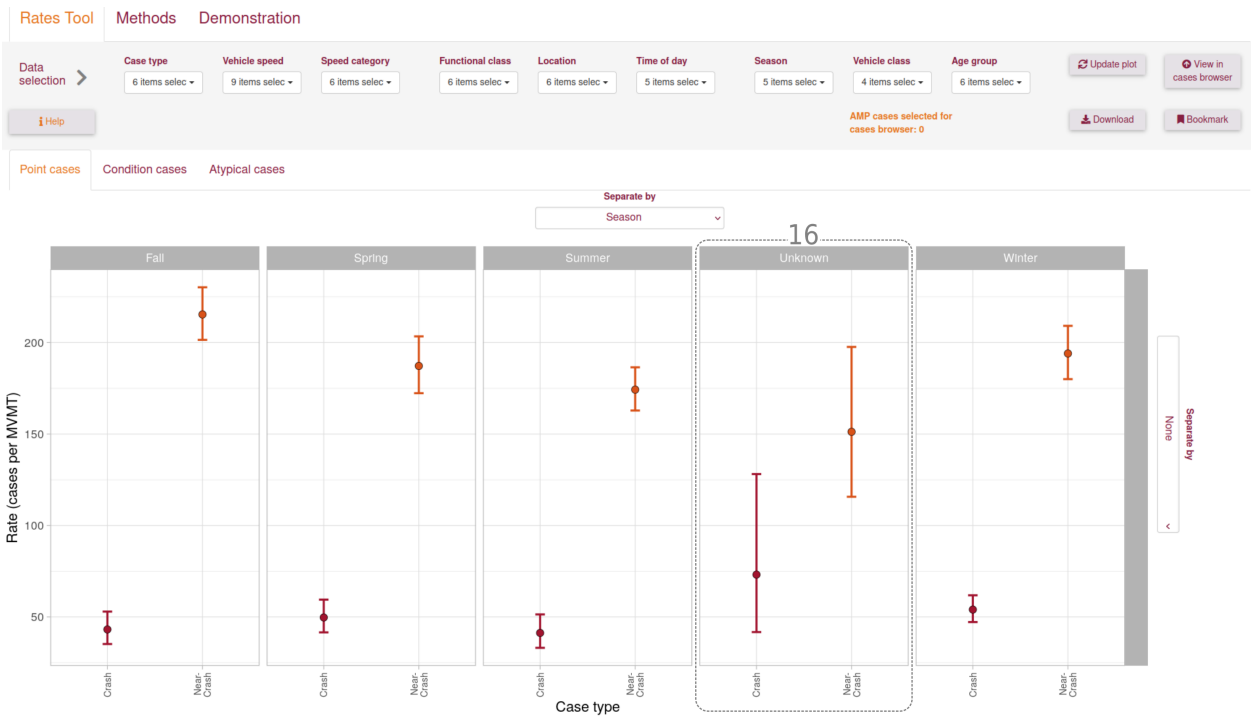


Figure 7.5: The faceting elements of the rates tool enabling users to see rates by different ODD segments.

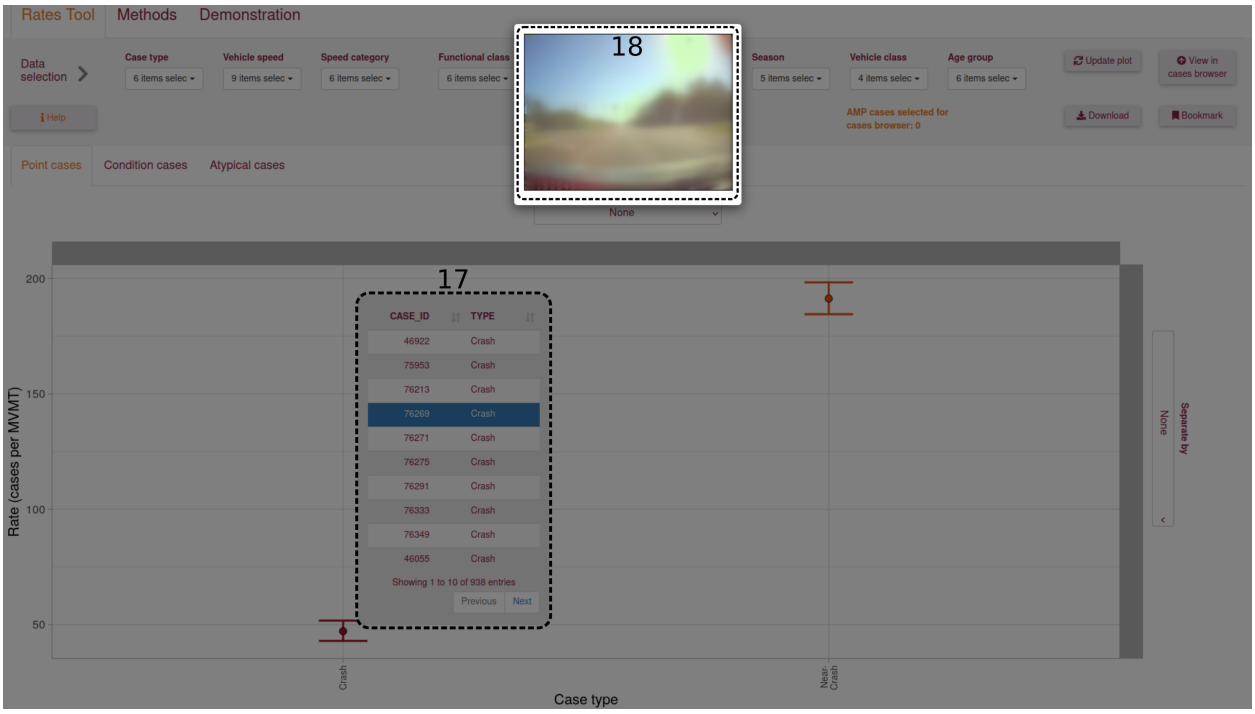


Figure 7.6: The interactive video elements of the rates tool.

4–6. **Secondary navigation tabs**: These secondary navigation tabs separate the rates by the type of case. Cases that occur as a point instant are calculated as number of cases per MVMT traveled and are shown in the first tab. Crashes and near-crashes are examples of such cases. Cases that occur as a condition are calculated as number of miles where the condition exists per MVMT and are presented in the second tab. Roadway type and environmental conditions are examples of such cases. Finally, cases that required unique approaches for rate calculation are presented in the Atypical cases tab.

7. **Data selection drop-down menus**: These menus let the user select the case types, roadway properties, environmental factors, and driver demographics. Therefore, not only does this tool provide rates of each case type, but also provides it in a dynamically selectable subset of the overall dataset.

8. **Plot update button**: The plot update button controls when the visualization updates after changes have been made to the data selection drop-down menus. This is a deliberate design choice to add an additional step for the user after changing the data selections. The alternative was to spontaneously update the plot as the user changed the data selection options. However, when making multiple changes, the plot would keep updating and would result in a suboptimal experience.

9–10. Facet drop-down menus: The horizontal and vertical facet drop-down menus enable the user to pick a category to separate the rates by. For example, in Figure 7.5 crash and near-crash rates are separated horizontally by season. This feature helps the user quickly compare how the rate is affected by various conditions.

11. **Rate glyph**: The symbol chosen for representing the rate is a circle with error bars. The circle in the center represents the estimated rate of occurrence, and the two horizontal error bars represent the ± 95% confidence limits.

12. **Hover information**: The hover information is activated by hovering the cursor over the rate glyph. The hover display gives information about the case type, the estimated rate, the ± 95% confidence intervals, the mileage analyzed to calculated the rate, and the estimated number of events observed.

13. **Download button**: The download button packages all the essential data, figures, and state of the tool into an Excel sheet ready to be saved on the user's computing device. The state of the tool is saved using a bookmark URL which can be used by the same or a different user to generate the same figures and populate the same data selection again. The download feature has three fundamental benefits for the end user. First, it enables convenient saving of the analysis. Second, it facilitates users to share their analysis with other team members easily. And finally, it provides a simple way for the

users to ingest the data and insights from the tool into the user's own data science tools such as MATLAB, Python, and R.

14. **Bookmark button**: The bookmark button saves the state of the analytics tool and provides the user with a URL that will open the tool in the exact same state.

15. **Help button**: The help button guides the user with the various features of the tool and helps them quickly get up to speed on how to use it.

16. **Plot facet**: The plot can be broken up into facets by using the "Separate by" drop-down menus. This allows simultaneous rate calculation and comparison for multiple subcategories.

17. **Case table**: Each rate glyph is an aggregate value representing several cases. A table containing all the cases represented by each glyph is activated by clicking on it as shown in Figure 7.6. This allows for the user to see each underlying case and update their mental models.

18. **Video viewer**: The video viewer is activated by clicking on the case row in the table and automatically starts playing a 6-second video of the driving segment from the ego vehicle front view perspective. The video snippet contains a significant amount of information about the scenario, roadway properties, and environmental conditions. This enables the user to quickly correlate the driving conditions to the visualized case type, conditions, etc., and update their mental models. The integration of video into traditional visualization methods accelerates the sense-making loop that is at the core of any interactive visualization task.

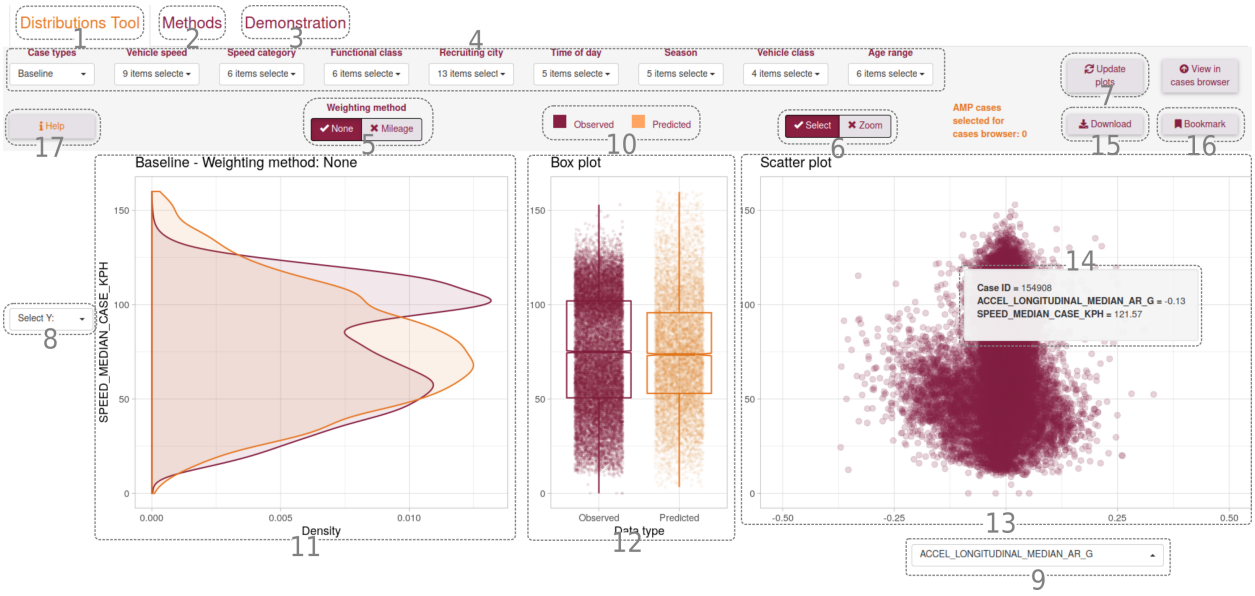


Figure 7.7: The major elements of the distributions tool.

7.5.3 Distributions Tool

7.5.3.1 Objective

The objective of the distributions tool is to illustrate the distribution of various kinematic, roadway, and environmental metrics for each case type in the AMP dataset. For example, what is the distribution of ego vehicle speed during crashes? This tool enables end users to understand the probability of experiencing a certain range of metrics such as speed and acceleration during a particular case type. Figures 7.7 and 7.8 show the various elements of the tool. A brief explanation of each element has been provided in the following sections.

7.5.3.2 Major Elements

1–3. **Primary navigation tabs**: The primary navigation tabs enable the user to navigate between the Distributions tool, the methods section, and the demonstration section. The various elements of the Analytics tab are explained below. The Methods tab

Figure 7.8: The interactive video elements of the distributions tool.

contains information about the methods used, the definition of the various metrics, and information about how to use the tool. The Demonstration tab contains a video explaining the various features of the tool. This high level navigation allows function-based compartmentalization of various aspects of the data analytics tools.

4. **Data selection drop-down menus**: The data selection drop-down menus enable the user to select the case type, ODD descriptors such as roadway speed category and functional class, environmental variables such as time of day and season, and driver demographics such as age and location. This ensures that the inferences can be drawn from data that is most appropriate for the user's needs.

5. **Weighting method selection**: The weighting method selection enables the user to appropriately weight the cases when fitting the data to the distributions.

6. **Click and drag selection**: This toggle switches the function of the click and drag functionality between selecting cases (brushing) and zooming into an area of the plot. The zoom in feature enables the user to enlarge an area of the plot to better visualize

or make more accurate selections in a region of interest.

7. **Plot update button**: The plot update button controls when the visualization updates after changes have been made to the data selection drop-down menus. This is a deliberate design choice to add an additional step for the user after changing the data selections. The alternative was to spontaneously update the plot as the user changed the data selection options. However, when making multiple changes, the plot would keep updating and would result in a suboptimal experience.

8. **Y-axis metric drop-down menu**: The y-axis drop-down menu lets the user select the y-axis variable for all three plots. Therefore, the y-axis is the primary metric choice for this tool. The y-axis drop-down menu gives the user a choice of hundreds of variables that are grouped by metric type and aggregation status. An example of what this drop-down menu looks like has been provided in Figure 7.10.

9. **X-axis metric drop-down menu**: The x-axis drop-down menu lets the user select the x-axis variable of the scatter plot. This choice does not affect the other plots. This drop-down menu has the same list of metrics as the last one. An example of what this drop-down menu looks like has been provided in Figure 7.10.

10. **Legend**: The legend illustrates that the maroon-colored marks are from the real data and the orange-colored marks are from the predicted or simulated data.

11. **Probability density function plot**: This plot shows two probability density functions. The orange function is a parametric density function created using by fitting the real-world data to a Weibull distribution or Gumbel distribution depending on the type of variable. The maroon function is created using a non-parametric kernel density function using the real data. Even though such plots are usually created with the variable on the x-axis and the density on the y-axis, we have chosen to plot the

variable on the y-axis and density on the x-axis. This allows the three plots to be analyzed together and lets the user derive more insight from the overall analytics tool. In case of categorical metrics, the distributions are replaced by the predicted and real frequencies of each category.

12. **Box plot**: The middle plot visualizes the real-world data in maroon and the simulated data in orange. The simulated data is created from the first 10,000 points predicted by the fitted distributions. The underlying data is plotted using a jitter, i.e. random noise along the x-axis and the actual value along the y-axis. This aids in visualization and gives the user a better idea about the density of the data. A box plot is visualized on top to let the users quickly glean information about the median and the interquartile range.

13. **Scatter plot**: The scatter plot enables the user to add another metric to the analysis and create a two-dimensional view. For example, as shown in Figure 7.7, the y-axis represents the median speed of the ego vehicle in kilometers/hour. The probability density function and the box plot only show this variable. In Figure 7.7, the x-axis of the scatter plot represents median longitudinal acceleration and therefore can help the user understand if the vehicle was accelerating, decelerating, or going at a steady state speed. Hovering over a data point displays the hover information, which contains the value of the two selected variables and the case ID. Clicking on a mark activates the video viewer as shown in Figure 7.8. Once activated, the video viewer automatically plays a 6-second video segment on loop around the time of interest. This helps the user to glean a large amount of information about the selected case and accelerates their sense-making process.

14. **Hover information**: The hover display is activated by hovering over a mark on the scatter plot. It contains information about the values of the two selected variables and

metadata such as the case ID.

15. **Download button**: The download button packages all the essential data, figures, and state of the tool into an Excel sheet ready to be saved on the user's computing device. The state of the tool is saved using a bookmark URL, which can be used by the same or a different user to generate the same figures and populate the same data selection again. The download feature has three fundamental benefits for the end user. First, it enables convenient saving of the analysis. Second, it facilitates users to share their analysis with other team members easily. And finally, it provides a simple way for the users to ingest the data and insights from the tool into the user's own data science tools such as MATLAB, Python, and R.

16. **Bookmark button**: The bookmark button saves the state of the analytics tool and provides the user with a URL that will open the tool in the exact same state.

17. **Help button**: The help button guides the user with the various features of the tool and helps them quickly get up to speed on how to use it.

18. **Video viewer**: The video viewer is activated by clicking on the scatter plot and automatically starts playing a 6-second video of the driving segment from the ego vehicle front view perspective. The video snippet contains a significant amount of information about the scenario, roadway properties, and environmental conditions. This enables the user to quickly correlate the driving conditions to the visualized metrics and update their mental models. The integration of video into traditional visualization methods accelerates the sense-making loop that is at the core of any interactive visualization task.

Figure 7.9: The major elements of the graphical filter tool.

7.5.4 Graphical Filter Tool

7.5.4.1 Objective

There are two primary objectives of the graphical filter tool. First, it enable users to develop a multivariate understanding of the case library by simultaneously visualizing six different variables. Second, it facilitates graphical filtering of cases through repeated paring down to achieve a desired subset of cases. Figures 7.9 and 7.10 show the various elements of the tool. A brief explanation of each element has been provided in the following sections.

7.5.4.2 Major Elements

1. **Graphical filter tool tab**: The main tab that contains the interactive plot elements.

2. **Methods tab**: This tab explains how to use the tool and provides a searchable table

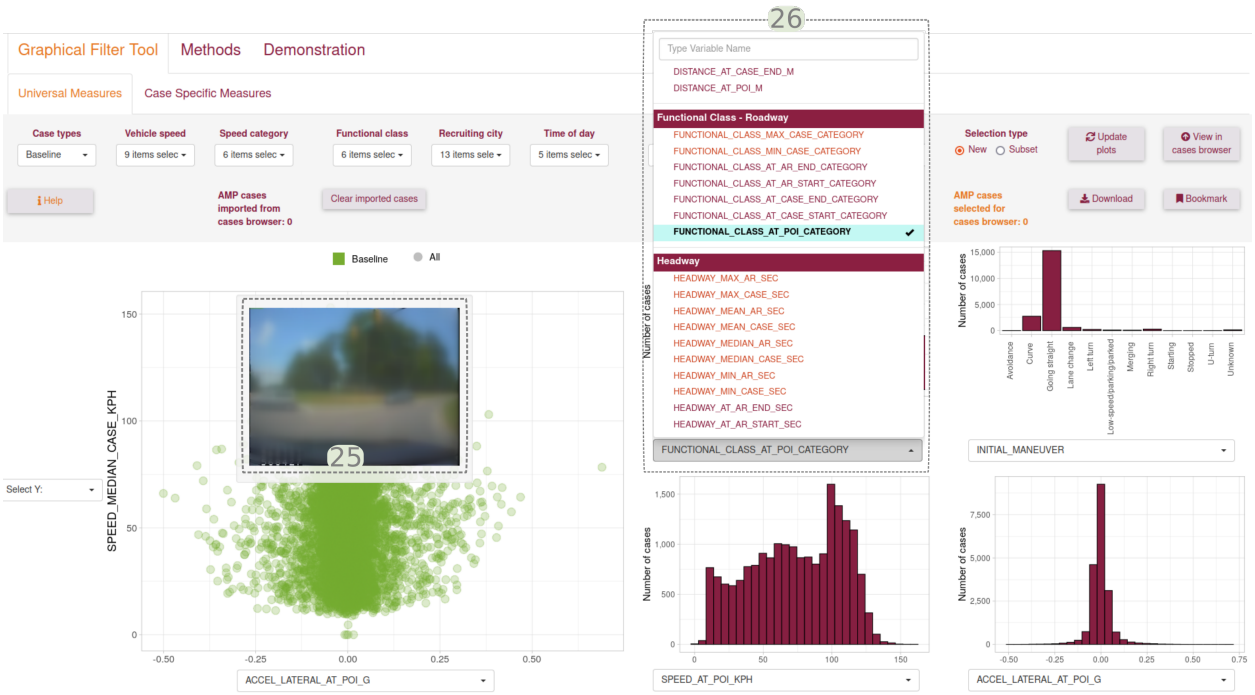


Figure 7.10: The interactive video elements of the graphical filter tool.

to look up metrics definitions, sources, and units.

3. **Demonstration tab**: This tab provides a video demonstrating the tool usage and basic features.

4. **Universal measures tab**: The universal measures tab provides the ability to simultaneously visualize multiple case types. The same metrics are available for all case types.

5. **Case specific measures tab**: This tab is similar to the universal measures tab but only allows visualizing one case type at a time. However, every case type has specific measures, which may not be available for other case types. This is the primary reason to create two types of visualization tabs, as having case-specific measures only available for some of the plotted cases could cause confusion.

6. **Data selection drop-down menus**: These menus enable the user to subset the data based on various data categories such case type, ego vehicle speed, roadway speed category and functional class, location, time of day, season, vehicle class, and age range. Such options allow the user to select the most appropriate cases based on scenario, roadway properties, environmental factors, and driver demographics.

7. **Selection type buttons**: This toggle button lets the user choose how the click and drag selection, also known as brushing, works. When the *New* radio button is selected, each click and drag operates as a new selection. When the *Subset* button is selected, each click and drag operates as an intersection on the previous selection, similar to how an AND operation works. This enables users to pare down the number of cases by making subsequent selections across different plots.

8. **Click and drag selection**: This toggle switches the function of the click and drag functionality between selecting cases (brushing) and zooming into an area of the plot. The zoom in feature enables the user to enlarge an area of the plot to better visualize or make more accurate selections in a region of interest.

9. **Plot update button**: The plot update button controls when the visualization updates after changes have been made to the data selection drop-down menus. This is a deliberate design choice to add an additional step for the user after changing the data selections. The alternative was to spontaneously update the plot as the user changed the data selection options. However, when making multiple changes, the plot would keep updating and would result in a suboptimal experience.

10–15. **Drop-down menus next to plots**: These drop-down menus control the metrics visualized on the plots. Each of the six drop-down menus contain the same list of variables, which are grouped into various categories such as roadway properties, vehicle

acceleration, etc. Clicking on a metric instantly updates the scatter plot while keeping the rest of the visualizations the same. The drop-downs have been placed right next to each axis to make the relationship obvious.

16. **Scatter plot**: The scatter plot displays all selected cases with color, shape, and transparency, conveying information about case type and selection status. The scatter plot enables the user to gain an idea about the relationship between the x and y variable. When the cursor is hovered over a mark, additional information about the case is displayed as shown by element 21 in Figure 7.9. This enables the user to interrogate the plot and extract additional information. The click and drag interaction either selects the brushed cases or zooms in to the region of the plot based on what is selected in element 8. Clicking on a mark results in a video player opening and playing a 6-second video snippet as shown by element 25 in Figure 7.10. As discussed previously, this interactive video on demand helps the user quickly update their mental model about the scenario.

17–20. **Distribution plots**: The four bar charts/histograms enable the user to quickly get an understanding of distributions of four variables at a time. The tool chooses between a bar chart and a histogram and automatically detects the appropriate type of figure for the selected variable. These plots also have click and drag brush functionality to either make new selections on further pare down already selected data.

21. **Hover information**: The hover information is triggered by hovering the cursor on a mark representing a data point. The hover overlay contains metadata about the case and the values of the metrics on the two axes.

22. **Download button**: The download button packages all the essential data, figures, and state of the tool into an Excel sheet ready to be saved on the user's computing

device. The state of the tool is saved using a bookmark URL, which can be used by the same or a different user to generate the same figures and populate the same data selection again. The download feature has three fundamental benefits for the end user. First, it enables convenient saving of the analysis. Second, it facilitates users to share their analysis with other team members easily. And finally, it provides a simple way for the users to ingest the data and insights from the tool into the user's own data science tools such as MATLAB, Python, and R.

23. **Bookmark button**: The bookmark button saves the state of the analytics tool and provides the user with a URL that will open the tool in the exact same state.

24. **Help button**: The help button guides the user with the various features of the tool and helps them quickly get up to speed on how to use it.

25. **Video viewer**: The video viewer is activated by clicking on the scatter plot and automatically starts playing a 6-second video of the driving segment from the ego vehicle front view perspective. The video snippet contains a significant amount of information about the scenario, roadway properties, and environmental conditions. This enables the user to quickly correlate the driving conditions to the visualized metrics and update their mental models. The integration of video into traditional visualization methods accelerates the sense-making loop that is at the core of any interactive visualization task.

26. **Metric selection drop-down menu**: This element shows the drop-down menu for selecting the metrics. Since there are hundreds of available metrics, they have been grouped in two ways. First, metrics representing similar physical qualities are grouped under one heading. Second, the metrics are color coded to show whether they represent a value taken at one timestamp or an aggregated value taken over a period. The

definitions, sources, and units of each metric are available in the Methods tab identified as element 2 in Figure 7.9.

7.6 Limitations of Analytics Tools and the Need for Guided Analytics

The interactive analytics tools successfully provide users four fundamental ways to analyze naturalistic driving data through the following lines of questioning:

1. How does the NDS crash dataset compare to the national crash dataset?
2. How often do specific scenarios occur in the NDS data?
3. What is the distribution of key metrics in each scenario?
4. What does a group of cases look like from a multidimensional perspective?

This information has been valuable for users to incorporate insights from the NDS data. However, these analytics tools usually require some training before new users can start benefiting from them. As explained in the previous section, several aids such as the help button, methods section, and the demonstration video were provided to help the user get familiar with the various functionalities. Nonetheless, it can still be challenging for a new user to fully understand all the context around the data used, the methods applied, and the various features of the analytics tool. This is further complicated by the nature of such tools, which do not always intuitively facilitate the user to ask the questions in a proper order.

Technical reports and papers have long been the standard for answering research questions through data analysis. As an information dissemination medium, they have been refined to

bring an unfamiliar reader up to speed through various sections like introduction, literature review, data, and methods. However, such reports lack interactivity and the open-ended nature of analytics tools that can often answer a whole set of research questions rather than just a few handpicked ones. They also lack the ability to incorporate more information-dense media such as video snippets that can be especially useful for transportation research. Therefore, there is a need for a new format that leverages the guided nature of reports and papers while retaining the interactivity and open-ended nature of interactive analytics tools. As a part of the AMP effort, such narrative-based interactive analytics tools have been created for analyzing various functional scenarios.

7.7 Narrative-based Interactive Analytics Tools for Analyzing Functional Scenarios

Functional scenarios offer an efficient way to develop and test ADS and ADAS systems. Large NDSs offer unique opportunities in creating functional scenarios by rooting test cases around real-world data. AMP has developed specific modules for analyzing various functional scenarios derived from NDSs. Analyzing such scenarios requires the user to understand various aspects of the data, such as how the scenarios were defined, what algorithms were used to detect them, and how the parameters were calculated. Using narrative-based interactive analytics tools are ideal for such problems as they help the user develop a nuanced understanding through descriptive text and figures while retaining the interactivity and open-ended nature of analytics tools.

Figures 7.11, 7.12, and 7.13 show screenshots of various elements of narrative-based analytics tool developed for analyzing various functional scenarios. All three figures show a navigation

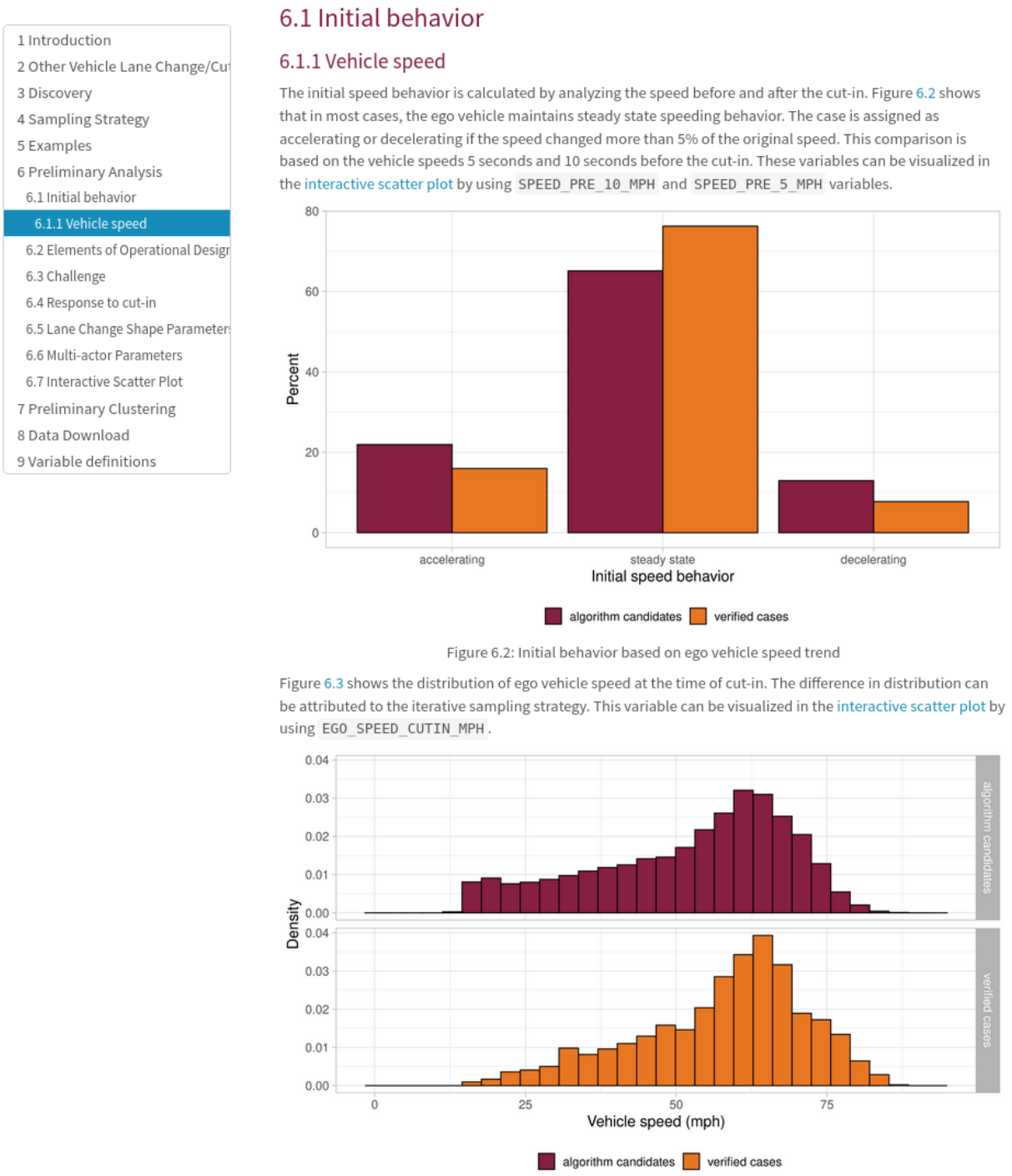


Figure 7.11: The functional scenario analytics tool provides descriptive narration that helps the user to quickly absorb context about the scenario.

Figure 7.12: The functional scenario analytics tool provides interactive plots with video overlay to accelerate the sense-making loop.

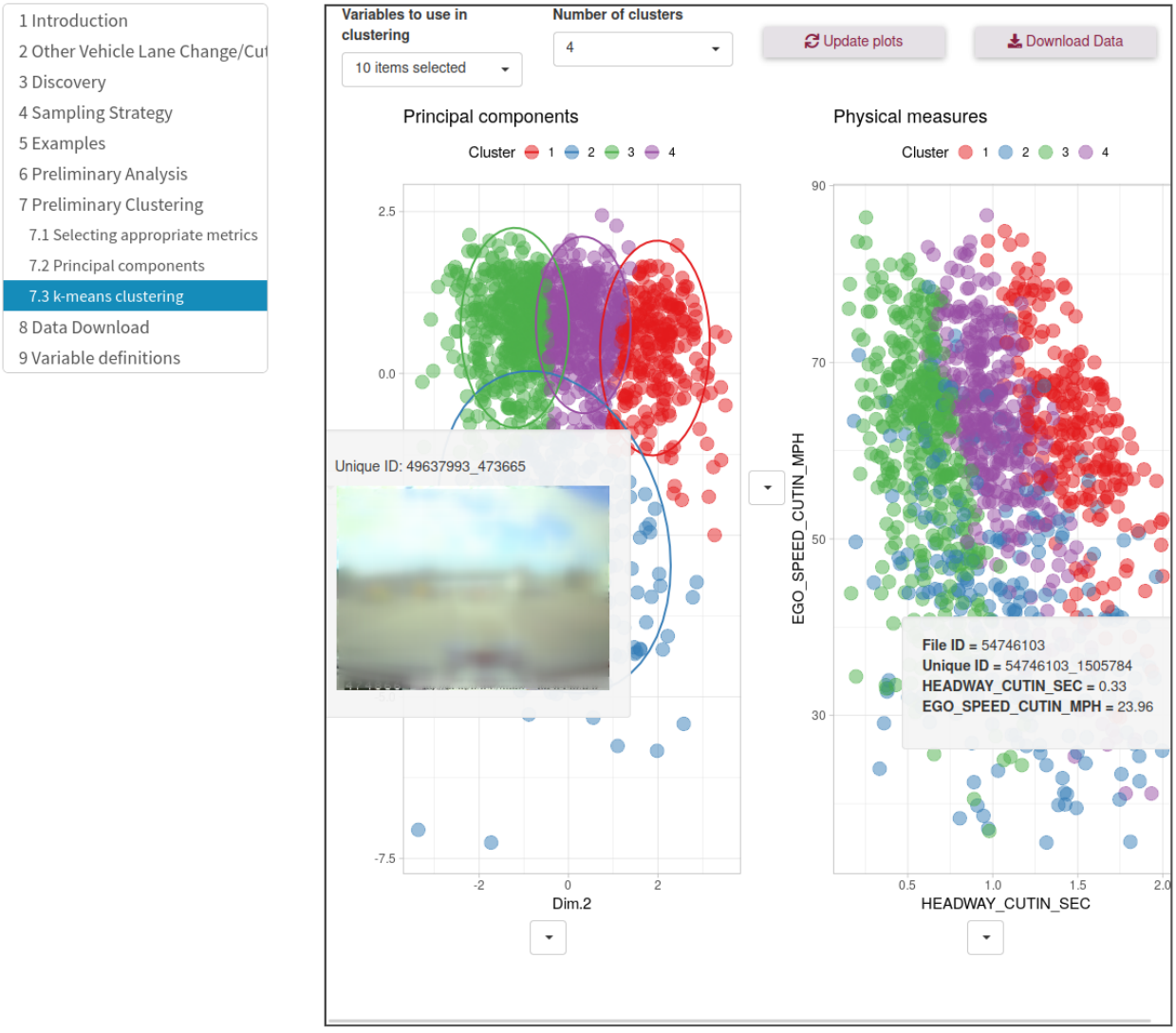


Figure 7.13: The functional scenario analytics tool has an embedded interactive clustering application to cluster cases and compare principal components with physical measures.

menu on the left that lets the user jump between various sections of the tool. Figure 7.11 shows descriptive text and figures being used to explain various aspects of the functional scenario. This is different from the previously discussed analytics tools which did not guide the user through the analysis. Various aspects of the functional scenario, such as the discovery, sampling strategy, and preliminary analysis, are explained through this medium.

Figure 7.12 shows one of the interactive applications embedded in the document that enables the user to select the variables representing the two axes of a scatter plot. The application also has an embedded video viewer that plays a 6-second video clip of a driving segment when activated by clicking on the mark representing it on the scatter plot. This sort of interactivity enables the user to quickly verify their assumptions about the selected metrics and update their mental models. As can be seen on the top of the scatter plot, the user can also download the underlying data, enabling them to continue their analysis using the tools of their choice.

Since one of the fundamental aims of functional scenario work is to find groups of representative cases, the interactive clustering tool shown in Figure 7.13 plays an important role in enabling the user to achieve their goals. The tool lets the user select the parameters of interest and the number of clusters required, and then performs a two-step clustering process to group the cases into the required number of clusters. In the first step, it performs a principal component analysis, and in the second step, it performs a k-means clustering process. As can be seen from the Figure 7.13, the two scatter plots let the user see the clusters plotted as the principal components and the physical metrics side by side. This helps the user correlate the physical metrics with the principal components, which otherwise do not have clear physical meaning. In addition to this, the on-demand video player that is activated by clicking on a glyph from either plot enables the user to see a 6-second video snippet of the case. This further accelerates the sense-making loop and provides the user an instantaneous method to

validate their mental models.

The narrative-based interactive analytics tools developed to analyze functional scenarios have proven to be a major improvement over existing methods of information dissemination. This format combines all the advantages of a technical report with the benefits of using interactive analytics tools. This has resulted in quicker on-boarding for new users as well as extracting deeper insights from the analyses.

7.8 Major Contributions

The discussion presented in this chapter describes the various aspects of the interactive analytics tools developed through the AMP program to extract insight from NDS data. Since these tools are used by engineers and researchers from several major OEMs and suppliers, they have the potential to positively influence the development and testing of ADAS and ADS. Our main focus has been to facilitate the use of NDS data for a broad range of user types by accelerating the sense-making loop. The major contributions can be summed up as follows:

1. A set of tools was created that enables users to derive insight from naturalistic driving studies in four fundamental ways. First, how do the NDS crashes compare with national crash datasets. Second, how often do certain scenarios occur in naturalistic driving data. Third, what is the distribution of important metrics in each scenario? And finally, what do the scenarios look like from a multidimensional perspective?

2. These tools were embedded with numerous interactive features that accelerate the sense-making loop. For example, hover information, on-click videos, and brushed selection enable the user to quickly glean a lot of information, update their mental

models, and propose new hypotheses.

3. A new paradigm for interactive data analysis was introduced that combines the salient features of descriptive technical reports and interactive analytics tools. This enabled us to guide new users through the various stages of scenario description, discovery, sampling, metric calculation, and preliminary analysis while retaining the ability to let them analyze the data on their own.

4. Numerous techniques were implemented to on-board new users through demonstration videos, product walkthrough guides, and detailed methods sections. Such tools enable users from diverse technical backgrounds to quickly get up to speed on how to use the various features of the analytics tools.

7.9 Author Statement on Credit

AMP has been a collaborative effort at VTTI with many contributors over several years. The author of this dissertation has played the following roles in the development of the tools described in this chapter.

- Leading the conception, development, and implementation of the analytics tools. This includes designing the interfaces, writing the software, and developing the various features such as user interactivity and data download.

- Leading the development of the data processing pipelines for the rates tool, distributions tool, graphical filter tool, and the functional scenarios. This includes designing the data structures, writing code to extract the various metrics, and packaging them into efficient formats. The author did not lead the development of the underlying

dataset for the database comparison tool. However, the author performed several processing tasks to make the data ready for the tool and was a part of the group that conceptualized the dataset.

- Leading the conceptualization and development of the narrative-based interactive analytics tools for the functional scenario analysis.
- Leading the conceptualization, development, and implementation of the video based interactive features.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In conclusion, this dissertation presents four major contributions towards furthering our understanding of driving behavior through the analysis of naturalistic driving data. These contributions can accelerate the development of automated driving systems and therefore help save human lives. The four major contributions are summarized in the following few paragraphs.

1. **The creation of the surface accelerations reference**. How vehicles speed up, slow down, or take turns is one of the most important aspects of driving behavior. However, comprehensive, large scale, and diverse datasets representing acceleration behavior were missing from literature. To fill this gap, over 34 million miles of driving data captured through the SHRP 2 NDS were analyzed to create acceleration-based driving profiles for the 3,500 participants. These profiles give information about how often acceleration events of various magnitudes occur on different roadway types in real world driving conditions. The profiles were made available for download and through an interactive web-based tool that lets users ask specific questions of the data. The standardized methodology for creating such a reference has also been made openly available so that similar profiles can be created for other driving cohorts.

2. **Quantifying the effect of various external factors on the rates of acceleration epochs.** Understanding how extrinsic factors affect acceleration rates is relevant for safety research as well as passenger comfort. However, there was a lack of studies that simultaneously modelled the effect of major factors based on large scale and real-world datasets. To fill this gap, the driving profiles provided in the surface accelerations reference were used to model the effect of roadway speed category, driver age, driver gender, vehicle class, and location. Roadway speed category was found to have the largest effect followed by driver age, vehicle class, and location. Driver gender was found to have the smallest effect on how people accelerated, braked, or took turns.

3. **Distilling acceleration based driving styles and understanding their relationship with crash risk.** After controlling for the effect of external factors on rates of acceleration, the intrinsic driving style of a person is the primary influence determining how they speed up, slow down, or take turns. However, there was a lack of studies extracting acceleration-based driving style using large real-world datasets. To fill this gap, a subset of the data from the accelerations reference representing driving on low-speed roads was used to extract driving styles. An unsupervised clustering algorithm was used to group 3,489 drivers into four driving styles called high, mid A, mid B, and low. Significant differences were observed when compared using crash rates, crash plus near-crash rates, and speeding behavior.

4. **Creating a set of interactive analytics tools that enable users to answer large sets of open-ended questions about real world driving.** The 70 million miles of naturalistic driving study data housed at VTTI contain valuable insight about real world driving that can be leveraged for the development and testing of ADS and ADAS. To facilitate this, interactive analytics tools were developed that offer users four types of insights. First, how does the naturalistic driving crash dataset compare

with national crash datasets? Second, how often do certain driving scenarios occur in real world driving conditions? Third, what is the distribution of key parameters such as deceleration or headway during a specific scenario? And fourth, how to gain a multidimensional understanding of a dataset? Novel techniques such as video on demand and guided analytics were developed to accelerate the sense-making process for users so that they can derive deeper insights in a shorter amount of time.

8.2 Future Work

The ideas presented in this dissertation have opened many possibilities of future work. Some of them that are currently being explored are:

1. Creating a maneuver-based database that describes the kinematic profiles of maneuvers while retaining key information about various ODD (Operational Design Domain) aspects. Maneuvers resulting in high g-force events will also be cataloged to help safety researchers better understand them. This is in direct continuation of the surface accelerations reference which has laid a foundation for this work.

2. Creating comprehensive driving profiles based on driving speed, speeding behavior, headway, and lane keeping habits. This too is in direct continuation of the surface accelerations reference which created a driving profile based purely on acceleration.

3. Extracting comprehensive driving styles based on metrics such as acceleration, speeding behavior, headway, lane keeping habits, and kinematic maneuver profiles. This will continue the driving style research presented in Chapter.

4. Continuing to integrate data from ongoing and new naturalistic driving studies into the surface accelerations reference and the interactive analytics tools.

5. Developing the next iteration of guided interactive analytics tools that enables users to extract insight about functional scenarios from multiple data sources simultaneously.

Bibliography

[1] *Policy on geometric design of highways and streets (5th edition).* American Association of State Highway and Transportation Officials (AASHTO), . ISBN 978-1-56051-263-9. URL http://www.knovel.com/knovel2/Toc.jsp?BookID=2528.

[2] Safety projects | pages | strategic highway research program 2 (SHRP 2), . URL http://www.trb.org/StrategicHighwayResearchProgram2SHRP2/Pages/Safety_Projects_304.aspx.

[3] CN Abernethy, GR Plank, E Donald Sussman, and HH Jacobs. *Effects of deceleration and rate of deceleration on live seated human subjects.* Urban Mass Transportation Administration.

[4] Rahmi Akçelik and Mark Besley. Acceleration and deceleration models. In *23rd conference of australian institutes of transport research (CAITR 2001), monash university, melbourne, australia*, pages 10–12.

[5] Giancarlo Alessandretti, Angelos Amditis, A. Etemad, and Christoph Kessler. EuroFOT: European large-scale field operational test on active safety systems. URL https://trid.trb.org/view/908400.

[6] Gibran Ali, Shane McLaughlin, and Mehdi Ahmadian. Quantifying the effect of roadway, driver, vehicle, and location characteristics on the frequency of longitudinal and lateral accelerations. 161:106356, . ISSN 0001-4575. doi: 10.1016/j.aap.2021.106356. URL https://www.sciencedirect.com/science/article/pii/S0001457521003870.

[7] Gibran Ali, Shane McLaughlin, and Mehdi Ahmadian. The surface accelerations reference — a large-scale, interactive catalog of passenger vehicle accelerations. .

[8] Yasir Ali, Michiel C. J. Bliemer, Zuduo Zheng, and Md. Mazharul Haque. Comparing the usefulness of real-time driving aids in a connected environment during mandatory and discretionary lane-changing manoeuvres. 121:102871, . ISSN 0968-090X. doi: 10.1016/j.trc.2020.102871. URL https://www.sciencedirect.com/science/article/pii/S0968090X20307713.

[9] Yasir Ali, Anshuman Sharma, Md. Mazharul Haque, Zuduo Zheng, and Mohammad Saifuzzaman. The impact of the connected environment on driving behavior and safety: A driving simulator study. 144:105643, . ISSN 0001-4575. doi: 10.1016/j.aap.2020.105643. URL https://www.sciencedirect.com/science/article/pii/S0001457520303092.

[10] Yasir Ali, Zuduo Zheng, Md. Mazharul Haque, and Meng Wang. A game theory-based approach for modelling mandatory lane-changing behaviour in a connected environment. 106:220–242, . ISSN 0968-090X. doi: 10.1016/j.trc.2019.07.011. URL https://www.sciencedirect.com/science/article/pii/S0968090X19302244.

[11] Jon Antin, Suzie Lee, Jon Hankey, and Tom Dingus. Design of the in-vehicle driving behavior and crash risk study: In support of the SHRP 2 naturalistic driving study, . URL http://www.nap.edu/catalog/14494.

[12] Jon Antin, Kelly Stulce, Lisa Eichelberger, Jon Hankey, Strategic Highway Research Program Safety Focus Area, Transportation Research Board, and National Academies of Sciences, Engineering, and Medicine. *Naturalistic driving study: descriptive comparison of the study sample with national data*. Transportation Research Board, . ISBN 978-0-309-43273-3. doi: 10.17226/22196. URL http://www.nap.edu/catalog/22196.

[13] Omar Bagdadi. Assessing safety critical braking events in naturalistic driving studies. 16:117–126. ISSN 1369-8478. doi: 10.1016/j.trf.2012.08.006. URL https://www.sciencedirect.com/science/article/pii/S1369847812000770.

[14] Omar Bagdadi and András Várhelyi. Jerky driving—an indicator of accident proneness? 43(4):1359–1363. ISSN 0001-4575. doi: 10.1016/j.aap.2011.02.009. URL https://www.sciencedirect.com/science/article/pii/S0001457511000236.

[15] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4.

[16] Mohammad Mahdi Bejani and Mehdi Ghatee. A context aware system for driving style evaluation by an ensemble learning on smartphone sensors data. 89:303–320. ISSN 0968-090X. doi: 10.1016/j.trc.2018.02.009. URL https://www.sciencedirect.com/science/article/pii/S0968090X18301931.

[17] Klaus Bengler, Klaus Dietmayer, Berthold Farber, Markus Maurer, Christoph Stiller, and Hermann Winner. Three decades of driver assistance systems: Review and future perspectives. 6(4):6–22. ISSN 1941-1197. doi: 10.1109/MITS.2014.2336271.

[18] Mohamed Benmimoun, Felix Fahrenkrog, Adrian Zlocki, and Lutz Eckstein. Incident detection based on vehicle CAN-data within the large scale field operational test "euroFOT". In *22nd enhanced safety of vehicles conference (ESV 2011), washington, DC/USA*.

[19] Gennaro Nicola Bifulco, Luigi Pariota, Mark Brackstione, and Michael Mcdonald. Driving behaviour models enabling the simulation of advanced driving assistance systems: revisiting the action point paradigm. 36:352–366. ISSN 0968-090X. doi: 10.1016/j.trc.2013.09.009. URL https://www.sciencedirect.com/science/article/pii/S0968090X13001897.

[20] Lawrence Blincoe, Ted R. Miller, Eduard Zaloshnja, and Bruce A. Lawrence. The economic and societal impact of motor vehicle crashes, 2010 (revised). URL https://trid.trb.org/view/1311862.

[21] Robert M Brooks. Acceleration characteristics of vehicles in rural pennsylvania.

[22] Jeremy Broughton. Car driver casualty rates in great britain by type of car. 40(4): 1543–1552. ISSN 0001-4575. doi: 10.1016/j.aap.2008.04.002. URL https://www.sciencedirect.com/science/article/pii/S0001457508000596.

[23] V. A. Butakov and P. Ioannou. Personalized driver assistance for signalized intersections using v2i communication. 17(7):1910–1919, . ISSN 1558-0016. doi: 10.1109/TITS.2016.2515023.

[24] V. A. Butakov and P. A. Ioannou. Driver/vehicle response diagnostic system for the vehicle-following case. 15(5):1947–1957, . ISSN 1558-0016. doi: 10.1109/TITS.2014.2305735.

[25] J. Bärgman, N. van Nes, M. Christoph, R. Jansen, V. Heijne, O. Carsten, M. Dotzauer, F. Utech, E. Svanberg, M. Pereira, F. Forcolin, J. Kovaceva, L. Guyonvarch, D. Hibberd, T. Lotan, M. Winkelbauer, F. Sagberg, E. Stemmler, H. Gellerman, C. Val, K. Quintero, H. Tattegrain, M. Donabauer, A. Pommer, I. Neumann, G. Albert, R. Welsh, and C. Fox. UDrive deliverable d41.1: The UDrive dataset and key analysis results. URL https://erticonetwork.com/wp-content/uploads/2017/12/UDRIVE-D41.1-UDrive-dataset-and-key-analysis-results-with-annotation-codebook.pdf.

[26] M. R. Carlos, L. C. González, J. Wahlström, G. Ramírez, F. Martínez, and G. Runger.

How smartphone accelerometers reveal aggressive driving behavior?—the key is the representation. 21(8):3377–3387. ISSN 1558-0016. doi: 10.1109/TITS.2019.2926639.

[27] G. Castignani, T. Derrmann, R. Frank, and T. Engel. Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. 7(1):91–102. ISSN 1941-1197. doi: 10.1109/MITS.2014.2328673.

[28] Winston Chang, Joe Cheng, Joseph J Allaire, Yihui Xie, and Jonathan McPherson. Shiny: web application framework for r. 1(4):106.

[29] Jaisung Choi, Richard Tay, Sangyoup Kim, and Seungwon Jeong. Turning movements, vehicle offsets and ageing drivers driving behaviour at channelized and unchannelized intersections. 108:227–233. ISSN 0001-4575. doi: 10.1016/j.aap.2017.08.029. URL https://www.sciencedirect.com/science/article/pii/S0001457517303093.

[30] S. G. Christopoulos, S. Kanarachos, and A. Chroneos. Learning driver braking behavior using smartphones, neural networks and the sliding correlation coefficient: Road anomaly case study. 20(1):65–74. ISSN 1558-0016. doi: 10.1109/TITS.2018.2797943.

[31] B. Ciuffo, M. Makridis, T. Toledo, and G. Fontaras. Capability of current car-following models to reproduce vehicle free-flow acceleration dynamics. 19(11):3594–3603. ISSN 1558-0016. doi: 10.1109/TITS.2018.2866271.

[32] Allison E. Curry, Melissa R. Pfeiffer, Dennis R. Durbin, and Michael R. Elliott. Young driver crash rates by licensing age, driving experience, and license phase. 80:243–250. ISSN 0001-4575. doi: 10.1016/j.aap.2015.04.019. URL https://www.sciencedirect.com/science/article/pii/S0001457515001566.

[33] T. A. Dingus, S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. D. Sudweeks, M. A. Perez, J. Hankey, D. J. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland,

and R. R. Knipling. The 100-car naturalistic driving study, phase II - results of the 100-car field experiment, . URL https://trid.trb.org/view/783477.

[34] Thomas A. Dingus, Feng Guo, Suzie Lee, Jonathan F. Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. 113(10):2636–2641, . ISSN 0027-8424. doi: 10.1073/pnas.1513271113. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4790996/. tex.pmcid: PMC4790996.

[35] Thomas A. Dingus, Jonathan M. Hankey, Jonathan F. Antin, Suzanne E. Lee, Lisa Eichelberger, Kelly Stulce, Doug McGraw, Miguel and Stowe Perez Loren, Strategic Highway Research Program Safety Focus Area, Transportation Research Board, and National Academies of Sciences, Engineering, and Medicine. *Naturalistic driving study: Technical coordination and quality control*. Transportation Research Board, . ISBN 978-0-309-43347-1. doi: 10.17226/22362. URL https://www.nap.edu/catalog/22362.

[36] Thomas A. Dingus, Vicki L. Neale, Sheila A. Garness, Richard J. Hanowski, Aysha S. Keisler, Suzanne E. Lee, Miguel A. Perez, G. S. Robinson, S. M. Belz, John G. Casali, E. F. Pace-Schott, R. A. Stickgold, and J. A. Hobson. Impact of sleeper berth usage on driver fatigue, final project report, . URL https://vtechworks.lib.vt.edu/handle/10919/55096.

[37] Yanchao Dong, Zhencheng Hu, Keiichi Uchimura, and Nobuki Murayama. Driver inattention monitoring system for intelligent vehicles: A review. 12(2):596–614. ISSN 1558-0016. doi: 10.1109/TITS.2010.2092770.

[38] Dominik Dörr, David Grabengiesser, and Frank Gauterin. Online driving style recognition using fuzzy logic. In *17th International IEEE Conference on Intelligent Trans-*

portation Systems (ITSC), pages 1021–1026. doi: 10.1109/ITSC.2014.6957822. ISSN: 2153-0017.

[39] Laura Eboli, Gabriella Mazzulla, and Giuseppe Pungillo. Combining speed and acceleration to define car users' safe or unsafe driving behaviour. 68:113–125. ISSN 0968-090X. doi: 10.1016/j.trc.2016.04.002. URL http://www.sciencedirect.com/science/article/pii/S0968090X16300067.

[40] Rune Elvik. Laws of accident causation. 38(4):742–747. ISSN 0001-4575. doi: 10.1016/j.aap.2006.01.005. URL https://www.sciencedirect.com/science/article/pii/S0001457506000145.

[41] H. Eren, S. Makinist, E. Akin, and A. Yilmaz. Estimating driving behavior by a smartphone. In *2012 IEEE intelligent vehicles symposium*, pages 234–239. doi: 10.1109/IVS.2012.6232298. tex.eventtitle: 2012 IEEE intelligent vehicles symposium.

[42] Daniel B Fambro, Kay Fitzpatrick, and Rodger J Koppa. *Determination of stopping sight distances*. Transportation Research Board.

[43] Federal Highway Administration. Section 1. introduction - highway functional classifications - related - statewide transportation planning - processes - planning - FHWA. URL https://www.fhwa.dot.gov/planning/processes/statewide/related/highway_functional_classifications/section01.cfm#Toc329359418.

[44] Fred Feng, Shan Bao, James R. Sayer, Carol Flannagan, Michael Manser, and Robert Wunderlich. Can vehicle longitudinal jerk be used to identify aggressive drivers? an examination using naturalistic driving data. 104:125–136. ISSN 0001-4575. doi: 10.1016/j.aap.2017.04.012. URL https://www.sciencedirect.com/science/article/pii/S0001457517301409.

[45] John B Ferris. Factors affecting perceptions of ride quality in automobiles. 65:6.

[46] GM Fitch, SE Lee, S Klauer, J Hankey, J Sudweeks, and T Dingus. Analysis of lane-change crashes and near-crashes.

[47] Inc Geotab. Geotab management by measurement.

[48] Ekaterina Gilman, Anja Keskinarkaus, Satu Tamminen, Susanna Pirttikangas, Juha Röning, and Jukka Riekki. Personalised assistance for fuel-efficient driving. 58: 681–705. ISSN 0968-090X. doi: 10.1016/j.trc.2015.02.007. URL https://www.sciencedirect.com/science/article/pii/S0968090X15000480.

[49] A. B. Rodríguez González, M. R. Wilby, J. J. Vinagre Díaz, and C. Sánchez Ávila. Modeling and detecting aggressiveness from driving signals. 15(4):1419–1428. ISSN 1558-0016. doi: 10.1109/TITS.2013.2297057.

[50] Feng Guo and Youjia Fang. Individual driver risk assessment using naturalistic driving data. 61:3–9. ISSN 0001-4575. doi: 10.1016/j.aap.2012.06.014. URL https://www.sciencedirect.com/science/article/pii/S0001457512002382.

[51] Feng Guo, Sheila G Klauer, Youjia Fang, Jonathan M Hankey, Jonathan F Antin, Miguel A Perez, Suzanne E Lee, and Thomas A Dingus. The effects of age on crash risk associated with driver distraction. 46(1):258–265. ISSN 0300-5771. doi: 10.1093/ije/dyw234. URL https://doi.org/10.1093/ije/dyw234.

[52] P. Handel, I. Skog, J. Wahlstrom, F. Bonawiede, R. Welch, J. Ohlsson, and M. Ohlsson. Insurance telematics: Opportunities and challenges with the smartphone solution. 6 (4):57–70. ISSN 1941-1197. doi: 10.1109/MITS.2014.2343262.

[53] Jonathan M Hankey, Miguel A Perez, and Julie A McClafferty. Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets, .

[54] Jonathan M. Hankey, Miguel A. Perez, and Julie A. McClafferty. Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets. . URL https://vtechworks.lib.vt.edu/handle/10919/70850. Accepted: 2016-04-25T21:11:58Z Publisher: Virginia Tech Transportation Institute.

[55] Richard J. Hanowski. The impact of local/short haul operations on driver fatigue. URL https://vtechworks.lib.vt.edu/handle/10919/28416.

[56] Richard J Hanowski, Walter Wierwille, Andrew Gellatly, Nancy Early, and Thomas Dingus. Impact of local short haul operations on driver fatigue.

[57] M. Mazharul Haque, Oscar Oviedo-Trespalacios, Ashim Kumar Debnath, and Simon Washington. Gap acceptance behavior of mobile phone–distracted drivers at roundabouts. 2602(1):43–51. ISSN 0361-1981, 2169-4052. doi: 10.3141/2602-06. URL http://journals.sagepub.com/doi/10.3141/2602-06.

[58] Here Technologies. FunctionalClassType - routing API - HERE developer, . URL https://developer.here.com/documentation/routing/topics/resource-type-functional-class.html.

[59] Here Technologies. SpeedCategoryType - geocoder API - HERE developer, . URL https://developer.here.com/documentation/geocoder/topics/resource-type-speed-category.html.

[60] B. Higgs and M. Abbas. Segmentation and clustering of car-following behavior: Recognition of driving patterns. 16(1):81–90. ISSN 1558-0016. doi: 10.1109/TITS.2014.2326082.

[61] L. L. Hoberock. A survey of longitudinal acceleration comfort studies in ground transportation vehicles. 99(2):76. ISSN 00220434. doi: 10.1115/1.3427093.

URL http://DynamicSystems.asmedigitalcollection.asme.org/article.aspx?articleid=1402622.

[62] C. Hyden. The development of a method for traffic safety evaluation: the swedish traffic conflicts technique. (70). URL https://trid.trb.org/view/239059.

[63] Alena Høye. Vehicle registration year, age, and weight – untangling the effects on crash risk. 123:1–11. ISSN 0001-4575. doi: 10.1016/j.aap.2018.11.002. URL https://www.sciencedirect.com/science/article/pii/S0001457518309242.

[64] Motonori Ishibashi, Masayuki Okuwa, Shun'ichi Doi, and Motoyuki Akamatsu. Indices for characterizing driving style and their relevance to car following behavior. In *SICE annual conference 2007*, pages 1132–1137. doi: 10.1109/SICE.2007.4421155. tex.eventtitle: SICE annual conference 2007.

[65] Jwan Kamla, Tony Parry, and Andrew Dawson. Analysing truck harsh braking incidents to study roundabout accident risk. 122:365–377. ISSN 0001-4575. doi: 10.1016/j.aap.2018.04.031. URL https://www.sciencedirect.com/science/article/pii/S0001457518301842.

[66] Nadezda Karginova, Stefan Byttner, and Magnus Svensson. Data-driven methods for classification of driving styles in buses. pages 2012–01–0744. doi: 10.4271/2012-01-0744. URL https://www.sae.org/content/2012-01-0744/.

[67] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, editors, *Information Visualization*, volume 4950, pages 154–175. Springer Berlin Heidelberg. ISBN 978-3-540-70955-8 978-3-540-70956-5. doi: 10.1007/978-3-540-70956-5_7. URL http:

//link.springer.com/10.1007/978-3-540-70956-5_7. ISSN: 0302-9743, 1611-3349 Series Title: Lecture Notes in Computer Science.

[68] Charlie Klauer, John Pearson, and Jon Hankey. An overview of the canada naturalistic driving and canada truck naturalistic driving studies. . URL https://www.vtti.vt.edu/PDFs/ndrs-2018/s4/Klauer.pdf.

[69] Sheila G. Klauer, Thomas A. Dingus, Vicki L. Neale, Jeremy D. Sudweeks, and D. J. Ramsey. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data, . URL https://vtechworks.lib.vt.edu/handle/10919/55090.

[70] Sheila G. Klauer, Thomas A. Dingus, Vicki L. Neale, Jeremy D. Sudweeks, and David J. Ramsey. Comparing real-world behaviors of drivers with high versus low rates of crashes and near crashes. . URL https://trid.trb.org/view/894387.

[71] Leanne Kmet and Colin Macarthur. Urban–rural differences in motor vehicle crash fatality and hospitalization rates among children and youth. 38(1):122–127. ISSN 0001-4575. doi: 10.1016/j.aap.2005.07.007. URL https://www.sciencedirect.com/science/article/pii/S0001457505001351.

[72] Scott Le Vine, Alireza Zolfaghari, and John Polak. Autonomous cars: The tension between occupant experience and intersection capacity. 52:1–14. ISSN 0968090X. doi: 10.1016/j.trc.2015.01.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0968090X15000042.

[73] G. Li, F. Zhu, X. Qu, B. Cheng, S. Li, and P. Green. Driving style classification based on driving operational pictures. 7:90180–90189, . ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2926494.

[74] Guofa Li, Shengbo Eben Li, Bo Cheng, and Paul Green. Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities. 74:113–125, . ISSN 0968-090X. doi: 10.1016/j.trc.2016.11.011. URL https://www.sciencedirect.com/science/article/pii/S0968090X16302273.

[75] Li Li, Ding Wen, Nan-Ning Zheng, and Lin-Cheng Shen. Cognitive cars: A new frontier for ADAS research. 13(1):395–407, . ISSN 1558-0016. doi: 10.1109/TITS.2011.2159493.

[76] Todd Litman. Pay-as-you-drive pricing and insurance regulatory objectives. 23(3).

[77] Dale Litwhiler and Delton Martin. An investigation of acceleration and jerk profiles of public transportation vehicles. page 13.

[78] Gary Long. Acceleration characteristics of starting vehicles. 1737(1):58–70. ISSN 0361-1981, 2169-4052. doi: 10.3141/1737-08. URL http://journals.sagepub.com/doi/10.3141/1737-08.

[79] Anders Longthorne, Rajesh Subramanian, and Chou-Lin Chen. An analysis of the significant decline in motor vehicle traffic fatalities in 2008. URL https://trid.trb.org/view/934337.

[80] Dominique Lord and Fred Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. 44(5):291–305. ISSN 0965-8564. doi: 10.1016/j.tra.2010.02.001. URL https://www.sciencedirect.com/science/article/pii/S0965856410000376.

[81] M. Van Ly, S. Martin, and M. M. Trivedi. Driver classification and driving style recognition using inertial sensors. In *2013 IEEE intelligent vehicles symposium (IV)*, pages 1040–1045. doi: 10.1109/IVS.2013.6629603.

[82] Nengchao Lyu, Yue Cao, Chaozhong Wu, Jin Xu, and Lian Xie. The effect of gender, occupation and experience on behavior while driving on a freeway deceleration lane based on field operational test data. 121:82–93. ISSN 0001-4575. doi: 10.1016/j.aap.2018.07.034. URL https://www.sciencedirect.com/science/article/pii/S0001457518303750.

[83] Clara Marina Martinez, Mira Heucke, Fei-Yue Wang, Bo Gao, and Dongpu Cao. Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. 19(3):666–676. ISSN 1558-0016. doi: 10.1109/TITS.2017.2706978.

[84] Dawn L. Massie, Kenneth L. Campbell, and Allan F. Williams. Traffic accident involvement rates by driver age and gender. 27(1):73–87. ISSN 0001-4575. doi: 10.1016/0001-4575(94)00050-V. URL https://www.sciencedirect.com/science/article/pii/000145759400050V.

[85] Anne T. McCartt, Daniel R. Mayhew, Keli A. Braitman, Susan A. Ferguson, and Herbert M. Simpson. Effects of age and experience on young driver crashes: Review of recent literature. 10(3):209–219. ISSN 1538-9588. doi: 10.1080/15389580802677807. URL https://doi.org/10.1080/15389580802677807.

[86] Daniel V. McGehee, Mireille Raby, Cher Carney, John D. Lee, and Michelle L. Reyes. Extending parental mentoring using an event-triggered video intervention in rural teen drivers. 38(2):215–227. ISSN 00224375. doi: 10.1016/j.jsr.2007.02.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0022437507000321.

[87] Shane B. McLaughlin and Jonathan M. Hankey. Matching GPS records to digital map data: Algorithm overview and application. URL https://vtechworks.lib.vt.edu/handle/10919/51585.

[88] Gys Albertus Marthinus Meiring and Hermanus Carel Myburgh. A review of intelligent driving style analysis systems and related artificial intelligence algorithms. 15(12): 30653–30682. doi: 10.3390/s151229822. URL https://www.mdpi.com/1424-8220/15/12/29822.

[89] John A. Michon. A critical view of driver behavior models: What do we know, what should we do? In Leonard Evans and Richard C. Schwing, editors, *Human behavior and traffic safety*, pages 485–524. Springer US. ISBN 978-1-4613-2173-6. doi: 10.1007/978-1-4613-2173-6_19. URL https://doi.org/10.1007/978-1-4613-2173-6_19.

[90] C. Miyajima, H. Ukai, A. Naito, H. Amata, N. Kitaoka, and K. Takeda. Driver risk evaluation based on acceleration, deceleration, and steering behavior. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1829–1832. doi: 10.1109/ICASSP.2011.5946860.

[91] Amin Mohammadnazar, Ramin Arvin, and Asad J. Khattak. Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning. 122:102917. ISSN 0968-090X. doi: 10.1016/j.trc.2020.102917. URL https://www.sciencedirect.com/science/article/pii/S0968090X20308160.

[92] Yi Lu Murphey, Robert Milton, and Leonidas Kiliaris. Driver's style classification using jerk analysis. In *2009 IEEE workshop on computational intelligence in vehicles and vehicular systems*, pages 23–28. tex.organization: IEEE.

[93] Oren Musicant and Liat Lampel. When technology tells novice drivers how to drive. 2182(1):8–15. ISSN 0361-1981, 2169-4052. doi: 10.3141/2182-02. URL http://journals.sagepub.com/doi/10.3141/2182-02.

[94] Oren Musicant, Hillel Bar-Gera, and Edna Schechtman. Temporal perspective on individual driver behavior using electronic records of undesirable events. 70:55–64. ISSN 00014575. doi: 10.1016/j.aap.2014.03.008. URL https://linkinghub.elsevier.com/retrieve/pii/S0001457514000748.

[95] Thomas Müller, Hermann Hajek, Ljubica Radić-Weißenfeld, and Klaus Bengler. Can you feel the difference? the just noticeable difference of longitudinal acceleration. 57(1):1219–1223. ISSN 1541-9312. doi: 10.1177/1541931213571271. URL http://journals.sagepub.com/doi/10.1177/1541931213571271.

[96] National Highway Traffic Safety Administration. National motor vehicle crash causation survey: Report to congress, .

[97] National Highway Traffic Safety Administration. Traffic safety facts annual report tables, . URL https://cdan.nhtsa.gov/tsftables/tsfar.htm.

[98] Vicki Neale, Sheila Klauer, Thomas Dingus, Jeremy Sudweeks, and Michael Goodman. An overview of the 100-car naturalistic study and findings. page 10.

[99] Marie Claude Ouimet, Thomas G. Brown, Feng Guo, Sheila G. Klauer, Bruce G. Simons-Morton, Youjia Fang, Suzanne E. Lee, Christina Gianoulakis, and Thomas A. Dingus. Higher crash and near-crash rates in teenaged drivers with lower cortisol response: An 18-month longitudinal, naturalistic study. 168(6):517–522. ISSN 2168-6203. doi: 10.1001/jamapediatrics.2013.5387. URL https://doi.org/10.1001/jamapediatrics.2013.5387.

[100] H Ozaki. Reaction and anticipation in the car-following behavior. In *Proc. of 12th international symposium on theory of traffic flow and transportation*, pages 349–366.

[101] Marika Paaver, Diva Eensoo, Katrin Kaasik, Mariliis Vaht, Jarek Mäestu, and Jaanus

Harro. Preventing risky driving: A novel and efficient brief intervention focusing on acknowledgement of personal risk factors. 50:430–437. ISSN 0001-4575. doi: 10.1016/j.aap.2012.05.019. URL https://www.sciencedirect.com/science/article/pii/S0001457512002035.

[102] Johannes Paefgen, Thorsten Staake, and Elgar Fleisch. Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. 61:27–40. ISSN 0965-8564. doi: 10.1016/j.tra.2013.11.010. URL https://www.sciencedirect.com/science/article/pii/S096585641300236X.

[103] Eun Sug Park, Kay Fitzpatrick, Subasish Das, and Raul Avelar. Exploration of the relationship among roadway characteristics, operating speed, and crashes for city streets using path analysis. 150:105896. ISSN 0001-4575. doi: 10.1016/j.aap.2020.105896. URL https://www.sciencedirect.com/science/article/pii/S0001457520317164.

[104] Miguel A. Perez, Jeremy D. Sudweeks, Edie Sears, Jonathan Antin, Suzanne Lee, Jonathan M. Hankey, and Thomas A. Dingus. Performance of basic kinematic thresholds in the identification of crash and near-crash events within naturalistic driving data. 103:10–19. ISSN 0001-4575. doi: 10.1016/j.aap.2017.03.005. URL https://www.sciencedirect.com/science/article/pii/S0001457517301033.

[105] Virginia Petraki, Apostolos Ziakopoulos, and George Yannis. Combined impact of road and traffic characteristic on driver behavior using smartphone sensor data. 144:105657. ISSN 0001-4575. doi: 10.1016/j.aap.2020.105657. URL https://www.sciencedirect.com/science/article/pii/S0001457519315933.

[106] Ting Qu, Hong Chen, Dongpu Cao, Hongyan Guo, and Bingzhao Gao. Switching-

based stochastic model predictive control approach for modeling driver steering skill. 16(1):365–375. ISSN 1558-0016. doi: 10.1109/TITS.2014.2334623.

[107] H. A. Rakha, K. Ahn, W. Faris, and K. S. Moran. Simple vehicle powertrain model for modeling intelligent vehicle applications. 13(2):770–780. ISSN 1558-0016. doi: 10.1109/TITS.2012.2188517.

[108] G. Anthony Ryan, Matthew Legge, and Diana Rosman. Age related changes in drivers' crash risk and crash type. 30(3):379–387. ISSN 0001-4575. doi: 10.1016/S0001-4575(97)00098-5. URL https://www.sciencedirect.com/science/article/pii/S0001457597000985.

[109] Fridulv Sagberg, Selpi, Giulio Francesco Bianchi Piccinini, and Johan Engström. A review of research on driving styles and road safety. 57(7):1248–1275. ISSN 0018-7208. doi: 10.1177/0018720815591313. URL https://doi.org/10.1177/0018720815591313.

[110] J. M. Scanlon, R. Sherony, and H. C. Gabler. Models of driver acceleration behavior prior to real-world intersection crashes. 19(3):774–786. ISSN 1558-0016. doi: 10.1109/TITS.2017.2699079.

[111] Chris Schwarz, Tad Gates, and Yiannis Papelis. Motion characteristics of the national advanced driving simulator. In *Proceedings of the driving simulation conference*.

[112] Anshuman Sharma, Zuduo Zheng, Jiwon Kim, Ashish Bhaskar, and Md Mazharul Haque. Is an informed driver a better decision maker? a grouped random parameters with heterogeneity-in-means approach to investigate the impact of the connected environment on driving behaviour in safety-critical situations. 27:100127. ISSN 2213-6657. doi: 10.1016/j.amar.2020.100127. URL https://www.sciencedirect.com/science/article/pii/S2213665720300178.

[113] Iván Silva and José Eugenio Naranjo. A systematic methodology to evaluate prediction models for driving style classification. 20(6):1692. ISSN 1424-8220. doi: 10.3390/s20061692. URL https://www.mdpi.com/1424-8220/20/6/1692.

[114] B. G. Simons-Morton, Z. Zhang, J. C. Jackson, and P. S. Albert. Do elevated gravitational-force events while driving predict crashes and near crashes? 175 (10):1075–1079, . ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwr440. URL https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwr440.

[115] Bruce G. Simons-Morton, Feng Guo, Sheila G. Klauer, Johnathon P. Ehsani, and Anuj K. Pradhan. Keep your eyes on the road: Young driver crash risk increases according to duration of distraction. 54(5):S61–S67, . ISSN 1054-139X. doi: 10.1016/j.jadohealth.2013.11.021. URL https://www.sciencedirect.com/science/article/pii/S1054139X13007799.

[116] Bruce G. Simons-Morton, Marie Claude Ouimet, Jing Wang, Sheila G. Klauer, Suzanne E. Lee, and Thomas A. Dingus. Hard braking events among novice teenage drivers by passenger characteristics. 2009:236–242, . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3019610/. tex.pmcid: PMC3019610.

[117] S. Singh. Critical reasons for crashes investigated in the national motor vehicle crash causation survey. URL https://trid.trb.org/view/1507603.

[118] C. C. Smith, D. Y. McGehee, and A. J. Healey. The prediction of passenger riding comfort from acceleration data. 100(1):34. ISSN 00220434. doi: 10.1115/1.3426338. URL http://DynamicSystems.asmedigitalcollection.asme.org/article.aspx?articleid=1402752.

[119] C. Sohn, J. Andert, and R. N. Nanfah Manfouo. A driveability study on automated

longitudinal vehicle control. 21(8):3273–3280. ISSN 1558-0016. doi: 10.1109/TITS.2019.2925193.

[120] Joshua Stipancic, Luis Miranda-Moreno, and Nicolas Saunier. Vehicle manoeuvers as surrogate safety measures: Extracting data from the gps-enabled smartphones of regular drivers. 115:160–169. ISSN 0001-4575. doi: 10.1016/j.aap.2018.03.005. URL https://www.sciencedirect.com/science/article/pii/S000145751830109X.

[121] STMicroelectronics. LIS3lv02dq STMicroelectronics. URL https://www.mouser.in/ProductDetail/511-LIS3LV02DQ.

[122] Radoslav Stoichkov. Android smartphone application for driving style recognition. page 59.

[123] Jonathan Sung, Krista Mizenko, Randolph Atkins Jr, and Heidi Coleman. A comparative analysis of state traffic safety countermeasures and implications for progress "toward zero deaths" in the united states. URL https://trid.trb.org/view/1466521.

[124] Evgenia Suzdaleva and Ivan Nagy. An online estimation of driving style using data-dependent pointer model. 86:23–36. ISSN 0968-090X. doi: 10.1016/j.trc.2017.11.001. URL https://www.sciencedirect.com/science/article/pii/S0968090X17302991.

[125] Andrew P. Tarko. Use of crash surrogates and exceedance statistics to estimate road safety. 45:230–240. ISSN 0001-4575. doi: 10.1016/j.aap.2011.07.008. URL http://www.sciencedirect.com/science/article/pii/S000145751100193X.

[126] Orit Taubman-Ben-Ari, Mario Mikulincer, and Omri Gillath. The multidimensional driving style inventory—scale construct and validation. 36(3):323–332. ISSN 00014575.

doi: 10.1016/S0001-4575(03)00010-1. URL https://linkinghub.elsevier.com/retrieve/pii/S0001457503000101.

[127] Brian Tefft. Rates of motor vehicle crashes, injuries, and deaths in relation to driver age, united states, 2014 – 2015. URL https://trid.trb.org/view/1566515.

[128] J.J. Thomas and K.A. Cook. A visual analytics agenda. 26(1):10–13. ISSN 0272-1716. doi: 10.1109/MCG.2006.5. URL http://ieeexplore.ieee.org/document/1573625/.

[129] W. A. Tillmann and G. E. Hobbs. The accident-prone automobile driver. 106 (5):321–331. ISSN 0002-953X. doi: 10.1176/ajp.106.5.321. URL https://ajp.psychiatryonline.org/doi/abs/10.1176/ajp.106.5.321.

[130] Tomer Toledo, Haris N. Koutsopoulos, and Moshe Ben-Akiva. Integrated driving behavior modeling. 15(2):96–112, . ISSN 0968-090X. doi: 10.1016/j.trc.2007.02.002. URL http://www.sciencedirect.com/science/article/pii/S0968090X07000046.

[131] Tomer Toledo, Oren Musicant, and Tsippy Lotan. In-vehicle data recorders for monitoring and feedback on drivers' behavior. 16(3):320–331, . ISSN 0968-090X. doi: 10.1016/j.trc.2008.01.001. URL https://www.sciencedirect.com/science/article/pii/S0968090X08000041.

[132] J. F. Torres. Acceleration noise, power spectra, and other statistics derived from instrumented vehicle measurements under freeway driving conditions. (308). URL https://trid.trb.org/view/116391.

[133] Dimitrios I. Tselentis, George Yannis, and Eleni I. Vlahogianni. Innovative motor insurance schemes: A review of current practices and emerging challenges. 98:139–148.

ISSN 0001-4575. doi: 10.1016/j.aap.2016.10.006. URL https://www.sciencedirect.com/science/article/pii/S0001457516303670.

[134] V. Vaitkus, P. Lengvenis, and G. Žylius. Driving style classification using long-term accelerometer information. In *2014 19th international conference on methods and models in automation and robotics (MMAR)*, pages 641–644. doi: 10.1109/MMAR.2014.6957429.

[135] Jarke J Van Wijk. The value of visualization. In *VIS 05. IEEE visualization, 2005.*, pages 79–86.

[136] Daniel Vankov, Ronald Schroeter, and Divera Twisk. Understanding the predictors of young drivers' speeding intention and behaviour in a three-month longitudinal study. 151:105859. ISSN 0001-4575. doi: 10.1016/j.aap.2020.105859. URL https://www.sciencedirect.com/science/article/pii/S0001457520316791.

[137] Fei-Yue Wang. Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications. 11(3):630–638. ISSN 1558-0016. doi: 10.1109/TITS.2010.2060218.

[138] Jun Wang, Karen K. Dixon, Hainan Li, and Jennifer Ogle. Normal acceleration behavior of passenger vehicles starting from rest at all-way stop-controlled intersections. 1883(1):158–166. ISSN 0361-1981, 2169-4052. doi: 10.3141/1883-18. URL http://journals.sagepub.com/doi/10.3141/1883-18.

[139] C. N. Webb. Motor vehicle traffic crashes as a leading cause of death in the united states, 2015. URL https://trid.trb.org/View/1503884.

[140] Hadley Wickham. *ggplot2: Elegant graphics for data analysis*. Use r! Springer-

Verlag. ISBN 978-0-387-98141-3. doi: 10.1007/978-0-387-98141-3. URL https://www.springer.com/gp/book/9780387981413.

[141] Hadley Wickham, Romain Francois, Lionel Henry, and K Müller. dplyr: A grammar of data manipulation. 3.

[142] Allan F Williams. Teenage drivers: patterns of risk. 34(1):5–15. ISSN 0022-4375. doi: 10.1016/S0022-4375(02)00075-0. URL https://www.sciencedirect.com/science/article/pii/S0022437502000750.

[143] Kun-Feng Wu and Paul P. Jovanis. Defining and screening crash surrogate events using naturalistic driving data. 61:10–22, . ISSN 0001-4575. doi: 10.1016/j.aap.2012.10.004. URL https://www.sciencedirect.com/science/article/pii/S0001457512003600.

[144] Kun-Feng Wu and Paul P. Jovanis. Screening naturalistic driving study data for safety-critical events. 2386(1):137–146, . ISSN 0361-1981. doi: 10.3141/2386-16. URL https://doi.org/10.3141/2386-16.

[145] Li Xu, Jie Hu, Hong Jiang, and Wuqiang Meng. Establishing style-oriented driver models by imitating human driving behaviors. 16(5):2522–2530. ISSN 1558-0016. doi: 10.1109/TITS.2015.2409870. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[146] Yilin Zhao. Telematics: safe and fun driving. 17(1):10–14. ISSN 1941-1294. doi: 10.1109/5254.988442.

[147] Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, and Cheng Chen. Data-driven intelligent transportation systems: A survey. 12(4):1624–1639. ISSN 1558-0016. doi: 10.1109/TITS.2011.2158001.

[148] Bing Zhu, Yuande Jiang, Jian Zhao, Rui He, Ning Bian, and Weiwen Deng. Typical-driving-style-oriented personalized adaptive cruise control design based on human driving data. 100:274–288. ISSN 0968-090X. doi: 10.1016/j.trc.2019.01.025. URL https://www.sciencedirect.com/science/article/pii/S0968090X18306521.

[149] Türker Özkan, Timo Lajunen, Joannes El. Chliaoutakis, Dianne Parker, and Heikki Summala. Cross-cultural differences in driving behaviours: A comparison of six countries. 9(3):227–242. ISSN 1369-8478. doi: 10.1016/j.trf.2006.01.002. URL https://www.sciencedirect.com/science/article/pii/S1369847806000039.

Appendices

Appendix A

Signal Processing

A.1 Introduction

For SHRP 2 NDS, each of the 3,500 vehicles were installed with identical data acquisition systems (DAS) that processed and stored signals from an extensive sensor suite consisting of accelerometer, gyroscope, GPS, vehicle CAN, and multiple video feeds. Since the DAS has limited storage space, it was necessary to down sample some signals such as the accelerometer output to ensure that the hard drives did not fill up in less than four months. The main purpose of the acceleration signal was to determine periods of high g-force events such as hard braking, acceleration, swerving, and turns. In this appendix, the accelerometer signal acquisition and processing will be discussed. It is essential to determine whether the acceleration signal has been appropriately processed to avoid artifacts such as peak distortion and aliasing. For this, the signal acquisition is described, the frequency response of the applied filter is examined, and the power spectral density of the unfiltered signal is checked to see if the the applied filter is adequately processing the signal.

A.2 Signal Ingestion and Processing

Each DAS used in SHRP 2 NDS had a 3-axis MEMS linear accelerometer with three integrated $\Sigma\Delta$ analog to digital converters (ADC). It has a selectable full scale of $\pm$ 2g or $\pm$

6g and capable of sampling acceleration at a rate of 640 Hz for all axes [121]. The sensing element consists of micro-machined suspended silicon structures attached to a substrate but free to move in the direction of the measured acceleration. When an acceleration is applied, it causes the suspended mass to displace producing a analog voltage. The three analog acceleration signals are input to a MUX (multiplexer) where they are each sampled at 20.5 kHz by a single charge amplifier. The charge amplifier is operating at 61.5 kHz, which is three times faster than the sample rate fo each analog input because there are three inputs and only one charge amplifier. It should be noted that the data is already sampled at this stage and is no longer analog. The demuxed outputs of the charge amplifier are then sent to three independent ΣΔ ADCs operating at 20.5 kHz. The output of the ΣΔ ADCs is input to a digital decimation filter which has one of four selectable output data rates (ODR) of 2560 Hz, 640 Hz, 160 Hz, or 40 Hz.

The purpose of the decimation filter is to insure that the subsampled output data is not aliased, i.e. it has no frequency components above the Nyquist frequency associated with the ODR. According to the data sheet, the cut-off frequency of the digital decimation filter is always $\frac{ODR}{4}$. We can assume that the "cut-off frequency" means the "-3 dB point". The data sheet does not provide any information about the decimation filter and therefore, we need to use engineering judgment to analyze what might be happening.

Typically, ΣΔ ADCs use either first order or third order Sinc filters for the digital decimation stage. An N^{th} order Sinc filter is given by:

$$H(z) = \left(\frac{1 - Z^{-OSR}}{1 - Z^{-1}} \right)^{N} \tag{A.1}$$

where OSR is the oversample rate and is given by $OSR = \frac{fs}{ODR}$ and $fs = 20.5kHz$. Normally, the OSR values are integers but based on the numbers provided in the datasheet, the

calculated OSR values for these accelerometers are fractional. This probably means that the manufacturer is not using a standard Sinc filter. If we assume that a conventional first-order low-pass decimation filter is used, we will only get about 7 dB rejection at Nyquist for the ODR, which is not adequate. If we assume a conventional second-order low-pass decimation filter was used, then we would get about -12 dB which is better but still not ideal.

Since it has been about 15 years since the data acquisition system was designed, the constraints and the reasoning behind some design decisions cannot be fully known. However, we can use our engineering judgment to determine if some design decisions are problematic and should be avoided in future data acquisition systems. For the purposes of SHRP 2 NDS, the DAS designers needed to capture the acceleration data continuously at a relatively low frequency (∼10 Hz) and at a higher frequency (∼500 Hz) during periods of high g-force events. For the data acquisition system, the accelerometer output data rate was set at 640 Hz. This 640 Hz signal was sub-sampled at 500 Hz, then decimated with a 50 point moving average filter and then down-sampled to 10 Hz.

There are two major problems with this approach. First, the 640 Hz signal should not have been sub-sampled at 500 Hz. This is because a jitter is introduced in the signal which means the samples are not collected at a uniform sample rate. They are either collected before or after the actual sample time with the worst case of jitter error of 1.5 ms or 75% of the total sample period. The high frequency signal should have been sampled and stored at the ODR of 640 Hz. The second major problem is the 50 point moving average filter used for decimating the 500 Hz signal. As is discussed in the next section, this filter has an inadequate rejection of 4 dB at the Nyquist frequency of 5 Hz. A good anti-alias filter should provide between 10 and 20 dB rejection at Nyquist.

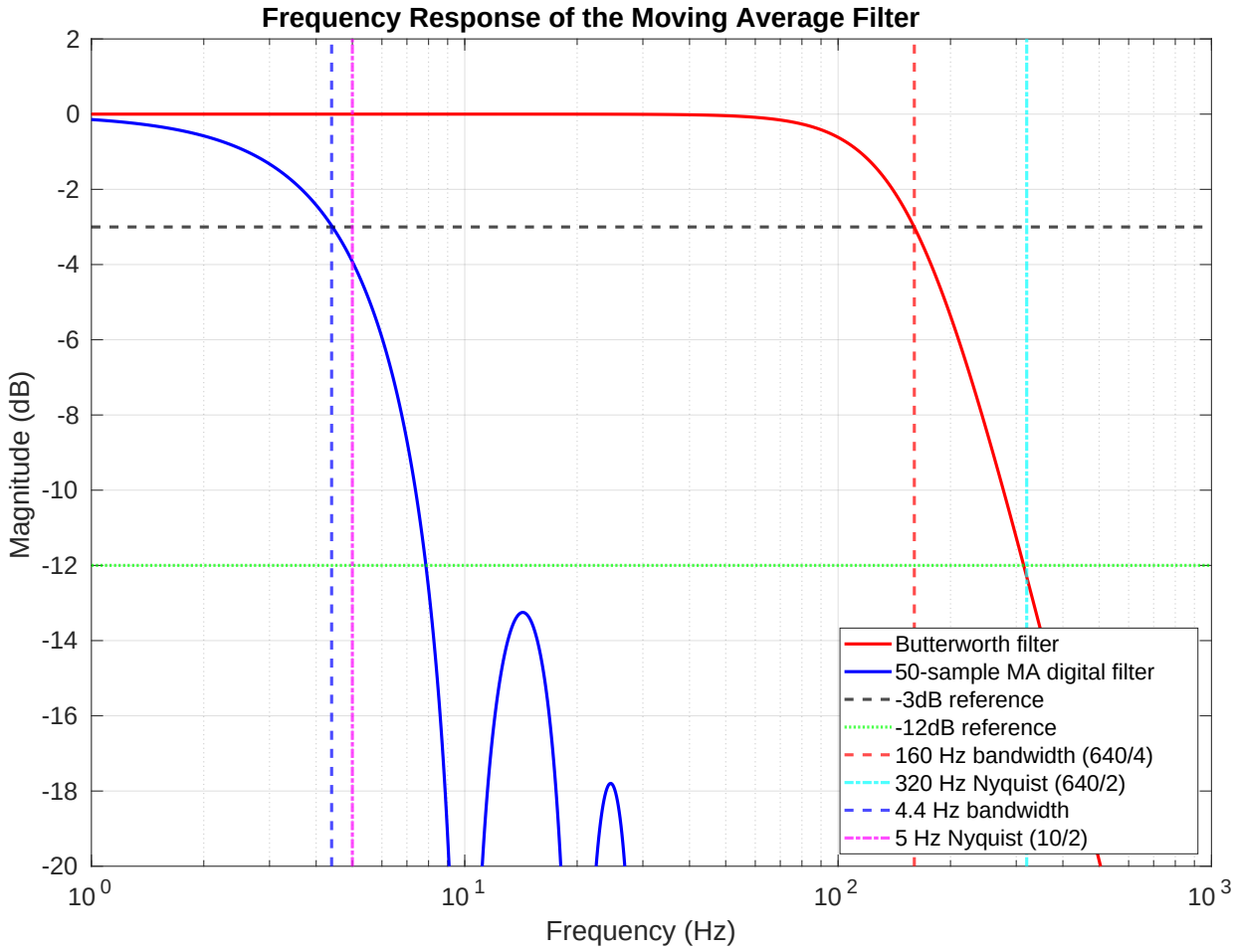


Figure A.1: Frequency response of the 50 point moving average filter.

A.3 Frequency Response of the Moving Average Filter

Figure A.1 illustrates the frequency response of the 50 point moving average filter that was used to decimate the 500 Hz acceleration signal. As the figure shows, the filter has an effective cut-off frequency of 4.4 Hz. The rejection at the Nyquist frequency of 5 Hz is ∼-4 dB. This is not sufficient rejection as ideally, a low-pass filter should provide between 10 to 20 dB rejection at the Nyquist frequency.

According to the specification data sheet of the IMU, the filter bandwidth is equal to $\frac{ODR}{4}$ which corresponds to 160 Hz for the selected ODR of 640 Hz [121]. Since no information

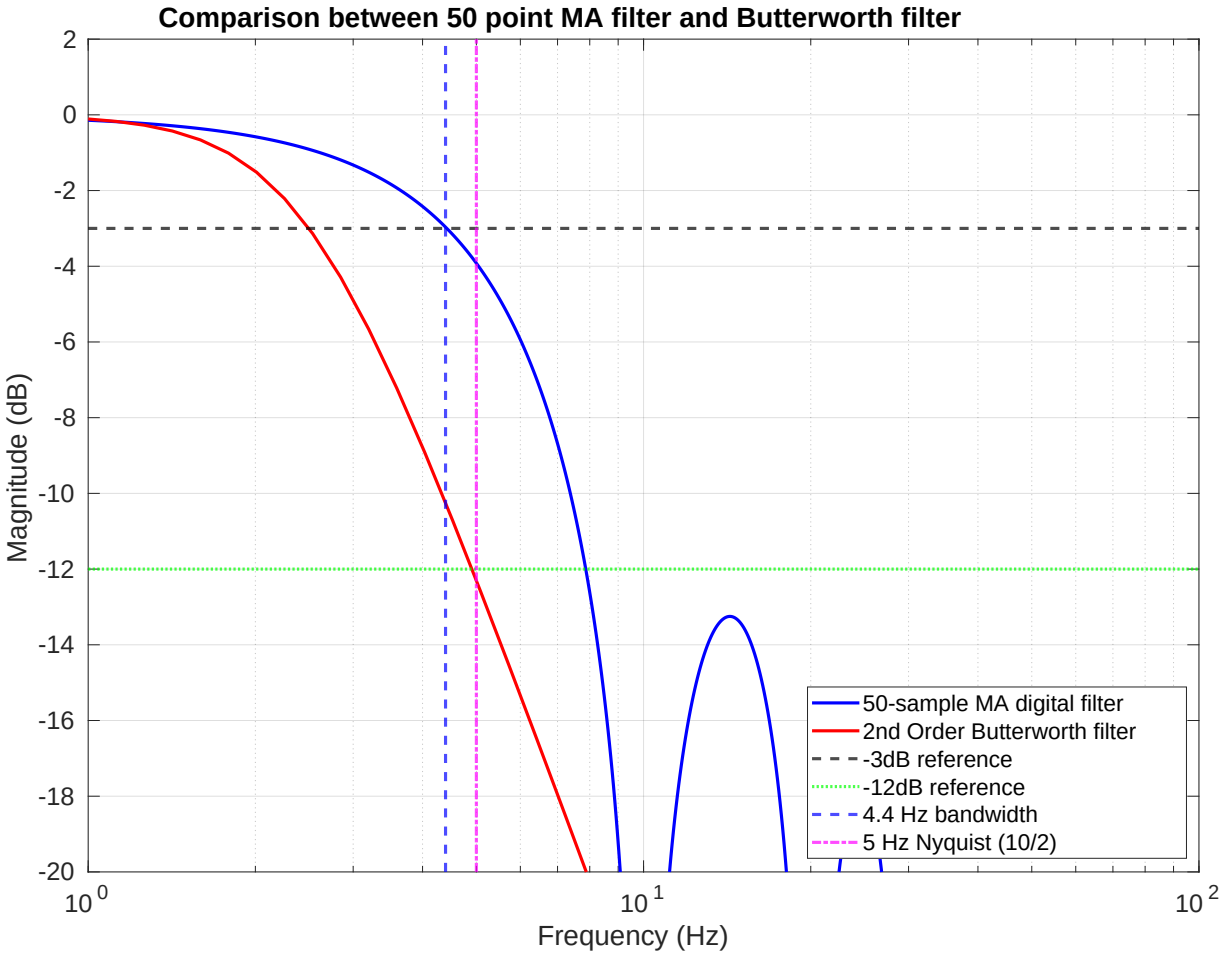


Figure A.2: Comparison of the moving average filter with a 2nd order Butterworth filter.

is provided about the exact nature of the filtering methods used in the accelerometer, a hypothetical second order Butterworth filter with a cut-off frequency of 160 Hz has been plotted to show that a -12 dB rejection occurs at the Nyquist frequency of 320 Hz which could be considered adequate. Even though this filter is possible in such accelerometers, the actual filtering mechanism used could be different and therefore this plot is not necessarily representative of the system. This is another shortcoming of the signal processing design that such details were not available when picking the accelerometer.

Figure A.2 compares the 50 point moving average filter with a possible alternative that

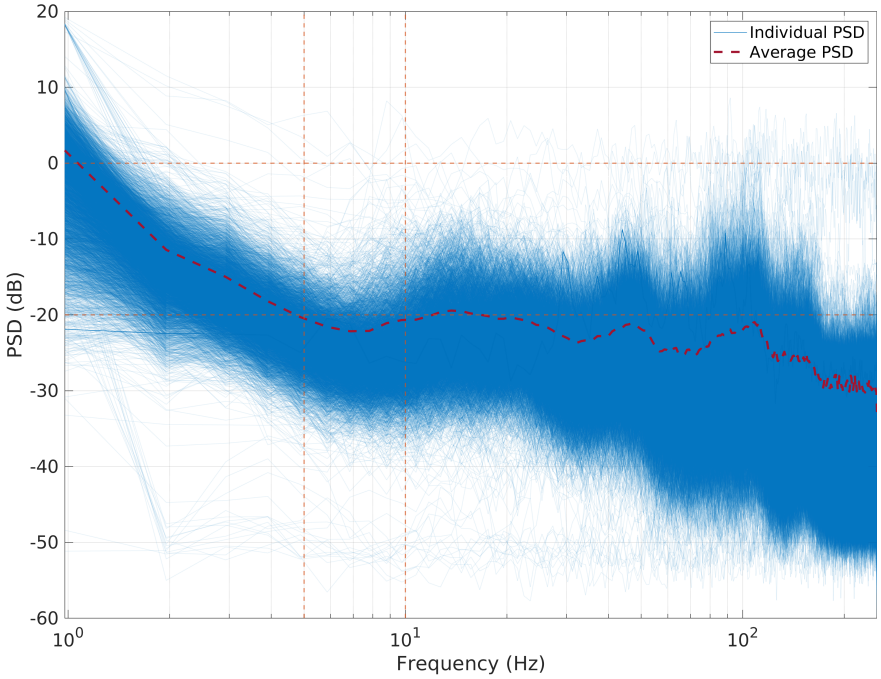
would have adequate rejection for the Nyquist frequency. The frequency response of the 50 sample moving average filter is compared to a 2^{nd} order Butterworth filter with a cut-off frequency of 2.5 Hz. Assuming that the accelerometer signal had to be saved at 10 Hz and adequate processing power was available, the 2^{nd} order Butterworth filter would allow for 12 dB rejection at Nyquist which can be considered acceptable. If the same rejection is needed at Nyquist frequency with a higher bandwidth, a higher order filter would be needed.

A.4 Power Spectral Density of the 500 Hz Signals

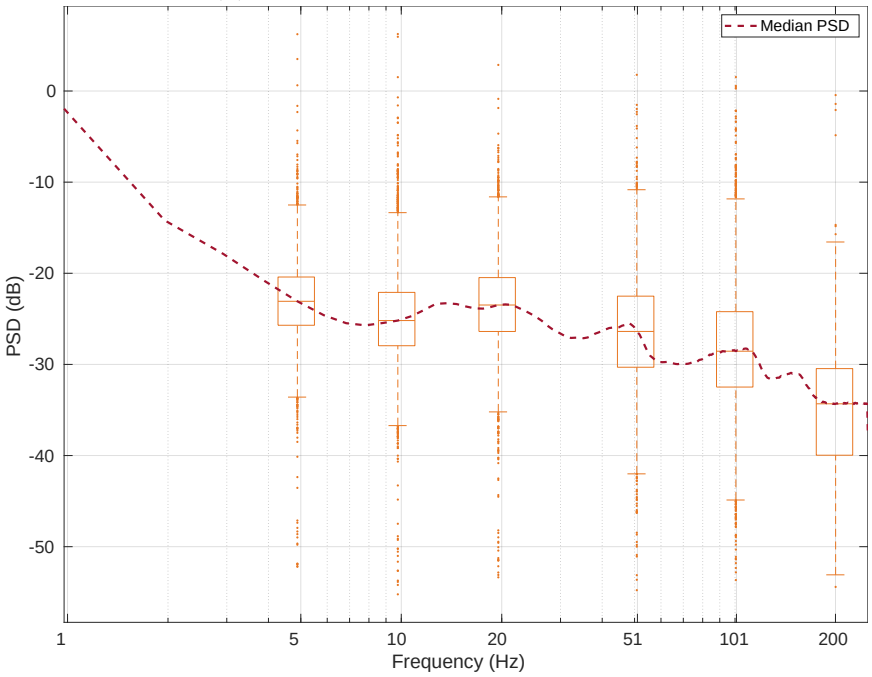
Every time the magnitude of the moving average 10 Hz signal crosses $\pm$ 0.4 g, a six second snippet of the 500 Hz signal is also saved on the hard drive in addition to the 10 Hz signal. Each snippet had about 3 – 4 seconds of pre-trigger data and 2 seconds of post-trigger data. Millions of such instances are available from SHRP 2 NDS. Since the signal decimation process includes non-ideal steps, these raw 500 Hz data snippets are helpful in understanding the power spectral density (PSD) of the raw versus filtered signal. To perform this analysis, about 6,650 snippets were randomly chosen from the various trips in SHRP 2 NDS where both the 500 Hz and decimated 10 Hz signals were available.

Based on Figures A.1 and A.2, the most critical frequency range is 5 Hz to 8 Hz. 5 Hz is the Nyquist frequency and anything beyond 8 Hz will be rejected by atleast 12 dB through the 50 point moving average filter. Therefore, we need to determine how much signal content might be expected in this frequency range that is not adequately filtered by the anti-alias filter.

Figure A.3a shows the power spectral density of each of the 6,650 longitudinal acceleration snippets with a transparency/alpha value of 0.1. This illustrates the spread of the PSD at the various frequencies. Each of the PSD curves is calculated using the pwelch function in

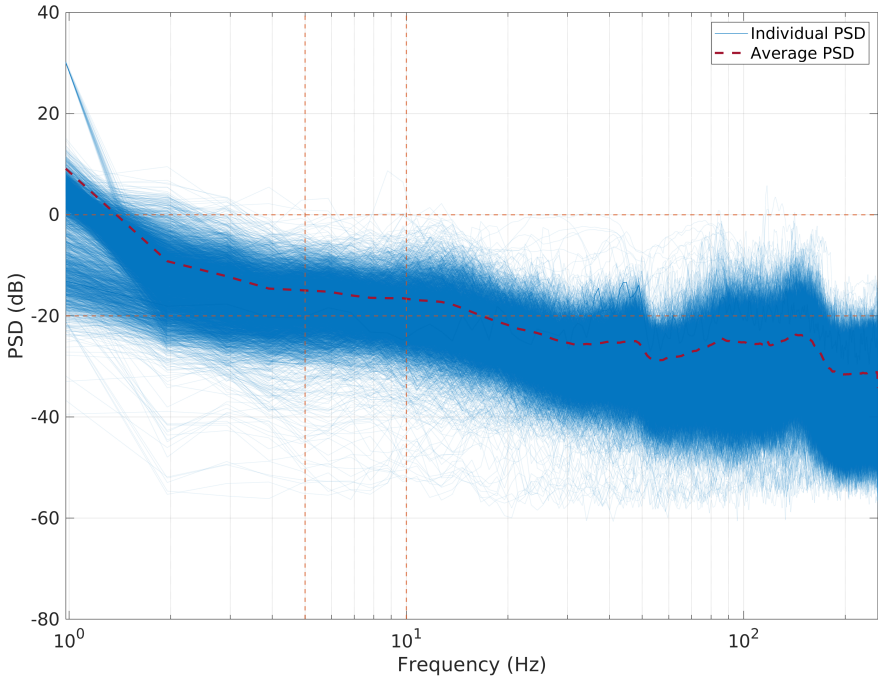


(a) Each PSD overlaid with mean.

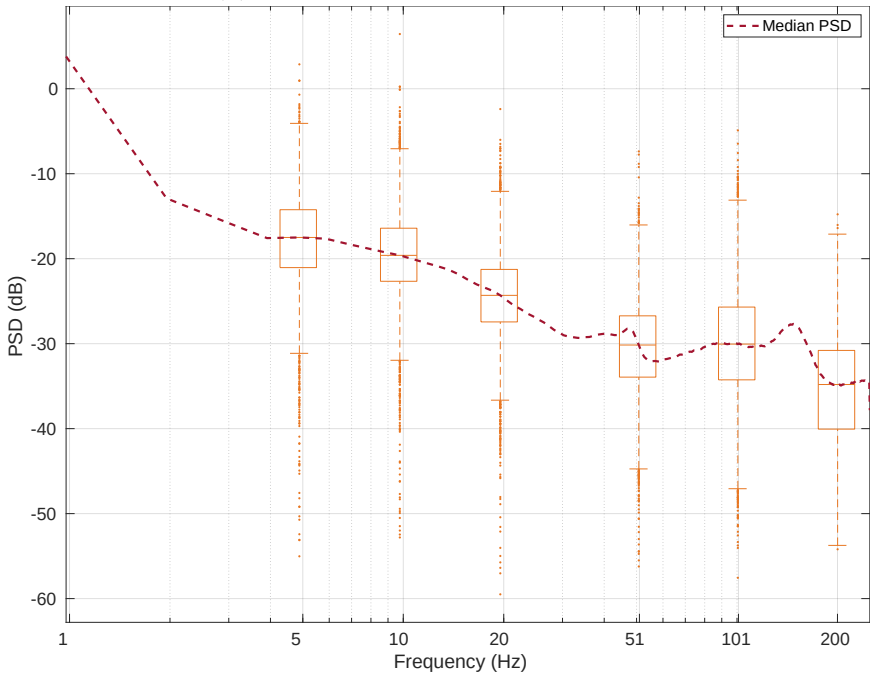


(b) Box plots showing distributions.

Figure A.3: The power spectral density for randomly sampled six second snippets of longitudinal acceleration data at 500 Hz.

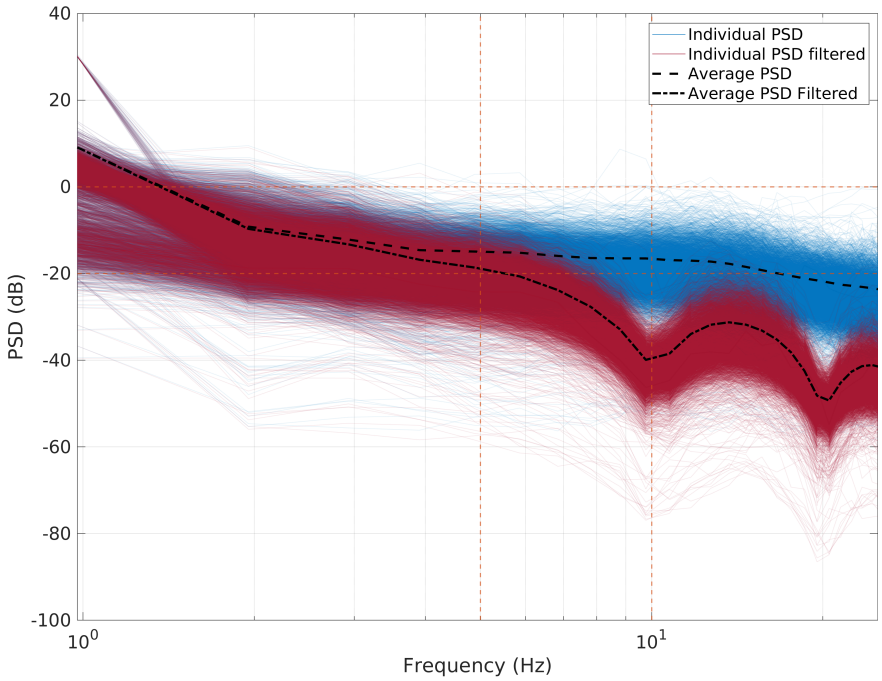


(a) Each PSD overlaid with mean.

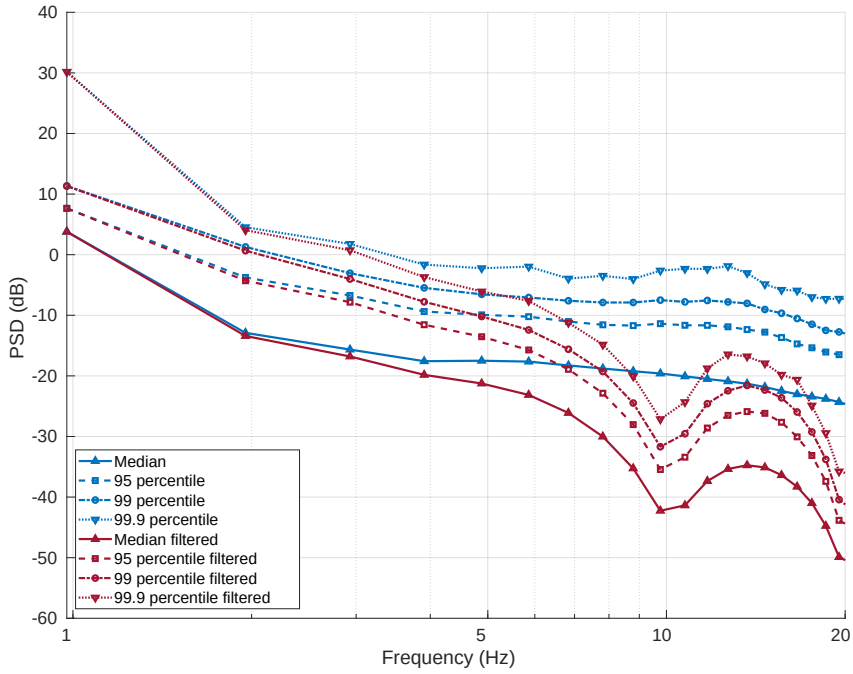


(b) Box plots showing distributions.

Figure A.4: The power spectral density for randomly sampled six second snippets of lateral acceleration data at 500 Hz.



(a) Comparing PSD of unfiltered (blue) versus filtered (red) signals.



(b) Comparing PSD of unfiltered (blue) versus filtered (red) signals for key percentiles.

Figure A.5: The power spectral density comparison for unfiltered versus 50 sample moving average filtered lateral acceleration data.

MATLAB using the same frequency vector. This creates a matrix which can be indexed to calculate aggregate PSD values for each frequency value in the frequency vector. The overlaid average PSD curve is calculated by averaging the PSD values of all 6,650 data snippets at each frequency.

Figure A.3b illustrates the same data as Figure A.3a but instead of plotting all the individual PSD curves, box plots are created at key frequencies to show the distribution of the PSD at these values. The overlaid maroon curve represents the median calculated for each value of the frequency vector used in the pwelch function. For example at 5 Hz, 25^{th} to 75^{th} percentile PSD values lie between -20 to -25 dB. Figure A.4 shows the same data for lateral acceleration snippets.

Figure A.5a compares the PSD plots of the raw 500 Hz lateral acceleration signal versus the PSD plots of the 50 sample moving average lateral acceleration signal (also at 500 Hz). Figure A.5b compares the same data but only plots key percentiles values of median (50^{th}), 95^{th}, 99^{th}, and 99.9^{th} PSD magnitude at all frequencies between 1 and 20 Hz. These key percentiles are chosen to represent average as well as outlier cases. Since Figure A.5a only shows frequencies between 1 to 20 Hz, the data may look different than Figure A.4 which plots frequencies between 1 to 200 Hz for the same data. Some key inferences that can be drawn from the plots are:

1. Figures A.3 to A.5 show that almost all signals have a DC or very low frequency component that is about 20 dB higher than it is at 5 Hz. This is a positive outcome as it means that the inadequate rejection of the moving average filter at the Nyquist frequency will not create a significant issue.

2. Figures A.3 to A.5 also show that there are outliers near the 5 Hz frequency that seem to have a much higher PSD value than the median. It is important to determine

whether these outliers show enough reduction between 1 Hz and 5 Hz to be deemed acceptable for use.

3. Figure A.5 compares the key percentiles values of PSD for raw versus filtered data. We can see that PSD values of filtered signals reduce between 1 and 5 Hz by:
    - $\sim$ 25 dB for the 50^{th} percentile (one in two samples)
    - $\sim$ 22 dB for the 95^{th} percentile (one in twenty samples)
    - $\sim$ 22 dB for the 99^{th} percentile (one in hundred samples)
    - $\sim$ 35 dB for the 99.9^{th} percentile (one in thousand samples)

Future users of SHRP 2 NDS data should perform their own analysis of the 500 Hz data using this technique to determine whether the specific signal has any significant spectral energy between 5 and 10 Hz.

A.5 Time Domain Peak Distortion

The timeseries output from the accelerometer at 640 Hz is subsampled to 500 Hz, then decimated using a 50 point moving average filter, and finally sub-sampled to 10 Hz for storage. During this transformation, it is important to understand the peak distortion of the time domain data. Figures A.6 and A.7 show two examples of time series comparison of the raw 500 Hz signal with the decimated 10 Hz high signal. The bottom portion of both plots show the PSD for the 500 Hz unfiltered signal.

The following conclusions can be drawn from these two plots:

1. Figure A.6 shows that the 500 Hz signal has a maximum value of 0.6 g and the 10 Hz signal has a maximum value of around 0.4 g experienced between 426 and 427 seconds.

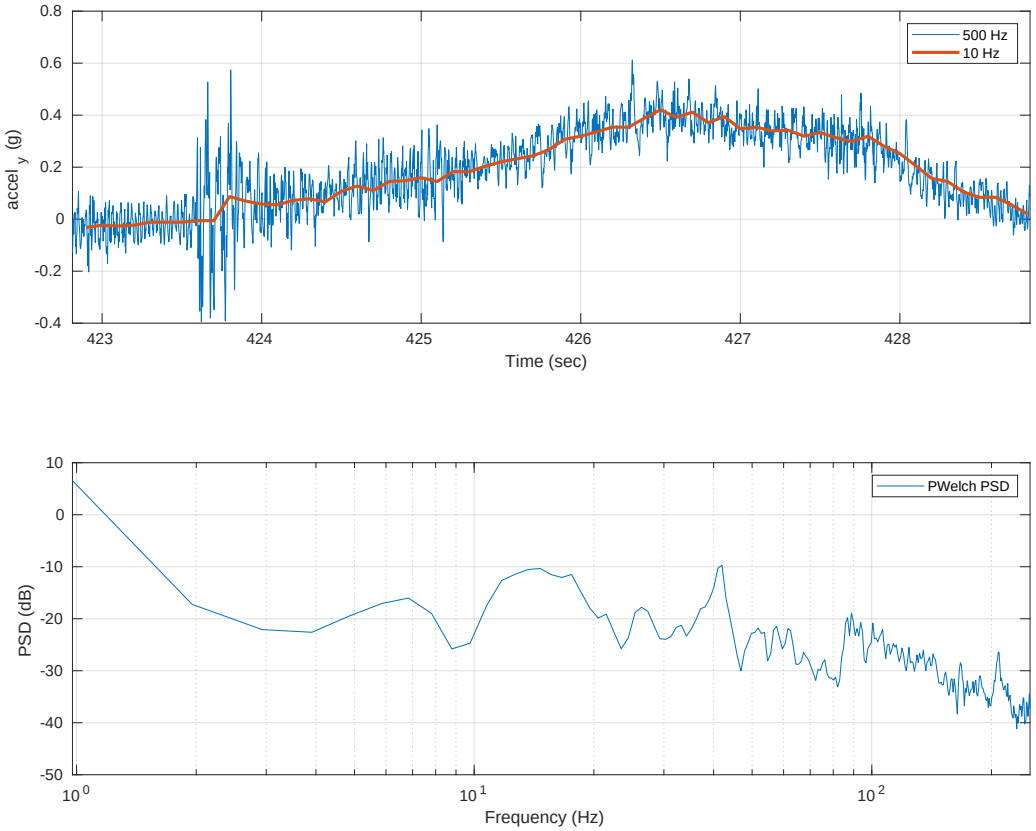


Figure A.6: First example comparing the time series 500 Hz lateral acceleration signal with the 10 Hz signal.

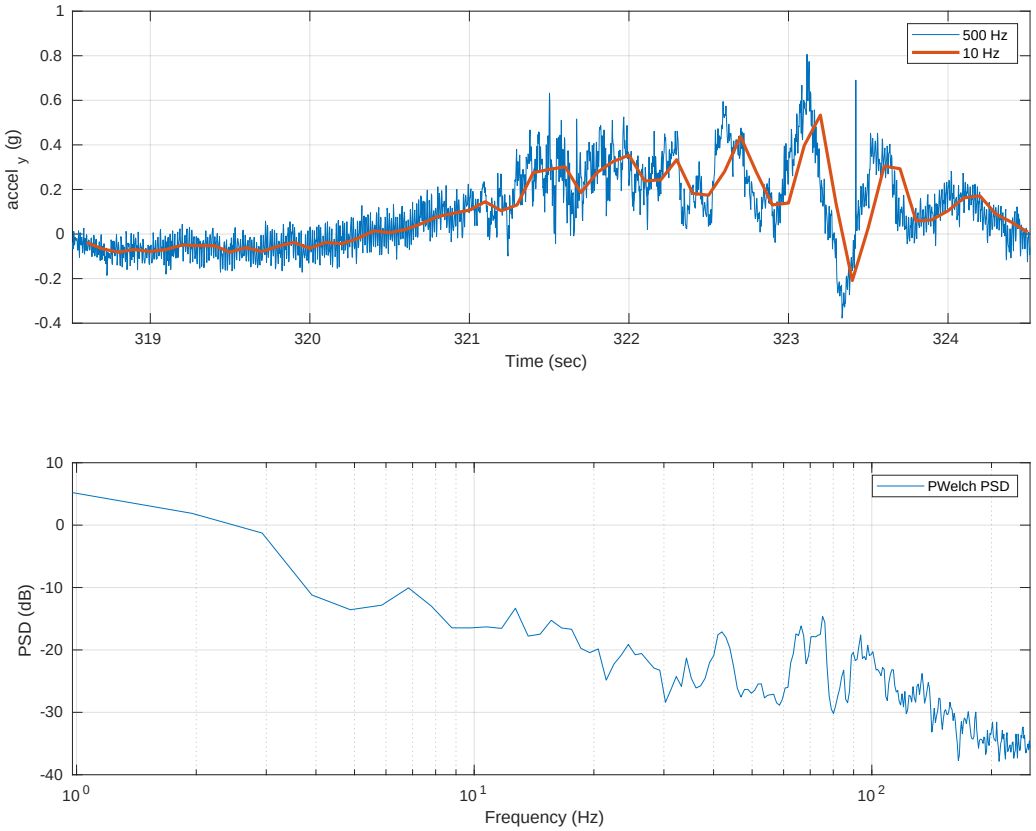


Figure A.7: Second example comparing the time series 500 Hz lateral acceleration signal with the 10 Hz signal.

There is also a high amplitude and high frequency fluctuation between 423.5 to 424 seconds noticeable in the 500 Hz signal that is not seen in the 10 Hz signal. Most of the analyses in this dissertation deal with human behavior which are appropriately captured in the 10 Hz signal. The fluctuations captured in the 500 Hz signal that are not captured in the 10 Hz signal do not represent human preferences. Therefore, in this particular case, even though there is considerable peak distortion, the 10 Hz signal is capturing the intended human behavior.

2. Figure A.7 shows that the 500 Hz signal has a maximum value of around 0.8 g whereas the 10 Hz signal has a maximum value of around 0.5 g. There is also a $\sim$ 2 Hz component prominent between 321 and 324 in both the 500 Hz and 10 Hz signals. However, in this case, further examination is required to determine whether the fluctuation is based on human behavior or other factors such as vehicle dynamics. Therefore, a case can be made that the 10 Hz signal is not adequately representing the true peak experienced by the human driver.

3. Figure A.7 example shows a higher PSD value between 1 to 3 Hz when compared with Figure A.6 example. This can be explained by the fluctuation of the lateral acceleration seen in time series plot of Figure A.7 between 321.5 seconds to 324 seconds. The peak distortion is higher in example 1 as compared to example 2.

4. The peak distortion is higher for signals with a high crest factor (i.e. the ratio of the peak to the rms value).

The primary purpose of the acceleration signals in SHRP 2 NDS was to detect periods of harsh longitudinal and lateral accelerations. These periods were then reviewed to identify crashes and near crashes. The thresholds for reviewing video cases were decided after analyzing the success rates at various different values of longitudinal and lateral acceleration

[104]. Therefore, the peak distortion of the timeseries signal will not affect the performance of the crash detection algorithms as they have been tuned for SHRP 2 NDS data specifically. Similarly, comparisons made between participants of the SHRP 2 NDS should also not be affected by peak distortion as the distortions will be common to all participants. However, when it comes to comparing maximum or minimum values across different studies, it is important to compare the filtering process used in each study before making comparisons.