EXTRACTING RURAL CRASH INJURY AND FATALITY PATTERNS DUE TO CHANGING CLIMATES IN RITI COMMUNITIES BASED ON ENHANCED DATA ANALYSIS AND VISUALIZATION TOOLS (PHASE I)

FINAL PROJECT REPORT

by

Guohui Zhang, Ph.D., P.E., Panos D. Prevedouros, Ph.D., P.E., David T. Ma, Ph.D., Hao Yu, Ph.D., Zhenning Li, Ph.D., Runze Yuan
Department of Civil and Environmental Engineering
University of Hawaii at Manoa

for

Center for Safety Equity in Transportation (CSET)
USDOT Tier 1 University Transportation Center
University of Alaska Fairbanks
ELIF Suite 240, 1764 Tanana Drive
Fairbanks, AK 99775-5910

In cooperation with U.S. Department of Transportation,
Research and Innovative Technology Administration (RITA)

DISCLAIMER

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

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
Extracting Rural Crash Injury and Fatality Patterns Due to Changing Climates in RITI Communities Based on Enhanced Data Analysis and Visualization Tools	September 30, 2021
	6. Performing Organization Code

7. Author(s) and Affiliations	8. Performing Organization Report No.
Guohui Zhang, Ph.D., P.E., Panos D. Prevedouros, Ph.D., P.E., David T. Ma, Ph.D. Hao Yu, Ph.D., Zhenning Li, Ph.D., Runze Yuan Department of Civil and Environmental Engineering University of Hawaii at Manoa	INE/CSET 21.10

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Center for Safety Equity in Transportation ELIF Building Room 240, 1760 Tanana Drive Fairbanks, AK 99775-5910	**11. Contract or Grant No.**
	Grant # 69A3551747129

12. Sponsoring Organization Name and Address	13. Type of Report and Period Covered
United States Department of Transportation Research and Innovative Technology Administration 1200 New Jersey Avenue, SE Washington, DC 20590	Research Report
	14. Sponsoring Agency Code

15. Supplementary Notes

Report uploaded to:

16. Abstract

Traffic crashes cause considerable incapacitating injuries and losses in Rural, Isolated, Tribal, or Indigenous (RITI) communities. Compared to urban traffic crashes, those rural crashes, especially for those occurred in RITI communities, are heavily associated with factors such as speeding, low safety devices application (for instance, seatbelt), adverse weather conditions and lacking maintenance and repairers for road conditions, inferior lighting conditions, and so on. Therefore, there exists an urgent need to investigate the unique attributes associated with the RITI traffic crashes based on numerous approaches, such as statistical methods, and data-driven approaches. This project focused on extracting rural crash injury and fatality patterns due to changing climates in RITI communities based on enhanced data analysis and visualization tools. Three new interactive graphic tools were added to the Rural Crash Visualization Tool System (RCVTS), to enhance the visualization approach. A Bayesian vector auto-regression based data analysis approach was proposed to enable irregularly-spaced mixture-frequency traffic collision data interpretation with missing values. Moreover, a finite mixture random parameters model was formulated to explore driver injury severity patterns and causes in low visibility related single-vehicle crashes. The research findings are helpful for transportation agencies to develop cost-effective countermeasures to mitigate rural crash severities under extreme climate and weather conditions and minimize the rural crash risks and severities in the States of Alaska, Washington, Idaho, and Hawaii.

17. Key Words	18. Distribution Statement
Traffic safety, Fatality pattern, Visualization, Changing climate, Rural area	

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified.	Unclassified.	80	N/A

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized.**

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
	NOTE: volumes greater than 1000 L shall be shown in m^3			
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	$candela/m^2$	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS				
Symbol	**When You Know**	**Multiply By**	**To Find**	**Symbol**
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	$candela/m^2$	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Traffic crashes have caused considerable incapacitating injuries and losses in rural, isolated, tribal, or indigenous (RITI) communities. Compared to urban traffic crashes, those rural crashes, especially for those occurred in RITI communities, are heavily associated with factors such as speeding, low safety devices application (for instance, seatbelt), adverse weather conditions, limited maintenance and repair for roads, inferior lighting conditions, and so on. Therefore, there exists an urgent need to investigate the unique attributes associated with the RITI traffic crashes based on numerous approaches, such as statistical methods, and data-driven approaches. However, it was found that crash data analysis suffers from not only the unobserved heterogeneities but also the temporal instability. What's worse, many related characteristics may have a different cycle, resulting in incomplete data records.

To address the research gap, the Year 2 project aims: 1) to enhance the interactive baseline crash data platform, which is capable of visualizing and analyzing rural crash in RITI communities, with more interactive graphs; 2) to investigate the Bayesian vector autoregression-based approach for mixed frequency crash data interpretations with missing values; and 3) to propose a finite mixture random parameter model to explore driver injury severity patterns and causes in low visibility conditions. This research effort has gathered and leveraged existing traffic crash databases with the state of Washington, Idaho, Alaska, and Hawaii. The proposed research enabled effective traffic safety program management at all levels in RITI communities to design and implement appropriate countermeasures to mitigate rural crash severities and risks.

The project updated the RCVTS, a web-based tool that aims to deal with visualization issues associated with various rural crash characteristics. The updated RCVTS features three new graph types. The RCVTS allows users access to traffic data stored in the database, and to create highly customized analytical graphs. Currently, traffic crash data collected in the northwest region— i.e., Alaska, Idaho, and Washington—were shared online through a MySQL database using the phpMyAdmin technique.

A novel Bayesian vector autoregression approach is proposed to address this problem. An unevenly spaced traffic collision dataset with missing values, containing all collisions in different severities that occurred on the state highways in Washington State from January 2006 to December 2016, is selected in this study of the impacts of transportation-, weather- and socioeconomic-related characteristics on traffic collisions. A Gibbs sampler is used to conduct Bayesian inference for model parameters and unobserved high-frequency variables. Results show that the model has a reasonably superior fit accuracy and can capture the unobserved heterogeneity in the dataset. The proposed VAR also demonstrates better performance than other missing value imputation techniques, including linear regression, predictive mean matching, k-nearest neighbors, and random forests.

Low visibility is consistently considered as a hazardous factor due to its potential to lead to severe fatal crashes. However, unlike the other inclement weather conditions that have attracted extensive research interests, only a few studies have been conducted to investigate the impacts of risk factors on driver injury severity outcomes in low visibility related crashes. A three-year crash dataset including all low visibility related crashes from 2010 to 2012 in four South Central states, i.e., Arkansas, Louisiana, Texas, and Oklahoma, is adopted in this study. A finite mixture random parameter approach is developed to interpret both within-class and between-class unobserved heterogeneity among crash data. After a careful comparison, a two-class finite mixture random parameter model with normal distribution

assumptions is selected as the final model. Estimation results show that three variables, including young (specific to injury, I), male (specific to serious injury and fatal, F), and a large truck (specific to serious injury and fatal, F), are found to be normally distributed and have significant impacts on driver injury severities. Variables with fixed effects including rural, wet, 60 mph or higher, no statutory limit, dark, Sunday, curve, rollover, light truck, old, and drug/alcohol-impaired also have significant influences on driver injury severities. This study provides an insightful understanding of the impacts of these variables on driver injury severity outcomes in low visibility related crashes, and a useful reference for developing countermeasures and strategies to mitigate driver injury severities under these conditions.

CHAPTER 1. INTRODUCTION

1.1. Problem Statement

Traffic crashes have caused considerable incapacitating injuries and losses in rural, isolated, tribal, or indigenous (RITI) communities. For instance, more than 50% of fatalities occurred on rural roadways, and more than 20, 000 people lost their lives annually in rural crashes. Moreover, US Department of Transportation (USDOT) declared that the fatality rate in rural areas is double the rate in urban areas in their 2013 National Highway Traffic Safety Administration (NHTSA) report, and the Hawaii Department of Transportation (HDOT) also reported 195% higher fatality rate in the rural areas of the state than in the urban areas in 2014. In the same document, the HDOT emphasized that the native Hawaiians or other pacific island residents were involved in more than 20% of the motor vehicle crashes in 2014.

Compared to urban traffic crashes, rural crashes, especially for those that occurred in RITI communities, are heavily associated with factors such as speeding, low safety devices application (for instance, seatbelt use), adverse weather conditions, lack of maintenance and repair for roads, inferior lighting conditions, and so on. Therefore, an urgent need exists to investigate the unique attributes associating with the RITI traffic crashes based on numerous approaches, including statistical methods, and data-driven approaches. However, it was found that crash data analysis suffers from not only the unobserved heterogeneities but also the temporal instability. What's worse, many related characteristics may have a different cycle, which results in incomplete data records.

To address the research gap, the Year 2 project aimed: 1) to enhance the interactive baseline crash data platform, which is capable of visualizing and analyzing rural crash in RITI communities, with more interactive graphs; 2) to investigate the Bayesian vector autoregression-based approach for mixed frequency crash data interpretations with missing values; and 3) to propose a finite mixture random parameter models to explore driver injury severity patterns and causes in low visibility conditions. This research effort has gathered and leveraged existing traffic crash databases with the state of Washington, Idaho, Alaska, and Hawaii.

The proposed research enabled effective traffic safety program management at all levels in RITI communities to design and implement appropriate countermeasures to mitigate rural crash severities and risks. The updated crash data platform would provide more interesting functions, and the proposed Bayesian approach and finite mixture random parameter models made fundamental contributions in the crash data analysis in RITI communities.

1.2. General Background

This project is well-aligned with the CSET Year 2 project themes on baseline data establishment by extracting rural crash injury and fatality patterns. Based on the research tasks, the project team acquired and obtained rural crash data related to RITI transportation safety. The data platform system built up the data infrastructure needed to measure CSET performance and overall contribution to RITI transportation safety over time. This project directly contributes to safety data collection, retrieval, management, visualization, and analysis in the rural and tribal areas. The research tasks address CSET baseline data needs, such as:

- Develop three 3D rural crash data visualization modules to interpret and visualize the rural crash data dynamically;

- Develop a new Bayesian vector auto-regression based data analytics approach to enable mixed-frequency rural crash data interpretations with missing values; and
- Develop a finite mixture random parameters model to explore driver injury severity patterns in low-visibility-related crashes.

The analytical results of the rural crash data records will greatly facilitate active countermeasure development to minimize crash risks and severities in RITI communities. To our best knowledge, based on a thorough literature search, there is no existing literature focusing on investigating the driver injury severity patterns in low-visibility-related crashes considering finite mixture random effects, and on interpreting with missing values, which motivated us to conduct fundamental methodological analysis for rural crash characteristics in RITI communities.

1.3. Research Objectives

This project aimed at improving the data-driven baseline crash data platform, developing a statistical model to handle the missing data issue, and proposing a novel finite mixture random parameter model for driver injury severity analysis in RITI communities. Towards this goal, the research objectives were as follows:

- Update rural crash data from multiple Departments of Transportation of the RITI communities for crash analysis.
- Develop the novel interactive crash analysis tools onto the onstreetmap-based online rural crash data platform for crash attribute interpretation and visualization.
- Develop a Bayesian vector auto-regression based data analytics approach to enable mixed-frequency rural crash data interpretations with missing values.
- Develop a finite mixture random parameters model to explore driver injury severity patterns in low-visibility-related crashes.

1.4. Report Organization

The remainder of this report is organized in the following manner.

Chapter 2 presents a comprehensive review of previous studies that are relevant to this study, including studies focusing on crash modeling, characteristics in crash modeling, and other critical issues, such as temporal instability and missing data. Chapter 3 briefly describes the rural crash data visualization platform proposed during Year 1 project and presented three new 3D rural crash data visualization modules interpret and visualize the rural crash data dynamically. Chapter 4 proposes a new Bayesian vector auto-regression based data analytics approach to enable mixed-frequency rural crash data interpretations with missing values. Chapter 5 presents a finite mixture random parameter model to explore driver injury severity patterns in low-visibility-related crashes. Finally, Chapter 6 presents the conclusion of this research and the recommendations for future research.

CHAPTER 2. LITERATURE REVIEW

2.1. Crash Data Modelling

In the past decades, numerous approaches have been taken to identify the contributing factors affecting driver injury severity in highway single-vehicle crashes. Due to the discrete nature of injury severity outcomes (i.e., no injury, possible injury, evident injury, severe injury, and fatality), multinomial logit models have been widely applied to investigate effects of significant factors in single-vehicle crashes (Savolainen and Mannering, 2007; Shankar and Mannering, 1996; Xie et al., 2012). Alternatively, nested logit models have also been employed to partially address the endogenous correlations among different severity outcomes (Nassar et al., 1994; Islam and Mannering, 2006; Wu et al., 2016b). Moreover, considering the intuitive ordering of injury outcomes (i.e., from no injury to severe injury and fatality), ordered logit and probit models are used (Rifaat and Chin, 2007; Lee and Li, 2014; Fountas and Anastasopoulos, 2018).

Above mentioned models provide a good understanding of contributing factors associated with injury severity outcomes in single-vehicle crashes. However, parameters in these models are estimated as constants and can hardly capture unobserved heterogeneity across observations. A significant number of factors affecting crash severity are not available in post-crash observation, such as the mental status of deceased drivers, or not included in the crash records, such as dynamic traffic flow conditions. Unobserved factors are correlated with both the crash outcome and observed factors. These factors thus lead to potential variations in the impacts of observed ones on crash severity, which constitute unobserved heterogeneity (Mannering et al., 2016). Recently, unobserved heterogeneity received growing concern. Random parameter approaches and their variants, such as random parameter models (Behnood and Mannering, 2015, 2016; Gong and Fan, 2017; Kim et al., 2013; Li et al., 2019b; Seraneeprakarn et al., 2017; Wu et al., 2014; Wu et al., 2016b), latent class models (Behnod et al., 2014; Shaheed and Gkritza, 2014; Xie et al., 2012; Li et al., 2019b), latent class models with random parameters within classes (Li et al., 2018b), random parameters ordered probability models (Fountas and Anastasopoulos, 2017, 2018; Fountas et al., 2018b; Fountas et al., 2019; Yu et al., 2019), and mixed logit models with heterogeneity in means and/or variances (Alnawmasi and Mannering, 2019; Behnood and Mannering, 2017a, 2017b; Seraneeprakarn et al., 2017), have been the most frequently used methods to cope with the unobserved heterogeneity in single-vehicle crash severity analysis (see Mannering et al. (2016)).

2.2. Impact Factors in Crash Data Analysis

Factors affecting the severity of single-vehicle crashes, such as crash exposures, road geometries, and driver features, have been explored extensively in previous studies (Behnood and Mannering, 2015; Gong and Fan, 2017; Kim et al., 2013; Lee and Mannering, 2002; Lee and Li, 2014; Li et al., 2018b; Wu et al., 2016a, 2016b; Xie et al., 2012). For example, Lee and Mannering (2002) modeled the severity of run-off-road crashes, considering a combination of temporal indicators, driver status, environmental characteristics, and roadway conditions. Xie et al. (2012) investigated the impact factors for rural single-vehicle crashes. Compared to Lee and Mannering (2002), additional information such as crash types, lighting conditions, and in-vehicle protections, were involved in Xie's model (2012). Kim et al. (2013) analyzed unobserved heterogeneous effects of drivers' age and gender on injury severities in single-vehicle crashes.

Moreover, previous studies showed that driving in the rain may be associated with higher crash risk than clear weather (Jung et al., 2010). A sizable portion of severe traffic crashes is brought about by these issues and induces significant fatalities and serious injuries. According to the Texas Department of Transportation (TxDOT, 2016), 16,818 rural crashes (159 fatal crashes) occurred under rain conditions in 2015, which is four times as many as those related to all other inclement weather conditions (e.g., blowing sand, sleet, and hail). Also, crash statistics from Arkansas and Oklahoma (Arkansas Department of Transportation, 2015; OKDOT, 2015) showed that single-vehicle crashes under rain conditions, especially those that occurred in rural areas, have a probability of drivers being seriously injured approximately twice as high as that for multi-vehicle crashes that occurred under the same or similar conditions. However, in most traffic safety studies, weather conditions have been considered as a contributing factor in crash cause-effect analysis, and only a limited number of studies directly focused on crashes under rain conditions. Andrey and Yagar (1993) analyzed the crash risk during and after rain events in urban areas. They discovered that the overall crash risk under rain conditions is 70% higher than that in average-day clear conditions. Jung et al. (2010) developed two types of polychotomous response models to analyze rain-related crashes in Wisconsin and concluded that rain-related factors could significantly affect injury severity. However, the safety impacts of rain and other variables in rain-related crashes are found unstable among different studies. For instance, a study examining the temporal and spatial distribution of rain-related crashes in Texas suggested that rain is a contributor to fatal crashes only in few dry counties but has no impacts on crashes in some of the wetter counties (Jackson and Sharif, 2014). Qiu and Nixon (2008) reported that rain is associated with higher injury severity and crash rates. Feng et al. (2016) concluded that severe accidents are about twice as likely to occur on curved roadways on rainy days, although straight and curved roadways have similar impacts on clear days. Shaheed et al. (2016) also reported that gender, seating position, road junction type, and other risk factors have different effects on injury severity in weather-related (rain, snow, blowing sand, etc.) and non-weather-related crashes. Whereas in the article by Lee et al. (2015), estimation results showed that injury severity is relatively lower under rain conditions in all crash types since drivers tend to reduce their speeds and be more careful on a wet surface. The sophisticated influences of rain on overall traffic safety indicate that there is a need for detailed analyses regarding external weather conditions and collision types.

2.3. Issues in Crash Analysis

Limited length of available data is a common issue that frequently appears in developing areas (and might affect developed regions as well). This problem arises when the information either has stopped being measured or has just recently started to be measured (Li et al., 2013; Tan et al., 2013; Van Lint et al., 2005). Additionally, the discontinuity problem arises when the data has its measurement approach changed without applying the new approach to historical observations. The data properties might be significantly affected due to the new approach. Moreover, when collecting and processing raw data, measurement defects may result in missing observations at irregular intervals. Series with data frequency switches can also be seen as a particular case of series with missing observations (Duan et al., 2016; Xiaolei Ma et al., 2015; Tang et al., 2015; Zhong et al., 2004). In the context of multivariate traffic safety analysis, these issues become further problematic, i.e., different variables may have dissimilar reporting frequencies (e.g., the unemployment rate is reported monthly, the total road length is reported annually, etc.) and even randomly reporting intervals, resulting in an irregularly-spaced mixed-frequency dataset.

Another troublesome problem in present traffic safety studies is that, as suggested in an abundance of relatively recent research, the influence of factors affecting crash occurrence may not be stable across both temporal and spatial domains (Bauer et al., 2016; Behnood and Mannering, 2016, 2015; Blazquez and Celis, 2013; Cheng et al., 2017; Dong et al., 2018; Feng et al., 2016; Khan et al., 2015; Kweon, 2011; Li et al., 2019b; Mannering, 2018; Peng et al., 2017; Xu et al., 2014; Yu et al., 2015). There is a growing body of studies using different methods to address the temporal instability issue. For instance, Malyshkina et al. (2009) used Markov switching models with estimated crash models alternating between two states overtime to capture for temporal instability in the dataset. Similar models have also been developed in other contemporaneous articles (Malyshkina and Mannering, 2010, 2009). Chen et al. (2018) proposed a modified mixed logit model to estimate a real-time refined-scale traffic crash dataset. Results underscored and confirmed the significant temporal impacts on crashes imposed by real-time traffic conditions and environment characteristics. Hierarchical Bayesian models have also broadly developed to address this (Dong et al., 2016; Khan et al., 2015; Li et al., 2019a, 2018a; Xu et al., 2014; Yang et al., 2018; Yu et al., 2013; Zeng et al., 2017). For example, Li et al. (2019) proposed a hierarchical Bayesian spatiotemporal random parameters approach to analyze the potential temporal instability in the crash dataset of Idaho. Ma et al. (2017) developed a Bayesian multivariate space-time model to study the model crash frequencies of different injury severity levels. Liu and Sharma (2018) used a multivariate spatiotemporal Bayesian model to investigate crashes with different severities and showed that temporal correlations were significant over time.

2.4. Summary

Recent studies on crash modeling, impact factors on crash injury severity, and some other critical issues in the crash analysis were reviewed in this section. It was found that crash analysis in adverse scenarios was in urgent need, as well as the temporal instability of crash characteristics. In this study, the project team will investigate the Bayesian vector autoregression-based approach for mixed frequency crash data interpretations with missing values; and develop a finite mixture random parameter models to explore driver injury severity patterns and causes in low visibility conditions. The proposed research enabled effective traffic safety program management at all levels in RITI communities to design and implement appropriate countermeasures to mitigate rural crash severities and risks.

CHAPTER 3. RURAL CRASH DATA VISUALIZATION WITH INTERACTIVE MODULES

This chapter of the report presents a brief description of the rural crash data visualization platform created during the Year 1 project and three different interactive rural crash data visualization modules will be emphasized to interpret and visualize the rural crash data dynamically.

3.1. Rural Crash Visualization Tool

The Rural Crash Visualization Tool System (RCVTS) starts with a login page, and users can log in or register for an account. Except for the administrator, three types of user authority were defined, as shown in Table 3-1. The design of RCVTS is quite straightforward, following the guidelines of "overview first, filter, visualization, details-on-demand, and then download" (Shneiderman, 1996). On this page, the description area is replaced with the functional area. The three primary functions—i.e., data visualization, data analysis, and data retrieval—are located under different tags. Under the data visualization tag, RCVTS provides the users with a comprehensive filter option including filter type, crash information, environmental condition, passenger condition, and a timeline. A significant feature of RCVTS is that all these seemingly independent components are tied together. Once the filtering condition is submitted under the data mapping tag, selected crashes records will be presented in the embedded map. The data analysis and retrieval process applied to the crash data set presented on the map directly, i.e., the filter results are shared within the three components.

Table 3-1 Authorities for different users

User Type	Target User	Available Data	Function
I	Public User	3-year-data (2010-2013)	Static Plot
II	Registered Researcher	All Data	Data Mapping Static & Interactive Graph
III	Authorized Researcher Related Officials	All Data	Data Mapping Static & Interactive Graph Data Retrieve

The RCVTS obtained three primary functions, i.e., crash visualization, crash data analysis, and crash data retrieval. For crash visualization, the RCVTS provides a rich set of filter options. As summarized in Table 3-2, the filter options are in the four categories. Firstly, users are required to choose the filter type, i.e., by area or by road. When users want by region, they have the option to query the database based on the state of the crash, county, and the city town the accident occurred. Otherwise, the user has the option to query the database based on the road type and road name. RCVTS provides three road types, i.e., city/street, state route, and county road. If a user chooses the city/street option, they can query based on the name of the primary roadway. If they choose the state route option, users can query based on the state route id. If the county road option is chosen, the user can query based on the county road number. The RCVTS populates the possibilities using a php program dynamically querying all the options based on the dataset.

Table 3-2 Summary of filter options provided in RCVTS

Filter Group	Subfilter	Filter Options
Map Filter[a]	By Area	State, County, City/Town, etc.
	By Road	Road type, Road name
Crash Information		Severity, First collision type, Second collision type, Number of involved vehicles, Number of involved users, Major contribution, etc.
Environmental Conditions		Weather, Road surface condition, Light condition
Passenger/User Condition	Driver 1	Gender, Age, Vehicle Type, Injury Type, Seat Position, Alcohol test result, etc.
	Driver 2[b]	Gender, Age, Vehicle Type, Injury Type, Seat Position, Alcohol test result, etc.
	Driver 3[b]	…

[a]Users shall choose a filter type from either “By Area” or “By Road.”

[b]The number of Driver information here depends on the number of involved vehicles entered in crash information filter.

Figure 3.1 illustrates typical query results. A successful query indicates that the database contains data that meets the filter conditions. As shown in Figure 3.1, in the desired area, 21046 crash records are identified. If no crash records satisfy the recent query, the interface generates a popup to inform the user. The popup presents a summary of submitted filter conditions for user’s convenience, as shown in Figure 3.2.

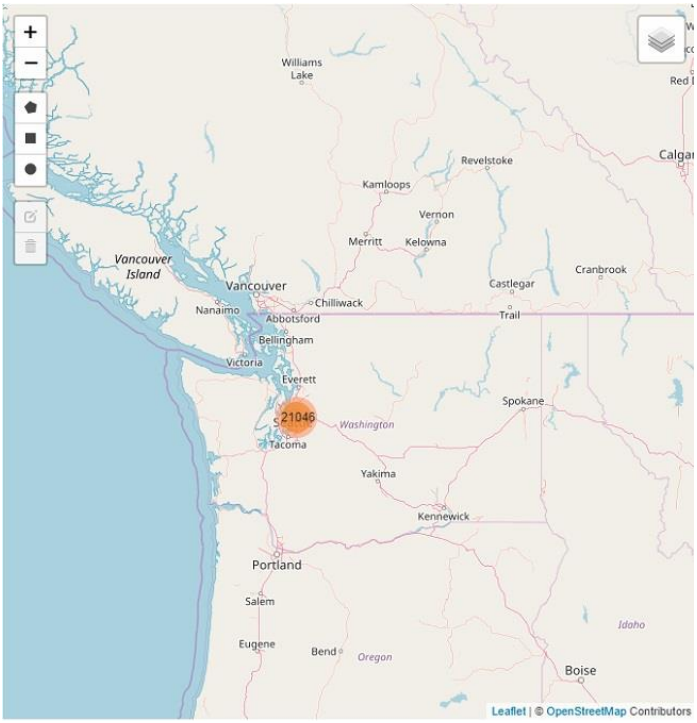


Figure 3.1 Successful query result for RCVTS

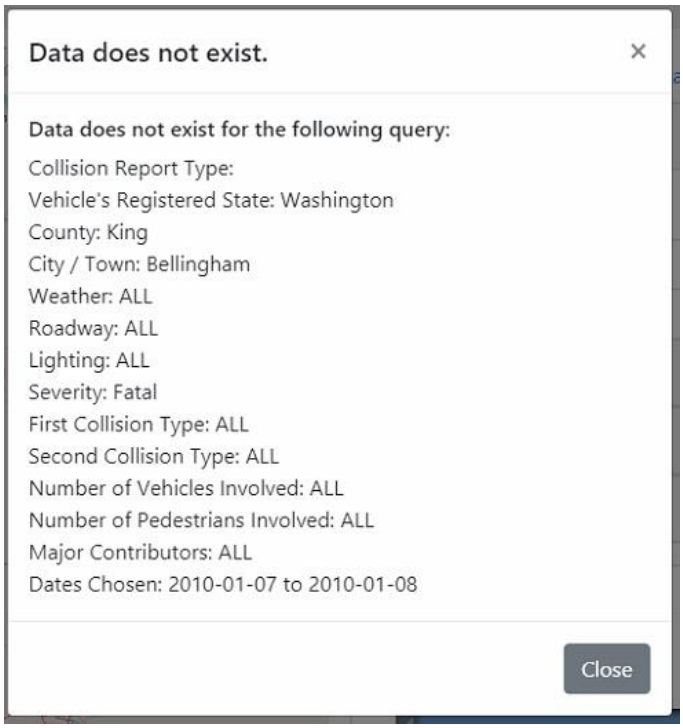


Figure 3.2 Pop-up with failure information.

Occlusion has been an issue when dealing with nearby crash records because it is challenging to count overlapping points. Wongsuphasawat solved this problem using the hot mode, in which regions with the most influence are colored in reds while less active areas are colored in blues (Wongsuphasawat, 2009).

In RCVTS, the solution is quite straightforward: crash records are grouped into specific clusters with a label indicating the total number of crashes in this cluster (see Figure 3.1). This solution is beneficial for the following reasons:

- It reduces the overhead cost for creating each marker on the map.
- It prevents the overlapping of multiple markers
- The labeled number illustrates the density of crashes directly.

In this case, the RCVTS proposes the zoom-in function and get a more detailed distribution of crashes in this area. As shown in Figure 3.3, the color for different cluster represents the crash counts in a hot mode. When it cannot be zoomed anymore, each marker presents a crash, and by clicking the crash mark, the interface provides crash-related information to the user, as shown in Figure 3.4. Note that, in RCVTS, zooming in can be achieved either by scroll or by double-clicks. To enhance the flexibility of crash data selection, RCVTS also provides a graphic query tool. More specifically, users can choose a specific marker on the map to reshape the area; this allows the user to select crashes in the designated area and remove all crash records outside of that area. There are three types of marker shapes—i.e., including the polygon, the square and the circle, as shown in Figure 3.5.

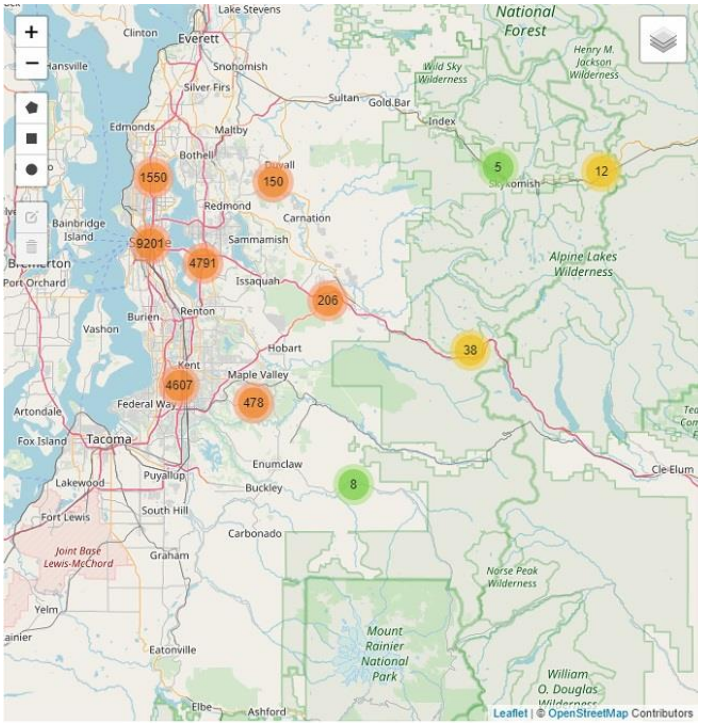


Figure 3.3 Zoom in result in crash query.

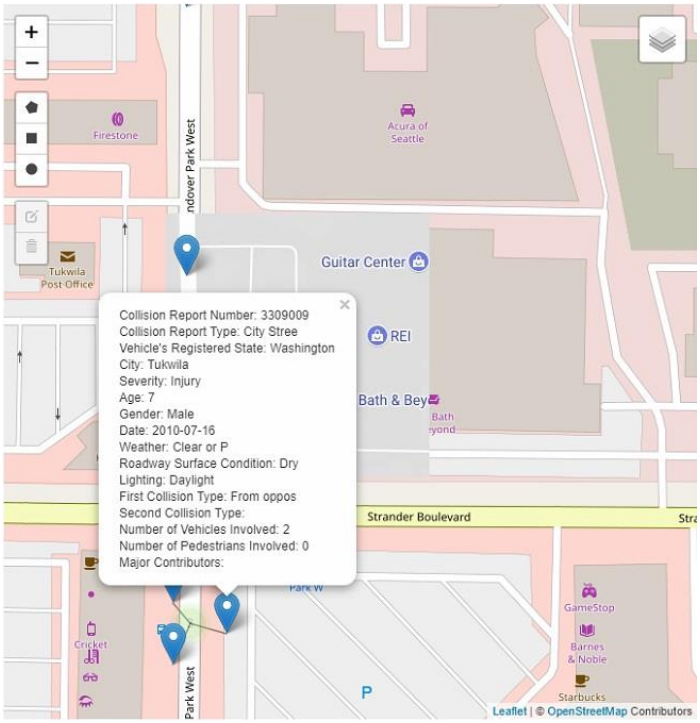
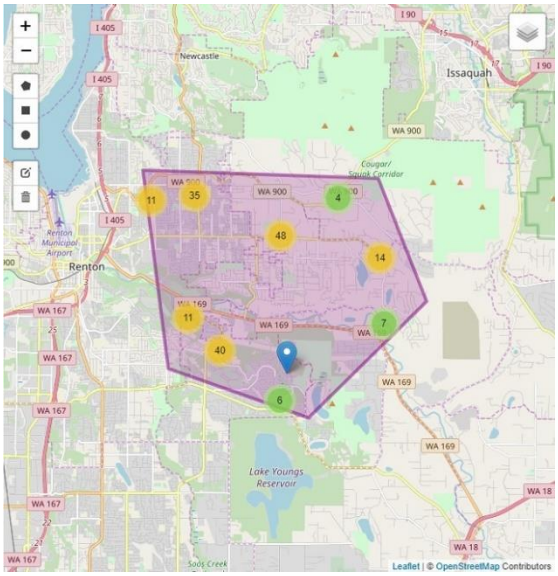
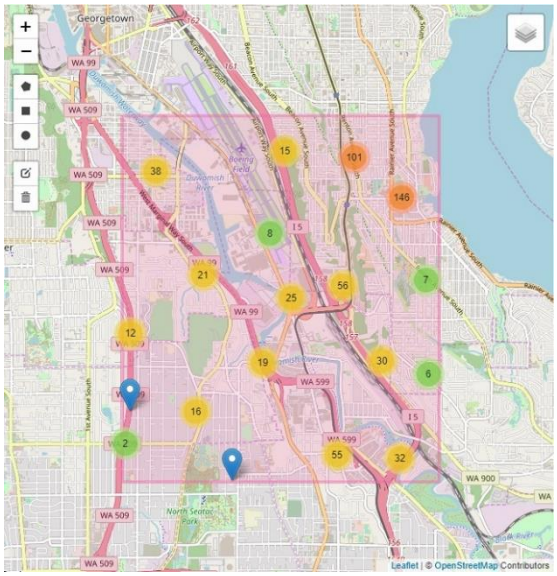


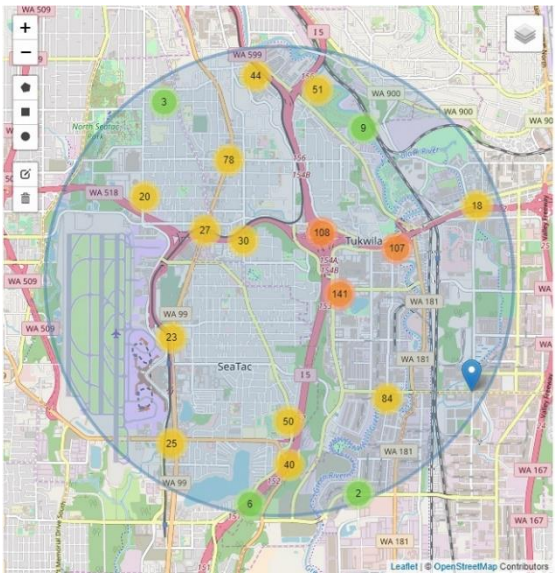
Figure 3.4 Crash detail shown in map-based interface.



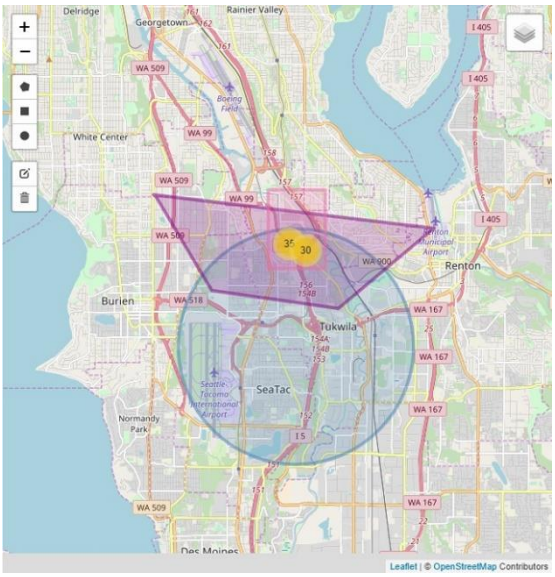
(a) Polygon



(b) Square



(c) Circle



(d) Intersection union

Figure 3.5 Result of graph query tool in RCVTS.

As for the crash data analysis part, the original RCVTS proposed a set of visualization approaches, including static charts—i.e., the scatter chart—the line chart, the area chart, the bar chart, and an interactive graph—i.e., the sunburst chart, as illustrated in Figure 3.6 to Figure 3.10. The interface allows users to generate customized analytical graphs by specifying the parameters and scale. These visualization tools—i.e., the scatter chart, the line chart, the area chart, and the bar chart—can be accessed by selecting the corresponding option located in the lower part under the data visualization tag. For example, to generate a line graph, users can examine the crash counts, fatality counts, and injury crashes. Users choose the parameter of interest; then they may select the time scale displayed on the graph, e.g., daily, monthly, or yearly. Moreover, in order to reduce the anxiety of waiting, after pressing the create button at the bottom, animation occurs for each point of the line graph.

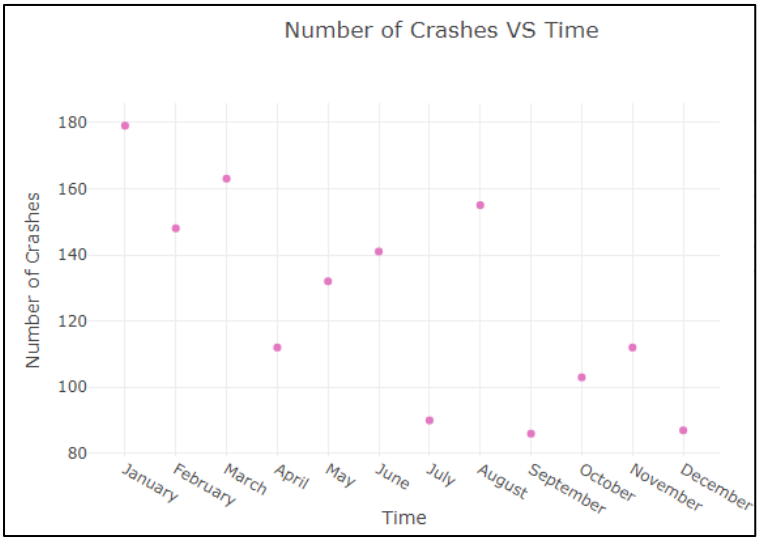


Figure 3.6 Scatter chart sample generated in RCVTS.

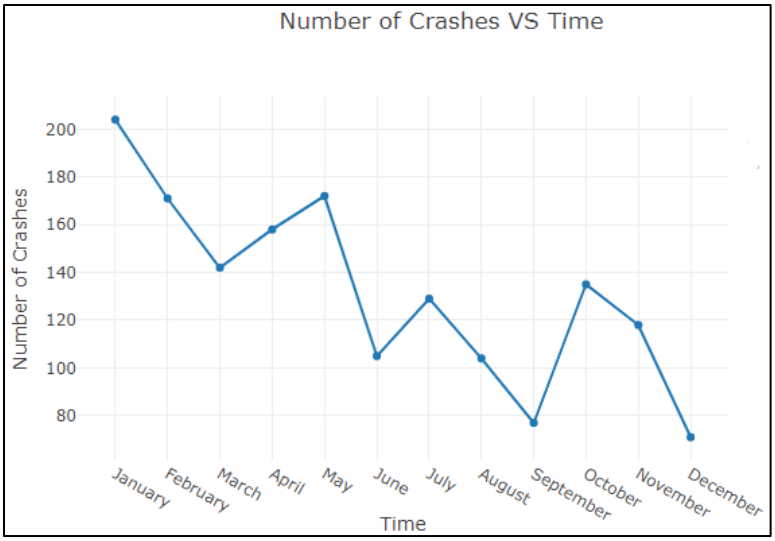


Figure 3.7 Line chart sample generated in RCVTS.

As mentioned before, in the proposed RCVTS, only authorized users can download the selected crash data in a comma-separated value (CSV) format with limitations, as shown in Table 3-3. Currently, access to the raw data is not provided even with authority. On top of the raw data, we plan to enable access to processed data generated in the visualization procedure.

Table 3-3 Data retrieval limitation

Limitation Type	Description
Frequency	5 queries per day
Quantity	maximum 50000 records per query
Accessible Information	time label, GPS, route name, crash type, severity level, weather condition, lighting condition, major contributing

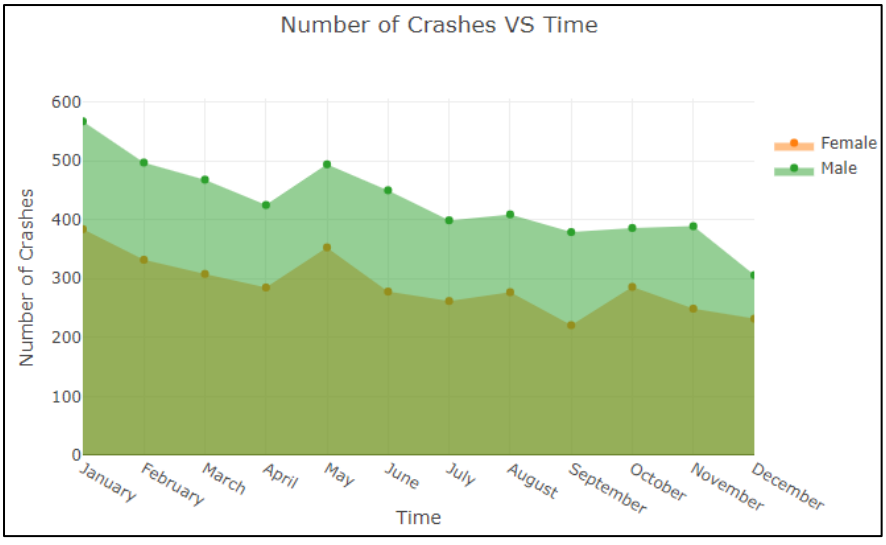
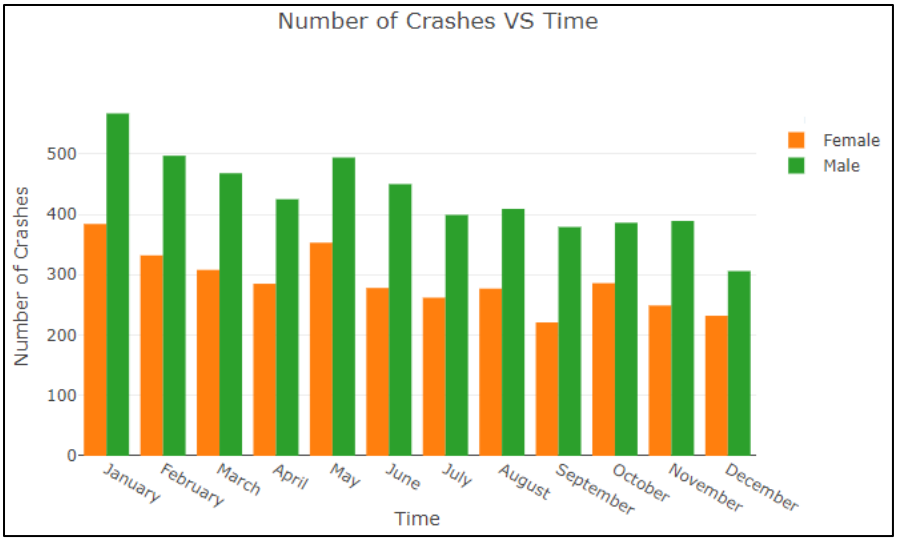
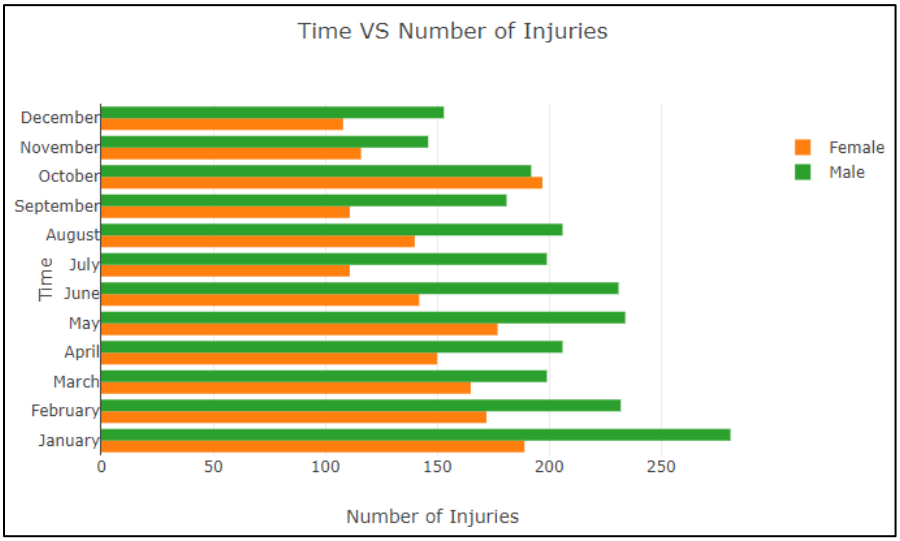


Figure 3.8 Area chart sample generated in RCVTS.



(a) Vertical bar chart



(b) Horizontal bar chart

Figure 3.9 Bar chart sample generated in RCVTS.

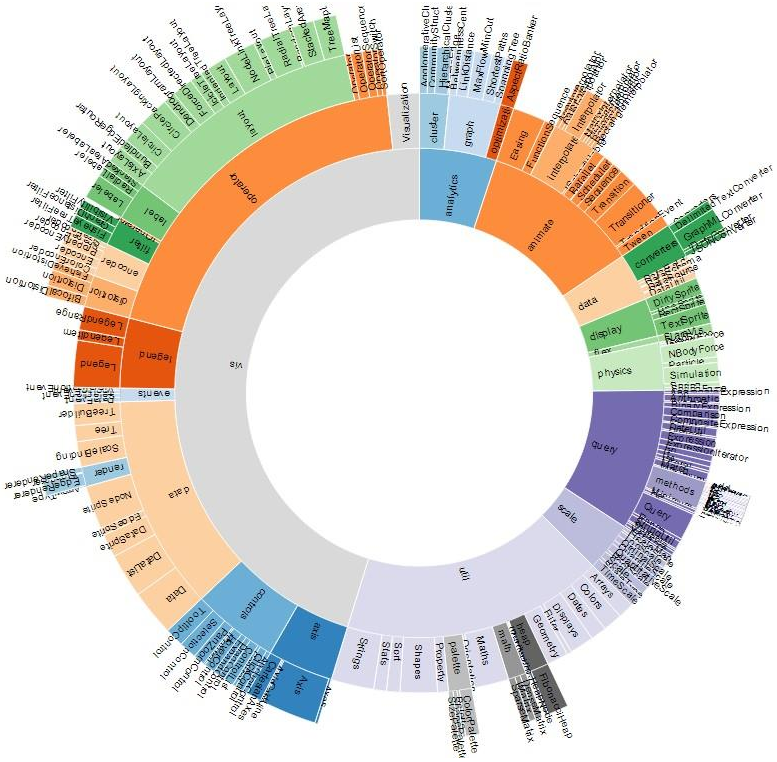


Figure 3.10 Sunburst chart sample generated in RCVTS.

3.2. Novel Interactive Rural Crash Data Visualization Modules

In the past year, the RCVTS was enhanced with three different interactive rural crash visualization modules: the double vertical graph, the collapsible force graph, and the interactive bubble graph.

3.2.1. Double Vertical Graph

The double vertical graph is a benefit for its flexibility in generating the relationships between multiple categories. For instance, as illustrated in Figure 3.11, the relationships among different collision severities and the collision locations are demonstrated in a single interactive graph. Moreover, by moving the mice to the specific collision severity or the particular county name, the collision information associated with the particular category is presented, as demonstrated in Figure 3.12.

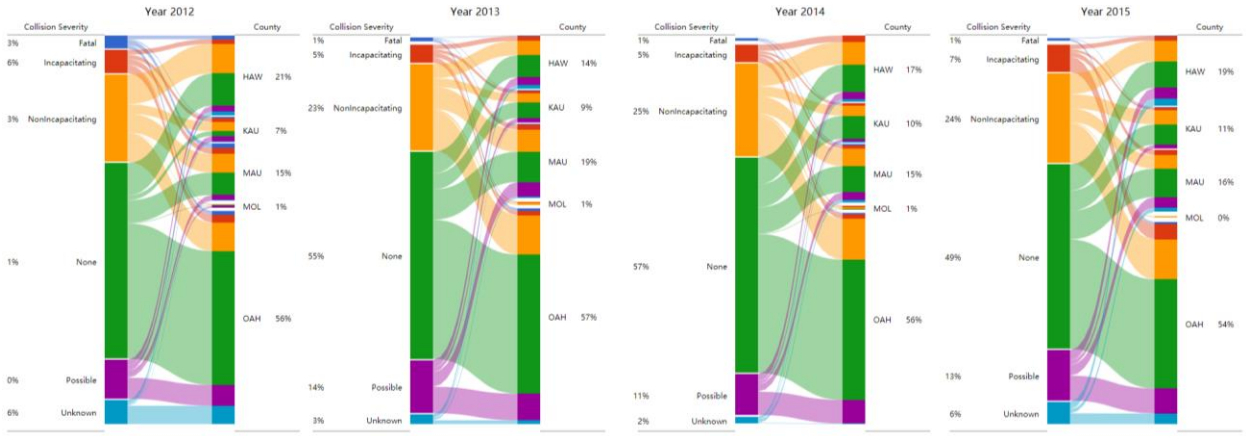
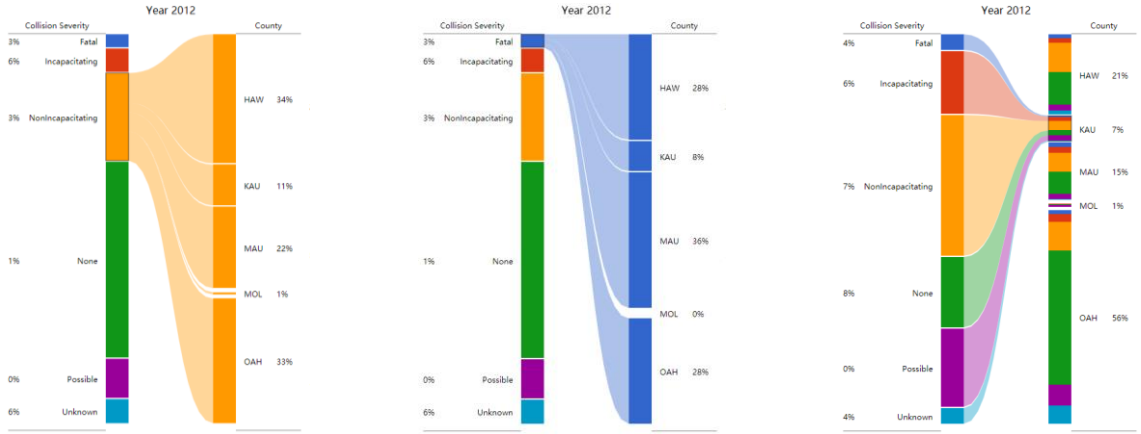


Figure 3.11 Double vertical graph sample generated in RCVTS.



(a) (b) (c)

Figure 3.12 Variations in double vertical graph

As illustrated in Figure 3.12(a), the non-incapacitating collision occupied 3% of the whole collisions, while the percentages in different islands are 34%, 11%, 22%, 1% and 33% in Hawaii, Kaui, Maui, Molokai, and Oahu, respectively.

3.2.2. Collapsible Force Graph

The proposed collapsible force graph is a special kind of chart used to display multi-item data related in a hierarchical, linear or mixed way, as a series of linked bubbles. The collision in different categories is represented using the circles, while the circle radiuses indicate the collision counts, or percentages, as shown in Figure 3.13. Meanwhile, these circles are expandable. As shown in Figure 3.14, the in-depth relationships among different categorical variables are illustrated. The force graph is a benefit for keeping the structure readable all the time. As shown in Figure 3.14(a) and (b), no matter how many circles are expanded, the overall structure remains stretched.

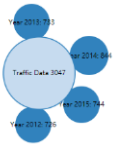
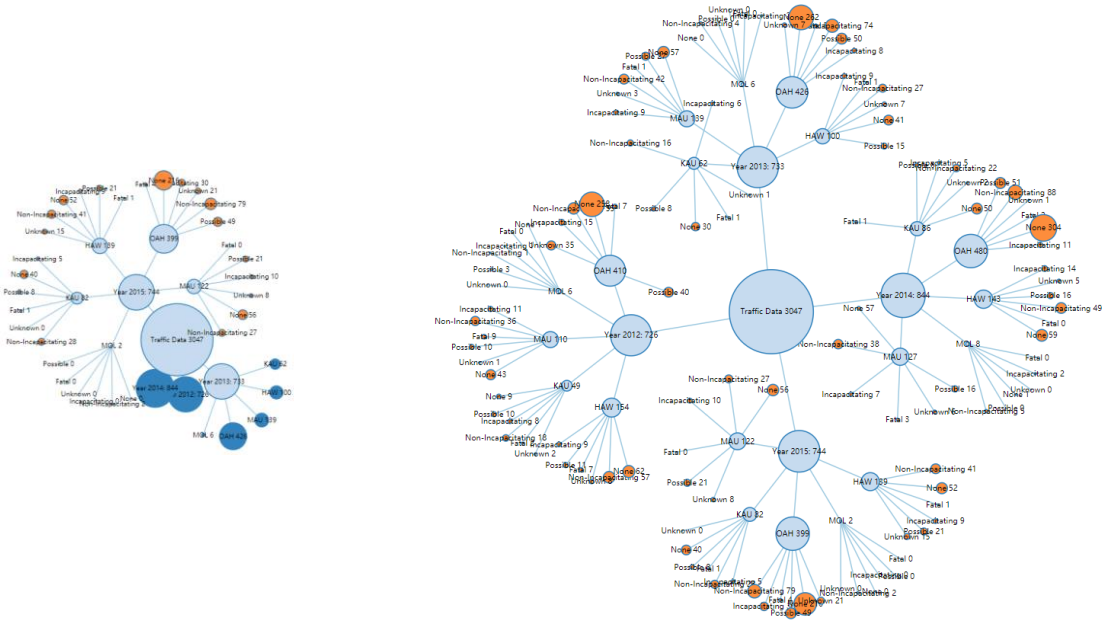


Figure 3.13 Collapsible force graph sample generated in RCVTS.



(a) (b)

Figure 3.14 Extended force graph sample generated in RCVTS.

3.2.3. Interactive Bubble Graph

The interactive bubble graph is, somehow, similar to the collapsible force graph. It also represents data related in a hierarchical or linear way, as a series of linked bubbles. The significant difference located that the collapsible force graph can involve more attributes, while the interactive bubble graph is a benefit for clearly demonstrating the relationship between limited characteristics.

As shown in Figure 3.15, crash records were separated into four parts, indicating different years. Figure 3.16 represents the expanded variation of the proposed bubble graph. Moreover, the detailed information is presented directly.

Figure 3.15 Interactive bubble graph sample generated in RCVTS.

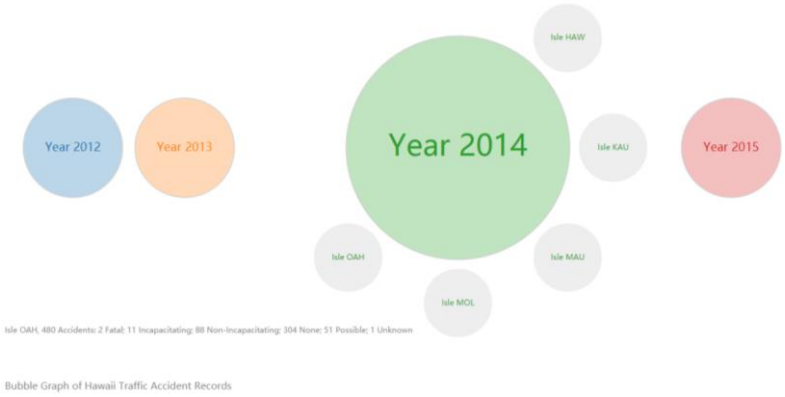


Figure 3.16 Variations in interactive bubble graph

3.3. Summary

The project updated the RCVTS, a web-based tool that aims to deal with visualization issues associated with various rural crash characteristics. The updated RCVTS features three new graph types. The RCVTS allows users access to traffic data stored in the database, and to create highly customized analytical graphs. Currently, traffic crash data collected in the northwest region— i.e., Alaska, Idaho, and Washington—were shared online through a MySQL database using the phpMyAdmin technique. RCVTS regulated three levels of users with different access to the database and visualization tools. The three significant functions provided in RCVTS were traffic data visualization, data analysis, and retrieval of corresponding data. More specifically, in traffic data visualization procedure, a combination of conditional filter and map-based graph query provided the users a flexible data query environment; in the analysis part, different tools were produced based on the type of data.

The researchers hope that the revised RCVTS application will help transportation professionals spend less time in crash data analysis and inspire their creativity to investigate the underlying relationships among various parameters. More endeavors are also underway to enhance both the depth and width of the RCTVS. In future updates, the tool users will be allowed to upload their crash data onto the new RCVTS. Accordingly, RCVTS would be able to help those professionals as a crash data visualization tool,

for not only the data provided in the database but also their data. In this case, the RCVTS can be used directly in their research or practical applications.

CHAPTER 4. A BAYESIAN VECTOR AUTOREGRESSION-BASED DATA ANALYTICS APPROACH

In this chapter, we propose a Bayesian vector autoregression-based data analytics approach to enable irregularly-spaced mixed-frequency traffic collision data interpretations with missing values. The proposed mixed-frequency VAR model is more innovative than prevailing models for analyzing collision data in the following aspects: (1) it can handle irregularly-spaced mixed-frequency data without simply filling time series gaps or reducing the sample size; (2) the model can extensively capture for unobserved heterogeneity within the unaggregated dataset by introducing the random effects term; (3) all contributing factors are considered as endogenous variables, and therefore the troublesome endogenous issue can be addressed. To the best of our knowledge, this is the first time such an econometric model is proposed and estimated in the field of traffic safety research. The rest of the chapter is organized as follows: Section 4.2 provides an explicit description of the dataset. The detailed methodology design is described in Section 4.3. The model analysis results and discussions are illustrated in Section 4.4. Finally, the entire research effort is concluded in Section 4.5.

4.1. General Background

Due to the inevitable casualties and economic losses caused by motor-vehicle crashes, numerous studies have been developed to figure out contributing factors to traffic crash occurrence and provide corresponding countermeasures to reduce the likelihood of traffic crashes (Chen et al., 2016; Ma et al., 2015). An abundance of relevant factors, including traffic (e.g., freeway mileage, daily vehicle miles traveled), road (e.g., total road length, road alignment, road profile), weather (e.g., precipitation, average temperature), demographic (e.g., population density, percent of male population, average education level), and macroeconomic (e.g., GDP growth, unemployment rate, median household income) characteristics, have been extensively examined using different analytic approaches (Anastasopoulos, 2016; Bhat et al., 2014; Chen et al., 2014; Chen and Chen, 2011; Chiou et al., 2014; Chiou and Fu, 2015; Eustace et al., 2015; Li et al., 2019a, 2018b; Mothafer et al., 2016; Venkataraman et al., 2014; Zeng et al., 2016). Thanks to numerous open data government agencies and organizations, for instance, Fatal Accident Reporting System (FARS), Highway Performance Monitoring System (HPMS), United States Census Bureau, etc., the majority of this data could be attained by request or even directly downloaded, providing convenience for the traffic safety study. However, these comprehensive sources of data may contain conflicting information and may adversely impact research results and even limit the accuracy of inferences and predictions as well. As introduced in Section 2.3, numerous studies have been proposed to investigate the temporal instability in crash analysis. Although these models are advanced compared to conventional models (for instance, Poisson (Miaou, 1994), negative binomial (Shankar et al., 1995), Gamma (Daniels et al., 2010), etc.,) for analyzing crash-occurrence data, none of them can be developed on an irregularly-spaced mixed-frequency dataset. Also, due to the structure of these count-data models, the modeling processes are not suitable to accommodate any endogenous-variables correction techniques, and therefore may induce endogenous problems and lead to bias estimates.

Under such a complicated framework, some existing models might compromise the statistical inference. Trimming out data with mechanical approaches without formal statistical support can undoubtedly produce erroneous observations that are contrary to the actual situation. Alternatively, and more commonly, some studies considered reducing the time series frequency of the whole dataset to avoid dealing with mixed-frequency issues. Coarsely decreasing the sample size of data with high rates to

accommodate data with lower frequencies can lead to the loss of relevant information in the more top frequency data. Furthermore, aggregating the high-frequency data into the lower rate may also introduce potential temporal instabilities, as variables (e.g., temperature, monthly vehicle miles of travel (MVMT), etc.) may shift significantly over time while they do not demonstrate significant trends in the aggregated period (Mannering, 2018).

Rather than merely pinpointing individual values to fill gaps or roughly aggerating data into a period, the literature in the economic field has evolved into the development of Bayesian Gibbs samplers to recover the entire joint distribution of the missing observations (Alves and Fasolo, 2015). In light of the articles of Schorfheide and Song (2015) and Eraker et al. (2014), we developed a Bayesian mixed-frequency vector autoregression (VAR) to deal with the irregularly-spaced mixed-frequency traffic collision dataset. The VAR model is a frequently used tool in applied macro-econometrics. A VAR is a multivariate time series model that can be used, for instance, to forecast individual time series, to analyze the sources of economic cycle fluctuations, or to assess the effects of policy interventions on the macroeconomy (Schorfheide and Song, 2015). The Bayesian mixed-frequency VAR is a recent extension of the traditional VAR by assuming that a VAR with unknown parameters can describe the dynamics of the multivariate time series. The proposed mixed-frequency VAR can be conveniently represented as a state-space model, in which a VAR gives the state-transition equations at high frequency. The state vector is utilized to reserve measurement equations related to the observed series to the underlying, potentially unobserved, monthly variables (Schorfheide and Song, 2015).

4.2. Data

The collision dataset contains all collisions of different severities that occurred on the state highways in Washington State from January 2006 to December 2016. As shown in Figure 4.1, the state highways of Washington comprise a network of over 7,000 miles of state highways, including all interstate and U.S. highways that traverse through the state, maintained by the Washington State Department of Transportation (WSDOT) (Washington State Department of Transportation, 2018). The state highway system spans 8.8% of the state's public road centerline miles; nevertheless, it carries 56.2% of vehicle miles traveled (VMT) and 44.7% of traffic collisions (Washington State Department of Transportation, 2016). Transportation characteristics, including centerline miles, lane miles, and monthly vehicle miles traveled (MMVT), are obtained from WSDOT. Weather data, including average precipitation and average temperature, are downloaded from the Office of the Washington State Climatologist (Office of the Washington State Climatologist, 2019). Other socioeconomic variables including total income, population, unemployment rate, and GDP are extracted from the Bureau of Economic Analysis (Bureau of Economic Analysis, 2019), Bureau of Labor Statistics (Bureau of Labor Statistics, 2019), and Federal Reserve Bank of St. Louis (Federal Reserve Bank of St. Louis, 2019), respectively. Table 4-1 presents the statistical results of each variable. The detailed data in the time series is illustrated in Figure 4.2.

Table 4-1 Summary of variables and descriptive statistics.

Variable	Time Series	Description	Mean	SD	Min	Max
Total Collisions	Monthly	Continuous from January 2006 to December 2016	4044.47	633.38	2884	5998
Fatal Collisions	Monthly	Continuous from January 2006 to December 2016	19.61	5.75	9	33

Variable	Time Series	Description	Mean	SD	Min	Max
Serious Injury Collisions	Monthly	Continuous from January 2006 to December 2016	68.03	14.54	41	111
Minor Injury Collisions	Monthly	Continuous from January 2006 to December 2016	1303.39	213.30	887	1845
PDO Collisions	Monthly	Continuous from January 2006 to December 2016	2653.36	448.61	1877	4081
MMVT (by natural logarithms)	Monthly	From January 2005 to December 2017, except for the missing values for the full year of 2007	21.70	0.10	21.46	21.92
Centerline Miles (thousand mile)	Annually	Continuous from 2005 to 2016	70.52	0.67	70.42	70.61
Lane Miles (thousand mile)	Annually	Continuous from 2005 to 2016	185.77	1.26	183.63	187.15
Temperature (F)	Monthly	Continuous from January 2005 to December 2018	47.48	12.52	25.11	68.97
Precipitation (in)	Monthly	Continuous from January 2005 to December 2018	3.81	2.75	0.04	14.10
Unemployment Rate (%)	Monthly	Continuous from January 2005 to December 2018	6.47	1.91	4.00	11.30
Total Income (USD, by natural logarithms)	Quarterly	Continuous from 2005 Q1 to 2018 Q4	26.50	0.19	26.16	26.86
GDP (USD, by natural logarithms)	Quarterly	Continuous from 2005 Q1 to 2018 Q4	26.71	0.18	26.39	27.07
Population (by natural logarithms)	Annually	Continuous from 2005 to 2018	15.74	0.06	15.65	15.84

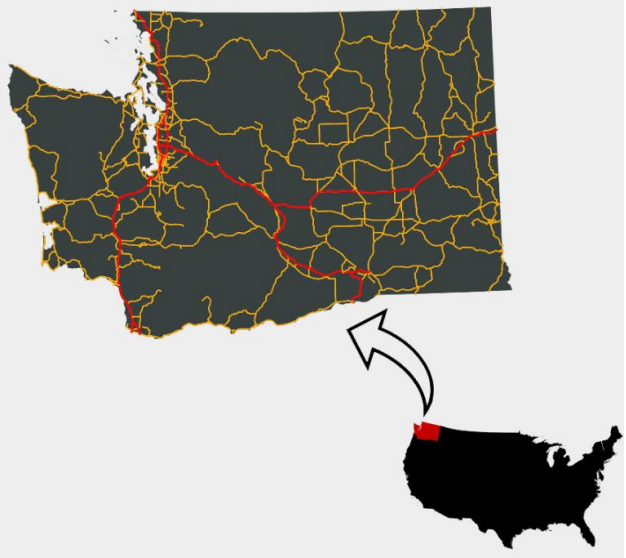


Figure 4.1 Geographical illustration of study area.

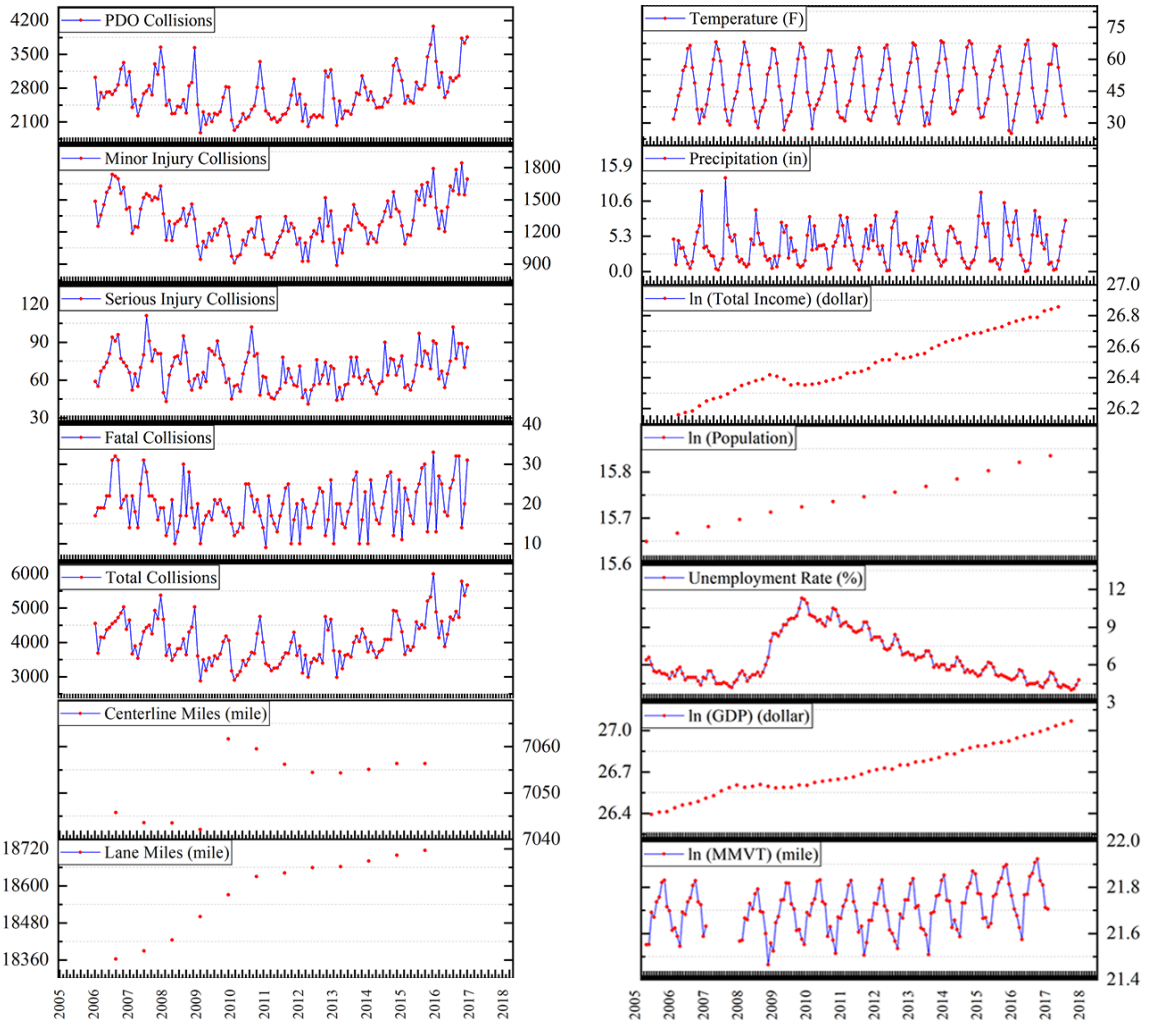


Figure 4.2 Illustration of dataset in time series.

As shown in Table 4-1, as well as in Figure 4.2, variables have different temporal trends in study periods. Some variables, such as total collisions, average temperature, unemployment rate, etc., are reported monthly, while others, for example, total income, population, etc., are reported quarterly or annually. Also, variables are not in the same length, i.e., some series start later or end earlier. Furthermore, the variable, MMVT, is interrupted through the time series, resulting in multiple missing observations. Because of the nature of these variables, the entire data set becomes an unevenly-spaced mixed-frequency data, which poses tremendous challenges to the estimation using conventional models.

4.3. Methodology

4.3.1. State-Transitions and Measurement

The mixed-frequency VAR developed in this study is grounded on the standard fixed-parameter VAR in which the length of the time is one month. In response to the collision-frequency dataset's mixed observed, we assume that the model evolves at the highest available rate, i.e., monthly, which means that many high-frequency observations for low-frequency variables are merely missing data.

Correspondingly, the underlying high-frequency series of the low-frequency variables can be considered as the latent states of the system, and this treatment is naturally applicable to the state space representation of the system high-frequency and low-frequency observed variables.

Firstly, let the $N \times 1$ vector $y_t \in Y \equiv \{y_1, \dots, y_T\}'$ denote the endogenous variable vector containing all observations at time t follow a VAR (p) dynamics:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \phi_c + \mu_t,\ \mu_t \sim \text{iid} N(0, \Sigma) \tag{4-1}$$

where N (N =14 in this study) is the total number of all endogenous variables listed in Table 4-2, and Y is the set of all endogenous variables. Through this paper, $Y_{t_0:t_1}$ is used to denote the sequence of observations or random variables $\{y_{t_0}, \dots, y_{t_1}\}$. If no ambiguity arises, for the sake of simplicity, we sometimes drop the time subscripts and abbreviate $Y_{1:T}$ by Y. y_{t-i} ($i \in \{1, \dots, p\}$) is the ith lag of y_t, ϕ_i is a time-invariant $N \times N$ matrix of coefficients, ϕ_c is a $N \times 1$ vector of constants, μ_t is a $N \times 1$ vector of error terms, and Σ is a $N \times N$ positive definite covariance matrix. As the error terms follow a multivariate normal distribution with zero mean and covariance matrix Σ, the model is able to capture the unobserved heterogeneity (Mannering et al., 2016).

Furthermore, y_t can be composed into $y_t = [y'_{t,h}, y'_{t,l}]'$, where the $N_h \times 1$ vector $y_{t,h}$ collects variables that are fully observed at the highest frequency, for example, the monthly unemployment rate, while the $N_l \times 1$ vector $y_{t,l}$ involves the variables with missing data that are counted at a lower frequency, for instance, annually reported total miles of state highways. Note that the time t here takes the highest frequency, and the dimensions N, N_h, and N_l are time-invariant, i.e., $N = N_h + N_l$. The partition between the fully observed variable $y_{h,t}$ and the variable with missing data $y_{l,t}$ is give by

$$y_t = \begin{bmatrix} y_{t,h} \\ y_{t,l} \end{bmatrix} = \begin{bmatrix} \phi_{1,hh} & \phi_{1,hl} \\ \phi_{1,lh} & \phi_{1,ll} \end{bmatrix} \begin{bmatrix} y_{t-1,h} \\ y_{t-1,l} \end{bmatrix} + \cdots + \begin{bmatrix} \phi_{p,hh} & \phi_{p,hl} \\ \phi_{p,lh} & \phi_{p,ll} \end{bmatrix} \begin{bmatrix} y_{t-p,h} \\ y_{t-p,l} \end{bmatrix} + \begin{bmatrix} \phi_{c,h} \\ \phi_{c,l} \end{bmatrix} + \begin{bmatrix} \mu_{t,h} \\ \mu_{t,l} \end{bmatrix} \tag{4-2}$$

where

$$\begin{bmatrix} \mu_{t,h} \\ \mu_{t,l} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{hh} & \Sigma_{hl} \\ \Sigma_{lh} & \Sigma_{ll} \end{bmatrix} \right)$$

It is clear that the number of regressors in Eq. (4-2) is $q \equiv Np + 1$, thus the VAR has Nq coefficients, and the whole system has $Nq + N^2$ parameters.

Note that for simplicity, it is convenient to assume that y_t is recorded at two frequencies, as has widely seen in previous macroeconomic studies (Schorfheide et al., 2014). However, as we mentioned in the last section, in reality, the time series at lower frequencies may not be observed at regular intervals. For convenience, we regard the yearly reported variables as irregularly spaced quarterly observed variables with missing values. Therefore, we assume the existence of the set of observable variables $Y^o \equiv \{y_1^o, \dots, y_T^o\}'$, where y_t^o is a $N_t^o \times 1$ vector of observed endogenous variables, whose dimension N_t^o changes over time due to irregularly-spaced mixed-frequency. Note that it is possible that at certain periods, $N_t^o = 0$, i.e., no observations are available. Naturally, $N_t^o \leq N$ ($N_t^o < N$, if $N_t^o = 0$). More specifically, in light of the articles of Canova (2011) and Schorfheide and Song (2015), we then represent the $VAR(p)$ dynamics in Eq. (4-1) in a dynamic linear model (DLM) form, and rewrite Eq. (4-1) in a companion form:

$$z_t = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p'-1} & \phi_{p'} \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & I & 0 \end{bmatrix} z_{t-1} + \varepsilon_t = A z_{t-1} + \varepsilon_t,\ \varepsilon_t \sim \text{iid}\ N(G, \Omega(\Sigma)) \qquad \text{(4-3a)}$$

and

$$y_t^o = \begin{pmatrix} y_{t,h}^o \\ y_{t,l}^o \end{pmatrix} = \begin{pmatrix} I_{N_h} & 0 \\ 0 & M_{t,l} \end{pmatrix} \begin{pmatrix} I_{N_h} & 0 \\ 0 & \Lambda_{z,l} \end{pmatrix} z_t = M_t \Lambda_z z_t \qquad \text{(4-3b)}$$

where Eqs. (4-3a) and (4-3b) are the transition equation and observation equation, respectively. This companion form representation transforms the $VAR(p)$ model in a larger scale $VAR(1)$ model and it is convenient for computing moments and deriving parameter estimates. $z_t \equiv \left[y_t', \dots, y_{t-p'+1}'\right]'$ is a $Np' \times 1$ vector of states, A is a $Np' \times Np'$ matrix of coefficients for endogenous variables, $G = [\phi_c, \mathbf{0}]'$ is a $Np' \times 1$ matrix for constants, $\Phi = \left[\phi_1, \dots, \phi_{\mathrm{p}'}, \phi_c\right]'$, Ω is a $Np' \times Np'$ the positive semi-definite covariance matrix, and the $N \times N$ upper-left submatrix of Ω equals Σ and all other elements are zero. M_t is a $N_t^o \times N$ deterministic selection matrix, and Λ_z is a $N \times Np'$ transformation matrix converting high-frequency values into low-frequency.

Note that p' may not always be equal to p, i.e., the observed interval p' of latent values in high-frequency does not necessarily match the lags of VAR, p. We assume that (1) if $p' < p$, let $p' = p$ and extend Λ_z with zero matrices, i.e., $\Lambda_z = \left[\Lambda_z, 0_{N \times N(p-p')}\right]$; (2) if $p' > p$, add more zero matrices $\phi_i = 0_{N \times N}$, for $i \in \{p+1, \dots, p'\}$. In the context, M_t is a time-varying selection matrix where the number of rows is adjusted to match the number of absent observations in each time point. In the ideal case, we let the $M_{t,l}$ be the N_l identity matrix if all low-frequency variables are observed at time t so that $y_{t,l}^o = \Lambda_{z,l} z_t$. In the remaining periods, $M_{t,l}$ is an empty matrix such that $y_t^o = y_{t,h}^o$. In a more complicated case, some of the variables are unexpected missed at time t, for instance, the MMVT in this study. It is straightforwardly accomplished by simply treating the missing observations at $t = \{T^* + 1, \dots, T\}$ as regular missing data, and forecasting conditional on the observations that do exist at $t > T^*$, where T^* is the most recent time point at which all variables are observed. Therefore, by dropping the row of M_t that corresponds to the variable, whether it is observed at high or low frequencies, we can make draws from the posterior distribution of the missing variable.

4.3.2. Bayesian Inference

The Bayesian Mixed Frequency (BMF) estimator is an application of Gibbs sampling, which requires iterating over objects of interest and sampling them from known distributions conditional on the remaining objects. In the current setup, the starting point of Bayesian inference for the VAR is a joint distribution of observations $Y_{1:T}^o$, latent states $Z_{0:T}$, and parameters (Φ, Σ), conditional on a pre-sample $Y_{-p+1:0}^o$ to initialize lags. In practical, a major challenge in previous studies with VARs is to find the proper priors for the coefficient matrix Φ to deal with the "curse of dimensionality" of VAR (Korobilis, 2013). For example, a $\mathrm{VAR}(4)$ with 5 endogenous variables contains 105 coefficients. Classical approaches always imposed strong *a*-prior restrictions on what variables and which lags should be in the

VAR, and purged "unimportant" variables and lags from the model using a t-test or similar procedures (Canova, 2011). Follow the article of Schorfheide and Song (2015), we use a non-informative prior, i.e., the Minnesota prior, to center the distribution of Φ at a value that implies a random-walk behavior of each of the components of y_t, and to reduce the dimensionality of the problem. More specifically, all coefficients have a zero-prior mean (except the first own lag), and prior distributions become more concentrated for coefficients on longer lags. We treat the prior restrictions on VAR coefficients as dummy observations and add them to the system of VAR equations to combine sample and prior information efficiently. The procedure is as follows.

For the sake of simplicity, we assume there are only two endogenous variables in a $VAR(2)$, i.e., $N = 2$, and $p = 2$. Then a transposed version of Eq. (4-1) can be rewritten as

$$y_t' = [y_{t-1}', y_{t-2}', 1]'\Phi + \mu_t' = \beta_t'\Phi + \mu_t',\ \mu_t \sim \text{iid}N(0, \Sigma)$$

The Minnesota prior can be generated by dummy observations (y_*, ω_*) that are indexed by $\boldsymbol{\lambda}_{5\times1} = \{\lambda_i\}$. Let $\bar{y}$ and σ be $n \times 1$ vectors of means and standard deviations, the computation of pre-sample moments for the variables observed at highest frequency is straightforward. For those variables observed at a lower frequency, we simply equate $\bar{y}_l$ with the pre-sample mean and set the σ_l equal to the pre-sample mean and standard deviation of the observed values, and

$$\begin{bmatrix} \lambda_{l,1}\sigma_{l,1} & 0 \\ 0 & \lambda_{l,1}\sigma_{l,2} \end{bmatrix} = \begin{bmatrix} \lambda_{l,1}\sigma_{l,1} & 0 & 0 & 0 & 0 \\ 0 & \lambda_{l,1}\sigma_{l,2} & 0 & 0 & 0 \end{bmatrix}\Phi + \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix}$$

where λ_1 controls the tightness of the prior. The first row of the above equation can be written as

$$\lambda_{l,1}\sigma_{l,1} = \lambda_{l,1}\sigma_{l,1}\phi_{11} + \mu_{11}$$
$$0 = \lambda_{l,1}\sigma_{l,1}\phi_{21} + \mu_{12}$$

Since μ_t is normally distributed, thus

$$\phi_{11} \sim N\left(1, \Sigma_{11}/\lambda_{l,1}^2\sigma_{l,1}^2\right)$$
$$\phi_{21} \sim N\left(1, \Sigma_{22}/\lambda_{l,1}^2\sigma_{l,1}^2\right)$$

where ϕ_{ij} denotes the element i, j of the matrix Φ, and Σ_{ij} corresponds to element i, j of Σ. The hyperparameter λ_2 which is used to scale the prior standard deviations can be obtained from the following formulation

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & \lambda_{l,1}\sigma_{l,1}2^{\lambda_{l,2}} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{l,1}\sigma_{l,2}2^{\lambda_{l,2}} & 0 \end{bmatrix}\Phi + U$$

The prior for the covariance matrix Σ which is diagonal with elements equal to the pre-sample variance of y_t is obtained by stacking the observations λ_3 times:

$$\begin{bmatrix} \sigma_{l,1} & 0 \\ 0 & \sigma_{l,2} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}\Phi + U$$

When lagged values of a variable $y_{i,t}$ are at the level $\bar{y}_i$, the same value $\bar{y}_i$ is a prior likely to be a good forecast of $y_{i,t}$, regardless of the value of other variables:

$$\begin{bmatrix} \lambda_4\bar{y}_1 & 0 \\ 0 & \lambda_4\bar{y}_2 \end{bmatrix} = \begin{bmatrix} \lambda_4\bar{y}_1 & 0 & \lambda_4\bar{y}_1 & 0 & 0 \\ 0 & \lambda_4\bar{y}_2 & 0 & \lambda_4\bar{y}_2 & 0 \end{bmatrix} \Phi + U$$

When all lagged y_t are at the level $\bar{y}$, a priori y_t trends to persist at that level:

$$[\lambda_5\bar{y}_1 \quad \lambda_5\bar{y}_2] = [\lambda_5\bar{y}_1 \quad \lambda_5\bar{y}_2 \quad \lambda_5\bar{y}_1 \quad \lambda_5\bar{y}_2 \quad \lambda_5]\Phi + U$$

After collecting the T^* dummy observations, the likelihood function associated with Eq. (A-1) can be used to relate the dummy observations to the parameters Φ and Σ. If we combine the likelihood function with the improper prior $p(\Phi, \Sigma) \propto |\Sigma|^{-(n+1)/2}$, we can deduce that the product $p(X^*|\Phi, \Sigma) \cdot |\Sigma|^{-(n+1)/2}$ can be interpreted as

$$(\Phi, \Sigma) \sim MNIW\left(\bar{\Phi}, \left(Y^{o*\prime}Y^{o*}\right)^{-1}, \sigma_l, T^* - k\right)$$

where $\bar{\Phi}$ and σ_l can be obtained as

$$\bar{\Phi} = \left(Y^{o*\prime}Y^{o*}\right)^{-1}Y^{o*\prime}Y^{o*}$$
$$\sigma_l = (Y^* - Y^{o*}\bar{\Phi})'(Y^* - Y^{o*}\bar{\Phi})$$

According to the Bayes rule, the joint distribution $p(Y^o_{1:T}, Z_{0:T}, \Phi, \Sigma|Y_{-p+1:0}, \lambda)$ can be factorized as follows:

$$p(Y^o_{1:T}, Z_{0:T}, \Phi, \Sigma|Y_{-p+1:0}, \lambda) = p(Y^o_{1:T}|Z_{0:T})p(Z_{1:T}|z_0, \Phi, \Sigma)p\left(z_0\middle|Y^o_{-p+1:0}\right)p(\Phi, \Sigma|\lambda) \qquad (4\text{-}4)$$

where the distribution of $(Y^o_{1:T}|Z_{0:T})$ is given by a point mass at the value of $Y^o_{1:T}$ that satisfies Eq.(4-3b), the density $p(Z_{1:T}|z_0, \Phi, \Sigma)$ is obtained from Eq.(4-1), and the conditional density $p\left(z_0\middle|Y^o_{-p+1:0}\right)$ is chosen to be Gaussian. Eventually, $p(\Phi, \Sigma|\lambda)$ represents the prior density of the VAR parameters.

In order to sample from the intractable posterior distribution of latent variables and parameters given the data, $p(\Phi, \Sigma, Z_{0:T}|Y^o_{-p+1:T})$, a Gibbs sampler is applied here which decomposes the posterior into two blocks of full conditional densities which is straightforward to sample from. The conditional posterior densities of VAR parameters and the latent states of the model can be expressed as:

$$p(\Phi, \Sigma|Z_{0:T}, Y^o_{-p+1:T}) \propto p(Z_{1:T}|z_0, \Phi, \Sigma)p(\Phi, \Sigma|\lambda) \qquad (4\text{-}5a)$$

and

$$p\left(Z_{0:T}\middle|\Phi, \Sigma, Y_{-p+1:T}\right) \propto p(Y^o_{1:T}|Z_{0:T})p(Z_{1:T}|z_0, \Phi, \Sigma)p\left(z_0\middle|Y^o_{-p+1}\right) \qquad (4\text{-}5b)$$

where it can be observed that the parameters (Φ, Σ) are independent of Y given Z. Conditional on the parameters, the unobservable variables can be sampled using a simulation smoother. The Gibbs sampler that iterates over the two conditional posterior distributions in Eq. (4-5) was well described in previous studies and is therefore omitted (Carter and Kohn, 1994). The interested reader is referred to Section 2 of the book by Del Negro and Schorfheide (2011) for the detailed procedure of the posterior inference for such a VAR. In addition, as the prior for parameters (Φ, Σ) belongs to the family of matrix- normal-inverse-Wishart (MNIW) distributions and is conjugate for the Gaussian likelihood, thus the conditional posterior is in the same family of distributions by standard results (Karlsson, 2013).

4.3.3. Hyperparameter Selection and Estimation of the Marginal Data Density

The choice of hyperparameters has a significant impact on the empirical performance of the VAR. The hyperparameter vector contains five different elements, including: λ_1, the overall tightness; λ_2, the decay rate of prior variance; λ_3, the dispersion of the prior on the covariance matrix; λ_4, the sum of coefficients on the lags; and λ_5, the persistence restrictions imposed on coefficients. In previous studies, the high-dimensional prior distributions are usually parameterized by a low-dimensional vector of hyperparameter s (Canova et al., 2007). A crude way to choose these auxiliary but important parameters is to use default values (Carriero et al., 2015). However, as the application changes, the hyperparameters may also need to be changed accordingly (Giannone et al., 2015).

In light of the article of Ankargren et al. (2018), an empirical Bayes approach by maximizing the marginal data density (MDD) is utilized to select these hyperparameters. The quantity of interest to estimate is the MDD can be given as

$$p(Y_{1:T}^o|Y_{-p+1:0}^o,\lambda) = \int p(Y_{1:T}^o, Z_{0:T}, \Phi, \Sigma|Y_{-p+1:0}^o,\lambda) \times d(Z_{0:T}, \Phi, \Sigma) \quad (4\text{-}6)$$

Notice that logMDD can be further expressed by decomposing of the one-step-ahead predictive densities $p(y_t^o, |Y_{-p+1:t-1}^o,\lambda)$, that is

$$\ln p(Y_{1:T}^o|Y_{-p+1:0}^o,\lambda) = \sum_{t=1}^{T} \ln \int p(y_t^o, |Y_{-p+1:t-1}^o, \Phi, \Sigma) \times p\big(\Phi, \Sigma\big|Y_{-p+1:t-1}^o,\lambda\big) d(\Phi, \Sigma) \quad (4\text{-}7)$$

We consider a grid search for λ and assign an equal prior probability to each value on the grid. Considering the initialization of the VAR, according to the Bayes rule, we have

$$p\big(Y_{1:T,l}, y_{0,l}, Y_{1:T}^o\big|Y_{-p+1:0}^o,\lambda\big) = p\big(Y_{1:T,l}, y_{0,l}\big|Y_{1:T}^o, Y_{-p+1:0}^o,\lambda\big)p\big(Y_{1:T}^o|Y_{-p+1:0}^o,\lambda\big) \quad (4\text{-}8)$$

where $Y_{1:T,l}$ stacks the missing values of the low-frequency variables $y_{t,l}$, and $y_{0,l}$ is the values for each initialization period $t = -p+1, \dots, 0$. By this means, the approximation of the MDD can be written as:

$$\tilde{p}\big(Y_{1:T}^o|Y_{-p+1:0}^o,\lambda\big) = a\left[\frac{1}{R}\sum_{i=1}^{R}\frac{g_0\big(y_{0,l}^{(i)}\big)g\big(Y_{1:T,l}^{(i)}\big)}{p\big(Z_{1:T}^{(i)}\big|z_0^{(i)},\lambda\big)p\big(z_0^{(i)}\big|Y_{-p+1:0}^o,\lambda\big)}\right]^{-1} = a\left[\frac{1}{R}\sum_{i=1}^{R}\frac{g\big(Y_{1:T,l}^{(i)}\big)}{p\big(Z_{1:T}^{(i)}\big|z_0^{(i)},\lambda\big)}\right]^{-1} \quad (4\text{-}9)$$

where the constant a is the Jacobian term associated with the change-of-variables from $\big(Y_{1:T,l}, y_{0,l}, Y_{1:T}^o\big)$ to $(z_0, Z_{1:T})$. Let $g_0\left(y_{0,l}^{(i)}\right) \equiv p\left(z_0^{(i)}\middle|Y_{-p+1:0}^o,\lambda\right)$ such that the two terms cancel out. The $g(Y_{1:T,l}^{(i)})$ function stands for the trimmed multivariate Gaussian distribution with mean $\bar{\mu}_{Y_{1:T,l}} = \frac{1}{R}\sum_{i=1}^{R} Y_{1:T,l}^{(i)}$ and variance $\sigma_{Y_{1:T,l}} = \frac{1}{R}\sum_{i=1}^{R} Y_{1:T,l}^{(i)} Y_{1:T,l}^{(i)'} - \bar{\mu}_{Y_{1:T,l}}\bar{\mu}'_{Y_{1:T,l}}$, and $\int Y_{1:T,l}^{(i)}\, dY_{1:T,l}^{(i)} = 1$. The draws from the distribution of $Y_{1:T,l}|(Y_{1:T}^o,\lambda)$ can be attained by converting the draws from $Y_{1:T,l}|(Y_{1:T}^o,\lambda)$ which are generated as a by-product of the posterior sampler.

4.3.4. Model Comparison and Estimation

The first step of a VAR is to figure out whether each variable has correlations with others. The Granger-causality test is the most frequently employed technique to cope with this issue. The idea of the Granger-causality test is that if a variable affects another variable, the former should help to improve the performance of the latter variable. Another critical issue in VAR analysis is how to find the optimal number of lags that yields the best results. Commonly, model comparisons are based on information

criteria such as AIC, HQ, or SC (Juselius, 2006). Due to its favorable small sample forecasting features, AIC is the most widely used criteria. In order to provide a comprehensive model comparison, other criteria are also adopted in this study. Overall, these information criteria can be computed as:

$$AIC(p) = \ln\det\left(\tilde{\Sigma}_\varepsilon(p)\right) + \frac{2}{T}pN^2 \quad (4\text{-}10a)$$

$$HQ(p) = \ln\det\left(\tilde{\Sigma}_\varepsilon(p)\right) + \frac{2\ln(\ln(T))}{T}pN^2 \quad (4\text{-}10b)$$

$$SC(p) = \ln\det\left(\tilde{\Sigma}_\varepsilon(p)\right) + \frac{\ln(\ln(T))}{T}pN^2 \quad (4\text{-}10c)$$

with $\tilde{\Sigma}_\varepsilon(p) = T^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t'$.

Once a final VAR model is determined, the estimated parameter values have to be interpreted. Since all variables in the VAR model are interdependent, each parameter value only provides limited information about the system. To acquire a better intuition of the model's dynamic behavior, impulse responses (IR) are adopted in this study (Lütkepohl, 2005). The idea of impulse responses is to calculate the effect of a unit change of the impulse variable (e.g., unemployment rate) on the response variable (e.g., total collisions). The impulse response of the jth variable to the ith variable at time t given at time $t-s$ with a one-unit change of standard deviation can be expressed as

$$\psi_{ij}^{s} = \frac{\partial y_{i,t}}{\partial \sigma_{j,t-s}}, i,j = 1,\dots,N \quad (4\text{-}11)$$

The reason for using the impulse of standard deviation instead of one-unit change is because the variables are not in the same scales. From the perspective of traffic safety study, we are more curious about the long-run effects of the impulse of a variable on another variable. Therefore, the following accumulated impulse response function can be used:

$$\Psi_{ij} = \sum_{\forall s}\psi_{ij}^{s} \quad (4\text{-}12)$$

Furthermore, if the accumulated impulse response Ψ_{ij} is positive, it indicates that the jth variable has a positive impact on the ith variable.

4.4. Estimation Results and Discussion

The model estimation is conducted in the R language using LaplacesDemon (Statisticat, 2015) and mfbvar (Ankargren et al., 2018) packages on a computer with an Intel Core i7-6700 CPU at 3.40 GHz processor and 16.0 GB RAM. For each estimation sample, we generate 20,000 draws from the posterior distribution of the VAR parameters using the Markov Chain Monte Carlo (MCMC) algorithm. The first 10,000 draws are discarded, and the remaining 10,000 draws are used to calculate Monte Carlo approximations of posteriors.

4.4.1. Model Comparison Results

As shown in Table 4-3, the Granger-causalities of all five injury severities is significant, indicating other variables listed in Table 4-1 have substantial impacts on them. Therefore, all variables are retained in the VAR. AIC, HQ, and SC are adopted for selecting optimal lag, p. As illustrated in Figure 4.3, the three criteria reach their minimal values when $p = 8$, indicating using eight lags can provide the best model

performance. Therefore, $VAR(8)$ with all endogenous variables in Table 4-1 are selected as the final model.

Table 4-3 Granger-causality test results.

Variable	**Chi-squared**	**p-value**
Total Collisions	67.99	<0.01
Fatal Collisions	58.85	<0.01
Serious Injury Collisions	65.42	<0.01
Minor Injury Collisions	65.59	<0.01
PDO Collisions	64.79	<0.01

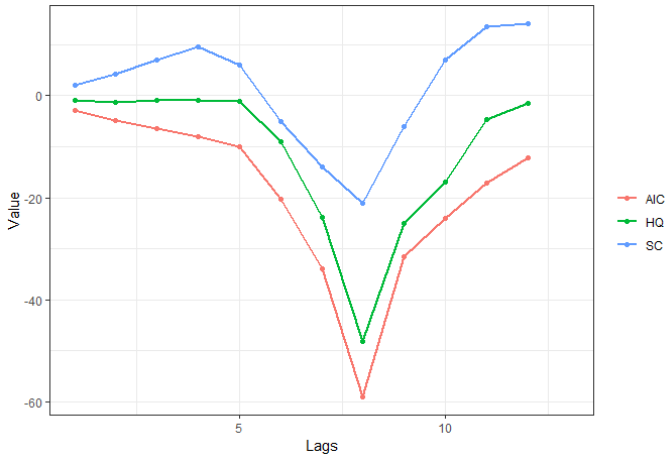


Figure 4.3 Criteria comparison in different lags.

In general, the larger hyperparameter the smaller the prior variance and the more informative the prior. For the relative less-important dispersion hyperparameter, λ_3, we follow the experience of previous studies and set it to be 1. $\hat{\lambda}_4$ and $\hat{\lambda}_5$ are selected based on the grid search on $\Lambda^{(1)}$ for the first recursive sample, and $\hat{\lambda}_4 = 2.2$, $\hat{\lambda}_5 = 4.0$, respectively. Conditioning on $\hat{\lambda}_3$, $\hat{\lambda}_4$, and $\hat{\lambda}_5$, λ_1 and λ_2 are obtained based on the grid search on $\Lambda^{(2)}$. The logMDD surface is illustrated in Figure 4.4 as a function of λ_1 and λ_2, holding the remaining three hyperparameters fixed as $\lambda_3 = 1$, $\lambda_4 = \hat{\lambda}_4$, and $\lambda_5 = \hat{\lambda}_5$. The surface is in a convex shape and reaches its maximum, 12246.2, at $\hat{\lambda}_1 = 0.07$ and $\hat{\lambda}_2 = 4.0$, and this hyperparameter vector is selected as the final model input.

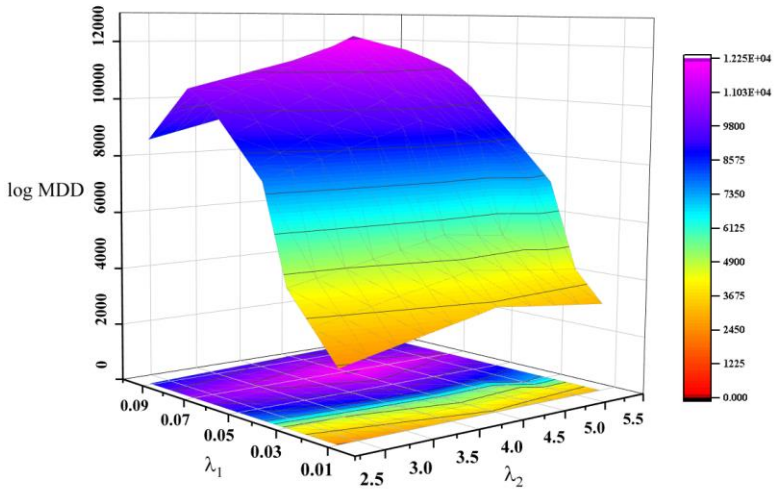


Figure 4.4 Log Marginal data density.

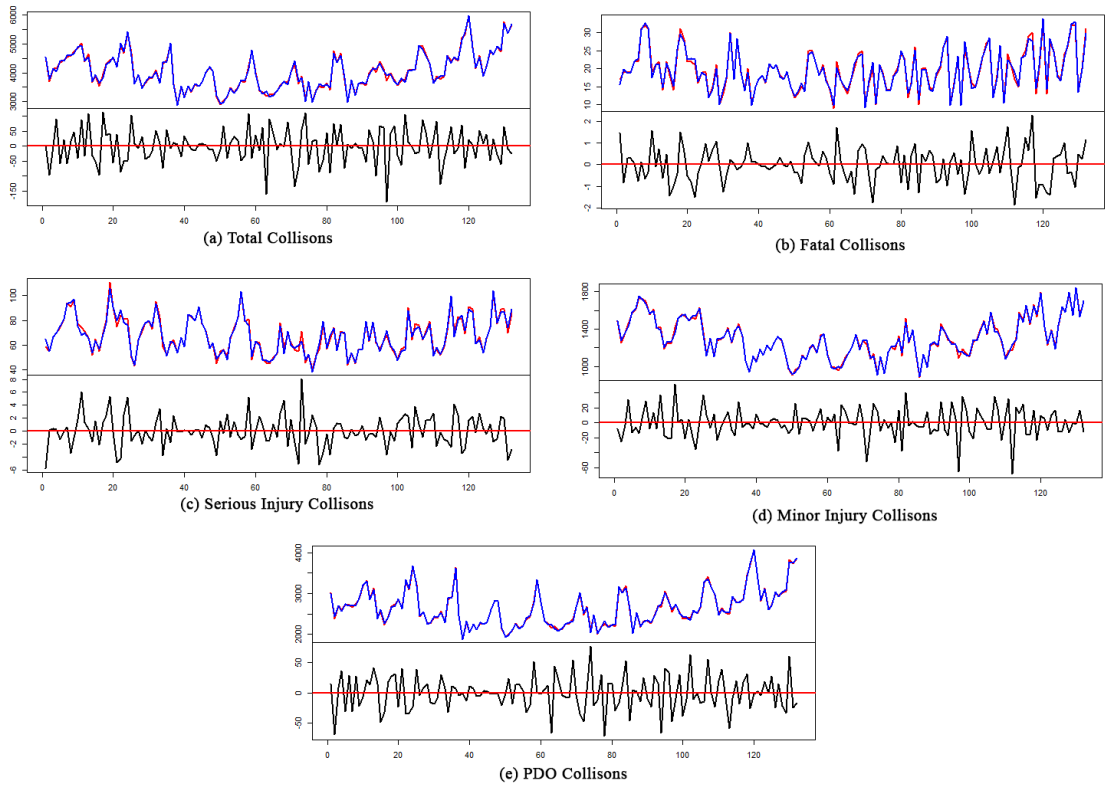


Figure 4.5 Diagram of fit and residuals for collisions in different severities (the red curve represents the real observations, the blue curve represents fitted values, and black curve represents residuals).

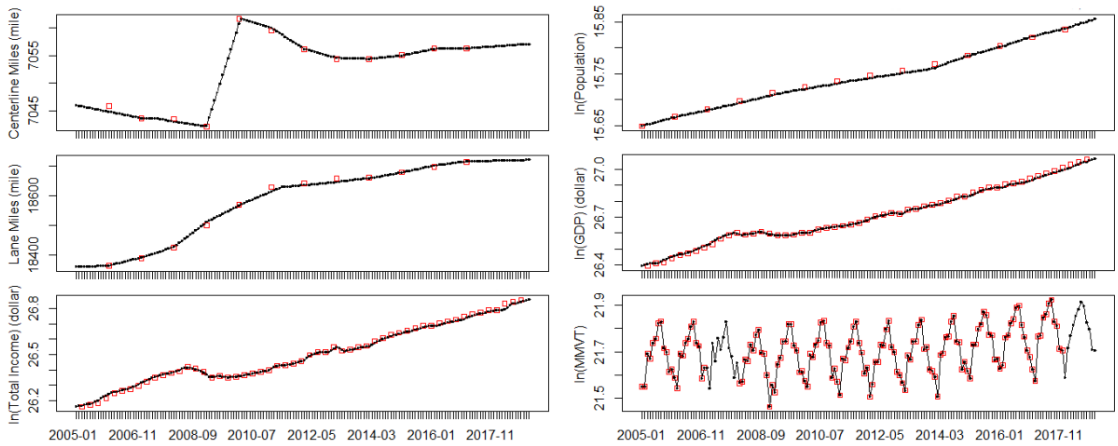


Figure 4.6 Imputation results of low-frequency observed data. (the black dots are the filling results in a monthly rate, and the red squares are the observed data in a low-frequency rate)

Table 4-4 Imputation results of different models

Variables	Linear	PMM	k-NN	RF	BVAR
Centerline Miles	10.24*/0.65**	2.21/0.82	1.23/0.91	0.88/0.97	0.91/0.96
Lane Miles	8.41/0.78	2.56/0.84	1.62/0.90	1.28/0.92	1.12/0.95
Total Income	0.24/0.86	0.12/0.90	0.09/0.92	0.04/0.96	0.04/0.96
Population	0.09/0.88	0.06/0.92	0.04/0.96	0.04/0.96	0.04/0.96
GDP	0.07/0.87	0.04/0.92	0.02/0.97	0.02/0.97	0.02/0.97
MMVT	0.25/0.54	0.16/0.72	0.08/0.88	0.06/0.90	0.02/0.95

[a] Linear regression
[b] Predictive mean matching
[c] k-Nearest neighbors
[d] Random forests
[e] Bayesian mixed frequency VAR
* Normalized root mean square error (NRMSE)
** Adjusted R-squared

Figure 4.5 presents the model fit and residuals of the proposed $\mathrm{VAR}(8)$. The VAR shows a relatively superior fit accuracy, as residuals of collisions in different severities are all small and around zero. The average fitting errors of the total collisions, fatal collisions, serious injury collisions, minor injury collisions, and POD collisions, are 1.12%, 3.08%, 2.65%, 1.12%, and 0.84%, respectively, also indicating the proposed VAR has a favorable performance.

Data filling results of low-frequency data and data with missing values are illustrated in Figure 4.6. It can be seen that most of the observations are on the curve of the proposed model, implying that the proposed method has high accuracy. The following methods including linear regression (Jones, 1996), predictive mean matching (Landerman et al., 1997), k-nearest neighbors (García-Laencina et al., 2009), and random forests (Hapfelmeier et al., 2014), are selected as benchmarks to evaluate the fitting results of the proposed Bayesian mixed-frequency VAR. It can be seen that the linear regression has the worst performance since it has the largest normalized root mean square error and least adjusted R-squared. Random forests, k-nearest neighbors, and the proposed VAR have comparative performance in terms of the variable goodness-of-fit of total income, population, and GDP. As shown in the lower right corner of Figure 4.6, VAR not only simply estimates the missing values, but takes the time trend of the data into

consideration. Therefore, the proposed Bayesian VAR has the overall finest performance and is capable of capturing temporal instability.

4.4.2. Model Estimation Results

The cumulative impulse response functions of the five different types of collisions are illustrated in Figure 4.7 to Figure 4.11. A variable is considered to raise the collision likelihood if the majority of its cumulative impulse response is larger than zero. Conversely, if most of the cumulative impulse response is less than zero, the variable is considered to moderate the probability of a collision. Therefore, as presented in Figure 4.7 to Figure 4.11, the variables, centerline miles, and line miles, are found to increase the possibility of collisions in all severities according to their significant positive cumulative impulse responses. This finding indicates that the increase of highway lengths could lead to more collisions in different severities. Besides the results show that MMVT only has significant impacts on the occurrence of serious injury collisions while it has no influence on other collision types.

Estimation results of the variable for temperature, demonstrate that the increase in temperature could lead to the decrease in the frequencies of total collisions, fatal collisions, serious injury collisions, and PDO collisions, respectively. These results are in line with previous studies. For instance, Usman et al., (2012) reported that a 1% increase from the mean values of temperature could lead to a 0.6% decrease in collisions. Brijs et al. (2008) also claimed that temperature has a significant and non-linear relationship with collision occurrence, and lower temperatures may lead to a greater number of collisions.

In terms of precipitation, this variable was found to be associated with an increased likelihood of collisions in all severities except for minor injury collisions. Previous studies and experiences also evidence similar findings. For example, Hermans et al. (2006) found that precipitation during the observation period could significantly increase the crash frequency. Andrey et al. (2001) figured out that precipitation is associated with more than a 70% collision increase. Usman et al. (2012) discovered that a 1% increase from the mean values of precipitation would cause the mean number of collisions to rise by 0.02%. The underlying mechanism of these two weather-related variables on the collision likelihood is somewhat complicated. First, as illustrated in Figure 4.2, months with lower average temperatures, i.e., November, December, January, and February, have higher precipitation. Heavy rainfall and even snow at low temperatures may unsurprisingly lead to slippery, icy, or even black ice road surface, which in turn reduces the pavement friction and skid resistance. Considering the relatively high travel speed on state highways, it is challenging for drivers to fully control their vehicles, e.g., steering, stopping, etc., under such adverse conditions. These inclement weather conditions may be associated with visibility reduction and lane obstruction, which also pose significant challenges to drivers. Furthermore, since these periods are among the holiday and touristic season, the traffic volume increases significantly, resulting in more occurrence of collisions (Bellos et al., 2019; Vahdati et al., 2016). The higher proportion of drug/alcohol-impaired driving drivers and fatigued drivers who traveled long distances or rushed to get somewhere during these periods may also contribute to an increased likelihood of collisions (Liu et al., 2005).

An increase of total income can lead to a decrease in occurrences of collisions in all severities. Previous studies also demonstrated similar conclusions (Li et al., 2019a; Yasmin and Eluru, 2018). The results may be that as personal wealth increases, people are more likely to purchase more advanced, sounder quality, safer vehicles. These vehicles may have more crash-prevention features, for instance, anti-lock braking systems (ABS), electronic stability control (ESC), tire pressure monitoring systems (TPMS),

adaptive cruise control, adaptive headlights, lane-departure warning, forward collision warning with auto-braking, etc., which can significantly decrease the possibilities of collisions (C. Chen et al., 2016b). In addition, some of these features, together with all-side-airbags, safety belts alerts, etc., can also contribute to the decrease of frequency of severe collisions (C. Chen et al., 2016a; Li et al., 2018a; Wu et al., 2014).

Similar to the total income, the variable GDP, is also found to have significantly favorable impacts on collision occurrence in all severities, while having no significant effect on fatal collisions. Abundant studies have revealed comparable conclusions (Yannis et al., 2014). For example, Page (2001) studied crash occurrence in OECD (Organization for Economic Co-operation and Development) countries and discovered that collisions per registered vehicle tend to decrease over time as GDP increases. The reason may be that the rise of macroeconomics leads to additional expenditures and investments in road infrastructure and maintenance, for instance, a more sophisticated weather warning system, faster road clutter cleaning speed, better speed monitoring equipment, etc. These advances and improvements may further promote road safety and reduce collisions.

The frequencies of total collisions and fatal collisions are also found to be influenced by the unemployment rate. Results show that an increase in the unemployment rate could increase the occurrence of the two collision types. In previous studies, the unemployment rate was considered to have mixed effects on traffic collision frequencies (Leigh and Waldon, 1991; Liu and Sharma, 2018). Some previous studies suggested a high unemployment rate may increase anxiety, psychological stress, and depression in the population and cause higher alcohol/drug consumption and more lethal DUI crashes (Freeman, 2007; Li et al., 2019a; Males, 2009; Traynor, 2009). Besides, mental stress in the population associated with both job loss and the anxiety of job loss may lead to more aggressive driving patterns and more collisions (Liu and Sharma, 2018). On the other hand, some studies claimed that the unemployment rate might bring about lower driving frequency and fewer traffic collision s (Brazil and Kirk, 2016; Scuffham and Langley, 2002). These complex effects may be the cause of our estimation results.

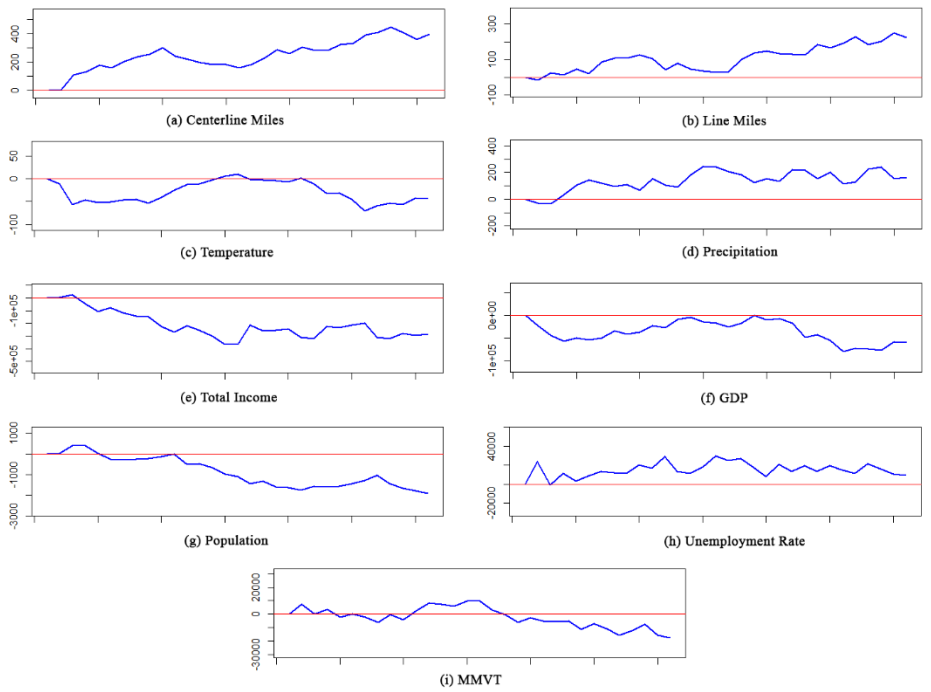


Figure 4.7 Cumulative impulse response for total collisions.

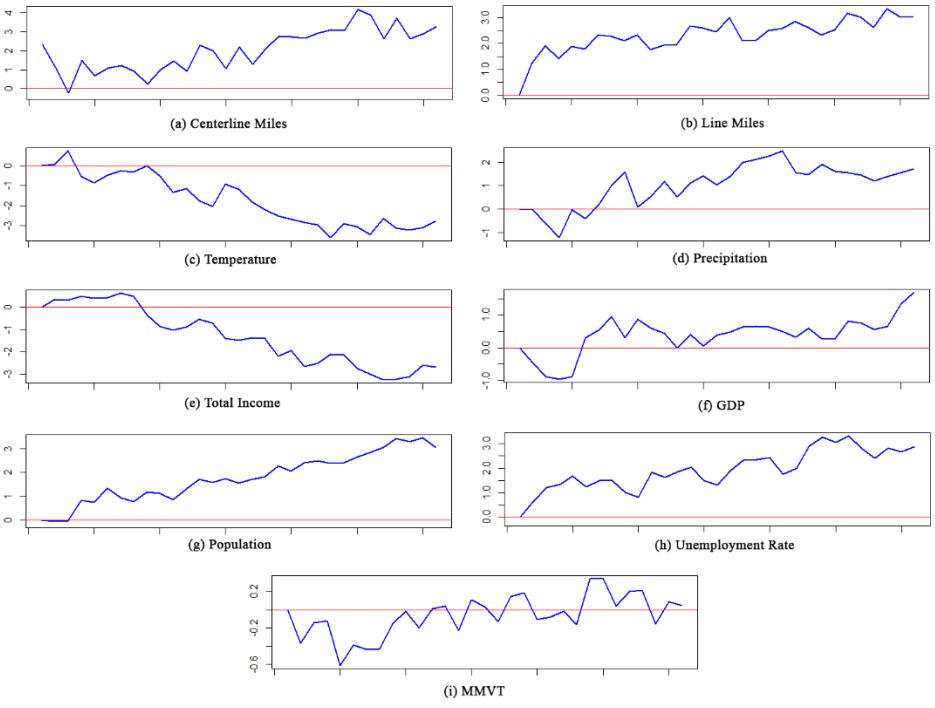


Figure 4.8 Cumulative impulse response for fatal collisions.

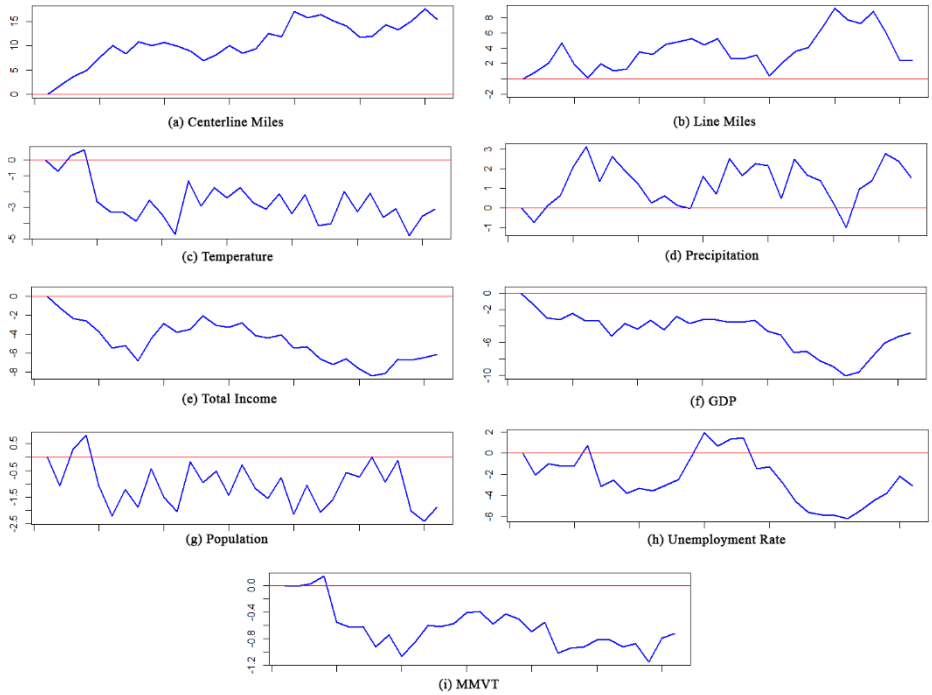


Figure 4.9 Cumulative impulse response for serious injury collisions.

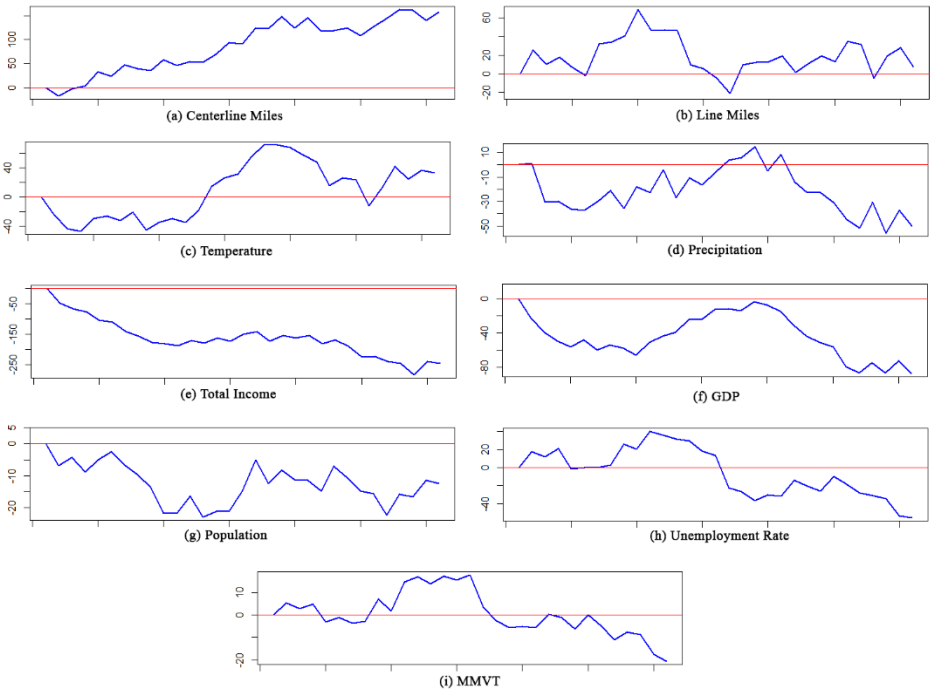


Figure 4.10 Cumulative impulse response for minor injury collisions.

The tendencies of collisions in different severities from January 2017 to December 2018 are illustrated in Figure 4.12. The results show that total collisions, serious injury collisions, and minor injury collisions have descending trends in the predicted time periods, while the average monthly fatal collisions and

PDO collisions have increasing tendencies in the forecast periods. November and December are the two months with highest total collisions, minor injury collisions, and PDO collisions, while fatal collisions and serious injury collisions are more likely to occur in July and August.

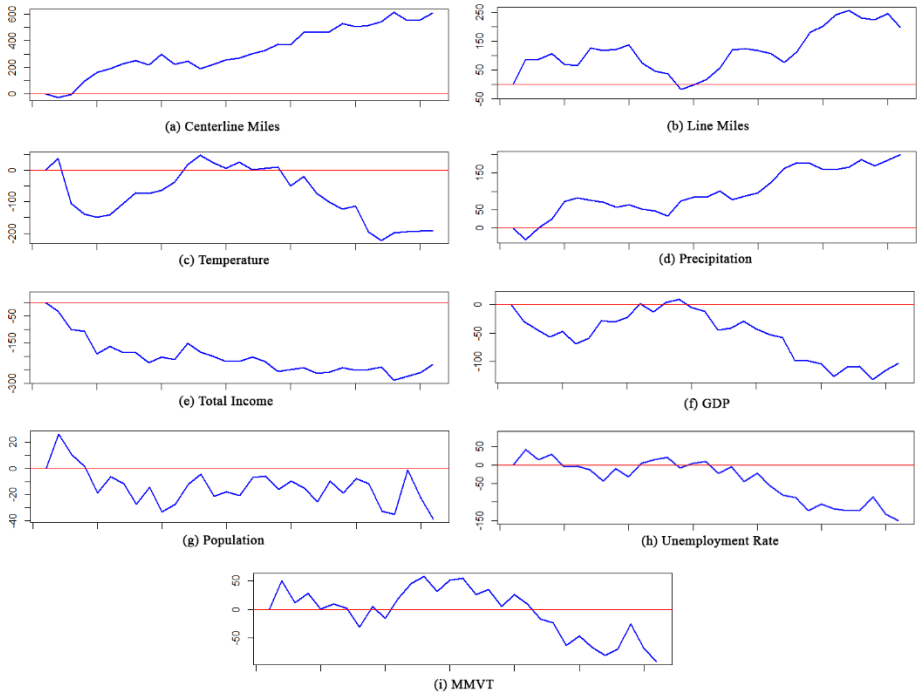


Figure 4.11 Cumulative impulse response for PDO collisions.

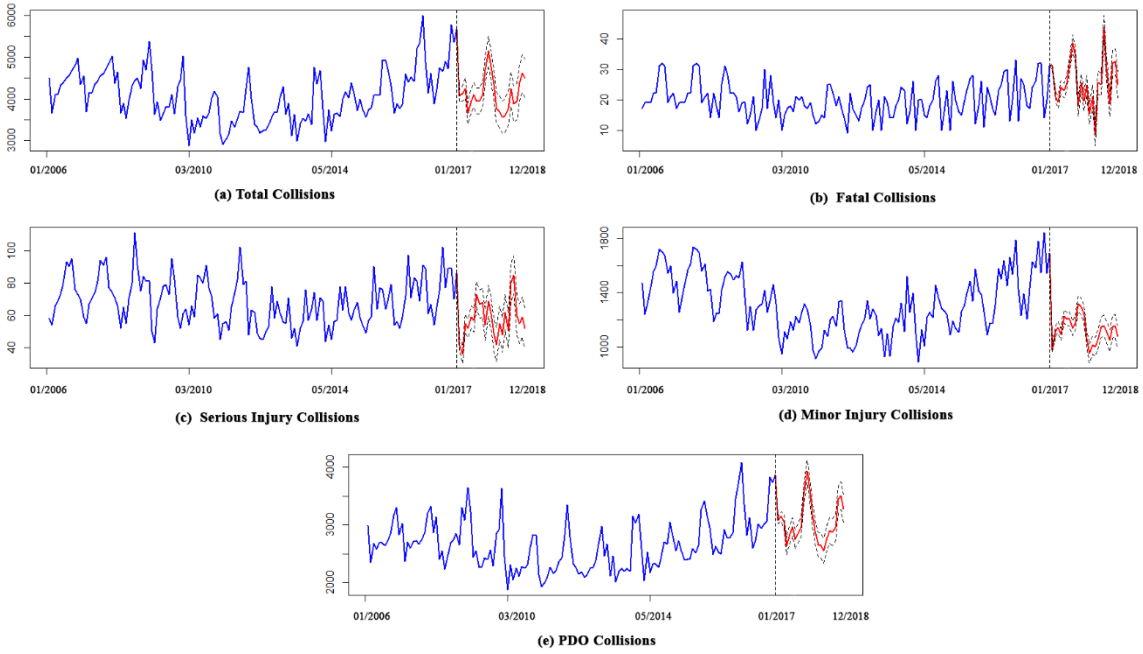


Figure 4.12 Prediction of collisions in different severities.

4.5. Summary

A mixed-frequency Bayesian vector autoregressive model is proposed to analyze the impacts of different transportation-, weather- and socioeconomic-related characteristics on traffic collisions. The selected dataset is unevenly-spaced traffic collision data with missing values, containing all collisions in different severities that occurred on the state highways in Washington State from January 2006 to December 2016. A Gibbs sampler is used to conduct Bayesian inference for model parameters and unobserved high-frequency variables. By assuming the error terms after a multivariate normal distribution, the model can capture unobserved heterogeneity. To cope with the dimensionality of the Bayesian VAR, a Minnesota prior is adopted to shrink the VAR coefficients toward univariate random-walk representations. The degree of shrinkage is decided in a data-driven technique, by maximizing the marginal data density concerning a low-dimensional vector of hyperparameters. Granger-causality tests demonstrate that all variables can Granger-cause collisions of different severities, and therefore they are retained in the model. The optimal lag of the VAR is selected with information criteria, and results show that eight lags can provide the best model performance. Therefore, $VAR(8)$ with all endogenous variables are designated as the final model. Cumulative impulse response results reveal that the increase of centerline miles can increase the possibilities of collisions in all severities. Also, the increases in precipitation and unemployment rate are found to raise the likelihood of collisions in some severities. On the contrary, the increases in total income, temperature, population, and GDP can moderate the probabilities of collisions. The forecast results of collisions in different severities from January 2017 to December 2018 demonstrate different collisions have various tendencies during the forecast periods, which provides beneficial references for proposing corresponding countermeasures to mitigate the likelihoods of collisions.

A potential disadvantage of the VAR approach is that, especially as the number of lags grows, the number of parameters to be estimated grows considerably. Besides, although the model can mitigate the temporal instability brought by aggregating data into a lower frequency, it cannot eliminate the influences of the unstable temporal issue. The time-variant Bayesian VAR is recommended to explore the temporal-related unobserved heterogeneity further.

CHAPTER 5. A FINITE MIXTURE RANDOM PARAMETERS MODEL

In this chapter, we present a finite mixture random parameters model to explore driver injury severity causes in low visibility related to single-vehicle crashes. The existing literature has provided insightful guidance regarding the impacts of reduced visibility on traffic safety performance and contributed to the analysis of crash injury severities. However, it is desirable to focus on driver injury severity formulation in low visibility related crashes through advanced discrete choice models. Therefore, we propose a finite mixture random parameters approach to analyze the risk factors and their impacts on driver injury severity outcomes in low visibility related crashes. A three-year crash dataset from 2010 to 2012 focusing on low visibility related crashes in four South Central states, including Arkansas, Louisiana, Texas, and Oklahoma, was utilized in this study. The rest of the chapter is organized as follows: Section 5.1 demonstrates a brief introduction and literature review for studies focusing on low visibility related single-vehicle crashes; Section 5.2 provides the explicit description of the dataset. The details of model development are illustrated in Section 5.3. In Section 5.4, the model analysis results are comprehensively presented and discussed regarding the implication of the proposed model and the impacts of different risk factors. Finally, the entire research effort is concluded in Section 5.5.

5.1. General Background

Driving under inclement weather conditions is more challenging, as the adverse weather may degrade significant safety performance present under normal driving conditions. Previous studies have demonstrated that severe injuries are more likely to occur under such weather conditions (Chiou et al., 2014; Shaheed et al., 2016; Behnood and Mannering, 2017). Among all the inclement weather conditions, low visibility, mainly associated with fog, dust, or smoke, is one of the most hazardous factors due to its considerable adverse impacts. According to National Highway Traffic Safety Administration, over 9% of weather-related crash fatalities during 2005-2014 occurred due to low visibility. However, low visibility-related crashes accounted for around 3% in all weather-related crashes (NHTSA, 2016). In addition, low visibility also plays a significant role in pedestrian- and cyclist-involved collisions in mixed traffic flows since it becomes much challenging for the nonmotorized traffic to be seen during such conditions, and may result in serious injuries. Thus, special attention is needed to investigate the underlying mechanisms of low visibility related crashes that contribute to such severe injury outcomes. It is noted that, although rain or snow may also cause low visibility, crashes that are influenced by these conditions mainly result from low skid resistance, and therefore are excluded in this study.

5.1.1. Related Work

When analyzing driver injury severity, most previous studies considered low visibility as an item of the weather variables and combined it with other conditions, e.g., clear, or rain, (Eluru et al., 2012; Zou et al., 2014; Haleem and Gan, 2015). Unlike other weather-related crashes (e.g., rain- or snow-related crashes), crashes that occurred under low visibility conditions were not thoroughly investigated in the existing literature. It might be due to the lack of clear documentation of such weather conditions in crash datasets. Based on our thorough literature review, some existing studies provide theoretical and empirical contributions to the body of knowledge in low visibility related crash modeling and analysis (Uc et al., 2009; Abdel-Aty et al., 2012; Shaheed and Gkritza, 2014; Norros et al., 2016). For instance, a study conducted by the Virginia Department of Transportation (VDOT) reported that almost all the

primary fog crashes occurred in fog-prone areas. Most of them involved secondary crashes leading to severe injuries and property damage (Lynn et al., 2002). Given that fog fades the colors and reduces the contrasts in the scene with respect to their distances from the driver, Tarel et al. (2012) proposed an innovative approach to facilitate camera-based Advanced Driver Assistance Systems (ADAS) on the processing of fog images and to enhance safety performance. A multilevel ordered logistic model was utilized to examine the effects of various risk factors using a low visibility crash dataset from Florida between 2003 and 2007. Results showed these crashes were more prevalent on high-speed roads, undivided roads, roads with no sidewalks and two-lane rural roads, and tended to involve more vehicles and more severe injuries (Abdel-Aty et al., 2011). Other similar studies using different approaches also provided meaningful insights for analyzing low visibility related crashes (McCann and Fontaine, 2016; Wu et al., 2018).

5.1.2. Limitations in previous studies

Based on our best understanding, only a few analytic methods have been proposed to investigate the contributing factors and their impacts on driver injury severity in low visibility related crashes. Therefore, new methodological approaches should be developed and tested in terms of appropriateness in analyzing low visibility associated crashes. The current studies have indicated that heterogeneity modeling is a promising means for traffic safety researchers to provide more accurate estimation when analyzing crash data extracted from police reports (Milton et al., 2008; Russo et al., 2014). Although the collected crash data are sufficient to provide all detailed attributes with multiple variables and descriptions, some unobserved factors cannot be fully addressed. For instance, occupants in the same age range (i.e., young, middle-aged, or old), may demonstrate significantly different attributes from each other, including perception/reaction time, physical conditions, etc., which may make the impacts of the age variable on injury severities different from one observation to the other. Interested readers can find detailed examples of explanatory variable analyses with possible heterogeneous effects conducted by Mannering et al. (2016). If the unobserved heterogeneity in the dataset is not fully addressed, the impacts of the observed variables on injury severities are then constrained to be constant across all the observations, which may result in biased estimation and erroneous predictions. The injury severities of affected occupants are often modeled as discrete severity outcomes (for instance, fatal injury, disabling injury, visible injury, a complaint of injury or possible injury, and no apparent injury), once the crash is observed. Therefore, discrete choice models accounting for unobserved heterogeneity are required for analyzing the commonly collected crash datasets. A mixed logit model, a type of random parameter model, has been widely adopted by all the various approaches that can meet the aforementioned requirements (Chen and Tarko, 2014; Russo et al., 2014; Ye and Lord, 2014; Coruh et al., 2015). For instance, Kim et al. (2010) applied a mixed logit model to analyze pedestrian injury severities in pedestrian-vehicle crashes. They discovered that the effect of pedestrian age was normally distributed across observations, and the probability of fatal injuries increased substantially with the increase of pedestrian ages. Ye and Lord (2014) verified that random parameters models outperformed traditional discrete choice models in crash severity modeling by allowing the same parameter to vary across observations based on the predefined distributions but might result in a complex multimodal distribution with unexpected shapes and skewness.

A finite mixture approach (also known as a latent class model) is another simplified approach to address the unobserved heterogeneity. It is designed to seek observations of similar characteristics and gather them into different groups. Shaheed and Gkritza (2014) utilized this approach to investigate the factors

that affect crash severity outcomes in single-vehicle motorcycle crashes based on the crash data in Iowa from 2001 to 2008, and the unobserved heterogeneity issue was addressed by two distinct crash data classes identified by the model. Considerable heterogeneity was also verified across the subtypes in a study conducted by Behnood et al. (2014) with a two-class finite mixture model that explored the differences in driver-injury severity between sober and alcohol-impaired drivers. Other studies that applied this approach provided an in-depth understanding of its applicability and effectiveness (Afghari et al., 2016; Behnood and Mannering, 2016; Yu et al., 2017). A limitation of the finite mixture approach is that it is difficult to determine the optimal number of subtypes, and the unobserved heterogeneity, although reduced, might still exist within each identified latent class. Previous applications also suggested that after specifying more than four subtypes, it becomes challenging to achieve model convergence and obtain accurate parameter estimation (Greene, 2012).

To overcome both the limitations of random parameter models and finite mixture models, a hybrid approach combining these two models was proposed in some previous research efforts (Xiong and Mannering, 2013; Buddhavarapu et al., 2016). This hybrid approach predefined the number of latent classes, allows the random parameters to vary across latent classes and observations within each identified potential class, and, therefore, can model more sophisticated unobserved heterogeneity than traditional discrete choice models. However, this hybrid approach was recently introduced to the traffic safety analysis domain and has not been used to analyze crashes under low visibility weather conditions. In addition, there are still issues remaining to be addressed regarding the model structure and parameter assumptions (i.e., how to decide a more reasonable number of latent class and proper distributions for random parameters).

5.2. Data

The crash dataset in this study was obtained from the state Departments of Transportation (DOTs) of Texas, Arkansas, Oklahoma, and Louisiana. The geographical locations of these states are illustrated in Figure 5.1. All the low visibility related single-vehicle crash data from 2010 to 2012 in these states were utilized in this research. The joint investigation of the four states based on their similar geographic features and demographic characteristics is appropriate and verified by many studies (Adams et al., 2016).

Given the different policies and standards of the crash reporting systems in the four states, only the common variables in the datasets of the four states are selected in this study. The integrated dataset contains critical information regarding low visibility related crashes and the associated vehicles and drivers. Driver injury severity, as the dependent variable, was initially classified into five categories: fatal injury, incapacitating injury, visible injury, the complaint of damage or possible harm, and no apparent injury. In this study, to maintain a statistically meaningful sample size and simplify the analysis procedure, three injury severity levels were defined, where no injury (N, no apparent injury in the original category system) is selected as the referenced severity, injury (I, visible injury and complaint of injury or possible injury) and serious injury and fatal (F, incapacitating injury and fatal injury). After carefully screening all the incomplete and erroneous records, 3,049 low visibility related single-vehicle crashes were analyzed. The variables of roadway geometries, vehicle information, driver demographics, and driver injury severities, are presented and summarized. The detailed information of the dataset is illustrated in Table 5-1.

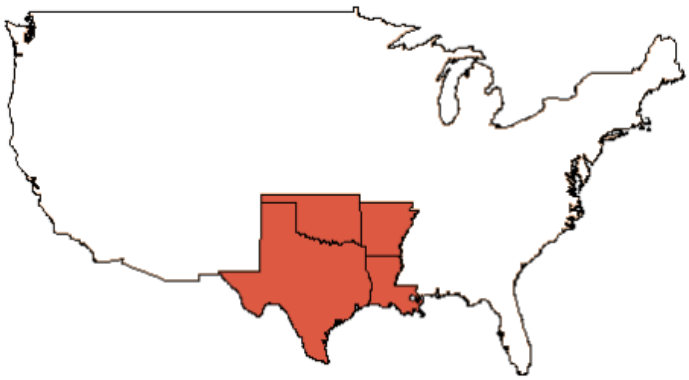


Figure 5.1 Location of study area

Table 5-1 Definitions and descriptions of variables

Variable	**Driver injury severity**						**Total**
	No injury (N)		*Injury* (I)		*Serious injury and fatal* (F)		
Severity	1820	59.69%	983	32.24%	246	8.07%	3049
Day of Week							
Sunday	277	56.30%	155	31.50%	60	12.20%	492
Monday	263	57.30%	166	36.17%	30	6.54%	459
Tuesday	277	66.59%	119	28.61%	20	4.81%	416
Wednesday	210	55.12%	147	38.58%	24	6.30%	381
Thursday	247	63.82%	106	27.39%	34	8.79%	387
Friday	259	58.86%	141	32.05%	40	9.09%	440
Saturday	287	60.55%	149	31.43%	38	8.02%	474
Light Condition							
Dark	921	57.31%	536	33.35%	150	9.33%	1607
Dawn	154	66.96%	66	28.70%	10	4.35%	230
Daylight	703	61.13%	369	32.09%	78	6.78%	1150
Dark with Light	42	67.74%	12	19.35%	8	12.90%	62
Area							
Rural	1204	58.11%	682	32.92%	186	8.98%	2072
Urban	616	63.05%	301	30.81%	60	6.14%	977
Road Character							
Straight	1191	60.00%	657	33.10%	137	6.90%	1985

Variable	Driver injury severity						Total
	No injury (N)		*Injury* (I)		*Serious injury and fatal* (F)		
Curve	629	59.12%	326	30.64%	109	10.24%	1064
Road Grade							
Level	1251	59.91%	678	32.47%	159	7.61%	2088
Hillcrest	59	57.84%	33	32.35%	10	9.80%	102
On grade	477	61.63%	231	29.84%	66	8.53%	774
Dip	33	38.82%	41	48.24%	11	12.94%	85
Road Surface Condition							
Dry	538	60.86%	280	31.67%	66	7.47%	884
Wet	1088	57.66%	635	33.65%	164	8.69%	1887
Ice	176	70.97%	56	22.58%	16	6.45%	248
Loose Material	18	60.00%	12	40.00%	0	0.00%	30
Road Pavement							
Paved Road	1801	59.66%	975	32.30%	243	8.05%	3019
Road not Paved	19	63.33%	8	26.67%	3	10.00%	30
Traffic Controls							
No Control	448	68.19%	189	28.77%	20	3.04%	657
Stop-Yield Sign	131	57.71%	84	37.00%	12	5.29%	227
Signal Control	1010	57.48%	557	31.70%	190	10.81%	1757
Other Control Methods	231	56.62%	153	37.50%	24	5.88%	408
Number of Lanes							
One Lane	56	71.79%	20	25.64%	2	2.56%	78
Two Lanes	1417	59.09%	796	33.19%	185	7.71%	2398
Multiple Lanes	347	60.56%	167	29.14%	59	10.30%	573
Speed Limit							
30 mph or Less	261	43.53%	272	45.30%	67	11.17%	600
35 or 40 mph	373	58.63%	224	35.13%	40	6.24%	636
45 or 50 mph	366	65.40%	158	28.25%	35	6.35%	559
55 mph	464	66.70%	177	25.51%	54	7.78%	696

Variable	Driver injury severity						Total
	No injury (N)		*Injury* (I)		*Serious injury and fatal* (F)		
60 mph or Higher	345	67.03%	131	25.55%	38	7.43%	514
No Statutory Limit	11	26.10%	21	47.65%	11	26.25%	44
Crash Type							
Collision with Fixed Object	873	58.19%	504	33.62%	123	8.18%	1500
Collision with Object Not Fixed	885	67.01%	352	26.67%	83	6.32%	1321
Rollover	61	27.05%	126	55.47%	40	17.48%	227
Vehicle Type							
Passenger Car	1237	59.36%	683	32.77%	164	7.87%	2084
Light Truck	254	58.80%	141	32.64%	37	8.56%	432
Bus	9	100.00%	0	0.00%	0	0.00%	9
Large Truck	184	58.04%	104	32.81%	29	9.15%	317
Motorcycle	136	65.70%	55	26.57%	16	7.73%	207
Action							
Going Straight	1012	57.51%	581	32.98%	168	9.51%	1761
Turning Left	181	54.83%	131	39.65%	18	5.52%	329
Stopped in Traffic Lane	270	70.28%	110	28.75%	4	0.96%	384
Turning Right	83	71.01%	32	27.19%	2	1.80%	117
Slowed in Traffic Lane	116	68.55%	51	30.20%	2	1.25%	169
Backing Up	51	87.23%	6	10.95%	1	1.81%	58
Negotiating Curve	108	46.54%	72	31.24%	51	22.23%	231
Age							
Young (<25 years)	660	56.60%	411	35.25%	95	8.15%	1166
Middle (25~64 years)	1026	61.70%	506	30.43%	131	7.88%	1663
Old (>64 years)	134	60.91%	66	30.00%	20	9.09%	220
Seat Belt used							
Used	1777	59.69%	961	32.28%	239	8.03%	2977
Not Used	43	59.72%	22	30.56%	7	9.72%	72
Drug/Alcohol Impaired	123	38.56%	117	36.68%	79	24.76%	319

Variable	Driver injury severity						Total
	No injury (N)		*Injury* (I)		*Serious injury and fatal* (F)		
Gender	2563	59.59%	1395	32.43%	343	7.97%	4301
Male	1077	59.93%	571	31.78%	149	8.29%	1797
Female	743	59.35%	412	32.91%	97	7.75%	1252

5.3. Methodology

5.3.1. Model development

As mentioned above, unobserved heterogeneity has been recognized as a critical issue in crash data analysis. The current paper adds to the growing literature of studies that address unobserved heterogeneity issues. Random parameters model (mixed logit model) and finite mixture model (latent class model) have been proven effective in treating unobserved heterogeneity issues by multiple studies on injury severities (Behnood et al., 2014; Russo et al., 2014; Shaheed and Gkritza, 2014; Wu et al., 2014; Ye and Lord, 2014; Barua et al., 2016; Behnood and Mannering, 2016; Heydari et al., 2017), but both the models have their drawbacks. When some observation groups have similar parameters, mixed logit models with conventional distributions (normal distribution, uniform distribution, etc.) may not be capable of tracking heterogeneity in the data. They may result in complex multimodal distributions with varying skewness and kurtosis (Mannering et al., 2016). On the other hand, the latent class model can identify subgroups that maximize the heterogeneity among these subgroups but have difficulties in tracking the remaining unobserved heterogeneity within each identified subgroup. To overcome these limitations, an approach derived from the latent class model and allowing random parameters within each class has been considered in previous studies (Xiong and Mannering, 2013; Buddhavarapu et al., 2016), and is also developed in this study.

Let us start with the standard finite mixture model. The underlying mechanism of the finite mixture model posits that individual behavior depends on observing attributes and on latent heterogeneity that varies with unobserved attributes. A finite number, *Q*, is predefined to classify the whole crash dataset into *Q* classes (subsets) that maximize the heterogeneity among these subsets. However, it is not clear which class contains any particular individual. Assuming that driver injury severities have *K* levels (in this study, *K*=3, indicating no injury, injury, and serious injury and fatal), the utility function determining the probability that the *i*th record (driver) belongs to class *q* ($q \in Q$) and has injury severity level *k* ($k \in K$), is the one with maximum utility as follows.

$$U_{ki|q} = \boldsymbol{\beta}_{kq}^{T} \boldsymbol{x}_{ki} + \varepsilon_{ki|q} \tag{5-1}$$

where $\boldsymbol{x}_{ki}$ is the union vector of all attributes that are included in the utility function (insignificant attributes will be eliminated in the final function), $\varepsilon_{ki|q}$ is the unobserved heterogeneity for the ith driver with the kth injury severity for class q, and $\boldsymbol{\beta}_{kq}$is the specific vector of parameters for class q.

A discrete nature of injury severity outcomes, a discrete choice model, i.e., the multinomial logit model, is assumed to generate the injury severity probability for each driver due to the discrete nature of injury

severity outcomes. Consequently, the conditional probability of the ith driver getting involved in the kth injury severity within the class q is given by

$$\text{Prob}[y_i = k|\text{class} = q] = \frac{\exp(\boldsymbol{\beta}_{kq}^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})}{\sum_{k=1}^{K} \exp(\boldsymbol{\beta}_{kq}^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})} \tag{5-2}$$

The prior probability (also known as the class probability) for the ith record (driver) in the class q, π_{iq}, is specified by the multinomial logit form, and is given by

$$\text{Prob}[\text{class} = q] = \boldsymbol{\pi}_{iq} = \frac{\exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)}{\sum_{q=1}^{Q} \exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)} \tag{5-3}$$

where $\boldsymbol{z}_\text{i}$ is a vector demonstrating the homogeneity among different individuals that reside in class q, and $\boldsymbol{\theta}_q$ is the specific vector for parameters accounting for the homogeneity within class q. The elements in z_i are a set of observed characteristics of each individual observation which enters the model for its class membership. Note that in the equation, the Qth parameter vector, $\boldsymbol{\theta}_Q$, is fixed as a constant, zero, to secure identification of the model. In addition, there may be no such homogeneity observed among individuals, and in this case, z_i turns into a 1×1 vector with only one element, one. The prior probabilities, $\boldsymbol{\pi}_{iq}$ $(\forall q \in Q)$, then become a set of simple functions of $\boldsymbol{\theta}_q$ which, by construction, can sum up to one. Finally, the probability of ith driver getting involved in the kth injury severity is the expectation (over all latent classes) of the class-specific probabilities given by

$$\text{Prob}(y_i = k) = \sum_{q=1}^{Q} \boldsymbol{\pi}_{iq} \times \text{Prob}[\text{y}_\text{i} = k|\text{class} = q] = \frac{\exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)}{\sum_{q=1}^{Q} \exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)} \times \frac{\exp(\boldsymbol{\beta}_{kq}^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})}{\sum_{k=1}^{K} \exp(\boldsymbol{\beta}_{kq}^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})} \tag{5-4}$$

Different from the standard finite mixture model, the proposed finite mixture random parameter model can account for the heterogeneity both within and across the classes, and thus can accommodate two layers of unobserved heterogeneity. The outside layer, i.e., the latent class model, assumes that individuals distinguished across the classes by different vectors of parameters that are the same for the individuals within each class. The inside layer captures unobserved heterogeneity by specifying several continuous distributions for model parameters within each class, which is the same with the framework of the standard random parameters model. Therefore, with this assumption, the conditional probability of the ith driver getting involved in the kth injury severity within the class q is given by

$$\text{Prob}[y_i = k|\text{class} = q] = \frac{\exp(\alpha_{ki} + \boldsymbol{\beta}_{i|q}^T \boldsymbol{x}_{ki}^* + \boldsymbol{\gamma}_q^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})}{\sum_{k=1}^{K} \exp(\alpha_{ki} + \boldsymbol{\beta}_{i|q}^T \boldsymbol{x}_{ki}^* + \boldsymbol{\gamma}_q^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})} \tag{5-5}$$

where α_{ki} is a specific fixed constant, and α_{Ki} is set to be zero as reference. $\boldsymbol{\beta}_{i|q}$ is a coefficient vector randomly distributed across individuals with respect to the vector of random attributes $\boldsymbol{x}_{ki}^*$, and is used to capture within-class heterogeneity. $\boldsymbol{\gamma}_q^T$ is a fixed parameter vector corresponding to the vector of fixed attributes $\boldsymbol{x}_{ki}$, and ε_{ki} is the idiosyncratic error term that is designed as an identical and independent standard normal distribution model (Koop, 2003). Specifically, the value of $\boldsymbol{\beta}_{i|q}$ equals to

$$\boldsymbol{\beta}_{i|q} = \boldsymbol{\beta}_q + \sigma_q \nu_{ki|q} \tag{5-6}$$

where $\boldsymbol{\beta}_q$ is the population mean, $\nu_{ki|q}$ is the random item with a certain continuous distribution. Here, we assume that the source of the heterogeneity, $\nu_{ki|q}$, follows a standard normal distribution with its mean of 0 and standard deviation of 1 (it can follow other distributions, and will be discussed in the

following sections). Hence, σ_q is the standard deviation of the marginal distribution of $\boldsymbol{\beta}_{i|q}$ around $\boldsymbol{\beta}_q$. Thus, as stated above

$$\boldsymbol{\beta}_{i|q} \sim \text{Normal}\,[\boldsymbol{\beta}_q, {\sigma_q}^2] \tag{5-7}$$

Then, by combining the conditional probability (Eq. (5-5)) and the prior probability (Eq. (5-3)) together, the unconditional probability of the ith driver getting involved in the kth injury severity in the framework of the finite mixture random parameter model is given by

$$\text{Prob}(y_i = k) = \frac{\exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)}{\sum_{q=1}^{Q} \exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)} \times \frac{\exp(\alpha_{ki} + \boldsymbol{\beta}_{i|q}^T \boldsymbol{x}_{ki}^* + \boldsymbol{\gamma}_q^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})}{\sum_{k=1}^{K} \exp(\alpha_{ki} + \boldsymbol{\beta}_{i|q}^T \boldsymbol{x}_{ki}^* + \boldsymbol{\gamma}_q^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})} \tag{5-8}$$

where the same notations are utilized as before. The unconditional probability is obtained by integrating $v_{ki|q}$ out of the conditional probability. The integral is approximated by sampling n replicated draws from the assumed populations and averaging at each step. In this study, maximum simulated likelihood estimation is utilized to evaluate the aforementioned parameters in the likelihood expression. The contribution of the ith driver to the total simulated likelihood is

$$f\left(\boldsymbol{y}_i \middle| \boldsymbol{\theta}_q^T, \boldsymbol{x}_{ki}^*, \boldsymbol{x}_{ki}, \boldsymbol{\beta}_{ri|q}^T, \boldsymbol{\gamma}_q^T\right) = \frac{\exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)}{\sum_{q=1}^{Q} \exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)} \times \frac{1}{R} \sum_{r=1}^{R} \frac{\exp(\alpha_{ki} + \boldsymbol{\beta}_{ri|q}^T \boldsymbol{x}_{ki}^* + \boldsymbol{\gamma}_q^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})}{\sum_{k=1}^{K} \exp(\alpha_{ki} + \boldsymbol{\beta}_{i|q}^T \boldsymbol{x}_{ki}^* + \boldsymbol{\gamma}_q^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})} \tag{5-9}$$

where R is the number of draws (replications), $\boldsymbol{\beta}_{ri|q}^T$ is the rth of R draws on the random vector $\boldsymbol{\beta}_{i|q}^T$. Collecting all terms, the simulated log likelihood function to be maximized is

$$\log L_s = \sum_{i=1}^{N} \log\left[\frac{\exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)}{\sum_{q=1}^{Q} \exp(\boldsymbol{\theta}_q^T \boldsymbol{z}_i)} \times \frac{1}{R} \sum_{r=1}^{R} \frac{\exp(\alpha_{ki} + \boldsymbol{\beta}_{ri|q}^T \boldsymbol{x}_{ki}^* + \boldsymbol{\gamma}_q^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})}{\sum_{k=1}^{K} \exp(\alpha_{ki} + \boldsymbol{\beta}_{i|q}^T \boldsymbol{x}_{ki}^* + \boldsymbol{\gamma}_q^T \boldsymbol{x}_{ki} + \varepsilon_{ki|q})}\right] \tag{5-10}$$

where N is the total number of drivers.

The conventional approach to simulate random parameter estimations prefers to use different random draws from specified distributions (Blackburn and Gaston, 2001; Cappellari and Jenkins, 2003). Generally, when the models become complex, or the number of parameters to be estimated within the model is large, the required number of random draws that can stabilize estimations and provide reasonable model convergence performance becomes large. However, a large number of draws will significantly increase computational complexity and make the model estimation time-consuming. One of the alternative approaches is to use Halton draws instead of random draws because it can produce the same level of performance with a much smaller number of draws (Train, 2000; Bhat, 2003). In this study, by balancing model goodness-of-fit and computing efficiency, the estimation of coefficients is conducted with the maximum likelihood estimation (MLE) method with 1,000 Halton draws. The estimated asymptotic covariance matrix is based on the second derivatives of the specific utility functions. If the matrix fails to be positive due to rounding errors, the Berndt–Hall–Hall–Hausman (BHHH) estimator is adopted in this study (Berndt et al., 1974).

5.3.2. Model Performance Measurement

One should note that our proposed hybrid model still has intrinsic problems that need to be carefully addressed. First, random parameters are assumed to follow certain continuous distributions to account for the within-class heterogeneity, as shown in Eq. (5-5). This technique naturally generates a modeling

issue to search for the optimal distribution for each random parameter. A feasible way to address this issue is to test the frequently used distributions on these random parameters and evaluate their performances in the modeling process based on certain performance indices. In this research, three widely used continuous distributions, including normal distribution, lognormal distribution, and uniform distribution, are selected and incorporated into the proposed finite mixture random parameter models to account for the within-class heterogeneity in the dataset. They are given by

$$\boldsymbol{\beta}_{i|q} = \boldsymbol{\beta}_q + \sigma_q \nu_{ki|q}\ ,\ \nu_{ki|q} \sim N[0,1] \tag{5-11}$$

$$\boldsymbol{\beta}_{i|q} = \exp(\boldsymbol{\beta}_q + \sigma_q \nu_{ki|q})\ ,\ \nu_{ki|q} \sim N[0,1] \tag{5-12}$$

$$\boldsymbol{\beta}_{i|q} = \boldsymbol{\beta}_q + \sigma_q \nu_{ki|q}\ ,\ \nu_{ki|q} \sim \mathrm{U}[-1,1] \tag{5-13}$$

where σ_q is a scaling parameter. The relative notations in Eqs. (5-7)- (5-10) should be simultaneously changed when these distributions are adopted.

An existing challenge of finite mixture models is that it is hard to find an optimal number of latent classes that can maximize between-class data heterogeneity and within-class data homogeneity. Both the discrepancies from different classification methods (finite mixture model part) and the different distributions of random parameters within the classes (random parameters model part) have critical contributions to the model performance since they both can change the model framework. Instead of having a finite mixture structure, some previous studies used user-defined classes to account for between-class unobserved heterogeneity. The random parameter models were separately designed in each class (Morgan and Mannering, 2011; Wu et al., 2018). Although there are differences in modeling structures, these studies also provide insightful references on the model comparisons. A series of likelihood ratio tests and model performance comparison techniques using different statistical indices are utilized in their studies to determine the optimal number of classes. In this study, a statistical accrual searching process is adopted to find the optimal number of latent classes, starting with 2 and increasing by 1 at each step up to the maximum plausible number of classes with the different distribution assumptions (e.g., a three-class model with normal distribution assumptions). Once the estimated latent class probability of each class is not significant at 5% significance level, the model is considered to reach its maximum plausible number of classes.

In order to evaluate the performance of different models, two parsimony indices, i.e., Akaike Information Criterion (AIC) (Yamaoka et al., 1978) and Bayesian Information Criterion (BIC) (Weakliem, 1999), are selected in this study. These two indices are defined in Eq. (5-14) and Eq. (5-15), respectively,

$$\mathrm{AIC} = -2\ln(L) + 2p \tag{5-14}$$

$$\mathrm{BIC} = -2\ln(L) + p \times \ln(N) \tag{5-15}$$

where $\ln(L)$ is the log-likelihood of the model, p is the number of estimated model parameters, and N is the total number of observations used to train the model. In general, lower AIC or BIC value indicates a better model fit on the studied dataset.

In addition, the McFadden Pseudo R-squared measurement is also applied to evaluate model fitness as follows

$$R^2 = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}(M_{Constant})} \quad (5\text{-}16)$$

where $\hat{L}$ is the estimated likelihood, $M_{Constant}$ is the intercept model only including the constant term, M_{Full} is the full model with the constant term and all predicting variables. The ratio of the likelihoods measures the level of improvements over the intercept model offered by the full model, and a larger McFadden Pseudo R-squared value indicates the full model has better goodness-of-fit (Domencich and McFadden, 1975).

5.3.3. Pseudo Elasticity Analysis

In a discrete choice model with a multinomial dependent variable, the sign of an estimated parameter does not necessarily indicate an increase or decrease on the probability of the response value (Kim et al., 2013; Wu et al., 2014; Osman et al., 2016). Therefore, an elasticity analysis is necessary to assess the impact of the explanatory variables in the proposed model. For continuous variables, the standard elasticity is calculated as follows (Washington et al., 2011),

$$E_{X_{kim}}^{P_{ki}} = \frac{\partial P_{ki}}{\partial X_{kim}} \frac{X_{kim}}{P_{ki}} \quad (5\text{-}17)$$

where $E_{X_{kim}}^{P_{ki}}$ is the elasticity outcome for driver i, X_{kim} is the value of the ith variable for a ith driver in the propensity function of the kth injury severity. However, Eq. (5-17) is not applicable for this study since the variables have been transformed into binary forms (with the values of 0 or 1), and the probabilities are not differentiable with respect to indicator variables. In order to deal with this problem, a direct pseudo elasticity analysis approach is proposed in this study for measuring the influence of the explanatory variables on driver injury severities, and is expressed as follows (Kim et al., 2007)

$$E_{(p)}{}_{X_{kim}}^{P_{ki}} = \frac{P_{ki}[given\ X_{kim}=1] - P_{ki}[given\ X_{kim}=0]}{P_{ki}[given\ X_{kim}=0]} \quad (5\text{-}18)$$

where the possibilities P_{ki} specific to the binary values of the attributor X_{kim}. The direct pseudo-elasticity in Eq. (5-18), $E_{(p)}{}_{X_{kim}}^{P_{ki}}$, is calculated for each record in the dataset, and the average pseudo-elasticity is calculated based on all data records to measure variable influence. In addition, when calculating the direct pseudo elasticity of random parameters, instead of using the fixed means of the parameters, the estimated distributions are adopted to generate the parameters of the corresponding variable in each record.

5.4. Model Estimation Results and Discussions

5.4.1. Model Comparison

The identified significant random parameters, estimated latent class probabilities, and model performance quantified by AIC and BIC, are provided for model comparisons and illustrated in Table 5-2. All estimated latent class probabilities are not significant at the significance level of p=0.05 when the number of classes is greater than four. It indicates that the maximum plausible number of classes for our dataset is four.

Table 5-2 Comparison results of models with different distributions and number of classes

Number of classes	**2**			**3**			**4**		
Distributions	Normal	Lognormal	Uniform	Normal	Lognormal	Uniform	Normal	Lognormal	Uniform
Significant random parameters	Young (I), Male (F), Large Truck (F)	None	Young (I), Male (F)	Young (I)	None	Young (I)	None	None	None
Log likelihood	-2022.7	-2025.6	-2024.7	-2044.1	-2044.3	-2045.4	-2052.7	-2052.7	-2052.7
AIC	4145.4	4139.2	4145.4	4180.2	4176.6	4182.8	4193.4	4193.4	4193.4
BIC	4446.5	4404.2	4434.5	4457.2	4441.6	4459.8	4458.4	4458.4	4458.4
Estimated latent class probabilities	40.18%** /59.82%**	40.18%** /59.82%**	40.18%** /59.82%**	40.18%** /26.15%* /33.67%	40.18%** /26.15%* /33.67%	40.18%** /26.15%* /33.67%	20.80%*/ 19.38%/ 26.15% /33.67%*	20.80%*/ 19.38%/ 26.15% /33.67%*	20.80%*/ 19.38%/ 26.15% /33.67%*

I = injury;

F = serious injury and fatal;

*significant at 5% significance level (p<0.05);

**significant at 1% significance level (p<0.01).

It should be noted that the other approaches using user-specified classes always have more classes than those in this study. For instance, according to Morgan and Mannering (2011), the dataset is classified into twelve classes, and the number of classes is seven in Wu et al. (2018)'s study. On the other hand, similar to this study, Xiong and Mannering (2013) developed a finite mixture structure as part of the model estimation. The results showed that a two-class-model best fitted their dataset. The significantly different classification results between the two approaches may be attributed to their structures. Furthermore, the distribution assumptions of random parameters also impact their model performance when developing the finite mixture models.

Results in Table 5-2 showed that as the number of classes continues to increase, the subtypes become assimilated. When the dataset is classified into three types, one of the three latent classes is not significant at the p=0.05 significance level. While in the four-class models, two latent classes become insignificant. On the other hand, many classes may induce adverse impacts on the estimation of random parameters. More specifically, three parameters are found normally distributed in the two-class model, while the number drops to zero in the four-class model. For models with uniform distribution assumptions, this trend is also verified; that is, the number of randomly distributed parameters decreases as the number of classes increases. In addition, the two-class models have relatively lower AICs and BICs, indicating that they are more appropriate given this dataset.

Furthermore, considering distribution assumptions alone, the normal distribution shows its superiority compared to the other two distributions. It reveals more randomly distributed parameters with much lower AIC and BIC values. Although the lognormal distribution can restrict the sign of a coefficient (keeping it positive or negative), it does not show sufficient advantages given our dataset. No random

parameters are found significant with this assumption. In general, considering all the assessment measurements, the two-class model with a normal distribution assumption is selected as the final model, and its detailed discussions are provided in the following sections.

5.4.2. Model Estimation

This section demonstrates the estimation results of the proposed two-class finite mixture random parameters model with a normal distribution assumption. The coefficients, standard errors, *p*-values, and as well as the confidence interval of each significant variable (in either latent class) of this model are all illustrated in Table 5-3.

Table 5-3 Estimation results of the finite mixture random parameter models

Variable	**Parameters in Latent Class 1**				**Parameters in Latent Class 2**			
	Coefficient	S.d.[a]	95%CI[b]		Coefficient	S.d.[a]	95%CI[b]	
			Lower	Upper			Lower	Upper
Intercept (I)	2.23	0.72	2.19	2.27	3.26	0.55	3.24	3.29
Intercept (F)	0.88*	0.25	0.87	0.89	2.72*	0.44	2.7	2.74
Mean of Random Parameters								
Young (I)	1.32**	0.19	1.31	1.33	0.99*	0.12	0.98	1.00
Male (F)	1.22**	0.05	1.22	1.22	2.35**	0.33	2.34	2.37
Large Truck (F)	-[c]	-	-	-	1.21*	0.12	1.2	1.22
Distributions of Standard Deviations of Random Parameters								
Young (I)	1.10**	0.36	1.08	1.12	1.73**	0.51	1.70	1.76
Male (F)	0.88*	0.12	0.87	0.89	2.02*	0.33	2.01	2.04
Large truck (F)	-	-	-	-	0.65**	0.12	0.64	0.66
Fixed Parameters								
Rural (I)	1.13**	0.11	1.12	1.14	0.92*	0.06	0.92	0.92
Dip (I)	-	-	-	-	1.04**	0.21	1.03	1.05
Wet (I)	-	-	-	-	-1.06**	0.27	-1.07	-1.05
60 mph or higher (I)	1.66*	0.24	1.65	1.67	-	-	-	-
No Statutory Limit (I)	-	-	-	-	2.23*	0.54	2.21	2.25
Rollover (I)	1.83*	0.11	1.82	1.84	-	-	-	-
Stopped in traffic lane (I)	-1.33**	0.24	-1.34	-1.32	-	-	-	-
Sunday (F)	1.48*	0.22	1.47	1.49	-	-	-	-
Dark (F)	-	-	-	-	1.25**	0.19	1.24	1.26

Variable	Parameters in Latent Class 1				Parameters in Latent Class 2			
	Coefficient	S.d.[a]	95%CI[b]		Coefficient	S.d.[a]	95%CI[b]	
			Lower	Upper			Lower	Upper
Curve (F)	1.26*	0.33	1.24	1.28	-	-	-	
Signal control (F)	-	-	-	-	0.87**	0.08	0.87	0.87
60 mph or higher (F)	1.53**	0.27	1.52	1.55	-	-	-	-
No Statutory Limit (F)	-	-	-	-	3.33**	0.65	3.3	3.36
Rollover (F)	2.36*	0.43	2.34	2.38	-	-	-	-
Light Truck (F)	1.25**	0.12	1.24	1.26	-	-	-	-
Drug/Alcohol impaired (F)	2.26**	0.71	2.22	2.3	-	-	-	-
Old (F)	1.37*	0.36	1.35	1.39	1.04**	0.09	1.04	1.04
Model Statistics								
Number of Observations	3049							
Estimated Class Probabilities	40.18%**				59.82%**			
Log-likelihood at constants	-2022.7							
Log-likelihood at convergence	-744.35							
McFadden Pseudo R-squared	0.632							

[a] Standard deviation;

[b] The 95% confidence interval of estimation results;

I = Injury;

F = serious injury and fatal;

* Significant at 5% significance level (p<0.05);

** Significant at 1% significance level (p<0.01);

[c] Not significant at 5% significance level (p<0.05).

As shown in Table 5-3, the entire dataset is classified into two classes, which contain 40.18% and 59.82% of total records, respectively. The McFadden Pseudo R-squared value is equal to 0.632, indicating that the model has reasonably acceptable performance compared to the intercept-only model. The two classes are remarkably different, and the variables that significantly influence driver injury severities are quite diversely distributed in these two classes. For instance, it is found that *rollover (I)* only has significant impacts on the drivers in Class 1, whereas it has no effects on the drivers in Class 2. In addition, three parameters are randomly distributed in Class 2, including *young (I)*, *male (F)*, and *large truck (F)*. Nevertheless, in Class 1, only two variables, *young (I)* and *male (F)*, have random effects. These differentiated outcomes indicate the proposed model is appropriate for analyzing the given dataset because it can capture both within- and between-class heterogeneity.

5.4.3. Pseudo Elasticity Analysis Results

As noted above, the sign of an estimated coefficient does not always represent the probability change of the injury severity outcome and is therefore not suitable for interpreting the actual impact of the variable. Consequently, pseudoelasticity estimation is adopted in this study to address this issue, and the results are illustrated in Table 5-4. Since the primary purpose of this study is to reduce the likelihood of injury severity, *serious injury and fatal* (*F*), therefore, the variables that have significant impacts on this injury severity will be carefully discussed in the following sections.

Table 5-4 Pseudo elasticity estimation results of the proposed model

Variables	**Latent Class 1**	**Latent Class 2**
Random Parameters		
Young (I)	25.73%	11.52%
Male (F)	8.03%	19.66%
Large truck (F)	-	33.25%
Fixed Parameters		
Rural (I)	22.27%	10.58%
Dip (I)	-	46.30%
Wet (I)	-	-12.24%
60 mph or Higher (I)	55.79%	-
No Statutory Limit (I)	-	47.79%
Rollover (I)	74.06%	-
Stopped in Traffic Lane (I)	-15.68%	-
Sunday (F)	31.15%	-
Dark (F)	-	25.69%
Curve (F)	16.97%	-
Signal Control (F)	-	16.83%
60 mph or Higher (F)	44.52%	-
No Statutory Limit (F)	-	31.33%
Rollover (F)	112.54%	-
Light Truck (F)	27.18%	-
Drug/Alcohol Impaired (F)	165.94%	-
Old (F)	12.68%	6.15%

I = Injury;

F = serious injury and fatal;

It is found that *day of week* plays a significant role in contributing to driver injury severities. The variable, *Sunday*, is found to increase the possibility of injury severity, *serious injury, and fatal* (*F*), in Class 1 by 31.15%, while it is not significant in Class 2. This finding is consistent with previous studies (Valent et al., 2002; Depaire et al., 2008). Such a severe injury outcome may be attributed to the high proportion of drivers under the influence of drugs and alcohol consumptions on Sunday. Besides, the driver's higher speeding possibility due to less traffic on Sunday may also contribute to the injury outcome.

The variable, *Dark*, is found to aggravate the driver injury severity outcome in Class 2, since it increases the possibility of *serious injury and fatal (F)* by 25.69%. When driving in the dark, especially in low visibility conditions, drivers are unable to recognize the roadway conditions clearly, and therefore are more likely to get involved in the crashes associated with severe injuries, such as head-on collisions with fixed objects, rollover crashes, etc. (Chow et al., 2016). Previous studies also showed similar findings (Chen et al., 2015, 2016a; Pour-Rouholamin and Zhou, 2016).

As shown in Table 5-3 and Table 5-4, the variable, *rural (I)*, is found to affect driver injury severities in both the latent classes significantly. In Class 1, *rural (I)* increases the possibility of *injury (I)* by 22.27%, while the one in Class 2 is 10.58%. Comprehensive analyses of this result indicate the rural roadways are usually associated with poor lighting conditions, defective traffic signs and markings, weak law enforcement, and other factors (Yasmin et al., 2014; Anarkooli and Hosseinlou, 2016; Chen et al., 2016b; NHTSA, 2016). Therefore, these results are consistent with the previous studies.

The variable, *Curve*, is also found to have adverse impacts on driver injury severities. It can increase the possibility of *serious injury and fatal (F)* in Class 1 by 16.97%. The result is consistent with previous studies that severe injuries are more likely to happen on the curved roads (Holdridge et al., 2005; Ye and Lord, 2014). Considering the impacts of low visibility, driving on curvy roads becomes more challenging, as the driver's sight distance becomes shorter and may not be sufficient to handle the potential hazards associated with the curved roadways.

The variable, *Wet*, is found typically associated with more serious injury severities in Class 2. Still, it is not significant in Class 1, because the estimation result shows that *injury (I)* crashes are less likely to occur on the wet road by 12.24% in Class 2. This finding is not intuitive, and some previous studies have similar conclusions (Quddus et al., 2002; Gray et al., 2008; Feng et al., 2016). This result may be attributed to the fact that drivers tend to slow down and be more cautious when driving on wet roads. Therefore, the possibilities of them suffering *injury (I)* crashes are more likely to decrease.

As shown in Table 5-4, the variable, *signal control*, can increase the possibilities of driver injury severity, *serious injury, and fatal (F)* in Class 2 by 16.83%. According to Wang and Abdel-Aty (2008), signal control methods are more likely to be implemented at intersections with complex traffic conditions. Traffic signals can provide an orderly movement of conflicting flows by alternately assigning right of way to various traffic movements. However, this method may also increase the number of conflict points and the crash potential on roadways (Huth et al., 2015). The situation may become more problematic under low visibility conditions. The available response distances for drivers at intersections may significantly decrease, and therefore crash injury possibilities may significantly increase.

The variable, *Speed limit*, is also found to have significant impacts on driver injury severity outcomes. The variable, *60 mph or higher*, is found to considerably increase the possibilities of *injury (I)* and *serious injury and fatal (F)* in Class 1 by 55.79% and 44.52%, respectively. The results are reasonable because

drivers are more likely to drive faster at a higher speed limit level. The higher the impact speed, the more serious the consequences in terms of injury and material damage. The variable, *no statutory limit*, also increases the probabilities of *injury (I)* and *serious injury and fatal (F)* in Class 2 by 47.79% and 31.33%, respectively. The reason may be that drivers are more likely to speed up when there are no clear speed limit signs and are more likely to suffer severe injuries.

The variable, *Large truck,* is found to be randomly distributed in Class 2, indicating that this variable has a possibility of heterogeneous effects across observations that cannot be observed explicitly. The variables, *Large truck,* and *Light truck* are both found to increase driver injury severity outcomes. They increase the possibilities of *serious injury and fatal (F)* by 33.25% and 27.18% in Class 2. The reasons could be that large trucks and light trucks are much heavier than passenger cars, and the forces of the impact from these vehicles are much greater in crashes (Zhu and Srinivasan, 2011; Behnood and Mannering, 2017; Ahmed et al., 2018).

The variable, *Young*, is found to be randomly distributed to affect driver injury severity outcomes under low visibility conditions. The random effects of *young* show that there is unobserved heterogeneity across observations in this variable. Although young drivers are all less than 25 years old, their physical characteristics, perception/reaction time, and risk-taking behavior are different. Therefore, crashes with the other similar attributes may have different injury outcomes (Amarasingha and Dissanayake, 2014; Weiss et al., 2014).

Older drivers in both two classes are more likely to get involved in serious injuries and fatal injuries. The variable, *old*, increases the probability of *serious injury and fatal (F)* in Class 1 by 12.68%, and the corresponding value in Class 2 is 6.15%. Donmez and Liu (2015) identified that most senior persons, as they are getting older, may have substantial difficulties in driving due to their gradual loss in visual and physical abilities. The low visibility condition can further exacerbate this issue by inducing serious challenges for senior drivers under some complex traffic circumstances, such as misjudging the time or distance, failing to stay on proper lanes, etc., and thus resulting in severe injury outcomes (Abdel-Aty, 2003; Lee and Li, 2014; Liu et al., 2015; Li et al., 2018).

The original variables, *Drug,* and *alcohol-impaired*, are combined as a single variable due to their limited number of records in the dataset. As shown in Table 5-4, this variable is expected to increase the possibility of having *serious injury, and fatal (F)* crashes in Class 1 by 165.94%. The results are reasonable because both drug and alcohol have compromising impacts on drivers' judgment, perception/reaction time, hearing range, version ability, etc. These influences may cause more severe injury outcomes under low visibility conditions where reaction acuity and timely judgment are necessary.

The variable, *Male*, is found significantly affecting drive injury severities as a random parameter. It indicates that there exist unobserved heterogeneous effects in terms of gender impacts. The pseudo elasticity analysis results show that *Male* has significant impacts on driver injury severity outcomes in both the two latent classes. It increases the possibility of having *serious injury and fatal (F)* severity outcomes in the two classes by 18.29% and 16.35%, respectively. Some studies stated that it could be partially because male drivers are more likely to get involved in crashes of DUIs (Driving Under the Influence) and traffic violations. They are also more prone to have aggressive behavior and risk-taking actions (Kim et al., 2013; Weiss et al., 2014; Li et al., 2018).

5.5. Summary

A three-year crash dataset from 2010 to 2012 is utilized to investigate low visibility related single-vehicle crashes and their significant contributing factors to driver injury severities in four South Central states, Arkansas, Louisiana, Texas, and Oklahoma. A finite mixture random parameter model is developed for analyzing this dataset. The developed model can interpret both within- and between-class unobserved heterogeneity. The model goodness-of-fit measurements, such as AIC, BIC, and McFadden pseudo-R-squared, are computed to compare the models with different numbers of latent classes and various distribution assumptions of random parameters. The two-class model with normal distribution assumptions for random parameters shows its significant superiority to the other models and is selected as the final model.

Three variables, including *young (I)*, *male (F)*, and a *large truck (F)*, are found to be normally distributed and have significant impacts on driver injury severities. The other fixed-parameter variables that have significant influences on driver injury severities include *rural*, *wet*, *60 mph or higher*, *no statutory limit*, *dark*, *Sunday*, *curve*, *rollover*, *light truck*, *old*, and *drug/alcohol-impaired*. To better interpret the model estimation results, the pseudoelasticity analysis is conducted on these significant parameters. Results show that the variables, *old* and *male*, increase the possibilities of having a *serious injury and fatal (F)* severity outcomes in both of the two latent classes. The variables, *Sunday*, *curve*, *60 mph or higher*, *rollover*, *light truck*, *old*, and *drug/alcohol impaired*, only increase the possibilities of having *serious injury and fatal (F)* severity outcomes in Class 1, while the variables, *dark*, *signal control*, and *no statutory limit*, are found to increase the likelihoods of having *serious injury and fatal (F)* severity outcomes in Class 2.

Based on the analysis results and previous engineering experience, some appropriate countermeasures and strategies could be implemented to improve traffic safety performance under low visibility conditions. First, an in-vehicle crash warning system directed toward recognizing certain adverse driving behavior (lane departure, fatigue driving, etc.) may be beneficial to drivers under low visibility conditions (Ohn-Bar et al., 2015). Besides, roadway information systems could alert the driver of poor visibility conditions ahead and advise them to appropriately reduced travel speeds or take alternate routes. Moreover, advisory and warning strategies that provide information on predicted and prevailing conditions, including object guidance, variable speed limit signs, pavement markings, etc., also have significant impacts on mitigating driver injury severity. The other general strategies including installing facilities to increase light intensity, implementing stronger sanctions for drivers with higher blood alcohol content (BAC), and developing more rigorous safety education programs can also decrease the possibilities of drivers being seriously injured.

Some limitations may affect result estimation and interpretations in this study. As aforementioned, the dataset utilized in this study is aggregated over three years to provide a sufficient number of observations. However, since crashes are rare events, driver behavior and evolution are not always constant and may change over time, resulting in potential temporal instability (Behnood and Mannering, 2015). Also, another potential problem that can generate temporal variability is that the dataset may be a non-random sample due to drivers' selectivity. Unlike travel mode choice during clear days, some of the more cautious drivers may choose other travel modes when the visibility is low. The drivers who want to drive themselves tend to get involved in more serious crashes and may be over-represented in the dataset. As further described in Mannering (2018), the consequences of ignoring possible temporal

effects and potential temporal shifts in estimated parameters, could adversely affect the conclusions drawn from our model estimations and their transferability to forecast and evaluate the effects of safety countermeasures. To address this issue, the models that allow the estimated parameters to change over time can be correspondingly developed (Bhat and Dubey, 2014; Seraneeprakarn et al., 2017). Xiong et al. (2014) adopted Markov-switching random parameter ordered probit model considering road-segment heterogeneity to accommodate both temporal instability and time-constant unobserved heterogeneity. Their estimation results illustrate that Markov switching models are appropriate to deal with the temporal instability issue and have a wide variety of applications (Malyshkina et al., 2009; Mannering and Bhat, 2014). However, the complexity of the model estimation process is quite cumbersome. Although our proposed model can account for some unobserved heterogeneity in the dataset, it should be noted that it is not able to distinguish the unobserved heterogeneity that is entirely induced by temporal variations or a combination of temporal shifts and other traditional sources of unobserved heterogeneity. This point should be kept in mind when reviewing our findings.

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusions

Traffic crashes have caused considerable incapacitating injuries and losses in rural, isolated, tribal, or indigenous (RITI) communities. It was found that crash data analysis suffers from not only the unobserved heterogeneities but also the temporal instability. What's worse, many related characteristics may have a different cycle, resulting in incomplete data records. To address the research gap, this project enhanced the interactive baseline crash data platform, which is capable of visualizing and analyzing rural crashes in RITI communities, with more interactive graphs, investigated the Bayesian vector autoregression-based approach for mixed frequency crash data interpretations with missing values, and proposed a finite mixture random parameter model to explore driver injury severity patterns and causes in low visibility conditions. This research effort has gathered and leveraged existing traffic crash databases with the state of Washington, Idaho, Alaska, and Hawaii. The proposed research enabled effective traffic safety program management at all levels in RITI communities to design and implement appropriate countermeasures to mitigate rural crash severities and risks.

The project updated the RCVTS, a web-based tool that aims to deal with visualization issues associated with various rural crash characteristics. The updated RCVTS features three new graph types. A novel Bayesian vector autoregression approach is proposed to address this problem. An unevenly spaced traffic collision data set with missing values, containing all collisions in different severities that occurred on the state highways in Washington State from January 2006 to December 2016, was selected to study the impacts of transportation-, weather- and socioeconomic-related characteristics on traffic collisions. A Gibbs sampler is used to conduct Bayesian inference for model parameters and unobserved high-frequency variables. Results show that the model has a fairly superior fit accuracy and can capture the unobserved heterogeneity in the dataset. The proposed VAR also demonstrates better performance than other missing value imputation techniques, including linear regression, predictive mean matching, k-nearest neighbors, and random forests. A three-year crash dataset including all low visibility related crashes from 2010 to 2012 in four South Central states, Arkansas, Louisiana, Texas, and Oklahoma, is adopted in this study. A finite mixture random parameter approach is developed to interpret both within-class and between-class unobserved heterogeneity among crash data. After a careful comparison, a two-class finite mixture random parameter model with normal distribution assumptions is selected as the final model. Estimation results show that three variables, including young (specific to injury, I), male (specific to serious injury and fatal, F), and a large truck (specific to serious injury and fatal, F), are found to be normally distributed and have significant impacts on driver injury severities. Variables with fixed effects including rural, wet, 60 mph or higher, no statutory limit, dark, Sunday, curve, rollover, light truck, old, and drug/alcohol-impaired also have significant influence on driver injury severities.

6.2. Recommendations

To facilities future research, the following recommendations are made:

(1) The updated RCVTS has fulfilled a lot of online crash analysis and visualization demand. However, there exists a drawback that self-defined crash records were not well supported in the system. In this case, future work would be conducted to address this issue.

(2) The proposed finite mixture random parameter model can investigate the unobserved heterogeneities for both within and between classes among crash data. However, temporal instability was not considered in the formulation. Future study to enhance the crash injury severity modeling with temporal influence is of significant interest.

REFERENCES

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34(5), 597–603.

Abdel-Aty, M., Ekram, A.-A., Huang, H., Choi, K., 2011. A study on crashes related to visibility obstruction due to fog and smoke. Accident Analysis and Prevention 43(5), 1730–1737.

Abdel-Aty, M., Hassan, H., Ahmed, M., Al-Ghamdi, A., 2012. Real-time prediction of visibility related crashes. Transportation Research Part C 24, 288–298.

Adams, K., Drenner, R., Chumchal, M., Donato, D., 2016. Disparity between state fish consumption advisory systems for methylmercury and US Environmental Protection Agency recommendations: A case study of the south central United States. Environmental Toxicology and Chemistry 35(1), 247–251.

Afghari, A., Haque, M., Washington, S., Smyth, T., 2016. Bayesian latent class safety performance function for identifying motor vehicle crash black spots. Transportation Research Record: Journal of the Transportation Research Board 2601, 90–98.

Ahmed, M., Franke, R., Ksaibati, K., Shinstine, D., 2018. Effects of truck traffic on crash injury severity on rural highways in Wyoming using Bayesian binary logit models. Accident Analysis and Prevention 117, 106–113.

Alves, L., Fasolo, A., 2015. Not Just Another Mixed Frequency Paper (No. 400), Central Bank of Brazil, Research Department.

Amarasingha, N., Dissanayake, S., 2014. Gender differences of young drivers on injury severity outcome of highway crashes. Journal of Safety Research 49, 113 -120.

Anarkooli, A., Hosseinlou, M., 2016. Analysis of the injury severity of crashes by considering different lighting conditions on two-lane rural roads. Journal of Safety Research 56, 57–65.

Anastasopoulos, P.C., 2016. Random parameters multivariate tobit and zero-inflated count data models: Addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. Analytic Methods in Accident Research 11, 17–32. doi:https://doi.org/10.1016/j.amar.2016.06.001

Andrey, J.C., Mills, B.E., Vandermolen, J., 2001. Weather information and road safety.

Ankargren, S., Unosson, M., Yang, Y., 2018. A mixed-frequency Bayesian vector autoregression with a steady-state prior. Department of Statistics, Uppsala University.

Barua, S., El-Basyouny, K., Islam, M., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research 9, 1–15.

Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z., Wang, Y., 2016. Bayesian penalized spline models for the analysis of spatio‐temporal count data. Statistics in medicine 35, 1848‐1865.

Behnood, A., Mannering, F., 2017. Determinants of bicyclist injury severities in bicycle-vehicle crashes: A random parameters approach with heterogeneity in means and variances. Analytic Methods in

Accident Research 16, 35–47.

Behnood, A., Mannering, F., 2016. An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models. Analytic Methods in Accident Research 12, 1–17.

Behnood, A., Mannering, F., 2015. The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: some empirical evidence. Analytic Methods in Accident Research 8, 7–32.

Behnood, A., Roshandeh, A., Mannering, F., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Analytic Methods in Accident Research 3, 56–91.

Berndt, E., Hall, B., Hall, R., Hausman, J., 1974. Estimation and inference in nonlinear structural models. Annals of Economic and Social Measurement 3(4), 653–665.

Bellos, V., Ziakopoulos, A., Yannis, G., 2019. Investigation of the effect of tourism on road crashes. Journal of Transportation Safety & Security 1–18.

Bhat, C., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transportation Research Part B 37(9), 837–855.

Bhat, C.R., Born, K., Sidharthan, R., Bhat, P.C., 2014. A count data model with endogenous covariates: formulation and application to roadway crash frequency at intersections. Analytic Methods in Accident Research 1, 53–71.

Bhat, C., Dubey, S., 2014. A new estimation approach to integrate latent psychological constructs in choice modeling. Transportation Research Part B 67, 68–85.

Blackburn, T., Gaston, K., 2001. Local avian assemblages as random draws from regional pools. Ecography 24(1), 50–58.

Blazquez, C.A., Celis, M.S., 2013. A spatial and temporal analysis of child pedestrian crashes in Santiago, Chile. Accident Analysis & Prevention 50, 304–311.

Brazil, N., Kirk, D.S., 2016. Uber and metropolitan traffic fatalities in the United States. American journal of epidemiology 184, 192–198.

Brijs, T., Karlis, D., Wets, G., 2008. Studying the effect of weather conditions on daily crash counts using a discrete time-series model. Accident Analysis & Prevention 40, 1180–1190.

Buddhavarapu, P., Scott, J., Prozzi, J., 2016. Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. Transportation Research Part B 91, 492–510.

Bureau of Economic Analysis, 2019. Washington State Income [WWW Document]. URL https://www.bea.gov/

Bureau of Labor Statistics, 2019. Washington State Economy at a Glance [WWW Document]. URL https://www.bls.gov/eag/eag.wa.htm

Canova, F., 2011. Methods for applied macroeconomic research. Princeton university press.

Canova, F., Ciccarelli, M., Ortega, E., 2007. Similarities and convergence in G-7 cycles. Journal of Monetary economics 54, 850–878.

Cappellari, L., Jenkins, S., 2003. Multivariate probit regression using simulated maximum likelihood. The Stata Journal 3(3), 278–294.

Carriero, A., Clark, T.E., Marcellino, M., 2015. Bayesian VARs: specification choices and forecast accuracy. Journal of Applied Econometrics 30, 46–73.

Carter, C.K., Kohn, R., 1994. On Gibbs sampling for state space models. Biometrika 81, 541–553.

Chen, C., Zhang, G., Huang, H., Wang, J., Tarefder, R., 2016a. Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. Accident Analysis and Prevention 96, 79–87.

Chen, C., Zhang, G., Tian, Z., Bogus, S., Yang, Y., 2015. Hierarchical Bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations. Accident Analysis and Prevention 85, 186–198.

Chen, C., Zhang, G., Yang, J., Milton, J., 2016b. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. Accident Analysis and Prevention 90, 95–107.

Chen, E., Tarko, A., 2014. Modeling safety of highway work zones with random parameters and random effects models. Analytic Methods in Accident Research 1, 86–95.

Chen, F., Chen, S., 2011. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accident Analysis & Prevention 43, 1677–1688. doi:https://doi.org/10.1016/j.aap.2011.03.026

Chen, F., Chen, S., Ma, X., 2018. Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. Journal of Safety Research 65, 153–159. doi:https://doi.org/10.1016/j.jsr.2018.02.010

Chen, F., Chen, S., Ma, X., 2016. Crash frequency modeling using real-time environmental and traffic data and unbalanced panel data models. International journal of environmental research and public health 13, 609.

Chen, F., Ma, X., Chen, S., 2014. Refined-scale panel data crash rate analysis using random-effects tobit model. Accident Analysis & Prevention 73, 323–332. doi:https://doi.org/10.1016/j.aap.2014.09.025

Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X., Zhou, J., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. Accident Analysis & Prevention 99, 330–341. doi:https://doi.org/10.1016/j.aap.2016.11.022.

Chiou, Y.-C., Fu, C., 2015. Modeling crash frequency and severity with spatiotemporal dependence. Analytic Methods in Accident Research 5, 43–58.

Chiou, Y.-C., Fu, C., Chih-Wei, H., 2014. Incorporating spatial dependence in simultaneously modeling crash frequency and severity. Analytic Methods in Accident Research 2, 1–11.

Chow, K., Manners, S., Meuleners, L., 2016. Risk factors for killed and serious injury intersection crashes in metropolitan Perth: 2006-2015. Accident Analysis and Prevention 42(6), 1908–1915.

Coruh, E., Bilgic, A., Tortum, A., 2015. Accident analysis with aggregated data: The random parameters negative binomial panel count data model. Analytic Methods in Accident Research 7, 37–49.

Daniels, S., Brijs, T., Nuyts, E., Wets, G., 2010. Explaining variation in safety performance of roundabouts. Accident Analysis & Prevention 42, 393–402.

Del Negro, M., Schorfheide, F., 2011. Bayesian macroeconometrics.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis and Prevention 40(4), 1257–1266.

Domencich, T.A., McFadden, D., 1975. Urban travel demand, A behavioral analysis. North-Holland Publishing, Oxford.

Dong, B., Ma, X., Chen, F., Chen, S., 2018. Investigating the Differences of Single-Vehicle and Multivehicle Accident Probability Using Mixed Logit Model. Journal of Advanced Transportation 2018.

Dong, N., Huang, H., Lee, J., Gao, M., Abdel-Aty, M., 2016. Macroscopic hotspots identification: A Bayesian spatio-temporal interaction approach. Accident Analysis & Prevention 92, 256–264. doi:https://doi.org/10.1016/j.aap.2016.04.001

Donmez, B., Liu, Z., 2015. Associations of distraction involvement and age with driver injury severities. Journal of Safety Research 52, 23–28.

Duan, Y., Lv, Y., Liu, Y.-L., Wang, F.-Y., 2016. An efficient realization of deep learning for traffic data imputation. Transportation research part C: emerging technologies 72, 168–181.

Eluru, N., Bagheri, M., Miranda-Moreno, L., Fu, L., 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. Accident Analysis and Prevention 47, 119–127.

Eraker, B., Chiu, C.W., Foerster, A.T., Kim, T.B., Seoane, H.D., 2014. Bayesian mixed frequency VARs. Journal of Financial Econometrics 13, 698–721.

Eustace, D., Aylo, A., Mergia, W.Y., 2015. Crash frequency analysis of left-side merging and diverging areas on urban freeway segments–A case study of I-75 through downtown Dayton, Ohio. Transportation research part C: emerging technologies 50, 78–85.

Federal Reserve Bank of St. Louis, 2019. Washington State Economic Data [WWW Document]. URL https://fred.stlouisfed.org/categories/27331

Feng, S., Li, Z., Ci, Y., Zhang, G., 2016. Risk factors affecting fatal bus accident severity: Their impact on different types of bus drivers. Accident Analysis and Prevention 86, 29–39.

Freeman, D.G., 2007. Drunk driving legislation and traffic fatalities: new evidence on BAC 08 laws. Contemporary Economic Policy 25, 293–308.

García-Laencina, P.J., Sancho-Gómez, J.-L., Figueiras-Vidal, A.R., Verleysen, M., 2009. K nearest neighbours

with mutual information for simultaneous classification and missing data imputation. Neurocomputing 72, 1483–1493.

Giannone, D., Lenza, M., Primiceri, G.E., 2015. Prior selection for vector autoregressions. Review of Economics and Statistics 97, 436–451.

Granger, C.W.J., Newbold, P., 2014. Forecasting economic time series. Academic Press.

Gray, R., Quddus, M., Evans, A., 2008. Injury severity analysis of accidents involving young male drivers in Great Britain. Journal of Safety Research 39(5), 483–495.

Greene, W., 2012. NLOGIT 5 Reference Guide. Econometric Software. Inc., Plainview, NY.

Haleem, K., Gan, A., 2015. Contributing factors of crash injury severity at public highway-railroad grade crossings in the US. Journal of Safety Research 53, 23–29.

Hapfelmeier, A., Hothorn, T., Ulm, K., Strobl, C., 2014. A new variable importance measure for random forests with missing data. Statistics and Computing 24, 21–34.

Hermans, E., Brijs, T., Stiers, T., Offermans, C., 2006. The impact of weather conditions on road safety investigated on an hourly basis. Transportation Research Board

Heydari, S., Fu, L., Miranda-Moreno, L., Jopseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: a joint analysis of pedestrian and cyclist injuries. Analytic Methods in Accident Research 13, 16–27.

Holdridge, J., Shankar, V., Ulfarsson, G., 2005. The crash severity impacts of fixed roadside objects. Journal of Safety Research 36(2), 139–147.

Huth, V., Sanchez, Y., Brusque, C., 2015. Drivers' phone use at red traffic lights: A roadside observation study comparing calls and visual-manual interactions. Accident Analysis and Prevention 74, 42–48.

Jones, M.P., 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. Journal of the American statistical association 91, 222–230.

Juselius, K., 2006. The cointegrated VAR model: methodology and applications. Oxford university press.

Karlsson, S., 2013. Forecasting with Bayesian vector autoregression, in: Handbook of Economic Forecasting. Elsevier, pp. 791–897.

Khan, G., Bill, A.R., Noyce, D.A., 2015. Exploring the feasibility of classification trees versus ordinal discrete choice models for analyzing crash severity. Transportation Research Part C: Emerging Technologies 50, 86–96.

Kim, J.-K., Kim, S., Ulfarsson, G., Porrello, L., 2007. Bicyclist injury severities in bicycle-motor vehicle accidents. Accident Analysis and Prevention 39(2), 238–251.

Kim, J.-K., Ulfarsson, G., Kim, S., Shankar, V., 2013. Driver-injury severity in single-vehicle crashes in California: a mixed logit analysis of heterogeneity due to age and gender. Accident Analysis and Prevention 50, 1073–1081.

Kim, J.-K., Ulfarsson, G., Shankar, V., Mannering, F., 2010. A note on modeling pedestrian-injury severity

in motor-vehicle crashes with the mixed logit model. Accident Analysis and Prevention 42(6), 1751–1758.

Koop, G., 2003. Bayesian Econometrics. Wiley, Chichester, UK.

Korobilis, D., 2013. VAR forecasting using Bayesian variable selection. Journal of Applied Econometrics 28, 204–230.

Kweon, Y.-J., 2011. Development of crash prediction models with individual vehicular data. Transportation research part C: emerging technologies 19, 1353–1363.

Landerman, L.R., Land, K.C., Pieper, C.F., 1997. An empirical evaluation of the predictive mean matching method for imputing missing values. Sociological Methods & Research 26, 3–33.

Leigh, J.P., Waldon, H.M., 1991. Unemployment and highway fatalities. Journal of health politics, policy and law 16, 135–156.

Lee, C., Li, X., 2014. Analysis of injury severity of drivers involved in single-and two-vehicle crashes on highways in Ontario. Accident Analysis and Prevention 71, 286–295.

Li, L., Li, Y., Li, Z., 2013. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. Transportation research part C: emerging technologies 34, 108–120.

Li, Z., Chen, C., Ci, Y., Zhang, G., Wu, Q., Liu, C., Qian, Z., 2018. Examining driver injury severity in intersection-related crashes using cluster analysis and hierarchical Bayesian models. Accident Analysis and Prevention 120, 139–151.

Li, Z., Chen, X., Ci, Y., Chen, C., Zhang, G., 2019a. A hierarchical Bayesian spatiotemporal random parameters approach for alcohol/drug impaired-driving crash frequency analysis. Analytic Methods in Accident Research. doi:https://doi.org/10.1016/j.amar.2019.01.002

Li, Z., Ci, Y., Chen, C., Zhang, G., Wu, Q., Qian, Z. (Sean), Prevedouros, P.D., Ma, D.T., 2019b. Investigation of driver injury severities in rural single-vehicle crashes under rain conditions using mixed logit and latent class models. Accident Analysis & Prevention 124, 219–229. doi:https://doi.org/10.1016/j.aap.2018.12.020

Liu, C., Chen, C.-L., Utter, D., 2005. Trend and pattern analysis of highway crash fatality by month and day.

Liu, C., Sharma, A., 2018. Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. Analytic methods in accident research 17, 14–31.

Liu, J., Khattak, A., Richards, S., Nambisan, S., 2015. What are the differences in driver injury outcomes at highway-rail grade crossings? Untangling the role of pre-crash behaviors. Accident Analysis and Prevention 85, 157–169.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice 44, 291–305.

Lütkepohl, H., 2005. New introduction to multiple time series analysis. Springer Science & Business Media.

Lynn, C., Schreiner, C., Campbell, R., 2002. Reducing fog-related crashes on the Afton and Fancy Gap

Mountain sections of I-64 and I-77 in Virginia. Virginia Transportation Research Center, Charlottesville, 2002.

Ma, X., Chen, F., Chen, S., 2015. Modeling crash rates for a mountainous highway by using refined-scale panel data. Transportation research record 2515, 10–16.

Ma, X., Chen, S., Chen, F., 2017. Multivariate space-time modeling of crash frequencies by injury severity levels. Analytic Methods in Accident Research 15, 29–40.

Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies 54, 187–197.

Males, M.A., 2009. Poverty as a determinant of young drivers' fatal crash risks. Journal of safety research 40, 443–448.

Malyshkina, N. V, Mannering, F.L., 2010. Zero-state Markov switching count-data models: An empirical assessment. Accident Analysis & Prevention 42, 122–130.

Malyshkina, N., Mannering, F., Tarko, A., 2009. Markov switching negative binomial models: An application to vehicle accident frequencies. Accident Analysis and Prevention 41(2), 217–226.

Malyshkina, N. V, Mannering, F.L., 2009. Markov switching multinomial logit model: an application to accident-injury severities. Accident Analysis & Prevention 41, 829–838.

Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research 17, 1–13.

Mannering, F., Bhat, C., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research 1, 1–22.

Mannering, F., Shankar, V., Bhat, C., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1–16.

McCann, K., Fontaine, M., 2016. Assessing driver speed choice in fog with the use of visibility data from road weather information systems. Transportation Research Record: Journal of the Transportation Research Board 2551, 90–99.

Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis & Prevention 26, 471–482.

Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40(1), 260–266.

Morgan, A., Mannering, F., 2011. The effects of road-surface conditions, age, and gender on driver-injury severities. Accident Analysis and Prevention 43(5), 1852–1863.

Mothafer, G.I.M.A., Yamamoto, T., Shankar, V.N., 2016. Evaluating crash type covariances and roadway geometric marginal effects using the multivariate Poisson gamma mixture model. Analytic Methods in Accident Research 9, 16–26. doi:https://doi.org/10.1016/j.amar.2015.11.001

NHTSA, 2016. 2015 motor vehicle crashes: overview. Traffic Safety Facts Research Note 2016, 1–9.

Norros, I., Kuusela, P., Innamaa, S., Pilli-Sihvola, E., Rajamäki, R., 2016. The Palm distribution of traffic conditions and its application to accident risk assessment. Analytic Methods in Accident Research 12, 48–65.

Office of the Washington State Climatologist, 2019. Climate Analysis Toolbox - Time Series [WWW Document]. URL https://cefa.dri.edu/Westmap/Westmap_home.php?page=timeseries.php

Ohn-Bar, E., Tawari, A., Martin, S., Trivedi, M., 2015. On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems. Computer Vision and Image Understanding 134, 130–140.

Oregon Department of Transportation, 2018. ODOT Statewide crash data system 2017 motor vehicle traffic crash analysis and code manual

Osman, M., Paleti, R., Mishra, S., Golias, M., 2016. Analysis of injury severity of large truck crashes in work zones. Accident Analysis and Prevention 97, 261–273.

Page, Y., 2001. A statistical model to compare road mortality in OECD countries. Accident Analysis & Prevention 33, 371–385.

Peng, Y., Abdel-Aty, M., Shi, Q., Yu, R., 2017. Assessing the impact of reduced visibility on traffic crash risk using microscopic data and surrogate safety measures. Transportation research part C: emerging technologies 74, 295–305.

Pour-Rouholamin, M., Zhou, H., 2016. Investigating the risk factors associated with pedestrian injury severity in Illinois. Journal of Safety Research 57, 9–17.

Prado, R., West, M., 2010. Time series: modeling, computation, and inference. Chapman and Hall/CRC.

Quddus, M., Noland, R., Chin, H., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. Journal of Safety Research 33(4), 445–462.

Russo, B., Savolainen, P., Schneider IV, W., Anastasopoulos, P., 2014. Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. Analytic Methods in Accident Research 2, 21–29.

Schorfheide, F., Song, D., 2015. Real-time forecasting with a mixed-frequency VAR. Journal of Business & Economic Statistics 33, 366–380.

Schorfheide, F., Song, D., Yaron, A., 2014. Identifying long-run risks: A bayesian mixed-frequency approach. National Bureau of Economic Research.

Scuffham, P.A., Langley, J.D., 2002. A model of traffic crashes in New Zealand. Accident Analysis & Prevention 34, 673–687.

Seraneeprakarn, P., Huang, S., Shankar, V., Mannering, F., Venkataraman, N., Milton, J., 2017. Occupant injury severities in hybrid-vehicle involved crashes: A random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research 15, 41–55.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis & Prevention 27, 371–389.

Shaheed, M., Gkritza, K., 2014. A latent class analysis of single-vehicle motorcycle crash severity outcomes. Analytic Methods in Accident Research 2, 30–38.

Shaheed, M., Gkritza, K., Carriquiry, A., Hallmark, S., 2016. Analysis of occupant injury severity in winter weather crashes: A fully Bayesian multivariate approach. Analytic Methods in Accident Research 11, 33–47.

Statisticat, L.L.C., 2015. Laplacesdemon: complete environment for Bayesian inference. R package version 15.

Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y.-J., Li, F., 2013. A tensor-based method for missing traffic data completion. Transportation Research Part C: Emerging Technologies 28, 15–27.

Tang, J., Zhang, G., Wang, Y., Wang, H., Liu, F., 2015. A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. Transportation Research Part C: Emerging Technologies 51, 29–40.

Tarel, J.-P., Hautiere, N., Caraffa, L., Cord, A., Halmaoui, H., Gruyer, D., 2012. Vision enhancement in homogeneous and heterogeneous fog. IEEE Intelligent Transportation Systems Magazine 4(2), 6–20.

Train, K., 2000. Halton sequences for mixed logit. Department of Economics, University of California, Berkeley.

Traynor, T.L., 2009. The impact of state level behavioral regulations on traffic fatality rates. Journal of safety research 40, 421–426.

Uc, E., Rizzo, M., Anderson, S., Dastrup, E., Sparks, J., Dawson, J., 2009. Driving under low-contrast visibility conditions in Parkinson disease. Neurology 73, 1103–1110.

Usman, T., Fu, L., Miranda-Moreno, L.F., 2012. A disaggregate model for quantifying the safety effects of winter road maintenance activities at an operational level. Accident Analysis & Prevention 48, 368–378.

Vahdati, S.S., Ala, A., Falaki, R., Fahimi, R., Safapour, A., 2016. Demographic Study of Maxillofacial Injury in Multiple Trauma Patients. Emerg Med (Los Angel) 6, 2.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S., Barbone, F., 2002. Risk factors for fatal road traffic accidents in Udine, Italy. Accident Analysis and Prevention 34(1), 71–84.

Van Lint, J.W.C., Hoogendoorn, S.P., van Zuylen, H.J., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transportation Research Part C: Emerging Technologies 13, 347–369.

Venkataraman, N., Shankar, V., Ulfarsson, G.F., Deptuch, D., 2014. A heterogeneity-in-means count model for evaluating the effects of interchange type on heterogeneous influences of interstate geometrics on crash frequencies. Analytic methods in accident research 2, 12–20.

Wang, X., Abdel-Aty, M., 2008. Analysis of left-turn crash injury severity by conflicting pattern using partial

proportional odds models. Accident Analysis and Prevention 40(5), 1674–1682.

Washington, S., Karlaftis, M., Mannering, F., 2011. Statistical and econometric methods for transportation data analysis. Chapman and Hall/CRC.

Washington State Departement of Transportation, 2010. Guide for interpreting short duration traffic count reports.

Washington State Department of Transportation, 2018. State Highway Log Planning Report 2017.

Washington State Department of Transportation, 2016. 2015 Annual Collision Summary.

Weakliem, D., 1999. A critique of the Bayesian information criterion for model selection. Sociological Methods and Research 27(3), 359–397.

Weiss, H., Kaplan, S., Prato, C., 2014. Analysis of factors associated with injury severity in crashes involving young New Zealand drivers. Accident Analysis and Prevention 65, 142–155.

Wu, Q., Chen, F., Zhang, G., Liu, X., Wang, H., Bogus, S., 2014. Mixed logit model-based driver injury severity investigations in single-and multi-vehicle crashes on rural two-lane highways. Accident Analysis and Prevention 72, 105–115.

Wu, Q., Ci, Y., Chen, C., Zhang, G., 2018. Examining driver injury severity in single-vehicle crashes: a two-step study using cluster analysis and mixed logit model. 97th Annual Meeting of the Transportation Research Board, CD-ROM, Washington, D.C.

Wu, Q., Zhang, G., Chen, C., Tarefder, R., Wang, H., Wei, H., 2016. Heterogeneous impacts of gender-interpreted contributing factors on driver injury severities in single-vehicle rollover crashes. Accident Analysis and Prevention 94, 28–34.

Wu, Y., Abdel-Aty, M., Lee, J., 2018. Crash risk analysis during fog conditions using real-time traffic data. Accident Analysis and Prevention 114, 4-11.

Xiong, Y., Mannering, F., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach. Transportation Research Part B 49, 39–54.

Xiong, Y., Tobias, J., Mannering, F., 2014. The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity. Transportation Research Part B 67, 109–128.

Xu, C., Wang, W., Liu, P., Guo, R., Li, Z., 2014. Using the Bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models. Transportation research part C: emerging technologies 38, 167–176.

Yamaoka, K., Nakagawa, T., Uno, T., 1978. Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations. Journal of Pharmacokinetics and Pharmacodynamics 6(2), 165–175.

Yang, K., Wang, X., Yu, R., 2018. A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. Transportation Research Part C: Emerging Technologies 96, 192–207. doi:https://doi.org/10.1016/j.trc.2018.09.020

Yannis, G., Papadimitriou, E., Folla, K., 2014. Effect of GDP changes on road traffic fatalities. Safety Science 63, 42–49. doi:https://doi.org/10.1016/j.ssci.2013.10.017

Yasmin, S., Eluru, N., 2018. A mixed grouped response ordered logit count model framework. Analytic Methods in Accident Research 19, 49–61. doi:https://doi.org/10.1016/j.amar.2018.06.002

Yasmin, S., Eluru, N., Bhat, C., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Analytic Methods in Accident Research 1, 23–38.

Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models. Analytic Methods in Accident Research 1, 72–85.

Yu, H., Li, Z., Zhang, G., Liu, P., 2019. A latent class approach for driver injury severity analysis in highway single vehicle crash considering unobserved heterogeneity and temporal influence. Analytic Methods in Accident Research 24, 100110.

Yu, H., Liu, P., Chen, J., Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. Accident Analysis & Prevention 66, 80-88.

Yu, H., Yuan, R., Li, Z., Zhang, G., Ma, D., 2020. Investigate Factors Affecting Driver Injury Severity in Snow-Related Rural Single-Vehicle Crashes. Accident Analysis & Prevention, In Press.

Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accident Analysis & Prevention 50, 371–376.

Yu, R., Xiong, Y., Abdel-Aty, M., 2015. A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway. Transportation research part C: emerging technologies 50, 68–77.

Yu, R., Wang, X., Abdel-Aty, M., 2017. A hybrid latent class analysis modeling approach to analyze urban expressway crash risk. Accident Analysis and Prevention 101, 37–43.

Zeng, Q., Huang, H., Pei, X., Wong, S.C., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Analytic methods in accident research 10, 12–25.

Zeng, Q., Sun, J., Wen, H., 2017. Bayesian hierarchical modeling monthly crash counts on freeway segments with temporal correlation. Journal of Advanced Transportation 2017.

Zhang, Q., Yu, H., Li, Z., Zhang, G., Ma, D., 2020. Assessing Potential Likelihood and Impacts of Landslides on Transportation Network Vulnerability. Transportation Research Part D: Transport and Environment 82, 102304.

Zhong, M., Lingras, P., Sharma, S., 2004. Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. Transportation Research Part C: Emerging Technologies 12, 139–166.

Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accident Analysis and Prevention 43(1), 49–57.

Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Analytic Methods in Accident Research 1, 39–52.