

**Modeling drivers' naturalistic driving behavior on rural two-lane curves**

by

**Bo Wang**

A dissertation submitted to the graduate faculty  
in partial fulfillment of the requirements for the degree of  
**DOCTOR OF PHILOSOPHY**

Major: Civil Engineering (Transportation Engineering)

Program of Study Committee:  
Shauna Hallmark, Co-Major Professor  
Jing Dong, Co-Major Professor  
Omar Smadi  
Peter Savolainen  
Yehua Li  
Dianne Cook

Iowa State University

Ames, Iowa

2015

Copyright © Bo Wang, 2015. All rights reserved.

## **DEDICATION**

To my wife and my parents who taught me the meaning of love.

格物致知，知行合一。

Truth can only be discovered by the unity of observing, thinking and practicing.

## TABLE OF CONTENTS

DEDICATION .....	ii
LIST OF FIGURES .....	vii
LIST OF TABLES .....	xi
NOMENCLATURE .....	xiii
ACKNOWLEDGMENTS .....	xv
ABSTRACT .....	xvi
CHAPTER 1.INTRODUCTION.....	1
1.1 The Scope of Traffic Safety Problem .....	2
1.2 Traditional Safety Analysis and Its Drawbacks.....	3
1.2.1 Police-reported crash data analysis .....	4
1.2.2 Driving simulator study .....	5
1.3 The Concept of Naturalistic Driving Study and Its Opportunities .....	6
1.4 Problem Statement .....	9
1.5 Research Questions.....	11
1.5.1 Research question 1 .....	12
1.5.2 Research question 2 .....	13
1.5.3 Research question 3 .....	14
1.6 Dissertation Structure.....	15
CHAPTER 2.LITERATURE REVIEW .....	16
2.1 Curve Perception.....	16
2.2 Vehicle Speed on Curves .....	19
2.3 Vehicle Lateral Position on Curves .....	25
2.4 Summary of Major Findings.....	28
CHAPTER 3.INTRODUCTION TO THE STRATEGIC HIGHWAY RESEARCH PROGRAM 2 NATURALISTIC DRIVING STUDY .....	31
3.1 Review of Previous Naturalistic Driving Studies .....	31
3.1.1 100-Car naturalistic driving study, U.S. ....	31
3.1.2 Naturalistic teenage driving study, U.S. ....	33
3.1.3 UDRIVE, Europe.....	35

3.1.4 Shanghai naturalistic driving study, China .....	36
3.1.5 Canada naturalistic driving study, Canada.....	36
3.2 Introduction to the SHRP2 Naturalistic Driving Study .....	37
3.2.1 Study background .....	37
3.2.2 Data acquisition system .....	39
3.2.3 Roadway information database .....	42
3.2.4 Driver data .....	44
3.2.5 Vehicle data .....	45
3.2.6 Time series DAS data .....	46
3.2.7 Crashes and near-crashes data .....	48
3.2.8 Institutional review board and other issues.....	49
3.2.9 The challenges of analyzing the SHRP2 NDS data .....	49
 CHAPTER 4.CRASHES AND NEAR-CRASHES ANALYSIS ON RURAL TWO- LANE CURVES .....	 52
4.1 Introduction.....	52
4.2 Data Collection .....	54
4.2.1 Definition of crash, near-crash, and baseline events.....	55
4.2.2 Query of curve-related crashes .....	56
4.2.3 Variables from SHRP2 NDS event table .....	57
4.2.4 Variables from forward video .....	59
4.2.5 Variables from Google Earth .....	60
4.3 Initial Screen of All crashes on All Curves .....	62
4.4 Description of Roadway Departure Crashes on Rural Two-Lane Curves .....	66
4.4.1 Speeding.....	66
4.4.2 Secondary tasks.....	67
4.4.3 Adverse surface conditions .....	68
4.5 Logistic Regression Analysis of Roadway Departure Crashes on Rural Two- Lane Curves.....	 69
4.5.1 Background .....	70
4.5.2 Data description .....	71
4.5.3 Logistic regression model .....	72
4.5.4 Model results.....	73
4.6 Discussion .....	76
4.7 Summary .....	79
 CHAPTER 5.MULTIVARAITE ANALYSIS OF DRIVER BEHAVIOR ON RURAL TWO-LANE CURVES.....	 80
5.1 Introduction.....	80
5.1.1 Vehicle speed .....	80
5.1.2 Lateral positions.....	81
5.1.3 Lateral acceleration .....	82
5.1.4 Curve-related crash analysis .....	83
5.1.5 Research objectives.....	85

5.2 Data Description .....	85
5.2.1 Data collection .....	86
5.2.2 Data sources .....	87
5.2.3 Batch process of time series data .....	92
5.2.4 Data quality assurance .....	93
5.3 Vehicle Lateral Acceleration Analysis .....	95
5.3.1 Lateral acceleration vs. curve radius .....	95
5.3.2 Lateral acceleration vs. speed .....	96
5.3.3 Drivers' comfortable lateral acceleration on curves .....	97
5.4 Vehicle Speed Analysis .....	100
5.4.1 Drivers' curve negotiation behavior on 55 MPH roadways .....	101
5.4.2 Drivers' curve negotiation behavior on 45 MPH roadways .....	103
5.4.3 Speed prediction model .....	105
5.5 Discussion .....	110
5.6 Summary .....	112
 CHAPTER 6. FUNCTIONAL DATA ANALYSIS OF TIME SERIES DATA ON RURAL TWO-LANE CURVES .....	 114
6.1 Introduction .....	114
6.1.1 Challenges of analyzing time series data in the SHRP2 NDS .....	114
6.1.2 Introduction to functional data analysis .....	115
6.1.3 Example of functional data analysis .....	117
6.2 Data Description .....	120
6.3 Methodology .....	124
6.3.1 Convert the discrete time series data to functional data .....	124
6.3.2 Calculate the mean, standard deviation, and the derivatives for a group of functional data .....	127
6.3.3 Functional principal component analysis .....	130
6.4 Analysis of Time Series Data using Functional Data Analysis .....	132
6.4.1 Plot of raw speeds .....	132
6.4.2 Plot of fitted functional data, mean speed, and the confidence interval ..	134
6.4.3 Examine the deceleration profile on the curves .....	137
6.4.4 Vehicle deceleration profile before the curve point of curvature .....	139
6.4.5 Phase-plan plot of vehicle dynamics .....	140
6.4.6 Functional principal component analysis of vehicle speed .....	142
6.5 Discussion .....	145
6.6 Conclusion .....	147
 CHAPTER 7. CONCLUSION .....	 149
7.1 Summary of Major Findings .....	149
7.1.1 Crashes and near-crashes analysis on rural two-lane curves .....	150
7.1.2 Multivariate analysis of driver behavior on rural two-lane curves .....	151
7.1.3 Functional data analysis of time series data on rural two-lane curves ....	153

7.2 Implications for Future Research.....	155
7.2.1 Implication for big data research .....	157
7.2.2 Implication for connected vehicle research .....	157
7.2.3 Implication for automated vehicle research.....	158
 BIBLIOGRAPHY.....	 159
 APPENDIX A. BACKGROUND OF THE SHRP2 NDS.....	 170
 APPENDIX B. SAMPLE R CODE.....	 175
 APPENDIX C. FUNCTIONAL DATA ANALYSIS MODEL OUTPUTS .....	 176

## LIST OF FIGURES

Figure 1.1 Fatalities and fatality rate per 100 million VMT from 1990 to 2012 .....	3
Figure 1.2 Typical variables collected in police-reported crash data .....	4
Figure 1.3 National advanced driving simulator at University of Iowa .....	6
Figure 1.4 Four camera views in SHRP2 NDS.....	7
Figure 1.5 Video camera locations on the horizontal curves.....	10
Figure 2.1 Gaze strategies on curves .....	17
Figure 2.2 Horizontal curves with vertical alignment .....	18
Figure 2.3 The distribution of wheel position on the study curves.....	25
Figure 2.4 Sketches of six types of trajectories .....	26
Figure 2.5 Six patterns of vehicle track behavior on curves simulated in Matlab.....	27
Figure 3.1 Data acquisition system in the 100-Car study .....	32
Figure 3.2 Extreme G-force events per 100 miles for teen drivers, the parents, and the teen driver with adult passengers.....	34
Figure 3.3 Data acquisition system in the UDRIVE project.....	35
Figure 3.4 Shanghai naturalistic driving study .....	36
Figure 3.5 Head unit, main unit, and front radar used in Canada naturalistic driving Study .....	37
Figure 3.6 SHRP2 NDS data collection sites and the number of participants.....	39
Figure 3.7 Data acquisition system components used in the SHRP2 NDS .....	40
Figure 3.8 SHRP2 data acquisition system installation schematic .....	40
Figure 3.9 Four camera views from the SHRP2 data acquisition system: forward view (upper left), in-cabin driver face view (upper right), instrument panel view	

(bottom left), rear view (bottom right). (The driver in the picture is not experimental participant and just for demonstration purpose only) .....	42
Figure 3.10 A site map shows the RID database in the SHRP2 NDS .....	43
Figure 3.11 Drivers demographics by age group and gender .....	44
Figure 3.12 Vehicle types in the SHRP2 NDS .....	46
Figure 3.13 Crash and near crashes by incident types .....	48
Figure 4.1 Venn diagram of crash causes by percentage .....	53
Figure 4.2 Filter criteria for curve-related crashes and near-crashes in the SHRP2 NDS .....	56
Figure 4.3 Curve radius measurement from chord length and offset distance .....	61
Figure 4.4 Crashes and near-crashes by severities .....	63
Figure 4.5 Distribution of curve-related crashes and near-crashes by incident types .....	64
Figure 4.6 Distribution of curve-related crashes and near-crashes by junction types .....	64
Figure 4.7 Subset the SHRP2 curve-related events .....	65
Figure 4.8 Distribution of driver behaviors for roadway departure events on rural two-lane curves .....	67
Figure 4.9 Distribution of secondary tasks for roadway departure events on rural two-lane roadways .....	68
Figure 4.10 Distribution of surface conditions for roadway departure crashes on rural two-lane roadways .....	69
Figure 5.1 Influence of tangent speeds on curves .....	81
Figure 5.2 Sketches of six types of trajectories .....	82
Figure 5.3 The relationship between lateral acceleration and speed .....	83
Figure 5.4 Example of segment, buffer, trace, and observations in ArcGIS .....	87
Figure 5.5 Integration of different data sources .....	92



Figure 5.6 Plot of lateral acceleration vs. curve radius .....	96
Figure 5.7 Plot of lateral acceleration vs. speeds.....	97
Figure 5.8 Cumulative distribution function of lateral acceleration by curve radius .....	98
Figure 5.9 Plot of 85 <sup>th</sup> , 50 <sup>th</sup> , and 15 <sup>th</sup> percentile of lateral acceleration by curve radius .....	99
Figure 5.10 Data quality assurance of vehicle speed data .....	100
Figure 5.11 Plot of mean speeds by curve radius on 55 MPH roadways .....	102
Figure 5.12 Boxplot of mean speeds by advisory speed limits on 55 MPH roadways .....	102
Figure 5.13 Plot of speeds by curve radius on 45 MPH roadways .....	104
Figure 5.14 Boxplot of mean speeds by advisory speed limits on 45 MPH roadways .....	104
Figure 5.15 Comparison of observed mean speeds to the predicted mean speeds on rural two-lane curves .....	109
Figure 5.16 Plot of random effects for the drivers.....	110
Figure 6.1 Example of converting discrete speeds profile to functional data on curve NY67a.....	117
Figure 6.2 Plot of 146 speed profiles on the same curve .....	118
Figure 6.3 Plot of vehicle speed (top panel) and the first derivative of the vehicle speed (bottom panel).....	119
Figure 6.4 The criteria used to select sample curves .....	121
Figure 6.5 Google street view of the four example curves .....	123
Figure 6.6 The 15 spline basis functions defined over the interval (-500, 500) by 11 interior knots. The polynomial segments has order four polynomials. The polynomial values and its derivatives were required to be smoothly connected at the interior knots. ....	126
Figure 6.7 Contour plot of correlation function across locations for vehicle speeds .....	129
Figure 6.8 Plot of raw speeds on the curves .....	133

Figure 6.9 Plot of mean speeds and 95% confidence intervals on the curves .....	135
Figure 6.10 Plot of vehicle acceleration on the curves .....	138
Figure 6.11 Plot of deceleration profile before the curve PC .....	139
Figure 6.12 Phase plane plot for the 95 driving traces on curve 1.....	140
Figure 6.13 Phase plane plot for the four sample curves.....	141
Figure 6.14 The first principal component explained 75.8% of the variability .....	143
Figure 6.15 The second principal component explained 10.3 % of the variability .....	143
Figure 6.16 The third principal component explained 8.6 % of the variability.....	144
Figure A.1 SHRP2 participant versus U.S. driving population percentages by age group ..	174
Figure C.1 Scree plot for choosing the optimal number of basis functions.....	176
Figure C.2 List of basis functions on the four example curves .....	177
Figure C.3 Scree plot for choosing the optimal turning parameters .....	178
Figure C.4 Plot of discrete raw time series speed data and best fitted functional data on curve 1.....	179
Figure C.5 Plot of discrete raw time series speed data and best fitted functional data on curve 2.....	180
Figure C.6 Plot of discrete raw time series speed data and best fitted functional data on curve 3.....	181
Figure C.7 Plot of discrete raw time series speed data and best fitted functional data on curve 4.....	182
Figure C.8 Scree plot for choosing optimal number of principal components. The first three principal components were chosen for analysis.....	183

## LIST OF TABLES

Table 2.1 Review of speed prediction models .....	24
Table 3.1 List of collected variables in the roadway information database (RID) .....	43
Table 3.2 List of driver surveys collected in the SHRP2 NDS .....	45
Table 3.3 Description of time series data collected from SHRP2 NDS .....	47
Table 4.1 List of variables from the event detail table .....	58
Table 4.2 Description of the variables from forward view video .....	60
Table 4.3 A list of variables included in logistic regression model .....	71
Table 4.4 Logistic regression model results .....	73
Table 4.5 Classification table for the logistic regression model .....	74
Table 4.6 The odds ratio estimate with 95% confidence interval .....	76
Table 5.1 Time series driving data .....	89
Table 5.2 Curve geometry data .....	90
Table 5.3 Driver demographics and vehicle types .....	91
Table 5.4 Summary of data quality for the rural two-lane curve dataset .....	94
Table 5.5 Fixed effects of the speed prediction model on rural two-lane curves with 45 MPH upstream speed limit .....	107
Table 5.6 Random effect estimate from the linear mixed effect model .....	108
Table 6.1 List of curve characteristics for the sample curves .....	122
Table 6.2 Summary statistics of the mean speed profile .....	136
Table A.1 Recruitment summary by method, age group, and site .....	170
Table A.2 Assessment questionnaires administered .....	171
Table A.3 Cognitive assessments .....	172
Table A.4 SHRP2 NDS vehicle classes .....	173

Table A.5 Quality assessment of select vehicle metrics .....	173
--	-----

**NOMENCLATURE**

AADT	Annual Average Daily Traffic
AIC	Akaike Information Criterion
ASSHTO	American Association of State Highway and Transportation officials
BIC	Bayesian Information Criteria
CAN	Controller Area Network
CDC	Center for Disease Control
CDF	Cumulative Distribution Function
CTRE	Center for Transportation Research and Education
DAS	Data Acquisition System
DOT	Department of Transportation
FARS	Fatality Analysis Reporting System
FDA	Functional Data Analysis
FHWA	Federal Highway Administration
FPCA	Functional Principal Component Analysis
GES	General Estimates System
GIS	Geographic Information System
GPS	Geographic Positioning System
IRB	Institutional Review Board

MAP-21	Moving Ahead for Progress in the 21 <sup>st</sup> Century
MPH	Mile Per Hour
MUTCD	Manual on Uniform Traffic Control Devices
NDS	Naturalistic Driving Study
NHTSA	National Highway Traffic Safety Administration
NTDS	Naturalistic Teenage Driving Study
PC	Point of Curvature
PT	Point of Tangent
RID	Roadway Information Database
SAFETEA-LU	Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users
SHRP2	Strategic Highway Research Program 2
SUV	Sport Utility Vehicle
TEA	Transportation Equity Act
UDRIVE	eUropean naturalistic Driving and Riding for Infrastructure & Vehicle safety and Environment
VMT	Vehicle Miles Traveled
VTTI	Virginia Tech Transportation Institute

## ACKNOWLEDGMENTS

I would like to express the deepest gratitude to my mentor and my major professor, Dr. Shauna Hallmark for her cheerful encouragement, tremendous patience, and generous support. She is such a great role model who demonstrates strong work ethics and commits to excellence in her everyday work. She always embraces new ideas and encourages me to explore new directions. Without her supervision and guidance, this dissertation would not have been accomplished. I would like to thank Dr. Jing Dong, Dr. Omar Smadi and Dr. Peter Savolainen for their supports and helps throughout the course of this research. Their insightful comments and constructive suggestions helped me understand my research from many different perspectives. Special thanks are also due to Dr. Yehua Li and Dr. Dianne Cook's support and assistance for statistical analysis.

In addition, I would also like to thank my friends, InTrans colleagues, the CCEE department faculty and staff for making my time at Iowa State University such an extraordinary experience. They provided a wonderful and productive working environments to let me complete my dissertation in a smooth and timely manner. Many thanks are given to Dr. Nadia Gkritza, Dr. Michael Pawlovich, Dr. Abhisek Mudgal, Dr. Nicole Oneyear, Zach Hans, Skylar Knickerbocker for providing helpful supports along the ways. My research would not have been possible without their helps.

Last but not least, I am greatly indebted to my wife who always stands by me and cheers for me in my ups and downs. I thank my parents for their unconditional love and caring since my day one. I would also like to thank my sister for taking care of our parents while I was studying abroad.

## **ABSTRACT**

This dissertation examined drivers' naturalistic driving behavior on rural two-lane curves using the Strategic Highway Research Program 2 Naturalistic Driving Study data. It is a state-of-the-art naturalistic driving study that collected more than 3,000 drivers' daily driving behavior over two years in the U.S. The major data sources included vehicle network, lane tracking system, front and rear radar, driver demographics, driver surveys, vehicle characteristics, and video cameras. This dissertation has three objectives: 1) examine the contributing factors to crashes and near-crashes on rural two-lane curves; 2) understand drivers' normal driving behavior on rural two lane curves; 3) evaluate how drivers continuously interact with curve geometries using functional data analysis.

The first study analyzed the crashes and near-crashes on rural two-lane curves using logistic regression model. The model was used to predict the binary event outcomes using a number of explanatory variables, including driver behavior variables, curve characteristics, and traffic environments. The odds ratio of getting involved in safety critical events was calculated for each contributing factor. Furthermore, the second study focused on the analysis of drivers' normal curve negotiation behavior on rural two-lane curves. Many important relationships were found among curve radius, lateral acceleration, and vehicle speeds. A linear mixed model was used to predict mean speeds based on curve geometry and driver factors. The third analysis applied functional data analysis method to analyze the time series speed data on four example curves. Functional data analysis was found to be a useful method to analyze the time series observations and understand driver's behavior from naturalistic driving study.



Overall, this dissertation is one of the first studies to investigate drivers' curve negotiation behavior using naturalistic driving study data, and greatly enhanced our understanding about the role of driver behavior in the curve negotiation process. This dissertation had many important implications for curve geometry design, policy making, and advanced vehicle safety system. This dissertation also discussed the opportunities and challenges of analyzing the Strategic Highway Research Program 2 Naturalistic Driving Study data, and the implications for future research.

## CHAPTER 1. INTRODUCTION

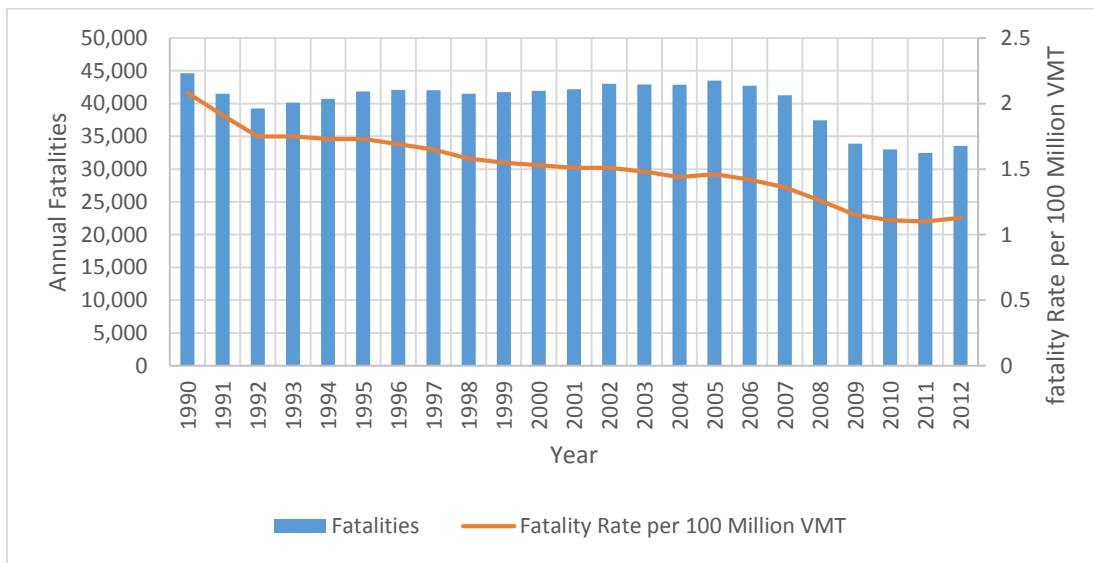
The past studies identified human factors as the major contributing factors to approximately 90% of crashes (Treat et al., 1979), but the traditional crash data analysis does not address human factor very well. Fortunately, the recent development of naturalistic driving study (NDS) allows transportation researchers to observe drivers' daily driving behaviors on roads. Naturalistic driving study provides us an unprecedented opportunity to understand how drivers interact with vehicle, roadway, and traffic environments. It will lead the future of transportation safety research in the next decade (SHRP2, 2015).

Crash statistics have shown drivers had higher probability getting involved in crashes on horizontal curves. According to Federal Highway Administration (FHWA, 2015), the number of fatal crashes on horizontal curves was disproportionally higher than tangent segments. Understanding the drivers' driving behavior on horizontal curves was the focus of transportation safety researchers in the past decades (Shinar et al., 1977; Levison, 1988; Bonneson et al., 2009), but this problem has not been satisfactorily investigated yet. In order to address this knowledge gap, this dissertation studied the drivers' curve negotiation behavior on rural two-lane curves using the state-of-the-art Strategic Highway Research Program 2 Naturalistic Driving Study (SHRP2 NDS) data. The findings from this dissertation will help us understand the causes of curve-related crashes and the drivers' normal driving behavior on curves. The findings provided important insights on better curve geometry design, sign placement, countermeasure selections, and transportation policy making.

## 1.1 The Scope of Traffic Safety Problem

The development of automobile enables people to move freely, but it also threatens people's health and lives on daily basis. Although different definitions had been used in the past, in this dissertation, crash is referred to as a vehicle contacts a moving or fixed object resulting in measurably transferred or dissipated kinetic energy, which may cause injury, death, and property damage (Dingus et al., 2006). According to National Safety Council (2014), the odds of being killed in a car crash in a life time is 1 in 112. The U.S. Center for Disease Control (CDC, 2010) identified car crash as the leading cause of death for people aged between 1 and 34 years old. According to the most recent published NHTSA report *Economic and Societal Impact of Motor Vehicle Crashes*, there were about 33,000 fatalities, 3.9 million injuries, and 24 million damaged vehicles in 2010. The economy cost of these crashes was estimated to be \$277 billion, which was 1.9% of U.S. Gross Domestic Product in 2010 (NHTSA, 2014). Therefore, crash is one of the biggest societal issues that directly affects everyone's quality of life.

Figure 1.1 plots the number of motor vehicle fatalities and fatality rates from 1990 to 2012 in the U.S. The fatality rate per 100 million VMT (shown in red line) dropped from 1990 to 2012. Extensive efforts were conducted by federal, state, local agencies, and Non-Government Organizations to reduce the number of crashes on roads. Transportation safety was listed as the top priority in the past federal transportation bills: TEA-21 (1998-2003), SAFETEA-LU (2005-2009), and MAP-21 (2012-2014). Because of the collaboration efforts from all sectors, transportation safety in the U.S. has improved steadily in the past many years.



**Figure 1.1 Fatalities and fatality rate per 100 million VMT from 1990 to 2012 (FHWA, 2014)**

Although the fatality rate was decreasing in the past twenty years, the cost of crashes was still a big burden to U.S. society, especially for those who lost their loved ones. Hence, there is still a long way to go before achieving zero fatality goal in the U.S. The development of naturalistic driving study opened a new research field to help transportation community understand the causes of crashes and drivers' daily driving behaviors.

## 1.2 Traditional Safety Analysis and Its Drawbacks

A number of methods has been used in the past to study transportation safety, including crash data analysis, driving simulator study, and instrumented cars. The crash data analysis method has been widely used in the past decades to understand the contributing factors to crashes, but they also suffered from several limitations to prevent us from further understanding the role of human factor in crashes. This section reviewed the pros and cons of the existing safety

research methods and why naturalistic driving study provides promising solution to address the traffic safety problems.

### 1.2.1 Police-reported crash data analysis

Police-reported crash data is one of the most important source of information for traffic safety analysis. For example, Fatality Analysis Reporting System (FARS), maintained by National Highway Traffic Safety Administration, is one of the most widely used crash database with focus on fatal crashes. In addition, General Estimates System (GES) is another popular crash data source that randomly sampled all crashes in the U.S. to represent the national crash profile. Furthermore, each state also maintained their own crash data program to identify the safety priority areas in each state. The typical variables collected in crash reports are listed in Figure 1.2. Those variables are commonly categorized into driver, roadway, vehicle, environment, and crash type.

Driver	Roadway	Vehicle	Environment	Crash Type
<ul style="list-style-type: none"> <li>•Person Type</li> <li>•Severity of Injuries</li> <li>•Traffic Law Violation</li> <li>•Alcohol or Drug Impairment</li> </ul>	<ul style="list-style-type: none"> <li>•Latitude and Longitude of Crash Location</li> <li>•Area types</li> </ul>	<ul style="list-style-type: none"> <li>•Vehicle Types</li> <li>•Number of Vehicles Involved</li> <li>•Direction of Travel</li> </ul>	<ul style="list-style-type: none"> <li>•Time of Day, Day of Week</li> <li>•Weather Conditions</li> </ul>	<ul style="list-style-type: none"> <li>•Manner of Collision</li> <li>•Crash Diagram</li> <li>•Narrative Description of Crash</li> </ul>

**Figure 1.2 Typical variables collected in police-reported crash data**

However, crash data analysis has several limitations to prevent us from further understanding the cause of crashes. Those limitations include underreporting issue and the lack

of driver behavior information in crash database (Ye, F. and Lord, D., 2011). Crash dataset did not include all crashes on the public roads. For example, the police officer in Iowa only files a crash report when an accident causes death, personal injury, or total property damage of \$1500 or more (Iowa DOT, 2015). The recent finding from the 100-CAR naturalistic driving study found 85% of the crashes were not reported to police, which revealed the magnitude of the under-reporting issue in crash dataset (Dingus, 2006). Additionally, the pre-crash driver behavior information was poorly collected in crash report. Most driver behavior information relied on the drivers' own statement or witness' testimony, which could be inaccurate or biased for various reasons. Therefore, drivers' contributing factor to crashes is still not well understood in transportation safety research community.

### 1.2.2 Driving simulator study

Driving simulator has been widely used in the past years to study drivers' behaviors in the simulated environments. It is a flexible research method that could be used to study a variety of topics in traffic safety, including driver's reaction to different roadway designs and driver performance under alcohol influence. The biggest advantage of driving simulator study is the ability to simulate dangerous driving scenarios without putting drivers at risks. The experiment environment can be relatively easily controlled, so that the cause-effect conclusion could be directly drawn from the studies (Fisher et al., 2011). Therefore, driving simulator study has been used by many researchers for transportation safety research. Figure 1.3 shows state-of-the-art national advanced driving simulator at the University of Iowa.



**Figure 1.3 National advanced driving simulator at University of Iowa (National Advanced Driving Simulator, 2015)**

Nevertheless, driving simulator study also suffers several limitations. The findings from driving simulator studies are often criticized for its validity in real-world driving conditions. Some studies found many drivers had false speed perception and distance perception in driving simulator. Most of driving simulator did not incorporate motion perception, which might conflict with visual cues. Some participants even experienced motion sickness and discomfort which could potentially bias the study results (Casali and Wierwille, 1980; Fisher et al., 2011).

### 1.3 The Concept of Naturalistic Driving Study and Its Opportunities

Naturalistic driving study was introduced in the past few years to better understand the causes of crashes and drivers' daily normal driving behavior. While there is no agreed definition for naturalistic driving study, it is often referred to as an unobtrusive observation method which studies the drivers' daily driving behavior in a natural setting without any experiment control (Dingus et al, 2014; Schagen et al., 2012; FOT- Wiki, 2014; PROLOGUE, 2009). In a naturalistic driving study, drivers do not receive instruction about when, where, and how to drive

their vehicles. This allows researchers, at the first time, to observe how drivers naturally interact with vehicle, roadway, and traffic environment in their everyday driving activities.

Recent development in data collection and storage technology allows researchers to conduct large-scale naturalistic driving study. The typical data acquisition system (DAS) in naturalistic driving study consists of vehicle network, lane tracking system, forward radars, video cameras (Figure 1.4), accelerometers, vehicle network information, Geographic Positioning System (GPS), eye-tracking system and data storage system. Most of the variables are collected at 10Hz which is every 0.1 second. Naturalistic driving study usually collects very large size of data. For instance, SHRP2 NDS collected 4 petabytes (4 million gigabytes) data from more than 3,000 drivers over 2 years.



**Figure 1.4 Four camera views in SHRP2 NDS (Note: Driver in image is a nonparticipant employed by coordination contractor) (Dingus et al., 2014)**

Naturalistic driving study provides many advantages over the traditional safety analysis methods. First and foremost, naturalistic driving study collects a variety of variables regarding



drivers' everyday driving behavior without any experimental control. This information can be used to understand how drivers naturally interact with vehicle, roadway, and traffic environments in their daily driving activities.

Second, naturalistic driving study usually collects very large dataset. For example, the SHRP2 NDS project collected 4 million gigabytes data from 3000 drivers over two years. It covers many roadway types, traffic conditions, and driver behaviors. The SHRP2 NDS is one of the largest dataset that had been collected in the transportation research community, but analyzing such a large dataset brings many challenges.

Third, human factor was known to contribute to 90% of crashes, but it is the least understood factor in crashes. The in-vehicle cameras in naturalistic driving study provide critical information regarding drivers' in-vehicle behaviors before, during, and after crashes. The forward view video also provides critical information about the traffic environments. Above all, naturalistic driving study is a new research method that will result in better understanding of the causes of crashes and drivers' everyday driving behaviors.

The potential users of the study results include transportation agencies, insurance companies, and car companies. For example, transportation agencies could use the information to develop new safety countermeasures and make better informed public policies. Insurance companies could use the information to identify the risky driver groups. Additionally, vehicle companies could learn from driver's interaction with vehicles and traffic environments to develop advanced safety systems.

## 1.4 Problem Statement

The previous section reviewed the scope of traffic safety issue and introduced the concept of naturalistic driving study. This section narrows down the focus to the specific traffic safety problem on rural two-lane curves. Horizontal curve has been the spotlight of transportation safety research for the past decades (Bonneson et al., 2009; Felipe et al., 2007; Schurr et al., 2002). Fatality Analysis Reporting System (FARS) reported 8,059 fatal crashes on horizontal curves, which accounted for one quarter of all motor vehicle fatalities in 2012. The majority of those crashes were located in rural area, especially on rural two-lane roadways. The crash rate on horizontal curves was found to be three times higher than the crash rate on tangent roadways (FHWA, 2014). Similarly, Farmer and Lund (2002) found the odds of a rollover crashes were 2.15 to 2.42 times higher on horizontal curves than tangent segments. Therefore, the crash rate was disproportionally higher on curves and the drivers were found to have difficulty negotiating curves.

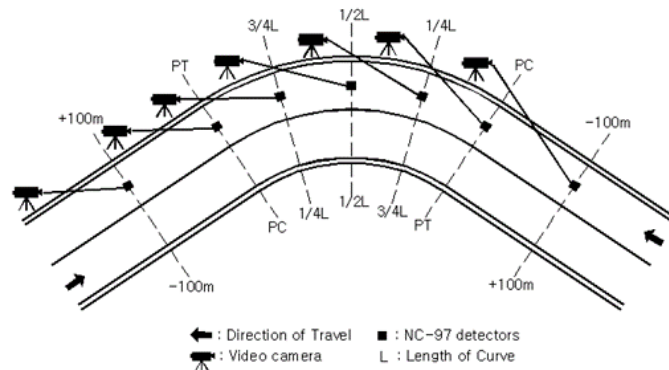
An extensive number of studies has been conducted to study drivers' behaviors on curves. First of all, several researchers found the number of crashes was correlated with curve characteristics, including curve radius, degree of curvature, average annual daily traffic (AADT), length of curve, shoulder width, grade and tangent distance. (Council et al., 1988; Milton and Mannering, 1998; Suh, 2006; Khan et. al., 2013; Schneider, 2010; Torbic, 2004; Zegger, 1991).

The second group of researchers attempted to explain the drivers' curve negotiation behavior from human factor perspective. They argued that the visually distorted travel lane made it more difficult for drivers to recognize the existence and the sharpness of a curve (Charlton S.

G., 2007; Hassan, 2003; Hong, 2006; Kandil, 2010; Mars, 2008; Shinar, 1977; Suh, 2006; Wooldridge, 2000, Zakowska, 2000).

Third, many studies predicted vehicle speeds on rural curves using generalized linear regression models. The 85<sup>th</sup> percentile speeds on curves were predicted based on curve radius, degree of curvature, curve length, tangent distance, and speed limits. It was found the drivers only reduced vehicle speed for curves with radius less than approximately 1000 feet (Anderson, 2000; Bella, 2007; Bella, 2013; Bonneson, 2009; Collins, 2007; Donnell, 2007; Fitzpatrick, 2007; Figueroa, 2005; Felipe, 1998; Montella, 2014; Richard, 2013; Schurr, 2007).

Fourth, the vehicle lateral position on curves was carefully examined by several researchers. Most of them observed curve cutting behavior at mid-point of a curve. Six patterns of curve trajectories were identified in those studies (Abele, 2011; Bella, 2013; Bertola, 2012; Charlton, 2007; Gunay, 2007; Hallmark, 2012; Reymond, 2001; Spacek, 2005; Suh, 2006; Taylor, 2005). Several low-cost countermeasures were proposed to help drivers keep their lateral positions, such as raised pavement marking, rumble strips, dynamics curve warning systems, and chevrons (Cheung, 2010; McGee, 2006).



**Figure 1.5 Video camera locations on the horizontal curves (Park et al., 2002)**

However, there were some research gaps in previous studies. First of all, most of previous studies collected vehicle speeds using roadside radar guns or road tubes at limited number of locations on the curves. One example of this type of experimental setting is illustrated in Figure 1.5. The researchers needed to make extrapolation of vehicle speeds between the measurement locations. With limited number of measurement locations, it was difficult to understand drivers' curve negotiation behavior as a continuous process. Second, driver distraction was cited as the major contributing factors to run-off-road crashes on curves, but driver behavior information was not collected in those in-field studies. Third, vehicle dynamics was highly correlated over time, but previous research failed to consider the temporal correlations of the driving data. The traditional statistical analysis usually summarizes time series data at event level, but many important variables were neglected in the aggregated data. It is interesting to evaluate driver's behavior directly from time series data. Those research gaps will be addressed in this dissertation using the SHRP2 NDS data.

## 1.5 Research Questions

The objective of this dissertation was to understand drivers' curve negotiation behavior through the analysis of SHRP2 NDS data. There were three research questions in this dissertation. The first research question focused on the analysis of cashes and near-crashes on rural two-lane curves using logistic regression model. The second research questions conducted multivariate analysis of drivers' daily normal driving behaviors on rural two-lane curves. The third research focused on analyzing a group of time series observations using functional data analysis (FDA) method. The three research questions are discussed in details in the following sections.

### 1.5.1 Research question 1: what are the contributing factors to crashes and near-crashes on rural two-lane curves?

The objective of this research question is to understand the contributing factors to crashes and near-crashes on rural two-lane curves using the SHRP2 NDS data. Previous studies conducted crash data analysis on curves, but most of the explanatory variables were curve geometries. The behavior variables were poorly collected in previous studies. Fortunately, a number of important driver behavior variables could be collected in naturalistic driving study. The crashes and near-crashes data from the SHRP2 NDS provided in-vehicle drivers' behavior, traffic environment, and roadway conditions. The crash and near-crash data was queried from the SHRP2 InSight website. A total number of 176 crashes, 210 near-crashes, and 2,729 balanced-sample baseline events were included in this analysis. The crashes and near-crashes were combined and defined as safety critical events in this study. The multivariate logit model was used to predict event outcomes from the explanatory variables. The model found the speeds over posted speed limits, wet surface, icy/snowy surface, presence of curb, and visual distraction increased the likelihood of safety critical events on rural two-lane curves. Larger radius and presence of shoulder were found to decrease the likelihood of safety critical events on rural two-lane curves. Overall, this is one of the first analyses on the crashes and near-crashes using the SHRP2 NDS data. This analysis successfully examined driver behavior variables, roadway characteristics, and traffic environments in the logistic regression model.

### 1.5.2 Research question 2: how do drivers normally negotiate different rural two-lane curves?

The objective of this research question is to understand how drivers normally negotiate rural two-lane curves. This analysis included more than 10,000 observations from 202 drivers on 219 curves. The sample size of this study is much larger than any of previous studies. Most of previous studies predicted the 85<sup>th</sup> percentile speeds based on curve geometries. The driver demographics information was not available in previous studies. Fortunately, it is possible to associate driving data to driver demographics in the SHRP2 NDS. The main data source of this study is the time series driving data summarized at event level. Other data sources included forward video, drive demographics, vehicle types, and curve geometries. The drivers' curve negotiation behaviors were examined from two aspects: lateral acceleration and vehicle speed. The vehicle lateral acceleration was found to be highly correlated with vehicle speeds and curve radius. Drivers' 85<sup>th</sup> percentile lateral acceleration was also summarized for different curve radius. Furthermore, the vehicle speeds were examined on both 45 MPH and 55 MPH roadways. The mixed linear model was used to predict vehicles' mean speeds inside the curves. The tangent speed, advisory speed limits, logarithm of radius, car following, and younger drivers were found to be statistically significant for predicting vehicle mean speeds on curves. The model successfully identified the effect of individual drivers' speeding behavior on curves. For future research, it is recommended to examine vehicle lateral position variable in the analysis.

### 1.5.3 Research question 3: how do drivers interact with curve geometry as a continuous process on rural two-lane curves?

The objective of this research question is to evaluate how drivers interact with curve geometry as a continuous process using functional data analysis method. Functional data analysis was a relatively new statistical method that was developed in the past twenty years. It allowed researchers to ask many interesting research questions directly from time series data, which could not be done in traditional statistical analysis. Many researchers are interested in directly analyzing the time series data and understanding how drivers react to roadway geometry as a continuous process. A majority of the variables collected in the SHRP2 NDS are time series data by its nature. This study took the challenge to analyze the time series speed data from the SHRP2 NDS and understand what driver behaviors can be learned from analyzing the time series data using functional data analysis.

Because building functional data models is a time consuming process, only the vehicle speeds from four sample curves were examined in this study. The four rural two-lane curves had radius ranged from 117 feet to 1288 feet. The overall goal was to examine the similarities and differences between the speed profiles collected on the same curve. This study first discussed how to build functional data from discrete time series data. After the functional data were created, the mean and confidence interval for a group of functional observations were calculated to find the typical driving behavior on a curve. The derivative information was also calculated to examine how drivers reacted to the curves differently. Lastly, the functional principal component analysis was used to identify different driving patterns on the same curve. In summary, the functional data analysis method was found to be a ground breaking research method to

summarize the features for a group of time series observations. It is a useful way to illustrate how drivers interact with roadway geometries continuously. This methodology has important implications for analyzing the time series data from naturalistic driving study.

## 1.6 Dissertation Structure

Chapter 2 reviews the previous studies on drivers' curve negotiation behavior from three aspects: curve perception, vehicle speed, and vehicle lateral position. The major findings and the knowledge gaps of previous studies are discussed in this chapter. Chapter 3 reviews the existing naturalistic driving studies and then introduces the experimental design, data acquisition system, and collected variables from the SHRP2 NDS. Chapter 4 addresses the first research question by analyzing crashes and near-crashes using logistic regression model. Chapter 5 focuses on the analysis of drivers' normal curve negotiation behavior using multivariate analysis method. Vehicle speed and lateral acceleration are examined in this analysis. The linear mixed model successfully predicts vehicle mean speeds on rural two-lane curves. Chapter 6 analyzes time series data using functional data analysis. This chapter first introduces the methodology of functional data analysis, and it then applies the functional data analysis on four sample curves. Finally, Chapter 7 summarizes the major findings and limitations of this dissertation, and discusses the recommendations for future research.



## CHAPTER 2. LITERATURE REVIEW

This chapter reviews the literature on drivers' curve negotiation behavior from three aspects: curve perception, vehicle speed, and vehicle lateral position. Section 2.1 discusses the human factor studies about how drivers perceive the existence and sharpness of a curve. Section 2.2 reviews the previous studies on predicting vehicle speeds on curves based on curve geometries. Section 2.3 discusses the previous studies on vehicle lateral positions and trajectories inside curves. The major findings and limitations of previous studies are also discussed at the end of this chapter.

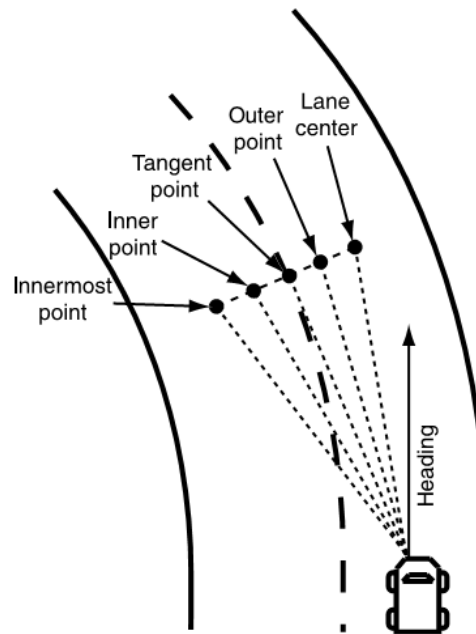
### 2.1 Curve Perception

Many researchers are interested in how drivers perceive and react to different curves. Cacciabue (2007) discussed driving tasks in his book *Modelling Driver Behavior in Automotive Environments* and defined driving tasks into three categories, which were drivers' physical ability (size, reach, force, and endurance), perceptual ability (vision, hearing, touch, and proprioception), and cognitive ability (memory, attention, and decision). This section discusses drivers' perception of curvature and their eye movement patterns in curve negotiation process.

Shinar et al. (1977) conducted one of the earliest studies on drivers' eye movement patterns on curves. Five drivers' gaze behaviors were recorded with a video camera on rural two-lane curves. The drivers were found to fix their gaze locations on the edge line marking, instead of focusing on the line of expansion as they did on tangent roads. Monitoring the edge line allowed drivers to quickly react to any lateral drifts on curves. The gaze patterns were also found to be different between left turn and right turn curves.

Similar conclusions were found in a recent study by Kandil et al. (2010). They examined two gaze strategies on curves: tangent points gaze strategy and sampling points gaze strategy. The drivers' eye movement directions and durations were recorded on a head-mounted eye tracker. They found half of the drivers used tangent point as their eye fixture point on curves. The eye movement patterns were also found to be different on left turn and right turn curves.

In another study, Mars (2008) instructed the drivers to negotiate curves using different eye fixation patterns as shown in Figure 2.1. The five eye fixation patterns included lane center, outer point, tangent point, inner point, and innermost point. The standard deviations of lateral positions were found to be lowest if the drivers fixed their eye sights at the innermost point.



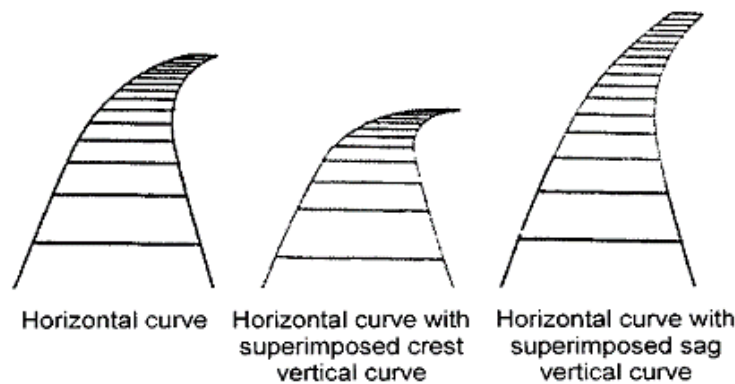
**Figure 2.1 Gaze strategies on curves (Mars, 2008)**

Suh et al. (2006) evaluated the relationship between curve geometries, vehicle speeds, lateral positions, and eye movement patterns with an on-board eye tracking system. The vehicle lateral positions and speeds were measured from roadside cameras. The eye gaze locations were

recorded on both tangent segments and curve segments. The eye movement was found to be more fixed at sharper curves and night time driving scenarios, which were correlated with higher cognitive and visual demands.

Wooldridge et al. (1981) used vision occlusion technique as a measure of drivers' visual demand to control vehicle's lateral position. The underlying assumption was that higher cognitive demands were correlated with higher visual demands. They tested 24 drivers on both isolated single curves and continuous curves. The visual demands were found to be significantly higher at curves with smaller radii.

Vertical slope tends to distort the shape of the curves and mislead the drivers' judgement on curve sharpness. Hassan and Easa (2002) studied the curves with a combination of different radii and vertical slopes. They found horizontal curves looked sharper with crest curves, but they looked flatter with sag curves as shown in Figure 2.2. Bauer et al. (2013) evaluated the effect of vertical curves on the number of crashes. They developed crash modification factors (CMF) for vertical grades on rural two-lane roads. The crash rates were found to increase linearly as vertical grade increased.



**Figure 2.2 Horizontal curves with vertical alignment (Hassan et al., 2002)**

Lastly, Bertola et al. (2012) tested the influence of roadway familiarity, driver inattention, and driver hurriiness on drivers' performance on curves. Fourteen participants' were asked to solve mathematics problems and perform classification tasks in a driving simulator. They used financial incentives to simulate the level of hurriiness in driving. It was found the drivers who were familiar with the roadways drove faster and also had larger lane deviations. They concluded that driver inattention degraded driver's lane keeping behavior. In summary, the human factors were found to play an important role in the curve negotiation process.

## 2.2 Vehicle Speed on Curves

Vehicle speeds on horizontal curves had been investigated extensively in the past studies (Anderson et al., 2000; Bella et al., 2006; Lamm et al., 1988; Levison et al., 1998; Torregrosa et al., 2013; and Wu et al., 2013). Anderson et al. (2000) evaluated the relationship between speed reductions and crash rates on 1,126 curves. They found statistically significant correlation between speed reductions and crash rates. The speed reduction was calculated as the speed difference between 85<sup>th</sup> percentile of vehicle speeds on tangent section and the vehicle speeds at midpoint of a curve.

In another study, Hauer (1999) evaluated the relationship between curve radius and the number of crashes. The study found smaller curve radius was correlated with higher number of crashes. The number of crashes was also in proportion to curve lengths. The speed reductions and crash rates were significantly higher on curves with radius less than 300 meters (1000 ft.).

Numerous studies predicted vehicle speeds based on a number of factors including vehicle type, curve geometry, driving comfort, safety, and law enforcement (Levison, 1998). The American Association of State Highway and Transportation officials (ASSHTO) published the

report *A Policy on Geometric Design of Highways and Streets* book in 2001. As shown in Formula 2.1, they predicted vehicle speeds on curves as a function of side friction demand, super elevation rate, gravitational acceleration, and curve radius.

$$v_c = \sqrt{gR \left( f_D + \frac{e}{100} \right)} \quad (\text{Formula 2.1})$$

$v_c$  = curve speed, ft/s

$f_D$  = side friction demand factor (or lateral acceleration);

$e$  = super elevation rate (percent);

$g$  = gravitational acceleration (32.2 ft/s<sup>2</sup>);

$R$  = radius of curve (ft);

In addition to the ASSHTO guideline, numerous studies built speed prediction models based on curve radius, tangent speed, deflection angel, curve length, sight distance, vertical slope, vehicle type, and lane width. Bonneson et al. (2009) predicted vehicle speeds on 55 horizontal curves on rural two-lane highways. A total of 6,677 passenger car observations were collected in this study. The speed prediction model included curve radius, curve length, width of traffic lanes and shoulders, super elevation, and vertical grade. The drivers were also found to tolerate higher lateral acceleration on low speed curves.

The tangent distance between consecutive curves was also found to have significant impact on vehicle speed. Findley et al. (2012) modeled the spatial relationship of the neighboring curves, such as tangent distance, curve direction, radius, and length. They concluded the distance to adjacent curves was a reliable indicator for predicting the number of crashes.

Schurr et al. (2002) examined the relationship of curve design, operating speed, and posted speed for 56 sites in Nebraska. They installed two detectors at the entry and mid-point of the curves. Several linear regression model were used to predict vehicle speeds based on the explanatory variables, including guardrail, bridge, traffic control signs, posted speed limit, intersection types, and lane width. They also found vehicle speeds were highly correlated with deflection angel, curve length, posted speeds, vertical grade, and AADT. They hardly observed any speed reduction when the curve radius was larger than 350 meters (1148 feet).

Fitzpatrick et al. (1999) predicted vehicle speeds as a function of roadway geometry on rural two-lane highway. The data was collected at six states, including Minnesota, New York, Pennsylvania, Oregon, Washington, and Texas. More than 100 observations were collected at each site. The speed data was collected with radar gun and on-pavement piezoelectric sensors. They did not find significant impact of spiral curves on the 85<sup>th</sup> percentile vehicle speed. They did not observe speed reduction if the curve radius was larger than 800 meters (2625 feet). Significant speed reduction was observed only for curves with radii less than 250 meters (820 feet).

In order to improve drivers' performance on curve negotiation, several researchers evaluated different types of speed reduction countermeasures on curves. Charlton (2007) tested two groups of treatments in a driving simulator. The first treatment group included different types of curve warning signs. The second treatment group included different types of on-pavement markings. In summary, chevron and rumble strip were found to reduce vehicle speeds effectively. The herringbones road marking significantly improved drivers' lane keeping behavior on curves. Advanced Curve Sign and Advisory Speed Warning were found to have

little impact on vehicle speeds. The conclusion was that more noticeable visual cues were often more effective to influence drivers' behaviors on curves. However, the study was conducted on driving simulator, which might not represent the real world driving scenarios.

Researchers were also interested in the influence of curve spiral design on drivers' curve negotiation behavior. Passetti et al. (1999) compared the 85<sup>th</sup> percentile speeds on 12 curves with spiral design and 39 curves without spiral design. Regression model was used to test the effect of spiral on vehicle mean speeds, but no statistically significant relationship was found between spiral design and vehicle speeds. In a similar study, Council (1998) examined crash data on 15,000 transition sections in the state of Washington and found spiral design only decreased crash rates on the sharp curves.

Bella et al. (2013) tested drivers' behavior on curves with different roadway configurations. They found the driver behavior was only influenced by cross-sections and geometric elements, rather than the roadside environments. They did not find statistically significant impact of guardrail on vehicle speeds and lane positions on curves. They also confirmed the drivers had tendency to cut the curves on both left turn and right turn curves.

More recently, the on-board data collection technology allowed researchers to collect vehicles' continuous speeds on curves with instrumented cars. Montella et al. (2015) measured the vehicle speeds on curves on 45 horizontal curves in Italy. They found the drivers' speeds were not constant over the curves. They also found significant differences between different drivers. This study pointed out the importance to study the continuous speed profile of vehicle speeds, which will be addressed later in Chapter 6.

Most studies on horizontal curves collected data on passenger cars. Donnell et al. (2001) built speed prediction model for trucks on 17 sites. Regression model was used to predict vehicle speeds on curves. They found the degree of curvature and vertical grades had statistically significant impact on speed reduction for trucks on curves.

Several studies also examined the relationship between vehicle speed and vehicle lateral acceleration on curves. Felipe et al. (2007) examined driver's comfortable lateral accelerations and vehicle speeds on curves. The drivers were asked to drive the vehicles in a test track at their most comfortable speeds and also at their maximum speeds. The test sites had curve radius between 16 meters to 100 meters. It was found the lateral acceleration for comfortable driving was within 0.35 to 0.40 g.

In summary, the relationship between curve geometries and vehicle speeds has been investigated extensively by previous researchers. The speed prediction models from previous studies are summarized in Table 2.1. Most of studies found curve radius, curve length, curve deflection angel, and vertical slope had significant impact on vehicle's speeds on curves. However, those studies only predicted vehicle speeds based on curve geometries. It was not clear how individual drivers' driving style influenced vehicle speeds on curves. Additionally, most previous studies measured vehicle speeds using roadside equipment, which only collected speed data on limited number of locations on curves. The recent development of in-vehicle data collection technologies allowed researchers to observe drivers' continuous speed profile at high frequency. The detailed speed data in naturalistic driving studies allows researchers to examine drivers' speeds as a continuous process.

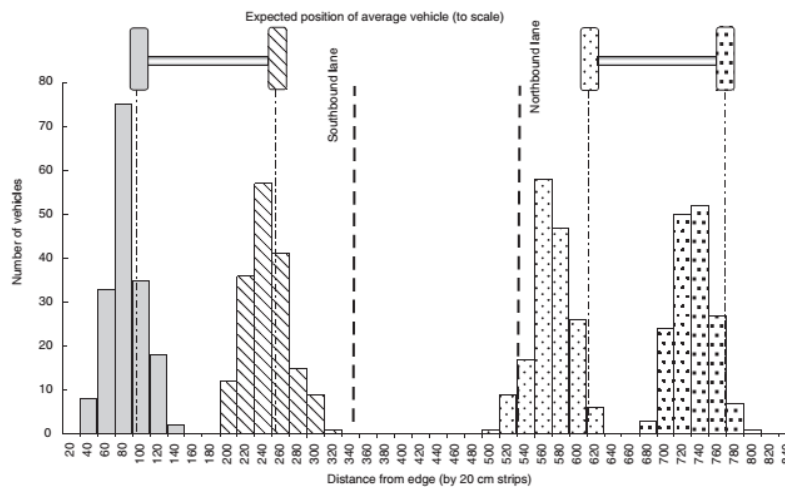


**Table 2.1 Review of speed prediction models**

Author	Sample Size	Formula	Independent Variables
Bonneson et al. (2009)	55 curves	$V_c = 11.0R \left( -b_2 + \sqrt{b_2^2 + \frac{4c}{32.2R}} \right)$	radius, super elevation
Schurr et al. (2002)	56 curves	$V_{md} = 67.4 - 0.1126\Delta + 0.02243L + 0.276V_p$	deflection angel ( $\Delta$ ), curve length (L), posted speed ( $V_p$ )
Fitzpatrick et al. (1999)	24 curves	$V_{85} = 56.34 + 0.808R^{0.5} + 9.34/AD$	curve radius (R), approach density(AD)
Krammes et al. (1995)	138 curves	$V_{85} = 102.45 - 1.57D + 0.0037L - 0.10I$	degree of curvature (D), length of curvature (L), deflection angle (I)
Taragin et al. (1954)	35 Curves	$V_{90} = 88.87 * \frac{2554.76}{R}$	radius (R)
Glennon & Weaver (1971)	56 curves	$V_{85} = 103.96 * \frac{4524.94}{R}$	radius (R)
Lamm et al. (1988)	261	$V_{85} = 94.39 * \frac{3189.94}{R}$	radius (R)
Ottesen & Krammes (2000)	138 curves	$V_{85} = 103.64 * \frac{3400.73}{R}$	Radius (R)

### 2.3 Vehicle Lateral Position on Curves

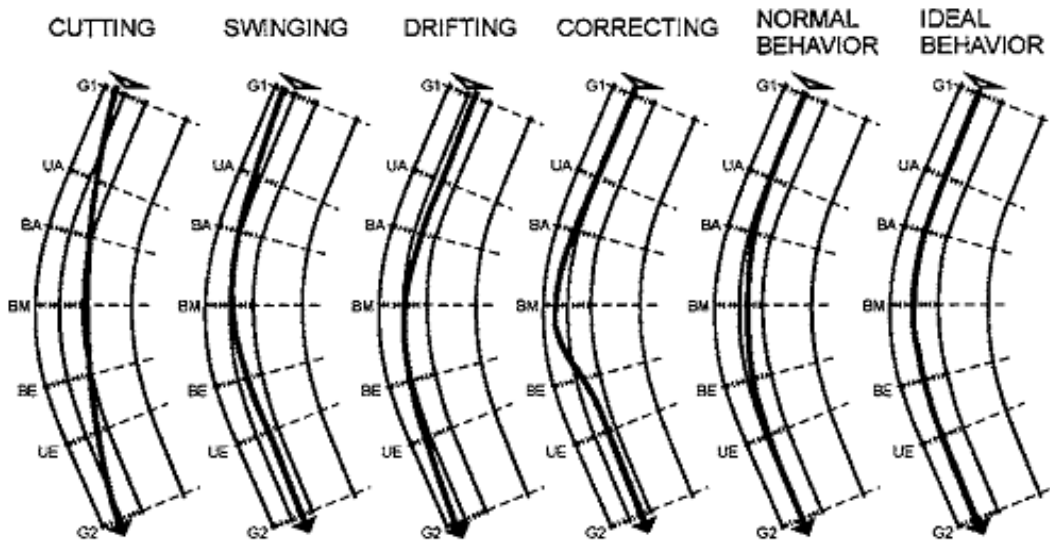
Vehicle lateral positions was often used as crash surrogate for roadway departure crashes. Many studies found the drivers had difficulty to keep their lane positions inside curves. Therefore, a number of researchers investigated how drivers kept their lane positions on different curves. Gunay et al. (2007) recorded vehicles' lateral positions on curves using video cameras. The lane deviation was calculated as the number of pixels between vehicle wheels and the edge line of lane marking. The distribution of wheel locations on curves is plotted in Figure 2.3. The vehicles were found to shift to the center of the curves at the midpoint of a curve, which is also known as curve-cutting behavior.



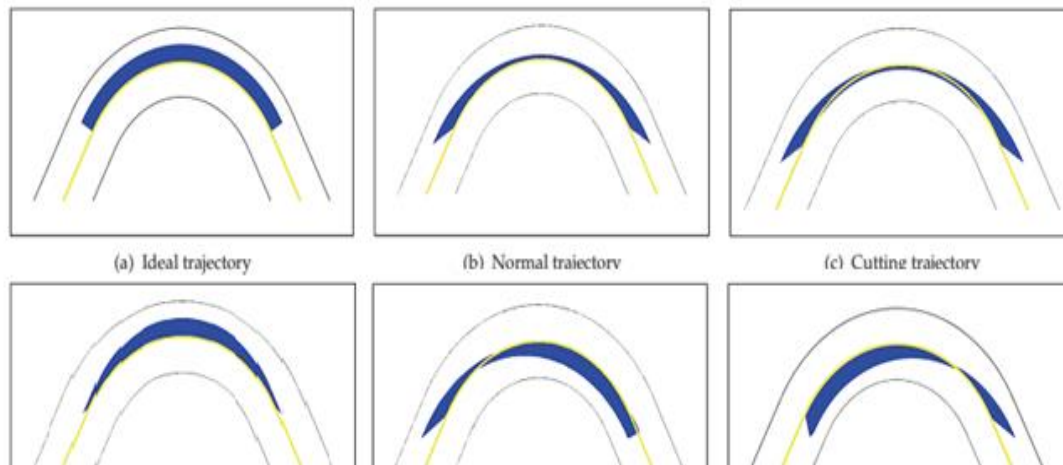
**Figure 2.3 The distribution of wheel position on the study curves (Gunay et al., 2007)**

In another study, Hallmark et al. (2012) investigated the relationship between curve speed and lateral positions at different locations on curves. Hallmark used the Z-configuration road tubes to measure vehicle lateral positions and speeds at three study sites in Iowa. The odds ratios of roadway departure events were 2.37 to 4.47 times higher for vehicle traveling 5 mph over speed limits.

Several researchers attempted to categorize vehicle's trajectory patterns on curves (Ren, 2012; Spacek, 2005; Verkehrssicherheit, A., 1980). In a field study, Spacek (2005) measured the vehicle lateral positions and speeds with twelve roadside measuring posts. The vehicle dynamics were collected on six left turn curves and six right turn curves with curve radius ranged between 65 meters and 195 meters. Six types of curve negotiation types were identified on left turn curves as illustrated in Figure 2.4. The six trajectory patterns included ideal behavior, normal behavior, cutting behavior, drifting behavior, swinging behavior, and correcting behavior. In a similar study, Ren et al. (2012) developed mathematical models for the six behaviors and simulated the results in Matlab. The vehicle trajectories was plotted in Figure 2.5., which was very similar to Spacek's (2005) proposed trajectories.



**Figure 2.4 Sketches of six types of trajectories (Spacek, 2005)**



**Figure 2.5 Six patterns of vehicle track behavior on curves simulated in Matlab (Ren et al., 2012)**

Some researchers evaluated the effects of countermeasures on vehicle lateral positions. Rumble stripes and rumble strips became more and more popular in recent years to address the roadway departure problem. Taylor et al. (2005) evaluated the effects of rumble strips with a single paved lane marking and double paved lane marking. The vehicle lateral position was measured with road tubes on curves. The rumble strips were found to make drivers move away from the edge lines and also resulted in smaller variance of lateral positions.

In another study, Rasanen et al. (2005) measured vehicle speeds and lateral positions on curves before and after milling the rumble strip. They did not find statistically significant differences in vehicle speeds, but the curve encroachment rate dropped from 9.2% to 2.5%. The standard deviation of lateral positions were also smaller with rumble strip installed on roads.

Bella (2011) examined different roadway cross-sectional designs and roadside environments in a driving simulator study. The driver's behaviors were found to be influenced

by geometric elements, rather than the roadside environments. They did not find the impact of guardrail on vehicle speed and lane position.

Charlton (2007) evaluated different roadway designs on drivers' speed and lateral position on curves. The chevron and rumble strip were found to reduce drivers' speed effectively. The herringbones road marking improved drivers' lane keeping behavior significantly. Interestingly, advanced curve sign and advisory speed warning sign did not have any significant impact on vehicles' speed.

## 2.4 Summary of Major Findings

This chapter reviewed previous studies on drivers' curve negotiation behavior from three aspects: curve perception, vehicle speed, and vehicle lateral position. A number of important findings were discovered in the previous studies. Some of the most important findings are summarized in the following paragraphs.

Section 2.1 reviewed the studies on drivers' perception of curves from human factor perspective. They explained why drivers have difficulty negotiating a curve. Several researchers evaluated drivers' eye movement patterns on curves and found drivers were likely to focus their eye sights on edge lines so that they can quickly react to the changes in lateral positions on curves (Kandil et al., 2010; Mars et al., 2008; Shinar et al., 1977). Several papers found cognitive and visual demand were much higher on sharp curves and night time driving (Suh et al., 2006; Wooldridge et al., 1981). The vertical slope was found to distort the shape of a curve severely and mislead drivers' judgement on curve sharpness (Bauer et al., 2013; Hassan and Easa, 2002). Lastly, familiarity with roadway, driver inattention, and driver hurriness were found to have

significant impact on driving performance (Bertola et al, 2012). Those studies found human factor played important roles in the curve negotiation process.

Section 2.2 focused on the analysis of vehicle speeds on curves and many speed prediction models were summarized from these studies. A few studies examined the relationship between curve radius and the number of crashes. They found the crash rate increased as curve radius decreased, especially for curves with radius less than 1000 feet (Anderson et al., 2000; Hauer et al., 1999). Additionally, several studies built speed prediction models with a number of explanatory variables as shown in Table 2.1. The important variables included tangent speed, vehicle type, curve radius, deflection angle, curve length, sight distance, vertical slope, lane width, and presence of shoulder (Bonneson et al., 2009; Findley et al., 2012; Schurr et al., 2002;). Several researchers examined the effectiveness of different countermeasures for vehicle speeds on curves. The rumble strips, chevron, and dynamic speed feedback signs significantly reduced drivers' speed on curves. On the contrary, the spiral curve, advanced curve sign, and guardrail did not have significant speed reduction on curves (Bella et al., 2013; Charlton 2007; Hallmark et al., 2013; Lamm et al., 1988; Torregrosa et al., 2013, and Wu et al., 2013). Many studies only observed significant speed reduction when curve radius was less than 1000 feet (Schurr et al., 2002; Fitzpatrick et al., 1999). Several studies examined the relationship between vehicle speeds and lateral accelerations on curves. The drivers were found to tolerate higher lateral acceleration on curves with smaller radius (Felipe et al., 2007; Raymond et al., 2001).

Section 2.3 discussed drivers' lane keeping behavior on curves and summarized several vehicle trajectory patterns. Most of the lateral position studies were measured with roadside equipment. Curve cutting was found to be a frequent behavior observed in several studies

(Gunay et al., 2007; Verkehrssicherheit, 1980; Spacek, 2005). The speeds over speed limits also had negative impact on drivers' lane keeping behavior (Hallmark et al., 2012). Several researchers categorized vehicle trajectories into six types, which were ideal behavior, normal behavior, cutting behavior, drifting behavior, swinging behavior, and correcting behavior (Verkehrssicherheit, 1980; Ren, 2012; Spacek, 2005). Additionally, rumble strip was found to be effective to reduce lane encroachment rate on curves (Taylor et al., 2005; Rasanen et al., 2005). Guardrail had little effect on vehicle speed and lane keeping behavior (Bella et al., 2011). The herringbones roadway marking was found to improve driver's lane keeping behavior significantly (Charlton, 2007).

In summary, significant amount of attentions was focused on drivers' behavior on curves in the past decades. However, there were still some shortcoming in previous studies. The major hurdle of previous studies was the difficulty to collect high quality data. The roadside equipment can only measure limited sample size at specific locations on a curve. Additionally, it was difficult to collect driver demographics and driver behavior information, so the influence of individual driver's driving style was unknown. Fortunately, the SHRP2 NDS data could be used to address these issues and help us better understand the curve negotiation process on rural two-lane curves.

## CHAPTER 3. INTRODUCTION TO THE STRATEGIC HIGHWAY RESEARCH PROGRAM 2 NATURALISTIC DRIVING STUDY

This chapter reviews the previous naturalistic driving studies and lays the foundation for the discussion in the next few chapters. Section 3.1 reviews the previous and on-going naturalistic driving studies inside and outside U.S. Section 3.2 introduces the experimental design, data acquisition system, and the variables collected in the SHRP2 NDS. The pros and cons of the SHRP2 NDS data are also discussed by the end of this chapter.

### 3.1 Review of Previous Naturalistic Driving Studies

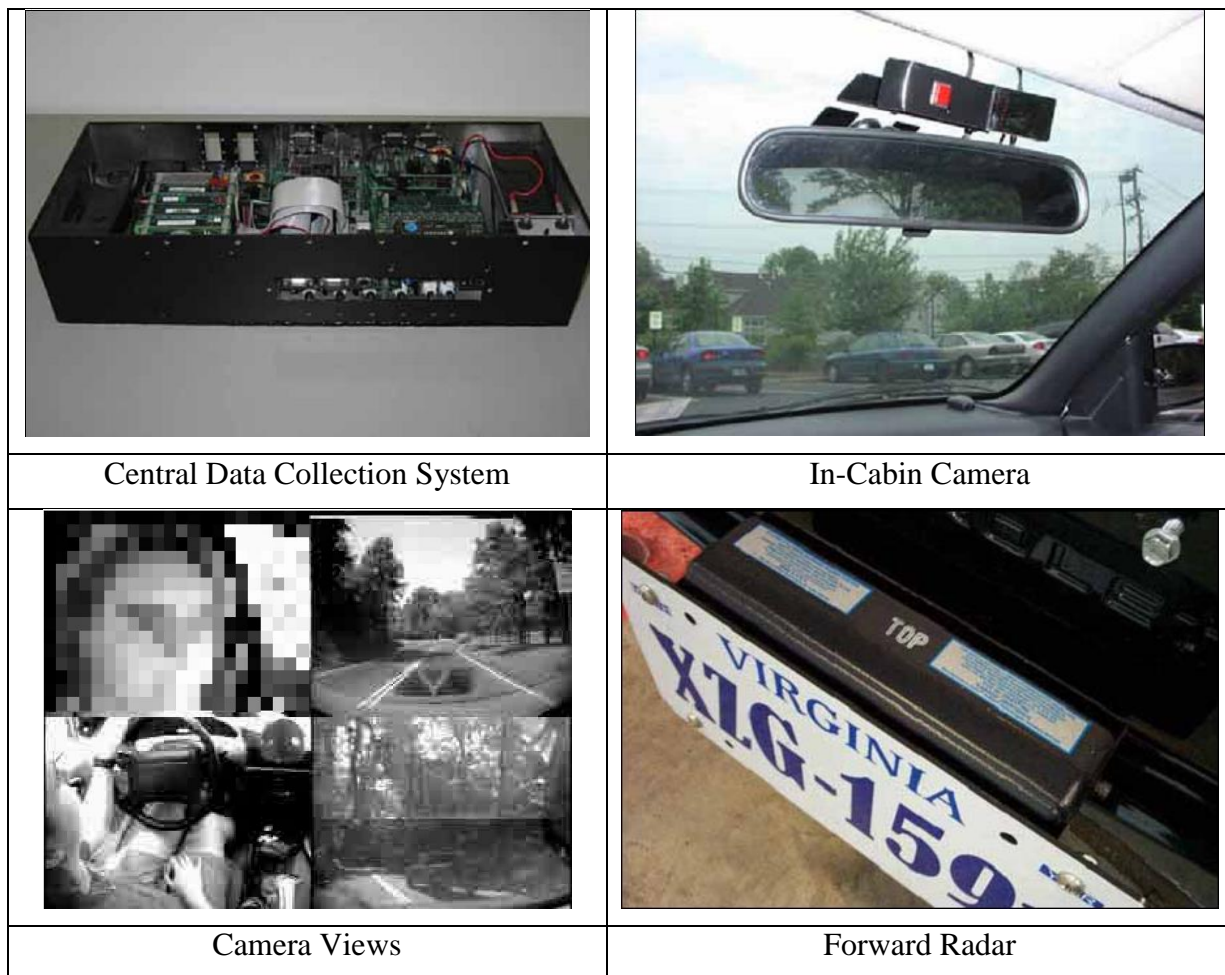
This section reviews five naturalist driving studies inside and outside U.S. The 100-Car Naturalistic Driving Study was one of the first large-scale naturalistic driving studies in the U.S. This project is the precursor for the large-scale SHRP2 NDS. Additionally, the Naturalistic Teenage Driving Study (NTDS) was conducted by U.S. Department of Health and Human Services with more focuses on teenagers' driving behavior. The three international studies in Europe, China, and Canada are still in the data collection phase by the time of writing this dissertation. These studies are briefly discussed in this section. It should be noted that there could be other naturalistic driving studies and this list might not be comprehensive.

#### 3.1.1 100-Car naturalistic driving study, U.S.

The 100-Car Naturalistic Driving Study collected one year's driving data from 241 primary and secondary drivers. This study included 43,000 hours' driving data for approximately 2 million traveled miles. The data acquisition system used in this study is shown in Figure 3.1. This graph shows the central data collection system, in-cabin camera, camera views, and forward



radar. The variables collected in this study were vehicle speed, vehicle headway, longitudinal and lateral acceleration, GPS, brake activation, turn signal, camera views, and lane position. Most of those variables were recorded at 10 Hz, which is every 0.1 second. In addition, the drivers were asked to answer questionnaires about their driving history, demographics, and driving skills before they entered the study. The variables provided rich information regarding how drivers interact with vehicle, roadway, and traffic environments.



**Figure 3.1 Data acquisition system in the 100-Car study (Dingus, 2014)**

One of the most frequently asked question was whether the drivers were affected by the presence of data acquisition system in their vehicles. This study investigated this issue and found

the drivers adapted to the presence of DAS within the first few hours of driving. During one year's study period, 82 crashes were recorded in this study, but only 21% of those crashes were reported to the police, which again confirmed the under-reporting issue in police-reported crash dataset.

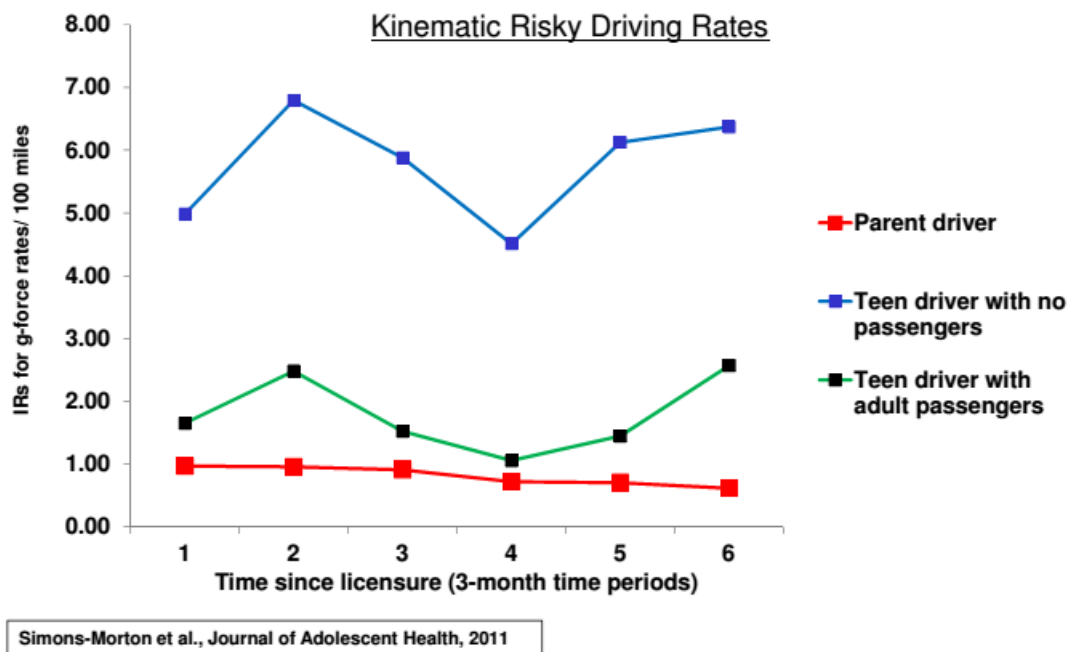
The most common types of crashes were single vehicle run-off road crashes and rear-ended crashes. The primary cause of those crash and near-crashes was driver distraction, which was a contributing factor to 78% of crashes and 65% of near-crashes. The younger driver group aged from 18 to 35 years was found to have disproportionally higher chance to get involved in distraction-related crashes than the older driver group (35 years old or above). The use of digital device, such as cellphone and music player, was the most frequent type of distraction. However, this study only collected drivers' behavior in Virginia and Washington D.C. areas, which might not represent the overall U.S. driver population. Another limitation was the data quality issue such as missing data and large noises in the raw data.

### 3.1.2 Naturalistic teenage driving study, U.S.

Naturalistic teenage driving study was conducted by U.S. Department of Health and Human Services. The objective of NTDS was to understand teenage drivers' driving performance and the risks associated with driver distractions. The DAS included cameras, radar, accelerometer, GPS, lane position sensor, and vehicle network information. This study installed data acquisition system on vehicles with 42 newly licensed teenage drivers over 18 months. The average age of those drivers was 16.4 years. Additionally, 54 parents participated in the study as the control group. This project collected approximately 102,000 trips over 500,000 miles. A total number of 40 crashes and 270 near-crashes were collected in this study. The main research

questions were to 1) understand the teenage drivers' exposure to crash and near-crash events; 2) identify the influence of secondary tasks on teenager's driving performance; 3) study the influence of teenage passengers on the teenage drivers' driving performance.

One of the most interesting results is shown in Figure 3.2, which plots the extreme G-force events per 100 miles driven by teenage drivers (blue line) and the adult drivers (red line). They found the teenage drivers have systematically higher crash risks compared to their parents. Interestingly, if the teenage drivers drove with adult passengers inside the vehicles, the crash risks dropped dramatically than those teenage drivers who drove themselves. In addition, the odds ratio for extreme event was 8.32 for dialing a cellphone, 3.87 for texting, 3.90 for looking at roadside object, and 2.99 for eating and drinking.



**Figure 3.2 Extreme G-force events per 100 miles for teen drivers, the parents, and the teen driver with adult passengers (Simons-Morton et al., 2014)**

### 3.1.3 UDRIVE, Europe

The eUropean naturalistic Driving and Riding for Infrastructure & Vehicle safety and Environment (UDRIVE) is the first large-scale naturalistic driving study conducted in Europe. The objective of this project is to understand drivers' daily driving behavior and its influence on transportation safety and environment. The UDRIVE focused on five research questions, which are 1) crash causation; 2) normal driving behavior; 3) distracted driving; 4) vulnerable road users; 5) driving style and eco-driving (Eenink et al., 2014).

In the UDRIVE Naturalistic Driving Study, the driving data were collected from 2012 to 2016 in seven countries in Europe. UDRIVE collected driving data from 290 participants with a variety of vehicle types, including passenger cars, trucks, and motorcycles. The data acquisition system included camera, CAN, GPS, accelerometer, and speed sensors as shown in Figure 3.3. The results from this study will guide the future transportation regulation, enforcement, driver awareness, driver training, and road design methods in Europe.



**Figure 3.3 Data acquisition system in the UDRIVE project (Pierre, G. S., 2014)**

### 3.1.4 Shanghai naturalistic driving study, China

The Shanghai Naturalistic Driving Study is the first naturalistic driving study conducted in China, which is a joint effort among Tongji Jiaotong University, Virginia Tech Transportation Institute, and General Motor Company. The primary goal of this study was to investigate similarities and differences between Chinese drivers and drivers from other countries. Another goal of this study was to learn how drivers interact with the advanced safety warning system. The study recruited 90 participants over 2 years. Each driver drove the instrumented vehicle for 2 months. A total of 5 vehicles (2 Buick Lacrosse, 2 Chevy Cruze, and 1 Cadillac) were used in this study. This study is still in the data collection phase.



**Figure 3.4 Shanghai naturalistic driving study (Fang, 2014)**

### 3.1.5 Canada naturalistic driving study, Canada

The Canada naturalistic driving study was conducted in Saskatoon, Canada sponsored by Canadian Deputy Ministers of Transport and Highway Safety. A number of 125 drivers participated in this study. The data acquisition system was designed by VTTI and thus the collected variables was very similar to the 100-Car study. The equipment includes head unit, main unit, and front radar. This project is still in the data collection phase.



**Figure 3.5 Head unit, main unit, and front radar used in Canada naturalistic driving Study (Hankey, 2014)**

In summary, naturalistic driving study is a relatively new research method that is getting more and more popular in recent years. Several studies had been conducted in the U.S. The three international naturalistic driving studies are still in the data collection phase. In addition to these naturalistic driving studies, the SHRP2 Naturalistic Driving Study is one of the largest naturalistic driving studies by far, which will be discussed in details in the next section.

## 3.2 Introduction to the SHRP2 Naturalistic Driving Study

This section introduces the Strategic Highway Research Program 2 Naturalistic Driving Study. The information includes study background, data acquisition system, roadway information system, driver data, vehicle data, and time series data. The crash and near-crash events were extracted from the SHRP2 NDS as a separate dataset. The challenges of working with the SHRP2 NDS dataset is discussed at the end of this section.

### 3.2.1 Study background

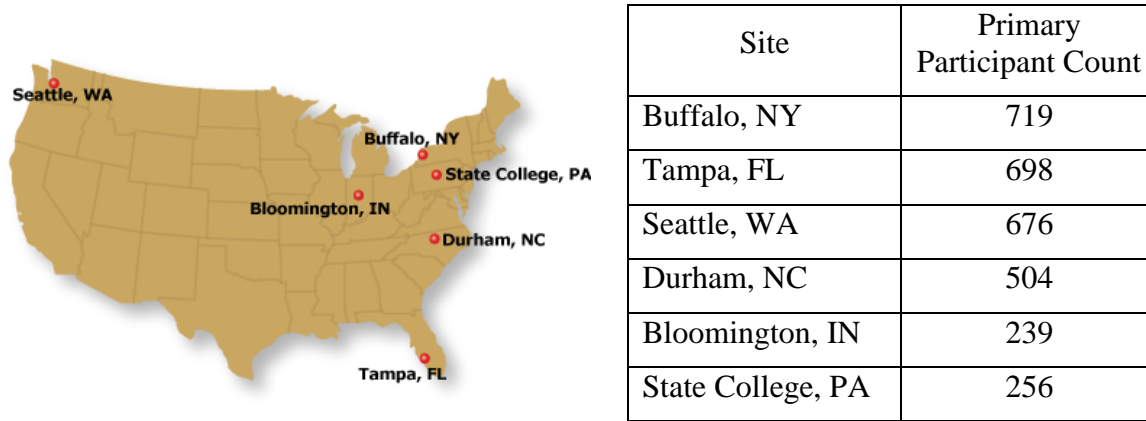
The Strategic Highway Research Program 2 Naturalistic Driving Study (SHRP2 NDS) is state-of-the-art naturalistic driving study and collected more than 4 petabytes (4 million gigabytes) data in total. The objective of the SHRP2 NDs is to investigate how drivers interact

with vehicle, roadway, traffic conditions, and traffic control devices. The project helps us understand the causes of crashes and the associated risks with different driving behaviors. The findings can be used to propose better roadway designs and develop effective safety countermeasures.

The study collected more than 3,092 drivers' real-world driving data from October 2010 to November 2013 (Dingus, 2014). A total of 5 million trips, more than 3,900 vehicle years, and one million hours of driving data were collected at six study sites in the United State. The study sites included Indiana, Pennsylvania, Florida, New York, North Carolina, and Washington State. The number of participants at each site is listed in Figure 3.6. Those sites and participants were selected as a nationally representative sample for a variety of roadway types, driver demographics, weather conditions, and state laws. The detailed data collection plan was explained in the report *Naturalistic Driving Study: Technical Coordination and Quality Control* (Dingus et al., 2015). Those participants were recruited with a variety of methods, such as call center, emails, websites, flyers, etc. A summary of the recruitment method, age, and sites is shown in Table A.1 (Appendix A). The drivers were compensated with \$300 each year. A certificate of confidentiality was initiated by the National Institute of Mental Health (NIMH) to protect the participant privacy.

In order to examine the representativeness of the participant group, Antin et al. (2015) compared the SHRP2 NDS driving profiles to the national traffic statistics in the report *Naturalistic Driving Study: Descriptive Comparison of the Study Sample with National Data*. They concluded that the six study sites were similar to the nation's temperature patterns, but the six sites had more annual rainfall and more urban development area. The comparison SHRP2

participants versus U.S. driving population percentages by age groups is shown in Figure A.1 (Appendix A).



**Figure 3.6 SHRP2 NDS data collection sites and the number of participants (VTTI, 2013)**

### 3.2.2 Data acquisition system

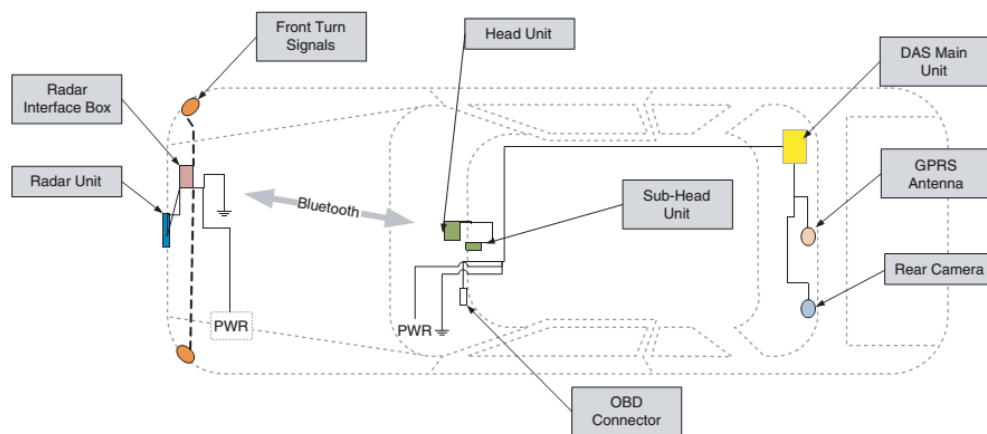
The data acquisition system (DAS) used in this study was the state-of-the-art data collection system, which included forward and rear radar, four video cameras, accelerometer, vehicle controller area network, GPS, lane-tracking system, alcohol sensor, incident button, illuminance sensor, and data storage system. The DAS equipment and the installation schematic is illustrated in Figure 3.7 and Figure 3.8. The forward radar collected vehicle headway and its closing rate to the front vehicle. The video cameras recorded in-vehicle driver behavior and roadway environments. The accelerometer recorded vehicle's acceleration and deceleration. The vehicle controller area network (CAN) provided a variety of vehicle operating status, such as brake activation, electronic stability control, steering wheel angel, etc. The GPS provided vehicle locations which is used to link the driving data to the roadway information system (RID). The lane-tracking system measured vehicles' lateral position. The alcohol sensors measured the



presence of alcohol to detect drunk driving activities. The illuminance sensor measured forward visibility from driver's view. Lastly, the data from different sources was synthesized into the main unit system (NextGen) for processing, recording, and communicating between different devices.



**Figure 3.7 Data acquisition system components used in the SHRP2 NDS (including NextGen, Head Unit, Radar, Network Box, Radar Interface Box (Dingus et al., 2014))**



**Figure 3.8 SHRP2 data acquisition system installation schematic (Antin et al., 2011)**

The four cameras played an important role in the SHRP2 NDS project to observe the external driving environments and the in-cabinet driver activities. There were four cameras

inside the vehicle, including Forward View, In-cabin Driver Face View, Instrument Panel View, and Rear View as shown in Figure. 3.9. Those information was extremely useful for researchers to understand the context of the driving events. The forward video (upper left quadrant of Figure 3.9) had a colored view that provided information regarding traffic conditions, roadway types, pavement conditions, highway signage location, area types, weather conditions, and other unexpected on-road hazard events. Additionally, it was critical to observe drivers' in-cabin eye-glance behavior or other secondary tasks using the driver face video (upper right quadrant of Figure 3.9). Driver's distraction behavior was coded form the driver face view, such as eating foods, texting, and talking to the passengers. However, a driver might wear sun glass and the eye sight direction had to be speculated from drivers' head positions. The driver face video could only be coded by trained data analyst at VTTI Data Secure Enclave. The instrument panel view (bottom left quadrant in Figure 3.9) was used to understand drivers' interaction with steering wheels and the infotainment system at the center console. The rear view (bottom right quadrant of Figure 3.9) was used to monitor the traffic density and the risks from the following vehicle. Lastly, a cabin snap shot was taken at every 10 minutes to record the number of passengers in the vehicles. The four camera views were found to be extremely helpful for researchers to understand the drivers' in-vehicle behavior and the context of driving events. The video information was lacking in the traditional crash data set.



**Figure 3.9 Four camera views from the SHRP2 data acquisition system: forward view (upper left), in-cabin driver face view (upper right), instrument panel view (bottom left), rear view (bottom right). (The driver in the picture is not experimental participant and just for demonstration purpose only) (Dingus et al., 2014)**

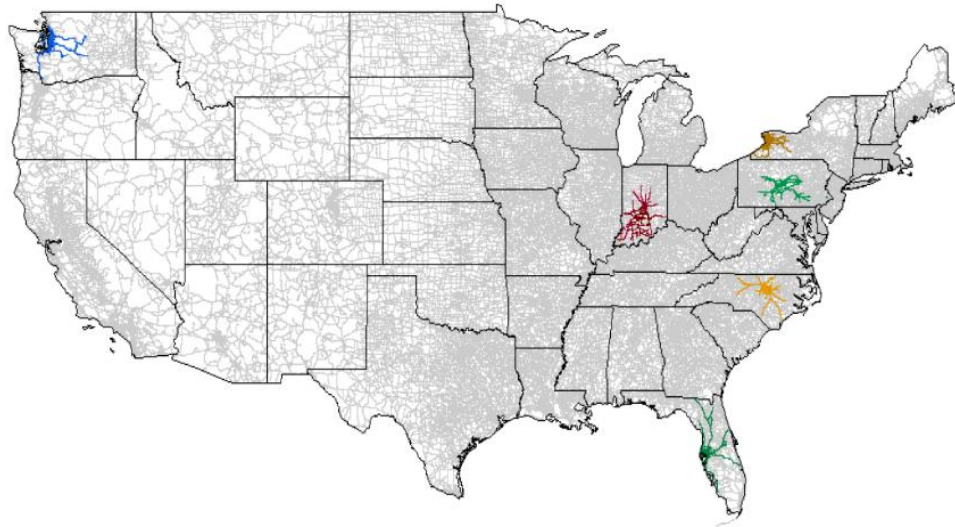
### 3.2.3 Roadway information database

Meanwhile, the roadway information database (RID) was also developed in the SHRP2 NDS project and could be linked to the driving data. The integration of the two data sources allowed researchers to identify the locations of interests and examine the driver interactions with roadway characteristics. The RID was collected and maintained by Center for Transportation Research and Education (CTRE) at Iowa State University. The roadway asset inventory was collected by instrumented mobile van driving at posted speed limits (Smadi, 2015). Due to the limited resources, only the roadway with high trip densities and greatest research interests were collected in this project. A list of the typical collected variables is shown in Table 3.1. As a

result, a total number of 25,076 miles of roadway data and more than 7 million of roadway assets were collected in the Geographic Information System (GIS). The colored roadway links in Figure 3.10 showed the RID data collected at the six sites.

**Table 3.1 List of collected variables in the roadway information database (RID)**

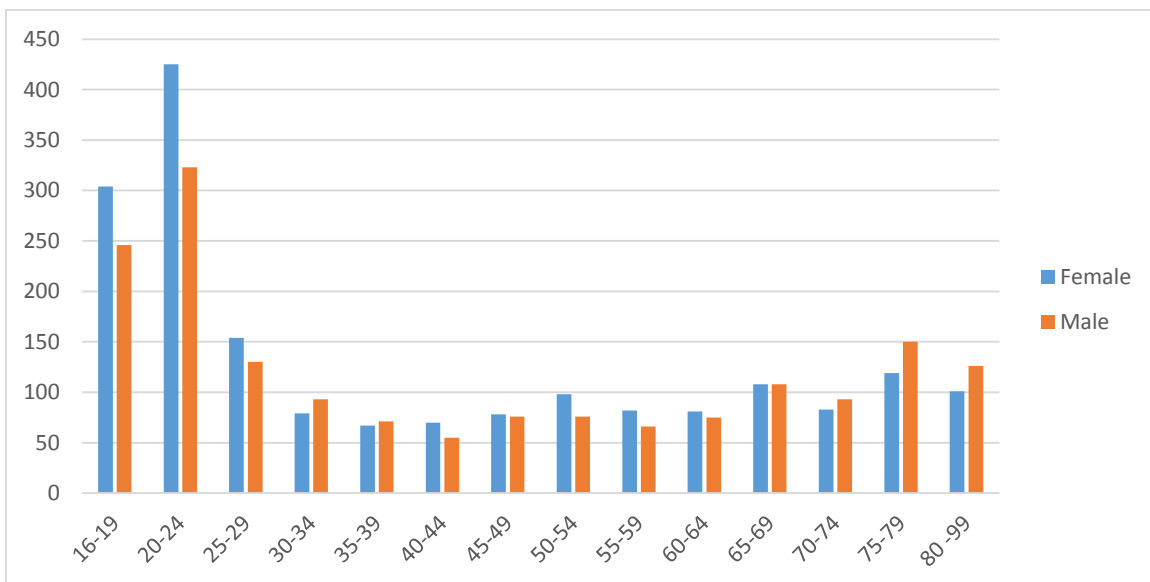
Curve Radius	Lane Types
Curve Length	Shoulder Types
Curve Point of curvature	All Signs on MUTCD
Curve Point of tangency	Guardrails/Barriers
Curve Direction	Intersection Geometry
Grade	Median Types
Cross-slope/Super elevation	Presence of Rumble Strips
Number of Lanes	Presence of Lighting
Lane Width	



**Figure 3.10 A site map shows the RID database in the SHRP2 NDS (Smadi, 2015)**

### 3.2.4 Driver data

According to the SHRP2 NDS Insights website, 3,092 drivers aged from 16 to 95 were recorded at six sites. Figure 3.11 shows the distribution of participants' ages and genders. A careful experimental design and recruiting process was conducted to represent the national drivers' demographics. Ideally, the participants should be randomly selected from the population of interests, who are all the licensed drivers in the United States. The recruitment methods were social media, posting ads on local newspapers, and distributing flyers. The younger drivers and older drivers were overrepresented in the study since they are more vulnerable drivers than other driver groups. Antin et al. (2014) compared the SHRP2 NDS participants to the national driver population and found there were slightly higher proportion of population identifying themselves as white and with college degree, and slightly lower proportion of married and full-time employed drivers.



**Figure 3.11 Drivers demographics by age group and gender (Insight Website, 2014)**

**Table 3.2 List of driver surveys collected in the SHRP2 NDS**

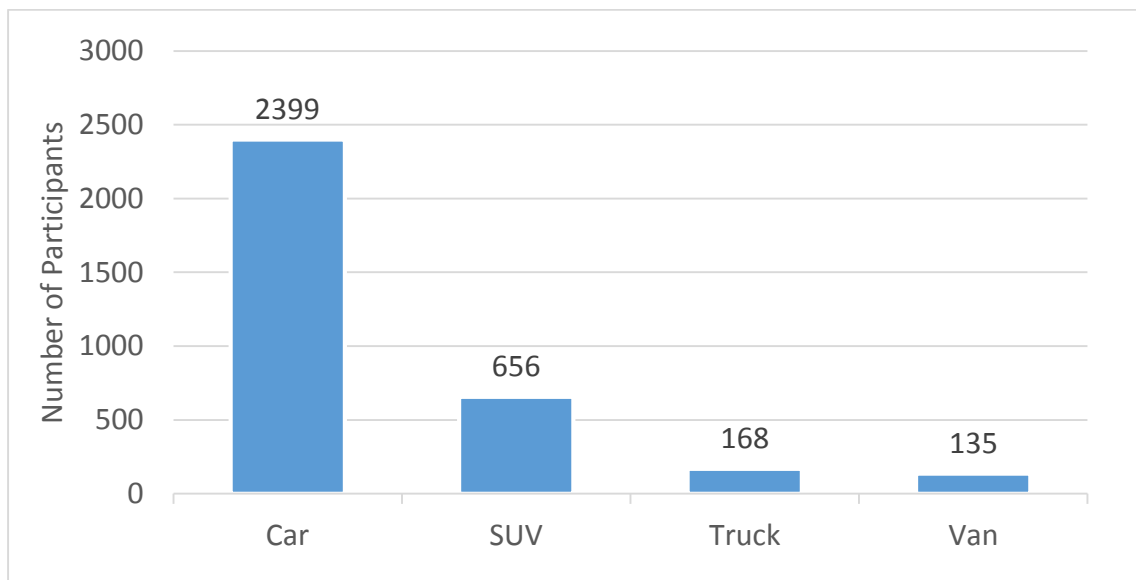
Driver Demographic Questionnaire	Driving history Questionnaire
Medical Conditions & Medications	Barkley's ADHD Screening Test
Driving Knowledge Survey	Risk Perception Questionnaire
Risk Taking Questionnaire	Sensation Seeking Scale Survey
Physical Strength Tests	Driver Behavior Questionnaire
Visual and Cognitive Tests	Sleep Habits Questionnaire
Medical Conditions and medications	Driver Exit Interview
Conner's Continuous Performance Tests	Clock Drawing Assessment

In order to better understand the drivers' driving conditions both physically and psychologically, the drivers were asked to fill out several questionnaires and conduct physical tests before entering the SHRP2 study. Those information included the driver demographics, vision tests, cognitive assessment, physical ability metrics, vehicle characteristics, and post-crash survey. The basic driver demographics included gender, driving history, employment status, household income, etc. Those information helped us identify the risky or vulnerable driver groups prone to crashes. A list of the participant questionnaires and tests are listed in Table 3.2. The detailed explanations for the assessment questionnaires and cognitive assessments are listed in Table A.2 and Table A.3 (Appendix A). However, none of those driver survey information was used in this dissertation.

### 3.2.5 Vehicle data

The SHRP2 vehicle fleet was recruited to represent the national light vehicle demographics and vehicle makes, including passenger car, sports utility vehicle, truck, and van. Those vehicles were owned by the participants for their daily driving purpose. The distribution

of vehicle types in the SHRP2 NDS is shown in Figure 3.12. The percentage of vehicles for passenger car, SUV, truck, and van were 71.4%, 19.5%, 5% and 4% respectively. The passenger cars are the predominant type of vehicle in the SHRP2 NDS, which followed by SUV, truck, and van. However, the SHRP2 NDS oversampled the newer cars between 2006 to 2011 years because some of the older cars do not provide vehicle network information. Therefore, the sample was slightly biased towards the newer generation of vehicles. A list of SHRP2 NDS vehicle by vehicle classes is shown in Table A.4 (Appendix A).



**Figure 3.12 Vehicle types in the SHRP2 NDS (SHRP2 InSight, 2014)**

### 3.2.6 Time series DAS data

Time series DAS data was collected from the GPS, video cameras, vehicle network, and vehicle sensors. A list of the most relevant time series variables is shown in Table 3.3. Speed is one of the most important variables in the time series dataset. The GPS location information is critical to link the driving data to RID data. The accelerometer, gas pedal, and brake pedal are

important indicators for drivers' speed control behavior and evasive maneuver in crashes and near-crashes. The lane marking distance can be used to study drivers' lane keeping behavior. The lane marking probability variable indicates the confidence of lane position values from in-field measurement. Forward radar also provides critical information to study drivers' car following behavior. Overall, it is a very comprehensive list of variables that could be potentially collected on roads.

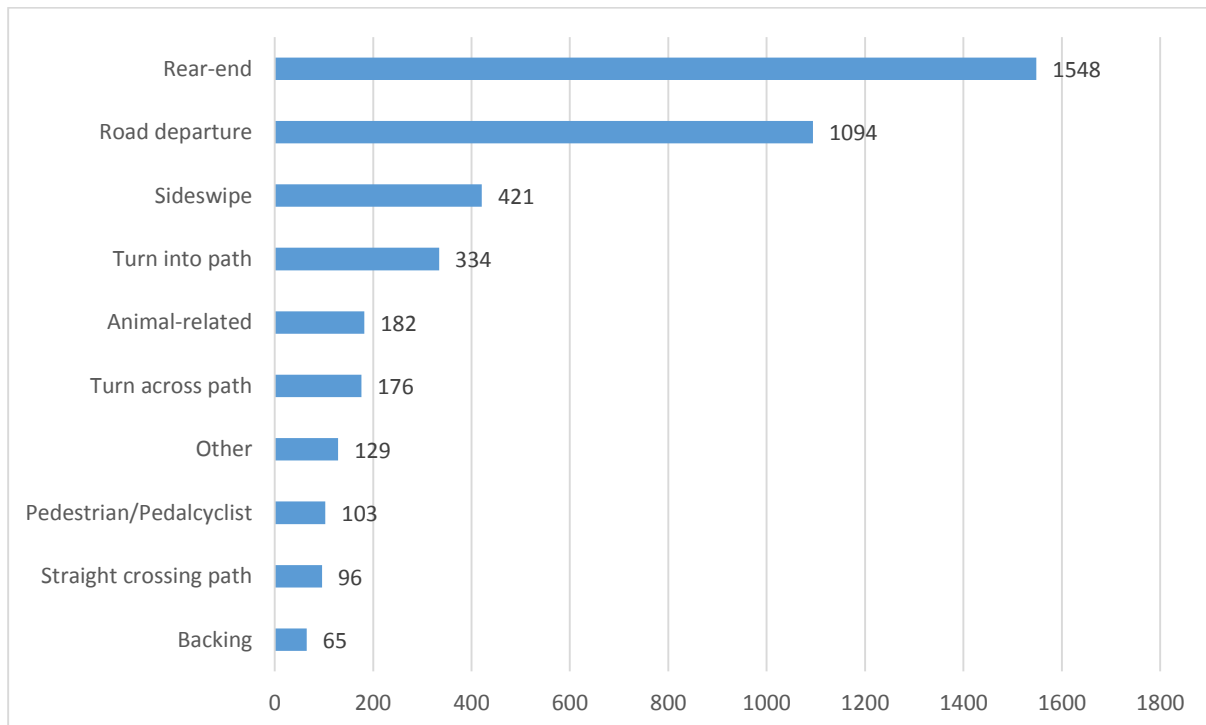
**Table 3.3 Description of selected time series data from the SHRP2 NDS (Dingus, 2013)**

<b>Variable</b>	<b>Description</b>	<b>Units</b>	<b>Rate</b>	<b>Accuracy</b>
Speed	Vehicle speed from GPS	meters/second	1Hz	+/-1km/h
Heading, GPS	Compass heading of vehicle from GPS	degrees	1Hz	+/-1.5deg
Acceleration	Vehicle acceleration in the longitudinal direction versus time.	meters <sup>2</sup> /second	10Hz	NA
Pedal & Accelerator Position	Position of the accelerator pedal collected from the vehicle network and normalized using manufacturer specs.	percentage	Varies	NA
Pedal, Brake	On or off press of brake pedal.	1/0	2 Hz	NA
Steering Wheel Position	Angular position and direction of the steering wheel from neutral position	degrees	10 Hz	NA
Lane Marking, Distance	Distance from vehicle centerline to inside of left side lane marker based on vehicle based machine vision.	cm	10 Hz	NA
Lane Marking, Probability	Probability that vehicle based machine vision lane marking evaluation is providing correct data for the right side lane markings.	1	10 Hz	NA
Radar, Range Rate	Range rate to forward radar targets measured longitudinally from radar.	meters/second	10 Hz	NA



### 3.2.7 Crashes and near-crashes data

A number of crash and near-crash events were identified from participant reports, crash notification algorithm, and the video reduction from forward view video. As of September 18th, 2015, a total of 1,465 crashes, 2,682 near-crashes, and 31,925 baseline events were presented on the SHRP2 InSight website. Crashes were further categorized into four severity levels: 1) Most severe; 2) Police-reportable crashes; 3) Minor Crashes; 4) Low-risk tire strike. The number of crashes and near-crashes for each category is illustrated in Figure 3.13. Near-crashes was defined as an event requires a rapid evasive maneuver. This is the major data source for the crashes and near-crashes analysis in Chapter 4. Additional events are likely to be added to the InSight website, but they won't be included in this dissertation.



**Figure 3.13 Crash and near crashes by incident types (SHRP2 InSight, July 2015)**

### 3.2.8 Institutional review board (IRB) and other issues

Since the SHRP2 data contains human subject information, it is important to protect participants' privacy and obtain the IRB approval from the home institution to properly use the SHRP2 NDS data. Protecting drivers' privacy issue was one of the highest priority in this study. The participant protection policy strictly followed the Code of Federal Regulations, Title 45 Public Welfare, Department of Health and Human Services, Part 46, Protection of Human Subjects (45 CFR 46). The analysis of SHRP2 data in this dissertation strictly followed the approved data analysis plan and data usage agreement submitted to IRB and VTTI.

### 3.2.9 The challenges of analyzing the SHRP2 NDS data

The dataset used in this dissertation was requested in summer 2013. By the time of receiving the data, only one third of the SHRP2 NDS data was available. The radar measurements as well as many other variables had not gone through data quality assurance process. Hence, it is important to understand the quality of the data before it is analyzed in this dissertation.

Even though naturalistic driving study has many advantages over traditional methods, it also had some shortcomings. First of all, it is difficult to draw casual-effect relationship from the complex driving environments. For example, the small change in vehicle speed could be caused by a variety of reasons in the complex and constantly changing driving environments. Drawing causal-effect relationship from the complex driving environments is a challenging task using NDS data.

Second, the data quality should be carefully controlled before conducting any statistical analysis. Some common data quality issues include missing data and large noise from sensor data. Missing data is a common issue for many variables, such as GPS locations, steering wheel position, and vehicle lateral positions. Some of the missing data were caused by equipment malfunction, but some were caused by the absence of detected objects. Outlier values were often observed in vehicle sensor data, especially for the front radar and lane position variables. Dingus et al. (2014) assessed the quality of the collected data in the SHRP2 NDS Study and the results are listed in Table A.5 (Appendix A).

Third, although one advantage of naturalistic driving study is the large size of the collected data, it could also bring challenges for data storage, data management, and statistical analysis method. For example, the SHRP2 NDS project collected approximately 4 million GB's data. Even analyzing a subset of the SHRP2 dataset could be a challenge and require certain level of database knowledge and programming skills.

Fourth, choosing appropriate statistical techniques for analyzing large-scale naturalistic driving data is a challenge. One way to address this problem is to conduct the multivariate analysis at summarized event level. However, most of naturalistic driving data are collected as time series data by its nature. Analyzing time series data from naturalistic driving study could be a challenging task, which will be addressed in Chapter 6 of this dissertation.

Last but not least, reducing the naturalistic driving data could be a time consuming process. For example, the chapter 5 of this dissertation contained 5 million rows of time series data with more 83 columns (variables) collected. Manually reducing the raw data would be time

consuming and even impractical sometimes. It is recommended to write programs to batch process the raw data in an efficient and consistent manner.

Overall, it is important to understand the quality of the collected SHRP2 NDS data. The missing data and outliers should be carefully processed before conducting any statistical analysis. Otherwise, it may come to the biased conclusion. The researchers should be careful about the data quality issue in future research.

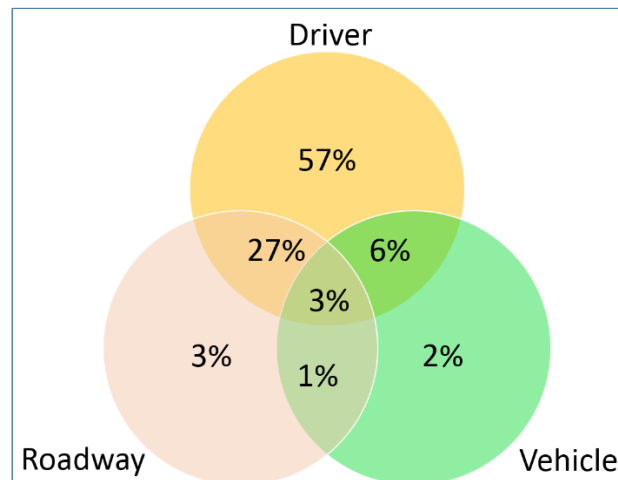
## CHAPTER 4. CRASHES AND NEAR-CRASHES ANALYSIS ON RURAL TWO-LANE CURVES

This chapter focuses on the analysis of crashes and near-crashes on rural two-lane curves using the SHRP2 NDS data. Section 4.1 introduces the background of the study. Section 4.2 discusses the data collection and data reduction process. Section 4.3 describes the initial screen of the SHRP2 NDS curve dataset. Section 4.4 focuses on the roadway departure crashes on rural two-lane curves. The logistic regression analysis of crashes and near-crashes is shown in Section 4.5. The discussion of the model results and summary of the findings are presented in Section 4.6 and Section 4.7.

### 4.1 Introduction

Several studies reviewed the causes of crashes and found human factor was the major contributing factor to approximately 90 percent of all crashes (NHTSA, 2008; Hendericks et al., 2001; Rumar et al., 1985; Salmon et al., 2005; Spainhour et al., 2005; Treat et al., 1979). In one of the earliest human factor study in transportation safety, Treat et al. (1979) reviewed 2,258 crash reports and found human factors contributed to 92.6% of the crashes; Environmental factors contributed to 33.8% of the crashes; Vehicle failure accounted for 12.6% of the crashes. In a similar study, Rumar et al.' study (1985) examined the causes of crashes and plotted the Veen diagram as shown in Figure 4.1. They found driver errors contributed to 93% of the crashes. Roadway factor accounted for 34% of the crashes. Vehicle factor contributed to 12% of crashes. It is possible that two or more factors existed in the same crash. In a more recent study, NHTSA (2008) again confirmed that human error was a major contributing factor for about 93%

of all crashes. Therefore, many previous studies revealed the important role of human factor in crashes, but it is also the least understood factor in the causes of crashes.



**Figure 4.1 Venn diagram of crash causes by percentage (Rumar, 1985)**

The traditional transportation safety analysis relied on police-reported crash data, but crash data is known to have a number of limitations. First of all, underreporting is a significant issue in crash data set. For example, Iowa only records any crashes causing death, personal injury, or total damages over \$1500. Many low-cost crashes were not recorded in the crash dataset. Second, human factor variables were poorly recorded in crash data. The drivers who involved in crashes might deliberately hide the true causes of crashes, or just simply did not remember what happened before the crashes. Hence, it is difficult to acquire accurate and reliable driver behavior information in crash report. Third, the near-crashes and unsafe driving behaviors were not recorded in crash dataset. Near-crash refers to the safety critical event that require abrupt evasive maneuver to avoid a crash (VTTI, 2015). Near-crash could provide important driver maneuver information for successful crash avoidance, but they were not recorded in crash data. Even though crash data is admittedly one of the most important data

sources for transportation safety research, it does not provide reliable and accurate driver behavior information.

In order to better understand the role of driver behavior in crashes, Strategic Highway Research Program 2 Naturalistic Driving Study (SHRP2 NDS) recorded more than 3,000 drivers' daily driving behavior for over two years in the U.S. Approximately 4 million gigabytes data were collected in this project. As the date of September 18<sup>th</sup>, 2015, a total of 1,465 crashes and 2,682 near-crashes, and 31,925 baseline events were recorded in the SHRP2 InSight website. The SHRP2 NDS data will greatly enhance our knowledge for the role of human factor in crashes.

## 4.2 Data Collection

This section discusses the data collection and data reduction process. The main data source of the crashes, near-crashes, and baseline events was the SHRP2 InSight website (<https://insight.shrp2nds.us/home>). This website served as the portal for researchers to review the collected SHRP2 NDS data. The variables in the SHRP2 NDS can be classified into four groups: vehicles, drivers, trips, and events. Vehicle information included vehicle type, make, age, and advanced technologies. Driver information included driver age, gender, traffic violation history, and driver behavior surveys. Trip information summarized the trips generated in SHRP2 NDS by average trip length, mean speed, acceleration, etc. Lastly, the event detail table summarized the SHRP2 NDS data at event level. The events included crash, near-crash, and baseline events. Additionally, forward view video for each event was also available on the SHRP2 InSight website. However, only qualified researchers, who obtained Institution Review Board certificate, have access to the event videos.

#### 4.2.1 Definition of crash, near-crash, and baseline events

The events were categorized into three types in this analysis, including crash, near-crash, and baseline events. The event types were manually reviewed and analyzed by data reductionist at VTTI. A total number of 1,465 crashes, 2,682 near-crashes, and 31,925 baseline events were presented on the SHRP2 InSight website as the date of September 18<sup>th</sup>, 2015. Furthermore, the crashes and near-crashes were treated as the same type of event in this study as safety critical events. The definitions for crash, near-crash, safety critical event, and baseline events were adopted from the SHRP2 InSight website and described in the following paragraphs:

**Crash:** *Any contacts that the subject vehicle has with an object, either moving or fixed, at any speed in which kinetic energy is measurably transferred or dissipated. Also includes non-premeditated departures of the roadway where at least one tire leaves the paved or intended travel surface of the road.*

**Near-Crash:** *Any circumstance that requires a rapid evasive maneuver by the subject vehicle or any other vehicle, pedestrian, cyclist, or animal to avoid a crash. Near Crashes must meet the following four criteria: 1) Not a Crash; 2) Not pre-meditated; 3) Evasion required; 4) Rapidity required.*

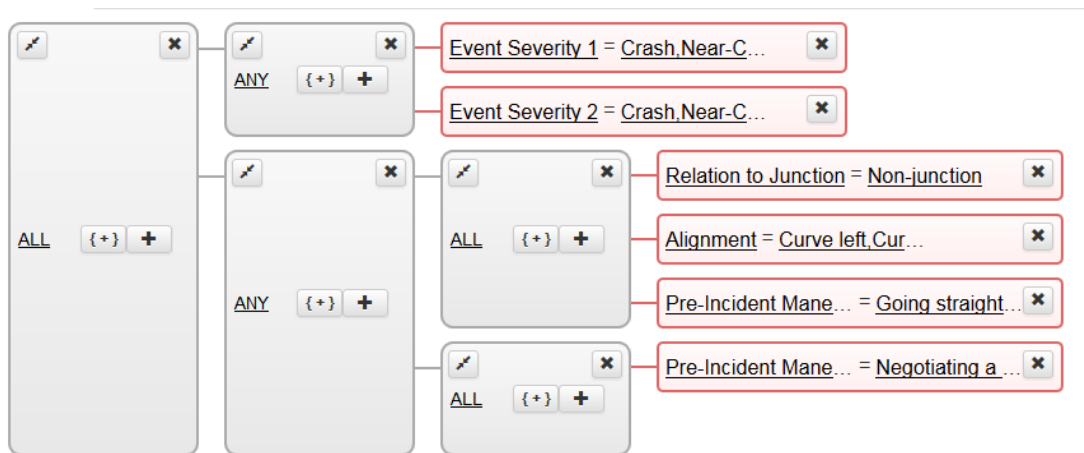
**Safety Critical Event:** *The combination of all crashes and near-crashes was referred as safety critical event in this study.*

**Baseline:** *An epoch of data selected for comparison to any of the conflict types listed above rather than due to the presence of conflict. For SHRP2, these baselines are 21 seconds long and were randomly selected with a goal of 20,000 baselines, a minimum of 1 baseline per driver (SHRP2 InSight Website, 2015).*



### 4.2.2 Query of curve-related crashes

The next step was to query all the curve-related crashes and baseline events from the SHRP2 InSight website. In this study, the curve-related crash was defined as any crashes occurred on curve alignment, or the pre-incident maneuver was negotiating a curve. The query of curve-related crashes from the InSight website is shown in Figure 4.2. As a result, 386 curve-related crashes (176 crashes and 210 near-crashes) were filtered from the InSight website. It should be noted that this step queried all crashes on all types of curves. The roadway departure crashes on rural two-lane curves are further filtered in Section 4.4.



**Figure 4.2 Filter criteria for curve-related crashes and near-crashes in the SHRP2 NDS**

In addition to crash and near-crash data, the baseline event data was also randomly sampled from the event table using similar criteria as shown in Figure 4.2, except that the event severity is baseline event. The goal of incorporating balanced-sample baseline was to represent the participants' exposure to different driving conditions. A total of 2,729 curve-related baseline events were found in the balanced-sample baseline events dataset. Similarly, any baseline events near intersections were excluded from the study. Only the events on rural-two lane curves with

paved surface were kept in the analysis. Finally, 386 curve-related crashes and 2,729 balanced-sample baseline events on curves were included in the initial analysis.

#### 4.2.3 Variables from SHRP2 NDS event table

The event table data was the main data source used in the crash and near-crash analysis. The majority of the event table data was reduced from forward view videos and driver face videos. The additional time series data, such as forward radar, accelerometer, gyro, and GPS, were also available to help understand the driver maneuver during the events. The variable definitions used in the SHRP2 NDS were based on the General Estimate System (GES) dataset developed by NHTSA, but modified for the naturalistic driving study data (Dingus et al, 2015).

Compared to the traditional crash data analysis, the most valuable information from NDS was drivers' in-vehicle driving behavior before the crashes and near-crashes. The video camera and vehicle sensors data provided critical information regarding drivers' distracted behavior, secondary tasks, reaction time, and impairments conditions. The information was not available in traditional crash data, but was collected in the SHRP2 NDS data. The precipitating event of the baseline events was defined as the event occurred at one second before the baseline event. This event table contains 75 variables about the event-related information, although some variables are not applicable to the baseline events. The detailed explanation for each variable is listed in the report *SHRP2 Research Dictionary for Video Reduction Data Version 3.4* (February 2015). A complete list of the variables in the event table is shown in Table 4.1. The following paragraphs discuss several important variables provided in the event table.

**Table 4.1 List of variables from the event detail table**

eventID	Roadway Departure	Curve-related
anonymousParticipantID	eventSeverity1	eventSeverity2
eventStart	subjectReactionStart	impactProximity
eventEnd	preIncidentManeuver	maneuverJudgment
precipitatingEvent	vehicle1SubjectConfig	vehicle2Config
vehicle3Config	eventNature1	incidentType1
crashSeverity1	vehicle1EvasiveManeuver1	vehicle1PostManeuver1
eventNature2	incidentType2	crashSeverity2
vehicle1EvasiveManeuver2	vehicle1PostManeuver2	airbagDeployment
vehicleRollover	driverBehavior1	driverBehavior2
driverBehavior3	driverImpairments	frontSeatPassengers
rearSeatPassengers	secondaryTask1	secondaryTask1StartTime
secondaryTask1EndTime	secondaryTask1Outcome	secondaryTask2
secondaryTask2StartTime	secondaryTask2EndTime	secondaryTask2Outcome
secondaryTask3	secondaryTask3StartTime	secondaryTask3EndTime
secondaryTask3Outcome	handsOnTheWheel	driverSeatbeltUse
vehicleContributingFactors	infrastructure	visualObstructions
lighting	weather	surfaceCondition
trafficFlow	trafficDensity	trafficControl
relationToJunction	intersectionInfluence	alignment
grade	locality	constructionZone
numberOfOtherMotorists	numberOfObjectsAnimals	fault
motorist2Location	motorist2Type	motorist2Maneuver
motorist2Reaction	motorist3Location	motorist3Type
motorist3Maneuver	motorist3Reaction	finalNarrative

The first group of variables described the nature of the events, including event severity, pre-incident maneuver, precipitating event, event nature, incident type, and crash severity.

Precipitating event start time and end time, and driver reaction time were also coded in the event

table. Drivers' evasive maneuver and crash outcomes were recorded in the table. The second group of variables described drivers' pre-crash behaviors, such as driving behavior, secondary tasks, and the duration of each task. The driver behavior included any pre-crash information, such as speeding and distracted driving. The secondary tasks specifies driver's distraction behavior, such as calling on a phone and reaching for objects inside vehicle. Some other driver behavior variables included hands on the steering wheel and wearing seatbelt. The third group of variables described the traffic and roadway environments, such as light, weather, surface condition, roadway alignment, traffic flow level, traffic density, etc. These variables helped us understand the context of the events. Lastly, a detailed narrative description of the event was also included in the table. They were written by the data reductionist at VTTI to describe the crashes and near-crashes. The information were critical for researchers to understand the context of the events. In addition to the variables that were listed in the event table, several other variables were also added into the event table for additional analysis. For example, posted speed limit and advisory speed limit were collected from Google Earth and added to the event table dataset.

#### 4.2.4 Variables from forward video

A large amount of useful information were manually coded from the forward view video. The data coded from the video includes traffic condition, roadway characteristics, and other environmental factors. The data were manually coded from the forward video from each event, which could be a very time consuming process. A list of variables collected from the forward videos is shown in Table 4.2.

**Table 4.2 Description of the variables from forward view video**

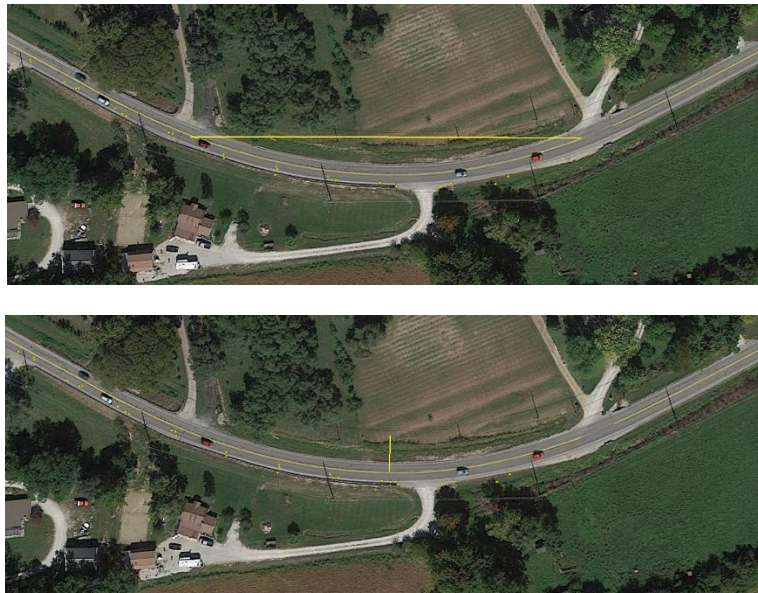
<b>Variables</b>	<b>Descriptions</b>
Pavement Condition	Pavement damage conditions. 0: normal surface condition; 1: moderate damage; 2: severe damage, presence of potholes
Shoulder Type	Types of shoulders. 0: Not Paved; 1: Curb; 2: Paved Shoulder; 3: On-street Parking
Guardrail	Guardrail presence. 0: No; 1: Yes
RPM	Raised pavement marking presence. 0: No; 1: Yes
Rumble Strips	Rumble strips presence. 0: No; 1: Centerline; 2: Edge line; 3: Both
Delineation	The visibility of lane delineation. 0: Visible; 1: invisible
Grade	Roadway vertical grade. 0: Flat; 1: Uphill; 2: Downhill
Chevrons	Chevron presence. 0: No; 1: Yes
Car-following	Presence of the following car in front. 0: No Car in front; 1: Front Car Presence (>100 ft); 2: Closely Followed (<100 ft)
Car-oncoming	Presence of on-coming car in the opposite lane. 0: No; 1: Yes

It should be noticed that the vehicle speed information was accessed from the time series data of these events. The concept of safety critical speed was introduced here. In the safety critical events, the safety critical speed was defined as the time at which the safety critical events occurred. For the baseline events, the critical safety speed was defined as the highest speed inside the curves. The speeds might be measured at different locations on the curves, but they all represented the safety critical speed in the events. The vehicle speed was defined as when the safety critical speed is at least 5 mph higher than posted or advisory speed limits. The definition is used consistently in this chapter.

#### 4.2.5 Variables from Google Earth

The curve radius variable was not available in the requested event table, but it is important to incorporate curve radius into the analysis. Curve radius needs to be measured from

Google Earth using the chord-offset method as shown in Figure 4.3. It is a method to calculate curve radius based on chord measurement and middle offset measurement method. The curve radius in this study was determined from the formula  $R = \frac{L^2}{8m} + \frac{m}{2}$ , where R is the radius in feet; L is the chord length in feet; and m is the middle offset in feet. A single curve was measured three times and the averaged value was used as curve radius.



**Figure 4.3 Curve radius measurement from chord length and offset distance (Google Map, 2015)**

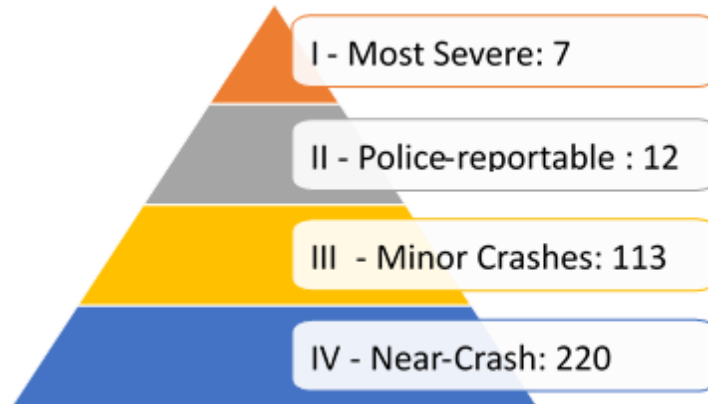
The variables collected in the SHRP2 NDS project is unique and different from the variables collected from traditional crash dataset. Drivers' in-vehicle driving behavior was coded from vehicle forward videos. Vehicle dynamics data was also available from vehicle sensors. This is a very detailed crash and near-crash dataset that provided many driver behavior variables before, during, and after the events. It will result in many new findings about the role of driver behavior in crashes and near-crashes on curves.

### 4.3 Initial Screen of All crashes on All Curves

The purpose of the exploratory analysis in this section is to summarize the collected SHRP2 NDS curve dataset. This initial curve dataset contained 132 crashes and 220 near-crashes on all types of curves. Additional 2,373 baseline events were also included in the analysis as the control group. It should be noted that the events this section included all crashes on all types of curves from the SHRP2 NDS project. In general, 45% percent of the all curve-related crashes occurred on left turn curves, while the other 55% of crashes occurred on right turn curves. The majority (84%) of crashes and near-crashes events were happened in rural area. Almost half of the crashes (46%) occurred under free flow conditions. It was found 93% of the drivers in safety critical events worn safety belt properly. Three quarters of the crash events occurred in day time. In order to better understand the requested data, the events were further summarized in the following paragraphs.

The crashes and near-crashes on all curves were further classified into four categories as shown in Figure 4.4. There are four levels of severities, including the most severe crashes (level 1), police-reportable crashes (level 2), minor crashes (level 3) and near-crashes (level 4). It was found the most severe type of crashes had the smallest number of events. The near-crash had the largest number of events. This phenomenon was referred to as Heinrich's law in the book *Industrial Accident Prevention, A scientific Approach* (Heinrich, 1931). Heinrich's law found the event frequency decreases as event severity increases. In this curve dataset, for every 7 severe crashes, there were 12 police-reportable crashes, 113 minor crashes, and 220 near-crashes as shown in Figure 4.4. In this dissertation, both crashes and near-crashes are referred as one event type as safety critical event. Guo et al. (2010) evaluated the causes of crashes and near-crashes

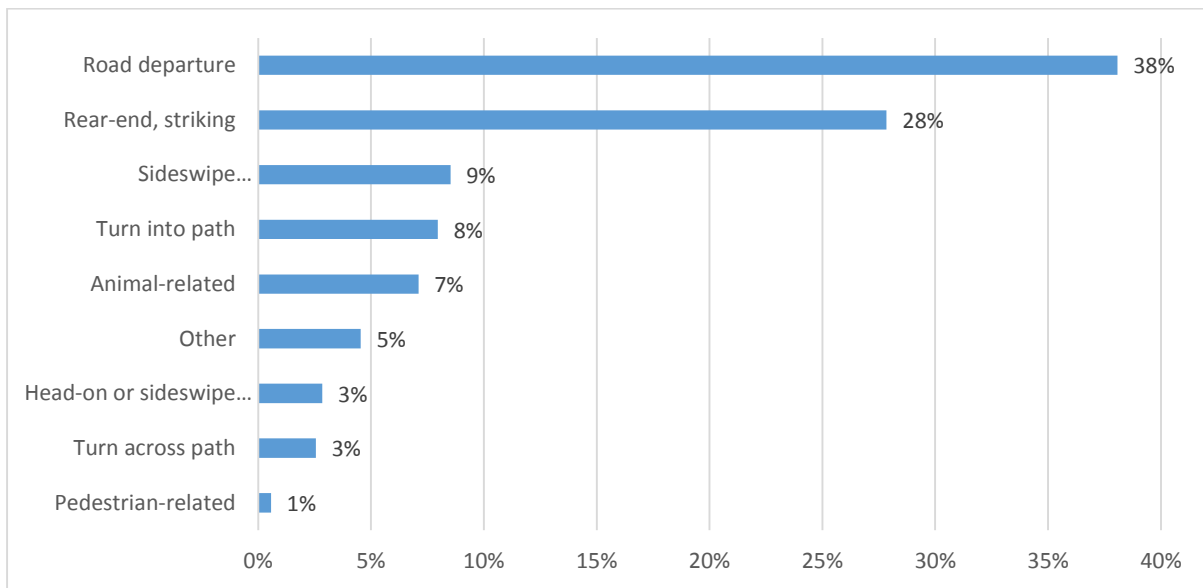
and concluded that the near-crashes could be used as crash surrogate for crashes in naturalistic driving study. This evidence supports the combination of crashes and near-crashes as one event category.



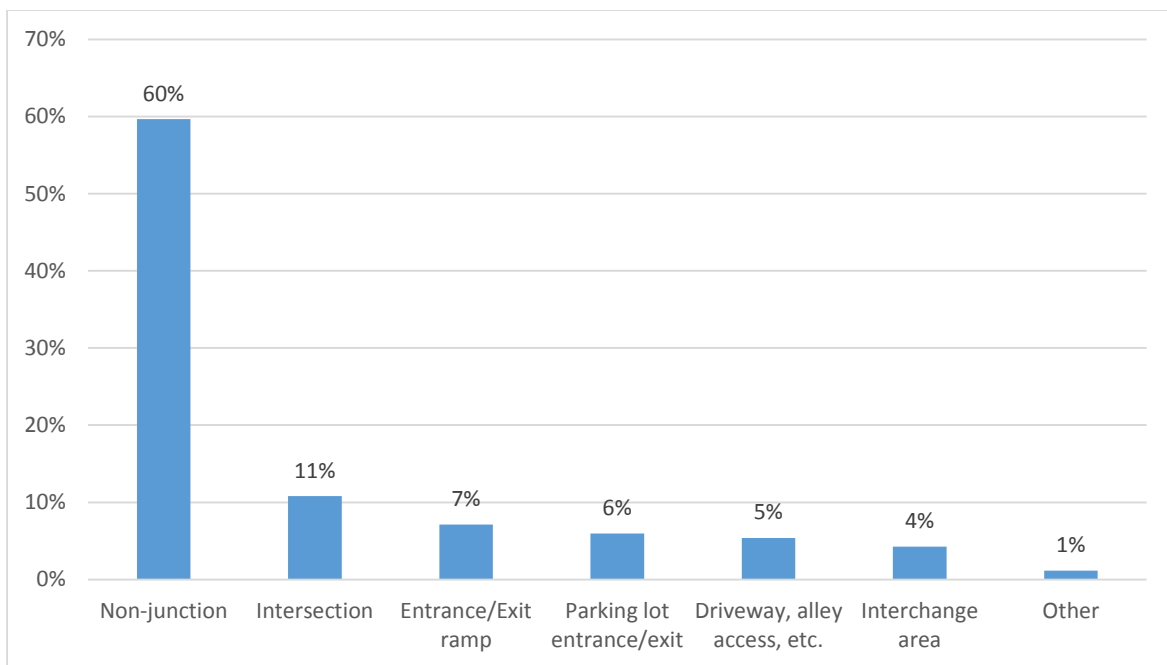
**Figure 4.4 Crashes and near-crashes by severities**

Although all 352 curve-related crashes were physically occurred on curves, the crashes were caused by a variety of reasons. The incident types of the curve-related crashes are shown in Figure 4.5. The most frequent incident type was roadway departure event. The rear-ended events were the second frequent type of events on curves. However, the rear-ended crashes were probably caused by traffic conditions rather than the curve geometry, so these events might not be included in the final analysis. Similarly, curve design also had little influence for animal-related crashes. Therefore, they were excluded from the final analysis. The pedestrian-related events were also irrelevant to the curve geometries, and they were excluded from the analysis.





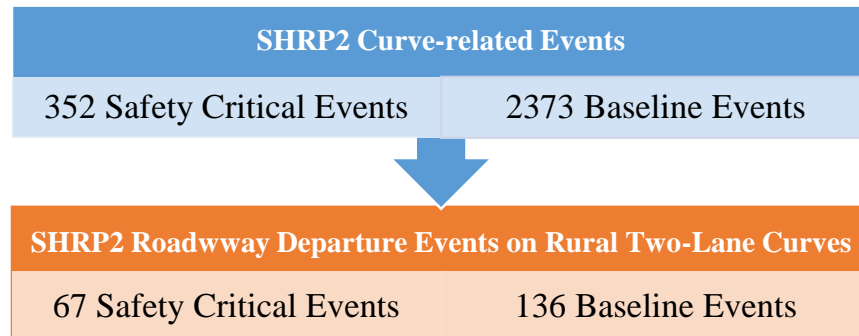
**Figure 4.5 Distribution of curve-related crashes and near-crashes by incident types**



**Figure 4.6 Distribution of curve-related crashes and near-crashes by junction types**

The curve-related crashes were further plotted by junction types as shown in Figure 4.6. Sixty percent of the curve-related crashes did not related to any junctions. However, the rest 40%

of curve-related events were related to certain types of junctions, such as intersection (11%), entrance or exit ramps (7%), parking lot (6%), and driveway (5%). The forward view video was manually checked to confirm the relation of the crashes to roadway junctions. Since the presence of junction could have significant impact on drivers' curve driving behavior, all the junction-related crashes were excluded from the final analysis.



**Figure 4.7 Subset the SHRP2 curve-related events**

The requested dataset found some types of crashes had little relationship with the curve geometry. For example, rear-ended crashes and animal-related crashes had minimum relationship with curve geometry. Those types of crashes should be treated separately. Hence, it was determined that only the roadway departure, heads-on collision, and sideswipe events were included in the final analysis. Additionally, the events near intersections, roundabout, construction zone and parking lots were excluded from the analysis. Only the roadway departure crashes on rural two-lane (one lane in each direction) curves were kept in the analysis. The forward view videos were carefully reviewed to make sure the selected events followed the definition of roadway departure crashes on rural two-lane roads in this study. The reduced rural two-lane curve event dataset is illustrated in Figure 4.7. The dataset contains 67 safety critical events on rural two-lane curves. For baseline events, the goal was to sample twice the number of

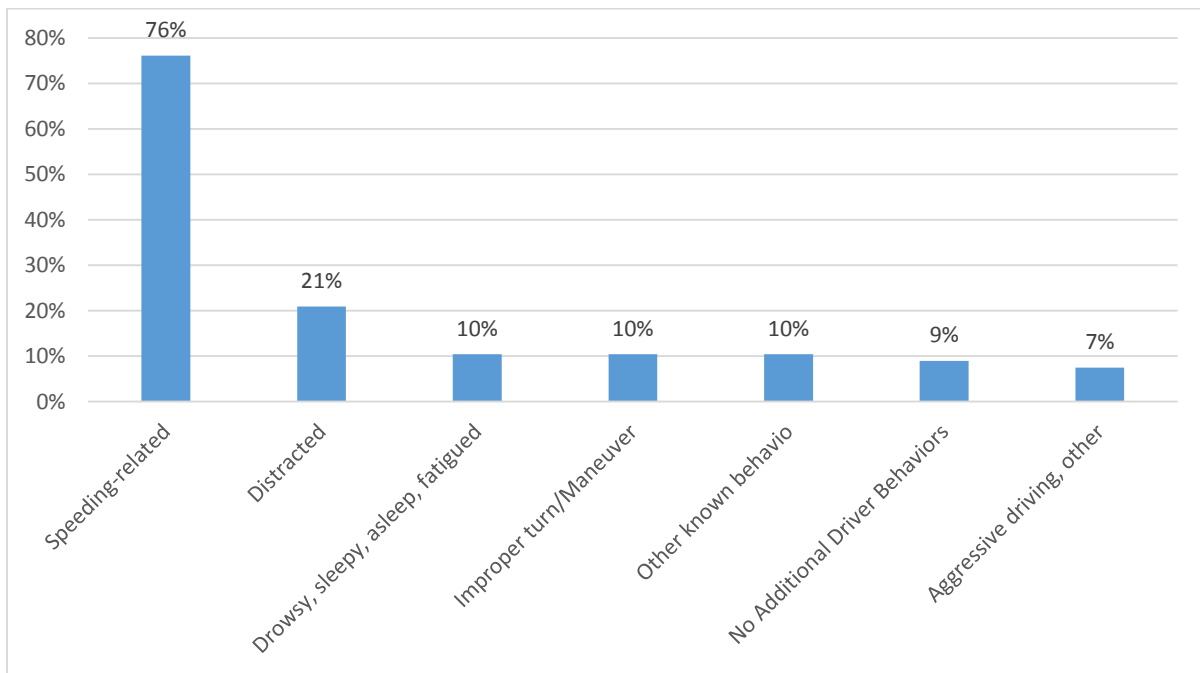
baseline events than the safety critical event, otherwise, the small ratio of safety critical events and baseline events could create instability issue for the logistic regression model. Hence, 136 baseline events were sampled from the 2,373 balanced-sample baseline events.

#### 4.4 Description of Roadway Departure Crashes on Rural Two-Lane Curves

This section described the reduced dataset for roadway departure crashes on rural two-lane curves. After the forward videos were manually reviewed, it was found vehicle speeding, drivers' engagement in secondary tasks, and slippery roadway surface were the major contributing factors to roadway departure crashes on rural two-lane curves.

##### 4.4.1 Speeding

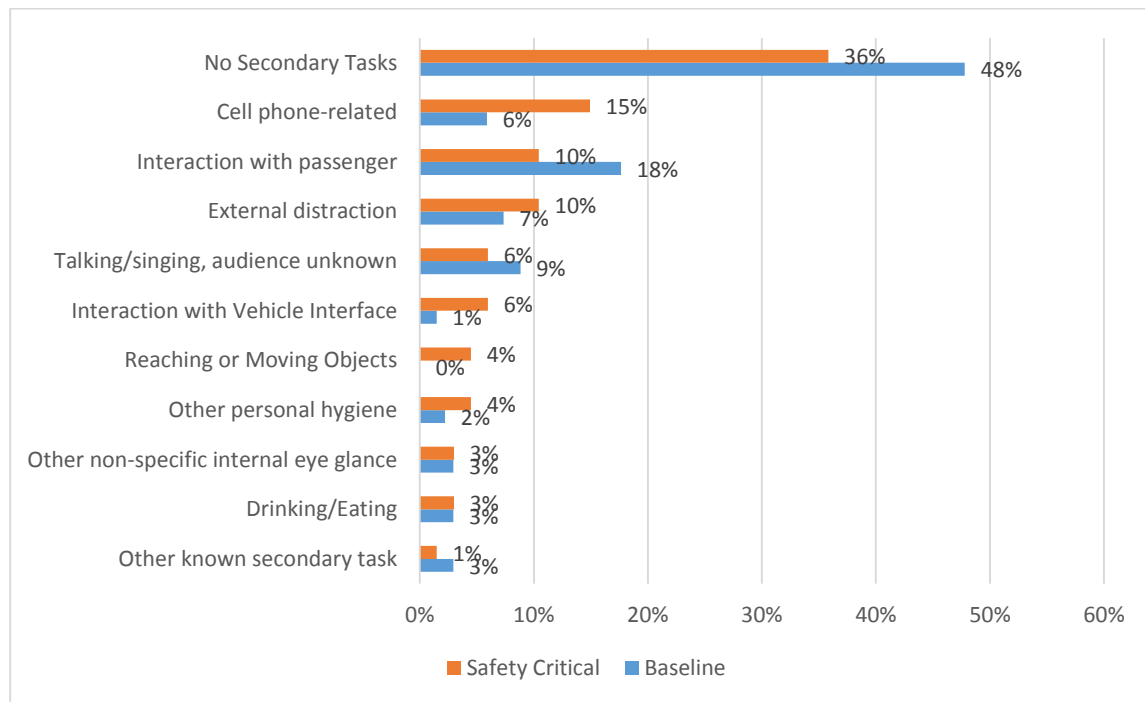
Speeding was identified as one of the major contributing factors to crashes on curves (Council, et al., 1988; Milton and Mannering, 1998; Suh, 2006; Khan et. al., 2013; Schneider, 2010; Torbic, 2004; Zegger, 1991). The percentage of curve-related crashes by driver behavior is illustrated in Figure 4.8. The sum of the percentages was over 100% because multiple driver behaviors could occur in the same event. In summary, speeding was a contributing factor to 76% of the curve-related crashes and near-crashes. The speeding was defined as driving too fast for conditions or driving over posted speed limits (FHWA, 2015). Driver distraction was a contributing factor to 21% of the roadway departure events on rural two-lane curves. Other types of driver behavior included fatigue driving, improper turn, and aggressive driving. In summary, driver behavior was a contributing factor to approximately 91% of the safety critical events on rural two-lane curves. This is similar to the findings from previous human factor studies. It revealed the important role of driver behavior in safety critical events on rural two-lane curves.



**Figure 4.8 Distribution of driver behaviors for roadway departure events on rural two-lane curves**

#### 4.4.2 Secondary tasks

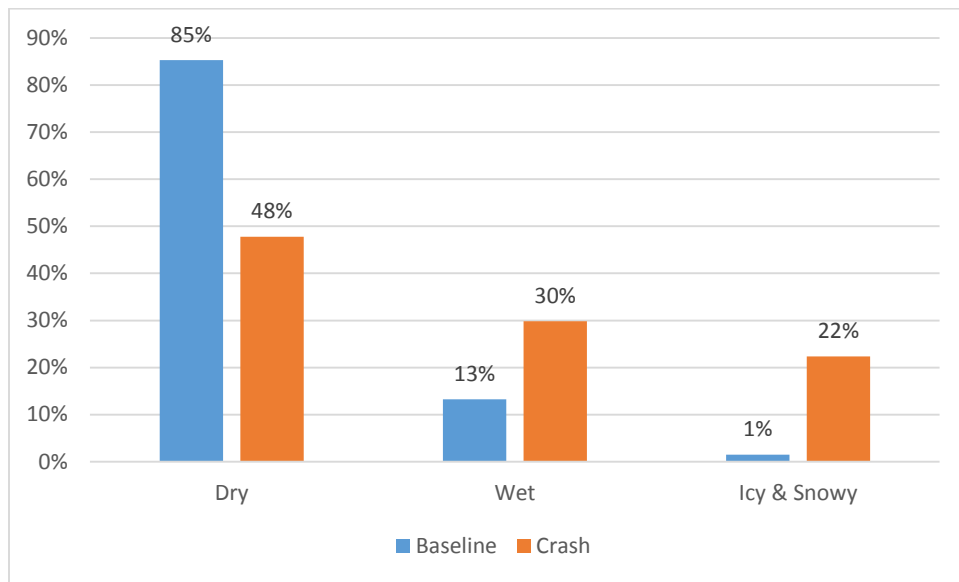
Another important variable reduced from event table was drivers' engagement in secondary tasks. The information was reduced from driver face video by data reductionists at VTTI and provided from the event detail table. Figure 4.9 indicated 64% of drivers engaged in secondary tasks in safety critical events. Cell phone-related events was the leading secondary behavior in safety critical events. Furthermore, external distraction and interacting with vehicle infotainment system had disproportionally higher percentage in safety critical events, which indicated higher risks associated with those events. This figure showed that drivers were very frequently engaged in secondary tasks whiling driving on rural two-lane curves.



**Figure 4.9 Distribution of secondary tasks for roadway departure events on rural two-lane roadways**

#### 4.4.3 Adverse surface conditions

Adverse surface condition was found to be an important contributing factor to the roadway departure events on rural two-lane curves. The percentages of surface conditions between baseline events and safety critical events are compared in Figure 4.10. Approximately 85% of baseline events were on dry surfaces, whereas only 48% of the safety critical events had dry surface. The safety critical events were disproportionately higher on wet surface and icy/snowy surface. Only 1% of baseline events had icy/snowy surface on rural two-lane curves, but it accounted for 22% of safety critical events. The summary statistics showed roadway surface friction is an important contributing factor to roadway departure events on rural two-lane curves.



**Figure 4.10 Distribution of surface conditions for roadway departure crashes on rural two-lane roadways**

Overall, this section summarized three important contributing factors and found speeding, secondary tasks, and adverse surface conditions were the major contributing factors to roadway departure events on rural two-lane curves. The next section examined multiple factors at the same time in the logistic regression model.

#### 4.5 Logistic Regression Analysis of Roadway Departure Crashes on Rural Two-Lane Curves

The objective of this study was to understand how driver behavior, roadway characteristics, and traffic environments affect the likelihood of roadway departure events on rural two-lane curves. As discussed before, the primary data source for roadway departure events was the SHRP2 InSight website. The other data sources included the variables coded from forward view video, Google Earth, driver demographics, and vehicle types. A total of 104 variables were collected for initial analysis. This study included 68 roadway departure safety

critical events and 136 baseline events on rural two-lane curves. This goal of this section is so predict the event outcomes as a function of explanatory variables using logistic regression model. The odds ratio was also developed for each contributing factor.

#### 4.5.1 Background

Due to the nature of the binary outcome (safety critical event vs. baseline event), logistic regression model was selected to predict event types based on explanatory variables, including driver behavior, roadway characteristics, and environmental factors. In an earlier study, Hallmark et al. (2011) evaluated several statistical methods to evaluate roadway departure events using naturalistic driving data. The candidate models included generalized linear model, Bayesian model, and regression tree model. The logistic regression model was found to be the most appropriate statistical method to analyze the event level of naturalistic driving study data. The model applies a logit transformation to the dependent variable  $Y$  and predict logit of  $Y$  from the explanatory variables  $X_i$ . The natural log transformation makes the relationship between  $\text{Logit}Y$  and  $X_i$  linear. Unlike linear regression model, logistic regression model does not assume normality, linearity, and homoscedasticity and it can handle many types of relationships between independent and dependent variables. However, the independence of observation is a key assumption. The model outputs can be interpreted by odds ratio, which indicates the probability of an event occurring relative to an event not occurring. The concept of odds ratio can be easily understood by transportation practitioners and stakeholders.

#### 4.5.2 Data description

This section included as list of variables that considered in the logistic regression model as shown in Table 4.3. Although the whole dataset contained over one hundred variables, only 23 most relevant variables were tested in the initial analysis.

**Table 4.3 A list of variables included in logistic regression model**

<b>Variables</b>	<b>Descriptions</b>
event ID	Unique Event ID
Participant ID	Unique Driver ID
Event Type	Event Types. 0: Baseline event; 1: Safety Critical Events.
Posted Speed Limit	Posted speed limits in the unit of 10 MPH.
Speeding Speed	Speed difference between in-curve speed and posted speed in MPH
Radius	Curve radius in feet
Pavement Condition	Pavement deterioration condition. 0:normal surface condition; 1:moderate damage; 2:severe damage, presence of potholes
Shoulder Type	Should Types on curves. 0:Not Paved; 1: Curb; 2: Paved Shoulder; 3: On-street Parking
Guardrail	Presence of guardrail. 0: No; 1: Yes
RPM	Presence of raised pavement marking. 0: No; 1: Yes
Rumble Strips	Presence of rumble strips. 0: No; 1: Centerline; 2: Edge line; 3: Both
Delineation	Visibility of lane delineation. 0: visible; 1: invisible
Grade	Roadway grade. 0:Flat; 1: Uphill; 2: Downhill
Chevrons	Presence of chevrons. 0: No; 1: Yes
Car-following	The subject was following another vehicle in front. 0: No Car in front; 1: Front Car Presence 2: Closely Followed
Car-oncoming	There was an oncoming car from opposite direction. 0: No; 1: Yes
Distraction Duration	The duration of distraction in seconds.
Secondary Task	Types of secondary tasks within 6 seconds before crashes.
Hands On The Wheel	Number of hands on the wheel. 0: Both Hands; 1: Left or right; 2: None
Driver Seatbelt Use	Seatbelt usage. 0: Not worn; 1: Properly worn
Lighting	Lighting types. 0: Daylight; 1: Dusk/Dawn; 2: Darkness
Surface Condition	Roadway surface conditions. 0: Dry; 1: Wet; 2: Icy/Snowy
Grade	Roadway vertical grade. 0: Flat; 1: Uphill; 2: Downhill
Driver Age	Driver age at the start of SHRP2 Study.
Driver Gender	Driver gender. 0: male; 1: Female
Vehicle Types	Vehicle types. 0: Passenger car; 1: Pickup; 2: SUV; 3: MiniVan/Van



### 4.5.3 Logistic regression model

Logistic regression was used to model the odds of roadway departure event (1 for safety critical events and 0 for baseline events). The logistic regression model for is shown as follows.

$$\log(Y) = \text{natural log(odds)} = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon_{ij}$$

$p_i$  is the probability of safety critical event occurring;

$Y$  is the odds of a safety critical event occurring;

$\beta_0$  is the intercept in the model;

$\beta_k$  is the regression coefficient;

$x_k$  is the categorical or continuous explanatory variables;

$\varepsilon_{ij}$  follows Bernoulli distribution.

Overall, the model coefficients were estimated by maximum likelihood method. The outcome variable was coded as binary values (0 and 1). The explanatory variables were coded as shown in Table 4.3. The value of coefficient  $\beta_k$  determined the positive or negative contribution to the probability of an event occurring. If the coefficient was positive, larger  $x_k$  was correlated to higher probability of roadway departure event. If the coefficient was negative, larger  $x_k$  was correlated to lower probability of roadway departure event. Several model diagnostics were used to check model goodness-of-fit and the statistical significance for each explanatory variable. The model was fitted with glm function in R software. The results were reported as odds ratio, which indicated the risks associated with each factor.

#### 4.5.4 Model results

The final logistic regression model predicted the likelihood of roadway departure events with 8 explanatory variables as shown in Table 4.4. The final logistic regression model is shown as follows.

Predicted logit of Roadway Departure Event

$$= -7.2178 + 0.9301 * (\text{Speeding Speed}) + 1.3378 * (\text{Wet Surface}) \\ + 3.5269 (\text{IcySnowySurface}) - 1.4473 * \log(\text{Radius}) + 1.5501 * (\text{Curb}) \\ - 1.3303 * (\text{PavedShoulder}) + 1.1213 * (\text{Visual Distraction})$$

**Table 4.4 Logistic regression model results**

Parameter	Estimate	SE	p-value	$e^{\beta}$
$\beta_0$ (intercept)	7.2178	1.8011	0.000	31745.24
$\beta_1$ (Speeding Speed)	0.9301	0.2639	0.000	2.53
$\beta_2$ (Wet Surface)	1.3378	0.5919	0.024	3.81
$\beta_3$ (Icy/Snowy Surface)	3.5269	1.0286	0.001	34.02
$\beta_4$ (Log Radius)	-1.4473	0.2939	0.000	0.24
$\beta_5$ (Curb)	1.5501	0.6905	0.025	4.71
$\beta_6$ (Paved Shoulder)	-1.3303	0.5268	0.012	0.26
$\beta_7$ (Visual Distraction)	1.1213	0.5868	0.056	3.07
Number of Observations	215			
Log Likelihood	-65.627			
Likelihood ratio test (p value)	142.91(0.000)			
Pseudo R-squared	0.625			
AIC	147.25			
Hosmer & Lemeshow (p value)	9.4898 (0.3027)			

The goodness-of-fit of the overall model was checked by the likelihood ratio chi-square test. The test compared the model with predictors to the model with intercept parameter only.

The p-value of the likelihood ratio test was statistically significant at 0.05 level, which indicates the overall model was significantly better than the null model with intercept term only. The Hosmer & Lemeshow test had p-value larger than 0.05, which indicated no evidence of poor fit. The Log likelihood, AIC, and R-square values were used in model selection. In order to check model prediction accuracy, a classification table was used to compare predicted outcomes to the observed outcomes. According to the classification table as shown in Table 4.5, the model successfully predicted 88% of the roadway departure events and 87% of the baseline events. It was concluded that the model fitted the data decently.

**Table 4.5 Classification table for the logistic regression model**

Observed Event Outcome	Predicted Event Outcome		% Correct
	Yes	No	
Yes	135	18	88%
No	8	54	87%
		Overall Correct %	87.6%

All eight independent variables were statistically significant predictors for roadway departure events with p-value less than 0.05. The intercept was statistically significant at 0.05 level, which means the intercept should be kept in the model to improve model fit. The driving speeds over posted speed limits were found to be statistically significant in the model. Both wet surface and icy/snowy surface contributed to higher likelihood of roadway departure crashes on curves. Smaller logarithm of radius increased the likelihood of roadway departure risks. Compared to no shoulder scenario, curb increased the likelihood of roadway departure crashes, and paved shoulder decreased the likelihood of roadway departure events. It should be noted that curb did not directly cause more roadway departure crashes, but the tire-strike with curb events

were more likely to get detected by extreme G-force from accelerometer and recorded in the event dataset. Finally, visual distraction events were also found to increase the likelihood of roadway departure crashes. The visual distraction was defined as visually distracted from monitoring roadway conditions. For example, texting on cellphone was defined as visual distraction, but talking on hands-free cell phone was not a visual distraction event.

The odds ratio and the 95% confidence interval are listed in Table 4.6. It is a measure of exposure and its associated risks. The odds ratio for speeding over posted speed limits was 2.54. In this case, the odds of getting into a roadway departure crashes on rural two-lane curves was 2.54 times higher for every 10 MPH above posted speed limits. The wet surface on curves also increased roadway departure risks by 3.81 times. The icy and snowy surface was more risky than wet surface. The odds ratio for icy/snowy surface was 34.08, which indicated the icy/snowy surface was a very significant contributing factor to roadway departure crashes on rural two-lane curves. In addition, the roadway departure risk was negatively correlated with curve radius. The larger logarithm of radius led to smaller chance of getting into a roadway departure events. The paved shoulder decreased the roadway departure crashes on curves by 0.26. However, the paved shoulder tend to have painted lane marking as well, which could help drivers' lane keeping in the curves. It is recommended to separate the effect of paved shoulder and painted lane marking in future research. The presence of visual distraction increased the likelihood of roadway departure crashes by 3.07. It is interesting to see curb increased the likelihood of roadway departure events by 4.71 times. However, curb probably did not cause the roadway departure events, but a departure event with curb could be more easily detected by accelerometer and saved in the event database. It is recommended to use lane encroachment as surrogate for roadway departure events in future study.

**Table 4.6 The odds ratio estimate with 95% confidence interval**

Parameter	Odds Ratio	5%	95%
$\beta_0$ (intercept)	1363.47	81.38	31745.24
$\beta_1$ (Speeding Speed)	2.54	1.679	4.031
$\beta_2$ (Wet Surface)	3.81	1.455	10.343
$\beta_3$ (Icy/Snowy Surface)	34.08	7.108	222.38
$\beta_4$ (Log Radius)	0.24	0.140	0.371
$\beta_5$ (Curb)	4.71	1.539	15.165
$\beta_6$ (Paved Shoulder)	0.26	0.108	0.618
$\beta_7$ (Visual Distraction)	3.07	1.166	8.128

Nevertheless, some variables had very large confidence interval, such as icy/snowy surface and the wet surface. The confident interval for icy/snowy surface was from 7.10 to 222.37. It was mainly caused by small number of icy/snowy conditions in safety critical events. It is recommended to add more safety critical events and baseline events in future study to reduce the width of the confidence intervals.

## 4.6 Discussion

The objective of this study was to understand the contributing factors to roadway departure crashes on rural two-lane curves using the SHRP2 Naturalistic Driving Study data. Previous studies found human factor contribute to 90% of crashes, but they were poorly studied in previous research. The analysis of naturalistic driving study data in this analysis revealed important information regarding the role of human factor in roadway departure events on rural two-lane curves. This study first collected all curve-related crashes, near-crashes, and baseline events from the SHRP2 InSight website. The event table included information regarding the types of crashes, driver distraction behavior, traffic environment, and weather conditions. The

forward view video was also available to understand the scenario before, during, and after the events. Speed limits and rumble strips information was retrieved from Google Earth. The driver demographics and vehicle types were also linked to each event.

A number of important findings was discovered in this research. The study first analyzed curve-related crashes by crash severity and found it followed the Heinrich's law. However, not all crashes on curves were related to the curve characteristics. Hence, a new dataset was created and only included the roadway departure events on rural two-lane curves. Any animal-related events or intersection-related events were excluded from the study. The summary statistics found driver behavior accounted for 91% of the safety critical events. Speeding and distraction were the top two reasons for the roadway departure events on curves. Furthermore, 64% of drivers were engaged in secondary tasks before the crashes and near-crashes. The wet surface and especially icy and snowy surface had disproportionally higher chance of getting involved in roadway departure events on curves.

The logistic regression model was then used to predict the event outcome (safety critical or baseline event) based on explanatory variables. A total of 23 variables were initially tested in the model and only 7 of them were found to be statistically significant. The speeding on curves, wet surface, icy and snowy surface, presence of curb, and visual distraction were positively correlated with higher likelihood of roadway departure crashes on rural two-lane curves. The logarithm of radius and paved shoulder contributed to lower likelihood of roadway departure crashes. Among the factors, icy and snowy surface had the highest odds ratio of 34.08 with confidence interval between 7.108 and 222.38. The overall model prediction accuracy is 87.6%, which indicates decent goodness-of-fit of the data.

Overall, this research demonstrated the analysis of SHRP2 NDS event data and draw safety implications from the crashes and near-crashes analysis. Logistic regression model was found to be a useful method to analyze the event level data from naturalistic driving study. The naturalistic driving study method had several advantages compared to previous research. First of all, it contained critical information about the drivers' in-vehicle distraction behavior before, during, and after the events. Second, SHRP2 NDS collected detailed time series data from vehicle sensors, which were not available in traditional crash data analysis. Third, the real-time traffic information was observed from the forward view video. The SHRP2 NDS allowed researchers to review how drivers interacted with other vehicles in traffic safety. Fourth, the detailed driver demographics and driving history information were also available for analysis. Lastly, SHRP2 NDS project also collected detailed information from Roadway Information Database (RID). The driving behavior data could be linked to roadway characteristics.

However, there were also several limitations in this study. The most significant limitation of the study was the limited number of crash and near-crash events collected in the SHRP2 NDS. The small sample size created difficulties for building the statistical inference from the model. Furthermore, it was very time-consuming to reduce roadway and driver behavior from forward view video and driver face video. The vehicle speed variables were defined as the difference between safety critical speed and posted/advisory speed limits. There are other ways to define vehicle speeding at fixed points, such as curve PC, but it is not discussed in this analysis. It is recommended to conduct sensitivity analysis using different speeding definitions. The sampling approach for the baseline events should be carefully interpreted. The sensitivity analysis could be conducted to test different sampling plan and see how they affected the odds ratios. Lastly, it is

recommended to use lane encroachment as crash surrogate, instead of relying on accelerometer data only.

## 4.7 Summary

This study successfully demonstrated the analysis of crashes and near-crashes using the SHRP2 NDS data. The newly developed naturalistic driving behavior dataset contained a number of invaluable information that was not available in traditional safety analysis. This study found speeding on curves, slippery surface, and visual distractions contributed to higher likelihood of roadway departure crashes on rural two-lane curves. Larger curve radius and paved shoulder decreased the likelihood of roadway departure crashes on rural two-lane curves.

The research findings had important implications for transportation agencies. First of all, it is recommended to remove the ice and snow on sharp curves promptly. Otherwise, the drivers should be advised to reduce their speeds significantly on sharp curves when wet or icy/snowy surfaces were presented. It is also important to make sure the drivers keep their focus on the roads and not visually distracted while negotiating a curve. The likelihood of roadway departure crashes on rural two-lane curves also exponentially increases as curve radius decreases. It indicates the speed management on the sharp curves with radius less than 1000 feet should be a priority for transportation agencies. It is suggested to design small curve radius as less frequent as possible. It is recommended to install paved shoulder on curves. Overall, this study found important findings about the causes of drivers' crashes and near-crashes on rural two-lane curves using the SHRP2 NDS data. The next chapter will analyze drivers' normal driving behavior on curves and understand how drivers interact with different curves in normal daily driving activities.



## CHAPTER 5. MULTIVARIATE ANALYSIS OF DRIVER BEHAVIOR ON RURAL TWO-LANE CURVES

### 5.1 Introduction

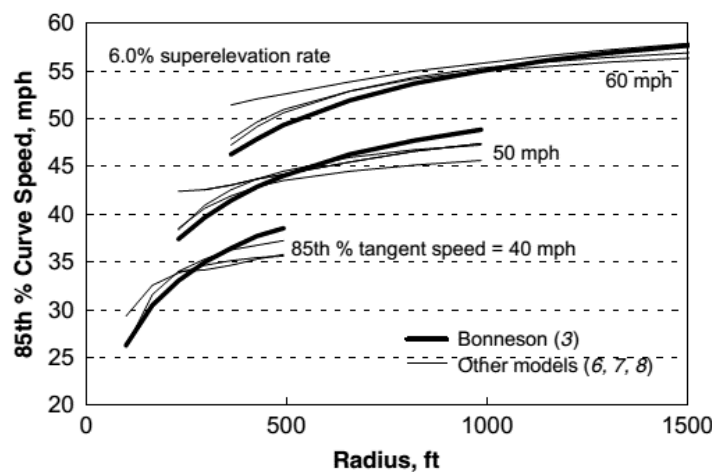
This chapter focuses on drivers' normal driving behavior on rural two-lane curves. The objective of this chapter is to understand how curve radius influences drivers' lateral acceleration and mean speed on rural two-lane curves using the SHRP2 NDS data. A large number of 11,691 observations on rural two-lane curves were summarized from 202 drivers on 219 curves. Understanding drivers' normal curve negotiation process can help transportation engineers improve curve design and develop safety countermeasures.

Section 5.1 briefly summarizes the findings of previous studies on rural curves in terms of vehicle speed, lateral acceleration, lateral position, and crash data analysis. Section 5.2 discusses the data collection, data quality assurance, and the descriptions of the collected variables. Section 5.3 focuses on the analysis of vehicle lateral accelerations, and Section 5.4 focuses on the analysis of vehicle speeds. The implications of the findings are discussed in Section 5.5, and the conclusion is included in Section 5.6.

#### 5.1.1 Vehicle speed

Vehicle speed was identified as one of the most important contributing factors to roadway departure crashes on horizontal curves (Fitzpatrick et al., 1999; Zegger et al., 2000). A lot of attentions has focused on predicting vehicle speeds based on curve geometries. In summary, previous studies found curve radius, deflection angle, curve length, and tangent speeds were the major influencing factors to drivers' speed choice on curves (Bonneson et al., 2009;

Fitzpatrick et al., 1999; Hallmark et al. 2013; Hauer, 1999; Krammes et al., 1995; Lamm and Choveini, 1988; Montella et al., 2015; Schurr et al., 2002). In one of the studies, Bonneson et al. (2009) found tangent speed, radius, deflection angel, and super elevation rate had significant impact on vehicle speeds on curves. The relationship between curve radius and 85<sup>th</sup> percentile speed is shown in Figure 5.1. Schurr et al. (2002) found the drivers only reduced their speeds for curves with radius smaller than 350 m (1146 ft).

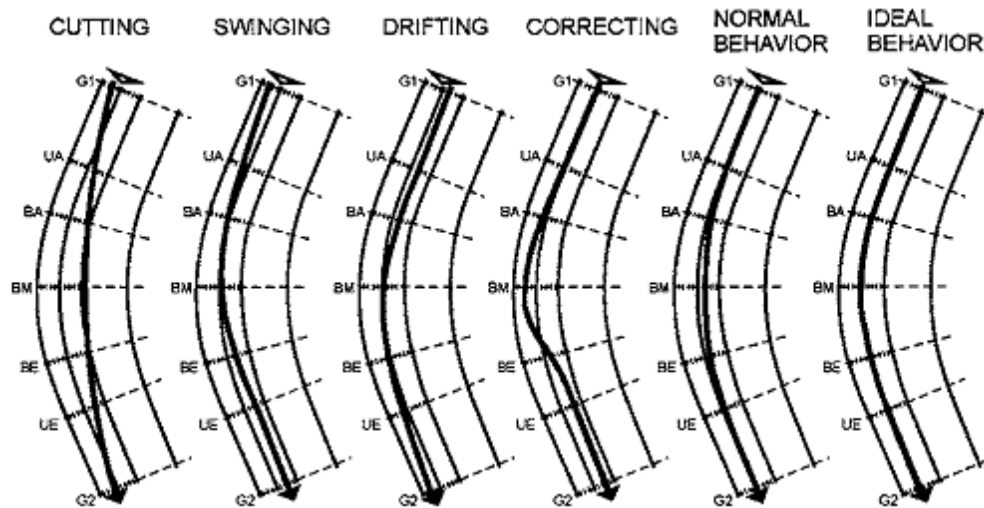


**Figure 5.1 Influence of tangent speeds on curves (Bonneson et al., 2009)**

### 5.1.2 Lateral positions

Vehicle lateral position was often used as a crash surrogate for roadway departure crashes. Most of previous studies measured vehicle lateral positions from roadside equipment. Hallmark et al. (2012) measured the vehicle lateral positions using Z configuration road tubes on rural curves. A statistically significant relationship was found between vehicle speeds over posted speed limits and lane deviations. Gunay et al. (2007) used roadside camera to study the distribution of wheel positions on curves and found the vehicles tended to “cut” the curve at mid-point of the curves. Furthermore, several researchers studied vehicle trajectories on curves and

proposed six patterns of curve negotiation behavior as shown in Figure 5.2. The six patterns included curve cutting, swinging, drifting, correcting, normal behavior, and ideal behavior (AGVC, 1980; Ren, 2012; Spacek, 2005).

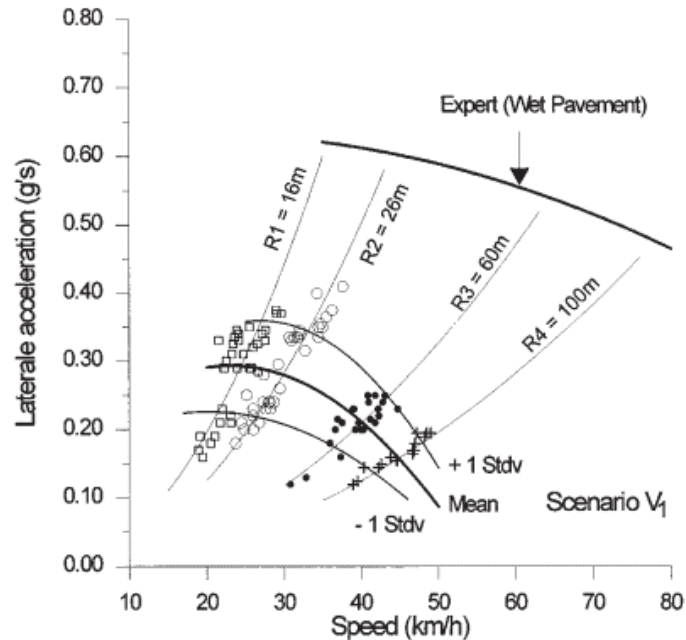


**Figure 5.2 Sketches of six types of trajectories (Spacek, 2005)**

### 5.1.3 Lateral acceleration

Vehicle lateral acceleration also played an important role in determining vehicle's speeds on curves. In one of an earliest studies in 1930s, the Driver Comfort Speed Method was used to establish the advisory speed limits based on drivers' subjective feeling of outward instability, but the method sometimes produces inconsistent and subjective results (FHWA, 2015b). In a more recent study, Felipe and Navin (2007) measured drivers' behavior on a test track. They investigated the influence of speed, pavement surface, gender, and curve radius on speed and lateral position. The lateral acceleration was found to be the dominant factor for speed selection on sharp curves, whereas drivers did not reduce speeds on the relatively flat curves. In this study,

the drivers were asked to drive the curves at their comfortable speeds and identified an important relationship among curve radius, vehicle speed, and lateral acceleration as shown in Figure 5.3.



**Figure 5.3 The relationship between lateral acceleration and speed (Felipe and Navin, 2007)**

#### 5.1.4 Curve-related crash analysis

Numerous studies conducted statistical analysis on the curve-related crashes. Many studies found statistically significant relationship between the number of curve-related crashes and small curve radius. Zegger et al. (2000) evaluated 104 fatal and non-fatal crashes and found curve radius was a critical geometry factor for increased crash rate. Lamm et al. (1988) examined 85<sup>th</sup> percentile speeds, accident rates, and curve radius on 261 rural two-lane curves. They found curvature had significant impact on the 85<sup>th</sup> percentile of speeds and accident rates. Findley et al. (2012) modeled the impact of tangent length on the crash rates and found the distance between two adjacent curves was a statistically significant predictor for the number of crashes.

In addition to curve geometry, curve design consistency is another important concept that had been extensively studied by several researchers (Fitzpatrick et al., 1999; Lamm et al., 1988; Torregrosa et al., 2013; and Wu et al., 2013). Roadway design consistency is defined as consistency between roadway designs and the drivers' expectation. Several papers found the inconsistency of roadway design often resulted in higher number of crashes (Donnell et al., 2009; Gibreel et al., 1999; Montella, 2015). Additionally, Anderson and Krammes (2000) investigated 1,126 rural two-lane curves and found statistically significant linear relationship between speed reductions and the accident rates.

Although previous studies had many important findings regarding drivers' curve negotiation behavior, several limitations were present. First of all, the sample sizes of the previous studies were usually smaller than the SHRP2 NDS data. Second, most of the previous studies were not able to incorporate driver demographics and vehicle information in the analysis. Third, most of previous study collected vehicle dynamics from roadside equipment at several locations on the curves. The speeds had to be interpolated between the measurement locations.

Fortunately, the SHRP2 NDS used advanced data collection technology to collect the time series data at high frequency (10 Hz) from vehicle sensors. This level of detailed information had never been collected in any previous studies. The SHRP2 NDS data overcomes the limitations in previous studies because multiple observations of driver behavior are available, driver demographics and vehicle information are included in the collected data, and continuous traces of driving behavior are available.

### 5.1.5 Research objectives

The objective of this study is to understand the drivers' curve negotiation process and the relationship among curve radius, vehicle speed, and vehicle lateral acceleration on rural two-lane curves. The SHRP2 Naturalistic Driving Study data provides an in-depth observation about drivers' curve negotiation behavior with detailed second-by-second data and other important driver information. This study only focused on the rural two-lane curves on paved roadways. Rural roadway was defined as at least one mile outside city limits. The requested driving data were available 500 feet before the curve, inside the curve, and 200 feet after the curve. This study only included roadways with posted upstream speed limits at either 45 mph or 55 mph. This chapter will focus on the following two research questions:

- 1) How do drivers maintain vehicles' lateral accelerations on rural two-lane curves?
- 2) What factors influence drivers' speeds on rural two-lane curves?

The lateral acceleration analysis in Section 5.4 examined the relationship between curve radius, vehicle speed, and vehicle acceleration. The vehicle speed analysis in Section 5.5 examined the drivers' speeds on 45 MPH roadways and 55 MPH roadways. A linear mixed model was used to predict the mean speed for each individual driving trace. The major findings are discussed in Section 5.6.

## 5.2 Data Description

One advantage of the SHRP2 Naturalistic Driving Study is the diversity of collected variables, such as vehicle sensor data, driver demographics, and vehicle types. To give an example of the magnitude of the SHRP2 NDS data, this rural two-lane curve dataset collected over 5 million rows and 83 columns of time series data. Moreover, the video data, driver

demographics, vehicle types, and curve geometry data were also available in this analysis. Nevertheless, cleaning and processing of such a large dataset is not an easy task. Automatic identification and batch processing was used to minimize the manual efforts for data reduction. This section discusses data collection, data sources, batch processing, and the data quality assurance.

### 5.2.1 Data collection

This study is one of the first analyses on the SHRP2 Naturalistic Driving Study data. This data was first requested as part of the effort for the project *Analysis of Naturalistic Driving Study Data: Roadway Departures on Rural Two-Lane Curves* (Hallmark et al., 2014). By the time of collecting this data in 2013, only one third of the SHRP2 dataset was collected and available for analysis. The five data collection sites included Florida, New York, Indiana, Pennsylvania, and North Carolina. The New York state was not included in the analysis because most of the driving data was collected in urban roadways in the state of New York.

In order to request the curve driving data from VTTI, the candidate curves were first manually identified from ArcGIS. A polygon was created around the selected study sites. VTTI filtered all the driving data within the created polygons. They sent all the time series data in 4,106 spreadsheets. Each spreadsheet represents one driver drove through one buffer once. One spreadsheet could contain multiple curves' driving data. The time series DAS variables includes GPS, accelerometer, event time, forward radar, lane tracking system, steering wheel, and brake pedal information. Most of the variables were collected at 10 HZ, which is every 0.1 seconds. The forward video data and driver face data were not included in the analysis due to the intense

The multivariate dataset consisted of three data sources, which include time series driving data, driver demographics data, and curve geometry data. The time series driving data contained the outputs from the vehicle network and sensors. The driver demographics data contained



drivers' age, gender, and traffic violation history information. The curve geometry dataset had curve radius, length, shoulder type, and presence of guardrail. The three datasets were merged together and the multivariate dataset allowed researchers to examine the relationship among drivers, vehicles, and roadway environments on rural two-lane curves. The variables collected from each data source are discussed in the following paragraphs.

### ***Time Series DAS Data***

The times series driving data included the variables from vehicle network and sensors. A large number of time series variables were collected in this project, but only the relevant variables were presented in Table 5.1. Most of the variables were collected at 10 HZ, which is every 0.1 seconds. Some of the variables were collected directly from vehicle sensors, but some variables were derived from other variables. It should be noted that vehicle speeds were collected from both GPS and vehicle network system. The speed measurement from vehicle network system was used in this study because it was more reliable and collected at higher frequency. Vehicle acceleration was collected from accelerometer. The distance from the wheels to the lane marking was collected from lane tracking system. The front radar was available to measure the distance to the front vehicle. Many other variables were collected from the vehicle sensors, but will not be discussed here.

**Table 5.1 Time series driving data**

<b>Variables</b>	<b>Description</b>	<b>Unit</b>
Event ID	Unique event ID	NA
Time Stamp	Integer used to identify one time sample of data	1/1000 Second
GPS Latitude/Longitude	Vehicle GPS Location	Degree
GPS Speed	Vehicle speed from GPS	MPH
Vehicle Network Speed	Vehicle speed from vehicle network	MPH
Mean Speed in Curve	Average speed inside the curve	MPH
Max Speed in Curve	Maximum speed inside the curve	MPH
Min Speed in Curve	Minimum speed inside the curve	MPH
Standard Deviation of vehicle Speed in Curve	Standard deviation of speeds inside the curve	MPH
Vehicle Speed at PC	Vehicle speed at beginning of the curve	MPH
Tangent Speed	Vehicle speeds at 5 seconds before the curve PC	MPH
Speed Reduction	Speed difference between tangent speed and the beginning of a curve	MPH
Acceleration in X Direction	Vehicle acceleration in the longitudinal direction versus time	g
Acceleration in Y Direction	Vehicle acceleration in the lateral direction versus time	g
Mean of Lateral Acceleration	Average value of lateral acceleration inside the curve	g
Max of Lateral Acceleration	Average value of lateral acceleration inside the curve	g
Left Lane Marker Probability	Probability that vehicle based machine vision lane marking evaluation is providing correct data for the left side lane markings	g
Left lane to Left Wheel Distance	Distance from vehicle centerline to inside of left side lane marker based on vehicle based machine vision	g
Maximum Left Distance	Maximum of lateral distance inside the curve	Meters
Standard Deviation of Left Distance	Standard Deviation of lateral distance inside the curve	Meters
Vehicle Front Radar	Range to forward radar targets measured longitudinally from the radar	Meters

### *Curve Features Data*

In addition to time series DAS data, curve geometry data were also reduced from Google Earth and Roadway Information Database. A list of the curve geometry variables is shown in Table 5.2.

**Table 5.2 Curve geometry data**

<b>Variables</b>	<b>Description</b>	<b>Unit</b>
Curve Direction	Curve Direction	Left/Right
Curve Radius	Curve radius	Feet
Curve Length	Curve length	Feet
Upstream Distance	Upstream distance before the curve	Feet
Downstream Distance	Downstream distance after the curve	Feet
Chevrons	Presence of chevrons	Yes/No
Large Arrow	Presence of large arrow sign (MUTCD W1.6) sign	Yes/No
Paved Shoulder	Presence of Paved Shoulder	Yes/No
Center Line Rumble Strip	Presence of center line rumble strip	Yes/No
Edge Line Rumble Strip	Presence of edge line rumble strip	Yes/No
Guardrail	Presence of guardrail	Yes/No
Curve Advisory Sign	Presence of curve advisory sign (MUTCD W1.2)	Yes/No
Advisory Speed Limit	Posted curve advisory speed limits	MPH
Upstream Speed Limit	Posted upstream speed limits before the curve	MPH
Recommend Speed Reduction	The speed difference between upstream posted speed limits and the advisory speed limits.	MPH
S Curve	A left or right turn curve is followed by another curve in the opposite turn direction	Yes/No
Compound Curve	A left or right turn curve is followed by another curve in the same turn direction	Yes/No
Curve Intersections	Number of intersections inside the curve	Integer
Curve Driveways	Number of driveways inside the curve	Integer

The curve radius was measured on Google Earth as described in section 4.2. The curve upstream distance was measured as the tangent distance from the curve PC to the nearest curve or the nearest intersection prior to the curve. Similarly, curve downstream distance was measured from curve PT to the nearest curve or the intersection after the curve. Other roadway

countermeasures included chevron, curve warning signs, paved shoulder, center line or edge line rumble strips, guardrail, posted speed limits, and advisory speed limits. The number of driveways or intersections were counted for each curve. The types of curves, such as S curve or compound curve, were coded in the dataset. The most relevant roadway characteristics were extracted from either Google Earth or Roadway Information Database. However, the super elevations of the curves were not collected in this study. After the curve geometry dataset was created, it was linked to the time series driving dataset.

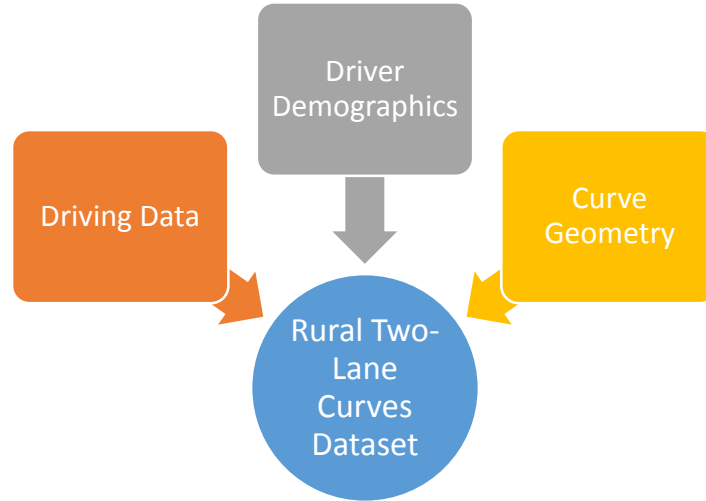
### ***Driver Demographics and Vehicle Types Data***

One advantage of naturalistic driving study data was the access to drivers' demographics and vehicle types. The NDS data contained a number of detailed driver demographics variables such as driver age and gender. The drivers were also asked to fill out several survey before they participated in the study, such as risk perception questionnaire and driving knowledge tests. Additional vehicle information was also available, such as vehicle class, vehicle make, and advanced vehicle technologies. Although a large number of driver and vehicle variables were collected in SHRP2 NDS, only relevant variables to this study are presented in Table 5.3.

**Table 5.3 Driver demographics and vehicle types**

<b>Variables</b>	<b>Description</b>	<b>Unit</b>
Gender	Driver gender at time of data collection	Male/Female
Driver Age	Driver age at time of data collection	Integer Number
Education	Driver education	NA
Years of Driving	Number of years driving experience	Integer Number
Number of Crashes	Number of previous crashes	Integer Number
Number of Violations	Number of previous violations	Integer Number
Vehicle Class	Classification of vehicle type based on body style	NA
Vehicle Make	The make of the participant's study vehicle	NA

Finally, the information from three different data sources were merged into one dataset. Each row in the dataset represented each driver drove through the curve once. This initial dataset contains 11,691 observations from 202 drivers on 219 rural two-lane curves.



**Figure 5.5 Integration of different data sources**

### 5.2.3 Batch process of time series data

Due to the large size of the dataset, it is very labor intensive and time consuming to reduce the dataset manually. Therefore, several programming procedures were written to reduce the 5 million rows of time series driving data automatically in R software. First of all, it was important to identify the time stamps at which the drivers drove through the beginning of a curve (PC) and end of a curve (PT). Manual identification of PC and PT from ArcGIS or forward video can be a very time-consuming process. In order to address these issues, several automatic identification algorithms were written to identify the time stamps at which the drivers drove through the PC and PT points. The distances between each GPS point to the known PC and PT points were calculated. The nearest GPS points were identified as curve PC or curve PT. The formula used to calculate the distance is illustrated in Formula 5.1. If the PC or PT points fell

between any two GPS points, an interpolation method was used to identify the nearest 0.1 second from the time stamp. Overall, the algorithm successfully identified the curve entering and curve exiting time stamp automatically.

$$Distance = \sqrt{(Latitude\ i - Latitude\ j)^2 + (Longitude\ i - Longitude\ j)^2} \quad (\text{Formula 5.1})$$

Additionally, it is necessary to identify travel direction from the time series data.

Normally, the traveled direction was manually coded from forward videos as either left turn or right turn curve, but it is too time consuming to manually reduce the 10,000 observations.

Another programming procedure was written in R to automatically identify the turn direction by judging whether the vehicle passed the PC point first or the PT point first. If the driver drove through the PC first, then the traveled direction was the same as the default curve direction.

Conversely, if the driver drove through PT point first, the driving direction was opposite to the default curve direction. The sample R code for the aforementioned batch processing is shown in Appendix B. Several other algorithms were also developed to identify the upstream speeds and summarize the in-curve driving data, but they are not discussed here due to limited space.

In summary, given the size of the SHRP2 NDS dataset, it is impractical to reduce all raw data manually. Even analyzing a subset of the SHRP2 NDS data could be time consuming or even impractical sometimes. The automated algorithms allowed researchers to batch process the large quantify data in a timely manner.

#### 5.2.4 Data quality assurance

Since most of the data were collected from vehicle sensors in real-world driving environments, many data quality issues emerged in this data set. The first issue was missing data.

For example, the lateral position measurement contained a large quantify of missing data. The missing data could be caused by several reasons, including no presence of pavement marking, unclear pavement marking, and malfunction of lane tracking system. The second issue was outlier measurements. The problem was more significant for forward radar and lane tracking system. The vehicle forward radar could wrongly detect a random object in front of the vehicle and report outlier values. They should be identified and excluded from the dataset before conducting any statistical analysis. In order to gain understanding of the quality of the SHRP2 NDS data, the percentage of missing data was summarized in Table 5.4. It should be noted that the probability for left lane and right lane marking is from 0 to 1024. The lower probability indicates less confidence in the measurement and vice versa. The threshold of 1000 is used here to summary the availability of high quality data.

**Table 5.4 Summary of data quality for the rural two-lane curve dataset**

<b>Variables</b>	<b>Data Availability</b>
GPS Speed	97.7%
Left Lane Marking Probability >1000	62.0%
Right Lane Marking Probability >1000	46.9%
Lighting Sensor	39.5%
Gas Pedal	68.3%
Brake Pedal	37.8%
Steering Wheel Position	22.4%
GPS Gyro in Z Direction	39.5%
Accelerometer X Direction	98.6%
Accelerometer Y Direction	98.6%
Accelerometer Z Direction	95.5%

In summary, vehicle speeds and accelerometer data had decent accuracy, but the lateral positions, gas pedal, brake pedal, and steering wheel position had relatively poor quality. In order to control the data quality in the analysis, several criteria were used to exclude the missing or incomplete data. The percentage of missing GPS data was summarized for each driving trace. The driving trace with more than 20% of missing GPS data were removed from the dataset. As a result, 9,912 out of 11,691 observations were included in the final analysis.

### 5.3 Vehicle Lateral Acceleration Analysis

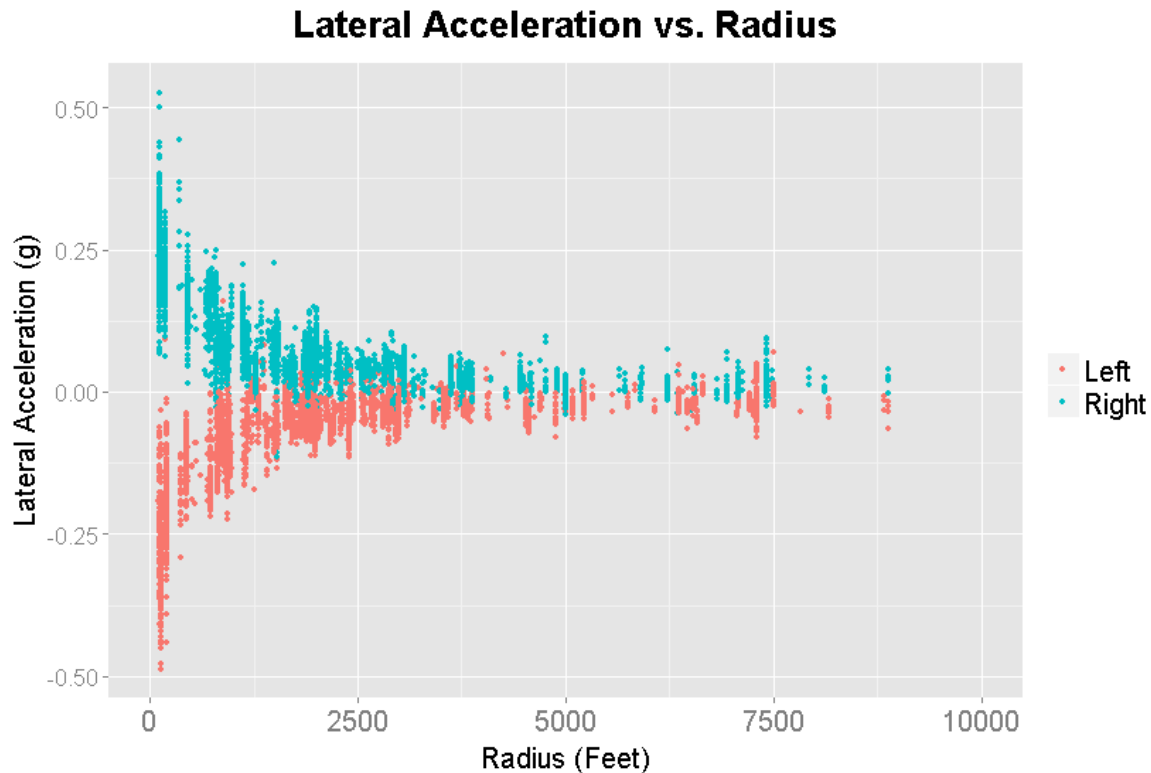
This section examines the relationship among vehicle lateral acceleration, curve radius, and vehicle speed on rural-two lane curves. A total of 9,587 observations were collected on rural two-lane curves with radius ranged from 116 feet to 13,640 feet. The first analysis examines how driver's lateral acceleration is affected by different curve radius. The second analysis investigates the relationship between lateral acceleration and vehicle speed. The third analysis develops cumulative distribution function of lateral acceleration for different curve radius.

#### 5.3.1 Lateral acceleration vs. curve radius

The vehicle lateral acceleration was plotted against curve radius on Figure 5.6. The x-axis is the curve radius and the y-axis is the lateral acceleration in the unit of gravity (g). The positive value indicates the lateral acceleration to the right side of the sensor, while the negative value indicates the lateral acceleration to the left side of the sensor. Figure 5.6 showed the lateral accelerations were symmetrical about the x-axis between the left turn curves and the right turn curves (both are from the perspective of the driver). Additionally, the magnitude of lateral acceleration was found to increase exponentially as the curve radius decreased. The rate of



increasing was much higher for the curves with radius less than 1000 feet. The findings indicated higher likelihood of loss of vehicle control and drifting off the roads on the curves with radius less than 1000 feet. This might explained why there was a large number of roadway departure events on curves with radius less than 1000 feet in Chapter 4.

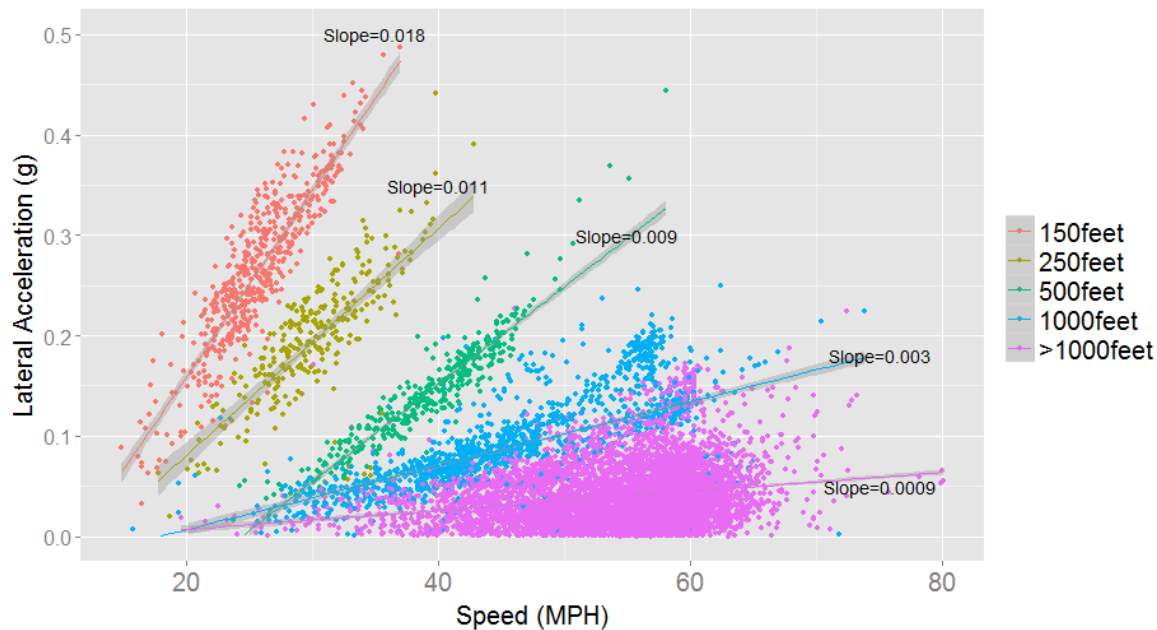


**Figure 5.6 Plot of lateral acceleration vs. curve radius**

### 5.3.2 Lateral acceleration vs. speed

The vehicle lateral acceleration was plotted against vehicle speeds as show in Figure 5.7. Several clusters of data points were found on this plot. Further investigation found the five clusters represented different ranges of curve radius. In order to better understand the relationship within each cluster, the simple linear regression model was fitted to each cluster. The 150-foot cluster had the steepest slope at 0.018. It means the lateral acceleration increased

0.018 g for every 1 mpg increases in vehicle speeds. Similarly, the lateral acceleration increased 0.011 g for every 1 MPH increases on the 250-foot curves. Interestingly, when the curve radius was larger than 1000 feet, every 1 MPH increase in vehicle speeds only increased lateral acceleration by 0.0009 g. This finding had important implication for why speeding on the sharp curves with radius less than 1000 feet was much more dangerous than the speeding on the curves with radius larger than 1000 feet. The findings also explained why there were disproportionately higher number of safety critical events on the curves with radius less than 1000 feet in Chapter 4.



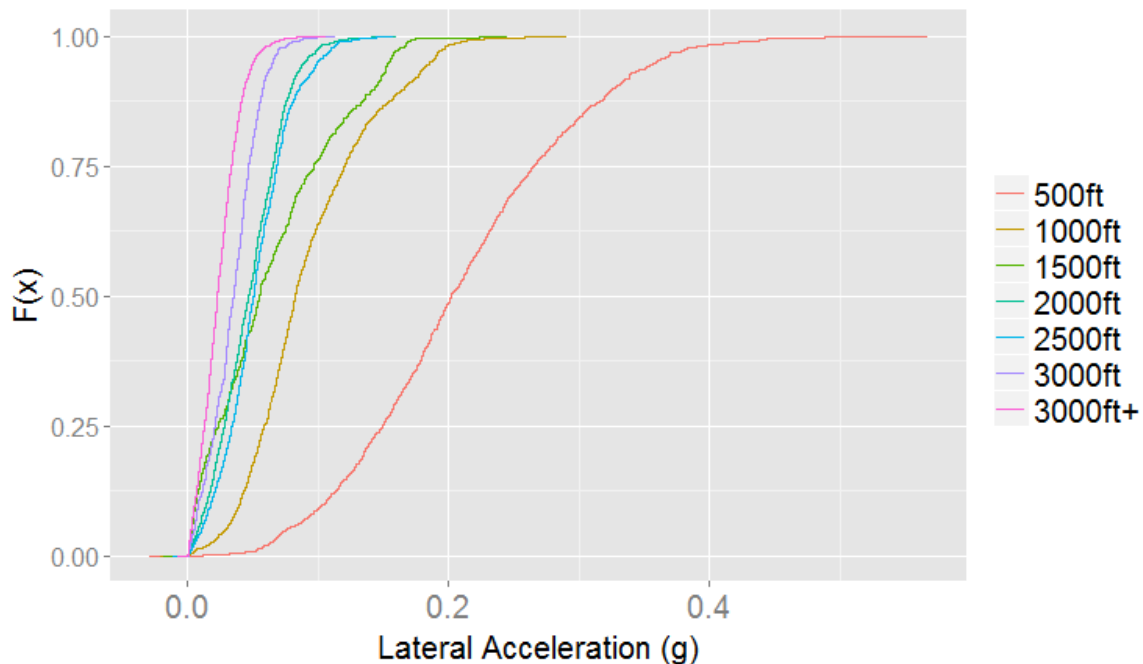
**Figure 5.7 Plot of lateral acceleration vs. speeds**

### 5.3.3 Drivers' comfortable lateral acceleration on curves

It is critical to understand the drivers' preferred lateral acceleration on different curves because transportation agencies relied on the lateral acceleration information to set advisory speed limits on the curves (FHWA, 2015b). In this study, the curves were categorized into six

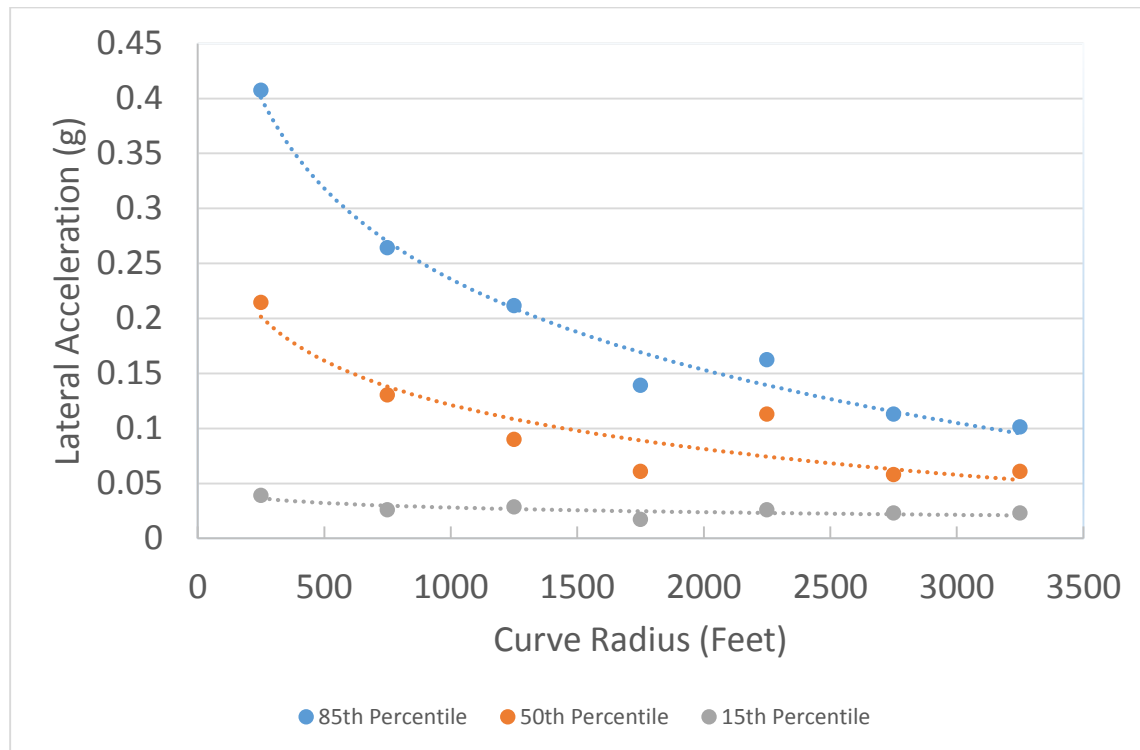
groups with different ranges of curve radius. The 500-foot curves indicated the curves with radius ranged from 0 feet to 500 feet. The 1000-foot curves indicated the curves radius ranged from 500 feet to 1000 feet, so forth and so on. The curves with radius larger than 4000 feet was put into one single category.

The cumulative distribution function (CDF) of lateral acceleration for each radius category was plotted on Figure 5.8. The distribution function calculated the probability that the value was less than or equal to the lateral acceleration value on the x-axis. For example, for those drivers who drove on the curves with radius less than 500 feet (the CDF in red color), 50% of the drivers had lateral acceleration less than 0.2 g. The CDF was found to be steeper for the group of curves with larger radius. On the contrary, the CDF was flatter for the curves with smaller radius. The flatter CDF curve indicated the drivers tend to take higher lateral acceleration on those sharp curves.



**Figure 5.8 Cumulative distribution function of lateral acceleration by curve radius**

The 85<sup>th</sup> percentile, 50<sup>th</sup> percentile, and 15<sup>th</sup> percentile of drivers' lateral acceleration by curve radius are plotted on Figure 5.9. The drivers were found to tolerate higher lateral acceleration on the sharp curves. The 85<sup>th</sup> percentile line indicates 85<sup>th</sup> percentile of the drivers have lateral acceleration below the blue dotted line. Similarly, the 50<sup>th</sup> percentile line indicates 50% of the drivers have lateral acceleration below the orange dotted line. This graphs showed the drivers' comfortable lateral acceleration derived from empirical SHRP2 NDS data. This information could be used by transportation agency to set appropriate advisory speed limits on curves.



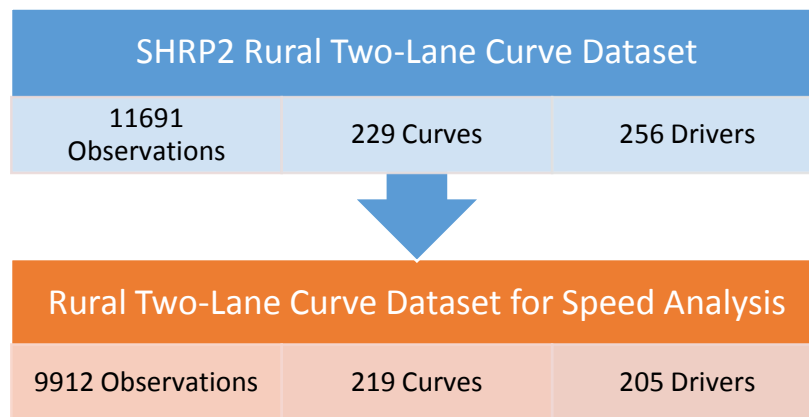
**Figure 5.9 Plot of 85<sup>th</sup>, 50<sup>th</sup>, and 15<sup>th</sup> percentile of lateral acceleration by curve radius**

Over all, this section investigated the relationship among vehicle lateral acceleration, curve radius, and vehicle speeds. The lateral acceleration was found to increase exponentially as

curve radius decreased. The 85<sup>th</sup> percentile of lateral acceleration was developed to help set appropriate advisory speed limits on curves.

## 5.4 Vehicle Speed Analysis

Speeding has been identified as one of the most critical contributing factor to crashes on curves. This section focused on the analysis of vehicle speeds on rural two-lane curves with 45 MPH upstream speed limit and 55 MPH upstream speed limit. The first research question was to understand drivers' mean speeds and their compliance to advisory speed limits on curves in Section 5.4.1 and 5.4.2. Additionally, a linear mixed model was used to predict vehicle speeds on 45 MPH roadways based on the explanatory variables in Section 5.4.3.



**Figure 5.10 Data quality assurance of vehicle speed data**

It is important to conduct quality assurance before any statistical analysis. Vehicle speeds from vehicle network system were used in this analysis because they were more reliable and also collected at higher frequency than the speed data collected from GPS. If more than 10% of the speed data were missing, the driving traces were excluded from the analysis. Additionally, any driving traces with more than 20% of missing GPS data were excluded from the analysis.

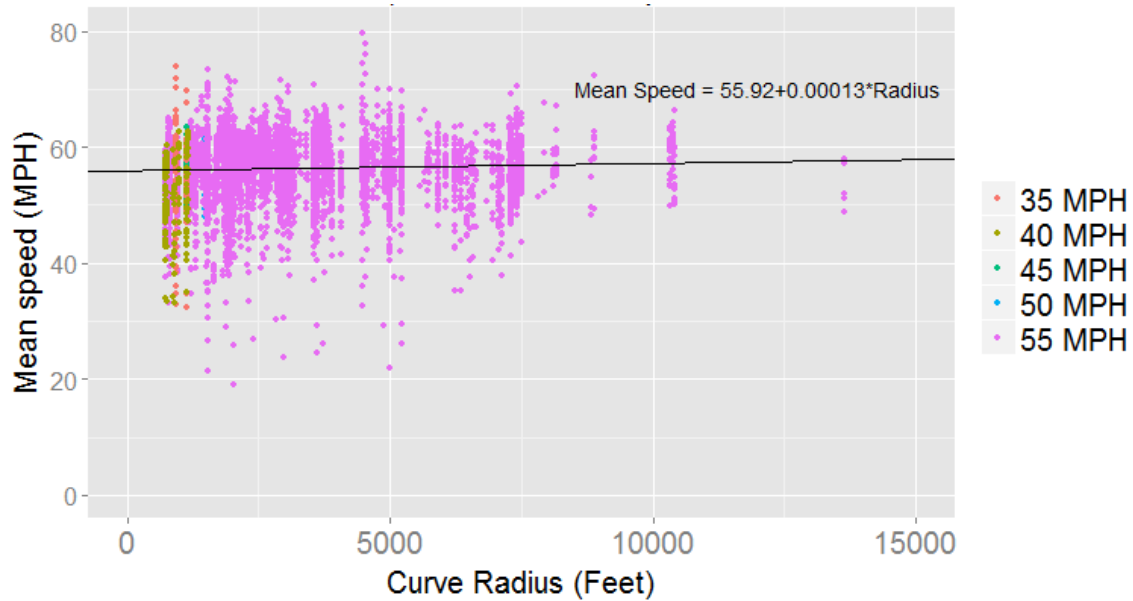
Sometimes the vehicle could be stopped on the roads because of congested traffic, so any driving traces with minimum speed less than 10 mph were excluded from the study. Finally, 9,912 out of 11,691 observations (84.8%) were kept in the dataset as shown in Figure 5.10. The vehicle speeds on 55 MPH roadways and 45 MPH roadways were discussed separately in the following sections.

#### 5.4.1 Drivers' curve negotiation behavior on 55 MPH roadways

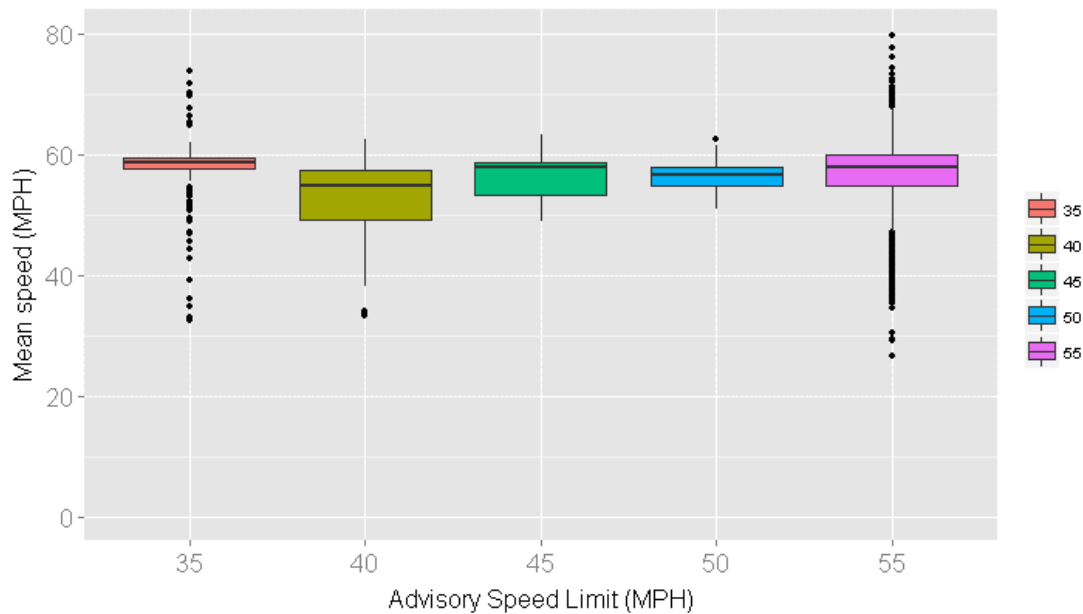
The mean speeds was plotted against curve radius on 55 MPH roadways in Figure 5.11. The y-axis is the mean speed on the curves and the x-axis is the curve radius. Each data point represents one driver drives through a curve once. There are 6,399 observations on the 55 MPH roadways. The majority of the curves had 55 MPH speed limits on curves as indicated by the purple color, but some curves had lower advisory speed limits as indicated by other colors. Nevertheless, the advisory speed limits were mostly installed on the curves with smaller radius. The observations collected in this study had radii ranged from 715 feet to 13,640 feet. None of the collected curves had radius smaller than 715 feet. One reason could be that the Manual on Uniform Traffic Control Devices (MUTCD) suggested transportation engineers not to design any curves less than 1,060 feet on 55 MPH roads.

A simple linear regression model was fitted to predict mean speed based on curve radius variable only. The fitted regression model is shown on Figure 5.11. The slope of the fitted regression line is around 0.00013, which indicated very little change in mean speeds as radius increased. It is an important finding that curve radius had little influence on vehicle speeds for the curves on 55 MPH roadways. Additionally, the variability of vehicle speeds were found to be very large even within the same curve radius. It might be caused by different individual driver's

behavior and various traffic conditions. Therefore, it is suggested to investigate how individual drivers' behavior affects vehicle speed on curves.



**Figure 5.11** Plot of mean speeds by curve radius on 55 MPH roadways



**Figure 5.12** Boxplot of mean speeds by advisory speed limits on 55 MPH roadways

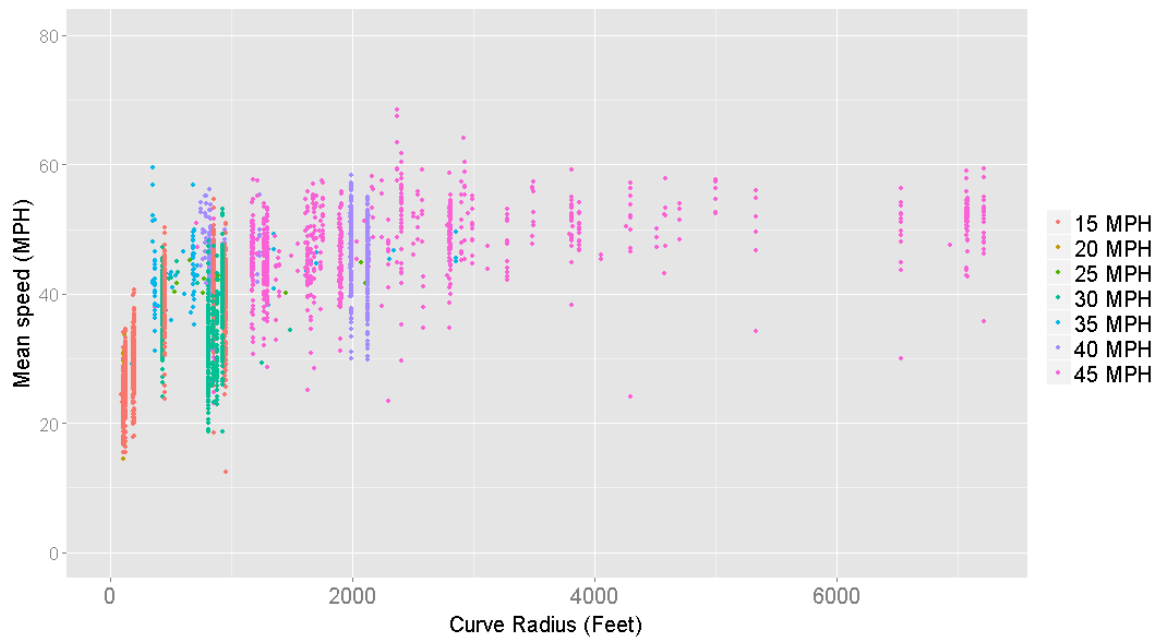
Furthermore, the boxplot of means speeds were summarized by different advisory speed limits on Figure 5.12. It was found the mean speeds were almost the same regardless of the advisory speed limits. It indicated drivers' poor compliance with advisory speed limits. Therefore, advisory speed limits were found to be ineffective to reduce drivers' speeds on 55 MPH curves. It is suggested to choose alternative countermeasures to reduce drivers' speed on curves.

#### 5.4.2 Drivers' curve negotiation behavior on 45 MPH roadways

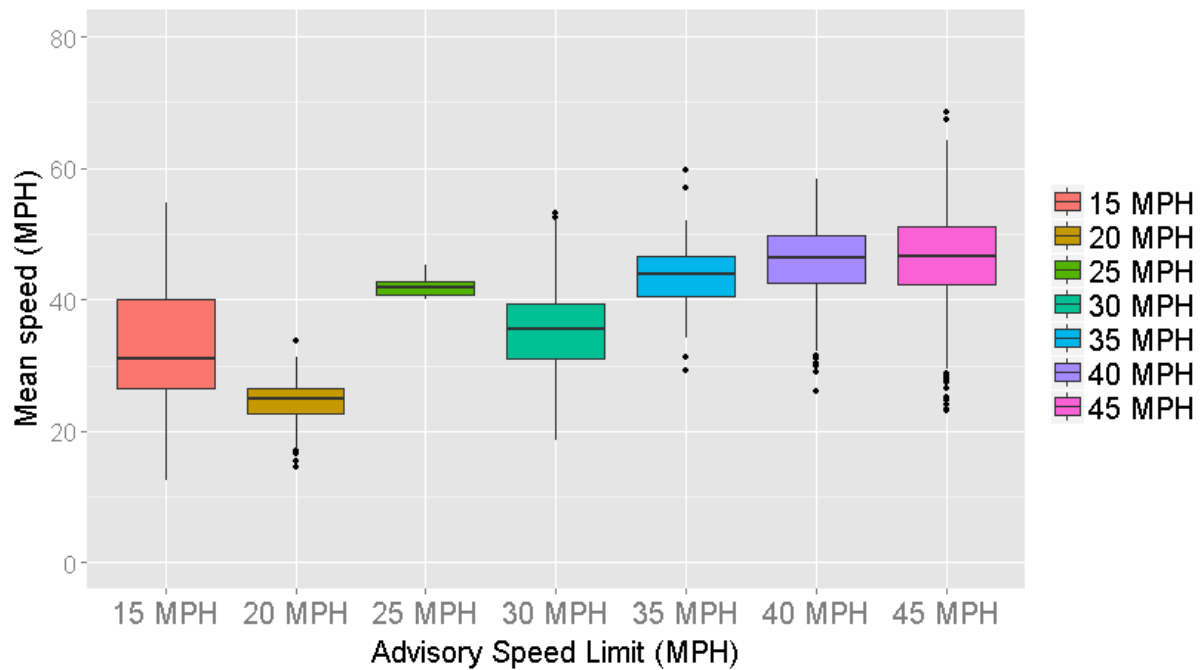
The mean speeds versus curve radius on 45 MPH roadways were plotted in Figure 5.13. The y-axis is the mean speed on the curves and the x-axis is the curve radius. There were 3,507 observations on the 45 MPH roadways. The advisory speed limits were indicated by different colors. The mean speeds were found to increase as the curve radius increased from 100 to 2000 feet, but the curve speeds seemed to level off after the curve radius was above 2000 feet. Again, large speed variability was found within the same curve. Therefore, it is important to understand how driver factors influenced vehicle speeds on curves.

The boxplot of mean speeds by advisory speed limits was plotted on Figure 5.14. It was found that the drivers reduced their speeds for the advisory speed limits on 45 MPH roads, but still drove much higher speeds than the advisory speed limits. However, smaller advisory speed limits were also correlated with smaller curve radius, so it was not clear if the speed reduction was caused by the advisory speed limits or the smaller curve radius. This problem is further investigated in the speed prediction model in the next section.





**Figure 5.13 Plot of speeds by curve radius on 45 MPH roadways**



**Figure 5.14 Boxplot of mean speeds by advisory speed limits on 45 MPH roadways**

### 5.4.3 Speed prediction model

Most of previous speed prediction models used linear regression model to predict the 85<sup>th</sup> percentile of vehicle speeds on curves (Bonneson, 2009; Krammes, 1996; Fitzpatrick, 1999). These studies found the factors influencing vehicle speeds included curve radius, deflection angle, curve length, and tangent speeds (Bonneson et al., 2009; Fitzpatrick et al., 1999; Hallmark et al. 2013; Hauer, 1999; Krammes et al., 1995; Lamm and Choveini, 1988; Montella et al., 2015; Schurr et al., 2002;). Instead of predicting the 85<sup>th</sup> percentile speed for each curve, this dissertation focuses on predicting the mean speed for each individual observation at each curve. It is a more difficult task than predicting the 85<sup>th</sup> percentile speed on a curve for a group of drivers. Additionally, some observations were collected from the same driver or the same curve, so the observations were interdependent from each other. In order to address the interdependency issue, linear mixed effect model was used to account for the within-driver correlation and within-curve correlation. The linear mixed effect model was a flexible approach to account for different types of correlations in the dataset. The details of the linear mixed effect model is discussed below.

#### ***Linear Mixed Effect Model***

Mixed effect model is a type of model included both fixed effect terms and random effect terms. Fixed effect refers to the effect that is identical across all groups, which has fixed number of possible values. For example, driver gender is usually treated as fixed effect. The random effect indicates the effect that varies from group to group. The effect is usually randomly drawn from a population. For example, the individual driver in the SHRP2 NDS is randomly sampled

from a larger driver population in the U.S., so the driver ID was treated as random effect in this model. Similarly, the curve ID was also treated as random effect in this model.

The linear mixed model is similar to a generalized linear regression model, but includes both fixed effect and random effect in the same model. The relationships of the independent variables are often assumed to be additive. The linear effect model has the form:

$$Y_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} \dots + \beta_n X_{nij} + b_{i1} Z_{1ij} + b_{i2} Z_{2ij} + \dots b_{in} Z_{nij} + \varepsilon_{ij}$$

Where  $Y_{ij}$  is the response variable for the observation  $j$  in group  $i$ .  $\beta_1$  to  $\beta_n$  are the fixed effect coefficients.  $X_{1ij}$  to  $X_{nij}$  are the fixed effect variables.  $b_{i1}$  to  $b_{in}$  are the random effects, which are assumed to be multivariate normally distributed.  $Z_{1ij}$  to  $Z_{nij}$  are the random effect variables. It is important to note that the fixed effects have fixed coefficient across groups, but the random effect has  $i$  number of random effects for each  $i$  number of groups.  $\varepsilon_{ij}$  is the error term for observation  $j$  in group  $i$ , which is assumed to be multivariate normally distributed. The model is estimated by Restricted Maximum Likelihood method. The variables are selected using the stepwise backward selection method. The Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and log likelihood tests were used to evaluate the model goodness-of-fit. The lme4 package in R 3.1.1 was used to fit the linear mixed model.

### ***Model Results***

The variables listed in Table 5.1, Table 5.2 and Table 5.3 were initially considered in the mixed linear model, but only statistically significant variables were kept in the final model. As discussed before, driver ID and curve ID were treated as random effect parameters to account for the interdependency within the same driver and the same curve. The AIC, BIC, and log-

likelihood ratio test were used for model comparison and model selection. The model with the smallest AIC and BIC values were selected as the final model. The quasi R-squared value is 0.87, which indicates decent fit of the data. The residual plot was found to be randomly scattered. The log-likelihood ratio test was used to test the statistical significance of the random effect terms. Both driver factor and curve factor were found to be statistically significant and the random effects should be kept in the model. The final model is shown in the following formula.

$$\text{Mean Speed}_{ij} = 0.447 * \text{Tangent Speed} + 0.1634 * \text{Advisory Speed Limits} + 4.634 * \text{Log Radius} - 2.782 * \text{Car Following} - 1.725 * \text{Younger Driver} + \text{Driver Effect } i + \text{Curve Effect } j$$

**Table 5.5 Fixed effects of the speed prediction model on rural two-lane curves with 45 MPH upstream speed limit**

Parameter	Estimate	S.E.	t Value	P-Value	5% CI	95% CI
Tangent Speed	0.447	0.0104	43.251	<0.0001	0.4271	0.4679
Advisory Speed Limits	0.1634	0.0181	9.053	<0.0001	0.1279	0.1989
Log Radius	4.634	0.2084	22.233	<0.0001	4.2253	5.0480
Car Following	-2.782	0.2962	-9.392	<0.0001	-3.3645	-2.1987
Younger Driver	-1.725	0.7103	-2.429	0.0162	-3.2668	-0.2705

Five fixed effect parameters were found to be statistically significant in the model. They were tangent speed, advisory speed limits, logarithm of radius, car following, and younger drivers. The estimates of the coefficients were included in Table 5.5. The tangent speed, advisory speed limits, and logarithm of radius were positively contributed to higher mean speeds. The car following and younger drivers were negatively contributed to the mean speeds on curves. For every 1 mph increase in tangent speed, the vehicle increased its mean speed by 0.447 MPH. For every 1 mph increase in the advisory speed limits, the mean speed increased 0.1634 MPH. The

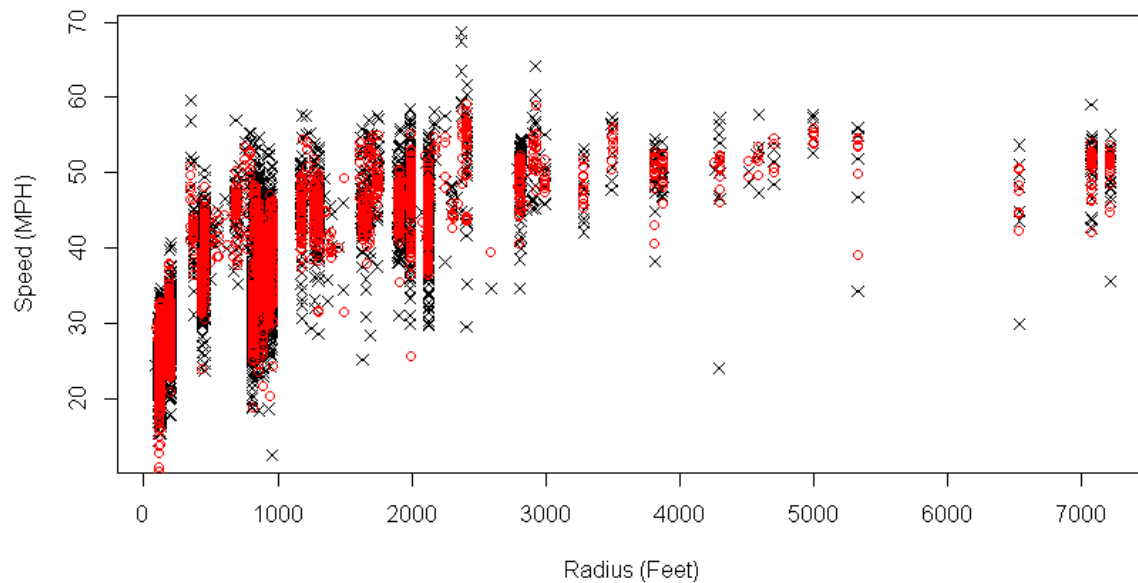
mean speed increased 4.634 MPH for every 1 unit increase in logarithm of radius. If the subject vehicle was following another vehicle in front, the mean speed was reduced by 2.782 MPH. It was also interesting to see younger driver from 16 to 24 years old had 1.725 MPH lower mean speed compared to other drivers in the study. It might be due to the lack of experience of driving on curves.

**Table 5.6 Random effect estimate from the linear mixed effect model**

<b>Group</b>	<b>Variance</b>	<b>S.D.</b>
Drivers (Intercept)	18.697	4.324
Curve (Intercept)	7.393	2.719
Residual	31.842	5.643

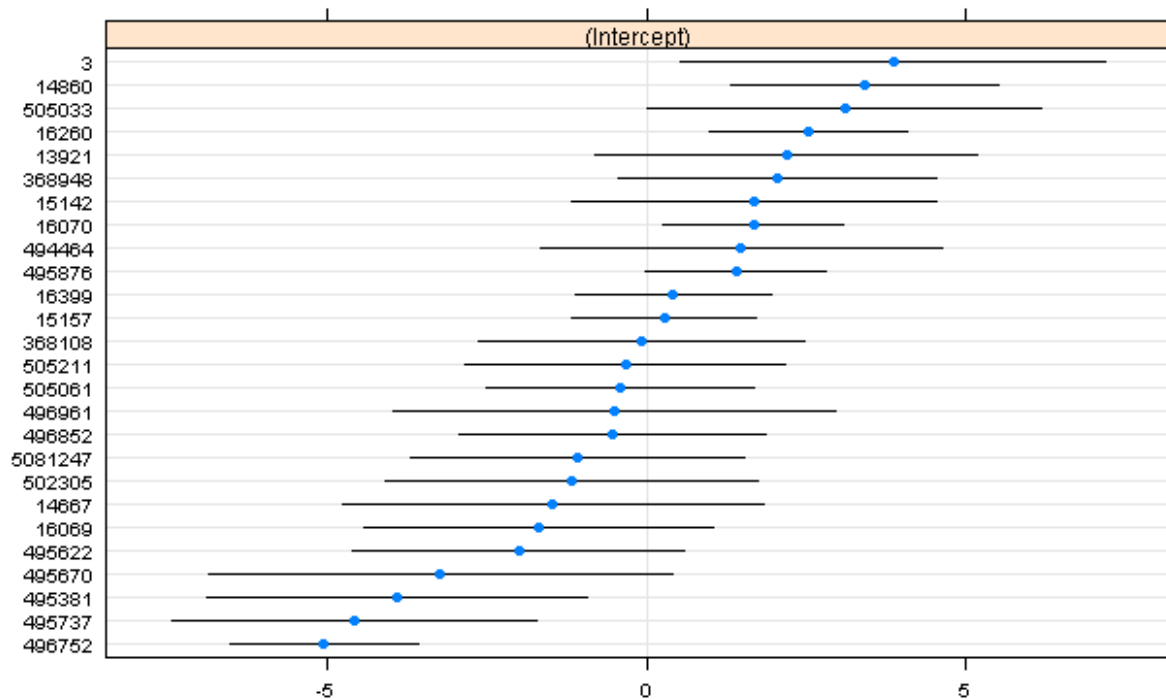
The random effect estimates from the linear mixed effect model is shown in Table 5.7. The mixed effects of driver factor and curve factor were found to be statistically significant, so the mixed effects should be kept in the model. The standard deviation of driver factor was 4.324, which represents the variability of speeds due to the drivers. The standard deviation for curve factor was 2.719, which represented the variability of speeds due to the curves. The standard deviation of the residual was 5.643, which represented the variability of speeds that could not be explained by the model. In summary, 34% of the variability was explained by the differences in drivers. The curve explained 21.4% of variability. The rest 44.6% of variability was the noises that was not explained by the model.

In order to check the model goodness-of-fit, the observed mean speeds was plotted against the predicted mean speeds on Figure 5.15. It can be seen that the predicted value followed the observed observations closely. It concluded that the model fits the observations decently.



**Figure 5.15 Comparison of observed mean speeds to the predicted mean speeds on rural two-lane curves (The red dots are predicted value and the black crosses are observed values).**

The random parameter assigned to each driver could be further investigated in Figure 5.16. Each row on the y-axis represents one driver. The x-axis represent the drivers' baseline speeds on curves in MPH. As shown in Figure 5.16, some drivers tend to drive faster than other drivers while some other drivers tend to drive slower than other drivers on curves. For example, the Driver 3 tend to drive approximately 4 mph faster on curves than the average drivers in this study. The Driver 496752 tends to drive 5 MPH lower than the average drivers in this study. The confidence interval for each driver was also plotted for each driver. Some drivers were found to have large confidence intervals, while some drivers had smaller confidence intervals. In general, the distribution of drivers' speeding behavior generally follows normal distribution. This methodology had important implications for identifying risky drivers who tended to be speeding on curves.



**Figure 5.16 Plot of random effects for the drivers**

Overall, the SHRP2 NDS allowed researchers to examine a number of variables that were not available in previous research. The linear mixed model successfully predicted vehicle mean speeds on two-lane tangent roadways as a function of driver behavior, roadway characteristics, and traffic environments. The individual drivers' speeding behavior could also be examined in random parameters. However, it is recommended to incorporate more drivers in future studies.

## 5.5 Discussion

The objective of this chapter was to understand how curve radius affect drivers' mean speed and lateral acceleration on rural two-lane curves using the SHRP2 NDS data. A total of 11,691 observations were collected from 202 drivers on 219 curves. This study built one of the largest driving dataset that has been collected on rural curve safety research. The findings could

be used to help transportation agencies to understand drivers' behavior on curves and set appropriate advisory speed limits for the curves. Two main analyses were presented in this chapter: the lateral acceleration analysis and the mean speed analysis.

The multivariate dataset consisted of data from three sources: time series DAS data, curve feature data, and driver demographics data. The time series data set was a big data set that contained more than 5 million rows with 83 columns. Manual reduction of such a large dataset would be almost impractical. Several batch processing programs were written in R to reduce the data for statistical analysis. The quality assurance procedures were also applied to ensure the quality of the data is reliable and accurate. In addition to the aggregated time series data, some other data sources, such as driver demographics and curve geometries, were also linked to the time series driving data.

The first analysis was vehicle's lateral acceleration on rural two-lane curves. It was found the lateral accretion increased exponentially as curve radius decreased. The relationship between lateral acceleration and vehicle speed was found to be dominant by curve radius. It again confirmed that the curve radius was a dominant factor for drivers' behaviors on curves. The 85<sup>th</sup> percentile of vehicle lateral acceleration on different curve radius were plotted in Figure 5.9. The magnitude of vehicle lateral acceleration was found to increase exponentially as curve radius decreased. This empirical data implied drivers' comfortable range of lateral acceleration on different curve radius. The transportation agencies could use this information to set appropriate advisory speed limits on curves.

The second analysis was vehicle's mean speeds on curves. There are two scenarios in this speed analysis: the roadways with 55 MPH upstream speed limit and the roadways with 45 MPH



upstream speed limit. For 55 MPH roadways, the curve radius were above 716 feet because the minimum speed limit recommended by MUTCD was 1060 feet, even though some smaller curve radius were found in this study. The advisory speed limits were found to be ineffective to decrease vehicle speeds. However, the vehicle speed was found to decrease as curve radius decreased from 2000 feet to 100 feet on 45 MPH roadways.

In order to better examine the relationship between curve radius and vehicle speed on 45 MPH roadways. A linear mixed model was used to predict the mean speeds on curves for each observation. The linear mixed model treated the driver factor and curve factor as random effects in the model. On one hand, the results found tangent speeds, advisory speed limits, logarithm of curve radius were positively correlated to higher mean speeds on curves. On the other hand, car following and younger drivers were found to have lower mean speeds on curves. The random parameters also revealed the drivers' baseline speeding behavior on the curves. The linear mixed model successfully predicted the mean speeds for each driving trace. The model has important implications for curve speed warning system and identifying risky driver groups.

## 5.6 Summary

In summary, this chapter examined the vehicle speed and vehicle lateral acceleration on rural two-lane curves. A large dataset was assembled from different sources. The multivariate analysis of vehicle speeds and vehicle lateral acceleration provided important insights on drivers' behavior on rural two-lane curves. The findings have important implications for improving curve design, developing curve warning system, and identifying risky driver groups. However, the lateral positions variables were not included in the analysis and it is recommended to investigate how curve radius, vehicle speed, and lateral acceleration affect drivers' lateral position on curves

in future research. The data quality assurance also summarized the data availability in the SHRP2 NDS project. It had some practical implications for future researchers to understand the quality of the data.

## CHAPTER 6. FUNCTIONAL DATA ANALYSIS OF TIME SERIES DATA ON RURAL TWO-LANE CURVES

### 6.1 Introduction

Previous chapters mainly focused on the analysis of the summarized event level data. However, the majority of the raw data collected in the SHRP2 NDS was time series data, such as vehicle speed, vehicle acceleration, and vehicle lateral position variables. The majority of the time series data in the SHRP2 NDS were collected at high frequency at 10 Hz (every 0.1 second), while some variables were collected at 1 Hz (every 1 second). This raised an interesting research question about what driver behavior information can be learned from analyzing the time series data. This chapter analyzed time series speed data using functional data analysis. Section 6.1 introduces the functional data analysis and explained why it is an appropriate method. Section 6.2 discusses the time series speed data and the selected sample curves in this study. Section 6.3 explains the methodology of functional data analysis, because it is a relative new statistical method. The main analyses and findings are presented in Section 6.4. The final discussion and conclusion are included in Section 6.5 and Section 6.6.

#### 6.1.1 Challenges of analyzing time series data in the SHRP2 NDS

Use of time series data was not a straight forward task and brought many challenges, including missing data, outliers, and the lack of well-established statistical method. First of all, the on-road data acquisition system have missing data due to the nature of field data collection. For example, the missing data in lateral position measurement could be caused by the malfunction of lane tracking system, discontinuities in pavement lane markings, or snow

coverage on the lane marking. Second, the collected data could be either inaccurate or contain outliers. Outliers are exceptionally high or low values occurring among data that does not fall into the normal range of values. The outliers should be spotted and removed before conducting any statistical analysis. Third, another challenge is that there was the lack of well-established statistical method to analyze time series data. The classical ARIMA (Autoregressive Integrated Moving Average) model is mainly used for forecasting purpose, which is very appropriate for describing drivers' behavior on roads. Therefore, analyzing the time series data from the SHRP2 NDS project is an both interesting and challenging task.

### 6.1.2 Introduction to functional data analysis

Functional data analysis emerged in early 1990s and quickly developed in the past twenty years. It is a branch of statistical method that focuses on the analysis of information from continuous curves or surfaces. The primary interests of functional data analysis is to understand the variations in the underlying process over a group of repeated measurements. In functional data analysis, a series of discrete time series observations is converted to a functional observation. Recently, this method started emerging in some research fields, such as medical field, environmental monitoring, and economic research (Ramsay et al., 2002). Nevertheless, this method has not been used in any transportation studies. Ramsay et al. (2005) discussed the functional data analysis in his book *Functional Data Analysis*, and pointed out the method is appropriate for the following types of data:

- **High Frequency Measurement.** The functional data analysis is appropriate for analyzing the time series data collected at high frequency, so that the features of the curve or surface are continuously captured in the time series data.

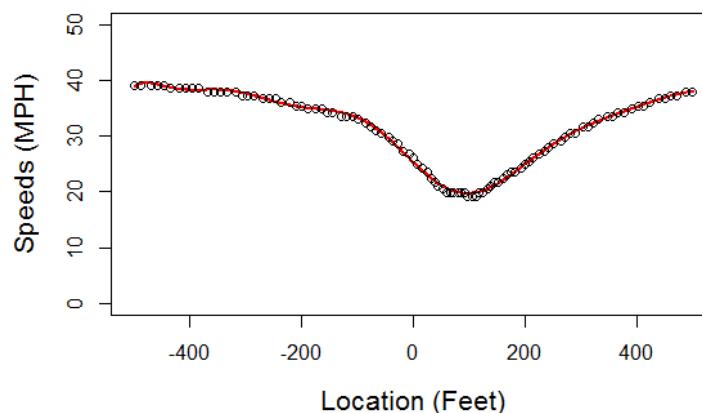
- **Smooth Process.** One important assumption of functional data analysis is that the underlying process is a smooth process. The main interest of FDA is to understand how the smooth process changes over time.
- **Complex Process that cannot be expressed in Parametric Model.** Functional data analysis is especially useful when the underlying process is difficult to be expressed in a parametric model. It can be used to examine complex process and explain how the process changes over time.
- **Repeated observations over the same process.** Functional data analysis is appropriate when a group of observations were repeatedly measured over the same process. FDA can be used to examine the similarity and difference between a set of repeated measures, which could not be done in traditional time series analysis.
- **High Dimensions.** Functional data analysis can be used to examine the correlation between multiple time series variables measured from the same process. For example, it is possible to identify the correlation between vehicle lateral acceleration and vehicle lateral position collected in the same time period.

Based on the discussion above, Functional data analysis is an appropriate method for analyzing the time series data from the SHRP2 NDS. First of all, the data collected in SHRP2 NDS has high frequency at 10 Hz, while some variables were collected at 1 Hz. The majority of the vehicle sensor data from the SHRP2 NDS followed a smooth underlying process. For example, vehicle speeds usually change smoothly over time. Additionally, drivers' behavior is a complex process that cannot be written in a parametric model. As discussed above, functional data analysis is appropriate for describing complex underlying process. Furthermore, FDA is a very useful tool to summarize the similarities and differences for a group of repeated time series

observations measured on the same roadway segment. Finally, many time series variables were collected simultaneously in the SHRP2 NDS. FDA can be used to examine how time series variables correlated with each other. In summary, functional data analysis is an ideal statistical method for analyzing times series data from the SHRP2 Naturalistic Driving Study.

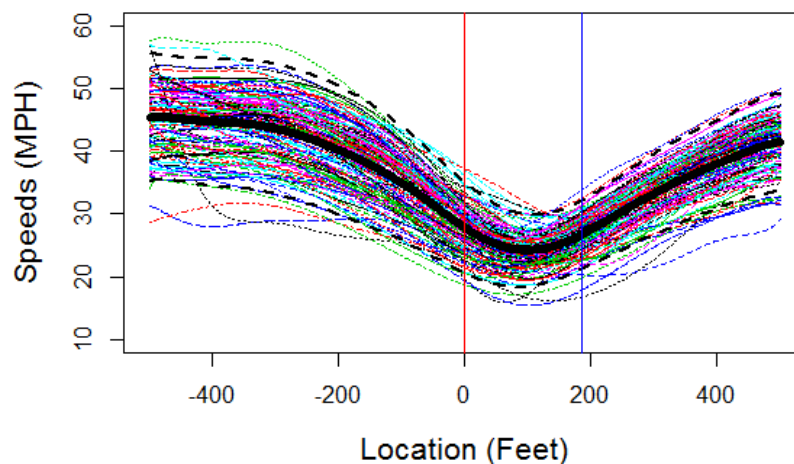
### 6.1.3 Example of functional data analysis

However, the analysis of functional data is not a straightforward task. It involves several steps to convert discrete time series data to functional data. Many methods could be used to build the best fitted spline function to the discrete time series data. The most popular two methods are the B-spline method and Fourier basis method. The development of B-spline function involves many steps, including choosing the number of knots, choosing the number of basic functions, and panelizing for smoothing curvature. Figure 6.1 plotted a sample of time series speed data and the best fitted B-spline function in red color. The development of function data from discrete time series data already smooths the noises as part of the first step in FDA. After the functional data is created from discrete data, the values can be evaluated at any points from the developed best fitted functions.



**Figure 6.1 Example of converting discrete speeds profile to functional data on curve NY67a**

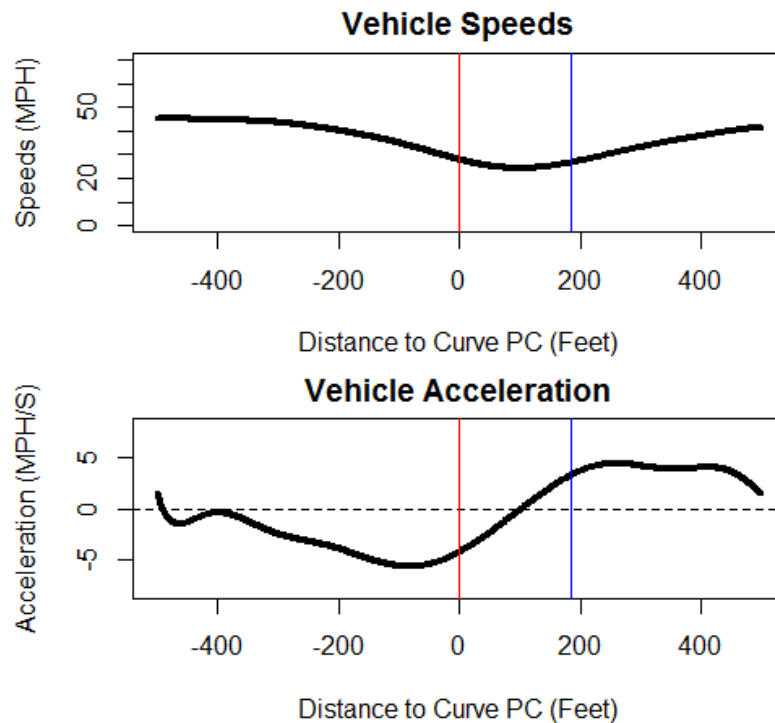
After the discrete raw data is converted into functional data, some functional analysis techniques could be used to summarize the mean and confidence interval for a group of functional data. The most popular type of functional data analyses is to calculate the average functional observation from a group of repeated functional data measured on the same process. For example, Figure 6.2 plotted 146 driving traces observed on the same curve. The x-axis is the distance to the beginning of the curve. The negative value indicates the distance before the beginning of the curve. The positive value indicates the distance after the beginning of the curve. The red vertical line indicates the beginning of the curve and the blue vertical line indicates the end of the curve. The 146 best fitted B-spline curves are plotted in different colors. The solid dark line is the average speed profile calculated from the 146 functional observations. The confidence interval of the mean speed profile is shown as the dashed black line. This plot is a useful technique to summarize the averaged behavior from a group of time series data and examine how drivers change vehicle speeds over time and locations.



**Figure 6.2 Plot of 146 speed profiles on the same curve**

Another popular technique with functional data analysis is the use of derivative information. It is particular useful for some of the SHRP2 NDS variables, because the derivative

of vehicle dynamics often has some meaning explanations. For example, the first derivative of vehicle speed indicates vehicle acceleration. A sample vehicle speed profile and the calculated first derivative of the vehicle speed are shown in Figure 6.3. Examine vehicle acceleration information is a useful way to illustrate how drivers change speeds before, during, and after the curves. In this case, the drivers started to slow down for the curves at 400 feet before the curve PC, and the deceleration rate increased as they moved closer to the beginning of the curve. The maximum deceleration rate occurred at approximately 30 feet before the curve PC. The acceleration changed from negative to positive at middle of a curve, which means the drivers started to accelerate back to tangent speeds after passing half of the curve. This example illustrated the use of derivative information in functional data analysis and found it was a very useful way to examine drivers' behavior as a continuous process.



**Figure 6.3 Plot of vehicle speed (top panel) and the first derivative of the vehicle speed (bottom panel)**



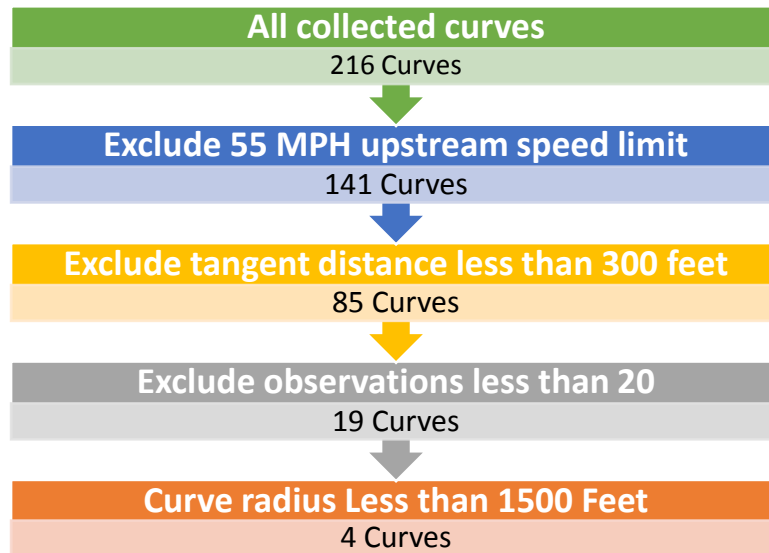
Additionally, it is possible to identify different patterns of driver behavior by conducting functional principal component analysis (FPCA). Similar to multivariate principal component analysis, FPCA examines the major modes of variations in a group of functional data. It is a useful way to explore different driving patterns from a large group of repeated time series measurements. This technique will be discussed in Section 6.4.6.

Overall, this chapter has three objectives. The first objective is to develop the best fitted B-spline functions from the observed discrete time series data, and summarize the mean speed and confidence interval for a group of functional data. The second research objective is to calculate the first derivative of vehicle speed and examine how drivers decelerate before the curves. The third research objective is to identify the major driving patterns using the functional principal component analysis. This dissertation sets an example about how to understand drivers' behavior by analyzing time series data using functional data analysis. The same methodology could be applied to examine drivers' behavior on other types of roadways in future studies

## 6.2 Data Description

This chapter focuses on the analysis of time series data collected from the SHRP2 NDS. The data acquisition system in the SHRP2 NDS collected hundreds of time series variables simultaneously. It is not feasible to analyze all the time series variables in this dissertation, so only the vehicle speed data was analyzed in this study. Speeding on curves has been identified as one of the most important contributing factors to roadway departure crashes on curves. It is important to understand how drivers manage vehicle speed and react to the curves. Hence, this chapter focuses on the analysis of vehicle speeds data on rural two-lane curves. The data used in chapter 6 went through the same data quality assurance process as discussed in Chapter 5.

A total of over 200 curves was collected in this dissertation. FDA usually involves large amount of model building procedures, so it is very time consuming to analyze the observations on all curves. Therefore, it was determined to conduct functional data analysis on several sample curves. The procedure to select the sample curves is shown in Figure 6.4.



**Figure 6.4 The criteria used to select sample curves**

The first criteria excluded the curves with 55 MPH upstream speed limits, because the analysis in chapter 5 showed the drivers rarely reduced vehicle speeds for curves on the 55 MPH roadways. Those curves do not contain the features of our interests. Therefore, the FDA only focuses on roadways with 45 MPH upstream speed limits. The second criteria excluded the curves with restricted upstream tangent distance. The overall goal of this study was to understand how drivers react to different curves, but sometimes the curves were located close to each other, so the tangent speed could be influenced by the adjacent curves. Therefore, any curves with upstream tangent distance smaller than 300 feet were excluded from the analysis. The third criteria excluded the curves with small number of observations. Any curves with less than 20

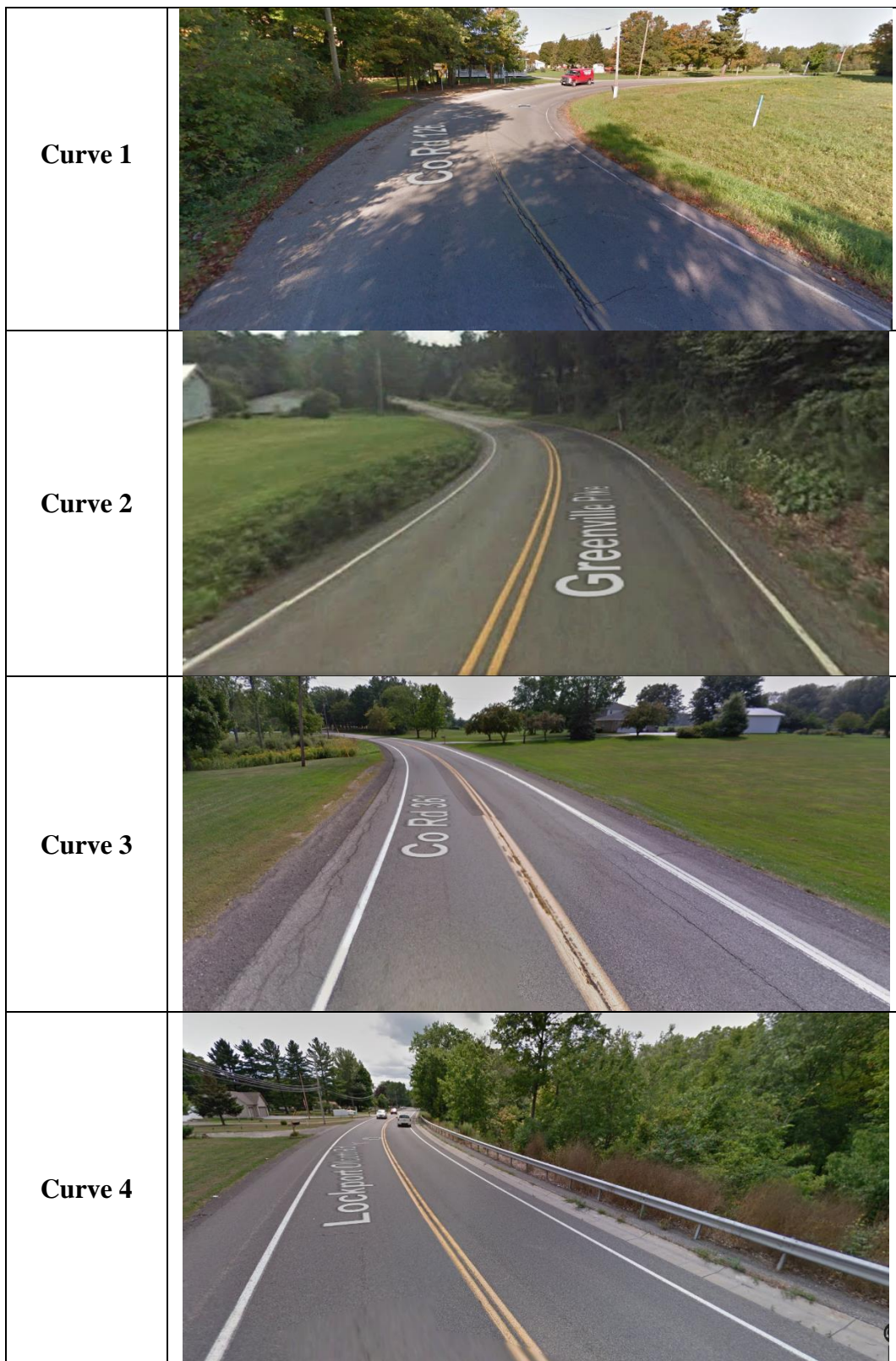
observations were excluded from the study. The last criteria only included curves with radius between 0 and 1500 feet. Previous analysis found the drivers only had significant speed reduction on the curves with radius roughly between 0 to 1500 feet. Finally, four example curves were chosen in this study.

The characteristics of the four sample curves are shown in Table 6.1. The curves radius ranged from 117 feet to 1,288 feet, and the curve length ranged from 186 feet to 584 feet. The tangent distance for all curves were greater than 300 feet, so there was no influence from adjacent curves or intersections. There were at least 28 observations on each curve. Two of them had left turn direction and the other two had right turn direction.

**Table 6.1 List of curve characteristics for the sample curves**

<b>Curve ID</b>	<b>Radius (Feet)</b>	<b>Length (Feet)</b>	<b>Curve Direction</b>	<b>Tangent Distance (Feet)</b>	<b>Number of Observations</b>	<b>Number of Drivers</b>
1	117	186	Right Turn	6875	95	8
2	361	276	Left Turn	392	28	1
3	828	568	Left Turn	11256	27	11
4	1288	584	Right Turn	3160	42	3

In order to better understand the characteristics of the selected curves, the Google Street View were taken from Google Map as shown in Figure 6.5. The pictures were taken from the beginning of the curves. It was concluded that the four selected curves had reasonable similar features, except for curve radius. This study examined how drivers reacted to the curves with different radius. The FDA package in R software was used to fit the models.



**Figure 6.5 Google street view of the four example curves (Google Map, 2014)**

### 6.3 Methodology

This section introduces the methodology for functional data analysis. Section 6.3.1 discusses how to convert discrete time series data to functional data. The development of functional data involves several steps, including choosing the number of basis function, selecting the number of knots, and setting the smoothing penalty parameter. Section 6.3.2 discusses how to calculate the mean, the standard deviation, and the first derivative of the functional data. Section 6.3.3 briefly describes the steps to conduct principal component analysis for functional data. The functional principal component analysis is mainly used to discover the underlying patterns in a group of functional observations.

#### 6.3.1 Convert the discrete time series data to functional data

The first step of functional data analysis is to find a continuous function  $x(t)$  to represent the discrete time series data  $y_1, y_2, \dots, y_n$  at any given time  $t$ . For any discrete observations  $y_i$  over time  $t_i$ , the data can be expressed in formula 6.1

$$y_i = x(t_i) + \varepsilon_i \quad (\text{Formula 6.1})$$

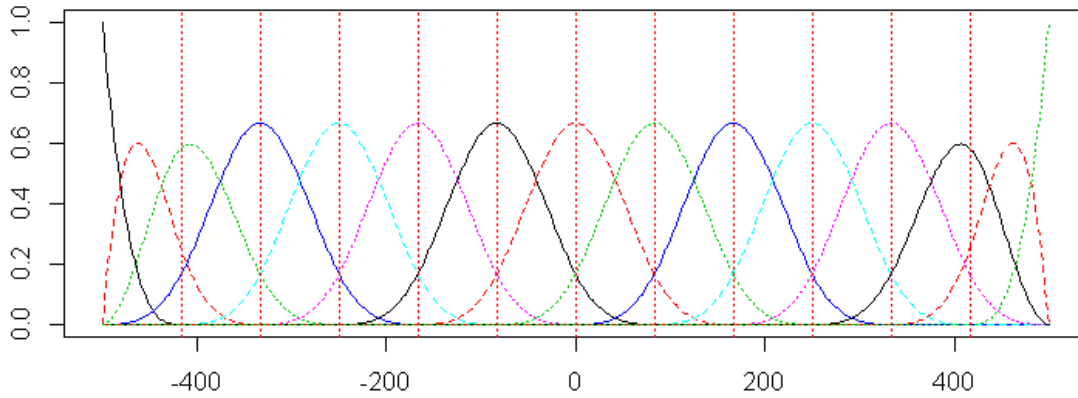
Formula 6.1 is an important assumption for function data analysis, which means the discrete observations  $y_i$  is a function of an underlying smooth process  $x(t_i)$  plus noises  $\varepsilon_i$ . Furthermore, the smoothing process  $x(t)$  can be expressed as a function of basis system  $\Phi(t)$  as shown in Formula 6.2.

$$x(t) = \sum_{j=1}^K C_j \phi_j(t) = \Phi(t)C \quad (\text{Formula 6.2})$$

Where  $\Phi(t)$  is the predefined basis system and  $C$  is the coefficients matrix for observation at time  $j$  from 1, 2, 3,...,  $K$ . The basis function  $\Phi(t)$  here can be expressed in several methods, including Fourier basis and B-spline basis. The Fourier basis fits better for periodic data and has excellent computational properties. The B-spline basis is flexible and appropriate for most types of data and the constraints are easily defined. In this dissertation, the B-spline was used as basis function for developing functional data, and the Fourier basis should have similar results.

### ***Basis Expansions***

The B-spline function is essentially piecewise polynomials and defined by two properties: the location of knots and the polynomial functions. The B-spline method first creates a number of knots to divide the time domain into equally spaced subintervals. The polynomial segments are fitted within each subinterval and the polynomial segments are required to be smoothly connected at the knots. The highest power of the polynomial function is called degree and the spline with degree of three is often used to make sure the first derivative and the second derivative are smoothly connected at the knots. The number of required basis function equals to the sum of polyline order  $m$  and the number of interior knots (Ramsay et al., 2015). The scree plots used to choose the optimal number of basis functions are shown in the Figure C.1 (Appendix C). Additionally, the created basis functions are plotted in Figure C.2 (Appendix C).



**Figure 6.6** The 15 spline basis functions defined over the interval  $(-500, 500)$  by 11 interior knots. The polynomial segments has order four polynomials. The polynomial values and its derivatives were required to be smoothly connected at the interior knots.

### *Smoothing Penalties*

The overall goal of fitting a B-spline function is to minimize the sum of squared errors between the observed value  $y_i$  and the estimated basis function as shown in Formula 6.3. The estimation process is essentially an ordinary least-squares estimate problem.

$$SSE = \sum_{i=1}^n (y_i - x(t_i))^2 = \sum_{i=1}^n (y_i - \Phi(t_i)C)^2 \quad (\text{Formula 6.3})$$

However, the method could potentially over fit the data and it is important to make sure the fitted line are smoothing. Hence, another term is added into Formula 6.3 to control the curve smoothness and now the estimation method is called penalized SSE (PENSSE) as shown in Formula 6.4.

$$PENSSE = \sum_{i=1}^n (y_i - \Phi(t_i)C)^2 + \lambda J[x] \quad (\text{Formula 6.4})$$

Where  $J[x]$  is a measure of the roughness of the fit and  $\lambda$  is a tuning parameter. The roughness measurement  $J[x]$  is often a measure of curvature of the fitted line, which

is  $\int [D^2x(t)]^2 dt$ . Now the penalized squared error becomes Formula 6.5, so that the coefficient matrix  $C$  can be solved from the formula by minimizing PENSSE.

$$PENSSE = \sum_{i=1}^n (y_i - \Phi(t_i)C)^2 + \lambda \int [D^2x(t)]^2 dt \quad (\text{Formula 6.5})$$

When  $\lambda$  gets larger, the roughness of the fit will be penalized heavier and the fitted line will be smoother and linear. When  $\lambda$  gets smaller, the roughness of the fit will be less penalized and the fitted line will fit closer to the actual observations. An appropriate tuning parameter should be selected to ensure the fitted function does not over fit the noises, and still capture the interesting features of the curve. Several cross validation methods were proposed to choose the optimal tuning parameter  $\lambda$ , such as ordinary cross validation (OCV), generalized cross validation (GCV), AIC and BIC, etc. The GCV is used choose the optimal smoothing parameters  $\lambda$  in this dissertation. The GCV can be calculate as Formula 6.6.

$$GCV(\lambda) = \left( \frac{n}{n-df(\lambda)} \right) \left( \frac{SSE}{n-df(\lambda)} \right) \quad (\text{Formula 6.6})$$

The scree plots used to choose optimal tuning parameters are shown in the Figure C.3 (Appendix C).

### 6.3.2 Calculate the mean, standard deviation, and the derivatives for a group of functional data

This section describes the methods for calculating the mean, standard deviation, and the derivatives for a group of functional data.



### ***Mean of Functional Data***

The functional mean and functional standard deviation are useful tools to examine the variations within a group of repeated time series observations. The functional mean is the pointwise average for a group of functional data. It represents the mode of variation shared by most curves. Let  $x_i, i=1,2,\dots,N$ , represents a set of fitted spline functions to the data. The sample mean can be simply expressed as  $\bar{x}(t)$  in Formula (6.7).

$$\bar{x}(t) = \frac{\sum x_i(t)}{n} \quad (\text{Formula 6.7})$$

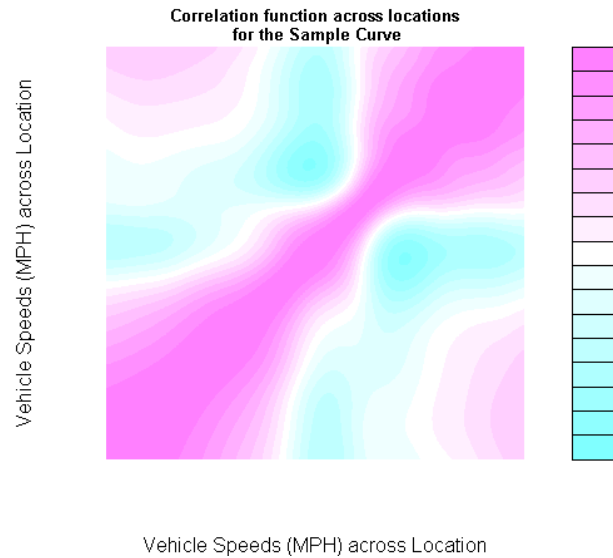
### ***Confidence Interval of Functional Data***

The variance-covariance matrix of the fitted value can be expressed as  $Var[\hat{y}] = \Phi C \Sigma C^T \Phi^T$ . The pointwise confidence interval can be derived from the variance-covariance matrix, and it is an effective way to examine the variability of data over the time domine.

$$\hat{y}(t) \pm 1.96\sqrt{Var[\hat{y}(t)]} \quad (\text{Formula 6.8})$$

### ***Covariance of Functional Data***

The covariance for a group of curves summarizes the dependence of records across different argument values, which is often referred to time  $t_i$ . The covariance can be expressed as  $cov_x(t_1 t_2) = (N - 1)^{-1} \sum_{i=1}^N \{x_i(t_1) - \bar{x}(t_1)\} \{x_i(t_2) - \bar{x}(t_2)\}$ . The variance-covariance is illustrated on a contour plot in functional data analysis. The diagonal running from lower left to upper right in the contour contains the unit values which indicates the correlation between the  $t_i$  and itself. As moving away from the diagonal line in the orthogonal direction for distance  $\delta$ , it indicates the correlation for the time pair  $((t - \delta, t + \delta))$ .



**Figure 6.7 Contour plot of correlation function across locations for vehicle speeds**

### ***Derivative Information of Functional Data***

Another popular application of functional data analysis is the use of derivative information. Calculating the derivatives is an effective way to examine the variations of the curves, which indicates the rates of change. The first derivative of the functional data  $x(t)$  is shown in Formula 6.9. The notion of  $D^1$  means the first derivative of function  $x(t)$ . The notion of  $L$  represents the differential operator  $L=D^1$  to the function  $x(t)$ . The  $L x(t)$  is called a forcing function.

$$L x(t) = D^1 x(t) \quad (\text{Formula 6.9})$$

Phase plane plot is often a useful way to investigate energy transfer and dissipating by plotting the relationship between the functional data and its derivatives. It will be discussed in details in Section 6.4.5.

### 6.3.3 Functional principal component analysis

Functional Principal Component Analysis (FPCA) can be used to identify the strongest and most important modes of variations with in a group of curves. It explains the underlying patterns in the data and provides different ways to look at the covariance structure. To solve the problem mathematically, the FPCA can be defined as the search for a set of mutually orthogonal and normalized weight functions  $\xi_m$ . Scree plot is used to find the number of principal components that contribute the most to the variance of the data. Sometime, VARIMAX method can be applied to attain an interpretable explanation of the dominant modes of variation.

In general, FPCA problem is equivalent to the numerical problem of solving matrix Eigenanalysis. Functional PCA is similar to multivariate PCA except that the summation sign is changed to integration sign. The functional principal components can be treated as a set of orthogonal basis functions to explain the variance as much as possible at each step. Each weight function defines the most important mode of variation in the curves. Smoothing algorithm is often applied on the estimated functional principal components, which is also called regularization. The goal of regularization is to remove the roughness in the raw PC curves. The steps to build regularized FPCA algorithm is briefly explain as follows:

**Step 1:** Build smooth function for the observed data. Expand the observed data  $x_i$  with respect to

the basis  $\phi$  to obtain coefficient vectors  $c_i$ . The simultaneous expansion of all N curves

can be expressed as  $x = C\phi$ . The variance-covariance function is  $v(s, t) =$

$$N^{-1} \sum_i^N x_i(s)x_i(t)$$

**Step 2:** The goal is to solve a set of mutually orthogonal and normalized weight functions by solving  $\int v(s, t) \xi(t) dt = \rho \xi(s)$ , where  $\rho$  is eigenvalues and  $\xi(s)$  is eigenfunctions of the variance covariance function.

**Step 3:** For the first PC, the main objective is to find a set of  $\xi_i(t)$  that maximizes

$\text{Var}[\int \xi_i(t) x_j(t) dt]$ , which is also subjected to the unit sum of square constraint.

**Step4:** The second PC should be orthogonal to the previous PC, so that the new information could be revealed. Therefore, another orthogonal constraint is added as  $\int \xi_i(t) \xi_l(t) dt = 0$  and  $\int \xi_l^2(t) dt = 1$ .

**Step 5:** It is optional to use the VARIMAX method to rotate the principal components and make it easier to interpret the PCs.

**Step 6:** Apply the smoothing operator  $S'$  to the resulting eigenvectors  $u$ . For example, if the functional PCA is too rough, the term  $\|D^2 \xi\|^2$  can be panelized so that  $\xi$  satisfies modified eigenequation  $\int v(s, t) \xi(t) dt = \rho[\xi(s) + \lambda D^4 \xi(s)]$ .

**Step 7:** Transform back to find the principal component function  $\xi$  with  $\xi(s) = \sum_v y_v \phi_v(s) = y' \phi(s)$

In summary, the theoretical background of functional data analysis is discussed in this section. The application of functional data analysis on the SHRP2 NDS data is shown in the following section.

## 6.4 Analysis of Time Series Data using Functional Data Analysis

The application of functional data analysis on the SHRP2 NDS data is presented in this section. Section 6.4.1 plotted the discrete raw data on the four sample curves. Section 6.4.2 plotted the best fitted spline functions for each curve. The mean and confidence interval for a group of repeated observations were also plotted on the same graph. Section 6.4.3 calculated the first derivative of the speed data, which was vehicle deceleration. Section 6.4.4 examined how drivers reduce their speeds for the curves. Another interesting way to examine vehicle dynamics on curves was the use of phase plane plot in Section 6.4.5. Finally, functional principal component analysis of vehicle speeds was discussed in Section 6.4.6. It showed the functional PCA was a very useful tool to discover the underlying patterns in a group of time series measurements.

### 6.4.1 Plot of raw speeds

First of all, the discrete time series data of vehicle speeds on the four curves are plotted in Figure 6.8. The x-axis is the distance to the beginning of a curve, which is also known as point of curvature (PC). The y-axis is the vehicle speeds in MPH. The discrete speed data was plotted on the four example curves. The beginning of the curves was labeled as the red vertical line, and the end of a curve was labeled as the blue vertical line. In general, the curve radius for Curve 1 is the smallest (117 feet) and there is clearly a reduction in speeds in curve 1. As the curve radius increases in Curve 2, Curve 3, and Curve 4, the speed reduction effect was less significant. The Curve 4 almost had no speed reduction. The curve radius, curve length, and turning direction are also labeled on the plot.

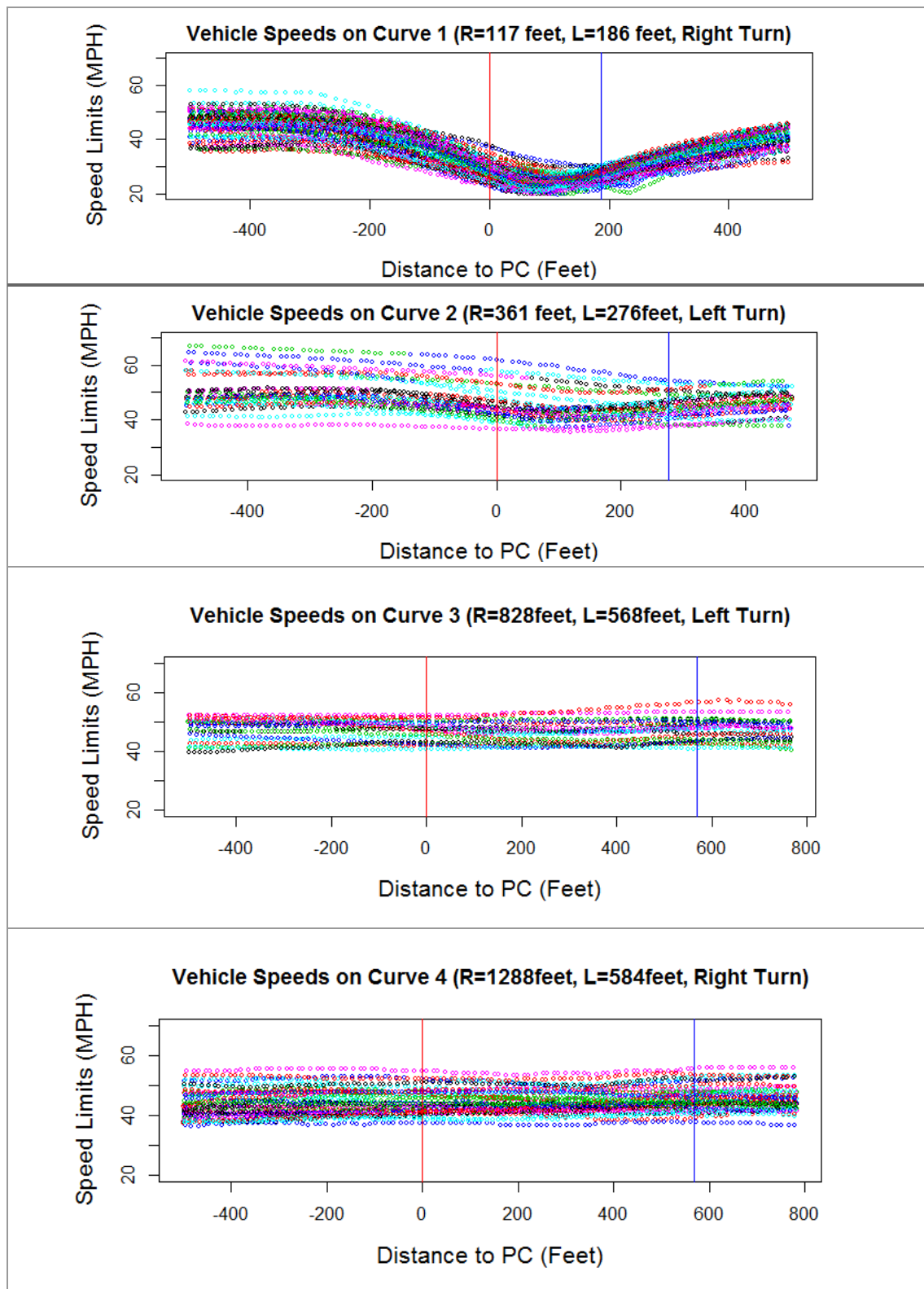
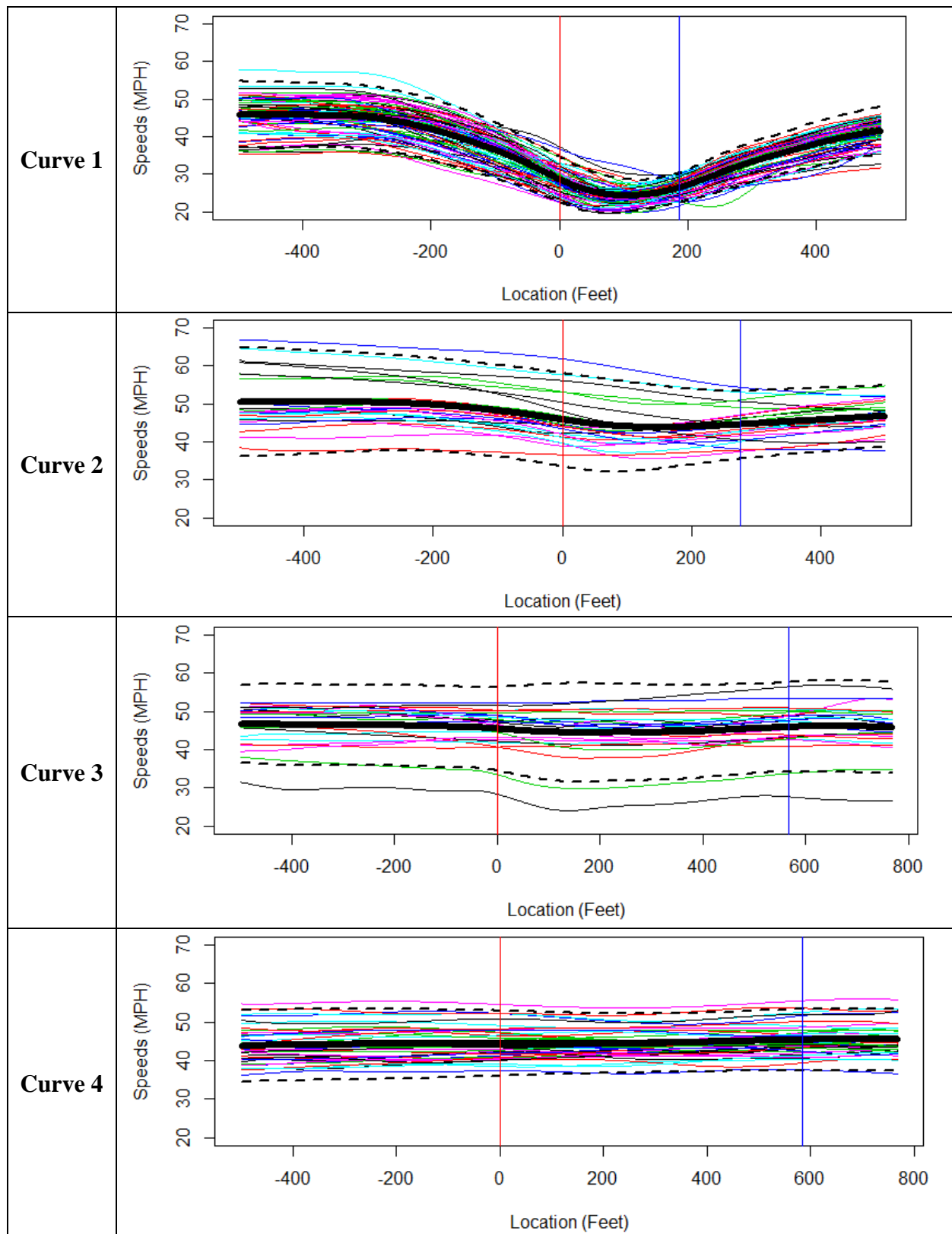


Figure 6.8 Plot of raw speeds on the curves (R=Radius, L=Length)

#### 6.4.2 Plot of fitted functional data, mean speed, and the confidence interval

The smoothed B-spline function was fitted for each driving trace on the same curve. The plots of best fitted functional data on the discrete raw data for the four curves are shown in Figure C.4, C.5, C.6, and C.7 in Appendix C. The most frequently used analysis in functional data analysis was to identify the average behavior from a group of functional observations. In this case, the average vehicle speeds on curve 1 was showed as the solid black line on Figure 6.9. The beginning (PC) point and ending (PT) point of a curve were labeled as the red color and blue color on the graphs. The pointwise 95% confidence intervals were also plotted as the black dashed lines on the graphs. It indicated the range of vehicle speeds at different locations on the curves.

The plot of vehicle mean speeds and the 95% confidence interval revealed important information regarding drivers' speed behaviors on the curves. The drivers were found to reduce their vehicle speeds significantly on curve 1 which has the smallest radius. Curve 2 had larger curve radius than curve 1 and less significant speed reduction was observed on curve 2. Curve 3 had very small speed reduction, and no obvious speed reduction was observed on Curve 4. Additionally, the confidence interval also revealed important information about the variability of vehicle speeds on the curves. For example, the confidence interval was wider on the tangent roadways on curve 1, but it became narrower inside the curve. It indicated the drivers had more freedom to choose their operating speed on tangent roadways, but the curve radius limited the vehicle speeds into a small range of vehicle speeds inside the curves. However, the curve 4 almost had constant confidence interval before and inside the curve, which indicated the curve had little effect on the variability of vehicle speeds for curve 4.



**Figure 6.9 Plot of mean speeds and 95% confidence intervals on the curves**



In order to better understand the driver behavior on the four example curves, the summary statistics of the mean speed profile is shown in Table 6.2. The tangent speed is summarized at 400 feet before the beginning of a curve. The speed reduction is calculated as the difference between tangent speed and curve midpoint speed. The width of the 95% confidence interval is also shown on the tangent segment and the curve segment.

The average tangent speeds was found to be different on the four curves. The curve 1, 3, and 4 had tangent speed around 45 MPH, which was very close to the posted speed limits. However, the curve 2 had tangent mean speed at 50.46, which was approximately 5 MPH above speed limits. The speed reduction for the curves decreased as curve radius increased. Curve 1 had highest speed reduction at 21.39 MPH, but curve 4 had speed reduction close to 0. The width of the 95% confidence interval was also calculated for each curve. The confidence interval on the tangent segment was between 8.608 to 11.130 MPH. The width of the confidence interval at midpoint of the curve is shown in the last column. The width of the confidence interval is only half of the confidence interval on the tangent segment for curve 1. The confidence intervals between the tangent and midpoint were similar for curve 2, 3, and 4.

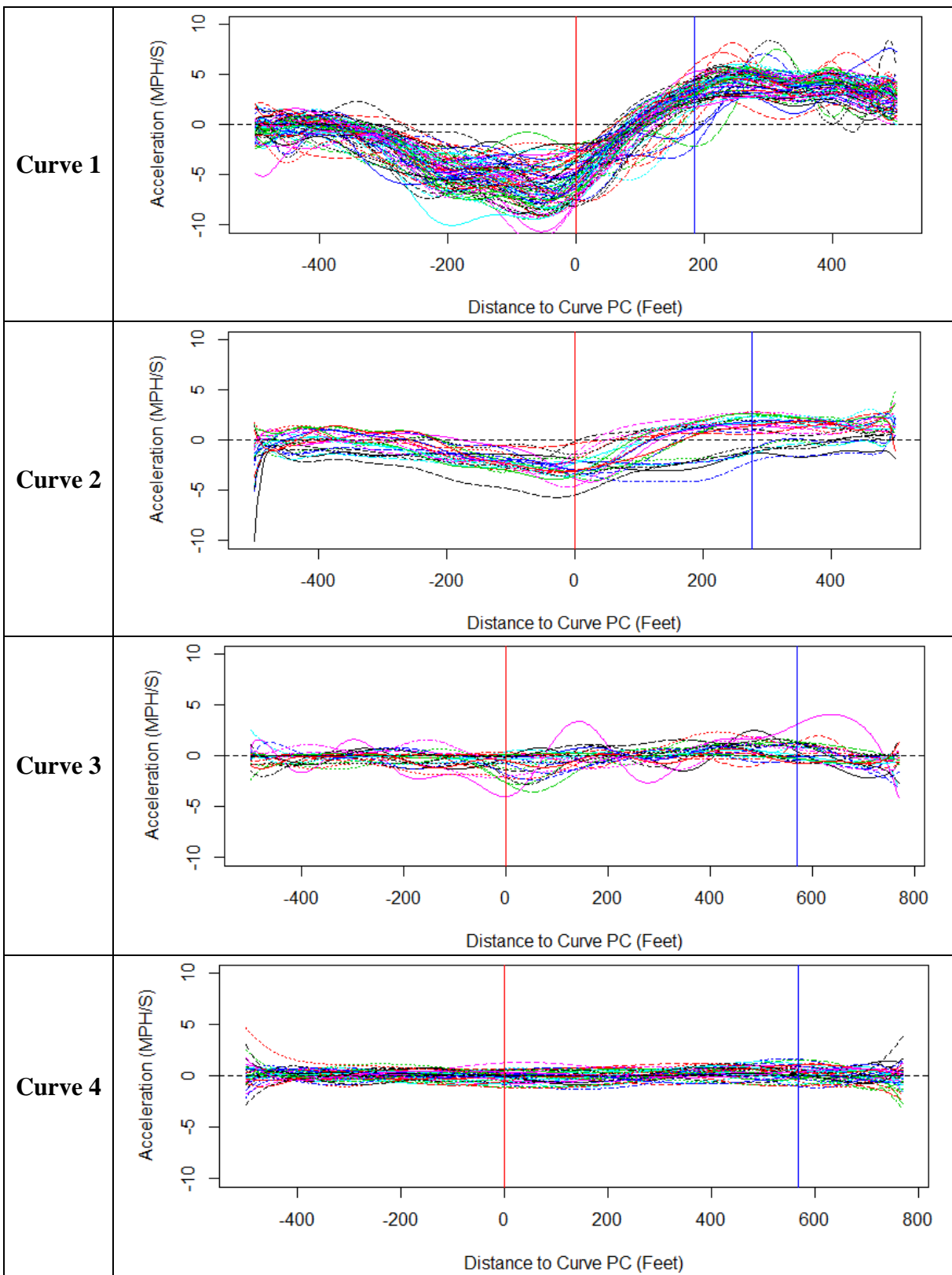
**Table 6.2 Summary statistics of the mean speed profiles**

<b>Curve ID</b>	<b>Tangent Speed</b>	<b>Curve Midpoint Speed</b>	<b>Speed Reduction</b>	<b>Width of Confidence Interval on Tangent</b>	<b>Width of Confidence Interval on Curve</b>
Curve 1	45.73	24.34	21.39	±8.608	±4.548
Curve 2	50.46	43.84	6.62	±13.674	±11.130
Curve 3	46.52	44.53	1.99	±10.638	±12.538
Curve 4	44.07	44.55	-0.47	±9.216	±7.578

### 6.4.3 Examine the deceleration profile on the curves

The first derivative of vehicle speed can be calculated from the functional data and the derivative information is useful to examine how drivers reduced their speeds for the four curves if at all. The first derivative was calculated with the methodology described in Section 6.3.2. The vehicle acceleration for each driving trace was calculated from the functional data and plotted in Figure 6.10. First of all, the acceleration on curve 1 had strong S-shape curves. The drivers decelerated vehicle speeds as they moved closer to the beginning of the curve 1 and reached the maximum deceleration rate at roughly 50 feet before the curve. Once the driver passed the beginning of the curve 1, the magnitude of deceleration reduced and the deceleration turned into acceleration after they passed half of the curve 1. The drivers increased their speeds back to tangent speeds after the end point of the curve (PT). However, this phenomenon was less significant on Curve 2. Some outliers were also observed with different deceleration behaviors. The vehicle deceleration on Curve 3 was not as significant, but one driving trace was found to be an outlier than other observations. The deceleration profiles were almost straight horizontal lines with deceleration equaled to zero on Curve 4. It showed the drivers did not make speed changes at all.

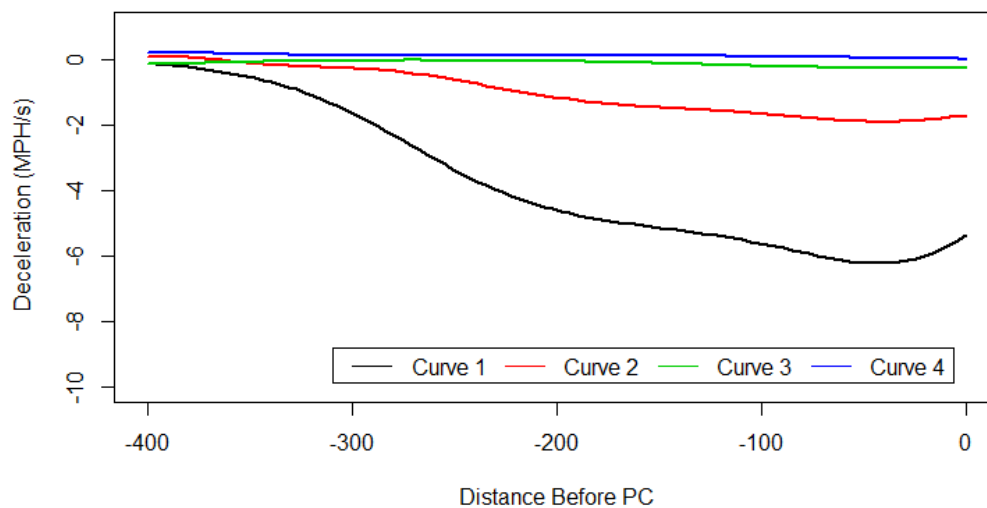
In summary, the derivative information was proved to be a useful way to understand drivers' interactions with different curves as continuous processes. However, the magnitude of deceleration was less significant as the curve radius increased. For curve 4 with radius at 1288 feet, there was no significant deceleration on the curve. Therefore, it was another evidence that the drivers did not reduce speed for the curves with radius approximately larger than 1000 feet.



**Figure 6.10 Plot of vehicle acceleration on the curves**

#### 6.4.4 Vehicle deceleration profile before the curve point of curvature

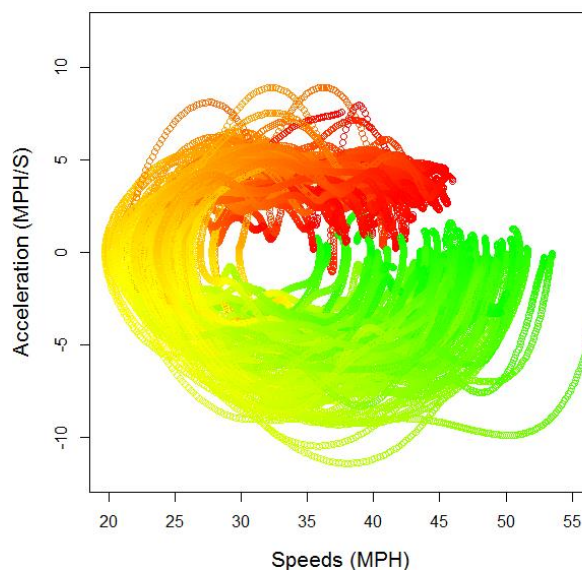
In order to further investigate how drivers reduced their speeds and reacted to the curves, the plot of the averaged deceleration profile before curve PC was shown on Figure 6.11. In general, the deceleration profiles became flatter when curve radius increased. It is clear that the curve 1 with the smallest radius had the largest deceleration rate. The magnitude of deceleration on curve 1 also increased as they move closer to the beginning of the curve. However, it seemed the deceleration profile has two stages with different slopes. The first deceleration slope was from -400 feet to -250 feet, which might reflect the drivers' speed reduction when they first saw the curve. The second stage was from -250 feet to -30 feet, which indicates the drivers' "fine tuning" stage to adjust vehicle speed to enter the curve. However, this phenomenon was only observed on this curve only and more curves should be included to confirm the speculation. The deceleration profile of Curve 2 was relatively flatter to Curve 1. The deceleration profile was almost flat on curve 3. The deceleration profile on curve 4 was nearly a straight line, which indicated no deceleration for the curve 4.



**Figure 6.11 Plot of deceleration profile before the curve PC**

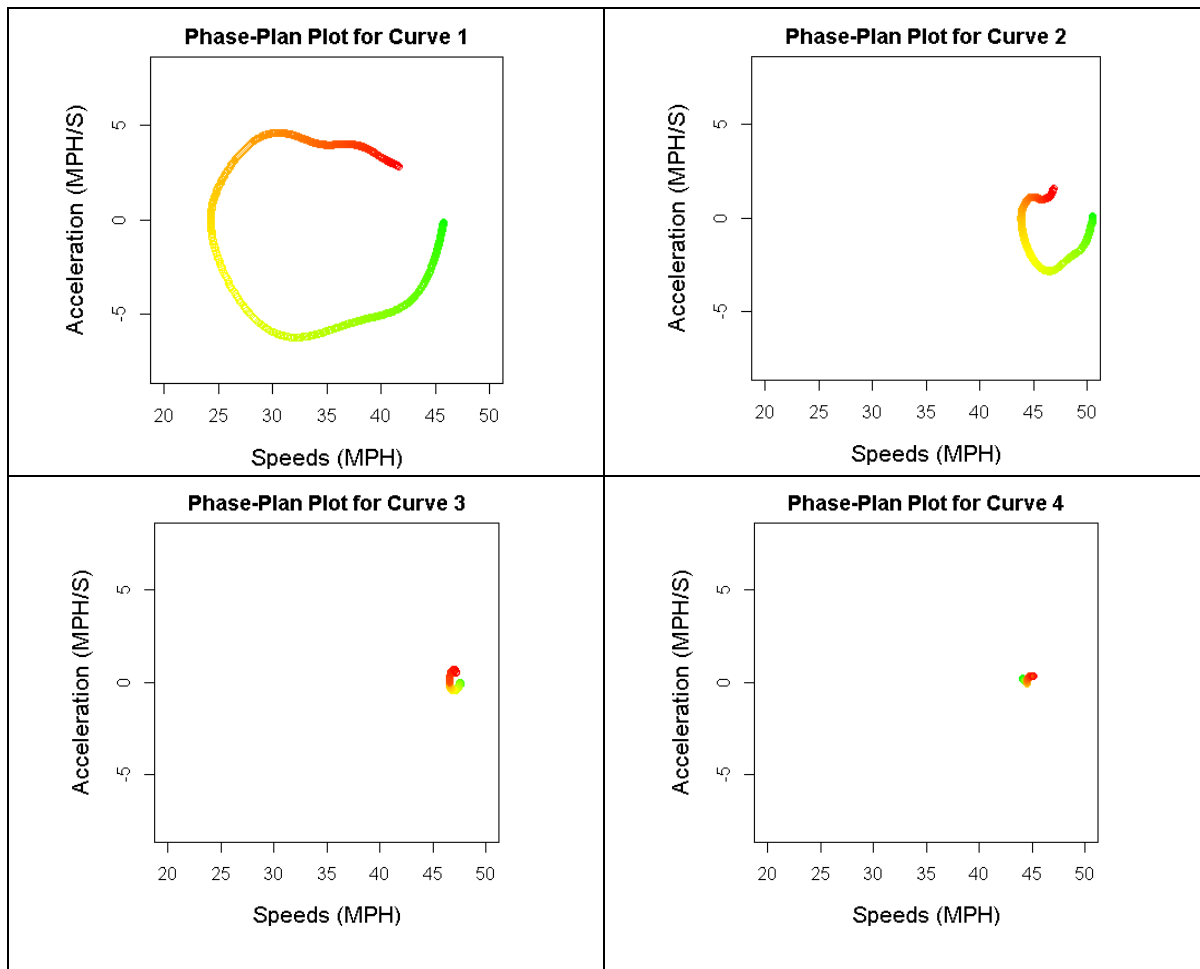
### 6.4.5 Phase-plan plot of vehicle dynamics

Phase plane plot is a useful visualization method to understand the energy transfer and dissipating in the curve negotiation process. The x-axis represents the vehicle speeds on the curves. The y-axis is the vehicle acceleration on the curve. The time dimension of the driving process is expressed by the change in colors. The events started with the green color and ends with the red color. The shape of the phase plane plot for Curve 1 had circular shape as shown in Figure 6.12. The bottom half of the phase plane plotted indicated the deceleration phase. The vehicles reached the minimum speed at the left side of the circle with acceleration equaled to zero. After the vehicle passed its minimum speed, the acceleration value turned into positive and the vehicle started to accelerate and get back to the tangent speeds. Many patterns of vehicle dynamics were found on the curve 1. Some driving traces had larger deceleration value than other drivers. There were also a large range of minimum vehicle speeds in the curve. The size of the circle indicated the change in vehicle dynamic energy. In conclusion, phase plane plot was a useful way to visualize the energy transfer and dissipating on curves.



**Figure 6.12 Phase plane plot for the 95 driving traces on curve 1**

Furthermore, the averaged profile for each curve is shown in Figure 6.13. It represented the average drivers' vehicle dynamics on each curve. It is interesting to see the curve 1 involved large speed reduction and also involved higher deceleration and acceleration in the negotiation process. The curve 2 had relatively smaller circle, which indicated less speed reduction. Smaller acceleration and deceleration involved in the curve negotiation process. The curve 3 had even smaller circles. The curve 4 almost had a single point, because it was not associated with any speed reduction. Above all, phase-plan plot is a useful way to illustrate vehicle dynamics involved in the curve negotiation process.

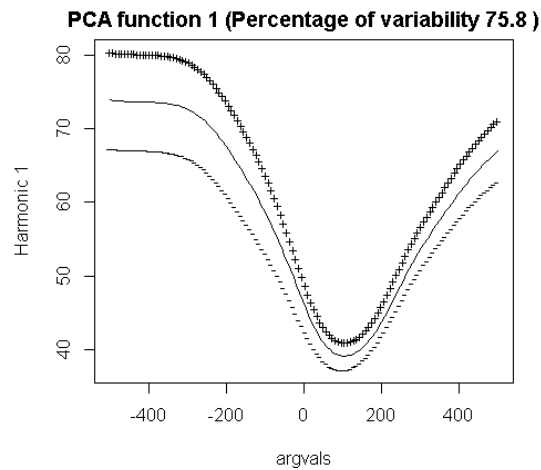


**Figure 6.13 Phase plane plot for the four sample curves**

#### 6.4.6 Functional principal component analysis of vehicle speed

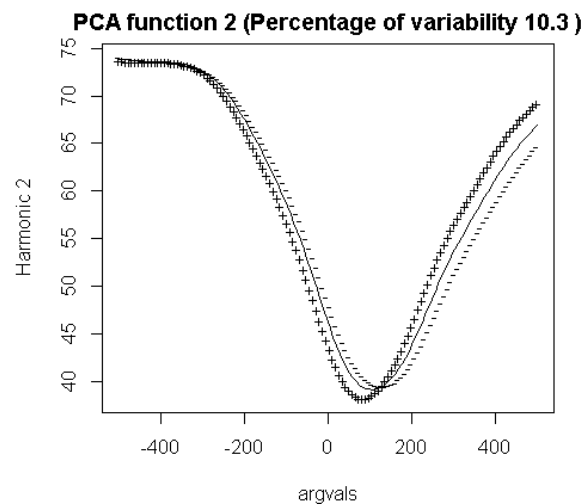
Functional principal component analysis (FPCA) is a data exploratory method to discover the underlying patterns in a group of repeated time series observations measured on the same process. The methodology of FPCA was discussed in details in Section 6.3.3. The FPCA can be applied to each curve, but this section only used curve 1 as the example curve because it had the most interesting features compared to other curves. Scree plot was used to determine the optimal number of principal components kept in the model (the scree plot was shown in Figure C.8 in Appendix C). The first three principal component were kept in the analysis and they explained 94.8% of the variability in vehicle speeds. The functional principal component analysis was performed using the `pca.fd` function in FDA package in R.

A common way to express the variability in each functional principal component is to plot the mean and plus and minus the principal component effect in the component direction. The first principal component found the contrast in vehicle speeds in vertical direction. The first principal components explained 75.8% of the variability in the functional data. The solid line in Figure 6.14 is the mean of the speed profile averaged from the 95 functional observations on curve 1. The line labeled as plus signs on the top indicated that any functional observation closed to this line will have higher scores in the first principal component. In contrast, the line labeled with negative signs on the bottom indicated the curves close to this line had lower values in the first principal component. In other words, the first principal component represented the differences in the magnitude of vehicle speeds on the curve.



**Figure 6.14 The first principal component explained 75.8% of the variability**

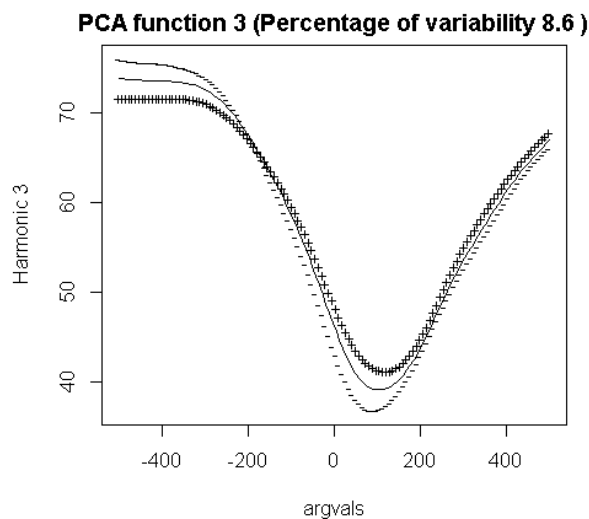
The second principal component explained 10.3% of the variability in the speed profiles. It mainly pointed out the contrast in locations in lateral direction. The driving traces with early deceleration had larger score on the second principal component. The driving traces with late deceleration had smaller score on the second principal component. This principal component differentiated the effects for the drivers who slowed down for the curve early compared to the drivers who slowed down for the curve later.



**Figure 6.15 The second principal component explained 10.3 % of the variability**



The third principal component explained only 8.6% of the variability in the driving traces. This principal component identified the drivers who had larger speed reduction for the curve compared to those who had smaller speed reduction for the curve. The drivers who drove higher speeds on tangent roadways and smaller speeds inside the curve had higher score on this principal component. The drivers who drove lower speeds on tangent roadways and higher speeds inside the curves had lower score on this principal component.



**Figure 6.16 The third principal component explained 8.6 % of the variability**

Overall, three major driving patterns were found in this analysis. The first mode of variation was the magnitude of vehicle speeds. The second major mode of variation was the difference between the drivers who slow down for the curve early and the drivers who slow down for the curve late. The third mode of variation was not as significant, but it indicated the magnitude of speed reduction for the curve. Overall, the FPCA successfully applied on the time series data from Curve 1. Three modes of variations were discovered in the analysis.

## 6.5 Discussion

The objective of this chapter is to analyze the time series data of vehicle speeds on rural two-lane curves. The traditional time series data analysis method was mainly used for forecasting purpose and it was not appropriate for summarizing the features for a group of time series observations. Functional data analysis was a relatively new statistical branch developed in the past twenty years. The method had not been used in transportation research field.

The functional data analysis is most appropriate for analyzing high frequency time series data with a smoothing underlying process. It is useful to summarize a group of time series data measured repeatedly over the same process. For example, the drivers might drive on the same curve for multiple times and it is interesting to summarize the average driving behavior from a group of repeated time series observations on the same roadway segments. Additionally, functional data analysis is very useful to describe complex process which could not be expressed in a simple parametric model. It is also possible to examine the high dimensional data and calculate the correlations between the time series variables. The time series data collected from the SHRP2 NDS project satisfied almost all criteria mentioned above. This chapter applied the functional data analysis to study drivers' speed behavior on rural two-lane curves using the SHRP2 NDS data. Many promising results had been found in this research.

However, functional data analysis involved several steps to convert discrete time series data to functional data. The process included choosing the number of basis function, selecting the number of knots, and setting the smoothing penalty parameters. After the functional data was built from the discrete time series data, several functional statistical analysis methods were used to analyze the created functional observations. The most common type of analyses was

calculating the mean and the confidence interval for a group of driving traces. The derivative information was also available for analyzing the rate of change for vehicle speeds on the curves. Phase plane plot was another way to visualize the relationship between vehicle speed and vehicle acceleration. Additionally, the functional principal component analysis was found to be an important tool to conduct exploratory analysis on the underlying patterns in a group of time series observations.

Although the driving data were collected on 219 curves, the functional data were not applied to all curves. Several criteria were used to select the curves on 45 MPH roadways with at least 20 observations on the curve. The chosen curves had at least 300 feet upstream distance. Finally, only four rural two-lane curves were chosen for functional data analysis in this dissertation.

First of all, the mean speed and confidence interval were calculated and presented in section 6.4.2. The four example curves had curve radius between 117 feet to 1288 feet. Significant speed reduction was found on curve 1, which had smallest radius. The speed reduction effect was less obvious on curve 2. Only slight speed reduction was observed on curve 3. No speed reduction was found on curve 4 with the largest radius.

Second, the first derivative of vehicle speed was discussed in section 6.4.3. The drivers were found to start decelerating at roughly 400 feet before the curves. The deceleration rate increased as the vehicles moved closer to curve PC. After driving half way through the curve, the drivers started accelerating back to tangent speed on curve 1. However, the deceleration profile was almost a straight line on curve 4. The phase plane plot was also found to be an effective way to visualize the vehicle dynamics on the curves.

The third analysis discussed the application of functional principal component analysis on curve 1. Similar to multivariate principal component analysis, FPCA is an important tool to investigate the underlying patterns in a group of time series observations. The developed principal components revealed the important underlying patterns. The first principal component explained 75.8% of the data and it indicated the overall magnitude of vehicle speeds on the curve. The second principal component explained 10.3% of the variability and it represented the drivers who reacted to the curves early compared to the drivers who reduced speed later. The third principal component explained the magnitude of speed reduction for the curve. Overall, the first three principal components explained 94.5% of the information in the driving data. It successfully revealed the important underlying patterns in a group of time series driving data. It is recommended to use FPCA to analyze driver behaviors on other curves.

## 6.6 Conclusion

Overall, the functional data analysis method was found to be a very useful and effective way to analyze time series data from naturalistic driving study. It successfully summarized the mean and confidence interval for a group of time series driving data. The derivative information was also available for analysis. Phase plane plot was a useful visualization technique to understand the vehicle dynamics on different curves. The FPCA was found to be a useful tool to discover the underlying trend the in a group of time series driving data. This dissertation demonstrated the functional data analysis as an important analytical tool to examine time series data in naturalistic driving study.

However, the study had some shortcomings. First of all, functional data analysis was a relatively complicated statistical method which required several steps to develop functional data

from discrete time series data. Because FDA is a relatively new method, it is not incorporated in many existing software. The analysis in this chapter used the FDA package in R, which required programming skills to apply the method. Second, the sample size included in the dissertation was very limited. Only four curves were included for functional data analysis. The general trend can be identified from the four curves with different radius. It is recommended to incorporate more curves in the analysis. Third, it is recommended to include more variables in the analysis. For example, vehicle lateral position variable could be used to study how drivers maintained their lane positions in the curves. In conclusion, functional data analysis was found to be a useful research tool to examine time series data in naturalistic driving study.

## CHAPTER 7. CONCLUSION

This chapter reviewed the main findings from this dissertation in Section 7.1 and discussed the implications for future research in Section 7.2.

### 7.1 Summary of Major Findings

The objective of this dissertation was to understand drivers' naturalistic driving behavior on rural two-lane curves using the state-of-the-art SHRP2 Naturalistic Driving Study data. Chapter 1 introduced the transportation safety issue and the background of naturalistic driving study. Chapter 2 reviewed the past studies on horizontal curves from three aspects: curve perception, vehicle speeds, and vehicle lateral positions. Chapter 3 reviewed some of the existing naturalistic driving studies and introduced the SHRP2 Naturalistic Driving Study. Chapter 4 analyzed the crashes and near-crashes on rural two-lane curves. It mainly focused on analyzing the contributing factors to the safety critical events using logistic regression model. Chapter 5 focuses on drivers' normal driving behavior on rural two-lane curves. The multivariate analysis method was used to analyze drivers' curve negotiation behavior from vehicle speed and lateral acceleration. Chapter 6 took the challenges to analyze time series data using functional data analysis and evaluated how drivers interacted with curve geometries as a continuous process. FDA was found to be a useful method to analyze time series data from naturalistic driving study. Overall, the SHRP2 Naturalistic Driving Study data helped us understand how drivers interact with vehicles, roadways, and traffic environments. The major findings and implications from this dissertation are discussed in the following sections.

### 7.1.1 Crashes and near-crashes analysis on rural two-lane curves

The objective of the first study was to understand the causes of crashes and near-crashes on rural two-lane curves using the SHRP2 NDS data. A total of 67 safety critical events and 136 baseline events were included in the final model. The preliminary analysis found speeding was a contributing factor to 78% of safety critical events; The engagement in secondary tasks were found in 64% of safety critical events; The wet and icy/snowy surface was also overrepresented in the safety critical events. The logistic regression model was used to predict the binary event outcomes. The model initially considered 24 variables, but only 8 of them were statistically significant in the model. The speeding, wet surface, icy/snowy surface, roadway curb, and visual distractions increased the likelihood of roadway departure events on rural two-lane curves. On the contrary, larger curve radius and paved shoulder decreased the likelihood of roadway departure events on rural two-lane curves.

The odds ratio was calculated for each contributing factor. The likelihood of safety critical event was 2.54 times higher if the drivers were driving 10 MPH above posted speed limit. The wet roadway surface increased the likelihood of roadway departure events by 3.81 times. The icy/snowy surface increased the likelihood by shockingly 34.08 times, but the confidence interval was very wide due to the limited sample size. Drivers' visual distraction also increased the crash likelihood by 3.07 times. Curb increased the likelihood of roadway departure events by 4.71 times. However, the tire-strike events were overrepresented in the dataset, so it is recommended to use lane encroachment as crash surrogate in the future study. Additionally, for every one unit increase in logarithm of curve radius, the likelihood of involving in safety critical events decreased by 0.24 times. The paved shoulder decreased the likelihood by 0.26 times. Another important observations in this study was the interaction between multiple factors. For

example, the most frequently observed interaction effect was speeding on sharp curves with wet or icy/snowy surface. Because of the limited sample size, it was difficult to examine the interaction effect in this analysis, but it is recommended to examine the interaction effect in future research.

In summary, this study was one of the first analyses of crashes and near-crashes using the large-scale SHRP2 NDS data. The state-of-the-art data set allowed researchers to examine many driver behavior variables that were not available in previous studies. This dissertation demonstrated the use of logistic regression to analyze event level crash and near-crash data and the results were interpreted as odds ratios. The findings suggested that the SHRP2 Naturalistic Driving Study provided invaluable information to help understand the role of human factor in crashes and near-crashes.

#### 7.1.2 Multivariate analysis of driver behavior on rural two-lane curves

The objective of the multivariate analysis was to understand drivers' normal curve negotiation behavior on rural two-lane curves using the SHRP2 NDS data. This study included two analyses: lateral acceleration analysis and mean speed analysis. The SHRP2 NDS driving data on rural two-lane curves were identified on ArcGIS and requested from VTTI. The major data sources included time series driving data, curve geometry, driver demographics, and vehicle types. After conducting data assurance, a total of 9,584 observations were included in the final dataset. It is one of the largest driving behavior datasets compared to previous studies.

The first analysis was to understand the relationship among lateral acceleration, vehicle speeds, and curve radius. The lateral accelerations were found to increase exponentially as curve radius decreased. Especially, the lateral accelerations increased quickly as the curve radius was



less than 1000 feet, which indicated higher roadway departure risks associated with these sharp curves. It explained why the safety critical events were overrepresented on the curves with radius less than 1000 feet as shown in Chapter 4. Additionally, the cumulative distribution of lateral acceleration were further examined on different ranges of curve radius. The drivers were found to tolerate higher lateral acceleration on the curves with smaller radius. The 85<sup>th</sup> percentile of lateral accelerations on different curve radius were found from the cumulative distribution functions. The findings could be used by transportation engineers to set the speed limits on curves to make sure the drivers feel comfortable on curves. For future research, it is suggested to investigate how lateral acceleration correlated with lane deviation on curves.

The second analysis was to understand the drivers' speed choice on rural two-lane curves. The analysis only focused on the curves with 45 MPH or 55 MPH upstream speed limits, even though the advisory curve speed limits could be different. The preliminary analysis found the drivers did not reduce their speeds on curves with 55 MPH upstream speed limits regardless of the curve radius. On the other hand, the drivers were found to reduce their speeds on 45 MPH upstream speed limits curves as curve radius decreased. In order to examine the contributing factors to the speed reduction, the linear mixed model was used to predict drivers' mean speed on curves 45 MPH upstream speed limits. The driver ID and curve ID were treated as random effects in the model to account for the interdependence in the observations.

The results found tangent speed, advisory speed limit, and higher logarithms of curve radius were positively correlated to higher mean speeds on curves. The car following and younger drivers were correlated with smaller mean speeds on the curves. Although advisory speed limits was found to be statistically significant in the model, the vehicle speeds only

reduced 1.63 mph for every 10 MPH suggested speed reduction in advisory speed limits. The preliminary analysis also found the drivers did not reduce speed at all on 55 MPH upstream speed limits. Hence, it was concluded that advisory speed limit was not effective to reduce mean speeds on curves. The younger drivers were found to drive slower on curves, which was contrary to our expectation, but it might be due to the lack of experience of driving on curves. The random effect of the drivers were further examined to identify the risky drivers who tend to drive faster than other drivers. The results successfully identified a group of drivers who drove faster than the other drivers on curves.

This analysis is one of the first studies to evaluate drivers' curve negotiation behavior using naturalistic driving study data. This study collected one of the largest multivariate curve driving behavior datasets than previous studies. The speed prediction model could be used to predict vehicles' speed based on the explanatory variables. The linear mixed model also identified the risky driver groups. This study also demonstrated how to batch process a large scale naturalistic driving dataset, which provided important insights for future researchers. However, this study also only focused on the rural two-lane curves. It is suggested to analyze drivers' curve negotiation behavior on other types of curves in future research.

### 7.1.3 Functional data analysis of time series data on rural two-lane curves

Many researchers were interested in understanding and summarizing driver behavior by analyzing the time series data from naturalistic driving study. This dissertation took the challenges to analyze the time series data using a relatively new branch of statistical method called functional data analysis. Since this is a recently developed research method in transportation research community, this study first reviewed the methodology of functional data

analysis. Four sample curves with different curve radius were used in this analysis. Only the vehicle speed data was analyzed in this study as an example. The functional data analysis involved several steps to convert the discrete time series data to functional data. After the functional data were created, the mean and 95% confidence interval for vehicle speed data were summarized for each curve. The average speed from a group of drivers were plotted on each of the four sample curves. The confidence interval were plotted on the same graph to indicate the variability of vehicle speeds on the four curves. This analysis illustrated the drivers' speed behavior on the rural two-lane curves as a continuous process. Curve 1 had the largest speed reduction, and curve 4 had no speed reduction at all. The plot of functional mean and confidence interval was a useful tool to examine how the drivers changed their speeds in relation to the curve geometries.

Another popular FDA technique was to examine the derivative information of the functional data. In this study, the first derivative of vehicle speeds was calculated from the functional data, which indicated the vehicle's deceleration. The plot of first derivative found the deceleration behaviors were different on the four sample curves. Significant decelerations were observed on curve 1 and curve 2. The drivers were found to slow down for the curves at roughly 400 feet before the curves on curve 1 and curve 2. The magnitude of deceleration rates also increased as the drivers moved closer to the beginning of the curve PC. However, the decelerations were less obvious on curve 3. No deceleration was found on curve 4. Phase plane plot was another useful visualization method to examine vehicles' energy changes on the curves. This study conformed to our expectation that the speed reduction was not significant for those curves larger than 1000 feet.

Finally, the functional principal component analysis was used to identify the major driving patterns in a group of time series data on curve 1. The first three principal components were identified in this analysis. The first PC represented the magnitude of vehicle speeds on the curve. The second PC represented the contrast between the drivers who slow down speeds early and the drivers who slowed down speeds later. The third principal component represented the magnitude of the speed reduction between the tangent speed and curve speed. Overall, the functional principal component analysis identified different patterns of drivers' behaviors from a group of time series observations measured on the same curve.

Overall, the functional data analysis was found to be a groundbreaking research tool that allowed researchers to analyze time series data collected from naturalistic driving study. The typical drivers' behavior could be easily calculated and visualized with the functional data analysis methods. The derivative information could provide important insights on the time series data. It is especially useful because it is sometimes more important to study the rate of change, instead of the absolute value of the variable. The functional principal component analysis discovered several important driving patterns on curve 1. A number of visualization tools were available to show how drivers interacted with the roadway features. Functional data analysis was proved to be an effective way to examine how drivers interact with roadway features. This method allowed researchers to examine how a group of drivers reacted to different roadway designs, which attracts great interests from transportation agencies.

## 7.2 Implications for Future Research

This dissertation analyzed the drivers' behavior on rural two-lane roads using the SHRP2 Naturalistic Driving Study. The crash and near-crash analysis on rural two-lane curves

successfully incorporated drivers' behavior into the analysis, and identified the odds ratio for each contributing factor. The results could be used to improve roadway design methods and develop safety countermeasures. Additionally, the multivariate analysis analyzed drivers' normal curve negotiation behavior with vehicle lateral accelerations and vehicle speeds. The relationship between curve radius, lateral acceleration, and vehicle speeds was carefully investigated in this analysis. The findings had important implications for setting curve speed limits, identifying dangerous driver groups, and developing crash warning system. Furthermore, functional data analysis was used to analyze the time series data from naturalistic driving study. Functional data analysis is a relatively new research method that hasn't been used in transportation research community. It was found to be a very useful method to analyze time series data collected from naturalistic driving study and help us understand how drivers interact with roadway features and traffic environments as a continuous process. Overall, the state-of-the-art SHRP2 naturalistic driving study was found to be an important research tool to understand the causes of crashes and how drivers interact with vehicle, roadway, and traffic environments.

Although the SHRP2 NDS data provided unprecedented opportunity to study drivers' naturalistic driving behaviors, there were also many challenges to analyze the SHRP2 NDS data. First of all, the quality of collected data should be carefully examined before conducting any statistical analysis. The common issues included large noises, sensor malfunction, missing data, outliers and omitted variables. Second, the large size of the collected data is an advantage of the SHRP2 NDS data, but it also brought many issues for data storage, data management, and data analysis. Third, the data collected in the SHRP2 NDS included both structured data (e.g. event table data) and unstructured data (e.g. video data). It is a big challenge to mining the diverse dataset and find correlations between different variables. Fourth, there are very limited previous

studies available in this research field. It is not clear which statistical method is most appropriate to analyze NDS data. Lastly, it is often not practical to reduce and clean the data manually. Some levels of programming skills would be greatly helpful to reduce a large dataset in a timely manner. In addition to transportation safety, this study also sheds lights on other aspects of transportation research, including big data, connected vehicles, and autonomous vehicles.

### 7.2.1 Implication for big data research

Due to the recent development in data collection and storage technology, the SHRP2 NDS collected probably one of the largest data set in transportation research community by far. A total of 4 million gigabytes data were collected in the SHRP2 NDS from 3000 drivers over two years. The collected variables were also very diverse, including vehicle sensor data, vehicle network system data, survey data, and video data. This project raised many issues regarding data storage, management, analysis, and visualization. Even analyzing a subset of the SHRP2 data could be a challenge for some of the existing software. For example, the multivariate analysis in this dissertation had 5 million rows and 83 columns of time series driving data. The big data topic is going to be a trend in transportation research, and it is a challenge to uncover the underlying valuable information from the large-scale, diverse, and complex dataset. The SHRP2 NDS is a forerunner for big data research in transportation field.

### 7.2.2 Implication for connected vehicle research

Connected vehicle was emerged as a popular research topic in the past few years. It had the potential to transform the way people travel in terms of safety, efficiency, and mobility. The SHRP2 NDS research and the connected vehicle research shared similar types of data, so that the

research results could benefit from each other. On one hand, the lessons learned from naturalistic driving study could be directly transferred to study drivers' behavior with connected vehicle. On the other hand, the collected driving data from connected vehicles could be used to expand the existing naturalistic driving study dataset. Therefore, the naturalist driving study research and connected vehicle research could benefit from each other.

### 7.2.3 Implication for automated vehicle research

Last but not least, the autonomous driving research could also benefit from the SHRP2 NDS data. An important topic in autonomous driving research is how to teach the autonomous vehicles make decisions in complex traffic environment, and how to create a comfortable riding experience for the passengers. The SHRP2 Naturalistic Driving Study dataset is an ideal dataset providing information about drivers' decision making in almost all aspects of on-road driving. The drivers' behaviors and decisions could be learned from the naturalistic driving study and implemented in autonomous driving scenarios.

Overall, despite of the challenges, the SHRP2 NDS data provided invaluable information to answer a number of critical research questions that could not be studied in traditional research methods. The SHRP2 NDS data provided unprecedented opportunity for researchers to learn the causes of crashes and how the drivers interact with vehicle, roadway, and traffic environments. Naturalistic driving study will lead the traffic safety research in the next decades. The SHRP2 NDS dataset is expected to lead the transportation safety research in the next decades.

## BIBLIOGRAPHY

- Abele, L., & Møller, M. (2011). Relationship between road design and driving behavior: a simulator study. *3rd International Conference on Road Safety and Simulation*, Transportation Research Board, 2011, 16p
- Antin, J; Stulce, K; Eichelberger, L; Hankey, J. (2015). *Naturalistic driving study: descriptive comparison of the study sample with national data*. SHRP 2 Report, Issue Report S2-S31-RW-1, 2015, 45p
- Antin, J., Lee, S., Hankey, J., Dingus, T. (2011). *Design of the in-vehicle driving behavior and crash risk study*. SHRP2 Report. S2-S05-RR-1. Transportation Research Board. Washington, D.C.
- Anderson, I B; & Krammes, R A. (2000). Speed reduction as a surrogate for accident experience at horizontal curves on rural two-lane highways. *Transportation Research Record, Transportation Research Board, Issue 1701*, p. 86-94
- Arbeitsgruppe Verkehrssicherheit (1980). Traffic safety by night, traffic- engineering- related improvements to the highway infrastructure. EJPD, Berne, Switzerland.
- Ayres, G; Wilson, B; LeBlanc, J. (2004). Method for identifying vehicle movements for analysis of field operational test data. *Transportation Research Record, Transportation Research Board, Issue 1886*, 2004, p. 92-100
- Bauer, Karin M; & Harwood, D. W. (2013). Safety Effects of Horizontal Curve and Grade Combinations on Rural Two-Lane Highways. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2398, pp. 37–49.
- Bella, F. (2013). Driver perception of roadside configurations on two-lane rural roads: Effects on speed and lateral placement. *Accident Analysis and Prevention* 50, P 251– 262.
- Bella, F. (2007). Parameters for evaluation of speed differential: contribution using driving simulator. *Transportation Research Record: Journal of the Transportation Research Board, Issue 2023*, 2007, pp 37-43
- Bella, F. (2013). Driver perception of roadside configurations on two-lane rural roads: Effects on speed and lateral placement. *Accident Analysis and Prevention* 50 (2013) pp.251-262
- Bertola, M. A., Balk, S. A., Shurbutt, J. (December, 2012). *Evaluating driver performance on rural two-lane horizontal curved roadways using a driving simulator*. Federal Highway Administration, McLean, VA
- Blincoe, L. J., Miller, T. R., Zaloshnja, E., Lawrence, B. A. (May, 2014). *The economic and societal impact of motor vehicle crashes, 2010*. (Report



- No. DOT HS 812 013). Washington, DC: National Highway Traffic Safety Administration.
- Bonneson, J. A., & Pratt, M. P. (2009). Model for predicting speed along horizontal curves on two-lane highways. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2092, pp. 19-27.
- Bosq, D. (2000). Linear processes in function spaces. theory and applications. Springer-Verlag New York.
- Cacciabue, P. C. (2007). Modelling driver behavior in automotive environments: critical issues in driver interactions with intelligent transport systems. Springer-Verlag New York, Inc. NJ, USA.
- Campbell, K. L. (October, 2012). The SHRP2 Naturalistic Driving Study. *Transportation Research News* 282.
- Camacho-Torregrosa, F. J., Pérez-Zuriaga, A. M. Campoy-Ungría, J. M., García-García, A. (2013). New geometric design consistency model based on operating speed profiles for road safety evaluation. *Accident Analysis & Prevention*, Volume 61, Issue 0, pp 33-42
- Casali, J. G., & Frank, L. H. (1986). Perceptual distortion and its consequences in vehicular simulation: Basic theory and incidence of simulator sickness. *Transportation Research Record* 1059.
- Charlton, S. G. (2007). The role of attention in horizontal curves: A comparison of advance warning, delineation, and road marking treatments. *Accident Analysis and Prevention*, Volume 39, Issue 5, pp 873-885
- Charlton, S. G., & De Pont, J. J. (2007). *Curve speed management*. Land Transport New Zealand Research Report 323.
- Cheung, J. (July 2010). *Horizontal Curve Safety*. Federal Highway Administration. Retrieved on September 23, 2015 at [http://safety.fhwa.dot.gov/roadway\\_dept/horcurves/cmhorcurves/horiz\\_curve.pdf](http://safety.fhwa.dot.gov/roadway_dept/horcurves/cmhorcurves/horiz_curve.pdf)
- Chin, H. C., & Quek, S. T. (1997). Measurement of Traffic Conflicts. *Safety Science* Vol. 26, No. 3, pp.169-185
- Chrysler, S. T., Cooper, J., McGehee, D. V., Yager, C. (2013). Comparison of driver distraction evaluations across two simulator platforms and an instrumented vehicle. *7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, University of Iowa, Iowa City.
- Council, F M. (1998). Safety benefits of spiral transitions on horizontal curves on two-lane rural roads. *Transportation Research Record, Transportation Research Board*, Issue 1635, pp. 10-17

- Collins, K. M., & Krammes, R. A. (2007). Preliminary validation of a speed-profile model for design consistency evaluation. *Transportation Research Record: Journal of Transportation Research Board*. Volume 1523, Pages 11-21
- Collins, J., Fitzpatrick, K., Bauer, K M., Harwood, D W. (2007). Speed variability on rural two-lane highways. *Transportation Research Record* 1658. Paper No. 99-0457.
- Dingus, T. A., Klauer, S.G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., Perez, M. A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z. R., Jermeland, J., Knippling, R.R. (April, 2006). *The 100-Car Naturalistic Driving Study, Phase II – Results of the 100-Car Field Experiment*. Virginia Tech Transportation Institute. Blacksburg, VA.
- Dingus, T. A., Hankey, J. M., Antin, J. F., Lee, S. E., Eichelberger, L., Stulce, K., McGraw, D., Perez, M., Stowe, L. (May, 2014). *Naturalistic Driving Study: Technical Coordination and Quality Control*. Virginia Tech Transportation Institute. Blacksburg, VA.
- Donnell, E T; Ni, Y; Adolini, M; Elefteriadou, Lily. (2001). Speed prediction models for trucks on two-lane rural highways. *Transportation Research Record, Transportation Research Board, Issue 1751*, pp. 44-55
- Donnell, E., Gross, F., Stodart, B., Opiela, K. (2009). Appraisal of the interactive highway safety design model's crash prediction and design consistency modules: case studies from Pennsylvania. *Journal of Transportation Engineering*, 135(2), pp. 62–73.
- Eenink, R., Barnard, Y., Baumann, M., Augros, X., Utesch, F. (2014). UDRIVE: the European naturalistic driving study. *Transport Research Arena 2014*, Paris
- European Naturalistic Driving Study. (2014). *UDRIVE NDS Website*. Retrieved on December 19<sup>th</sup> at <http://www.udrive.eu/>
- Fang, S. (2014). Shanghai Naturalistic Driving Study. *Fourth International Symposium on Naturalistic Driving Research*. Virginia Tech Transportation Institute. Blacksburg, VA.
- Farmer, C. M., & Lund, A. K. (2002). Rollover risk of cars and light trucks after accounting for driver and environmental factors. *Accident Analysis and Prevention*, Vol. 34, pp. 163–173.
- Federal Highway Administration (FHWA). (1995). Interactive Highway Safety Design Model: Accident Predictive Module. *Public Roads*. Vol 50. No 2. Washington, D.C.
- Federal Highway Administration. (2012). Horizontal Curve Safety. Retrieved on FHWA Safety Website at [http://safety.fhwa.dot.gov/roadway\\_dept/horcurves/](http://safety.fhwa.dot.gov/roadway_dept/horcurves/)
- Federal Highway Administration. (2015a). *Speed Management Safety*. Retrieved on August 8<sup>th</sup>, 2015 at <http://safety.fhwa.dot.gov/speedmgt/>

- Federal Highway Administration. (2015b). *Methods for Establishing Advisory Speed*. Retrieved on August 8<sup>th</sup>, 2015 at [http://safety.fhwa.dot.gov/speedmgt/ref\\_mats/fhwasal122/ch3.cfm](http://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwasal122/ch3.cfm)
- Ferraty, F., & Vieu, P. (2002). *Nonparametric Functional Data Analysis*. Springer-Verlag New York.
- Felipe, E., & Navin, F. (2007). Automobiles on Horizontal Curves: Experiments and Observations. *Transportation Research Record 1628*. Paper No. 98-0482
- Figuerola, A. M. (2005). Speed factors on two-lane rural highways in free-flow conditions. *The 84th Annual Meeting of the Transportation Research Board*. Washington D.C.
- Findley, D. J., Hummer, J. E., Rasdorf, W., Zegeer, C. V. Fowler, T. J. (March 2012). Modeling the impact of spatial relationships on horizontal curve safety. *Accident Analysis & Prevention*, Volume 45, March 2012, Pages 296-304
- Fisher, D. L., Rizzo, M., Caird, J. K., Lee, J. D. (eds.) (2011). *Handbook of driving simulation for engineering, medicine, and psychology*. Boca Raton, FL: CRC Press/Taylor and Francis.
- Fitzpatrick, K., Elefteriadou, L., Harwood, D. W., Collins, J. M., McFadden, J., Anderson, I. B., et al. (1999). *Speed Prediction for Two-Lane Rural Highways*. Texas Transportation Institute. The Texas A&M University System, College Station, Texas.
- Fitzpatrick, K; Shamburger, C B; Krammes, R A; Fambro, D B. (1997). Operating Speed on suburban arterial curves. *Transportation Research Record, Issue 1579*, pp. 89-96
- FOT-NET Wiki. (2014). *Category: NDS*. Retrieved on December 12th, 2014 on <http://wiki.fot-net.eu/index.php?title=Category:NDS>
- General Motor Media. (2012). *First Driving Study Launched in China*. Retrieved on December 18<sup>th</sup>, 2014 at [http://media.gm.com/media/cn/en/gm/news.detail.html/content/Pages/news/cn/en/2012/Sep/0912\\_Naturalistic\\_Driving.html](http://media.gm.com/media/cn/en/gm/news.detail.html/content/Pages/news/cn/en/2012/Sep/0912_Naturalistic_Driving.html)
- Gibreel, G., Easa, S., Hassan, Y., El-Dimeery, I. (1999). State of the art of highway geometric design consistency. *Journal of Transportation Engineering*. 125(4), pp. 305–313.
- Glennon, J. C., & Weaver, G. D. (1971). *The relationship of vehicle paths to highway curve design*. Report 134 - 5. Texas Transportation Institute, Texas A&M University, College Station.
- Gorman, T., Stowe, L., Hankey, J. (May, 2015). *NDS data dissemination activities task 1.6: radar post-processing*. Task Report. Virginia Tech Transportation Institute. Blacksburg, VA.

- Google Map (2015). *Now York State. Google Street View*. Retrieved from <https://www.google.com/maps>
- Guo, F., Klauer, S. G., Hankey, J. M., & Dingus, T. A. (2010). Near crashes as crash surrogate for naturalistic driving studies. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2147, pp. 66–74.
- Gunay, B.; Woodward, D. (2007). Lateral position of traffic negotiating horizontal bends. *Proceedings of the Institution of Civil Engineers, Transport. Volume 160, Issue 1*, pp. 1-11
- Hallmark, S. L., Smadi, O., Hawkins, N. (2013). Speed reduction impacts of dynamic speed feedback signs on rural two lane curves. *Transportation Research Board 92nd Annual Meeting, Transportation Research Board*.
- Hallmark, S. L. (2014). Relationship between speed and lateral position on curves. *Accident Reconstruction Journal, Volume 24, Issue 6*, pp 12-15.
- Hallmark, S., Hsu, Y., Boyle, L., Carriquiry, A., Tian, Y., Mudgal, A. (2011). SHRP 2 Report S2-S01E-RW-1: Evaluation of Data Needs, Crash Surrogates, and Analysis Methods to Address Lane Departure Research Questions Using Naturalistic Driving Study Data. Transportation Research Board of the National Academies, Washington, D.C., 2011. <http://www.trb.org/Publications/Blurbs/166050.aspx>.
- Hankey M. J. (2014). *International light vehicle naturalistic driving*. Retrieved on December 18<sup>th</sup> at <http://onlinepubs.trb.org/onlinepubs/shrp2/2014SafetySymposium/3-CAAUCH-Hankey.pdf>
- Hankey M. J. (2014). Canadian Naturalistic Driving Study. *Fourth International Symposium on Naturalistic Driving Research*. Virginia Tech Transportation Institute. Blacksburg, VA.
- Hassan, Y., & Easa, S. M. (2003). Effect of Vertical Alignment on Driver Perception of Horizontal Curves. *Journal of Transportation Engineering, Vol. 129, No. 4*
- Hauer, E. Safety and the choice of degree of curve. *Transportation Research Record, Transportation Research Board*, Issue 1665, 1999, pp. 22-27
- Helander, M., & Hagvall, B. (1976). An instrumented vehicle for studies of driver behavior. *Accident Analysis & Prevention. Volume 8, Issue 4*, pp. 271–277
- Hendricks, D. L., Freedman, M., Zador, P. L., Fell, J. C. (January, 2001). *The relative frequency of unsafe driving acts in serious traffic crashes*. Veridian Engineering, Inc.
- Hong, I., Iwasaki, M., Furuichi, T., Kadoma, T. (January 2006). Eye movement and driving behavior in curved section passages of an urban motorway. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 220:1319

- Iowa Department of Transportation (Iowa DOT). (2015). *Accident Reports*. Retrieved on October 20<sup>th</sup>, 2015 at <http://www.iowadot.gov/mvd/driverslicense/accidents.htm>
- Kandil, F. I., Rotter, A., and Lappe, M. (2010). Car drivers attend to different gaze targets when negotiating closed vs. open bends. *Journal of Vision*. 10(4):24, pp. 1-11
- Khan, G., Bill, A. R., Chitturi, M. V., Noyce, D. A. (2013). Safety Evaluation of Horizontal Curves on Rural Undivided Roads. *Transportation Research Record: Journal of the Transportation Research Board*, Issue 2386, pp 147–157
- Krammes, R. A., Brackett, R. Q., Shafer, M. A., Ottesen, J. L., Anderson, I. B., Fink, K. L., Collins, K. M., Pendleton, O. J., Messer, C. J. (1995). *Horizontal Alignment Design Consistency for Rural Two Lane Highways*. Report FHWA-RD-94-034. FHWA, U.S. Department of Transportation, 1995.
- Lamm, R., Choueiri, E. M., Hayward, J. C., Paluri, A. (1988). Possible design procedure to promote design consistency in highway geometric design on two-lane rural roads. *Transportation Research Record*, Issue 1195, pp. 111-122
- Lee, S. E., Simons-Morton, B. G., Klauer, S. E., Ouimet, M. C., Dingus, T. A. (February 2010). Naturalistic assessment of novice teenage crash experience. *Accident Analysis and Prevention* 43. pp. 1472–1479
- Levison, W. H. (1998). Interactive Highway Safety Design Model: Issues Related to Driver Modeling. *Transportation Research Record*, *Transportation Research Board*, Issue 1631, pp. 20-27
- Macadam, C. C. (2003). Understanding and modeling the human driver. *Vehicle System Dynamics*. Vol. 40, Nos. 1-3, pp. 101-134.
- Mars, F. (2008). Driving around bends with manipulated eye-steering coordination. *Journal of Vision*. 8(11):10 pp. 1-11
- McGee, H. W., & Hanscom, F. R. (December, 2006). *Low-Cost Treatments for Horizontal Curve Safety*. U.S. Department of Transportation. Washington, D.C.
- McFadden, J. and Elefteriadou, L. Evaluating horizontal alignment design consistency of two-lane rural highways: development of new procedure. *Transportation Research Record*, *Transportation Research Board*, Issue 1737, pp. 9-17
- Montella, A., Pariota, L., Galante, F., Imbriani, L. L., Mauriello, F. (2014). Prediction of Drivers' Speed Behavior on Rural Motorways Based on an Instrumented Vehicle Study. *Transportation Research Record: Journal of the Transportation Research Board*, Issue 2434, pp. 52–62

- Montella, A., Imbriani, L. L. (2015). Safety performance functions incorporating design consistency variables. *Accident Analysis & Prevention, Volume 74, Issue 0*, pp 133-144
- Milton, J., & Mannering, F. (1998). The Relationship among highway geometric, traffic-related elements, and motor-vehicle accident frequencies. *Transportation*, Vol. 25, 1998, pp. 395–413.
- National Safety Council. (2014). *The odds of dying from*. Retrieved on December 7th, 2014 at [http://www.nsc.org/NSCDocuments\\_Corporate/2014-Injury-Facts-Odds-Dying-43.pdf](http://www.nsc.org/NSCDocuments_Corporate/2014-Injury-Facts-Odds-Dying-43.pdf)
- National Highway Traffic Safety Administration (NHTSA). (2014). *Traffic Safety Facts 2012*. Washington, D.C.
- National Highway Traffic Safety Administration (NHTSA). (2014). *Fatality Analysis Reporting System*. Retrieved on December 8th, 2014 at <http://www.nhtsa.gov/FARS>
- National Highway Traffic Safety Administration (2014). *National Automotive Sampling System (NASS General Estimates System (GES) Analytical User's Manual 1988-2012*. Highway Traffic Safety Administration. Washington, DC.
- National Highway Traffic Safety Administration (NHTSA). (July 2008). *National Motor Vehicle Crash Causation survey*. Report to Congress. Springfield, Virginia.
- National Center for Injury Prevention and Control. (2014). *10 Leading Causes of Death by Age Group, United States – 2010*. Retrieved on December 7th, 2014 at [http://www.cdc.gov/injury/wisqars/pdf/10lcid\\_all\\_deaths\\_by\\_age\\_group\\_2010-a.pdf](http://www.cdc.gov/injury/wisqars/pdf/10lcid_all_deaths_by_age_group_2010-a.pdf)
- National Advanced Driving Simulator. (2015). *NADS-I*. Retrieved on January 2nd, 2015 at [http://www.nads-sc.uiowa.edu/sim\\_nads1.php](http://www.nads-sc.uiowa.edu/sim_nads1.php)
- Ottesen, J. L., & Krammes, R. A. (2000). Speed profile model for a design consistency evaluation procedure in the United States. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1701, pp. 76–85
- Park Y. J., Park C. H., and Chon, K. S. (2003). Development of potential measures to evaluate the safety of curves in rural highways. *KSCE Journal of Civil Engineering. Volume 7, No. 3*, pp. 233~242
- Pierre, G. S. (2014). From FOT to NDS Recent developments in UDRIVE. French Institute of Science and Technology for Transport, Development and Network. *The 4<sup>th</sup> Naturalistic Driving Study Symposium*. Blacksburg, Virginia.
- Polus, A; Fitzpatrick, K; Fambro, D. B. (2000). Predicting operating speeds on tangent sections of two-lane rural highways. *Transportation Research Record, Transportation Research Board, Issue 1737*, pp. 50-57

- PROLOGUE. (2009). *Promoting real life observations for gaining understanding of road user behavior in Europe*. Retrieved on December 12th, on <http://www.prologue-eu.eu/prologue/about/>
- Preston, H., & Schoenecker, T. (1999). *Potential safety effects of dynamic signing at rural horizontal curves*. BRW, Incorporated; Minnesota Department of Transportation, 1999, 82 p.
- Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and Matlab*. Springer-Verlag New York.
- Ramsay, J., & Silverman, B. W. (2002). *Applied functional data analysis*. Springer-Verlag New York.
- Ramsay, J., & Silverman, B. W. (2005). *Functional data analysis*. Springer-Verlag New York.
- Rasanen, M. (2005). Effects of a rumble strip barrier line on lane keeping in a curve. *Accident Analysis and Prevention* 37 (2005). pp.575-581.
- Raymond, G., Kemeny, A., Droulez, J., and Berthoz, A. (2001). Role of Lateral Acceleration in Curve Driving: Driver Model and Experiments on a Real Vehicle and a Driving Simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 2001.
- Ren, Y., Zhao, H., Li, X., Zheng, X. (2012). Study on vehicle track model in road curved section based on vehicle dynamic characteristics. *Mathematical Problems in Engineering*, vol. 2012, Article ID 818136, 17 pages, doi:10.1155/2012/818136
- Reymond, G., Kemeny, A., Droulez, J., and Berthoz, A. (2001). Role of lateral acceleration in curve driving: driver model and experiments on a real vehicle and a driving simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 43(3), pp.483-95
- Rumar K., Evans L. & Schwing R. (1985). The Role of Perceptual and Cognitive Filters in Observed Behavior. *Human Behavior in Traffic Safety*, eds. Plenum Press.
- Sagberg, F., Eenink, R., Hoedemaeker, M., Lotan, T., van Nes, N., Smokers, R., Welsh, R., Winkelbauer, M. (2011). *Recommendations for a large-scale European naturalistic driving observation study*. PROLOGUE Deliverable D4.1. TØI Institute of Transport Economics, Oslo, Norway.
- Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. 48:362
- Schneider, W., Zimmerman, K., Savolainen, P., and Moore, D. (2010). Effects of Horizontal Curvature on Single-Vehicle Motorcycle Crashes along Rural Two-Lane Highways. *Journal of the Transportation Research Board*, No. 2194, TRB, National Research Council, Washington, D.C., pp. 91 -98.

- Schagen, I. V. & Sagberg, F. (2012). The potential benefits of naturalistic driving for road safety research: Theoretical and empirical considerations and challenges for the future. *Proceeding: Social and Behavioral Sciences* 48 (2012) 692-701
- Schurr, K. S., McCoy, P. T., Pesti, G., and Huff, R. (2002). Relationship of Design, Operating, and Posted Speeds on Horizontal Curves of Rural Two-Lane Highways in Nebraska. *Transportation Research Record* 1976. Paper No. 02-2363.
- Salmon, P. M., Regan, M. A., Johnston, I. (2005). Human error and road transport: Phase one- a framework for an error tolerant road transport system. Monash University Accident Research Centre.
- Strategic Highway Research Program 2 (SHRP 2). (2014). The SHRP2 Naturalistic Driving Study. Retrieved December 7th, 2014 from <http://www.trb.org/StrategicHighwayResearchProgram2SHRP2/Pages/The-SHRP-2-Naturalistic-Driving-Study-472.aspx>
- Shinar, D., Mcdowell, D., Rockwell, T. (1977). Eye movements in curve negotiation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 1977.
- SHRP2. (2014). *Recruitment Website*. Retrieved on December 15<sup>th</sup> at <http://www.shrp2nds.us/>
- SHRP2 InSight. (2014). *Second Strategic Highway Research Program 2 InSight Website*. Retrieved on December 15<sup>th</sup> at <https://insight.shrp2nds.us/home/index>
- Simons-Morton, B., Klauer, C., Guo, F., Lee, S., Ouimet, M. C., Albert, P., Dingus, T. (2014). The naturalistic Teenage driving study: Methods & Selected Findings. *Naturalistic Driving Study Symposium*.
- Smadi, O. (2015). *SHRP2 Roadway Information Database*. Institute for Transportation. Ames, IA.
- Smith, R. C., Doerzaph, Z., Hankey, J. (2015). *Naturalistic Driving Study: Alcohol Sensor Performance*. SHRP 2 Report, Issue Report S2-S31-RW-2, 36p
- Spainhour, L. K., Brill, D., Sobanjo, J. O., Wekezer, J., Mtenga, P. V. (2005). *Evaluation of Traffic Crash Fatality Causes and Effects*. Florida A&M University-Florida State University. Tallahassee, Florida.
- Suh, W., Park, P., Park, C., and Chon, K. (2006). Relationship between Speed, Lateral Placement, and Drivers' Eye Movement at Two-Lane Rural Highways. *Journal of Transportation Engineering*. 132(8), pp. 649–653.
- Spacek, P. (2005). Track Behavior in Curve Areas: Attempt at Typology. *Journal of Transportation Engineering, Volume 131, Issue 9*, pp. 669-676



- Taragin, A. (1954). Driver performance on horizontal curves. *Proceedings, Vol 33, Highway Research Board*, 1954.
- Taylor, W. C., Abu-Lebdeh, G., Rai, S. (2005). Effect of continuous shoulder rumble strips and pavement marking on lateral placement of vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1911, pp. 105–112
- Torbic, D. J., Harwood, D. W., Gilmore, D. K., Pfefer, R., Neuman, T. R., Slack, K. L., Hardy, K. K. (2004). Guidance for Implementation of the AASHTO Strategic Highway Safety Plan. Volume 7: A Guide for Reducing Collision on Horizontal Curves. *NCHRP Report, Transportation Research Board; CH2M Hill, Issue 500*
- Transportation Research Board of the National Academies of Science. (2013). *The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset*. Retrieved on May 5<sup>th</sup>, 2015 at <https://insight.shrp2nds.us>.
- Treat, H. R., Tumbas, N. S., McDonald, S. T., Shinar, D., Hume, R. D., Mayer, R. E., Stansider, R. L., Castellan, N. J. (1979). *Tri-level study of the causes of traffic accidents: Final report – executive summary*. Institute for Research in Public Safety. Indiana University. Bloomington, Indiana.
- Virginia Tech Transportation Institute (February 2015). *SHRP2 Researcher Dictionary for Video Reduction Data*. Blacksburg, Virginia.
- Virginia Tech Transportation Institute (VTTI) (2013). *SHRP2 Naturalistic Driving Study (SHRP2 NDS)*. Reprieved on May 5<sup>th</sup>, 2015 at <http://www.shrp2nds.us/>
- Wierwille, W. W., Hanowski, R. J, HANKEY, J. M., Kieliszewski, C. A., Lee, S. E., Medina, A., Keisler, A. S., Dingus, T. A. (2002). *Identification and Evaluation of Driver Errors: Overview and Recommendations*. Virginia Polytechnic Institute and State University, Blacksburg; Federal Highway Administration.
- Wooldridge, M. D., Fitzpatrick, K., Koppa, R., Bauer, K. (2000). Effects of horizontal curvature on driver visual demand. *Transportation Research Record, Transportation Research Board, Issue 1737*, pp. 71-77
- Wu, K. & Jovanis, P. P. (2013). *Screening Naturalistic Driving Study Data for safety-critical Events*. Transportation Research Board 2013 Annual Meeting, Washington D.C.
- Wu, K., Donnell, E., Himes, S., Sasidharan, L. (2013). Exploring the Association between Traffic Safety and Geometric Design Consistency Based on Vehicle Speed Metrics. *Journal of Transportation Engineering*, 139(7), pp. 738–748.
- Ye, F., & Lord, D. (2011). Investigation of Effects of Underreporting Crash Data on Three Commonly Used Traffic Crash Severity Models: Multinomial Logit, Ordered Probit, and Mixed Logit. *Transportation Research Record: Journal of the Transportation Research Board, Issue 2241*, pp. 51-58

Zakowska, L. Road curve evaluation based on road view perception study. *Transportation Research Record, Transportation Research Board, Issue 1689*, pp. 68-72

Zegeer, C. V., Stewart, J. R., Council, F. M., Reinfurt, D. W., Hamilton, E. (1992). Safety effects of geometric improvements on horizontal curves. *Transportation Research Record, Issue 1356*, pp. 11-19

## APPENDIX A. BACKGROUND OF THE SHRP2 NDS

**Table A.1 Recruitment summary by method, age group, and site (Dingus, 2014)**

Method	Age Group	State College	Bloomington	Durham	Seattle	Tampa	Buffalo	Total
Calls Out	16–17	1	0	3	1	1	0	6
	18–20	1	5	6	4	4	9	29
	21–25	13	10	13	11	17	19	83
	26–35	47	45	93	56	86	94	421
	36–50	138	109	284	191	222	283	1,227
	51–65	208	168	284	243	297	370	1,570
	66–75	86	68	142	190	213	191	890
	76+	37	29	42	91	111	88	398
	Subtotal	531	434	867	787	951	1054	4624
Calls In	16–17	3	4	23	0	19	5	54
	18–20	11	15	42	0	68	17	153
	21–25	6	9	44	5	75	19	158
	26–35	7	13	49	10	41	32	152
	36–50	9	17	78	21	97	62	284
	51–65	15	16	81	15	148	76	351
	66–75	8	5	72	7	108	48	248
	76+	4	9	95	2	49	37	196
	Subtotal	63	88	484	60	605	296	1596
WBST	16–17	27	44	172	255	282	125	905
	18–20	92	80	188	398	840	413	2,011
	21–25	142	140	290	671	650	584	2,477
	26–35	138	61	339	517	503	508	2,066
	36–50	127	75	427	532	662	396	2,219
	51–65	125	56	220	414	456	269	1,540
	66–75	42	24	100	102	214	70	552
	76+	27	15	84	117	68	76	387
	Subtotal	720	495	1820	3006	3675	2441	12157
	Total	1314	1017	3171	3853	5231	3791	18377

**Table A.2 Assessment questionnaires administered (Dingus, 2014)**

<b>Name</b>	<b>Description</b>
Sleep Questionnaire	A questionnaire designed to determine the participant's sleeping patterns, habits, and level of fatigue (Appendix B).
Perception of Risk Survey and Frequency of Risky Behavior Questionnaires	A questionnaire designed to gauge the participant's perception of dangerous or unsafe driving behaviors or scenarios and a questionnaire designed to gauge the frequency and a participant's willingness to engage in dangerous, unsafe, or risky behaviors (Appendix C).
Barkley's ADHD Quick Screen	A short, clinical ADHD screening assessment. This screening instrument operationalizes ADHD symptoms in terms of specific behaviors (Appendix D).
Sensation Seeking Scale	A survey comprising questions to gauge the degree to which the participant engages in sensation seeking behavior. The test measures the participant's sensory stimulation preferences (Appendix E).
Driving Knowledge	A test of knowledge of driving laws and appropriate behaviors (Appendix F).
Medical Conditions and Medications Survey and Exit Survey	Questionnaires designed to obtain participants' self-reported medical history. The questions focus on the identification of conditions that could affect driving performance and safety (Appendix G).
Modified Manchester Driver Behavior	A self-reported driver behavior survey. The participant is asked to indicate how often he/she commits each described error (accidental) or violation (deliberate) (Appendix H).

**Table A.3 Cognitive assessments (Dingus, 2014)**

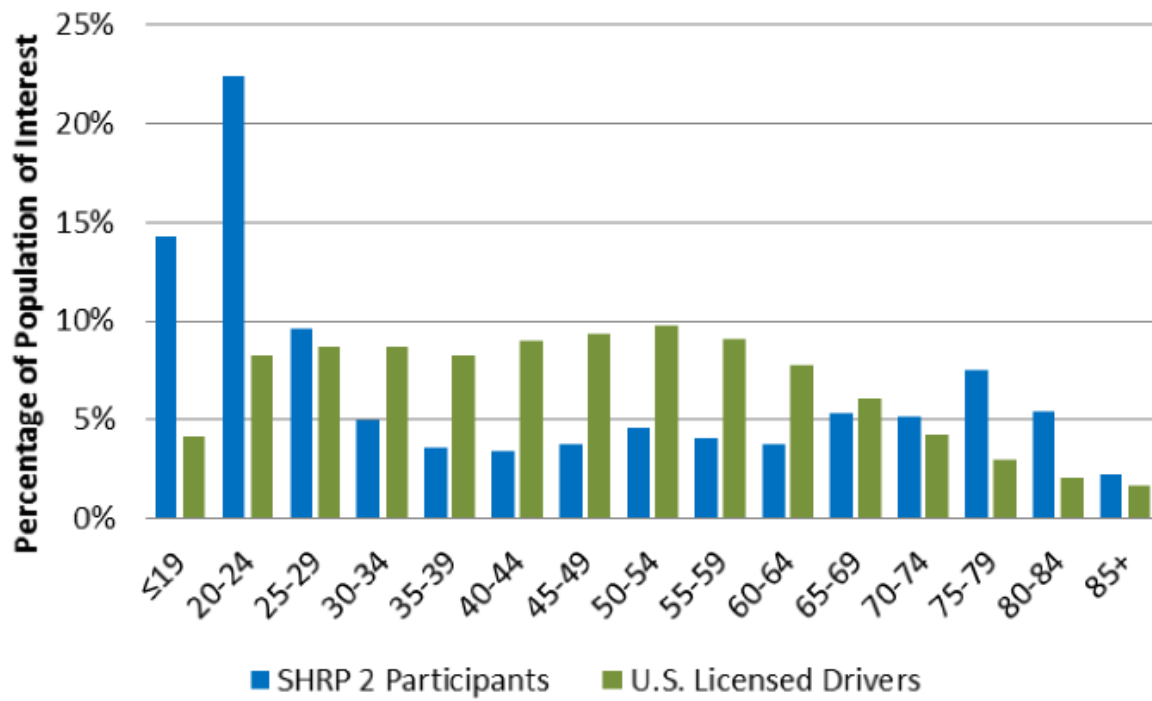
<b>Assessment</b>	<b>Description</b>
Clock drawing test	The participant is presented with pencil and paper; on the paper is a circle and nothing else. The participant is asked to draw numbers in the circle to make the circle look like the face of a clock and then draw the hands of the clock to read “10 after 11.”
Conners’ Continuous Performance Test Version 5 (CPT II)	The CPT II is a task-oriented computerized assessment of attention disorders and neurological functioning. Results indicate the likelihood that an individual has an attention disorder.
Visualizing missing information—Motor-Free Visual Perception Test	Participants are shown a reference image and four similar but incomplete figures. Participants are instructed to indicate which incomplete figure could be completed to duplicate the target figure; only one of the incomplete figures can be completed in such a way as to form an exact duplicate of the target figure.
Visual information processing speed—Useful Field of View (UFOV)	Participants are briefly presented one of two very similar target stimuli in the center of the display. Simultaneously, a second target icon—the same as the central target—is presented in one of eight possible peripheral locations at varying eccentricities in a 35-degree region around the central visual field. Participants have to identify both what the central target is and the location of the peripheral target.
Trail making (Parts A and B)	In Part A, participants use a touch screen to connect in order (i.e., 1-2- . . . n) a series of randomly arranged numbers, then in Part B they connect a series of randomly arranged numbers and letters in alternating progressing sequences (i.e., 1-A-2-B-3 . . . n). Time-to-completion of the entire series is recorded.

**Table A.4 SHRP2 NDS vehicle classes (Dingus, 2014)**

<b>Class</b>	<b>Definition</b>	<b>Vehicle Count</b>
<b>Prime</b>	Vehicles included on the original eligible vehicle list for which the Coordination Contractor procured PIDs	1717
<b>Subprime</b>	Vehicles manufactured primarily after 2009 for which Coordination Contractor was not able to procure PIDs but was able to obtain information through the CAN communication protocol, an industry standard after 2009	488
<b>Legacy</b>	Vehicles manufactured between 1996 and 2008, employing an older network for vehicle communications	736
<b>Basic</b>	Vehicles manufactured before 1996	421

**Table A.5 Quality assessment of select vehicle metrics (07/17/2013) (Dingus, 2014)**

<b>Data Item</b>	<b>Good Quality (%)</b>
Network Speed	97.35%
Accelerator Position	97.65%
Turn Signal Status	94.29%
Brake Pedal	96.15%
Usable Face Video	99.08%
Usable Forward Video	97.61%
Usable Rear Video	94.07%
Usable Lap Video	99.26%
Usable IMU (Accelerator, x-axis)	99.27%
Usable GPS (speed only)	99.06%



**Figure A.1 SHRP2 participant versus U.S. driving population percentages by age group (Dingus et al, 2014)**

## APPENDIX B. SAMPLE R CODE

```

#Loop for Curve PC/PT Identification
#Loop for Interpolation of Time Stamps
#Loop for Curve Direction Identification

for (i in 1:length(output.list)){
  event<-output.list[[i]][c("vtti.latitude", "vtti.longitude")]
  w <- find.matches(curvepcptlist[c("PC_Lat", "PC_Long")], event,
maxmatch=2, tol=c(.001,.001))

  pcpt<-na.omit((data.frame(w$matches)))#identify matched cases
  pcpt<-cbind(pcpt, na.omit((data.frame(w$distance))))#add match distance
  pcpt$curveid<-as.numeric(rownames(pcpt))

  #interpolation of the accurate position
  pcpt$id<-ifelse(pcpt$Match..1<pcpt$Match..2,

pcpt$Match..1+as.integer((pcpt$Distance..1)/(pcpt$Distance..1+pcpt$Distance..
2)*10),
                pcpt$Match..1-
as.integer((pcpt$Distance..1)/(pcpt$Distance..1+pcpt$Distance..2)*10))

  #merge the pcpt information with curve information
  pcpt.merge<-merge(curvepcptlist, pcpt, by.x=c("pcptid"), by.y=c("curveid"),
all.y=TRUE)
  pcpt.merge<-pcpt.merge[order(pcpt.merge$Match..1),]

  #create directional label
  if (nrow(pcpt.merge)%2==0){
    direction<-pcpt.merge[c(TRUE, FALSE),]$Direction_
    pcpt.merge[!c(TRUE, FALSE),]$Direction_<-direction}else{
    error.obs<-rbind(error.obs, i)
  }

  #Create PCPT
  if (nrow(pcpt.merge)>=2){
    #create PC/PT label
    pcpt.merge$pcpt<-NA
    pcpt.merge[c(TRUE, FALSE),]$pcpt<-'PC'
    pcpt.merge[!c(TRUE, FALSE),]$pcpt<-'PT'
  }else{
    error.obs<-rbind(error.obs, i)
  }

  output.list[[i]]$id<-as.numeric(rownames(output.list[[i]]))

  data_list.new[[i]]<-merge(output.list[[i]], pcpt.merge, by.x=c("id"),
all.x=TRUE)
}

```



## APPENDIX C. FUNCTIONAL DATA ANALYSIS MODEL OUTPUTS

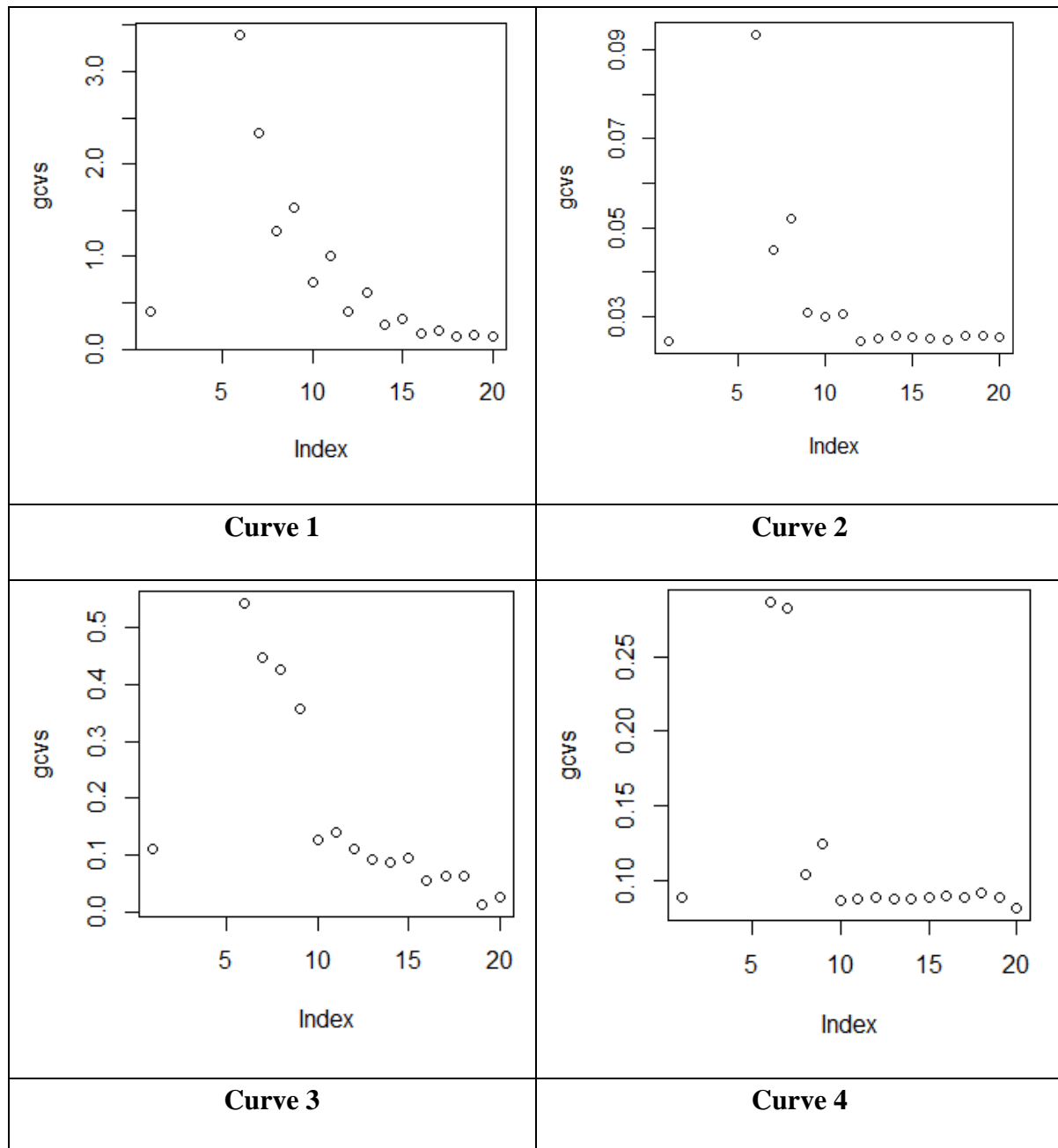
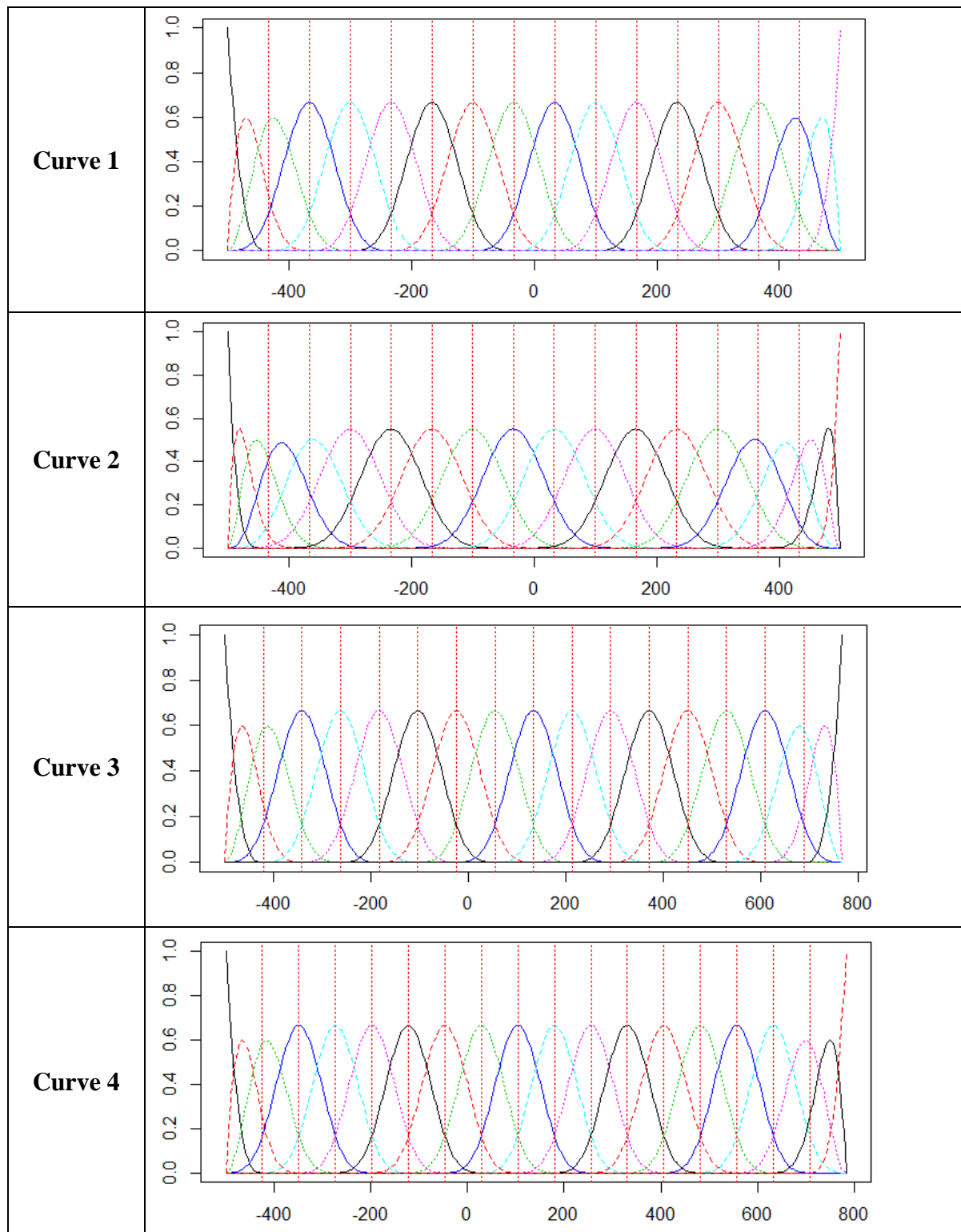
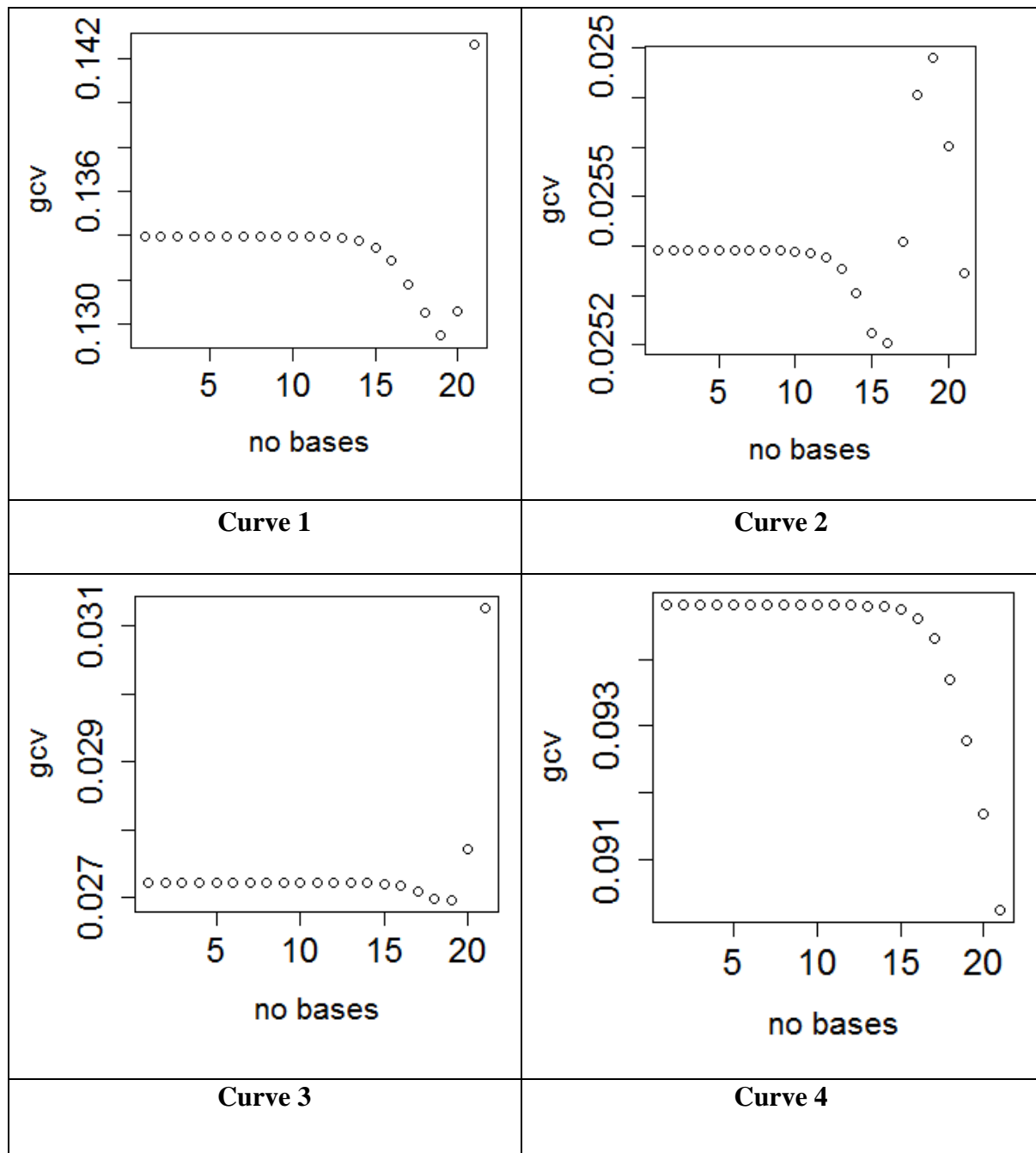


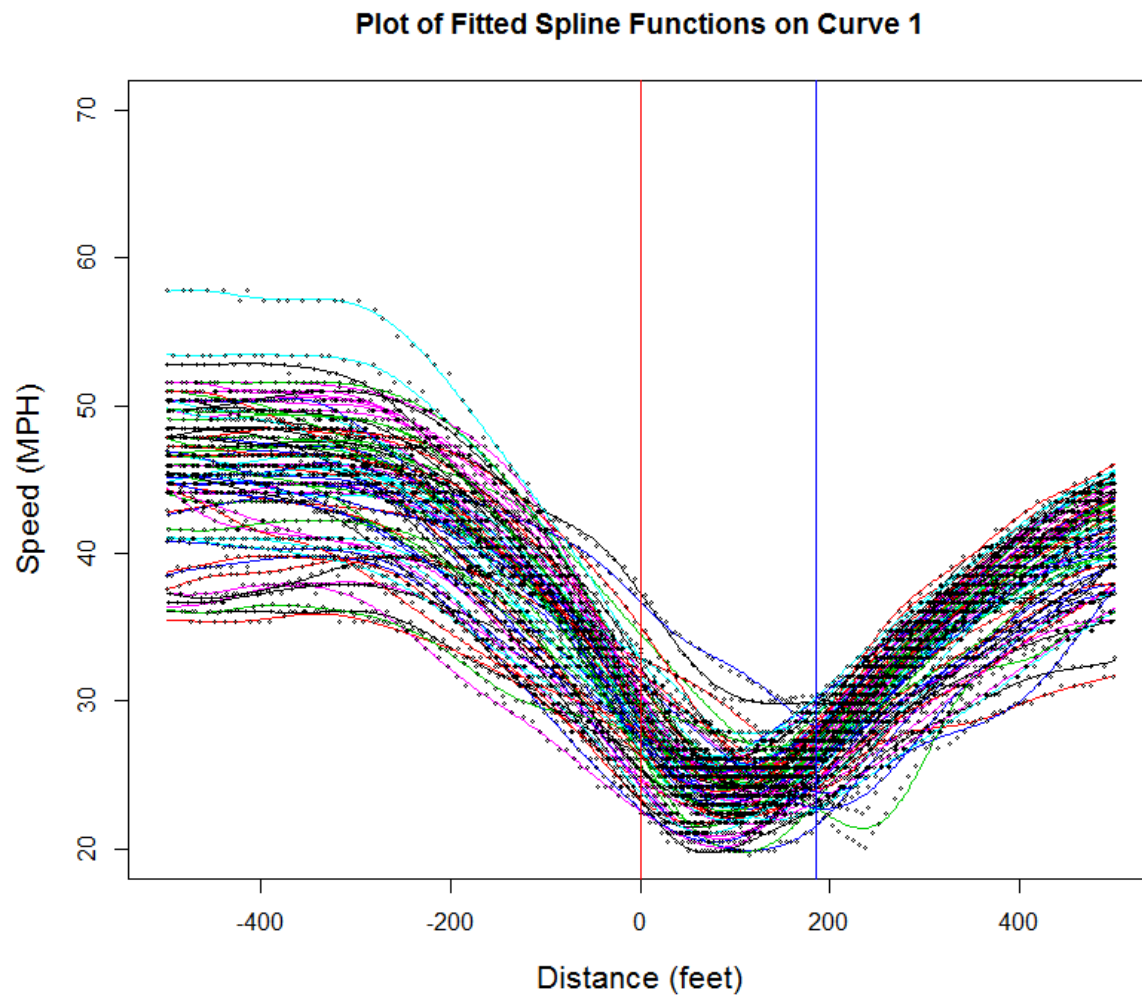
Figure C.1 Scree plot for choosing the optimal number of basis functions



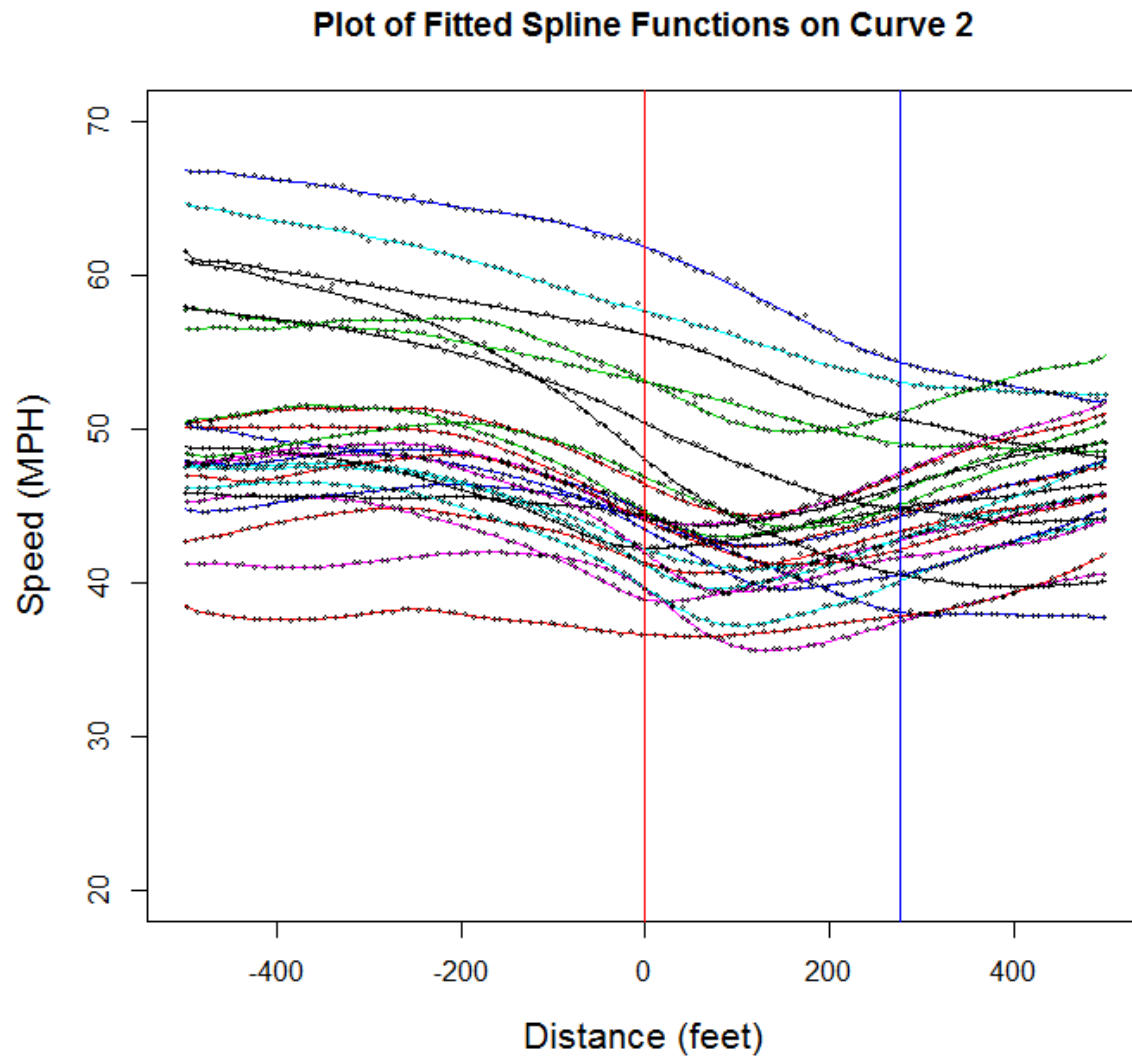
**Figure C.2 List of basis functions on the four example curves**



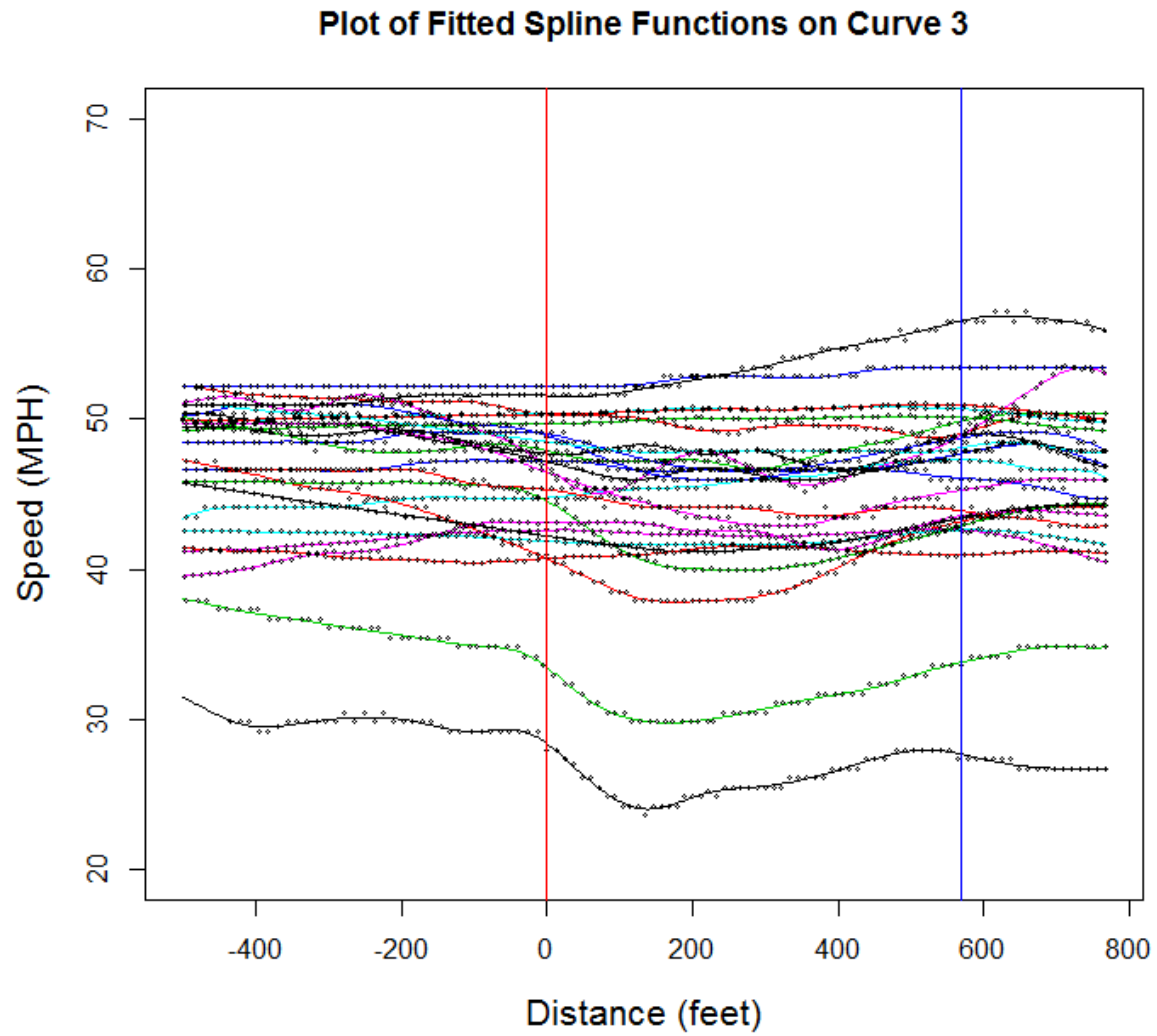
**Figure C.3** Scree plot for choosing the optimal turning parameters



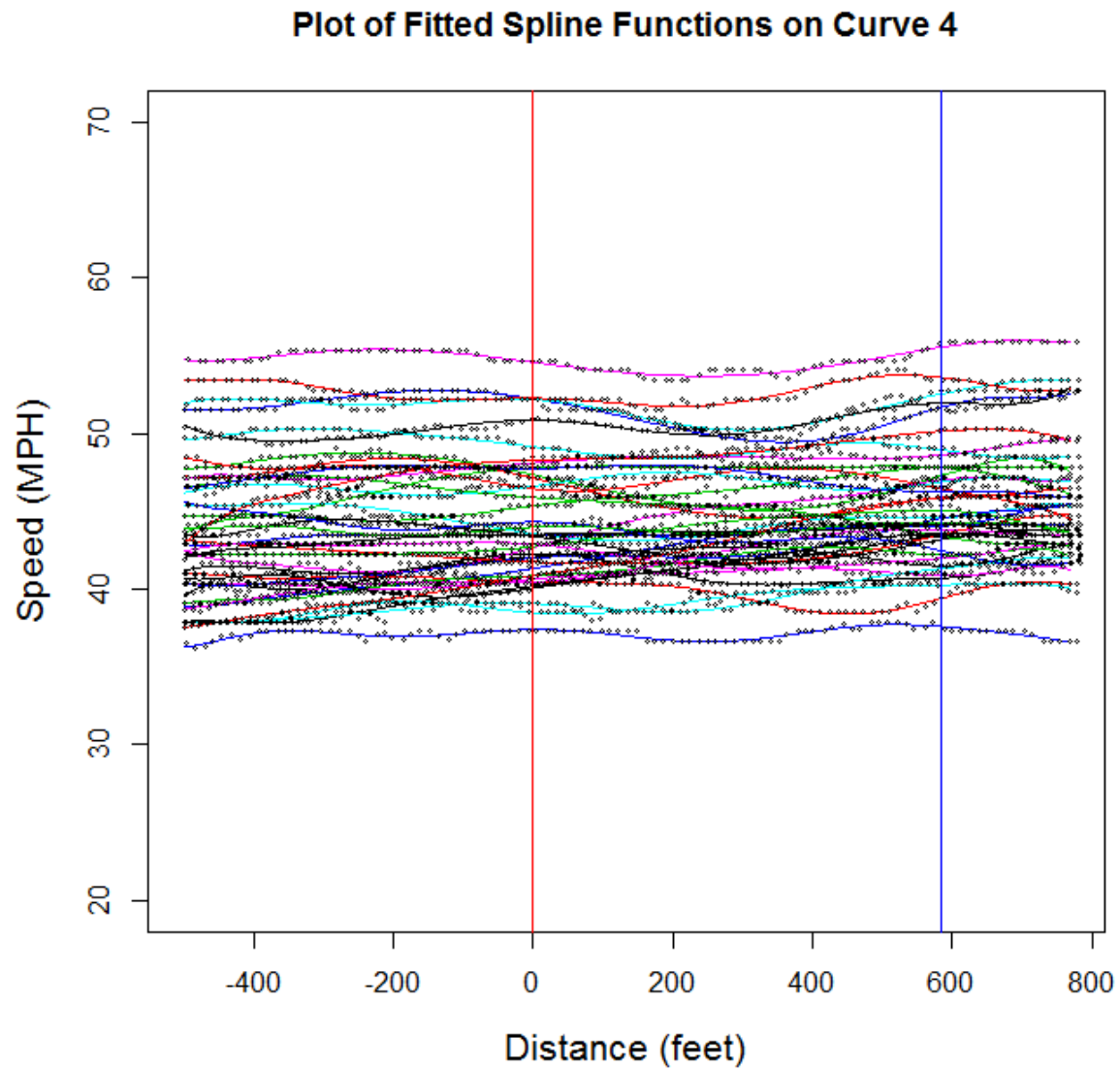
**Figure C.4 Plot of discrete raw time series speed data and best fitted functional data on curve 1**



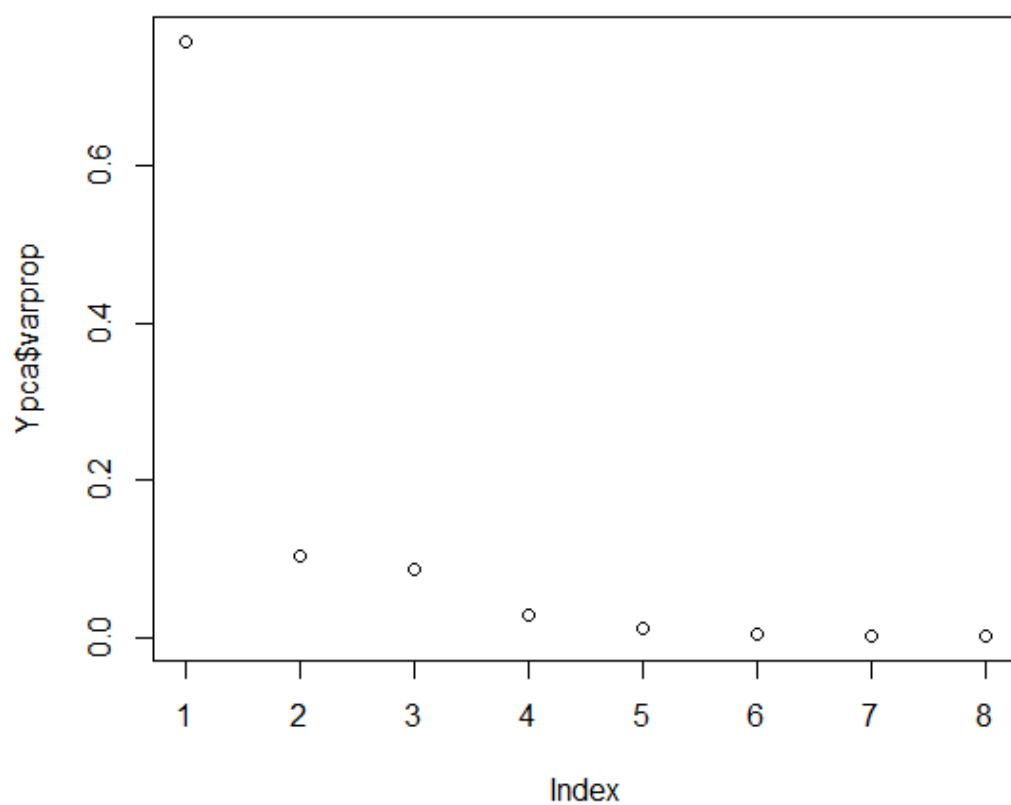
**Figure C.5 Plot of discrete raw time series speed data and best fitted functional data on curve 2**



**Figure C.6 Plot of discrete raw time series speed data and best fitted functional data on curve 3**



**Figure C.7 Plot of discrete raw time series speed data and best fitted functional data on curve 4**



**Figure C.8** Scree plot for choosing optimal number of principal components. The first three principal components were chosen for analysis.