ANALYZING MOTORCYCLE CRASHES ON RURAL UNDIVIDED ROADS: A DATA-DRIVEN APPROACH

by

Swastika Barua, B.S.

A thesis submitted to the Graduate Council of

Texas State University in partial fulfillment

of the requirements for the degree of

Master of Science

with a Major in Engineering

August 2025

Chair of the Committee: Dr. Subasish Das, Assistant Professor

Committee Member: Dr. Anandi Dutta, Assistant Professor

Committee Member: Dr. Krishna Kisi, Associate Professor

**COPYRIGHT**

by

Swastika Barua

2025

**ACKNOWLEDGMENTS**

May the wisdom of Lord Buddha grant me strength and clarity as I reflect on this journey. I extend my sincere gratitude to the Artificial Intelligence in Transportation (AIT) Lab for their invaluable support throughout my research. My two years at Texas State University have been a profoundly transformative experience, both intellectually and personally. I am especially thankful to my supervisor, Dr. Subasish Das, for his outstanding mentorship, steadfast encouragement, and dedicated guidance, all of which have been instrumental in shaping my development as a researcher. I am also deeply appreciative of my committee members, Dr. Anandi Dutta and Dr. Krishna Kisi, for their insightful feedback during my proposal defense and for their continued support throughout this work. Their expertise and thoughtful recommendations have significantly contributed to the depth and quality of this research. I gratefully acknowledge the support of my friends and peers, whose encouragement, collaboration, and willingness to exchange ideas provided both academic and emotional sustenance. I owe special thanks to my husband, whose unwavering love, patience, and belief in my abilities served as a constant source of strength during the most challenging phases of this journey. Last but not least, I am profoundly indebted to my parents for their enduring faith, encouragement, and sacrifices. Their unconditional support and prayers have been the foundation of this accomplishment.

# TABLE OF CONTENTS

**Page**

[**TABLE OF CONTENTS** v](#_Toc201217821)

[**LIST OF TABLES** 1](#_Toc201217822)

[**LIST OF FIGURES** 2](#_Toc201217823)

[**LIST OF ABBREVIATIONS** 5](#_Toc201217824)

[**ABSTRACT** 6](#_Toc201217825)

[I. **INTRODUCTION** 7](#_Toc201217826)

[1.1. Research Motivation 9](#_Toc201217827)

[1.2. Study Objective and Research Questions 10](#_Toc201217828)

[1.3. Chapter Organization 12](#_Toc201217829)

[II. **LITERATURE REVIEW** 14](#_Toc201217830)

[2.1. Contributing Factors in Motorcycle Crash Risk and Severity 14](#_Toc201217831)

[2.2. Methodological Approaches to Analyze Crash Severity 19](#_Toc201217832)

[2.3. Research Gap 25](#_Toc201217833)

[2.4. Chapter Summary 25](#_Toc201217834)

[III. **STUDY DESIGN** 26](#_Toc201217835)

[3.1. Structure Data 27](#_Toc201217836)

[3.2. Narrative Data 29](#_Toc201217837)

[3.3. Descriptive Statistics 30](#_Toc201217838)

[3.4. Chapter Summary 35](#_Toc201217839)

[IV. **METHODOLOGICAL FRAMEWORK** 36](#_Toc201217840)

[4.1. Cluster Correspondence Analysis 36](#_Toc201217841)

[4.1.1. Cross-Tabulation and Contingency Analysis 37](#_Toc201217842)

[4.1.2. Steps in the CCA Algorithm 37](#_Toc201217843)

[4.1.3. Biplot Generation and Scaling Adjustments 38](#_Toc201217844)

[4.2. SHapley Additive exPlanations (SHAP) 38](#_Toc201217845)

[4.3. Latent Dirichlet Allocation 39](#_Toc201217846)

[4.4. Random Parameter Logit Model 42](#_Toc201217847)

[4.4.1. Mixed Logit Modeling Framework 43](#_Toc201217848)

[4.4.2. Marginal Effects 46](#_Toc201217849)

[4.5. Chapter Summary 47](#_Toc201217850)

[V. **DATA MINING ON TABULAR CRASH DATA** 48](#_Toc201217851)

[5.1. Cluster Correspondence Analysis 48](#_Toc201217852)

[5.1.1. Variable Importance 48](#_Toc201217853)

[5.1.2. Cluster Correspondence Analysis 51](#_Toc201217854)

[5.2. Random Parameter Logit Model 62](#_Toc201217855)

[5.2.1. Rider Behavior 62](#_Toc201217856)

[5.2.2. Rider Demographic 65](#_Toc201217857)

[5.2.3. Crash Characteristics 66](#_Toc201217858)

[5.2.4. Traffic Condition 68](#_Toc201217859)

[5.2.5. Roadway Characteristics 69](#_Toc201217860)

[5.2.6. Temporal Factors 71](#_Toc201217861)

[5.2.7. Lighting Condition 72](#_Toc201217862)

[5.2.8. First Harmful Event 74](#_Toc201217863)

[5.2.9. Random parameter and Heterogeneity in means 76](#_Toc201217864)

[5.3. Chapter Summary 89](#_Toc201217865)

[VI. **ADVANCED TOPIC MODELING ON CRASH NARRATIVES** 90](#_Toc201217866)

[6.1. LDA Topic Modeling Results 90](#_Toc201217867)

[6.2. Chapter Summary 96](#_Toc201217868)

[VII. **CONVERGENT VALIDITY** 97](#_Toc201217869)

[7.1. Synthesis of Analytical Approaches and Policy Recommendations 99](#_Toc201217870)

[7.2. Implications for Rural Road Safety 100](#_Toc201217871)

[7.2.1. Safe Speeds 100](#_Toc201217872)

[7.2.2. Safe Roads 101](#_Toc201217873)

[7.2.3. Safe Road Users 101](#_Toc201217874)

[7.2.4. Safe Vehicles and Post-Crash Response 101](#_Toc201217875)

[7.2.5. Environmental and Wildlife Management 101](#_Toc201217876)

[7.3. Chapter Summary 102](#_Toc201217877)

[VIII. **CONCLUSION** 103](#_Toc201217878)

[8.1. Limitations of the Research 104](#_Toc201217879)

[8.2. Future Scope of Research 105](#_Toc201217880)

[**REFERENCES** 106](#_Toc201217881)

[**APPENDIX** 121](#_Toc201217882)

# LIST OF TABLES

**Page**

[Table 2.1. Summary of studies using cluster-based methods 19](#_Toc201208980)

[Table 2.2. Summary of Studies Using Random Parameter Models to Analyze Crash Injury Severity 23](#_Toc201208981)

[Table 3.1. Motorcycle injury severity by year on rural undivided roadway 28](#_Toc201208982)

[Table 3.2. Fatal motorcycle crashes in rural undivided roadways 29](#_Toc201208983)

[Table 3.3. Data descriptive statistics of motorcycle crashes in rural undivided roadways 30](#_Toc201208984)

[Table 5.1. Centroids and Size of Clusters 52](#_Toc201208985)

[Table 5.2. Cluster based injury levels for motorcyclist crashes on rural undivided roads 62](#_Toc201208986)

[Table 5.3. Model Estimation Results from Cluster 1 to Cluster 5 78](#_Toc201208987)

[Table 5.4. Marginal Effect Results from Cluster 1 to Cluster 5 83](#_Toc201208988)

[Table 7.1. Comparative Table: Clusters, Model Results, and Narrative Themes 97](#_Toc201208989)

# LIST OF FIGURES

**Page**

[Figure 1.1. Motorcyclist fatalities in the U.S. (a) motorcyclist vs passenger car fatalities (b) motorcycle fatalities frequency by the time of the day 8](#_Toc201326488)

[Figure 1.2. Percent change in motorcycle crash occurrences by year for urban and rural roadways in Texas (2017–2023) 9](#_Toc201326489)

[Figure 1.3. Distribution of motorcycle crash injury severity by location (rural vs. urban) in Texas 10](#_Toc201326490)

[Figure 2.1. Crash types involving motorcyclists (a) turning collisions (b) head-on (c) lane changing (d) horizontal curve crashes (e) fixed object and (f) rear-end crashes 18](#_Toc201326491)

[Figure 3.1.Flow chart describing research approach 26](#_Toc201326492)

[Figure 3.2.Distribution of crashes by crash severity 28](#_Toc201326493)

[Figure 4.1. Graphical model representation of LDA. The boxes are “plates” representing replicates (Blei et al., 2003). 40](#_Toc201326494)

[Figure 5.1. Variable selection process with (a) XGBoost (b) Cramér’s V and (c) Cramér’s V values for pairwise variables 51](#_Toc201326495)

[Figure 5.2. Clusters produced from motorcyclist injury severity crash data 54](#_Toc201326496)

[Figure 5.3. Barplot for Cluster 1 55](#_Toc201326497)

[Figure 5.4. Barplot for Cluster 2 56](#_Toc201326498)

[Figure 5.5. Barplot for Cluster 3 57](#_Toc201326499)

[Figure 5.6. Barplot for Cluster 4 58](#_Toc201326500)

[Figure 5.7. Barplot for Cluster 5 59](#_Toc201326501)

[Figure 5.8. SHAP summary plot of (a) cluster 1 (b) cluster 2 (c) cluster 3 (d) cluster 4 and (e) cluster 5 61](#_Toc201326502)

[Figure 5.9. Marginal effects of the variable helmet worn but damaged 64](#_Toc201326503)

[Figure 5.10. Marginal effects of the variable crash speed greater than 65 mph 67](#_Toc201326504)

[Figure 5.11. Marginal effects of the variable marked lane 69](#_Toc201326505)

[Figure 5.12. Marginal effects of the variable curve graded roadway 70](#_Toc201326506)

[Figure 5.13. Marginal effects of the variable afternoon 71](#_Toc201326507)

[Figure 5.14. Marginal effects of the variable daylight 73](#_Toc201326508)

[Figure 5.15. Marginal effects of the variable overturned 76](#_Toc201326509)

[Figure 6.1. Bigrams for topic 1 91](#_Toc201326510)

[Figure 6.2. Bigrams for topic 2 92](#_Toc201326511)

[Figure 6.3. Bigrams for topic 3 93](#_Toc201326512)

[Figure 6.4. Bigrams for topic 4 94](#_Toc201326513)

[Figure 6.5. Bigrams for topic 5 95](#_Toc201326514)

[Figure 9.1. Marginal effect of the variable helmet, not worn 121](#_Toc201326515)

[Figure 9.2. Marginal effect of the variable helmet, damaged 121](#_Toc201326516)

[Figure 9.3. Marginal effect of the variable morning 122](#_Toc201326517)

[Figure 9.4. Marginal effect of the variable afternoon 122](#_Toc201326518)

[Figure 9.5. Marginal effect of the variable fall 123](#_Toc201326519)

[Figure 9.6. Marginal effect of the variable daylight 123](#_Toc201326520)

[Figure 9.7. Marginal effect of the variable dark, not lighted 124](#_Toc201326521)

[Figure 9.8. Marginal effect of the variable curve, graded 124](#_Toc201326522)

[Figure 9.9. Marginal effect of the variable straight, level 125](#_Toc201326523)

[Figure 9.10. Marginal effect of the variable center/stripe/divider 125](#_Toc201326524)

[Figure 9.11. Marginal effect of the variable marked lane 126](#_Toc201326525)

[Figure 9.12. Marginal effect of the variable stop/yield/warning sign 126](#_Toc201326526)

[Figure 9.13. Marginal effect of the variable crash speed 30 to 40 mph 127](#_Toc201326527)

[Figure 9.14. Marginal effect of the variable crash speed greater than 65 mph 127](#_Toc201326528)

[Figure 9.15. Marginal effect of the variable unsafe speed 128](#_Toc201326529)

[Figure 9.16. Marginal effect of the variable fixed object 128](#_Toc201326530)

[Figure 9.17. Marginal effect of the variable overturned 129](#_Toc201326531)

[Figure 9.18. Marginal effect of the variable rear-end 129](#_Toc201326532)

[Figure 9.19. Marginal effect of the variable run-off-road 130](#_Toc201326533)

[Figure 9.20. Marginal effect of the variable rider aged greater than 65 years old 130](#_Toc201326534)

# LIST OF ABBREVIATIONS

**Abbreviation Description**

|  |  |
| --- | --- |
| CCA | Cluster Correspondence Analysis |
| RPL | Random Parameter Logit |
| NLP | Natural Language Processing |
| LDA | Latent Dirichlet Allocation |
| RPLHM | Random Parameter Logit with Heterogeneity in Means |
| SHAP | SHapley Additive exPlanations |
| CRPL | Correlated Random Parameters Logit |
| IIA | Independence of Irrelevant Alternatives |
| WCSS | Within-Cluster Sum of Squares |

# ABSTRACT

Motorcycle crashes on rural undivided roadways remain a significant safety concern due to the high incidence of severe injuries and fatalities associated with these environments. This study analyzes a comprehensive dataset of 12,753 motorcycle crashes from rural undivided roads in Texas, integrating structured crash records and narrative reports from the Texas CRIS database. Employing a multi-method approach, the research first applies Cluster Correspondence Analysis (CCA) to identify underlying patterns in crash characteristics, followed by the estimation of cluster-based Random Parameter Logit Models to quantify the effects of rider behavior, roadway, and environmental factors on injury severity. Latent Dirichlet Allocation (LDA) topic modeling of crash narratives further supplements the analysis by uncovering key crash scenarios and thematic trends. The results consistently show that excessive speed, loss of control, helmet non-use, and hazardous roadway geometry, such as curves and grades, are primary contributors to severe and fatal crashes, particularly in high-speed run-off-road, overturn, and intersection-related incidents. Narrative analysis also highlights risks associated with nighttime visibility and animal-related crashes. The findings inform a suite of policy interventions grounded in the Safe System Approach, recommending context-sensitive speed management, rural infrastructure upgrades, helmet use promotion, and improved emergency and wildlife response as essential strategies for reducing injury severity and enhancing motorcycle safety on rural undivided highways.

# INTRODUCTION

**I.**

The motorcycle rider refers to the individual controlling the motorcycle, while the passenger is someone seated on the motorcycle but not operating it. The term motorcyclist broadly includes both the rider and the passenger. In 2022, motorcyclists accounted for 6,218 fatalities in motor vehicle traffic crashes, representing 15% of all traffic deaths (USDOT, 2024). Motorcycles, including 2- and 3-wheeled motorcycles, mopeds, scooters, minibikes, and pocket bikes, had a fatality rate of 26.16 per 100 million vehicle miles traveled, nearly 22 times higher than passenger cars. Notably, 35% of motorcycle riders involved in fatal crashes lacked valid licenses, and 28% were alcohol-impaired, the highest among vehicle types. Helmet use significantly influenced survival rates, 54% of riders killed in states without universal helmet laws were not helmeted, compared to 11% in states with such laws. Speeding was a factor in 35% of fatal motorcycle crashes, especially among younger riders aged 21-24. Around 37% of motorcyclist fatalities occurred in single-vehicle crashes, while 63% involved multiple vehicles (IISH, 2024) (see Figure 1.1). Most deaths (60%) happened between May and September, with June experiencing the highest fatalities. Additionally, 46% of motorcyclist deaths occurred on weekends, with a peak in crashes after 6 p.m.

Given these statistics, it is crucial to focus on the specific risks motorcyclists face on rural undivided roadways (Lemonakis et al., 2021; Stamey et al., 2024; Xin et al., 2017b), where unique roadway features further elevate crash risks. Rural undivided roadways, including two-way, two-lane, and four or more-lane undivided roads, present heightened risks for motorcyclists due to higher travel speeds, limited traffic control, and reduced visibility around curves and hills. These roads often lack physical barriers to separate opposing traffic flows, increasing the likelihood of head-on collisions and run-off-road crashes, which are particularly dangerous for motorcyclists.

|  |
| --- |
| A graph with numbers and lines  AI-generated content may be incorrect. |
| (a) |
| A graph showing time of day  AI-generated content may be incorrect. |
| (b) |

Figure .. Motorcyclist fatalities in the U.S. (a) motorcyclist vs passenger car fatalities (b) motorcycle fatalities frequency by the time of the day

## Research Motivation

Rural two-way undivided roadways are selected for this study because the data clearly shows that motorcycle crashes on these roads are more likely to result in serious or fatal injuries compared to urban roads. The Figure 1.2 from the Crash Records Information System (CRIS). database (2017–2023) (TxDOT, 2024) shows that rural areas experienced larger increases in motorcycle crashes over time, especially after 2020, while urban areas saw smaller or more stable changes. The Figure 1.3 highlights that a much higher proportion of fatal and incapacitating injuries happen on rural roads, even though urban areas have more crashes overall. For example, nearly 40% of fatal motorcycle injuries occur on rural roads, even though rural areas have fewer total crashes. This pattern is important because rural two-way undivided roads often have higher speeds, fewer safety features, and longer emergency response times, which all increase the risk and severity of crashes.

Figure .. Percent change in motorcycle crash occurrences by year for urban and rural roadways in Texas (2017–2023)

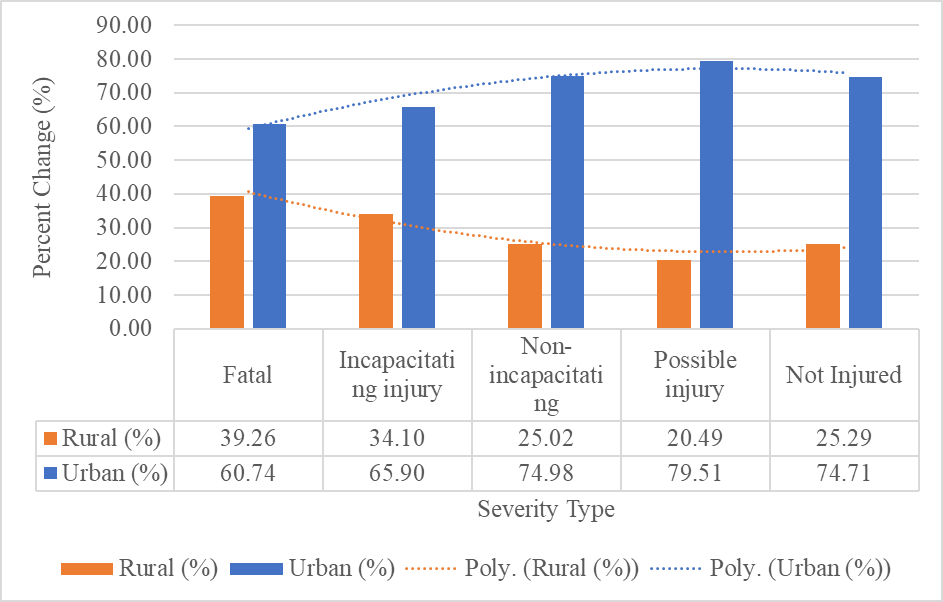


Figure .. Distribution of motorcycle crash injury severity by location (rural vs. urban) in Texas

## Study Objective and Research Questions

This study will adopt a comprehensive, multi-method approach to analyze motorcycle crashes on rural undivided roads, beginning with Cluster Correspondence Analysis (CCA) on the structured Texas crash dataset (2017–2023, 58,221 observations). The CCA method will be used to uncover key associations between categorical variables, allowing the study to group crashes with similar characteristics. By reducing dimensionality and identifying dominant patterns in the data, CCA will help form distinct crash clusters based on shared risk factors such as roadway features, crash timing, rider demographics, and environmental conditions. These clusters will serve as a foundation for further analysis of the main factors driving crash outcomes.

Next, the study will use cluster-based Random Parameter Logit Models (RPL) with heterogeneity in means to analyze crash severity within each identified cluster. This modeling approach will capture unobserved differences in the data and allow the influence of important variables to vary across clusters, providing a clearer understanding of how different risk factors affect injury severity. In addition, the study will examine unstructured narrative data using Natural Language Processing (NLP) and bigram topic modeling. By extracting and grouping key terms and themes from the narrative reports, topic modeling will help identify additional factors and causes of severe crashes that may not appear in the structured data. Together, these methods will provide a data-driven framework to support the development of targeted countermeasures for improving motorcycle safety on rural undivided roadways. This study will focus on addressing the following research questions:

**RQ1.** What behavioral, environmental, and roadway characteristics are most strongly associated with severe and fatal motorcycle crashes on rural undivided roadways, as identified through cluster correspondence analysis?

**RQ2.** How does unobserved heterogeneity influence the relationship between crash risk factors and injury severity across different crash clusters on rural undivided roadways?

**RQ3.** What key themes and topics emerge from the narrative crash data, and how do these narrative-driven factors relate to crash severity outcomes?

**RQ4.** How do the type and nature of the first harmful event influence injury severity in motorcycle crashes, and what factors explain variability in their impact?

## Chapter Organization

The Thesis has been consisting of eight chapters.

**Chapter One** introduces the study, outlining motivation, research questions, and objectives related to motorcycle crashes on rural two-way undivided roadways.

**Chapter Two** presents a review of the literature, summarizing existing research on motorcycle crash risks, severity analysis, and relevant analytical methods.

**Chapter Three** details the study design, including data sources, the structure of both tabular and narrative data, descriptive statistics, and the overall analytical approach.

**Chapter Four** explains the methodological framework used in the study. It provides a step-by-step description of the main analytical methods: CCA, SHAP analysis for variable importance, Latent Dirichlet Allocation (LDA) for topic modeling, and Random Parameter Logit Models. The chapter outlines how these methods are applied to identify patterns and risk factors associated with crash severity.

**Chapter Five** covers the analysis of the structured (tabular) crash data. It first describes the results from the CCA and then presents findings from the Random Parameter Logit Models, examining variables such as rider behavior, demographics, roadway features, crash characteristics, and the impact of first harmful events.

**Chapter Six** focuses on the analysis of narrative crash data using topic modeling, highlighting the main themes and patterns that emerge from unstructured text descriptions.

**Chapter Seven** addresses convergent validity, comparing and integrating findings from both structured and narrative data analyses to provide a comprehensive understanding of motorcycle crash severity.

**Chapter Eight** summarizes the study's key findings, discusses their implications for motorcycle safety, and provides recommendations for interventions and future research.

# LITERATURE REVIEW

**II.**

In this chapter, the main factors influencing motorcycle crash risk and severity are reviewed, with a focus placed on findings from recent studies addressing rider characteristics, roadway and environmental conditions, and advancements in safety technology. Methodological approaches used to analyze crash severity including cluster-based models, topic modeling, and random parameter techniques will be summarized. Research gaps identified in the current literature are highlighted, providing the basis for the multi-modal analytical approach adopted in this study.

## Contributing Factors in Motorcycle Crash Risk and Severity

Motorcycle crash risk and severity are shaped by a wide array of factors related to the rider, the environment, the road, traffic conditions, policy context, and vehicle-specific characteristics. Rider demographics and experience significantly influence motorcycle crash risk and injury severity. Age, riding history, and skill level are key factors that determine a motorcyclist’s vulnerability on the road. Younger and less experienced riders tend to exhibit riskier behaviors, while older riders face physical limitations that can affect their riding capabilities. Islam (2021) highlighted that younger riders, particularly those under 30, are more prone to crashes due to risk-taking behaviors and limited defensive riding skills. Similarly, Goodwin et al. (2022) reported that novice motorcyclists in their first year post-licensing have a heightened risk of crashes as they adapt to real-world riding conditions. De Rome et al. (2016) found that younger riders reported a higher number of near-miss incidents, indicating inadequate hazard detection skills. In contrast, Phillips et al. (2013) noted that older riders, while generally more cautious, are more likely to sustain severe injuries in crashes due to age-related frailty.

Helmet use remains one of the most effective protective measures for motorcyclists, substantially reducing the risk of head injuries and fatalities. However, helmet usage varies depending on legal mandates, rider compliance, and helmet type, directly affecting injury outcomes in crashes. Khor et al. (2017) demonstrated that helmets are approximately 37% effective in preventing motorcycle fatalities and 67% effective in reducing traumatic brain injuries. Christian et al. (2014) observed that helmeted riders were less likely to sustain facial injuries compared to non-helmeted riders, highlighting the additional protective benefits. De Rome et al. (2012) emphasized the importance of helmet type, revealing that non-standard helmets, such as half-shell designs, provided minimal protection during crashes. Zhao et al. (2011) conducted a meta-analysis showing that helmet use reduced fatal head injuries by 42% and brain injuries by 69%. Kashani et al. (2014) highlighted the critical role of passenger helmet use, noting a 28% reduction in fatality risk when passengers wore helmets.

Substance use, particularly alcohol and drugs, is a major contributor to motorcycle crashes. Impaired riders exhibit reduced reaction times, impaired judgment, and poor motor coordination, significantly increasing crash risks and severity. Carvalho et al. (2016) found that alcohol-positive motorcyclists were nearly three times more likely to be at fault in crashes compared to sober riders. Meanwhile, Sarmiento et al. (2020) observed that riders under the influence of alcohol or drugs were more frequently involved in single-vehicle crashes and were less likely to wear helmets, compounding injury risks. Similarly, Maistros et al. (2014) reported that intoxicated motorcyclists were disproportionately represented in fatal single-vehicle crashes, underscoring the heightened dangers of impaired riding.

Fatigue and engagement in risky riding behaviors significantly increase the likelihood of motorcycle crashes. Fatigued riders often exhibit slower reaction times and impaired decision-making, while risky behaviors, such as speeding and aggressive maneuvers, further elevate crash risks. Truong et al. (2020) studied motorcycle taxi drivers and found that prolonged working hours and physical exhaustion directly contributed to higher crash rates. Stephens et al. (2017) highlighted the prevalence of risky behaviors, such as speeding and running red lights, among riders involved in crashes. Cheng et al. (2011) reported that crash-prone riders demonstrated poorer hazard perception skills and were more likely to engage in aggressive riding (Islam and Brown, 2017).

Roadway conditions and environmental factors play a critical role in motorcycle crashes. Poor infrastructure, adverse weather, and inadequate traffic control measures can create hazardous conditions for motorcyclists, leading to increased crash risks and severity. Islam and Brown (2017)found that rural roads with limited lighting and higher speed limits experienced more severe motorcycle crashes compared to urban roads. Meanwhile, Haque et al. (2012) identified intersections as high-risk locations for motorcyclists, especially those involving uncontrolled left turns. Jeon and Woo (2024) further explored how streetscape features, primarily designed for bicycles, could inform safer urban planning for motorcycles by promoting traffic-calming measures.

The relationship between motorcycle crashes and the characteristics of rural undivided roads has been extensively studied, revealing significant safety concerns. It has been established that horizontal curves on rural undivided roads are major contributors to motorcycle crashes, particularly when curves have tighter radii and lack adequate signage, which increase the risk of accidents (Khan et al., 2013). Similarly, the design of horizontal curves has been shown to play a critical role in crash frequency, with sharper curves and poor pavement conditions contributing to higher rates of motorcycle crashes on rural two-lane undivided highways (Xin et al., 2017b). Speeding-related motorcycle crashes have also been frequently associated with rural undivided roads, especially in areas with poor geometric designs and insufficient traffic control measures, leading to higher frequencies of severe and fatal crashes (Das et al., 2022). Rider behavior has also been identified as a contributing factor, as studies have shown that motorcyclists often maintain higher speeds and follow trajectories close to the road centerline on curved sections, particularly under poor lighting conditions, which increases the likelihood of lane departures and collisions (Lemonakis et al., 2021). Figure 2.1 illustrates motorcyclists involved crash types.

|  |  |
| --- | --- |
| A crosswalk with cars and a motorcycle  AI-generated content may be incorrect. | A crosswalk with cars and a motorcycle  AI-generated content may be incorrect. |
| (a) | (b) |
| A road with cars and a arrow  AI-generated content may be incorrect. | A motorcycle on a road  AI-generated content may be incorrect. |
| (c) | (d) |
| A crosswalk with a street light and a scooter  AI-generated content may be incorrect. | A video game screen of a crosswalk  AI-generated content may be incorrect. |
| (e) | (f) |

Figure .. Crash types involving motorcyclists (a) turning collisions (b) head-on (c) lane changing (d) horizontal curve crashes (e) fixed object and (f) rear-end crashes

In addition to road design and rider behavior, the severity of motorcyclist injuries has been found to be notably higher in single-vehicle crashes on rural undivided roads, often resulting from high speeds, loss of control, and collisions with fixed objects (Schneider and Savolainen, 2011). Further attention has been drawn to motorcycle crashes involving fixed objects, where poor roadside infrastructure, such as inadequate lighting and ineffective guardrails, combined with risky behaviors like speeding and alcohol use, significantly increase the severity of injuries (Das et al., 2024a). These findings emphasize that, beyond roadway geometry, environmental conditions and rider behaviors substantially influence crash outcomes.

Technological innovations have played a growing role in motorcycle safety through the development of crash prevention systems and rider-assist technologies. Advances such as anti-lock braking systems (ABS), connected vehicle technologies, and wearable safety devices have contributed to reducing crash risks and injury severity. Song et al. (2017)) demonstrated that connected vehicle technologies, which provide real-time alerts to motorcyclists, effectively reduced crash involvement. Jiang et al. (2020) utilized GIS mapping to identify motorcycle crash hotspots, enabling targeted infrastructure improvements.

## Methodological Approaches to Analyze Crash Severity

Recent literature underscores the value of correspondence analysis (CA) and clustering techniques for advancing the study of crash causation in traffic safety. Unlike conventional regression models that primarily focus on estimating the individual effects of variables, CA-based methods such as CCA (Chakraborty et al., 2024; Das, 2021a; Rahman et al., 2023), multiple correspondence analysis (MCA) (Das et al., 2018; Liu et al., 2022), and taxicab correspondence analysis (TCA) (Das, 2021a) are particularly adept at revealing the complex interplay among multiple categorical factors present in crash datasets. These approaches enable the visualization of associations between variables in a low-dimensional space, which makes it possible to identify latent groupings and interaction effects that are often overlooked with standard statistical techniques. Through these methods, intricate links among crash characteristics such as time of day, environmental conditions, barrier types, and road user behavior can be revealed, facilitating a more holistic exploration of how driver behavior, roadway environments, and vehicle characteristics combine to produce distinct crash patterns (Chakraborty et al., 2024; Das et al., 2018; Das and Sun, 2014; Rahman et al., 2023). Table 2.1 presents the summary of cluster-based crash severity analysis methods.

Table .. Summary of studies using cluster-based methods

|  |  |  |
| --- | --- | --- |
| **Author(s) / Year** | **Methodology** | **Key Findings** |
| (Rahman et al., 2024) | CCA | Clustered pedestrian-involved hit-and-run crashes; clusters linked to nighttime, alcohol, high-speed environments; distinct risk patterns and tailored countermeasures. |
| (Liu et al., 2022) | Latent Class Clustering + Mixed Logit | VRU (pedestrian & cyclist) crash severity segmented by cluster; membership altered predictor impact. |
| (Kim et al., 2024) | MCA | Road infrastructure and traffic density patterns for high-risk pedestrian locations; built environment and traffic clusters were dominant. |
| (Moskal et al., 2012) | MCA | Helmet use and road environment were defining factors across clusters in pedestrian/bike crashes. |
| (Chakraborty et al., 2024) | CCA | Barrier-related motorcycle crash clusters varied by barrier type, road geometry, and driver risk; recommendations for targeted barrier placement and design. |
| (Das, 2021b) | TCA | Motorcycle roadway departure crash clusters; random parameters found cluster-specific risk factors. |
| (Islam and Brown, 2017b) | MCA | Motorcycle-barrier crash severity: clusters by barrier type, surface, and environment; infrastructure and awareness recommended. |
| Qin et al. (2013) | Hierarchical Clustering, MCA | Single-vehicle motorcycle crash clusters shaped by rider age, environment, and surface; tailored safety programs suggested. |
| (Das, 2021a) | CCA | Key factors in Louisiana motorcycle crashes: clusters defined by crash type, environment, and demographics. |
| (Islam, 2021b) | TCA | Taxicab clusters in motorcycle crashes tied to road, rider, and vehicle factors; each cluster showed unique risk patterns. |
| (Nguyen et al., 2021) | Cluster Analysis + Regression | Cluster-specific variables (age, gender, passenger) altered crash severity in Hanoi motorcycle crashes. |
| (Sivasankaran et al., 2021) | MCA | Severe motorcycle injury clusters in India: infrastructure and environment dominant factors. |
| (Moskal et al., 2012) | MCA | Environmental factors and helmet use varied across motorcycle crash clusters. |

Regarding crash narrative analysis, recent advancements in traffic safety research have highlighted the value of topic modeling as an innovative approach for analyzing crash causation, particularly when working with large collections of narrative crash reports. Topic modeling techniques, such as LDA and related probabilistic algorithms, are adept at extracting latent thematic structures from unstructured textual data. For instance, studies leveraging LDA have demonstrated the capacity to identify clusters of risk factors related to rider actions, adverse weather, and roadway geometry in motorcycle crashes, yielding insights not readily available from structured datasets alone (Das et al., 2021). Additionally, the use of topic modeling to analyze autonomous vehicle crash narratives has enabled the identification of key situational triggers such as crosswalk presence, turning maneuvers, and signalized intersections that are crucial for understanding crash risk among vulnerable road users (Boniphace Kutela et al., 2025). combining topic models with Bayesian networks or text network analysis have been able to quantify the relationships between narrative-derived factors and crash outcomes, improving both risk prediction and the identification of high-priority intervention points (Zhao et al., 2018). This methodological synergy is particularly beneficial for extracting crash contexts, including the roles of driver fatigue, distraction, and even regulatory or policy debates, as observed in the application of topic modeling to trucking industry crash reports (Levy and Franklin, 2014). This shift toward the integration of advanced text analytics marks a significant progression in the field’s capacity to address the multifaceted nature of crash risk and safety (Das et al., 2021; Boniphace Kutela et al., 2025; Levy and Franklin, 2014; Zhao et al., 2018).

The development of random parameter modeling approaches has marked a significant methodological shift in crash severity analysis. The initial adoption of the random parameter logit (RPL), or mixed logit model, addressed a central limitation of traditional fixed-parameter models: the inability to account for unobserved heterogeneity among crash events and road users. By allowing model coefficients to vary randomly across observations, the RPL framework enables analysts to capture the inherently diverse influences of factors such as roadway design, weather, driver characteristics, and vehicle type on crash outcomes. Early applications of RPL, such as those by (Anastasopoulos and Mannering, 2011; Eluru et al., 2007; Xin et al., 2017a). They consistently demonstrated improved fit and interpretability for injury severity modeling compared to fixed-parameter approaches (Eluru et al., 2007; Xin et al., 2017a). Recognizing that even greater realism could be achieved by acknowledging systematic variation in the mean effects of random parameters, researchers advanced to random parameter models with heterogeneity in means (RPLHM). This approach allows the mean of each random parameter to depend on observed covariates, further capturing how contextual factors modulate crash risk. For example, Anastasopoulos and Mannering (2011) and Das et al. (2024) leveraged RPLHM to show that not only do certain factors exert variable influence, but that this variability itself is systematically related to roadway environment or user characteristics. Such models have been particularly useful for contexts like motorcycle and toll road crashes, where local factors or population subgroups meaningfully alter the baseline risk.

Table 2.2 presents a summary of studies that applied random parameter-based models to examine crash injury severity, highlighting methodological approaches, data sources, and key findings related to heterogeneity in crash outcomes.

Table .. Summary of Studies Using Random Parameter Models to Analyze Crash Injury Severity

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Author (Year)** | **Study Focus** | **Data Source** | **Methodology** | **Notable Random Effects / Heterogeneity** | **Key Findings** |
| (Anastasopoulos and Mannering, 2011) | Crash injury severities on interstates | Indiana interstates, 5 years | RPL | Roadway geometrics, weather, pavement | Random parameter models outperform fixed; unobserved heterogeneity important. |
| (Das et al., 2025c) | Pedestrian hit-and-run injury severity | Louisiana, 2017–2021 | RPL | Driver condition, pedestrian actions | Bayesian approach reveals high injury when driver condition is 'normal'. |
| (Islam, 2021b) | Motorcycle injury severity by age group | US national crash data, 2017–2019 | RPLMV | Rider age, time, region, environment | Older motorcyclists more likely to suffer severe/fatal injuries. |
| (Das et al., 2024b) | Toll road crash severity | UK motorway toll crashes | Mixed Logit | Time of day, lighting, weather | Night, adverse weather, and curves increase injury severity risk. |
| (Ukkusuri et al., 2012) | Rear-end crash severity by vehicle type | Chinese expressway crash data | RPLMV | Vehicle type, roadway type, season | Heterogeneity-in-means vital to modeling severity; bus crashes riskier. |
| (Hossain et al., 2025) | Ambulance crash risk: pre- vs. pandemic era | Texas, 2017–2022 | RPL | Crash context, lighting, pandemic period | COVID-19 era changed injury patterns; more risk in dark/wet conditions. |
| (Wang, 2022) | Rider & pillion passenger injury, MC crashes | Taiwan, 2007–2010 | Random-parameters bivariate probit (RPBP) | Rider & pillion factors, crash location | Joint estimation improves insights into paired outcomes. |
| (Eluru et al., 2007) | MC crash severity | US motorcycle crash data | RPL | Rider behavior, environment | Accounting for random effects reveals unobserved risk factors. |
| (Kim et al., 2024) | Mixed vehicle crash injury severity | South Korea, 2012–2015 | RPLMV | Vehicle type, road/traffic conditions | Different vehicle types require tailored safety policies. |

## Research Gap

Despite considerable research on motorcycle crash risk and severity, several key gaps remain. First, most previous studies focus either on structured crash data or narrative text, but rarely integrate both to provide a holistic view of crash causation. While cluster-based methods (such as CCA and MCA) and random parameter models have revealed valuable insights into patterns and heterogeneity in crash severity, these approaches are usually applied separately and not in a complementary fashion. There is also limited research specifically examining rural two-way undivided roadways, even though data shows these roads are associated with higher proportions of fatal and severe injury outcomes. In addition, while topic modeling of narrative crash reports is increasingly recognized as a useful tool for extracting themes and risk factors, few studies combine these unstructured insights with advanced econometric models to explore the underlying mechanisms of crash severity in detail. Finally, most existing research does not fully address how unobserved heterogeneity and complex interactions among behavioral, environmental, and roadway factors collectively influence crash outcomes, especially for high-risk rural environments.

## Chapter Summary

This chapter reviews the wide range of factors influencing motorcycle crash risk and severity, including rider behavior, environmental and roadway characteristics, and technological interventions. It highlights recent methodological advances, such as cluster correspondence analysis, topic modeling, and random parameter logit models, used to analyze crash data. The review identifies key research gaps, particularly the lack of studies that combine structured and unstructured data and focus on rural two-way undivided roads. These insights provide the foundation for the integrated, multi-modal analysis undertaken in this thesis.

# STUDY DESIGN

**III.**

Motorcycle crashes on rural undivided roads present significant safety concerns due to factors such as high speeds, challenging roadway geometry, limited visibility, and impaired driving, all of which contribute to severe and fatal outcomes. Given these risks, a data-driven approach is essential to identify key crash patterns, assess contributing factors, and support the development of effective countermeasures to enhance motorcycle safety on these roadways.

The diagram in Figure 3.1 outlines the analytical framework for studying motorcycle crashes on rural undivided roads using the Texas CRIS database (2017–2023) with 12,753 observations and 21 variables. The process begins with data preprocessing, followed by three modeling approaches: Variable importance analysis using XGBoost to identify key factors influencing crash severity, NLP with Topic Modeling to extract themes from crash narratives, and a mixed logit model incorporating heterogeneity in means and variance to capture variations in crash outcomes. This integrated approach aims to provide a comprehensive understanding of crash factors and dynamics.

A computer screen shot of a computer

AI-generated content may be incorrect.

Figure ..Flow chart describing research approach

## Structure Data

This study utilizes crash data collected from the Texas CRIS database, which includes detailed records of motor vehicle crashes across Texas. From an initial dataset of 58,221 unique crashes, a total of 12,753 crashes were filtered to focus specifically on motorcycle crashes occurring on rural undivided roadways between 2017 and 2023. The KABCO scale classifies crash injury severity as killed (K) fatal injury within 30 days, incapacitating injury (A) severe injuries preventing normal activities, non-incapacitating injury (B) visible but non-disabling injuries, possible injury (C) complaints of pain without visible injury, and no injury/property damage only (O) no apparent harm but property damage occurred.

The map in Figure 3.2 illustrates the distribution of motorcycle crashes across Texas. The map illustrates crash severity distribution across Texas, with points representing various severity levels: fatal, incapacitating injury, non-incapacitating injury, possible injury, and non-injured cases. Clusters of higher-severity crashes, notably fatal and incapacitating injuries, are concentrated in urban regions like Dallas-Fort Worth, Houston, and San Antonio, indicating areas with increased crash risks.

A map of the united states

AI-generated content may be incorrect.

Figure ..Distribution of crashes by crash severity

Injury prevalence varied over the years, with incapacitating injuries being the most frequent each year, peaking in 2022 with 668 cases, while fatal crashes were highest in 2023 with 203 fatalities as shown in Table 3.1.

Table .. Motorcycle injury severity by year on rural undivided roadway

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Year** | **K** | **A** | **B** | **C** | **O** | **Grand Total** |
| 2017 | 146 | 579 | 607 | 224 | 253 | 1809 |
| 2018 | 133 | 464 | 532 | 256 | 274 | 1659 |
| 2019 | 125 | 517 | 551 | 232 | 247 | 1672 |
| 2020 | 182 | 541 | 535 | 256 | 264 | 1778 |
| 2021 | 177 | 657 | 572 | 279 | 242 | 1927 |
| 2022 | 177 | 668 | 572 | 272 | 213 | 1902 |
| 2023 | 203 | 654 | 663 | 283 | 203 | 2006 |
| Grand Total | 1143 | 4080 | 4032 | 1802 | 1696 | 12753 |

## Narrative Data

In Table 3.2, images from Google Earth Pro offer visual context of crash locations on rural undivided roadways, helping assess road geometry, surroundings, and visibility issues. The crash scenarios highlight key risk factors such as impaired driving in all three cases, unsafe speeds on curves and hillcrests leading to loss of control, and poor visibility, particularly in dark, unlit conditions, increasing crash severity.

Table .. Fatal motorcycle crashes in rural undivided roadways

|  |  |
| --- | --- |
| An aerial view of a road  AI-generated content may be incorrect. | **Crash Scenario 1** The crash happened around midday on July 18, 2020, on SH-16, a rural road with a 60 mph speed limit, characterized by a curve and a hillcrest. "*In this case, the driver of Unit 1, who was under the influence, lost control and veered into oncoming traffic, striking a group of motorcyclists riding in a staggered formation.”* The impact was devastating, resulting in the deaths of four motorcyclists and injuries to several others. Post-crash toxicology revealed that the Unit 1 driver had a BAC of 0.210 and tested positive for marijuana, highlighting the significant influence of substance use in this event. |
| An aerial view of a road in a forest  AI-generated content may be incorrect. | **Crash Scenario 2** The third crash took place on October 8, 2021, at 6:05 PM on FM 306, another rural road featuring a curve and grade with a 55 mph speed limit. “*During daylight and clear weather, two motorcycles traveling at unsafe speeds lost control, leading to a chain-reaction collision in which one motorcycle struck a tree.”* Both motorcyclists died as a result of the crash, while a third unit's driver survived. Toxicology results for the deceased motorcyclists showed BAC levels of 0.177 and 0.122, indicating that alcohol impairment played a significant role in the crash dynamics. |

## Descriptive Statistics

The descriptive statistics in Table 3.3 reveal key patterns across crash severities. Unsafe speed was a leading contributing factor, particularly in fatal (29.5%) and incapacitating injury crashes (32.8%). Most crashes occurred on straight, level roads, with dry surface conditions dominating across all severities. Clear weather was the most common during crashes, while poor lighting (e.g., “dark, not lighted”) was notable in severe cases. Run-off-road events were frequent, especially in non-fatal crashes. Helmet usage varied, with a significant proportion of fatalities involving riders without helmets (49.5%). Alcohol involvement was highest in fatal crashes (39.8%).

Table .. Data descriptive statistics of motorcycle crashes in rural undivided roadways

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables and Attributes** | **K N=1,143** | **A  N=4,080** | **B  N=4,032** | **C  N=1,802** | **O  N=1696** |
| ***Contributing factor*** | | | | | |
| Animal | 55 (4.81%) | 407 (9.98%) | 429 (10.6%) | 101 (5.60%) | 124 (7.31%) |
| Failed to yield/signal | 275 (24.1%) | 588 (14.4%) | 413 (10.2%) | 190 (10.5%) | 174 (10.3%) |
| Inattention/fatigued | 13 (1.14%) | 157 (3.85%) | 196 (4.86%) | 76 (4.22%) | 78 (4.60%) |
| Unsafe speed | 337 (29.5%) | 1337 (32.8%) | 1273 (31.6%) | 476 (26.4%) | 562 (33.1%) |
| NA | 290 (25.4%) | 929 (22.8%) | 922 (22.9%) | 604 (33.5%) | 476 (28.1%) |
| Other | 173 (15.1%) | 662 (16.2%) | 799 (19.8%) | 355 (19.7%) | 282 (16.6%) |
| ***Crash hour*** | | | | | |
| Dawn | 52 (4.55%) | 168 (4.12%) | 172 (4.27%) | 71 (3.94%) | 68 (4.01%) |
| Morning | 237 (20.7%) | 849 (20.8%) | 834 (20.7%) | 364 (20.2%) | 373 (22.0%) |
| Afternoon | 291 (25.5%) | 1021 (25.0%) | 992 (24.6%) | 467 (25.9%) | 421 (24.8%) |
| Evening | 131 (11.5%) | 513 (12.6%) | 527 (13.1%) | 223 (12.4%) | 200 (11.8%) |
| Night | 132 (11.5%) | 521 (12.8%) | 495 (12.3%) | 217 (12.0%) | 228 (13.4%) |
| Midnight | 300 (26.2%) | 1008 (24.7%) | 1012 (25.1%) | 460 (25.5%) | 406 (23.9%) |
| **Season** | | | | | |
| Fall | 336 (29.4%) | 1123 (27.5%) | 1073 (26.6%) | 494 (27.4%) | 484 (28.5%) |
| Spring | 310 (27.1%) | 1209 (29.6%) | 1198 (29.7%) | 491 (27.2%) | 479 (28.2%) |
| Summer | 319 (27.9%) | 1146 (28.1%) | 1176 (29.2%) | 465 (25.8%) | 458 (27.0%) |
| Winter | 178 (15.6%) | 602 (14.8%) | 585 (14.5%) | 352 (19.5%) | 275 (16.2%) |
| ***Posted Speed Limit*** | | | | | |
| Less than 25 | 4 (0.35%) | 78 (1.91%) | 118 (2.93%) | 74 (4.11%) | 56 (3.30%) |
| 30 to 45 | 260 (22.7%) | 1249 (30.6%) | 1438 (35.7%) | 887 (49.2%) | 720 (42.5%) |
| 50 to 65 | 644 (56.3%) | 2033 (49.8%) | 1882 (46.7%) | 651 (36.1%) | 707 (41.7%) |
| Greater than 65 | 233 (20.4%) | 689 (16.9%) | 564 (14.0%) | 176 (9.77%) | 195 (11.5%) |
| Other | 2 (0.17%) | 31 (0.76%) | 30 (0.74%) | 14 (0.78%) | 18 (1.06%) |
| ***Weather*** | | | | | |
| Clear | 942 (82.4%) | 3388 (83.0%) | 3346 (83.0%) | 1495 (83.0%) | 1402 (82.7%) |
| Cloudy | 163 (14.3%) | 594 (14.6%) | 537 (13.3%) | 214 (11.9%) | 207 (12.2%) |
| Fog | 10 (0.87%) | 14 (0.34%) | 15 (0.37%) | 9 (0.50%) | 8 (0.47%) |
| Rain | 20 (1.75%) | 65 (1.59%) | 117 (2.90%) | 78 (4.33%) | 69 (4.07%) |
| Severe crosswinds | 3 (0.26%) | 16 (0.39%) | 11 (0.27%) | 3 (0.17%) | 5 (0.29%) |
| Other | 5 (0.44%) | 3 (0.07%) | 6 (0.15%) | 3 (0.17%) | 5 (0.29%) |
| ***Lighting condition*** | | | | | |
| Daylight | 651 (57.0%) | 2658 (65.1%) | 2834 (70.3%) | 1233 (68.4%) | 1201 (70.8%) |
| Dark, lighted | 74 (6.47%) | 237 (5.81%) | 243 (6.03%) | 191 (10.6%) | 153 (9.02%) |
| Dark, not lighted | 368 (32.2%) | 1051 (25.8%) | 838 (20.8%) | 332 (18.4%) | 299 (17.6%) |
| Dawn | 8 (0.70%) | 35 (0.86%) | 36 (0.89%) | 19 (1.05%) | 21 (1.24%) |
| Dusk | 40 (3.50%) | 97 (2.38%) | 77 (1.91%) | 23 (1.28%) | 19 (1.12%) |
| Unknown | 2 (0.17%) | 2 (0.05%) | 4 (0.10%) | 4 (0.22%) | 3 (0.18%) |
| ***Roadway alignment*** | | | | | |
| Curve, grade | 130 (11.4%) | 465 (11.4%) | 422 (10.5%) | 121 (6.71%) | 144 (8.49%) |
| Curve, hillcrest | 23 (2.01%) | 79 (1.94%) | 67 (1.66%) | 17 (0.94%) | 33 (1.95%) |
| Curve, level | 285 (24.9%) | 975 (23.9%) | 864 (21.4%) | 288 (16.0%) | 347 (20.5%) |
| Straight, grade | 124 (10.8%) | 351 (8.60%) | 351 (8.71%) | 122 (6.77%) | 126 (7.43%) |
| Straight, hillcrest | 38 (3.32%) | 105 (2.57%) | 103 (2.55%) | 35 (1.94%) | 35 (2.06%) |
| Straight, level | 543 (47.5%) | 2096 (51.4%) | 2221 (55.1%) | 1215 (67.4%) | 1009 (59.5%) |
| Other | 0 (0.00%) | 9 (0.22%) | 4 (0.10%) | 4 (0.22%) | 2 (0.12%) |
| ***Surface condition*** | | | | | |
| Dry | 1098 (96.1%) | 3878 (95.0%) | 3741 (92.8%) | 1649 (91.5%) | 1535 (90.5%) |
| Ice | 1 (0.09%) | 1 (0.02%) | 1 (0.02%) | 3 (0.17%) | 0 (0.00%) |
| Sand, mud, dirt | 8 (0.70%) | 44 (1.08%) | 50 (1.24%) | 17 (0.94%) | 29 (1.71%) |
| Standing water | 1 (0.09%) | 2 (0.05%) | 17 (0.42%) | 6 (0.33%) | 4 (0.24%) |
| Wet | 27 (2.36%) | 131 (3.21%) | 186 (4.61%) | 113 (6.27%) | 110 (6.49%) |
| Other | 8 (0.70%) | 24 (0.59%) | 37 (0.92%) | 14 (0.78%) | 18 (1.06%) |
| ***Traffic condition*** | | | | | |
| Center stripe/divider | 222 (19.4%) | 828 (20.3%) | 770 (19.1%) | 258 (14.3%) | 295 (17.4%) |
| Marked lanes | 365 (31.9%) | 1186 (29.1%) | 1201 (29.8%) | 595 (33.0%) | 579 (34.1%) |
| No passing zone | 246 (21.5%) | 708 (17.4%) | 598 (14.8%) | 154 (8.55%) | 191 (11.3%) |
| Signal light | 32 (2.80%) | 115 (2.82%) | 127 (3.15%) | 142 (7.88%) | 86 (5.07%) |
| Stop/yield/warning | 144 (12.6%) | 461 (11.3%) | 416 (10.3%) | 251 (13.9%) | 203 (12.0%) |
| None | 120 (10.5%) | 711 (17.4%) | 853 (21.2%) | 368 (20.4%) | 299 (17.6%) |
| Other | 14 (1.22%) | 71 (1.74%) | 67 (1.66%) | 34 (1.89%) | 43 (2.54%) |
| ***Harmful event*** | | | | | |
| Animal | 48 (4.20%) | 346 (8.48%) | 338 (8.38%) | 80 (4.44%) | 96 (5.66%) |
| Fixed object | 290 (25.4%) | 741 (18.2%) | 562 (13.9%) | 222 (12.3%) | 218 (12.9%) |
| Motor vehicle in transport | 530 (46.4%) | 1372 (33.6%) | 1167 (28.9%) | 963 (53.4%) | 618 (36.4%) |
| Overturned | 259 (22.7%) | 1533 (37.6%) | 1851 (45.9%) | 471 (26.1%) | 708 (41.7%) |
| Parked car | 2 (0.17%) | 27 (0.66%) | 30 (0.74%) | 40 (2.22%) | 16 (0.94%) |
| Other | 14 (1.22%) | 61 (1.50%) | 84 (2.08%) | 26 (1.44%) | 40 (2.36%) |
| ***Intersection related*** | | | | | |
| Driveway access | 113 (9.89%) | 351 (8.60%) | 301 (7.47%) | 204 (11.3%) | 161 (9.49%) |
| Intersection | 182 (15.9%) | 498 (12.2%) | 402 (9.97%) | 252 (14.0%) | 227 (13.4%) |
| Intersection related | 74 (6.47%) | 306 (7.50%) | 425 (10.5%) | 275 (15.3%) | 214 (12.6%) |
| Non intersection | 774 (67.7%) | 2925 (71.7%) | 2904 (72.0%) | 1071 (59.4%) | 1094 (64.5%) |
| ***First harmful event*** | | | | | |
| Angle collisions | 116 (10.1%) | 348 (8.53%) | 291 (7.22%) | 213 (11.8%) | 178 (10.5%) |
| Head-on collisions | 233 (20.4%) | 390 (9.56%) | 246 (6.10%) | 142 (7.88%) | 106 (6.25%) |
| Rear-end collisions | 109 (9.54%) | 350 (8.58%) | 358 (8.88%) | 341 (18.9%) | 193 (11.4%) |
| Run-off-road | 605 (52.9%) | 2613 (64.0%) | 2711 (67.2%) | 776 (43.1%) | 1018 (60.0%) |
| Sideswipe collisions | 18 (1.57%) | 85 (2.08%) | 108 (2.68%) | 140 (7.77%) | 52 (3.07%) |
| Turning collisions | 60 (5.25%) | 281 (6.89%) | 303 (7.51%) | 165 (9.16%) | 139 (8.20%) |
| Other | 2 (0.17%) | 13 (0.32%) | 15 (0.37%) | 25 (1.39%) | 10 (0.59%) |
| ***Object struck*** | | | | | |
| Ditch | 22 (1.92%) | 81 (1.99%) | 97 (2.41%) | 36 (2.00%) | 42 (2.48%) |
| Hit fixed object | 240 (21.0%) | 622 (15.2%) | 391 (9.70%) | 172 (9.54%) | 136 (8.02%) |
| Hit median barrier | 32 (2.80%) | 57 (1.40%) | 60 (1.49%) | 18 (1.00%) | 28 (1.65%) |
| Hit tree, shrub, landscaping | 64 (5.60%) | 104 (2.55%) | 79 (1.96%) | 22 (1.22%) | 25 (1.47%) |
| Overturned | 337 (29.5%) | 1766 (43.3%) | 2015 (50.0%) | 506 (28.1%) | 747 (44.0%) |
| NA | 428 (37.4%) | 1379 (33.8%) | 1288 (31.9%) | 1014 (56.3%) | 662 (39.0%) |
| Other | 20 (1.75%) | 71 (1.74%) | 102 (2.53%) | 34 (1.89%) | 56 (3.30%) |
| ***Day*** | | | | | |
| Weekday | 660 (57.7%) | 2229 (54.6%) | 2304 (57.1%) | 1122 (62.3%) | 1025 (60.4%) |
| Weekend | 483 (42.3%) | 1851 (45.4%) | 1728 (42.9%) | 680 (37.7%) | 671 (39.6%) |
| ***Age*** | | | | | |
| Less than 15 | 2 (0.17%) | 38 (0.93%) | 33 (0.82%) | 12 (0.67%) | 9 (0.53%) |
| 16 to 25 | 145 (12.7%) | 615 (15.1%) | 827 (20.5%) | 337 (18.7%) | 364 (21.5%) |
| 26 to 45 | 427 (37.4%) | 1706 (41.8%) | 1633 (40.5%) | 683 (37.9%) | 705 (41.6%) |
| 46 to 65 | 458 (40.1%) | 1407 (34.5%) | 1250 (31.0%) | 506 (28.1%) | 492 (29.0%) |
| Greater than 65 | 111 (9.71%) | 314 (7.70%) | 289 (7.17%) | 264 (14.7%) | 126 (7.43%) |
| ***Ethnicity*** | | | | | |
| White | 913 (79.9%) | 3177 (77.9%) | 3004 (74.5%) | 1146 (63.6%) | 1165 (68.7%) |
| Asian | 8 (0.70%) | 40 (0.98%) | 52 (1.29%) | 29 (1.61%) | 20 (1.18%) |
| Black | 63 (5.51%) | 251 (6.15%) | 297 (7.37%) | 164 (9.10%) | 164 (9.67%) |
| Hispanic | 150 (13.1%) | 568 (13.9%) | 623 (15.5%) | 285 (15.8%) | 299 (17.6%) |
| Other | 9 (0.79%) | 44 (1.08%) | 56 (1.39%) | 178 (9.88%) | 48 (2.83%) |
| ***Gender*** | | | | | |
| Female | 44 (3.85%) | 245 (6.00%) | 283 (7.02%) | 72 (4.00%) | 120 (7.08%) |
| Male | 1099 (96.2%) | 3835 (94.0%) | 3744 (92.9%) | 1596 (88.6%) | 1573 (92.7%) |
| Unknown | 0 (0.00%) | 0 (0.00%) | 5 (0.12%) | 134 (7.44%) | 3 (0.18%) |
| ***Helmet*** | | | | | |
| Not worn | 566 (49.5%) | 1890 (46.3%) | 1535 (38.1%) | 580 (32.2%) | 600 (35.4%) |
| Worn, damaged | 445 (38.9%) | 1286 (31.5%) | 1110 (27.5%) | 228 (12.7%) | 458 (27.0%) |
| Worn, not damaged | 74 (6.47%) | 474 (11.6%) | 833 (20.7%) | 636 (35.3%) | 411 (24.2%) |
| Unknown | 58 (5.07%) | 430 (10.5%) | 554 (13.7%) | 358 (19.9%) | 227 (13.4%) |

## Chapter Summary

This chapter describes the data sources, case selection, and key variables used to analyze motorcycle crashes on rural undivided roads in Texas. It details the filtering process to focus on 12,753 relevant crashes, explains the structure of the dataset, and provides maps and descriptive statistics to highlight trends in crash severity, contributing factors, and demographic characteristics. The analysis shows that unsafe speed, impaired driving, poor lighting, and lack of helmet use were frequent in severe crashes. These insights set the stage for advanced modeling and in-depth exploration of crash patterns in later chapters.

# METHODOLOGICAL FRAMEWORK

**IV.**

In this chapter, the analytical methods that will be used to investigate motorcycle crash severity on rural undivided roadways will be outlined. CCA will be employed to identify patterns and groupings within the categorical crash data. SHapley Additive exPlanations (SHAP) will be used to interpret the influence of key factors on model predictions. LDA will be applied to extract major topics from narrative crash data. Advanced discrete outcome models including multinomial logit, mixed logit, and their extensions will be developed to capture unobserved heterogeneity and quantify the effects of explanatory variables on injury severity.

## Cluster Correspondence Analysis

The CCA method combines dimension reduction with cluster analysis for categorical data, improving cluster convergence compared to earlier methods (Markos et al., 2019). In this approach, a binary super indicator matrix is utilized, where the matrix dimensions are , where representing the number of observations and denoting the total number of categories across all variables. Each observation encodes a single category as one, with the remaining elements set to zero. The matrix is structured as , where is an indicator matrix for the -th categorical variable, and indicates the total categorical variables. Similar to the loading matrices in continuous data analysis, the matrix captures category quantifications across all categorical variables. The equation defines the quantifications for each j-th categorical variable, where has dimensions . Additionally, , an matrix, where reduced space coordinates of observations (rows of ) are represented. , , indicates cluster membership, and , , denotes the matrix of cluster centroids (Chakraborty et al., 2025b, 2025a; Markos et al., 2019; Rahman et al., 2024).

### Cross-Tabulation and Contingency Analysis

To establish the relationship between clusters and categorical variables, a cross-tabulation of the indicator matrix () and membership matrix () is constructed. This results in a K × p\_j contingency matrix, denoted as (Markos et al., 2019). Applying CCA to this contingency matrix (F) determines scaling values that maximize variance between clusters and categories. This process ensures that clusters effectively distinguish observations by optimizing variance across categorical variables, while also enhancing the distribution of categories within each variable.

### Steps in the CCA Algorithm

The CCA algorithm unfolds through several steps.

* Cluster Initialization: The cluster allocation matrix is initialized by randomly assigning objects to clusters. Simultaneously, category quantifications are estimated.
* Coordinate Computation: The coordinates for observations are determined by computing the averaged centered scores using the previously obtained category quantifications. This is represented by the equation: .
* Cluster Update: The K-means clustering algorithm is applied to the matrix, updating the cluster allocation matrix
* Iterative Refinement: The process is repeated, returning to Step 2, where the updated matrix is used for the next iteration.
* Convergence: The algorithm continues iterating until stabilizes, and consequently, and remain unchanged. Once this stability is achieved, the optimal cluster centroid matrix and category quantification matrix are finalized (Chakraborty et al., 2025b, 2025a; Markos et al., 2019; Rahman et al., 2024).

### Biplot Generation and Scaling Adjustments

These matrices, and derived from the clustering process are utilized to generate a biplot that visually represents clusters and category relationships. Due to dimensional disparities, standard CA normalizations may fail to maintain an even distribution of row and column points. To resolve this, a scaling adjustment is recommended, ensuring that the average squared deviation from the origin remains uniform across clusters and categories. This is achieved using a scaling constant (γ), calculated as . The new scaled measures, and , maintain the same average squared deviation from the origin, ensuring consistency for biplot presentations in the analysis (Chakraborty et al., 2025b, 2025a; Markos et al., 2019; Rahman et al., 2024).

## SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) was utilized in this study to examine the influence of key crash-related factors on motorcyclist injury severity on the rural undivided roadways. As a model interpretation approach, SHAP helps reveal the contribution of individual features to a model’s predictions. The SHAP values, adapted from Shapley values, initially formulated for equitable distribution in cooperative game theory, ensure a fair assessment of each feature’s impact within the predictive framework. For a given model , the SHAP value for a feature is computed as (Chakraborty et al., 2025b; Lundberg and Lee, 2017):

|  |  |
| --- | --- |
|  | () |

where:

* is the full set of features,
* is a subset of features excluding ​,
* is the model’s prediction using only the features in SSS,
* is the prediction when is added,
* ​ represents the SHAP value for feature , indicating its marginal contribution to the prediction.

## Latent Dirichlet Allocation

LDA is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei et al., 2003).

LDA assumes the following generative process for each document **w** in a corpus *D*:

1. Choose ∼ Poisson ().
2. Choose ∼ Dir ().
3. For each of the words

Choose a topic ∼ Multinomial ().

Choose a word from p , a multinomial probability conditioned on the topic .

Several simplifying assumptions are made in this basic model, some of which were removed in subsequent sections. First, the dimensionality of the Dirichlet distribution (and thus the dimensionality of the topic variable ) is assumed known and fixed. Second, the word probabilities are parameterized by a matrix where = *p*( = 1| = 1), which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows, and more realistic document length distributions can be used as needed. Furthermore, note that is independent of all the other data generating variables ( and ). It is thus an ancillary variable, and will generally ignore its randomness in the subsequent development (Blei et al., 2003).

A -dimensional Dirichlet random variable can take values in the ()-simplex (a -vector lies in the ()-simplex if ≥ 0, , and has the following probability density on this simplex:

|  |  |
| --- | --- |
|  | () |

where the parameter is a -vector with components > 0, and where, Γ(x) is the Gamma function. Dirichlet is a convenient distribution on the simplex is in the exponential family, has finite dimensional sufficient statistics, and is conjugated to the multinomial distribution. These properties will facilitate the development of inference and parameter estimation algorithms for LDA (Blei et al., 2003). Given the parameters α and β, the joint distribution of a topic mixture Γ(x), a set of topics , and a set of words is given by:

|  |  |
| --- | --- |
|  | () |

A black background with a white circle

AI-generated content may be incorrect.

Figure .. Graphical model representation of LDA. The boxes are “plates” representing replicates (Blei et al., 2003).

where *p*( | ) is simply for the unique such that = 1. Integrating over and summing over , it is obtained the marginal distribution of a document:

|  |  |
| --- | --- |
|  | () |

Finally, taking the product of the marginal probabilities of single documents, it is obtain the probability of a corpus:

|  |  |
| --- | --- |
|  | () |

The LDA model is represented as a probabilistic graphical model in Figure 4.1. As the figures makes clear, there are three levels to the LDA representation. The parameters and are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables are document-level variables, sampled once per document. Finally, the variables and are word-level variables and are sampled once for each word in each document (Blei et al., 2003).

It is important to distinguish LDA from a simple Dirichlet-multinomial clustering model. A classical clustering model would involve a two-level model in which a Dirichlet is sampled once for a corpus, a multinomial clustering variable is selected once for each document in the corpus, and a set of words are selected for the document conditional on the cluster variable. As with many clustering models, such a model restricts a document to being associated with a single topic. LDA, on the other hand, involves three levels, and notably the topic node is sampled repeatedly within the document. Under this model, documents can be associated with multiple topics (Blei et al., 2003).

Structures similar to that shown in Figure 4.1. Graphical model representation of LDA. The boxes are “plates” representing replicates (Blei et al., 2003). are often studied in Bayesian statistical modeling, where they are referred to as hierarchical models (Gelman et al., 1995), or more precisely as conditionally independent hierarchical models (Kass and and Steffey, 1989). Such models are also often referred to as parametric empirical Bayes models, a term that refers not only to a particular model structure, but also to the methods used for estimating parameters in the model (Morris, 1983).

## Random Parameter Logit Model

In this study, injury severity was classified into three categories: KA (fatal and incapacitating injuries), BC (non-incapacitating and possible injuries), and O (no injury). Due to the categorical nature of the outcome variable, a MNL model was initially applied to both the full dataset and yearly subsets to assess the influence of various factors over time. The base model included all variables, with backward elimination used to retain only those significant at a 90% confidence level. As the intercept was found to be insignificant, it was excluded from the final models. Recognizing the limitations of the MNL model in capturing unobserved heterogeneity, more advanced models were developed, including the Mixed Logit (RPL), Correlated Random Parameters Logit (CRPL), RPLHM. Each modeling step progressively allowed for greater flexibility by accounting for parameter correlations and systematic variations, thereby providing a more nuanced understanding of how explanatory variables influence injury severity in motorcycle crashes. Model fit and selection were evaluated using the likelihood ratio test, AIC, and McFadden’s Pseudo R-squared (Jafari et al., 2025b).

### Mixed Logit Modeling Framework

Police-reported crash data, collected at the scene, may omit key variables influencing injury severity, and the effects of observed factors can vary across individuals. This introduces unobserved heterogeneity, variation arising from factors not captured in the dataset, which can lead to biased parameter estimates and misleading conclusions if ignored (Behnood and Mannering, 2017; Das et al., 2023b, 2024b, 2024b, 2025b; Hossain et al., 2024; Jafari et al., 2025a, 2025c). Advanced mixed logit models address this limitation by allowing the effects of explanatory variables to vary across observations, thereby capturing individual-level heterogeneity. While the MNL model is commonly used for analyzing discrete outcomes, it assumes independently and identically distributed error terms and adheres to the independence of irrelevant alternatives (IIA) property, limiting its flexibility. The mixed logit model overcomes these restrictions by incorporating random parameters, enabling more realistic substitution patterns and accommodating panel structures with correlated choices over time (Jafari et al., 2025b).

The goal of this study is to investigate injury severity probability with a logit model with random parameters that can accommodate varying degrees of variability. This method, which is being used more often in empirical research, takes into consideration variances of random factors in addition to mean fluctuations (Washington et al., 2020). This study specifically investigates three possible outcomes when analyzing injury severity in crashes: KA, BC, and O. The modeling technique, which builds on previous research, starts by creating a function to determine the degree of injury (Washington et al., 2020) as outlined by the following equation:

|  |  |  |
| --- | --- | --- |
|  |  | () |

The probability that a motorcycle with the label *j* will sustain an injury with level *i* is represented by in the equation. On the other hand, stands for the variables affecting this degree of severity. These factors are associated with the estimable parameters, and the error term is represented by . The typical multinomial logit model is based on this residual term when it follows an extreme value distribution (Jafari et al., 2025b).

|  |  |  |
| --- | --- | --- |
|  |  | () |

is the likelihood that a motorcycle *j* will experience a certain injury severity level, represented by *i*, in a set of three possible severity outcomes. equation (8) can be written as follows to enable flexible estimate of one or more parameters within the set across various crash scenarios (McFadden and Train, 2000).

|  |  |  |
| --- | --- | --- |
|  |  | () |

The probability density function connected to in this case is represented by the function , where is a set of parameters that define this function. The variances and the mean value are two examples of these factors. The meanings of all other terms stay the same. To allow parameters to vary across observations, represents a set of quantifiable parameters that may differ between crash scenarios due to the possibility of differences in the means and variances of random parameters (Seraneeprakarn et al., 2017; Yu et al., 2020).

|  |  |  |
| --- | --- | --- |
|  |  | () |

where, stands for the average mean parameter estimates over all crashes. represents a vector of attributes that captures heterogeneity in the mean of random parameters and denotes the corresponding vector of estimable parameters. With corresponding estimable parameters labeled as , represents an additional set of observation-specific variables that explain variations in the standard deviation which captures heterogeneity in variances. In addition, the residual term is represented by . This equation is used in models that account for (RPLHM. Additionally, the possible association between random parameters is considered in this approach (Saeed et al., 2019; Washington et al., 2020).

As may be shown below (Se et al., 2021), the correlated random parameters models (CRPL) was investigated in situations when two or more random parameters are identified:

|  |  |  |
| --- | --- | --- |
|  |  | () |

In this case, is the random parameter vector’s mean, is the vector of explanatory variables that affect the mean of , η is the vector of estimable parameters that correspond to , and is the Cholesky matrix that is used to calculate the random parameter vector’s standard deviation. Furthermore, the random distributed term with a zero mean and variance is shown by . Based on both diagonal and off-diagonal elements of the symmetric Cholesky matrix, the standard deviation of the correlated random parameter *r* is defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | () |

where the diagonal element of the Cholesky matrix is denoted by , and the off-diagonal lower triangular matrix components of the random parameter *r* are represented by , , . The mean standard error of the standard deviation, or , can be found if is the standard deviation of observation-specific and is the number of crash observations. According to Fountas et al. (2018), the correlation coefficient between any two random parameters, and , is expressed as follows:

|  |  |  |
| --- | --- | --- |
|  |  | () |

where, is the covariance between the two indicator variables, and , where and stand for the random parameters. It is crucial to remember that the correlation coefficient, not the linear relationship between risk factors, indicates the correlation between the unobserved components recorded by the random parameters (Fountas et al., 2018).

In line with earlier studies (Behnood and Mannering, 2017; Das et al., 2023b, 2024b, 2024b, 2025b; Hossain et al., 2024; Jafari et al., 2025a, 2025c), the analysis’s methodology involved estimating models using simulated maximum likelihood with 1,000 Halton draws. The normal distribution was selected for the random parameters since numerous studies have demonstrated that it offers a better fit than alternative distributions (Damsere-Derry et al., 2019; Li et al., 2021). The study calculated marginal effects to assess the influence of a one-unit change in a particular explanatory variable on the likelihood of an outcome related to injury severity in order to aid in interpretation (Washington et al., 2020). To provide thorough insights, these unique marginal effects were calculated for every observation and then averaged throughout the whole dataset.

### Marginal Effects

The influence of explanatory variables on the probabilities of injury severity outcomes is evaluated using marginal effects. A marginal effect represents the change in the probability of a specific injury severity level resulting from a change in a binary explanatory variable from 0 to 1. It is calculated as follows (Washington et al., 2020):

|  |  |
| --- | --- |
|  | () |

All models are estimated using simulated maximum likelihood with 1,000 Halton draws to ensure greater result stability (Bhat, 2001; Train, 2009). A normal distribution is assumed for the random parameters, as previous studies have demonstrated that it often provides a better fit compared to alternative distributions (Damsere-Derry et al., 2019; Li et al., 2021). Marginal effects were calculated for the best-fitting model in each year as well as for the full dataset. The likelihood ratio test was used to determine the overall optimal model, and marginal effects were subsequently derived from this selected model for both individual years and the combined dataset.

## Chapter Summary

This chapter describes the full range of methodological approaches adopted for analyzing motorcycle crash data on rural undivided roadways. The use of Cluster Correspondence Analysis, SHAP, LDA, and a suite of logit-based models has been explained. The chapter establishes the analytical foundation for integrating structured and narrative data and set the stage for the interpretation of results in subsequent chapters.

# DATA MINING ON TABULAR CRASH DATA

**V.**

This chapter presents the results of a comprehensive analysis of motorcyclist injury severity on rural undivided roads by using structured or tabular crash data. The section begins with variable importance assessment and cluster correspondence analysis, followed by SHAP-based validation and a detailed mixed logit model evaluation. The findings identify key risk factors, show how these factors influence injury severity across distinct crash clusters.

## Cluster Correspondence Analysis

The results section presents a comprehensive variable importance analysis, identifying key factors influencing crash severity. This is followed by detailed CCA results, categorizing crash data into significant groups and a validation process executed by SHAP analysis. The section concludes with key findings that summarize critical insights and propose targeted countermeasures to enhance road safety and reduce crash occurrences.

### Variable Importance

Prior to utilizing the CCA, variable importance analysis utilizing both XGBoost (Chen et al., 2024), was conducted on the dataset with crash severity as the dependent variable (Ashifur Rahman et al., 2022; Chakraborty et al., 2025b; Das et al., 2023a). The dataset, which includes 18 variables related to motorcyclist injury severity on rural undivided roads, was analyzed using these machine-learning models to identify the most influential factors contributing to crash severity. Following this, a ranked list of important variables in descending order was created, as illustrated in Figure 5.1. The variables are displayed on the y-axis, with their corresponding importance scores on the x-axis. The variable importance plot from the XGBoost model shows that Helmet use has the highest impact on predicting crash severity, followed by First Harmful Event (FHE), Crash Speed (CrSpeed), and Contributing Factors (ConFac) each contributing notably to the model’s predictions. Variables like Gender, Weather, and Day had relatively low importance, suggesting they were less influential in the model’s decision-making process.

To identify the most appropriate variables for CCA, an initial set of 18 candidate variables was systematically evaluated. The selection process commenced with an assessment of variable importance derived from a feature importance plot as shown in , in which *Helmet*, *FHE*, *CrSpeed*, *ConFac*, and *ObjStrk* were identified as the most influential predictors. To mitigate issues related to multicollinearity, a Cramér’s V correlation matrix and corresponding correlation values were derived to examine associations among categorical variables. Variables exhibiting substantial intercorrelations, such as *HarmEvent* with *ObjStrk* and *FHE*, or *Intersection* with *TrafficCon*, were examined, and redundant variables were subsequently excluded. Moreover, a ranked cross-correlation analysis was employed to further identify pairs of variables with high dependency, thereby informing the removal of overlapping or less informative features. Following this comprehensive, multi-stage selection procedure, a final set of 12 variables was retained for inclusion in the model: *Helmet*, *CrSpeed*, *ConFac*, *TrafficCon*, *Age*, *RdAlgn*, *Crash Hour*, *LightCon*, *Season*, *FHE*, *Severity*, and *ObjStrk*. This subset was selected to optimize predictive accuracy while minimizing redundancy, thereby ensuring the development of a robust and interpretable modeling framework.

|  |  |
| --- | --- |
| A graph of a number of words  AI-generated content may be incorrect. | |
| * 1. XGBoost | |
| A graph with blue squares  AI-generated content may be incorrect. | A screenshot of a computer  AI-generated content may be incorrect. |
| (b) Cramér’s V | (c) Cramér’s V values for pairwise variables |
| A graph with text and numbers  AI-generated content may be incorrect. | |
| (d) Ranked cross-correlation table | |

Figure .. Variable selection process with (a) XGBoost (b) Cramér’s V and (c) Cramér’s V values for pairwise variables

### Cluster Correspondence Analysis

The CCA method was employed to examine crashes involving motorcyclist on the rural undivided roads using the top 12 variables identified from the variable importance analysis. The ‘clustrd’ package (Markos et al., 2019) in the R software (The R Foundation, 2024) was utilized in this analysis. Table 5.1 presents the clusters formed using the CCA using the K-means algorithm (Ashifur Rahman et al., 2022; Das et al., 2023a), which identifies associations between categorical variables by analyzing their relationships in a contingency table. The method first constructs a table where crash-related attributes, such as road conditions, traffic control, and environmental factors, are cross-tabulated. Then, the relative frequencies of these attributes are calculated, and a statistical transformation is applied to reduce dimensionality while preserving key associations. The data points are projected into a coordinate system where similar crash characteristics appear closer together, forming natural groupings (Mathai et al., 2022).

The results from the clustering analysis in Table 5.1 highlight the unique characteristics of each group of child pedestrian-involved crashes based on their centroids, size, and within-cluster sum of squares (WCSS). A centroid represents the average position of all observations within a cluster for each dimension, while the size indicates the number and percentage of observations in each cluster. The dimensions (Dimension 1 and Dimension 2) are the primary axes that explain variability in the dataset, while dimension 1 explains the most variance and dimension 2 the second most; they help to visualize relationships between categorical variables. The WCSS measures the compactness of each cluster, where a lower WCSS indicates a tighter grouping of data points.

Table .. Centroids and Size of Clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **Size and Percentage** | **Dimension 1** | **Dimension 2** | **Within cluster sum of squares by cluster** |
| Cluster 1 | 3382 (26.5%) | 0.0025 | -0.0005 | 0.0201 |
| Cluster 2 | 3354 (26.3%) | 0.0078 | -0.0031 | 0.0148 |
| Cluster 3 | 2577 (20.2%) | -0.0046 | -0.0010 | 0.0168 |
| Cluster 4 | 2324 (18.2%) | -0.0112 | -0.0009 | 0.0152 |
| Cluster 5 | 1116 (8.8%) | 0.0030 | 0.0149 | 0.0147 |

*Note: Dim. denotes dimensions or axis*

Cluster 1 is the largest group, containing 3,382 crashes (26.5%), and its centroid is located near the origin (Dim. 1 = 0.0025, Dim. 2 = -0.0005), suggesting that it represents common crash circumstances with a broad, moderate distribution of cases (WCSS = 0.0201). Cluster 2 includes 3,354 crashes (26.3%), with its centroid slightly further along Dimension 1 (0.0078) and Dimension 2 (-0.0031), indicating a profile that, while similar in size to Cluster 1, is more distinct in its attribute associations and exhibits a tighter grouping (WCSS = 0.0148). Cluster 3 comprises 2,577 crashes (20.2%) and is shifted negatively along Dimension 1 (-0.0046) and slightly on Dimension 2 (-0.0010), suggesting a different crash pattern with moderate compactness (WCSS = 0.0168). Cluster 4 contains 2,324 crashes (18.2%), with its centroid at (-0.0112, -0.0009), representing another distinct profile positioned further from the origin, and a compact internal structure (WCSS = 0.0152). Cluster 5 is the smallest cluster, with 1,116 crashes (8.8%), but stands out due to its displacement upwards on Dimension 2 (0.0149) and slightly positive on Dimension 1 (0.0030), indicating a unique crash pattern characterized by a tight grouping of cases (WCSS = 0.0147).

The variation in centroid positions and cluster sizes highlights the diverse nature of crash circumstances present in the dataset as shown in Figure 5.2. Larger clusters, such as Clusters 1 and 2, tend to capture more general and widely occurring crash scenarios, while smaller clusters, particularly Cluster 5, reflect more specific, tightly defined crash patterns along the two dimensions. The differences in WCSS values provide further insight into the relative homogeneity of each group, with lower WCSS indicating clusters where crashes share greater similarity in their key characteristics.

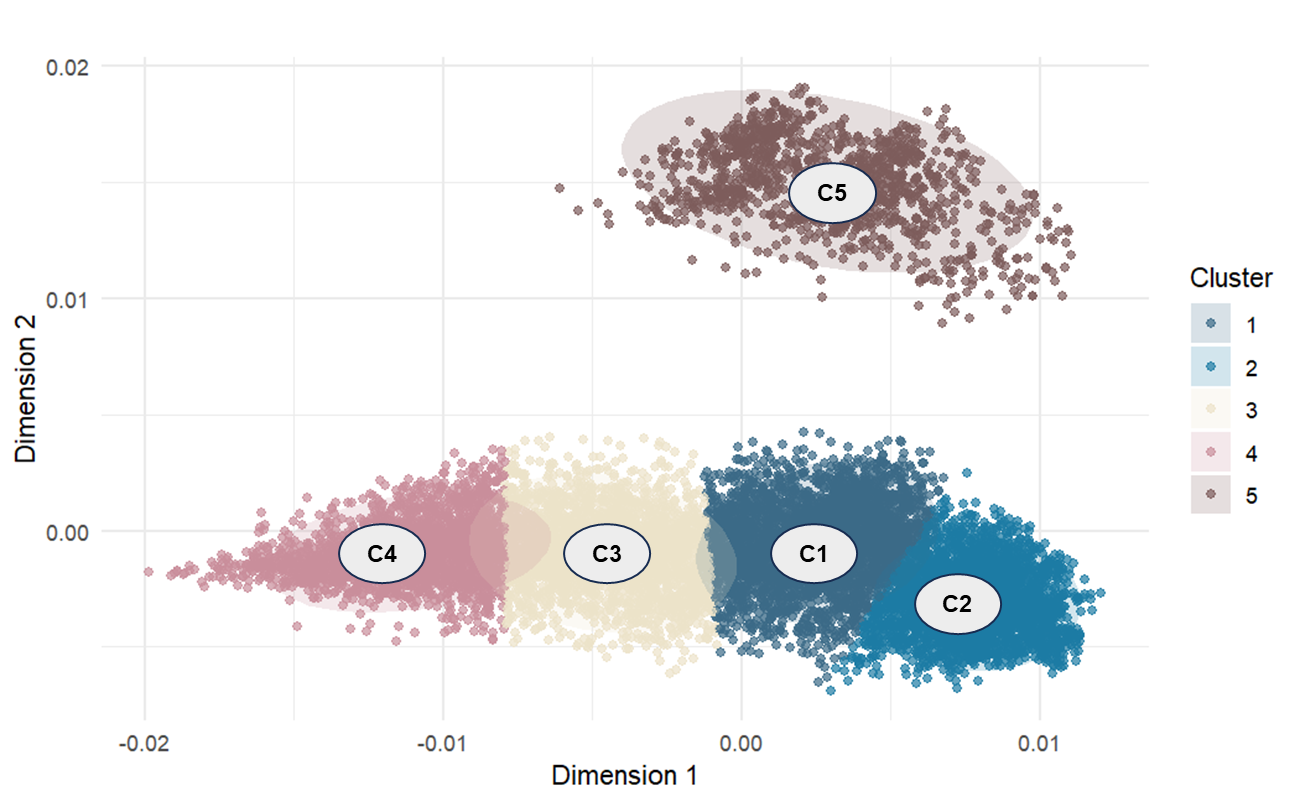
****

Figure .. Clusters produced from motorcyclist injury severity crash data

*Cluster 1 (C1: 26.5%)- Overturns and run-off-road crashes at high speeds*

Most of the crashes in Cluster 1 were associated with overturned vehicles, the presence of other contributing factors, and run-off-road incidents (see Figure 5.3). A significant number of these cases involved situations where the object struck was classified as "overturned," highlighting the severe consequences of loss of vehicle control. The presence of other contributing factors, including environmental and operational variables, further exacerbated injury risks. Run-off-road crashes featured prominently, suggesting that these incidents often occurred on road segments lacking adequate containment or recovery zones. Additional factors contributing to severity included straight, level road alignments and higher crash speeds (greater than 65 mph). This pattern indicates that many severe injuries occur when motorcyclists lose control at high speeds on straight road segments, leading to rollovers or run-off-road events (Effati and Ramezanpoor, 2025). Such circumstances are especially hazardous on rural undivided roads, where the absence of median barriers or rumble strips can fail to prevent or mitigate severe outcomes. A typical scenario for this cluster might involve a motorcyclist traveling at high speed on a rural straightaway, unexpectedly encountering an obstacle or losing control, resulting in a high-severity overturn or run-off-road crash (Das et al., 2025b).

A graph with red lines

AI-generated content may be incorrect.

Figure .. Barplot for Cluster 1

*Cluster 2 (C2: 26.3%) - Crashes with fixed objects on curves at unsafe speeds*

Cluster 2 is characterized by crashes that occurred predominantly on curved, level road sections and involved unsafe speed as a major contributing factor (see Figure 5.4). These crashes frequently resulted in run-off-road outcomes and often involved motorcyclists striking fixed objects, such as roadside barriers or trees. Notably, the cluster includes a substantial presence of "curve, grade" road alignments and incidents of unsafe speed, underscoring the elevated risk posed by excessive speed on rural curves. The marked prevalence of run-off-road crashes and collisions with fixed objects suggests that motorcyclists navigating curves at unsafe speeds face significant challenges in maintaining control, especially when roadside infrastructure is unforgiving. The combination of adverse geometry and speed amplifies the risk of losing traction or veering off the roadway, leading to severe or fatal injuries. In this context, a common scenario would involve a rider approaching a curve too quickly, failing to negotiate the turn, and ultimately striking a fixed object or running off the road (Das et al., 2025b).

A chart with text and numbers

AI-generated content may be incorrect.

Figure .. Barplot for Cluster 2

*Cluster 3 (C3: 20.2%) - Severe Straight Road Crashes with Riding Control Issues*

Crashes in Cluster 3 were dominated by riding control related events and were further marked by the presence of rear-end collisions, head-on collisions, and angle collisions as shown in Figure 5.5. The contributing factors of overturned vehicles and hitting fixed object or trees on the road also played significant roles, with many crashes taking place on curved, level alignments. Severe injury outcomes, including fatal crashes, were prominent in this cluster. The overlap of multiple high-severity crash types and challenging roadway or environmental conditions highlights the compounded risks faced by motorcyclists on rural roads (Arnadottir et al., 2025; Jafari et al., 2025c). A representative scenario would involve a motorcyclist encountering a sudden obstacle, such as an fixed object or trees, in a curved roadway segment, leading to a turning or head-on collision with serious or fatal injuries.

A screenshot of a graph

AI-generated content may be incorrect.

Figure .. Barplot for Cluster 3

*Cluster 4 (C4: 18.2%) - Crashes at intersections under dark, lighted conditions*

Most of the crashes in Cluster 4 in Figure 5.6 involved run-off-road and overturned vehicle scenarios, with unsafe speed emerging as a critical contributing factor. Crashes frequently occurred on curved, level roads and were associated with the presence of angle collisions and incidents at signalized intersections or in the presence of traffic control devices. Head-on collisions were also notable in this group, with non-passing zones and the striking of fixed objects further contributing to injury severity. The interplay of unsafe speed, complex roadway geometry, and insufficient traffic control creates an environment where the likelihood of high-severity crashes is significantly increased (Das et al., 2025a; B. Kutela et al., 2025). An illustrative scenario would involve a rider losing control at excessive speed while navigating a curve in a non-passing zone, resulting in an overturn or collision at an intersection.

A white and blue graph with black text

AI-generated content may be incorrect.

Figure .. Barplot for Cluster 4

*Cluster 5 (C5: 8.8%) - Run-off-road crashes at night with animals present*

Cluster 5, while comprising a smaller share of crashes (see Figure 5.7), is characterized by a concentration of incidents involving dark, unlighted road segments and the presence of animals on the road. Additional contributing factors include unsafe speed, run-off-road events, and failed attempts to yield or signal. The limited visibility, coupled with unexpected obstacles and inadequate traffic controls, increases the risk of severe outcomes for motorcyclists. The interplay of environmental and behavioral factors, such as riding at night without proper lighting or encountering animals, makes crash avoidance and recovery especially challenging (Song et al., 2025). In this cluster, a likely scenario would be a motorcyclist riding at night on a rural road, suddenly facing an animal crossing, and being unable to react in time due to poor visibility and high speed.

A screenshot of a graph

AI-generated content may be incorrect.

Figure .. Barplot for Cluster 5

The analysis of SHAP summary plots in Figure 5.8 across all clusters underscores the pivotal role of unsafe speed, helmet use, and object struck in determining motorcyclist injury severity on rural undivided roads. Unsafe speed, either as a direct contributing factor or as reflected in elevated crash speeds, consistently emerges as the most influential predictor of severe outcomes. This is evident in Clusters 1 and 2, where high speeds precipitate run-off-road incidents, overturns, and collisions with fixed objects, but the trend is also validated in the remaining clusters. Not wearing a helmet or using a damaged helmet substantially increases the risk of fatal or incapacitating injuries in every cluster, reinforcing the critical need for universal helmet use. Likewise, the type of object struck whether overturning, hitting trees, fixed objects, or animals directly amplifies injury severity, as confirmed by consistently high SHAP values and cross-cluster agreement on their impact.

In addition to these core variables, the type of harmful event (such as run-off-road, angle, and head-on crashes), road alignment, and the presence or absence of traffic controls further shape injury severity profiles across clusters. Run-off-road and angle crashes are particularly prominent in Clusters 1 and 4, while Cluster 3 highlights the severe consequences of head-on collisions. Road geometry also plays a defining role, with curved segments increasing risk through loss of control (Cluster 2) and high-speed straight segments compounding crash severity when paired with human or environmental risk factors. These findings, validated by the SHAP plots, indicate that both infrastructure characteristics and traffic management strategies such as improved delineation of curves, enhanced roadside safety treatments, and effective enforcement at high-risk locations are essential for injury prevention.

Finally, environmental and demographic variables add important nuance to the injury severity landscape. Poor lighting and adverse seasonal conditions, especially darkness and winter months, are associated with heightened crash severity, particularly in animal-related incidents (Cluster 5). The timing of crashes, such as afternoon and evening hours, and the involvement of vulnerable age groups both younger and older riders are also linked to more severe outcomes, reflecting patterns of exposure and risk-taking behavior. Collectively, these results highlight the multifactorial nature of motorcyclist injury severity on rural undivided roads and emphasize the value of integrated safety strategies, including speed management, helmet campaigns, targeted infrastructure improvements, and tailored educational programs for at-risk demographics. The strong agreement between cluster-based findings and SHAP interpretations lends further confidence to these recommendations and supports a comprehensive, data-driven approach to rural motorcycle safety.

|  |  |  |  |
| --- | --- | --- | --- |
| A graph of different colored dots  AI-generated content may be incorrect. | A diagram of a graph  AI-generated content may be incorrect. | | A graph of different colored dots  AI-generated content may be incorrect. |
| (a) | (b) | | (c) |
| A diagram of different colored dots  AI-generated content may be incorrect. | | A graph of different colored dots  AI-generated content may be incorrect. | |
| (d) | | (e) | |

Figure .. SHAP summary plot of (a) cluster 1 (b) cluster 2 (c) cluster 3 (d) cluster 4 and (e) cluster 5

## Random Parameter Logit Model

Following the CCA, the injury severity levels in this study were combined into three categories: fatal and incapacitating injuries (KA), non-incapacitating injuries (BC), and possible injuries (O). Based on this classification, a dataset comprising 12,753 motorcycle crashes across five clusters was prepared for the random parameter logit model. This approach allowed for the modeling of unobserved heterogeneity in injury outcomes. The process of selecting variables and attributes for further analysis was guided by the factors most strongly associated with each cluster. Through this targeted selection, a total of 20 key variables were identified for inclusion in the subsequent random parameter logit modeling. Five separate datasets were then prepared, corresponding to each cluster, to facilitate a cluster-based random parameter logit models analysis. Table 5.2 presents the distribution of injury severity levels among motorcyclist crashes on rural undivided roads, segmented by cluster.

Table .. Cluster based injury levels for motorcyclist crashes on rural undivided roads

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Injury Levels** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** | **Total** |
| KA | 1,329 | 1,627 | 1,108 | 697 | 462 | 5,223 |
| BC | 1,714 | 1,502 | 1,048 | 911 | 553 | 5,728 |
| O | 339 | 225 | 421 | 716 | 101 | 1,802 |
| Total | 3,382 | 3,354 | 2,577 | 2,324 | 1,116 | 12,753 |

In the following section, the effects of key variables on motorcycle injury severity will be explained in detail, including how their impacts differ across clusters. This discussion will show how the influence of factors such as helmet use, daylight, and specific crash types varies according to other crash conditions, offering a more nuanced understanding of injury risk on rural undivided highways.

### Rider Behavior

***Helmet not worn (1 if not worn, 0 otherwise) [KA]***

The variable “Helmet not worn” significantly influenced crash severity outcomes across several clusters in the analysis, particularly highlighting the increased probability of KA injuries. In Cluster-1 (RPLHM), this variable had a substantial positive coefficient of 1.583, strongly indicating that motorcyclists not wearing helmets experienced notably higher crash severity (see Table 5.3). Similar positive relationships were observed in Cluster-2 (RPLHM) and Cluster-3 (RPLHM), with coefficients of 0.865 and 0.644 respectively, consistently reinforcing the finding that not wearing helmets elevates severe injury risks. In Cluster-5 (RPL), this relationship persisted, albeit with a slightly lower coefficient of 0.581.

The marginal effect analysis further clarified these relationships. In Cluster-1, helmet non-use increased the likelihood of KA injuries by 0.0625, while decreasing BC injuries by 0.0536 and other O injuries by 0.0089 (see Table 5.4). A similar but stronger pattern emerged in Cluster-2, with helmet non-use raising KA injury probabilities by 0.0710, alongside reductions of 0.0589 for BC injuries and 0.0121 for O injuries. In Cluster-3, helmet non-use increased KA injuries by 0.0373, with corresponding reductions in probabilities of moderate (0.0290) and minor injuries (0.0083). Lastly, Cluster-5 showed helmet non-use increased the chance of KA injuries by 0.0261, decreasing BC injuries by 0.0207 and O injuries by 0.0054.

Helmets play the direct protective role helmets play in reducing severe crash outcomes for motorcyclists. The significant and consistent influence of helmet non-use across multiple clusters implies that, in the absence of helmets, motorcyclists are less equipped to withstand crashes, particularly on rural undivided highways, where higher travel speeds, limited emergency response times, and challenging road conditions amplify injury risks. These findings are in line with the findings of Kang et al. (2021).

***Helmet worn, damaged (1 if damaged, 0 otherwise) [BC]***

The analysis revealed that the variable “Helmet worn, damaged” had a significant influence on crash severity outcomes, presenting distinct relationships across multiple clusters. In Cluster-1 (RPLHM), a negative coefficient of -0.343 indicates that motorcyclists wearing helmets, even if damaged, generally experienced lower severity injuries (see Table 5.3). This protective trend was further supported by stronger negative coefficients in Cluster-2 (RPLHM, -0.486) and Cluster-3 (RPLHM, -0.526). However, in contrast, Cluster-4 (MNL) exhibited a positive coefficient of 0.233, implying scenarios where helmet damage corresponded to more severe impacts. Marginal effect analysis as shown in Figure 5.9 provided additional clarity regarding these severity patterns (see Table 5.4). In Cluster-1, damaged helmets slightly increased the probability of KA injuries by 0.0096, reduced BC injuries by 0.0141, and increased O injuries by 0.0045. Cluster-2 showed an increased likelihood of severe injuries (0.0240), alongside reductions in moderate injuries (-0.0294) and slight increases in minor injuries (0.0054). Similarly, Cluster-3 revealed a minor increase in KA injuries (0.0104), reduced BC injuries by 0.0164, and slightly increased O injuries (0.0060). Conversely, Cluster-4 indicated a decrease in KA injury probability (-0.0260), a notable rise in BC injuries (0.0536), and a reduction in O injuries (-0.0276). These patterns suggest that helmets substantially reduce crash severity, though helmet damage often indicates high-impact events.

A graph of clustering data

AI-generated content may be incorrect.

Figure .. Marginal effects of the variable helmet worn but damaged

***Unsafe speed (1 if rider riding with an unsafe speed, 0 otherwise) [O]***

The analysis identified "Unsafe speed" as a significant factor influencing crash severity with varying effects across clusters. Negative coefficients in Cluster-1 (-0.502) and Cluster-2 (-0.262) indicated that unsafe speeds were generally associated with less severe injury outcomes, potentially reflecting riders' evasive maneuvers or controlled falls during crashes (Jafari et al., 2025c). Conversely, Cluster-4 showed a positive coefficient (0.641), suggesting scenarios where unsafe speeds directly increased injury severity (see Table 5.3). Marginal effects revealed slight increases in KA injuries in Clusters 1 and 2 (0.0062 and 0.0066, respectively) and minor reductions in O injuries, whereas Cluster-4 showed significant reductions in severe (KA, -0.0590) and moderate (BC, -0.0760) injuries, but a pronounced increase in minor injuries (O, 0.1350) (see Table 5.4).

### Rider Demographic

***Rider age (1 if rider's age is greater than 65 years old, 0 otherwise) [O]***

The variable “Rider age (greater than 65 years old)” was notably associated with increased crash severity, as indicated by positive coefficients in Cluster-1 (RPLHM: 0.780), Cluster-3 (RPLHM: 0.965), and Cluster-4 (MNL: 0.696) (see Table 5.3). These coefficients highlight that older motorcyclists generally faced higher crash severity outcomes compared to younger riders, potentially due to age-related vulnerability, slower reaction times, or reduced physical resilience (Jafari et al., 2025a). Marginal effect analysis provided further insights, revealing that in Cluster-1, the probability of O injuries increased by 0.0090, while probabilities for KA and BC injuries decreased slightly (-0.0048 and -0.0042, respectively) (see Table 5.4). Cluster-3 showed a similar trend, with minor injuries increasing (0.0131) and severe and moderate injuries declining (-0.0079 and -0.0052, respectively). In Cluster-4, this pattern was pronounced, showing a notable increase in minor injuries (0.1465) alongside considerable reductions in severe (-0.0641) and moderate injuries (-0.0824). These findings suggest that older riders involved in crashes on rural undivided highways might adopt more cautious riding behaviors, thus experiencing less severe injuries but a higher frequency of minor injuries. Enhancing targeted safety measures, rider training, and age-specific awareness programs can significantly benefit this vulnerable group.

### Crash Characteristics

***Crash speed 30-45 mph (1 if speed between 30 to 45 mph, 0 otherwise) [O]***

The variable “Crash speed 30–45 mph” demonstrated varying associations with injury severity across the clusters, as indicated by negative coefficients in Cluster-1 (RPLHM: -0.330), Cluster-3 (RPLHM: -0.329), and Cluster-5 (RPL: -0.538), and a positive coefficient in Cluster-4 (MNL: 0.279) (see Table 5.3). These results suggest that, in most clusters, crashes occurring at moderate speeds were linked to lower injury severity among motorcyclists, possibly because higher speeds above this range tend to produce more severe impacts, while moderate speeds may allow for greater rider control or more effective collision avoidance. Marginal effect analysis supports this interpretation; for example, in Cluster-1, crash speeds of 30–45 mph slightly increased the probability of KA and BC injuries (0.0048 and 0.0046, respectively) (see Table 5.4), while decreasing the likelihood of O injuries by 0.0094. Cluster-3 reflected a similar pattern, with increases in KA (0.0075) and BC (0.0061) injuries and a decrease in minor injuries (-0.0136). In contrast, Cluster-4 showed reductions in severe (-0.0257) and moderate (-0.0331) injuries, but a pronounced increase in minor injuries (0.0588), indicating that crashes in this speed range within this cluster were more likely to result in less severe outcomes. Cluster-5 also reflected a protective effect, with minor increases in severe and moderate injuries (0.0038 and 0.0091) and a reduction in minor injuries (-0.0129).

***Crash speed greater than 65 mph(1 if speed greater than 65 mph, 0 otherwise) [KA]***

The variable “Crash speed greater than 65 mph” was strongly associated with increased crash severity across all clusters, as shown by consistently positive coefficients in Cluster-1 (RPLHM: 0.718), Cluster-2 (RPLHM: 0.347), Cluster-3 (RPLHM: 1.035), Cluster-4 (MNL: 0.933), and Cluster-5 (RPL: 0.693) (see Table 5.3). These results indicate that high-speed crashes significantly elevate the risk of severe and fatal outcomes for motorcyclists on rural undivided highways, likely due to the greater impact forces and reduced chances for successful evasive actions at such speeds (Jafari et al., 2025a). Marginal effect analysis as shown in Figure 5.10 further emphasizes this relationship: in Cluster-1, a crash speed above 65 mph increased the probability of KA injuries by 0.0232, while decreasing the likelihood of BC and O injuries by -0.0154 and -0.0078, respectively (see Table 5.4). Similar trends were observed in Cluster-3 (KA: 0.0233; BC: -0.0123; O: -0.0110) and Cluster-4, where the increase in severe injuries (0.1899) was particularly pronounced, accompanied by substantial decreases in moderate (-0.1040) and minor injuries (-0.0859). Cluster-5 also reflected an elevated risk for severe injuries (KA: 0.0266), with declines in BC (-0.0208) and O (-0.0058) injuries.

A graph with numbers and a bar chart

AI-generated content may be incorrect.

Figure .. Marginal effects of the variable crash speed greater than 65 mph

### Traffic Condition

***Center stripe or divider (1 if the center stripe/ divider is present, 0 otherwise) [BC]***

The variable “Center stripe or divider present” was found to be associated with injury severity outcomes only in Cluster-5 (RPL), as indicated by a positive coefficient (0.283) (see Table 5.3). This suggests that the presence of a center stripe or divider may influence crash dynamics, potentially by altering vehicle trajectories or limiting certain types of collisions on rural undivided highways. Marginal effect analysis in Cluster-5 revealed that having a center stripe or divider slightly decreased the probability of KA injuries by -0.0120 and O injuries by -0.0062, while increasing the likelihood of BC injuries by 0.0182 (see Table 5.4). These findings imply that while center stripes or dividers may offer some protection against the most severe and minor injuries, they could be associated with a greater share of moderate injuries.

***Marked lane (1 if the marked lane is present, 0 otherwise) [O]***

The variable “Marked lane present” was associated with reduced injury severity across multiple clusters, as evidenced by negative coefficients in Cluster-1 (RPLHM: -0.643), Cluster-2 (RPLHM: -0.509), and Cluster-3 (RPLHM: -0.455) (see Table 5.3). These findings suggest that the presence of marked lanes may contribute to lower crash severity among motorcyclists on rural undivided highways, possibly by providing clearer roadway guidance and improving traffic organization (Jafari et al., 2025c). The marginal effects as shown in Figure 5.11 for stop, yield, or warning signs, which may co-occur with marked lanes, showed that in Cluster-2, the presence of such signage slightly increased the probability of KA injuries by 0.0049, but decreased the likelihood of BC and O injuries by -0.0038 and -0.0011, respectively (see Table 5.4). In Cluster-4, signage presence led to a more notable increase in severe injury probability (KA: 0.0569) and decreases in both moderate (-0.0312) and minor (-0.0257) injuries.

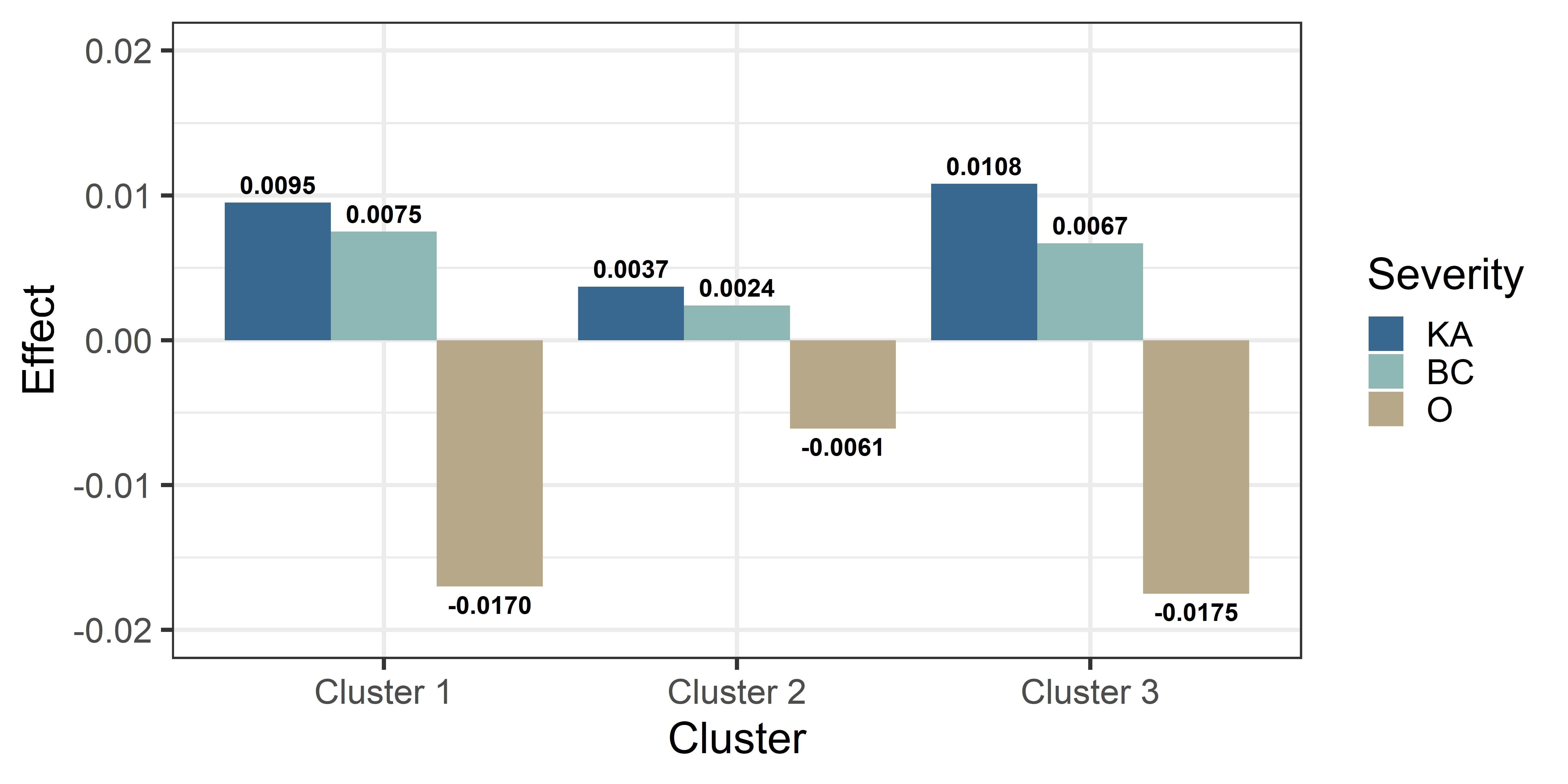
******

Figure .. Marginal effects of the variable marked lane

***Stop or yield or warning sign (1 if stop/yield/warning signs are present, 0 otherwise) [KA]***

The variable “Stop or yield or warning sign present” was associated with increased crash severity in certain clusters, as reflected by positive coefficients in Cluster-2 (RPLHM: 0.249) and Cluster-4 (MNL: 0.279) (see Table 5.3). These results suggest that the presence of regulatory or warning signs at crash locations may be linked to elevated injury severity for motorcyclists, potentially due to abrupt maneuvers, last-moment stops, or misjudgment of road conditions at these sites. Marginal effect analysis further revealed that in Cluster-2, the presence of such signage slightly increased the probability of KA injuries by 0.0049, while reducing BC and O injuries by -0.0038 and -0.0011, respectively (see Table 5.4). In Cluster-4, this pattern was even more pronounced, with a notable increase in the probability of severe injuries (KA: 0.0569) and reductions in moderate (-0.0312) and minor (-0.0257) injuries.

### Roadway Characteristics

***Curve grade (1 if the roadway alignment is curve and graded, 0 otherwise) [O]***

The variable “Curve grade” was associated with crash severity in opposite directions across clusters, as reflected by a positive coefficient in Cluster-1 (RPLHM: 0.582) and a negative coefficient in Cluster-2 (RPLHM: -0.584) (see Table 5.3). These results suggest that curved and graded road segments may be linked to elevated injury severity for motorcyclists in some settings possibly due to increased difficulty in handling and a higher risk of losing control while in other contexts, curves may encourage lower speeds and more cautious riding, thus reducing injury severity (Ye et al., 2025). Marginal effect analysis as shown in Figure 5.12 further revealed that in Cluster-1, crashes on curves resulted in very small decreases in the probabilities of severe (KA: -0.0015) and moderate (BC: -0.0017) injuries, but a slight increase in minor injuries (O: 0.0032) (see Table 5.4). In Cluster-2, the pattern was reversed, with curves associated with a slight increase in severe (KA: 0.0054) and moderate injuries (BC: 0.0036), and a reduction in minor injuries (O: -0.0090).

A graph showing a number of clusters

AI-generated content may be incorrect.

Figure .. Marginal effects of the variable curve graded roadway

***Straight level (1 if the roadway alignment is straight and leveled, 0 otherwise) [KA]***

The variable “Straight level” was associated with increased crash severity in Cluster-1 (RPLHM: 0.549), indicating that crashes on straight and level road segments may result in more severe injuries for motorcyclists (see Table 5.3). This could be related to higher typical travel speeds or overconfidence leading to riskier riding behaviors on straightaways. Marginal effect analysis showed that the presence of straight and level alignment in Cluster-1 substantially increased the probability of severe injuries (KA: 0.0559) (see Table 5.4), while reducing the likelihood of moderate (BC: -0.0383) and minor (O: -0.0176) injuries. These findings are found in line with the study of (Jafari et al., 2025a).

### Temporal Factors

***Afternoon (1 if crash happened in the afternoon, 0 otherwise) [O]***

The variable “Afternoon” was associated with reduced crash severity across all clusters, as indicated by consistently negative coefficients in Cluster-1 (RPLHM: -0.373), Cluster-2 (RPLHM: -0.370), and Cluster-3 (RPLHM: -0.393) (see Table 5.3). This suggests that motorcycle crashes occurring in the afternoon are generally linked to less severe injuries, potentially due to higher visibility and more stable traffic conditions during this time of day (Jafari et al., 2025a). Marginal effect analysis as shown in Figure 5.13 further revealed that in each of these clusters (see Table 5.4), afternoon crashes slightly increased the probability of KA and BC injuries by small margins (KA: 0.0040 to 0.0067; BC: 0.0023 to 0.0041), but more notably decreased the likelihood of O injuries (from -0.0057 to -0.0108).

A graph showing a number of clusters

AI-generated content may be incorrect.

Figure .. Marginal effects of the variable afternoon

***Morning (1 if crash happened in the morning, 0 otherwise) [KA]***

The variable “Morning” was linked to increased crash severity in Cluster-3 (RPLHM: 0.573) (see Table 5.3), indicating that motorcycle crashes occurring in the morning are more likely to result in severe injuries. This may be related to factors such as early-morning fatigue, lower light levels, or heightened rush-hour traffic. Marginal effect analysis for Cluster-3 showed that morning crashes increased the probability of KA injuries by 0.0203, while decreasing the likelihood of moderate (BC: -0.0099) and minor (O: -0.0104) injuries (see Table 5.4).

***Fall (1 if crash happened in fall, 0 otherwise) [KA]***

The variable “Fall” was associated with higher crash severity in Cluster-3 (RPLHM: 0.427) (see Table 5.3), suggesting that motorcycle crashes occurring in the fall season are more likely to lead to severe injuries. This may be due to environmental factors such as slippery roads from wet leaves, reduced daylight, or changing weather conditions that impact rider safety (Jafari et al., 2025a). Marginal effect analysis in Cluster-3 indicated that fall crashes increased the probability of KA injuries by 0.0175, while decreasing the likelihood of moderate (BC: -0.0088) and minor (O: -0.0087) injuries (see Table 5.4).

### Lighting Condition

***Dark, not lighted (1 if crash happened in dark, not lighted condition, 0 otherwise) [O]***

The variable “Dark, not lighted” was consistently associated with reduced crash severity across all clusters, as indicated by negative coefficients in Cluster-1 (RPLHM: -0.237), Cluster-2 (RPLHM: -0.498), Cluster-3 (RPLHM: -0.415), Cluster-4 (MNL: -0.545), and Cluster-5 (RPL: -1.298) (see Table 5.3). These results suggest that crashes occurring in unlit dark conditions tend to result in less severe injuries for motorcyclists, which may be due to lower vehicle speeds, heightened caution, or fewer vehicles present during nighttime hours. Marginal effect analysis further revealed that, across clusters, dark and unlit conditions slightly increased the probability of severe (KA: up to 0.0502) and moderate (BC: up to 0.0646) injuries in some clusters, but more notably and consistently decreased the probability of minor injuries (O: as much as -0.1148 in Cluster-4) (see Table 5.4). These findings are in line with the study by Ye et al. (2025).

***Daylight (1 if crash happened in daylight, 0 otherwise) [BC]***

The variable “Daylight” was associated with increased crash severity, particularly for moderate injuries, across all clusters, as reflected by positive coefficients in Cluster-1 (RPLHM: 0.951), Cluster-2 (RPLHM: 1.320), Cluster-3 (RPLHM: 0.454), Cluster-4 (MNL: 0.505), and Cluster-5 (RPL: 0.671) (see Table 5.3). This indicates that motorcycle crashes occurring during daylight hours are more likely to result in BC injuries, potentially due to higher exposure, increased traffic, or faster speeds during the day. Marginal effect analysis showed in Figure 5.14 imply that daylight conditions decreased the probability of severe (KA: from -0.0313 to -0.0632) and minor (O: from -0.0034 to -0.0598) injuries across clusters (see Table 5.4), while consistently increasing the probability of moderate injuries (BC: from 0.0358 to 0.1160). These findings are in line with the study by Ye et al. (2025).

A graph with numbers and a bar chart

AI-generated content may be incorrect.

Figure .. Marginal effects of the variable daylight

### First Harmful Event

***Rear-end (1 if the crash type is rear-end, 0 otherwise) [BC]***

The variable “Rear-end” was associated with increased moderate injury severity in Cluster-3 (RPLHM: 0.642) (see Table 5.3), indicating that motorcyclists involved in rear-end collisions are more likely to sustain moderate injuries compared to other crash types. Marginal effect analysis for Cluster-3 showed a modest increase in the probability of moderate (BC: 0.0236) injuries, with minor (O: -0.0055) and severe (KA: -0.0181) injuries being less likely (see Table 5.4). These findings suggest that rear-end collisions on rural undivided highways are often less catastrophic for motorcyclists than other crash scenarios, yet they still pose a significant risk for non-fatal injury, emphasizing the need for safe following distances and greater awareness of stopping behaviors.

***Run-off-road (ROR) (1 if the crash type is ROR, 0 otherwise) [KA]***

The variable “Run-off-road” demonstrated notable variability in its relationship with crash severity. In Cluster-1 (RPLHM: -0.392) and Cluster-3 (RPLHM: -1.868) (see Table 5.3), negative coefficients suggest that ROR events may be associated with reduced probabilities of severe injuries, potentially reflecting scenarios where riders are able to avoid high-impact crashes by leaving the roadway. However, in Cluster-2 (RPLHM: 0.818), the positive coefficient indicates higher severity, perhaps due to hazardous roadside features or loss of control at high speeds. Marginal effect analysis in Cluster-1 showed a reduction in the probability of severe (KA: -0.0481) injuries, with slight increases in moderate (BC: 0.0319) and minor (O: 0.0162) injuries (see Table 5.4). Conversely, Cluster-2 showed an increase in KA (0.1660), and decreases in BC (-0.1317) and O (-0.0343).

***Hit fixed object (1 if the crash happened involving hitting fixed object, 0 otherwise) [BC]***

Crashes involving a fixed object were associated with slightly lower injury severity in Cluster-3 (RPLHM: -0.885) (see Table 5.3), with marginal effect analysis indicating a very small increase in the probability of severe (KA: 0.0022) injuries, a decrease in moderate (BC: -0.0042) injuries, and a slight increase in minor (O: 0.0020) injuries (see Table 5.4). These results may reflect instances where fixed object collisions are less likely to be fatal for motorcyclists than anticipated, possibly due to reduced impact speeds or the type of roadside object struck. However, such events still pose significant risks, supporting continued efforts to remove or shield hazardous roadside features along rural highways.

***Overturned (1 if crash happened involving overturning, 0 otherwise) [BC]***

The variable “Overturned” showed strong positive associations with crash severity in Cluster-1 (RPLHM: 1.078), Cluster-2 (RPLHM: 1.454), and Cluster-5 (RPL: 0.481), while Cluster-4 (MNL: -0.956) had a negative association (see Table 5.3). These results indicate that overturning crashes are typically linked to a higher likelihood of moderate and severe injuries for motorcyclists (Jafari et al., 2025c). Marginal effect analysis in Figure 5.15 revealed that in Clusters 1 and 2, overturning substantially increased the probability of moderate (BC: 0.1018 and 0.1750) injuries, while reducing the probability of minor (O: -0.0307 and -0.0261) and severe injuries (KA: -0.0711 and -0.1489) (see Table 5.4). In Cluster-4, the effect was reversed, with an increase in the probability of severe injuries (KA: 0.1065) and minor increases in moderate (BC: -0.2197) and minor (O: 0.1132) injuries.

A graph of clustering data

AI-generated content may be incorrect.

Figure .. Marginal effects of the variable overturned

### Random parameter and Heterogeneity in means

The variables “Helmet not worn,” “Daylight,” and “Rear-end” are each modeled as random parameters in specific clusters, capturing significant heterogeneity in their effects on motorcycle injury severity across rural undivided highway crashes. For helmet non-use, the estimated means and standard deviations are 2.158 (3.56), 2.068 (3.75), and 3.550 (1.75) in Clusters 1, 3, and 5, respectively, with positive effects observed for 76.8%, 62.2%, and 95.6% of the distributions indicating that helmet non-use increases injury severity for most but not all riders (see Table 5.3). The daylight variable, with means and standard deviations of 2.847 (3.98), 1.123 (1.89), and 3.483 (3.53) in Clusters 1, 2, and 3, shows positive effects in 80.2%, 76.7%, and 79.5% of cases, suggesting that daylight generally elevates the likelihood of moderate injuries, yet a notable minority of crashes may not follow this pattern. Likewise, the random parameter for rear-end crashes in Cluster 3 (mean 0.642, std. dev. 2.292) is positive in about 61.0% of the distribution, reflecting that while rear-end collisions typically raise injury severity, considerable variation exists across incidents.

Heterogeneity in means explains how the impact of key risk factors like “helmet not worn” and “daylight” on injury severity can shift depending on other crash circumstances. In this analysis, the presence of unsafe speed reduces the average effect of daylight (–0.528) and helmet not worn (–0.994) (see Table 5.3), suggesting that when riders are speeding, the added risk from poor visibility or helmet non-use becomes less influential likely because speeding itself dominates the risk profile. Likewise, when the roadway is curve-leveled, the effect of daylight (–0.976) and helmet not worn (–0.768) also decreases, indicating that the complex geometry of curves is a stronger determinant of crash outcomes than either helmet use or time of day. Conversely, the absence of traffic control devices increases the mean effect of daylight (0.465), meaning daylight matters more for injury outcomes when intersections or signs are missing, likely because drivers and riders rely more on visibility to navigate these uncontrolled locations. Overturning (–1.185) and hitting fixed objects (–0.690) both reduce the effect of daylight, reflecting that these severe crash types are so hazardous that the benefit of daylight is diminished. Notably, when a no passing zone is present, the mean effect of helmet not worn increases (1.173), showing that helmet use is especially critical on these high-risk segments, while the effect of daylight decreases (–0.924), possibly because safe maneuvering is already restricted.

Table .. Model Estimation Results from Cluster 1 to Cluster 5

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **Cluster-1 (RPLHM)** | | **Cluster-2 (RPLHM)** | | **Cluster -3 (RPLHM)** | | **Cluster-4**  **(MNL)** | | **Cluster-5**  **(RPL)** | |
| **Coeff.** | **t-stat** | **Coeff.** | **t-stat** | **Coeff.** | **t-stat** | **Coeff.** | **t-stat** | **Coeff.** | **t-stat** |
| **Rider Behavior** | | | | | | | | | | |
| Helmet not worn (1 if not worn, 0 otherwise) [KA] | 1.583 | 8.12 | 0.865 | 8.14 | 0.644 | 4.33 | - | - | 0.581 | 2.41 |
| *Standard deviation of the random parameter helmet not worn* | 2.158 | 3.56 | - | - | 2.068 | 3.75 | - | - | 3.550 | 1.75 |
| Helmet worn, damaged (1 if damaged, 0 otherwise) [BC] | -0.343 | -2.75 | -0.486 | -4.63 | -0.526 | -3.23 | 0.233 | 2.35 | - | - |
| Unsafe speed (1 if rider riding with an unsafe speed, 0 otherwise) [O] | -0.502 | -3.56 | -0.262 | -2.24 | - | - | 0.641 | 3.79 | - | - |
| **Rider Demographic** | | | | | | | | | | |
| Rider age (1 if rider's age is greater than 65 years old, 0 otherwise) [O] | 0.780 | 4.32 | - | - | 0.965 | 4.72 | 0.696 | 4.42 | - | - |
| **Crash Speed** | | | | | | | | | | |
| Crash speed 30-45 mph (1 if speed between 30 to 45 mph, 0 otherwise) [O] | -0.330 | -2.59 | - | - | -0.329 | -2.73 | 0.279 | 4.08 | -0.538 | -2.63 |
| Crash speed greater than 65 mph (1 if speed greater than 65 mph, 0 otherwise) [KA] | 0.718 | 5.55 | 0.347 | 2.53 | 1.035 | 6.23 | 0.933 | 5.60 | 0.693 | 3.9 |
| **Traffic Condition** | | | | | | | | | | |
| Center stripe or divider (1 if the center stripe/ divider is present, 0 otherwise) [BC] | - | - | - | - | - | - | - | - | 0.283 | 2.01 |
| Marked lane (1 if the marked lane is present, 0 otherwise) [O] | -0.643 | -5.13 | -0.509 | -2.97 | -0.455 | -3.84 | - | - | - | - |
| Stop or yield or warning sign (1 if stop/yield/warning signs are present, 0 otherwise) [KA] | - | - | 0.249 | 1.79 |  |  | 0.279 | 2.91 | - | - |
| **Roadway Characteristics** | | | | | | | | | | |
| Curve grade (1 if the roadway alignment is curved and graded, 0 otherwise) [O] | 0.582 | 2.3 | -0.584 | -3.77 | - | - | - | - | - | - |
| Straight level (1 if the roadway alignment is straight and leveled, 0 otherwise) [KA] | 0.549 | 4.52 | - | - | - | - | - | - | - | - |
| **Temporal Factors** | | | | | | | | | | |
| Afternoon (1 if crash happened in the afternoon, 0 otherwise) [O] | -0.373 | -2.65 | -0.370 | -2.39 | -0.393 | -2.78 | - | - | - | - |
| Morning (1 if crash happened in the morning, 0 otherwise) [KA] | - | - | - | - | 0.573 | 4.01 | - | - | - | - |
| Fall (1 if crash happened in fall, 0 otherwise) [KA] | - | - | - | - | 0.427 | 3.35 | - | - | - | - |
| **Lighting Condition** | | | | | | | | | | |
| Dark, not lighted (1 if crash happened in dark, not lighted condition, 0 otherwise) [O] | -0.237 | -1.74 | -0.498 | -3.03 | -0.415 | -2.81 | -0.545 | -3.7 | -1.298 | -7.79 |
| Daylight (1 if crash happened in daylight, 0 otherwise) [BC] | 0.951 | 5.85 | 1.320 | 7.82 | 0.454 | 2.45 | 0.505 | 7.96 | 0.671 | 5.3 |
| *Standard deviation of the random parameter daylight* | 2.847 | 3.98 | 1.123 | 1.89 | 3.483 | 3.53 | - | - | - | - |
| **First Harmful Event** | | | | | | | | | | |
| Rear-end (1 if the crash type is rear-end, 0 otherwise) [BC] | - | - | - | - | 0.642 | 2.79 | - | - | - | - |
| *Standard deviation of the random parameter rear-end* | - | - | - | - | 2.292 | 1.82 | - | - | - | - |
| Run-off-road (ROR) (1 if the crash type is ROR, 0 otherwise) [KA] | -0.392 | -3.4 | 0.818 | 8.04 | -1.868 | -9.07 | - | - | - | - |
| Hit fixed object (1 if the crash happened involving hitting fixed object, 0 otherwise) [BC] | - | - | - | - | -0.885 | -2.48 | - | - | - | - |
| Overturned (1 if crash happened involving overturning, 0 otherwise) [BC] | 1.078 | 8.78 | 1.454 | 10.71 | - | - | -0.956 | -2.26 | 0.481 | 3.71 |
| **Heterogeneity in mean, Parameter:Variable** | | | | | | | | | | |
| Effect of unsafe speed on the mean of random parameter daylight | -0.528 | -2.35 | - | - | - | - | - | - | - | - |
| Effect of curve-leveled roadway alignment on the mean of random parameter daylight | -0.976 | -2.88 | - | - | - | - | - | - | - | - |
| Effect of unsafe speed on the mean of random parameter helmet not worn | -0.994 | -3.81 | - | - | - | - | - | - | - | - |
| Effect of curve-leveled roadway alignment on the mean of random parameter helmet not worn | -0.768 | -1.92 | - | - | - | - | - | - | - | - |
| Effect of absent of traffic control devices on the mean of random parameter daylight | - | - | 0.465 | 2.85 | - | - | - | - | - | - |
| Effect of overturning on the mean of random parameter daylight | - | - | -1.185 | -5.74 | - | - | - | - | - | - |
| Effect of hitting fixed object on the mean of random parameter daylight | - | - | -0.690 | -3.24 | - | - | - | - | - | - |
| Effect of no passing zone on the mean of random parameter hemet not worn | - | - | - | - | 1.173 | 2.69 | - | - | - | - |
| Effect of no passing zone on the mean of random parameter daylight | - | - | - | - | -0.924 | -1.97 | - | - | - | - |
| **Statistics** | | | | | | | | | | |
| Number of observartions | 3382 | | 3354 | | 2577 | | 2324 | | 1116 | |
| K | 20 | | 16 | | 19 | | 9 | | 8 | |
| Log likelihood at convergence | -3111.53308 | | -2892.93482 | | -2538.99162 | | -2474.44496 | | -1047.8101 | |
| Restricted log likelihood | -3715.50676 | | -3684.74562 | | -2831.12387 | | -2553.17496 | | -1226.05131 | |
| McFadden Pseudo R-squared | 0.1625549 | | 0.2148889 | | 0.103186 | | 0.0309 | | 0.1453783 | |
| AIC | 6263.1 | | 5817.9 | | 5116 | | 4966.9 | | 2111.6 | |
| AIC/N | 1.852 | | 1.735 | | 1.985 | | 2.137 | | 1.892 | |

Table .. Marginal Effect Results from Cluster 1 to Cluster 5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **Injury Levels** | **Cluster-1 (RPLHM)** | **Cluster-2 (RPLHM)** | **Cluster -3 (RPLHM)** | **Cluster-4 (MNL)** | **Cluster-5 (RPL)** |
| **Rider Behavior** | | | | | | |
| Helmet not worn (1 if not worn, 0 otherwise) [KA] | KA | 0.0625 | 0.0710 | 0.0373 | - | 0.0261 |
| BC | -0.0536 | -0.0589 | -0.0290 | - | -0.0207 |
| O | -0.0089 | -0.0121 | -0.0083 | - | -0.0054 |
| Helmet worn, damaged (1 if damaged, 0 otherwise) [BC] | KA | 0.0096 | 0.0240 | 0.0104 | -0.0260 | - |
| BC | -0.0141 | -0.0294 | -0.0164 | 0.0536 | - |
| O | 0.0045 | 0.0054 | 0.0060 | -0.0276 | - |
| Unsafe speed (1 if rider riding with an unsafe speed, 0 otherwise) [O] | KA | 0.0062 | 0.0066 | - | -0.0590 | - |
| BC | 0.0052 | 0.0045 | - | -0.0760 | - |
| O | -0.0114 | -0.0111 | - | 0.1350 | - |
| **Rider Demographic** | | | | | | |
| Rider age (1 if rider's age is greater than 65 years old, 0 otherwise) [O] | KA | -0.0048 | - | -0.0079 | -0.0641 | - |
| BC | -0.0042 | - | -0.0052 | -0.0824 | - |
| O | 0.0090 | - | 0.0131 | 0.1465 | - |
| **Crash Speed** | | | | | | |
| Crash speed 30-45 mph (1 if speed between 30 to 45 mph, 0 otherwise) [O] | KA | 0.0048 | - | 0.0075 | -0.0257 | 0.0038 |
| BC | 0.0046 | - | 0.0061 | -0.0331 | 0.0091 |
| O | -0.0094 | - | -0.0136 | 0.0588 | -0.0129 |
| Crash speed greater than 65 mph(1 if speed greater than 65 mph, 0 otherwise) [KA] | KA | 0.0232 | 0.0074 | 0.0233 | 0.1899 | 0.0266 |
| BC | -0.0154 | -0.0060 | -0.0123 | -0.1040 | -0.0208 |
| O | -0.0078 | -0.0014 | -0.0110 | -0.0859 | -0.0058 |
| **Traffic Condition** | | | | | | |
| Center stripe or divider (1 if the center stripe/ divider is present, 0 otherwise) [BC] | KA | - | - | - | - | -0.0120 |
| BC | - | - | - | - | 0.0182 |
| O | - | - | - | - | -0.0062 |
| Marked lane (1 if the marked lane is present, 0 therwise) [O] | KA | 0.0095 | 0.0037 | 0.0108 | - | - |
| BC | 0.0075 | 0.0024 | 0.0067 | - | - |
| O | -0.0170 | -0.0061 | -0.0175 | - | - |
| Stop or yield or warning sign (1 if stop/yield/warning signs are present, 0 otherwise) [KA] | KA | - | 0.0049 | - | 0.0569 | - |
| BC | - | -0.0038 | - | -0.0312 | - |
| O | - | -0.0011 | - | -0.0257 | - |
| **Roadway Characteristics** | | | | | | |
| Curve grade (1 if the roadway alignment is curve and graded, 0 otherwise) [O] | KA | -0.0015 | 0.0054 | - | - | - |
| BC | -0.0017 | 0.0036 | - | - | - |
| O | 0.0032 | -0.0090 | - | - | - |
| Straight level (1 if the roadway alignment is straight and leveled, 0 otherwise) [KA] | KA | 0.0559 | - | - | - | - |
| BC | -0.0383 | - | - | - | - |
| O | -0.0176 | - | - | - | - |
| **Temporal Factors** | | | | | | |
| Afternoon (1 if crash happened in the afternoon, 0 otherwise) [O] | KA | 0.0040 | 0.0034 | 0.0067 | - | - |
| BC | 0.0034 | 0.0023 | 0.0041 | - | - |
| O | -0.0074 | -0.0057 | -0.0108 | - | - |
| Morning (1 if crash happened in the morning, 0 otherwise) [KA] | KA | - | - | 0.0203 | - | - |
| BC | - | - | -0.0099 | - | - |
| O | - | - | -0.0104 | - | - |
| Fall (1 if crash happened in fall, 0 otherwise) [KA] | KA | - | - | 0.0175 | - | - |
| BC | - | - | -0.0088 | - | - |
| O | - | - | -0.0087 | - | - |
| **Lighting Condition** | | | | | | |
| Dark, not lighted (1 if crash happened in dark, not lighted condition, 0 otherwise) [O] | KA | 0.0024 | 0.0044 | 0.0050 | 0.0502 | 0.0191 |
| BC | 0.0031 | 0.0022 | 0.0039 | 0.0646 | 0.0302 |
| O | -0.0055 | -0.0066 | -0.0089 | -0.1148 | -0.0493 |
| Daylight (1 if crash happened in daylight, 0 otherwise) [BC] | KA | -0.0324 | -0.0632 | -0.0408 | -0.0562 | -0.0313 |
| BC | 0.0358 | 0.0750 | 0.0537 | 0.1160 | 0.0577 |
| O | -0.0034 | -0.0118 | -0.0129 | -0.0598 | -0.0264 |
| **First Harmful Event** | | | | | | |
| Rear-end (1 if the crash type is rear-end, 0 otherwise) [BC] | KA | - | - | -0.0181 | - | - |
| BC | - | - | 0.0236 | - | - |
| O | - | - | -0.0055 | - | - |
| Run-off-road (ROR) (1 if the crash type is ROR, 0 otherwise) [KA] | KA | -0.0481 | 0.1660 | -0.0292 | - | - |
| BC | 0.0319 | -0.1317 | 0.0133 | - | - |
| O | 0.0162 | -0.0343 | 0.0159 | - | - |
| Hit fixed object (1 if the crash happened involving hitting fixed object, 0 otherwise) [BC] | KA | - | - | 0.0022 | - | - |
| BC | - | - | -0.0042 | - | - |
| O | - | - | 0.0020 | - | - |
| Overturned (1 if crash happened involving overturning, 0 otherwise) [BC] | KA | -0.0711 | -0.1489 | - | 0.1065 | -0.0272 |
| BC | 0.1018 | 0.1750 | - | -0.2197 | 0.0415 |
| O | -0.0307 | -0.0261 | - | 0.1132 | -0.0143 |

## Chapter Summary

The results reveal that helmet use, crash speed, and specific crash types had the strongest influence on injury severity. Cluster analysis uncovers distinct crash patterns, which were further validated by SHAP interpretation. The mixed logit models demonstrate how the effects of key variables vary across clusters, providing a detailed understanding of risk and guiding practical countermeasures for improving motorcycle safety on rural roads

# ADVANCED TOPIC MODELING ON CRASH NARRATIVES

**VI.**

To further verify and enrich the insights drawn from the tabular crash data, a text analysis was performed using NLP techniques. By applying bigram-based LDA topic modeling to the crash narrative texts, the analysis revealed thematic patterns such as speeding, roadway departures, intersection conflicts, and lane-change maneuvers that were consistent with the trends identified in the quantitative dataset. This integration of structured and unstructured data allowed for a more comprehensive understanding of the factors contributing to crash risk. The process of generating bigrams and extracting dominant topics from narrative descriptions provided valuable context that supported and validated the key variables and crash patterns captured in the tabular results, reinforcing the overall conclusions of the study.

## LDA Topic Modeling Results

The application of LDA to the corpus of crash narratives facilitated the identification of distinct thematic structures underlying the data. Each topic, as revealed by the most influential bigrams, represents a unique crash scenario or set of contributing factors, offering insight into the complex mechanisms driving roadway incidents.

***Topic 1: Speeding-Related Roadway Departure***

The first topic is strongly characterized by bigrams such as “left side”, “unsafe speed”, “right side” and “lost control” which collectively point to crashes precipitated by excessive speed and subsequent loss of vehicle control (see Figure 6.1). Additional terms such as “barrow ditch”, “final rest” and “left curve” suggest a recurrent pattern where vehicles, unable to negotiate curves or lane boundaries, leave the roadway and come to rest in roadside ditches or medians. This topic underscores the critical role of speed management and lateral control in preventing run-off-road crashes, especially on segments with challenging geometry.

A graph of a bar graph

AI-generated content may be incorrect.

Figure .. Bigrams for topic 1

***Topic 2: Single-Vehicle Roadway Departure***

This topic predominantly highlights scenarios involving single-vehicle departures from the roadway (see Figure 6.2), as evidenced by bigrams like “loss of control”, “right side”, “unsafe speed”, and “improved shoulder”. Other key terms, including “hand curve”, “single lane”, and “grassy ditch” indicate that these incidents frequently occur on narrow, curvilinear road sections with limited recovery space. The narratives represented by this topic emphasize the compounded risk posed by road geometry and insufficient speed adaptation, leading to single-vehicle run-off-road events.

A graph of a bar graph

AI-generated content may be incorrect.

Figure .. Bigrams for topic 2

***Topic 3: Turn-Related Crash at Intersection***

The third topic is distinguished by bigrams such as “left turn”, “stop sign”, “turn left”, “turn onto”, and “evasive action” all of which signal intersection-related crashes, particularly those involving turning maneuvers (see Figure 6.3). Bigrams like “yield right”, “turn lane” and “private drive” further point to the prominence of right-of-way violations, improper driveway entries or exits, and left-turn conflicts. This topic highlights the ongoing challenges of intersection safety, with failures to yield and misjudged maneuvers emerging as frequent crash precursors.

A graph of a bar chart

AI-generated content may be incorrect.

Figure .. Bigrams for topic 3

***Topic 4: Run-Off-Road Crashes on County Roads***

In the fourth topic, the dominance of bigrams such as “right side”, “lost control”, “county road” and “wire fence” identifies run-off-road events on rural or county roadways as a distinct narrative theme (see Figure 6.4). The presence of terms like “barbed wire”, “minor damage”, and “minor injuries” suggests that while these crashes often result in relatively minor injury outcomes, property damage from fencing and roadside infrastructure is a common consequence. This topic underscores the heightened vulnerability of drivers on rural roads, where loss of control events often interact with roadside features.

A graph of a bar chart

AI-generated content may be incorrect.

Figure .. Bigrams for topic 4

***Topic 5: Lane-Change and Speed Control Conflicts***

The fifth topic is defined by bigrams such as “left lane”, “right lane”, “private drive”, “control speed”, “changed lanes” and “lost control” all of which are indicative of lane-change-related conflicts and speed mismanagement (see Figure 6.5). The inclusion of “inside lane”, “outside lane”, and “center median” highlights the spatial complexity of these incidents, where crashes frequently occur during lane transitions, particularly in multi-lane environments or near driveway access points.

A graph of a bar chart

AI-generated content may be incorrect.

Figure .. Bigrams for topic 5

The LDA-derived topics provide a comprehensive view of the principal mechanisms underlying crash causation in the studied dataset. The recurrent themes of speeding, loss of control, complex roadway geometry, intersection conflicts, and lane-change maneuvers demonstrate the multifaceted nature of roadway safety challenges. These findings illustrate the value of topic modeling in uncovering nuanced, context-specific patterns within narrative crash data, thereby informing the development of targeted interventions such as enhanced speed enforcement, improved signage and roadway design, intersection safety treatments, and measures to reduce run-off-road and lane-change-related crashes. By elucidating the latent structure of crash narratives, this approach supports data-driven policymaking and prioritization of roadway safety initiatives.

## Chapter Summary

This chapter combines variable-based modeling and LDA topic modeling to analyze factors contributing to crash severity. The key results from tabular and text data were consistent, with both highlighting speeding, roadway departures, and intersection-related risks. Integrating NLP findings with structured data offered strong validation and a more complete understanding of crash mechanisms.

# CONVERGENT VALIDITY

**VII.**

This section brings together the main findings from the cluster analysis, marginal effects modeling, and narrative topic modeling to provide a comprehensive view of motorcycle crash severity on rural undivided roadways. Key risk factors including speed, loss of control, roadway geometry, poor lighting, and helmet use are highlighted, and targeted policy recommendations are proposed using the Safe System Approach. Table 7.1is presented at the end of the chapter to summarize the clusters, model results, and narrative themes, reinforcing the core patterns and priority interventions discussed here.

Table .. Comparative Table: Clusters, Model Results, and Narrative Themes

|  |  |  |
| --- | --- | --- |
| **Cluster Findings** | **Marginal Effects** | **LDA Topics** |
| **C1: Overturns and run-off-road crashes at high speeds**  Overturned vehicles, run-off-road events, high speeds, other contributing factors, straight level roads. | **Helmet not worn [KA]:** 0.0625 ↑  **Crash speed >65 mph [KA]:** 0.0232 ↑  **Run-off-road [KA]:** -0.0481 ↓  **Straight, level [KA]:** 0.0559 ↑  **Overturned [BC]:** 0.1018 ↑  **Marked lane [O]:** -0.0170 ↓  **Afternoon [O]:** -0.0074 ↓  **Daylight [BC]:** 0.0358 ↑ | **Topic 1:** Speeding-related roadway departure  **Topic 2:** Single-vehicle roadway departure |
| **C2: Crashes with fixed objects on curves at unsafe speeds**  Curve, grade alignments, high speed, run-off-road into fixed objects, unsafe speed major factor. | **Helmet not worn [KA]:** 0.0710 ↑  **Crash speed >65 mph [KA]:** 0.0074 ↑  **Run-off-road [KA]:** 0.1660 ↑  **Curve, grade [KA]:** 0.0054 ↑  **Marked lane [O]:** -0.0061 ↓  **Overturned [BC]:** 0.1750 ↑  **Afternoon [O]:** -0.0057 ↓  **Daylight [BC]:** 0.0750 ↑ | **Topic 2:** Single-Vehicle Roadway Departure  **Topic 1:** Speeding-related roadway departure |
| **C3: Severe straight road crashes with riding control issues**  Straight road segments, fixed object, trees or obstacles, mix of crash types related to turning issues. | **Helmet not worn [KA]:** 0.0373 ↑  **Crash speed >65 mph [KA]:** 0.0233 ↑  **Run-off-road [KA]:** -0.0292 ↓  **Morning [KA]:** 0.0203 ↑  **Rear-end [BC]:** 0.0236 ↑  **Daylight [BC]:** 0.0537 ↑  **Dark, not lighted [KA]:** 0.0050 ↑ | **Topic 3:** Turn-Related Crash at intersection  **Topic 5:** Lane-change and speed control conflicts |
| **C4: Crashes at intersections under dark, lighted conditions**  Unsafe speed, signalized intersections, angle/head-on, complex geometry, non-passing zones, fixed objects. | **Helmet worn, damaged [KA]:** -0.0260 ↓  **Crash speed >65 mph [KA]:** 0.1899 ↑  **Unsafe speed [KA]:** -0.0590 ↓  **Crash speed 30–45 mph [KA]:** -0.0257 ↓  **Rider age >65 [KA]:** -0.0641 ↓  **Overturned [KA]:** 0.1065 ↑  **Daylight [BC]:** 0.1160 ↑  **Dark, not lighted [KA]:** 0.0502 ↑ | **Topic 3:** Turn-related crash at intersection |
| **C5: Run-off-road crashes at night with animals present**  Small cluster, dark roads, animals, run-off-road, unsafe speed, failed yielding/signaling. | **Helmet not worn [KA]:** 0.0261 ↑  **Crash speed >65 mph [KA]:** 0.0266 ↑  **Daylight [BC]:** 0.0577 ↑  **Dark, not lighted [KA]:** 0.0191 ↑  **Center divider [BC]:** 0.0182 ↑ | **Topic 4:** Run-off-road crashes on county roads |

## Synthesis of Analytical Approaches and Policy Recommendations

Across all analytical approaches cluster analysis, marginal effects modeling, and narrative topic modeling, excessive speed and loss of control on rural undivided roadways emerge as the leading contributors to severe motorcycle crashes. Clusters representing the most severe outcomes (Clusters 1, 2, and 3) consistently highlight scenarios involving high speeds, run-off-road events, overturns, and collisions with fixed roadside objects or obstacles. These findings are reinforced by the random parameter logit models, which identify helmet non-use, speeds over 65 mph, and certain road alignments (straight, level, or curved with poor geometry) as major risk factors for fatal or serious injury. LDA topic modeling adds further insights, with common themes centered on “unsafe speed,” “loss of control,” and “roadway departure,” confirming that the mechanisms of high-speed, uncontained travel are central to the most critical crash scenarios.

Intersection-related crashes, particularly those involving turning maneuvers (Cluster 4), present another significant area of risk. Both quantitative models and topic modeling reveal that these crashes frequently occur under dark or poorly lit conditions at rural intersections, where improper yielding, signal violations, and misjudged turns lead to severe angle and head-on collisions. The challenges are compounded by the limited presence of channelization, insufficient lighting, and absence of advanced warning systems that are common on rural undivided roads.

For Cluster 5, crashes that occur at night and involve animals highlight the unique vulnerability of rural roadways. These environments are often characterized by inadequate lighting, frequent wildlife crossings, and limited visibility, which, when combined with high speeds and failed attempts to yield or signal, substantially increase the risk of fatal run-off-road incidents. Both the model coefficients and the topics extracted from narrative data confirm that poor visibility and unexpected animal encounters are especially dangerous for motorcyclists in rural areas.

## Implications for Rural Road Safety

The convergence of findings from all three methods underscores a multifaceted risk environment on rural undivided highways. High speeds, limited road separation, minimal recovery zones, insufficient lighting, and challenging roadway geometry all combine to produce conditions where even minor errors or lapses in judgment can lead to catastrophic outcomes. The use of helmets especially undamaged, properly fitted ones remains a critical mitigating factor. The following sections will provide rural-focused policy interventions integrating Safe System Approach (USDOT, 2023).

### Safe Speeds

To address the central issue of excessive speed, speed limits on rural undivided roads should be based on detailed, context-sensitive assessments that take into account curve radii, shoulder widths, and roadside hazards (Gross et al., 2009). Speed enforcement strategies such as mobile speed cameras and high-visibility patrols should target identified high-risk segments, particularly those frequently involved in run-off-road and curve-related crashes (NHTSA, 2024a). Geometric improvements, like the installation of rumble strips and dynamic warning signs, can further prompt riders to reduce speed in hazardous zones (Liu, 2015).

### Safe Roads

Rural undivided roads would benefit from upgrades to roadside safety infrastructure. This includes expanding clear zones, installing energy-absorbing barriers, and replacing rigid roadside objects with breakaway supports (Dobrovolny et al., 2021). Enhanced pavement markings, edge lines, and reflective signage can improve both day and night visibility (FHWA, 2024). At intersections, treatments such as roundabouts, protected turning lanes, and better lighting should be prioritized, especially where analysis identifies frequent conflicts and high injury severity.

### Safe Road Users

Targeted education programs tailored for rural motorcyclists are essential. These should emphasize hazard perception, animal avoidance techniques, the dangers of speeding, and the importance of always wearing a helmet. Community campaigns can help shift cultural attitudes toward risk-taking and encourage adoption of safe riding behaviors. Helmet rebate programs and roadside safety checks may also improve compliance with helmet laws (NHTSA, 2024b).

### Safe Vehicles and Post-Crash Response

Incentivizing the use of advanced protective gear, such as high-visibility clothing and motorcycles equipped with ABS, could help reduce injury severity. Additionally, improvements to post-crash response such as installing emergency call boxes in areas with poor cellular coverage and enhancing EMS response times are vital, particularly in rural settings where immediate help is often not available.

### Environmental and Wildlife Management

Coordination with wildlife agencies to implement fencing, animal detection systems, and wildlife crossing signage can address the persistent issue of animal-related crashes (Grace et al., 2017). These interventions should be targeted based on the location data from cluster and topic analyses to maximize their effectiveness.

## Chapter Summary

By integrating cluster analysis, advanced modeling, and crash narrative insights, this study demonstrates that the most severe motorcycle crashes on rural undivided highways are a product of high speeds, poor geometry, limited containment, and behavioral risk factors. A Safe System Approach, incorporating engineering, enforcement, education, and emergency response applicable to rural roadways, offers the most comprehensive strategy for reducing fatal and serious motorcycle injuries in these challenging environments.

# CONCLUSION

**VIII.**

The persistent challenge of motorcycle crash severity on rural undivided roadways has been recognized as a significant safety concern due to disproportionately high rates of fatal and incapacitating injuries. This issue has drawn increasing attention because these environments combine high operating speeds, roadway geometries, limited traffic separation, and minimal safety infrastructure factors that collectively elevate crash risks for motorcyclists. In this research, the scope of the study was systematically investigated using comprehensive crash data from the Texas CRIS between 2017 and 2023. The dataset included both structured variables and narrative text fields, encompassing around 12,753 motorcycle crashes filtered specifically for rural undivided road settings.

Structured data were analyzed through CCA to uncover major crash typologies, followed by cluster-based RPLHM to explore heterogeneous effects of risk factors on injury severity. Simultaneously, unstructured crash narratives were examined using LDA topic modeling, allowing the extraction of prevalent themes and underlying crash scenarios not otherwise captured in coded variables. This multi-method approach was intended to ensure a more complete and validated understanding of the factors contributing to motorcycle crash severity in rural settings.

Key findings revealed that severe and fatal outcomes on rural undivided roadways were most strongly associated with excessive speeds (especially over 65 mph), run-off-road incidents, overturns, and impacts with fixed roadside objects. Helmet non-use, nighttime or poor lighting conditions, and hazardous roadway features such as curves and grades consistently amplified injury risks across multiple analytical approaches. Intersection-related and turning crashes were also identified as critical scenarios, particularly where improper yielding, inadequate signalization, or darkness further increased the likelihood of severe injuries. The thematic analysis of crash narratives reinforced the statistical results, with frequent mentions of speeding, loss of control, failure to yield, and animal encounters in rural environments. Together, the synthesis of cluster patterns, model parameters, and narrative themes emphasized the need for targeted, context-specific countermeasures.

This research demonstrated that the integration of statistical modeling, machine learning, and natural language processing provides a comprehensive and nuanced framework for identifying, prioritizing, and addressing the principal risk factors contributing to severe motorcycle crashes on rural undivided roadways. The research outcomes highlight practical directions for transportation policy, infrastructure design, and safety education tailored to rural environments, supporting more effective interventions and safer outcomes for vulnerable road users.

## Limitations of the Research

This study is subject to several important limitations. First, the mixed logit models were estimated using variables selected through cluster analysis and variable importance methods, which means that some potentially relevant variables were excluded from the random parameter analysis. This selective approach may limit the overall robustness and comprehensiveness of the model results. Second, the research did not provide a detailed examination of barrier types or roadside safety features, despite their established importance in influencing crash severity and preventing roadway departures. The omission of a thorough barrier analysis represents a critical gap in the current study. Finally, the quality and consistency of crash narrative data in the CRIS database varied considerably, which may have impacted the reliability of the topic modeling and limited the ability to capture all relevant crash factors in the analysis.

## Future Scope of Research

Future research should focus on the application of these analytical approaches within urban roadway environments, where the complexity of traffic operations, higher vehicle densities, and mixed road user interactions present unique safety challenges. There is considerable value in expanding the scope to systematically compare single-vehicle and multi-vehicle motorcycle crashes, as the mechanisms, risk factors, and severity outcomes may differ substantially across these contexts. Additionally, in-depth studies on the role of physical barriers including medians, guardrails, and crash cushions are warranted to better understand their effectiveness in mitigating crash severity and preventing roadway departures in diverse settings.

# REFERENCES

Anastasopoulos, P.Ch., Mannering, F.L., 2011. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. Accident Analysis & Prevention 43, 1140–1147. https://doi.org/10.1016/j.aap.2010.12.024

Arnadottir, A.T., Ulfarsson, G.F., Kim, S., 2025. Single-vehicle injury crashes on rural roads in Iceland: contribution of unforgiving roadsides to fatal and serious injuries of vehicle occupants. Traffic Safety Research 9, e000083–e000083. https://doi.org/10.55329/bkrk9175

Ashifur Rahman, M., Das, S., Sun, X., 2022. Using Cluster Correspondence Analysis to Explore Rainy Weather Crashes in Louisiana. Transportation Research Record 2676, 159–173. https://doi.org/10.1177/03611981221082582

Behnood, A., Mannering, F., 2017. The effect of passengers on driver-injury severities in single-vehicle crashes: A random parameters heterogeneity-in-means approach. Analytic Methods in Accident Research. https://doi.org/10.1016/j.amar.2017.04.001

Bhat, C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B: Methodological 35, 677–693. https://doi.org/10.1016/S0191-2615(00)00014-X

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research.

Carvalho, H.B.D., Andreuccetti, G., Rezende, M.R., Bernini, C., Silva, J.S., Leyton, V., D’Andréa Greve, J.M., 2016. Alcohol and drug involvement in motorcycle driver injuries in the city of Sao Paulo, Brazil: Analysis of crash culpability and other associated factors. Drug and Alcohol Dependence 162, 199–205. https://doi.org/10.1016/j.drugalcdep.2016.03.007

Chakraborty, R., Das, S., Mimi, M.S., Kutela, B., 2025a. Investigating Factor Associations in Barrier Crashes through Cluster Correspondence Analysis. Transportation Research Record 2679, 860–879. https://doi.org/10.1177/03611981241297976

Chakraborty, R., Das, S., Mimi, M.S., Kutela, B., 2024. Investigating Factor Associations in Barrier Crashes through Cluster Correspondence Analysis. Transportation Research Record: Journal of the Transportation Research Board. https://doi.org/10.1177/03611981241297976

Chakraborty, R., Mills, D., Das, S., 2025b. Children on wheels: Identifying crash determinants using cluster correspondence analysis. Accident Analysis & Prevention 216. https://doi.org/10.1016/j.aap.2025.108025

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., implementation), Xgb. contributors (base Xgb., 2024. xgboost: Extreme Gradient Boosting.

Cheng, A.S.K., Ng, T.C.K., Lee, H.C., 2011. A comparison of the hazard perception ability of accident-involved and accident-free motorcycle riders. Accident Analysis & Prevention 43, 1464–1471. https://doi.org/10.1016/j.aap.2011.02.024

Christian, J.M., Thomas, R.F., Scarbecz, M., 2014. The Incidence and Pattern of Maxillofacial Injuries in Helmeted Versus Non-Helmeted Motorcycle Accident Patients. Journal of Oral and Maxillofacial Surgery 72, 2503–2506. https://doi.org/10.1016/j.joms.2014.07.015

Damsere-Derry, J., Ebel, B.E., Mock, C.N., Afukaar, F., Donkor, P., Kalowole, T.O., 2019. Evaluation of the effectiveness of traffic calming measures on vehicle speeds and pedestrian injury severity in Ghana. Traffic Injury Prevention. https://doi.org/10.1080/15389588.2019.1581925

Das, S., 2021a. Identifying key patterns in motorcycle crashes: findings from taxicab correspondence analysis. Transportmetrica A: Transport Science 17, 593–614. https://doi.org/10.1080/23249935.2020.1802362

Das, S., 2021b. Identifying key patterns in motorcycle crashes: findings from taxicab correspondence analysis. Transportmetrica A: Transport Science 17, 593–614. https://doi.org/10.1080/23249935.2020.1802362

Das, S., Avelar, R., Dixon, K., Sun, X., 2018. Investigation on the wrong way driving crash patterns using multiple correspondence analysis. Accident Analysis & Prevention 111, 43–55. https://doi.org/10.1016/j.aap.2017.11.016

Das, S., Dutta, A., Tsapakis, I., 2021. Topic Models from Crash Narrative Reports of Motorcycle Crash Causation Study. Transportation Research Record: Journal of the Transportation Research Board 2675, pp-449-462. https://doi.org/10.1177/03611981211002523

Das, S., Dzinyela, R., Liu, J., Dadashova, B., Silvestri-Dobrovolny, C., 2024a. Understanding patterns of factor influences in motorcycle crashes with fixed objects. Journal of Transportation Safety & Security 1–27. https://doi.org/10.1080/19439962.2024.2429077

Das, S., Dzinyela ,Richard, Liu ,Jinli, Dadashova ,Bahar, and Silvestri-Dobrovolny, C., 2025a. Understanding patterns of factor influences in motorcycle crashes with fixed objects. Journal of Transportation Safety & Security 17, 483–509. https://doi.org/10.1080/19439962.2024.2429077

Das, S., Jafari, M., Dzinyela, R., Khan, M.N., 2025b. Applying hybrid dimension reduction and econometric model to investigate rider behaviors in roadway departure motorcycle crashes. Transportation Letters 1–21. https://doi.org/10.1080/19427867.2025.2516422

Das, S., Jafari, M., Hossain, A., Chakraborty, R., Mimi, M.S., 2024b. Toll road crash severity using mixed logit model incorporating heterogeneous mean structures. Transportmetrica A: Transport Science 1–21. https://doi.org/10.1080/23249935.2024.2343755

Das, S., Mousavi, S.M., Shirinzad, M., 2022. Pattern recognition in speeding related motorcycle crashes. Journal of Transportation Safety & Security 14, 1121–1138. https://doi.org/10.1080/19439962.2021.1877228

Das, S., Sakib, N., Geedapally, S., Wei, Z., 2025c. Understanding Pedestrian Hit and Run Crash Patterns using Louisiana Data. Transportation Safety and Environment. https://doi.org/10.1093/tse/tdaf039

Das, S., Sun, X., 2014. Exploring Clusters of Contributing Factors for Single-Vehicle Fatal Crashes Through Multiple Correspondence Analysis.

Das, S., Tabesh, M., Dadashova, B., Dobrovolny, C., 2023a. Diagnosis of Encroachment-Related Work-Zone Crashes by Applying Pattern Recognition. Transportation Research Record: Journal of the Transportation Research Board 2677, 222–236. https://doi.org/10.1177/03611981231152254

Das, S., Tamakloe, R., Zubaidi, H., Obaid, I., Ashifur Rahman, M., 2023b. Bicyclist injury severity classification using a random parameter logit model. International Journal of Transportation Science and Technology 12, 1093–1108. https://doi.org/10.1016/j.ijtst.2023.02.001

de Rome, L., Fitzharris, M., Baldock, M., Fernandes, R., Ma, A., Brown, J., 2016. The prevalence of crash risk factors in a population-based study of motorcycle riders. Injury 47, 2025–2033. https://doi.org/10.1016/j.injury.2016.03.033

De Rome, L., Ivers, R., Fitzharris, M., Haworth, N., Heritier, S., Richardson, D., 2012. Effectiveness of motorcycle protective clothing: Riders’ health outcomes in the six months following a crash. Injury 43, 2035–2045. https://doi.org/10.1016/j.injury.2011.10.025

Dobrovolny, C.S., Geary, G., Dixon, K., Manser, M., Chauhan, J., Texas A&M Transportation Institute, 2021. Addressing the Motorcyclist Advisory Council Recommendations: Synthesis on Barrier Design for Motorcyclists Safety (No. FHWA-SA-21-069).

Effati, M., Ramezanpoor, A., 2025. Examining the role of random parameters and unobserved heterogeneity in the frequency-severity of rural freeway run-off-road and fixed-object crashes: A Bayesian hierarchical-geospatial approach. Accident Analysis & Prevention 215, 108005. https://doi.org/10.1016/j.aap.2025.108005

Eluru, N., Bhat, C.R., Hensher, D.A., 2007. A Mixed Generalized Ordered Response Model for Examining Pedestrian and Bicyclist Injury Severity Level in Traffic Crashes. University of Sydney.

FHWA, 2024. Pavement marking safety study | FHWA [WWW Document]. URL https://highways.dot.gov/safety/other/visibility/benefits-pavement-markings-renewed-perspective-based-recent-and-ongoing-5 (accessed 6.18.25).

Fountas, G., Anastasopoulos, P.Ch., Abdel-Aty, M., 2018. Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. Analytic Methods in Accident Research. https://doi.org/10.1016/j.amar.2018.04.003

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9780429258411

Goodwin, A.H., Wang, Y.C., Foss, R.D., Kirley, B., 2022. The role of inexperience in motorcycle crashes: Crash trends of novice and returning motorcycle riders. Journal of Safety Research 82, pp 371-375. https://doi.org/10.1016/j.jsr.2022.07.003

Grace, M.K., Smith, D.J., Noss, R.F., 2017. Reducing the threat of wildlife-vehicle collisions during peak tourism periods using a Roadside Animal Detection System. Accident Analysis & Prevention 109, 55–61. https://doi.org/10.1016/j.aap.2017.10.003

Gross, F., Jovanis, P.P., Eccles, K., Chen, K.-Y., 2009. Safety Evaluation of Lane and Shoulder Width Combinations on Rural, Two-Lane, Undivided Roads - FHWA-HRT 09-031 (No. FHWA-HRT-09-031).

Haque, M.M., Chin, H.C., Debnath, A.K., 2012. An investigation on multi-vehicle motorcycle crashes using log-linear models. Safety science.

Hossain, A., Barua, S., Das, S., Starewich, M., 2025. Ambulance crash risk dynamics: a baseline (2017–2019) vs. pandemic-era (2020–2022) comparative study using a random parameter logit model. Transportmetrica A: Transport Science 1–39. https://doi.org/10.1080/23249935.2025.2481578

Hossain, A., Sun, X., Das, S., Jafari, M., Codjoe, J., 2024. Investigating older driver crashes on high-speed roadway segments: a hybrid approach with extreme gradient boosting and random parameter model. Transportmetrica A: Transport Science 0, 1–35. https://doi.org/10.1080/23249935.2024.2362362

IISH, 2024. Fatality Facts 2022: Motorcycles and ATVs [WWW Document]. IIHS-HLDI crash testing and highway safety. URL https://www.iihs.org/topics/fatality-statistics/detail/motorcycles-and-atvs (accessed 2.24.25).

Islam, M., 2021a. The effect of motorcyclists’ age on injury severities in single-motorcycle crashes with unobserved heterogeneity. Journal of Safety Research 77, 125–138. https://doi.org/10.1016/j.jsr.2021.02.010

Islam, M., 2021b. The effect of motorcyclists’ age on injury severities in single-motorcycle crashes with unobserved heterogeneity. Journal of Safety Research 77, 125–138. https://doi.org/10.1016/j.jsr.2021.02.010

Islam, S., Brown, J., 2017a. A comparative injury severity analysis of motorcycle at-fault crashes on rural and urban roadways in Alabama. Accident Analysis & Prevention 108, pp 163-171. https://doi.org/10.1016/j.aap.2017.08.016

Islam, S., Brown, J., 2017b. A comparative injury severity analysis of motorcycle at-fault crashes on rural and urban roadways in Alabama. Accident Analysis & Prevention 108, pp-163-171. https://doi.org/10.1016/j.aap.2017.08.016

Jafari, M., Das, S., Barua, S., Mimi, M.S., Starewich, M., 2025a. Crash outcomes of yellow school buses: Random parameter and correlated random parameter logit models with heterogeneity in means. Accident Analysis & Prevention 219, 108109. https://doi.org/10.1016/j.aap.2025.108109

Jafari, M., Starewich, M., Hossain, A., Barua, S., Alnawmasi, N., Ye, X., Das, S., 2025b. Assessing motorcyclist injury severity on curved road segments with temporal dynamics and unobserved heterogeneity. Sci Rep 15, 13110. https://doi.org/10.1038/s41598-025-97972-7

Jafari, M., Starewich, M., Hossain, A., Barua, S., Alnawmasi, N., Ye, X., Das, S., 2025c. Assessing motorcyclist injury severity on curved road segments with temporal dynamics and unobserved heterogeneity. Scientific Reports 15, 13110.

Jeon, J., Woo, A., 2024. The effects of built environments on bicycle accidents around bike-sharing program stations using street view images and deep learning techniques: The moderating role of streetscape features. Journal of Transport Geography 121, 103992. https://doi.org/10.1016/j.jtrangeo.2024.103992

Jiang, F., Yuen, K.K.R., Lee, E.W.M., 2020. Analysis of motorcycle accidents using association rule mining-based framework with parameter optimization and GIS technology. Journal of safety research.

Kang, L., Vij, A., Hubbard, A., Shaw, D., 2021. The unintended impact of helmet use on bicyclists’ risk-taking behaviors. Journal of Safety Research 79, 135–147. https://doi.org/10.1016/j.jsr.2021.08.014

Kashani, A.T., Rabieyan, R., Besharati, M.M., 2014. A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. Journal of safety research.

Kass, R.E., and Steffey, D., 1989. Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models). Journal of the American Statistical Association 84, 717–726. https://doi.org/10.1080/01621459.1989.10478825

Khan, G., Bill, A.R., Chitturi, M.V., Noyce, D.A., 2013. Safety Evaluation of Horizontal Curves on Rural Undivided Roads. Transportation Research Record: Journal of the Transportation Research Board 2386, 147–157. https://doi.org/10.3141/2386-17

Khor, D., Inaba, K., Aiolfi, A., Delapena, S., Benjamin, E., Matsushima, K., Strumwasser, A.M., Demetriades, D., 2017. The impact of helmet use on outcomes after a motorcycle crash. Injury 48, 1093–1097. https://doi.org/10.1016/j.injury.2017.02.006

Kim, Y., Yeo, H., Lim, L., Noh, B., 2024. Integrating visual and community environments in a motorcycle crash and casualty estimation. Accident Analysis & Prevention 208. https://doi.org/10.1016/j.aap.2024.107792

Kutela, B., Mihayo, M.P., Khalaf, H.M.M.B., Kidando, E., Kitali, A.E., 2025. Investigating the influence of lighting conditions on pre-crash vulnerable road users’ visibility. Advances in Transportation Studies 65, 103–118. https://doi.org/10.53136/97912218174927-pag.103-118

Kutela, Boniphace, Shita, H., Mbuya, C., Chimba, D., 2025. Prediction of sensor damage in automated vehicle involved collisions using parametric and non-parametric approaches. International Journal of Crashworthiness 30, 337–349. https://doi.org/10.1080/13588265.2024.2406607

Lemonakis, P., Eliou, N., Karakasidis, T., 2021. Investigation of speed and trajectory of motorcycle riders at curved road sections of two-lane rural roads under diverse lighting conditions. Journal of Safety Research 78, 138–145. https://doi.org/10.1016/j.jsr.2021.05.009

Levy, K.E.C., Franklin, M., 2014. Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry. Social Science Computer Review 32, 182–194. https://doi.org/10.1177/0894439313506847

Li, X., Liu, J., Zhang, Z., Parrish, A., Jones, S., 2021. A spatiotemporal analysis of motorcyclist injury severity: Findings from 20 years of crash data from Pennsylvania. Accident Analysis & Prevention.

Liu, C., 2015. Design Directional Raised Rumble Aggregates and Strips for Awakening Wrong-Way Drivers. International Journal of Transportation Science and Technology 4, 151–156. https://doi.org/10.1260/2046-0430.4.2.151

Liu, S., Lin, Z., Fan, W. (David), 2022. Investigating contributing factors to injury severity levels in crashes involving pedestrians and cyclists using latent class clustering analysis and mixed logit models. Journal of Transportation Safety & Security 14, 1674–1701. https://doi.org/10.1080/19439962.2021.1958037

Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. https://doi.org/10.48550/arXiv.1705.07874

Maistros, A., Schneider, W.H., Savolainen, P.T., 2014. A comparison of contributing factors between alcohol related single vehicle motorcycle and car crashes. Journal of Safety Research 49, 129.e1-135. https://doi.org/10.1016/j.jsr.2014.03.002

Markos, A., D’Enza, A.I., Van De Velden, M., 2019. Beyond Tandem Analysis: Joint Dimension Reduction and Clustering in R. J. Stat. Soft. 91. https://doi.org/10.18637/jss.v091.i10

Mathai, A., Provost, S., Haubold, H., 2022. Chapter 15: Cluster Analysis and Correspondence Analysis, in: Mathai, A.M., Provost, S.B., Haubold, H.J. (Eds.), Multivariate Statistical Analysis in the Real and Complex Domains. Springer International Publishing, Cham, pp. 845–886. https://doi.org/10.1007/978-3-030-95864-0\_15

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. J. Appl. Econ. 15, 447–470. https://doi.org/10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1

Morris, C.N., 1983. Parametric Empirical Bayes Inference: Theory and Applications. Journal of the American Statistical Association 78, 47–55. https://doi.org/10.1080/01621459.1983.10477920

Moskal, A., Martin, J.-L., Laumon, B., 2012. Risk factors for injury accidents among moped and motorcycle riders. Accident Analysis & Prevention 49, 5–11. https://doi.org/10.1016/j.aap.2010.08.021

Nguyen, D.V.M., Vu, A.T., Polders, E., Ross, V., Brijs, T., Wets, G., Brijs, K., 2021. Modeling the injury severity of small-displacement motorcycle crashes in Hanoi City, Vietnam. Safety Science 142. https://doi.org/10.1016/j.ssci.2021.105371

NHTSA, 2024a. Speed Safety Camera Enforcement | NHTSA [WWW Document]. URL https://www.nhtsa.gov/book/countermeasures-that-work/speeding-and-speed-management/countermeasures/enforcement/speed-safety-camera-enforcement (accessed 6.18.25).

NHTSA, 2024b. Promote Bicycle Helmet Use with Education | NHTSA [WWW Document]. URL https://www.nhtsa.gov/book/countermeasures-that-work/bicycle-safety/countermeasures/other-strategies-behavior-change (accessed 6.18.25).

Phillips, J., Walford, N., Hockey, A., Foreman, N., Lewis, M., 2013. Older people and outdoor environments: Pedestrian anxieties and barriers in the use of familiar and unfamiliar spaces. Geoforum 47, 113–124. https://doi.org/10.1016/j.geoforum.2013.04.002

Rahman, M.A., Das, S., Hossain, A., Codjoe, J., Mitran, E., Sun, X., 2024. Exploring Attribute Associations in Pedestrian-Involved Hit-and-Run Crashes through Cluster Correspondence Analysis. Transportation Research Record 2678, 1079–1098. https://doi.org/10.1177/03611981241242751

Rahman, M.A., Das, S., Sun, X., 2023. Understanding the Drowsy Driving Crash Patterns from Correspondence Regression Analysis. Journal of Safety Research 84, 167–181. https://doi.org/10.1016/j.jsr.2022.10.017

Saeed, T.U., Hall, T., Baroud, H., Volovski, M.J., 2019. Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: An empirical assessment of multilane highways. Analytic Methods in Accident Research. https://doi.org/10.1016/j.amar.2019.100101

Sarmiento, J.M., Gogineni, A., Bernstein, J.N., Lee, C., Lineen, E.B., Pust, G.D., Byers, P.M., 2020. Alcohol/Illicit Substance Use in Fatal Motorcycle Crashes. Journal of Surgical Research 256, 243–250. https://doi.org/10.1016/j.jss.2020.06.036

Schneider, W.H., Savolainen, P.T., 2011. Comparison of Severity of Motorcyclist Injury by Crash Types. Transportation Research Record: Journal of the Transportation Research Board 2265, 70–80. https://doi.org/10.3141/2265-08

Se, C., Champahom, T., Jomnonkwao, S., Karoonsoontawong, A., Ratanavaraha, V., 2021. Temporal stability of factors influencing driver-injury severities in single-vehicle crashes: A correlated random parameters with heterogeneity in means and variances approach. Analytic Methods in Accident Research 32, 100179. https://doi.org/10.1016/j.amar.2021.100179

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. https://doi.org/10.1016/j.amar.2017.05.003

Sivasankaran, S.K., Rangam, H., Balasubramanian, V., 2021. Investigation of factors contributing to injury severity in single vehicle motorcycle crashes in India. International Journal of Injury Control and Safety Promotion 28, pp-243-254. https://doi.org/10.1080/17457300.2021.1908367

Song, L., Lin ,Yixuan, Chen ,Guojun, Zhao ,Xin, Zhang ,Xuequan, and Fan, W. (David), 2025. Exploring behavior shifts and sample selectivity issues among speeding single-vehicle crash-injury severities before-and-after the stay-at-home order. International Journal of Injury Control and Safety Promotion 0, 1–13. https://doi.org/10.1080/17457300.2025.2496346

Song, M., McLaughlin, S., Doerzaph, Z., 2017. An on-road evaluation of connected motorcycle crash warning interface with different motorcycle types. Transportation Research Part C: Emerging Technologies 74, 34–50. https://doi.org/10.1016/j.trc.2016.11.005

Stamey, H.M., Brou, L.I., Meyers, K.R., Yang, H., Thackeray-Pflughoft, K., Spilman, S.K., 2024. Getting patients to the right level of trauma center after motorcycle crashes in a rural trauma system. The American Journal of Emergency Medicine 78, 8–11. https://doi.org/10.1016/j.ajem.2023.12.045

Stephens, A.N., Bishop, C.A., Liu, S., Fitzharris, M., 2017. Alcohol consumption patterns and attitudes toward drink-drive behaviours and road safety enforcement strategies. Accident Analysis & Prevention 98, 241–251. https://doi.org/10.1016/j.aap.2016.10.011

The R Foundation, 2024. R: The R Project for Statistical Computing [WWW Document]. URL https://www.r-project.org/ (accessed 12.21.24).

Train, K.E., 2009. Discrete Choice Methods with Simulation, Second Edition.

Truong, L.T., Nguyen, H.T., Tay, R., 2020. A random parameter logistic model of fatigue-related motorcycle crash involvement in Hanoi, Vietnam. Accident Analysis & Prevention.

TxDOT, 2024. CRIS Query [WWW Document]. URL https://cris.dot.state.tx.us/public/Query/app/home (accessed 6.18.25).

Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G., Isa-Tavarez, J., 2012. The Role of Built Environment on Pedestrian Crash Frequency. Safety Science, First International Symposium on Mine Safety Science and Engineering 2011 50, 1141–1151. https://doi.org/10.1016/j.ssci.2011.09.012

USDOT, 2024. Traffic Safety Facts | 2022 Data (No. DOT HS 813 589).

USDOT, 2023. What Is a Safe System Approach? | US Department of Transportation [WWW Document]. URL https://www.transportation.gov/safe-system-approach (accessed 6.18.25).

Wang, M.-H., 2022. Investigating the Difference of Factors Contributing to Motorcyclist Fatalities in Single Motorcycle and Multiple Vehicle Crashes in Taiwan. p. 17p.

Washington, S., Karlaftis, M.G., Mannering, F., Anastasopoulos, P., 2020. Statistical and Econometric Methods for Transportation Data Analysis. CRC Press.

Xin, C., Wang, Z., Lee, C., Lin, P.-S., 2017a. Modeling Safety Effects of Horizontal Curve Design on Injury Severity of Single-Motorcycle Crashes with Mixed-Effects Logistic Model. Transportation Research Record: Journal of the Transportation Research Board pp-38-46. https://doi.org/10.3141/2637-05

Xin, C., Wang, Z., Lin, P.-S., Lee, C., Guo, R., 2017b. Safety Effects of Horizontal Curve Design on Motorcycle Crash Frequency on Rural, Two-Lane, Undivided Highways in Florida. Transportation Research Record: Journal of the Transportation Research Board 2637, 1–8. https://doi.org/10.3141/2637-01

Ye, Y., He, J., Yan, X., Wang, C., Qin, P., 2025. Exploring determinants of motorcyclist non-violation crash injury severities on suburban roads of China: a random parameter logit model with heterogeneity in means and variances. Transportation Letters 0, 1–12. https://doi.org/10.1080/19427867.2025.2474317

Yu, M., Zheng, C., Ma, C., 2020. Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research. https://doi.org/10.1016/j.amar.2020.100126

Zhao, H., Chen, R., Deng, G., Yin, Z., Yang, G., Liu, S., Chen, H., Wang, Z., 2011. Comparison of injuries sustained by drivers and pillion passengers in fatal head-on motorcycle collision accidents. Forensic Science International 207, 188–192. https://doi.org/10.1016/j.forsciint.2010.10.003

Zhao, X., Zhan, M., Jie, C., 2018. Examining multiplicity and dynamics of publics’ crisis narratives with large-scale Twitter data. Public Relations Review 44, 619–632. https://doi.org/10.1016/j.pubrev.2018.07.004

# APPENDIX

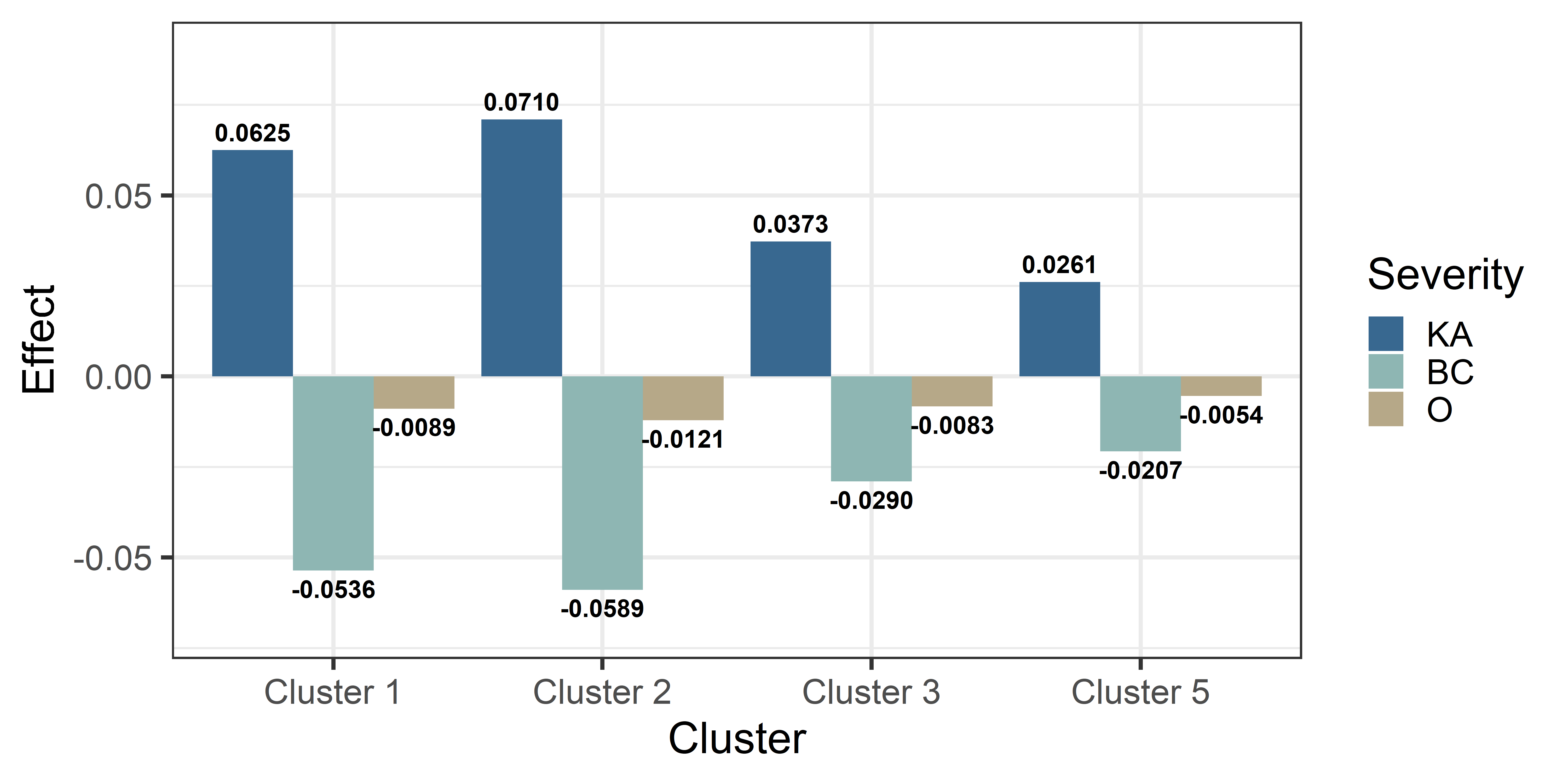


Figure .. Marginal effect of the variable helmet, not worn

A graph of clustering data

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable helmet, damaged

A graph with a bar graph

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable morning

A graph showing a number of clusters

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable afternoon

A graph with a bar graph

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable fall

A graph with numbers and a bar chart

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable daylight

A graph with different colored bars

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable dark, not lighted

A graph showing a number of clusters

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable curve, graded

A graph with a number of bars

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable straight, level

A graph with a bar graph

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable center/stripe/divider

A graph showing a number of clusters

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable marked lane

A graph with a bar graph and numbers

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable stop/yield/warning sign

A graph with different colored squares

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable crash speed 30 to 40 mph

A graph with numbers and a bar chart

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable crash speed greater than 65 mph

A graph with different colored squares

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable unsafe speed

A graph with a bar chart

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable fixed object

A graph of clustering data

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable overturned

A graph of clustering data

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable rear-end

A graph of a cluster

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable run-off-road

A graph with different colored squares

AI-generated content may be incorrect.

Figure .. Marginal effect of the variable rider aged greater than 65 years old