Scientific Reports

https://doi.org/10.1038/s41598-026-40755-5

Article in Press

# A multimodal data framework for motorcyclist injury severity on rural undivided roads

Received: 4 January 2026

Accepted: 16 February 2026

Published online: 01 March 2026

Cite this article as: Barua S., Dutta A.K. & Das S. A multimodal data framework for motorcyclist injury severity on rural undivided roads. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-40755-5

Swastika Barua, Anandi K. Dutta & Subasish Das

# A Multimodal Data Framework for Motorcyclist Injury Severity on Rural Undivided Roads

**Swastika Barua[1,*], Anandi K Dutta[2], Subasish Das[3]**

[1,3]Department of Civil Engineering, Texas State University, San Marcos, TX-78666, USA
[2]Department of Electrical Engineering, Texas State University, San Marcos, TX-78666, USA

*Corresponding author: swastika.barua.purna@gmail.com, qwx11@txstate.edu

**ABSTRACT**

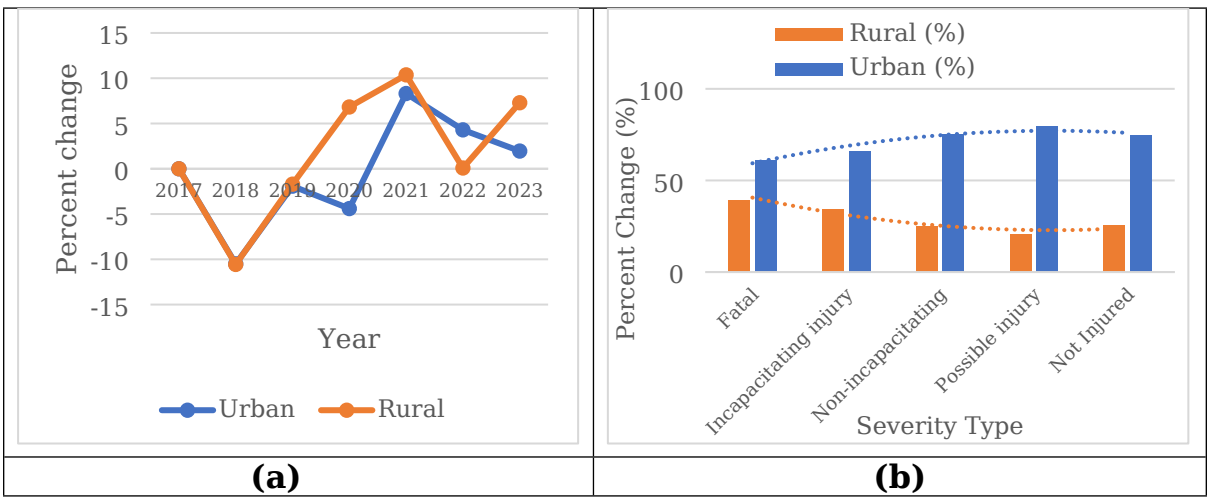
Motorcycle crashes on rural undivided roads remain a significant safety concern due to the high incidence of severe injuries and fatalities. This study analyzes a comprehensive dataset of 12,753 motorcycle crashes from rural undivided roads in Texas. Employing a multi-method approach, the research first applies Cluster Correspondence Analysis (CCA) to identify underlying patterns in crash characteristics, followed by the estimation of cluster-based Random Parameter Logit Models. Latent Dirichlet Allocation (LDA) topic modeling of crash narratives further supplements the analysis by uncovering key crash scenarios and thematic trends. The findings indicate that severe and fatal motorcycle injuries on rural undivided roads are primarily driven by high-speed loss-of-control events, particularly run-off-road and overturn crashes occurring on straight and curved segments. Crashes at intersections represent a severity mechanism, where inadequate lighting and turning or yielding conflicts combine to increase injury risk. Additionally, nighttime crashes on rural segments, especially those involving fixed objects or animals, emerge as a distinct high-risk scenario, reflecting the compounded effects of limited visibility, high operating speeds, and reduced reaction time on motorcyclist injury severity. The findings inform a suite of policy interventions grounded in the Safe System Approach (SSA), recommending context-sensitive speed management, rural infrastructure upgrades, helmet use promotion, and improved emergency and wildlife response as essential strategies.

*Keywords: Motorcycle crash severity, rural undivided roads, cluster correspondence analysis, RPL, SHAP, Topic modeling.*

## Introduction

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[1]. 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[1]. 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[1]. 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[2]. 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[2].

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 (a) from the Crash Records Information System (CRIS) database (2017–2023)[3] shows that rural areas experienced larger increases in motorcycle crashes over time, especially after 2020, while urban areas saw smaller or more stable changes. Figure 1 (b) 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[3]. 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.




**(a)** **(b)**

**Figure 1.** Percent change in motorcycle crash occurrences by (a) year

for urban and rural roadways (b) by injury severity by location (rural vs. urban) in Texas

While motorcycle crash severity has been examined extensively, this study makes several distinct contributions that extend existing research. Unlike prior studies that rely on either structured crash variables or narrative text in isolation, this study applies a rigorous multi-method approach to analyze 12,753 motorcycle crashes on rural undivided Texas roadways from 2017 to 2023. Cluster Correspondence Analysis (CCA) identifies key crash patterns by grouping cases based on roadway features, timing, rider demographics, and environmental factors. Random Parameter Logit Models then examine how the effects of risk factors on injury severity vary across clusters, accounting for unobserved heterogeneity. Additionally, Natural Language Processing (NLP) and bigram topic modeling of crash narratives reveal further themes and influences beyond structured data. By identifying empirically derived crash typologies on rural undivided roads and estimating cluster-specific random parameter logit models, the analysis captures context-dependent heterogeneity that is obscured in conventional pooled models. In addition, the convergence of econometric results with narrative-based themes strengthens the interpretability and validity of the findings. These methodological advances enable the development of scenario-specific, Safe System Approach–aligned countermeasures tailored to the dominant rural motorcycle crash mechanisms rather than generalized safety recommendations.

## Literature review

### 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[4] 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.[5] reported that novice motorcyclists in their first year post-licensing have a heightened risk of crashes as they adapt to real-world riding conditions. In contrast, Phillips et al.[6] 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. Christian et al.[7] 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.[8] emphasized the importance of helmet type, revealing that non-standard helmets, such as half-shell designs, provided minimal protection during crashes. 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.[9] 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.[10] 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.[11] reported that intoxicated motorcyclists were disproportionately represented in fatal single-vehicle crashes, underscoring the heightened dangers of impaired riding.

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[12] found that rural roads with limited lighting and higher speed limits experienced more severe motorcycle crashes compared to urban roads. Meanwhile, Haque et al.[13] identified intersections as high-risk locations for motorcyclists, especially those involving uncontrolled left turns. 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[14]. 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[15]. Collision with vulnerable road users and crashes on rural undivided roads have been shown to exhibit strong temporal and contextual heterogeneity in injury severity, and motorcyclists, also classified as vulnerable road users, are subject to similar influences related to roadway geometry, lighting conditions, and operating speed[16–18].

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[19]. 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[20]. 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[21]. 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[22]. Figure 2 illustrates motorcyclist involved crash types.

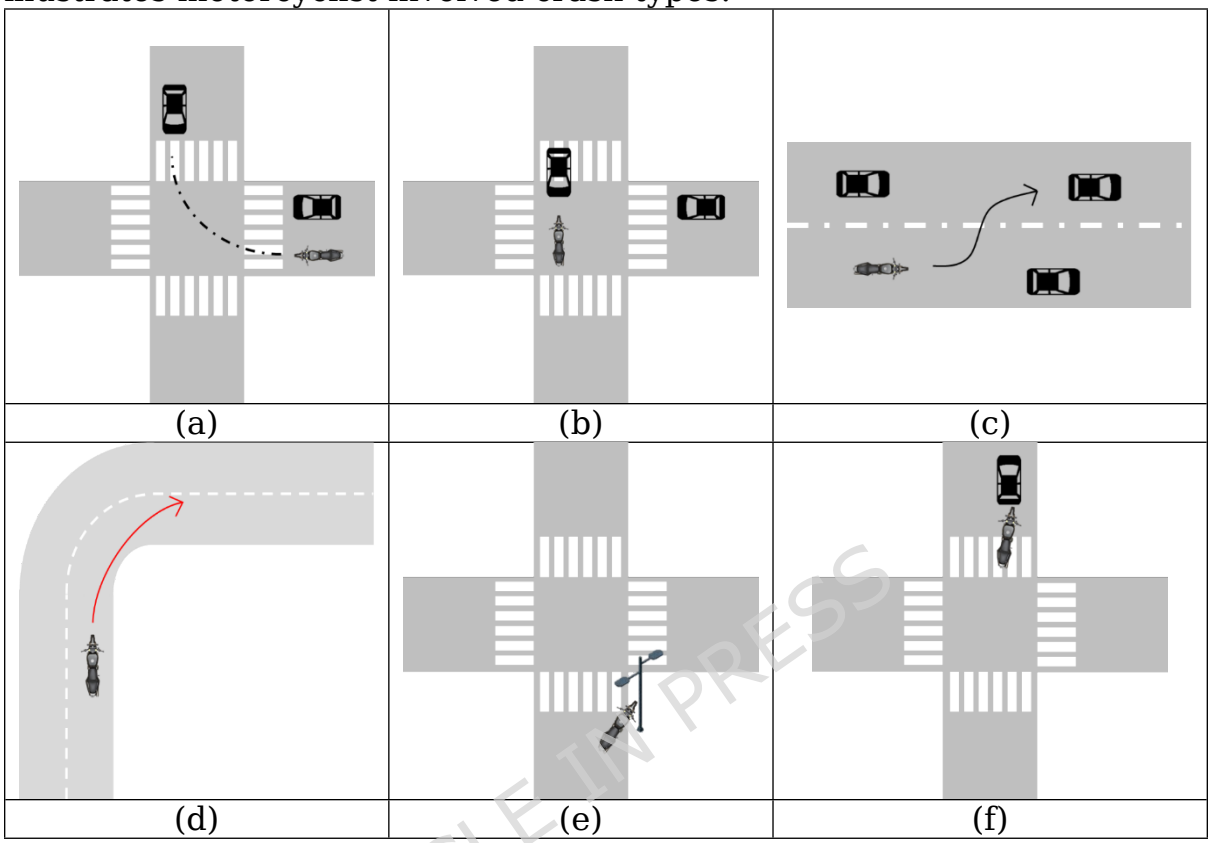



**Figure 2. Crash types involving motorcyclists (a) turning collisions (b) head-on (c) lane changing (d) horizontal curve crashes (e) fixed object and (f) rear-end crashes**



### 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[23–25], multiple correspondence analysis (MCA)[26,27], and taxicab correspondence analysis (TCA)[24] 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[23,25,26,28].

Recent studies show that topic modeling, such as LDA, can extract clusters of risk factors like rider actions, weather, and roadway geometry from motorcycle crash narratives, providing insights beyond structured data[29]. 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[30]. 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[31]. 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[29–32].

The development of random parameter modeling approaches has marked a significant methodological shift in crash severity analysis. 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 demonstrated its effectiveness in revealing unobserved heterogeneity in crash injury severity and showed that traditional fixed-parameter models often mask important variability in risk effects across crashes.[33–35]. Table 1 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.

**Research Gap**

Despite extensive research on motorcycle crash severity, key gaps persist. Most studies analyze either structured data or narrative text separately, rarely integrating both for a complete picture. Cluster-based and random parameter models are often used in isolation, and little research focuses on rural two-way undivided roadways, which have high rates of severe crashes. Additionally, few studies combine topic modeling of crash narratives with advanced econometric models to examine the mechanisms behind crash severity, and the influence of unobserved heterogeneity and complex factor interactions remains underexplored, especially in high-risk rural settings.

**Table 1.** 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)[33] | 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. (2025)[36] | 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 (2021)[37] | 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. (2024)[38] | 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)[39] | 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)[40] | 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)[41] | 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)[34] | MC crash severity | US motorcycle crash data | RPL | Rider behavior, environment | Accounting for random effects reveals unobserved risk factors. |
| Kim et al. (2024)[42] | Mixed vehicle crash injury severity | South Korea, 2012–2015 | RPLMV | Vehicle type, road/traffic conditions | Different vehicle types require tailored safety policies. |

## Study design and methodology

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 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.

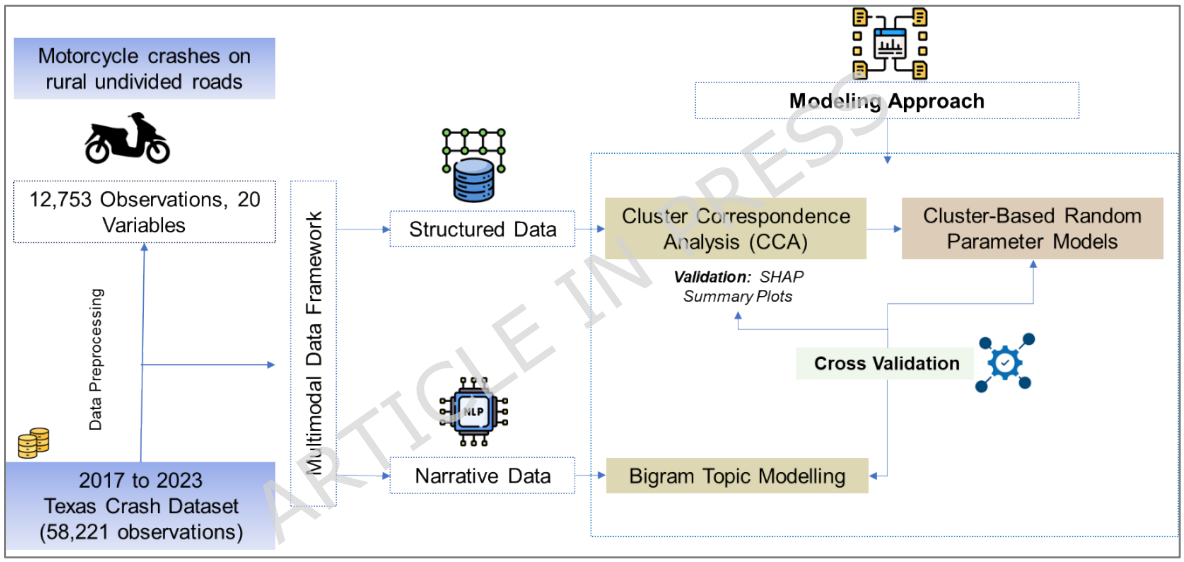



**Figure 3.** Flow chart describing research approach

Some assumptions and estimation procedures were followed throughout the analysis to ensure transparency and reproducibility. Random parameters were specified using normal distributions and estimated via simulated maximum likelihood with Halton draws, with model selection guided by likelihood ratio tests, Akaike Information Criterion (AIC), and McFadden's pseudo R-squared. The number of clusters in the CCA was selected based on within-cluster compactness and interpretability, and the topic number in the LDA analysis was specified accordingly.

### Cluster Correspondence Analysis

The CCA method combines dimension reduction with cluster analysis for categorical data, improving cluster convergence compared to earlier methods [43]. Figure 4 illustrates the stepwise process of Cluster Correspondence Analysis (CCA), detailing how cross-tabulation, clustering, and biplot generation are integrated to identify clusters within

categorical crash data[43–46].

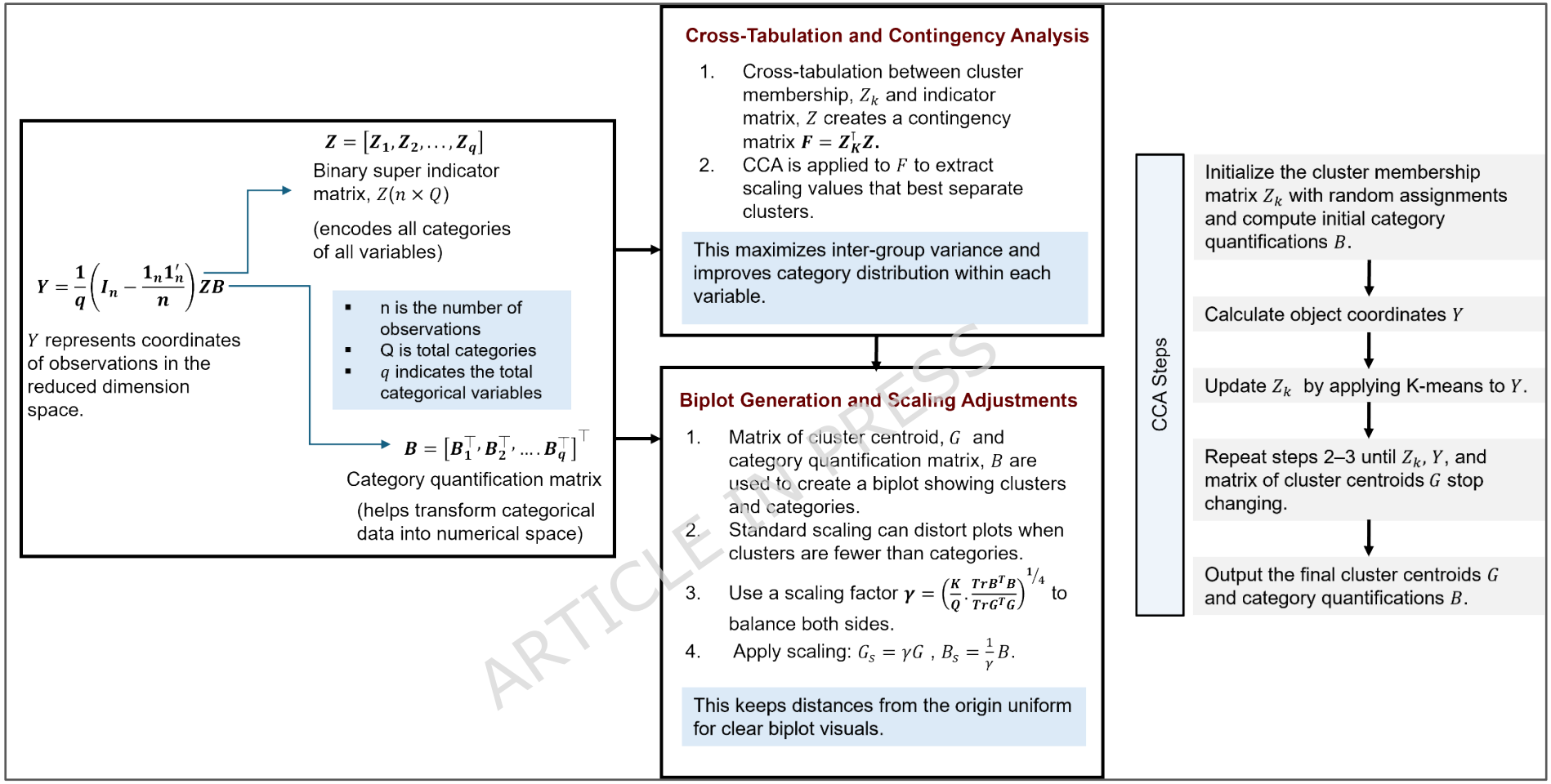



**Figure 4.** Stepwise process of CCA

**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 roads. 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 $f(x)$, the SHAP value for a feature $x_i$ is computed as[44,47]:

$$f(x) = \sum_{S \subseteq F \setminus \{i\}}^{\infty} \left( \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(s)] \right) \quad (1)$$

where:

$F$ is the full set of features,

$S$ is a subset of features excluding $x_i$ ,

$f(S)$ is the model's prediction using only the features in SSS,

$f(S \cup \{i\})$ is the prediction when $x_i$ is added,

$\phi_i$ represents the SHAP value for feature $x_i$ , indicating its marginal contribution to the prediction.

**Latent Dirichlet Allocation**

LDA is a generative probabilistic model of corpus-based analysis. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words[48].

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

1. Choose $N \sim$ Poisson ($\xi$).
2. Choose $\theta \sim$ Dir ($\alpha$).
3. For each of the $N$ words $w_n$

Choose a topic $z_n \sim$ Multinomial ($\theta$).

Choose a word $w_n$ from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic $z_n$.

A $k$ -dimensional Dirichlet random variable $\theta$ can take values in the ( $k - 1$)-simplex (a $k$ -vector $\theta$ lies in the $(k - 1)$-simplex if $\theta_i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$), and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_k - 1} \ldots \theta_k^{\alpha_k - 1}, \quad (2)$$

where the parameter $\alpha$ is a $k$ -vector with components $\alpha_i > 0$, and where, $\Gamma(x)$ is the Gamma function. Given the parameters $\alpha$ and $\beta$, the joint distribution of a topic mixture $\Gamma(x)$, a set of $N$ topics $z$, and a set of $N$ words $w$ is given by:

$$p(\theta, z, w \mid \alpha, \beta = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad (3)$$

where $p(z_n \mid \theta)$ is simply $\theta_i$ for the unique $i$ such that $z_n^i = 1$. Integrating over $\theta$ and summing over $z$, it is obtained the marginal distribution of a document:

$$p(w \mid \alpha,\beta) = \int p(\theta \mid \alpha) \left( \prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta \quad (4)$$

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

$$p(D \mid \alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d \quad (5)$$

**Random Parameter Logit Model**

In this study, injury severity was classified into three categories: fatal and incapacitating injuries (KA), non-incapacitating and possible injuries (BC), and no injury (O). Due to the categorical nature of the outcome variable, a Multinomial Logit (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), and Random Parameters Logit model with Heterogeneity in Means (RPLHM). Each modeling step progressively allowed for greater flexibility by accounting for parameter correlations and systematic variations. Model fit and selection were evaluated using the likelihood ratio test, AIC, and McFadden's Pseudo R-squared[49].

*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[38,38,50–55]. In Figure 5, the sequential steps of the mixed logit modeling framework are presented, illustrating how observed and unobserved factors are integrated to calculate injury severity scores for each crash event. The diagram highlights the use of random parameters to account for unobserved heterogeneity, the conversion of severity scores into outcome probabilities, and the interpretation of how changes in explanatory variables influence the likelihood of different injury severity levels.

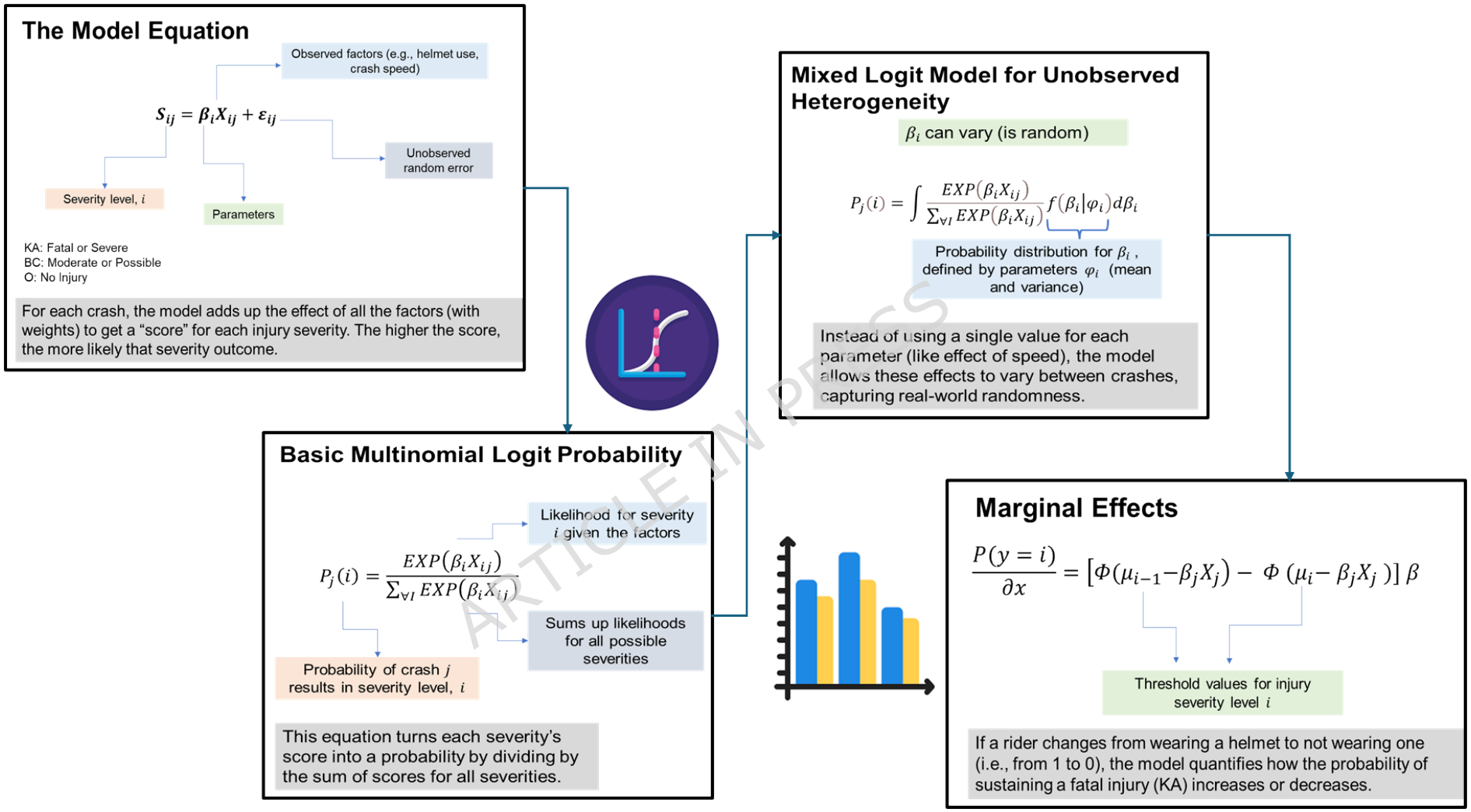



**Figure 5.** Stepwise process of CCA

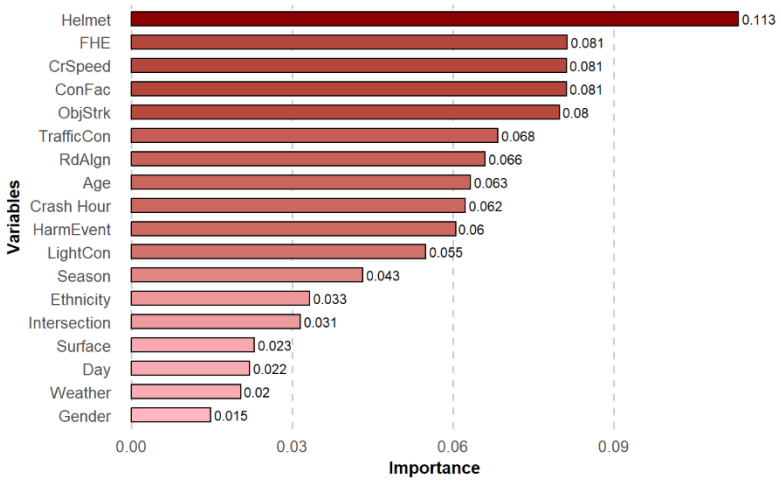
## Results and discussions

### Cluster Correspondence Analysis

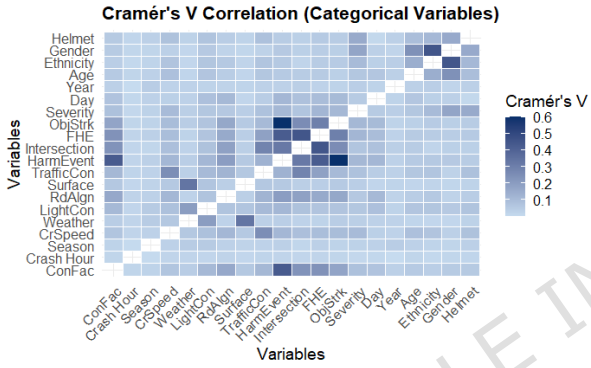
*Variable Importance*

Prior to utilizing the CCA, variable importance analysis utilizing XGBoost[56], was conducted on the dataset with crash severity as the dependent variable[44,57,58]. 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 6. To select variables for the CCA, 18 candidates were first assessed using an XGBoost importance plot, which identified Helmet use, First Harmful Event (FHE), Crash Speed (CrSpeed), Contributing Factor (ConFac), and Object Struck (ObjStrk) as top predictors of crash severity. *Contributing Factor* denotes the factor or combination of factors reported as contributing to the occurrence of a crash. In rural settings, these factors commonly include unsafe speed, animal presence on the roadway, failure to yield or signal, and driver inattention or fatigue. Another variable, *First Harmful Event,* describes the initial event in the crash sequence that produced damage or injury, such as a run-off-road departure, overturn, or collision. The variable *Object Struck* specifies the physical object or entity impacted during the crash, including roadside features, animals, other vehicles, or traffic control devices, providing additional context for understanding injury severity on rural undivided roads.

To select a consistent and non-redundant set of variables for the CCA, a structured multi-step screening process was applied. First, an XGBoost-based variable importance analysis was used to rank predictors according to their contribution to crash injury severity, and variables with importance scores below 0.04 were excluded. Next, Cramér's V correlation coefficients were computed for the remaining categorical variables, with values greater than 0.5 indicating a strong association. In addition, a ranked pairwise cross-correlation analysis was conducted, and only statistically significant relationships ($p < 0.05$) were considered. When strong or moderate correlations were identified (e.g., between Harmful Event and First Harmful Event, or between Intersection and Traffic Control), one variable was retained based on clarity and relevance. This process resulted in a final set of 12 variables, improving interpretability while reducing multicollinearity in the clustering analysis.



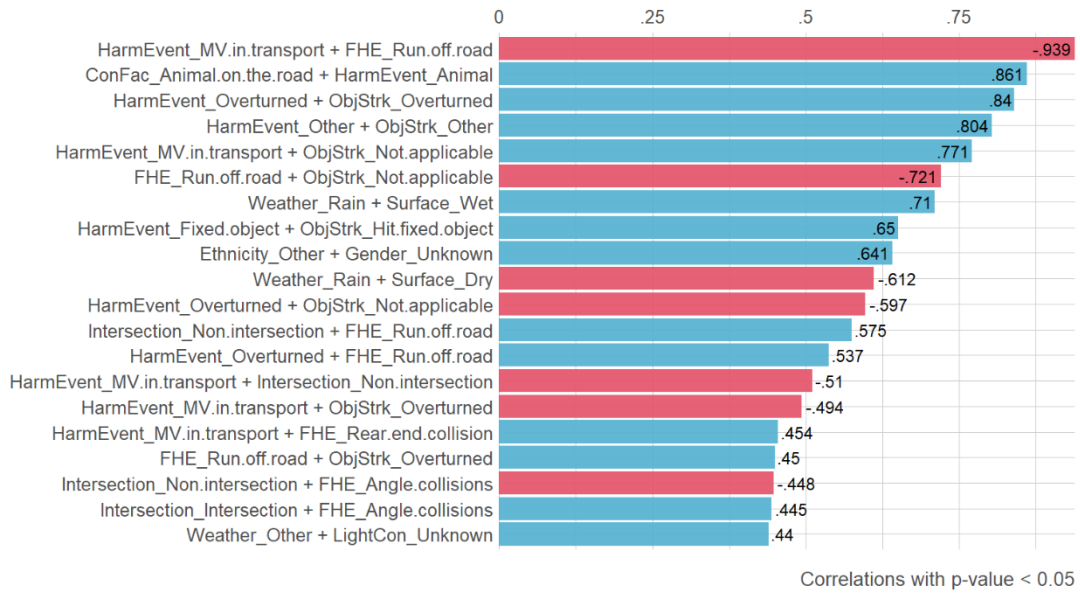

(a) XGBoost




(b) Cramér's V

| Variable 1 | Variable 2 | Cramér's V |
|---|---|---|
| HarmEvent | ObjStrk | 0.602 |
| Intersection | FHE | 0.454 |
| Ethnicity | Gender | 0.454 |
| ConFac | HarmEvent | 0.435 |
| HarmEvent | FHE | 0.432 |
| Weather | Surface | 0.341 |
| HarmEvent | Intersection | 0.325 |
| FHE | ObjStrk | 0.302 |
| TrafficCon | Intersection | 0.289 |
| Intersection | ObjStrk | 0.259 |

(c) Cramér's V values for pairwise variables




(d) Ranked cross-correlation table

**Figure 6.** Variable selection process with (a) XGBoost, (b) Cramér's V, and (c) Cramér's V values Cross-Correlations for pairwise variables and (d) ranked cross-correlation table

The CCA method was employed to examine crashes involving motorcyclists on the rural undivided roads using the top 12 variables identified from the variable importance analysis. The 'clustrd' package[43] in the R software[59] was utilized in this analysis. Table 2 presents the clusters formed using the CCA using the K-means algorithm[57,58], which identifies associations between categorical variables by analyzing their relationships in a contingency table. The results from the clustering analysis in Table 2 highlight the unique characteristics of each group of child pedestrian-involved crashes based on their centroids, size, and within-cluster sum of squares (WCSS). For example, Cluster 1 consists of 3,382 crashes, which is 26.5% of all child pedestrian-involved crashes in the dataset. The centroid for Cluster 1 is located at 0.0025 on Dimension 1 and -0.0005 on Dimension 2, representing the average position of all crashes in this group on the main axes of variation. The WCSS for Cluster 1 is 0.0201, indicating how tightly the crashes in this cluster are grouped around the centroid a lower WCSS like this suggests a relatively compact and well-defined cluster. The variation in centroid positions and cluster sizes highlights the diverse nature of crash circumstances present in the dataset as shown in Figure 7.

**Table 2.** 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*

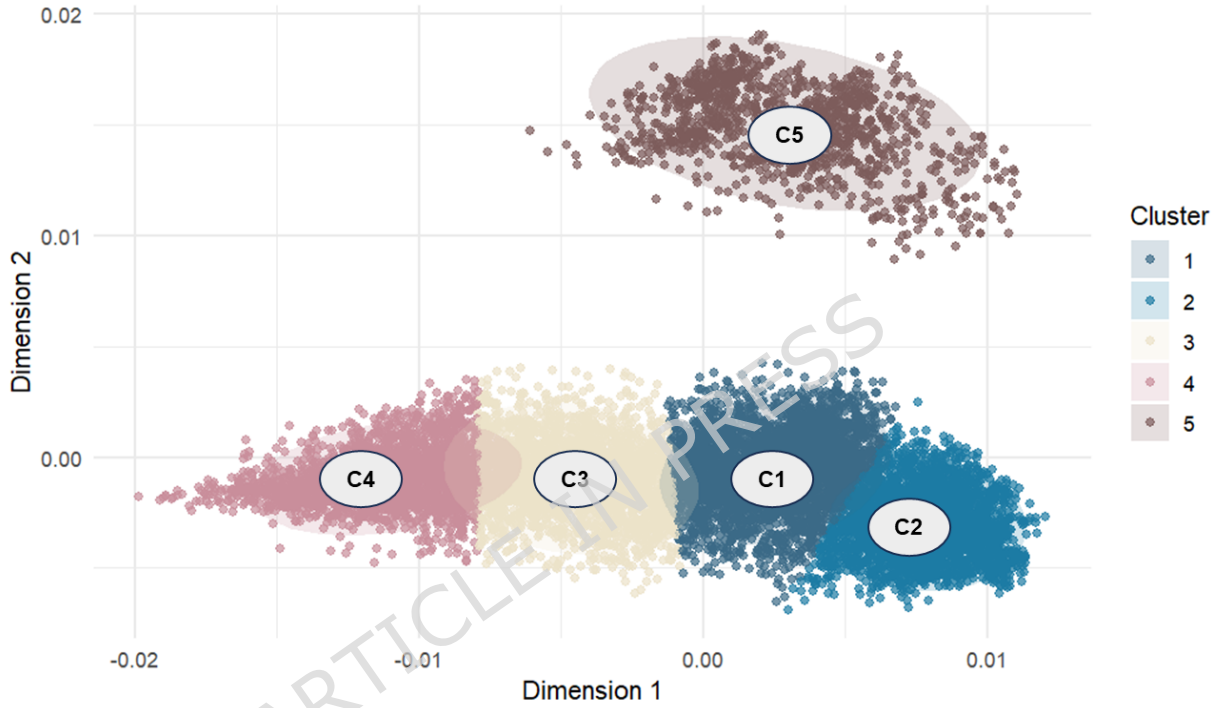



**Figure 7.** Clusters produced from motorcyclist injury severity crash data

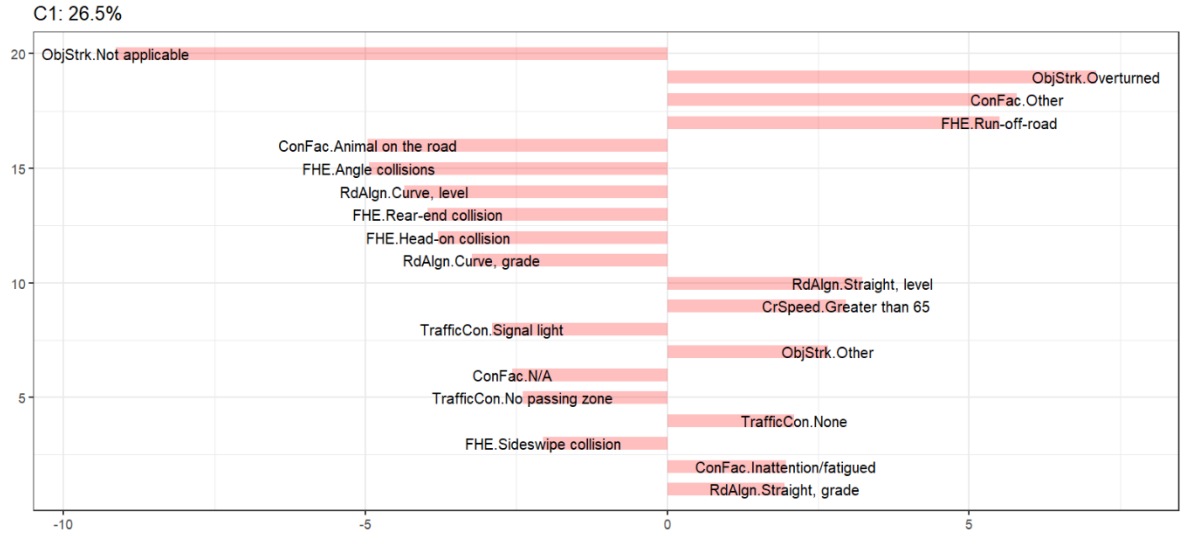


**Findings of the Cluster Correspondence Analysis**

*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 8). 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, which indicates the 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[60]. 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[51].

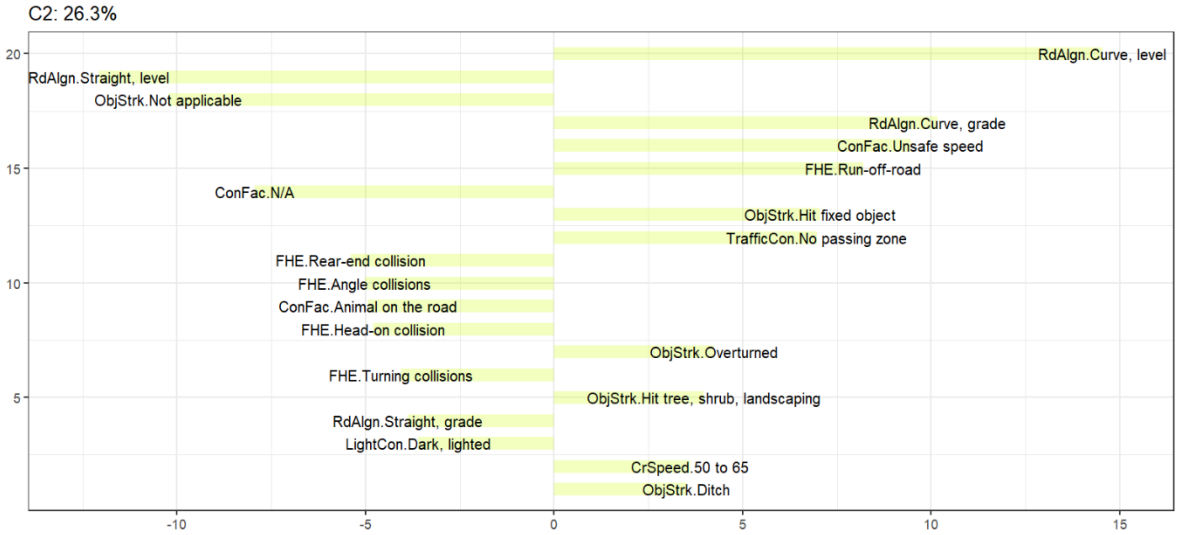



**Note:** *The x-axis shows the CCA category scores indicating each attribute's association with the cluster, and the y-axis lists the top 20 most frequent crash attribute levels within that cluster.*

**Figure 8.** Bar plot 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 9). 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. The combination of adverse geometry and speed amplifies the risk of losing traction or veering off the roadway. 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[51].

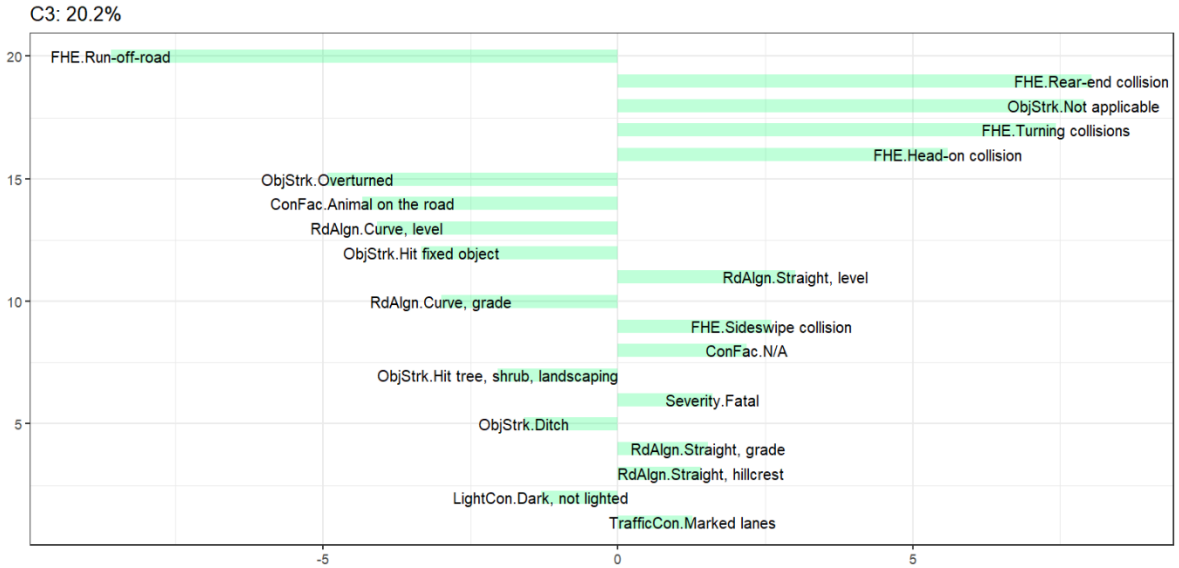



**Note:** *The x-axis shows the CCA category scores indicating each attribute's association with the cluster, and the y-axis lists the top 20 most frequent crash attribute levels within that cluster.*

**Figure 9.** Bar plot 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 10. The contributing factors of overturned vehicles and hitting fixed objects 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[55,61]. A representative scenario would involve a motorcyclist encountering a sudden obstacle, such as a fixed object or trees, in a curved roadway segment, leading to a turning or head-on collision.

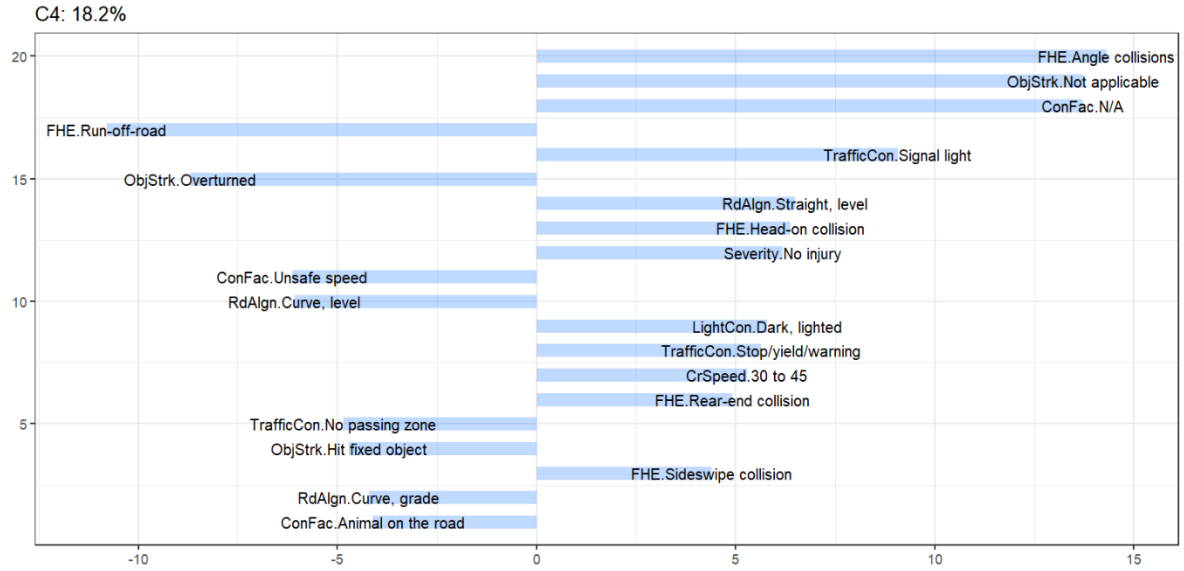



**Note:** *The x-axis shows the CCA category scores indicating each attribute's association with the cluster, and the y-axis lists the top 20 most frequent crash attribute levels within that cluster.*

**Figure 10.** Bar plot for cluster 3

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

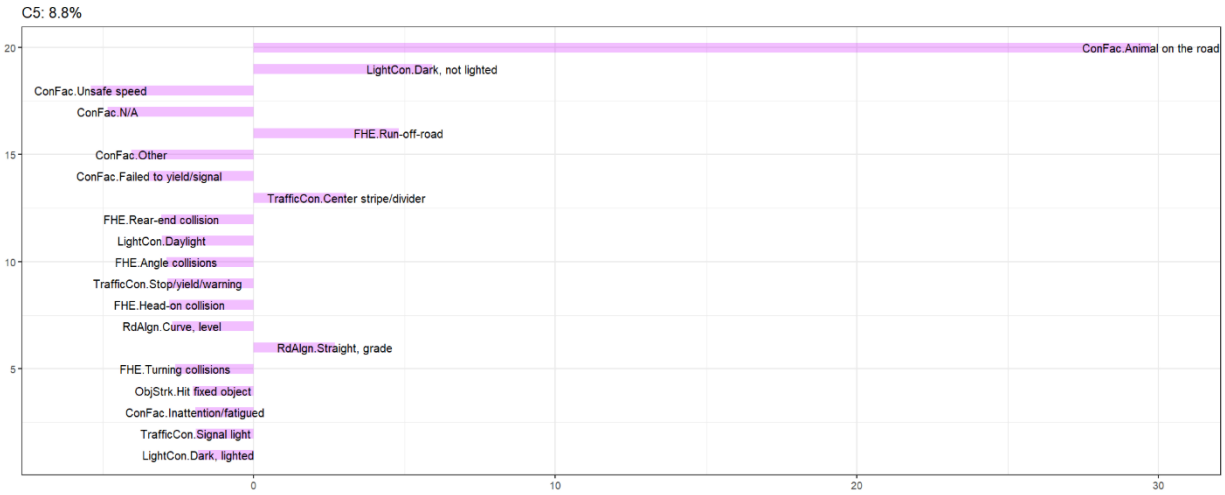
Most of the crashes in Cluster 4 in Figure 11 involved run-off-road and overturned vehicle scenarios, with unsafe speed emerging as a critical contributing factor. Crashes frequently occurred on curved, level roads and were associated with the presence of angle collisions and incidents at signalized 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[62,63]. 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.




**Note:** *The x-axis shows the CCA category scores indicating each attribute's association with the cluster, and the y-axis lists the top 20 most frequent crash attribute levels within that cluster.*

**Figure 11.** Bar plot 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 12), is characterized by a concentration of incidents involving dark, unlit 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[64]. 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.



**Note:** *The x-axis shows the CCA category scores indicating each attribute's association with the cluster, and the y-axis lists the top 20 most frequent crash attribute levels within that cluster.*

**Figure 12.** Bar plot for cluster 5

### Validation of Cluster Findings by SHAP

The analysis of SHAP summary plots in Figure 13 across all clusters underscores the pivotal role of unsafe speed, helmet use, and object struck in determining motorcyclist injury severity on rural undivided roads. For this validation, separate XGBoost models were trained for each crash cluster using cluster-specific datasets, and SHAP analyses were conducted independently to capture cluster-level feature contributions. For each cluster, the script encodes categorical variables, fits an XGBoost multi-class severity model using an 80/20 stratified split, and then computes SHAP values on the test data to produce the cluster-specific summary plot.

Unsafe speed is the most influential predictor of severe outcomes, especially in Clusters 1 and 2, where SHAP plots confirm that high speeds and loss of control are strongly linked to run-off-road incidents, overturns, and collisions with fixed objects. Across all clusters, helmet non-use and striking hazardous objects consistently raise the risk of fatal and serious injuries, with SHAP values validating these effects in each group. Cluster 3 is marked by severe head-on collisions and control-related issues, as highlighted by high SHAP values, while Clusters 1 and 4 are dominated by run-off-road and angle crashes patterns directly mirrored in their respective SHAP plots. Environmental and demographic factors, including poor lighting, adverse seasons, and rider age, are especially important in Cluster 5, where SHAP analysis underscores the severity of animal-involved nighttime crashes.

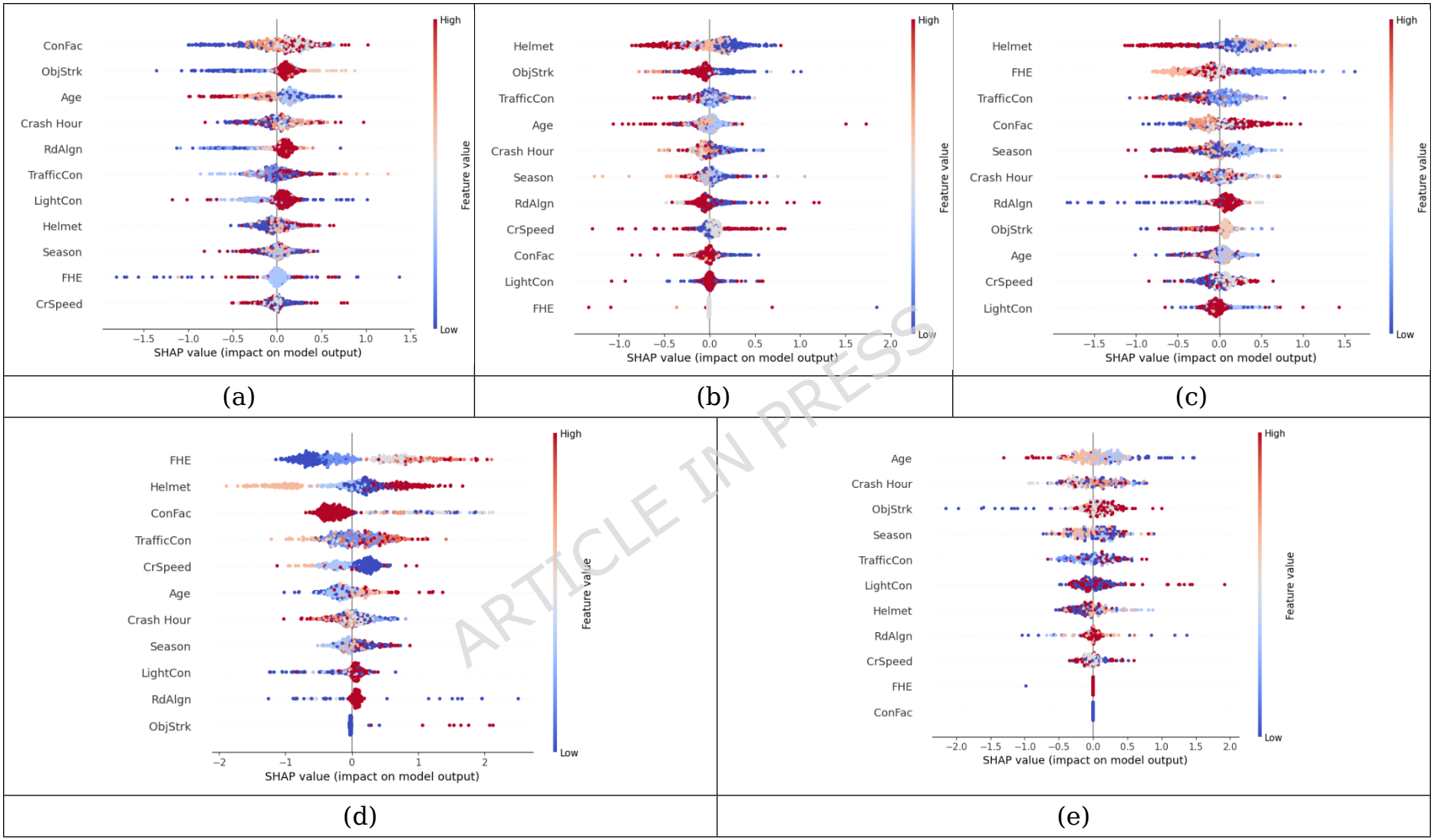



**Figure 13.** SHAP summary plot of (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4, and (e) cluster 5

**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. Figure 14 presents the distribution of injury severity levels among motorcyclist crashes on rural undivided roads, segmented by cluster.

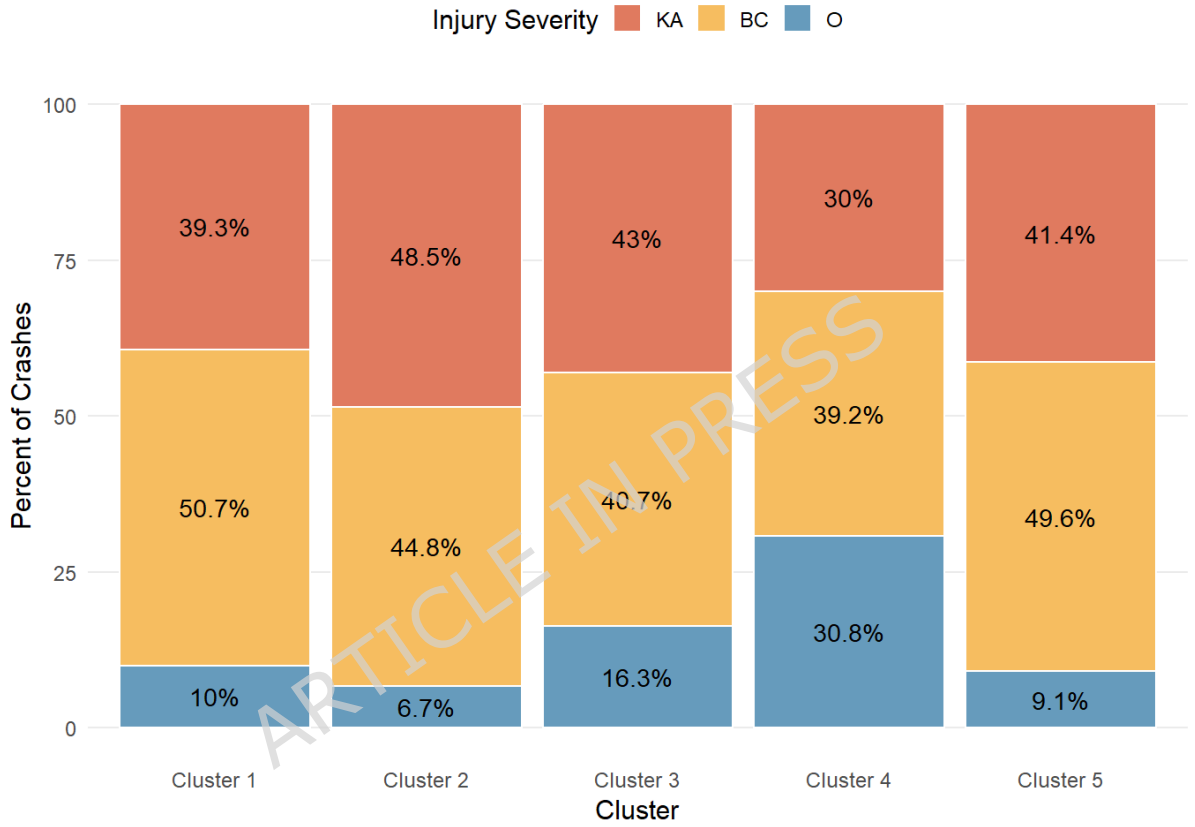



**Figure 14.** Cluster-based crash distribution

The descriptive statistics in Table 3 highlight several notable patterns across the five clusters. The highest rate of helmet non-use is seen in Cluster 1, while helmet damage is most frequent in Cluster 2. Cluster 4 records the greatest proportion of crashes at speeds between 30–45 mph, whereas Cluster 1 and Cluster 5 share the highest percentage of crashes occurring above 65 mph. Unsafe speed is most prevalent in Cluster 2, and the presence of a center stripe or divider peaks in Cluster 5. Marked lanes are most common in Clusters 1 and 3, and no passing zones are reported most frequently in Cluster 2. Clusters 1 and 5 have the highest proportion of crashes with no traffic control devices, but stop/yield/warning signs are most frequently present in Cluster 4. The occurrence of crashes involving riders over age 65 is highest in Cluster 2, while curve and grade alignments are notably more common in Cluster 2 as well. Straight and level roadway alignment is most prevalent in Cluster 4. Regarding lighting, dark, not lighted

conditions are most common in Cluster 5, while daylight crashes peak in Cluster 2. Rear-end crashes are reported most frequently in Cluster 3, whereas run-off-road crashes overwhelmingly dominate Cluster 2 and Cluster 5. Turning maneuvers and hitting fixed objects occur most often in Cluster 3 and Cluster 2, respectively. Overturned crashes are most common in Cluster 1.

**Table 3.** Cluster-Based Descriptive Statistics

| Variables | Cluster 1 | | Cluster 2 | | Cluster 3 | | Cluster 4 | | Cluster 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Helmet not worn (1 if not worn, 0 otherwise) | 0.438 | 0.50 | 0.400 | 0.49 | 0.412 | 0.49 | 0.355 | 0.48 | 0.414 | 0.49 |
| Helmet (1 if damaged, 0 otherwise) | 0.275 | 0.45 | 0.314 | 0.46 | 0.271 | 0.44 | 0.217 | 0.41 | 0.305 | 0.46 |
| Crash speed 30-45 mph (1 if speed between 30 to 45 mph, 0 otherwise) | 0.336 | 0.47 | 0.266 | 0.44 | 0.338 | 0.47 | 0.584 | 0.49 | 0.265 | 0.44 |
| Crash speed greater than 65 mph(1 if speed greater than 65 mph, 0 otherwise) | 0.213 | 0.41 | 0.103 | 0.30 | 0.157 | 0.36 | 0.065 | 0.25 | 0.212 | 0.41 |
| Unsafe speed (1 if rider riding with an unsafe speed, 0 otherwise) | 0.280 | 0.45 | 0.629 | 0.48 | 0.302 | 0.46 | 0.066 | 0.25 | 0.000 | 0.00 |
| Center stripe or divider (1 if the center stripe/ divider is present, 0 otherwise) | 0.209 | 0.41 | 0.185 | 0.39 | 0.188 | 0.39 | 0.086 | 0.28 | 0.323 | 0.47 |
| Marked lane (1 if the marked lane is present, 0 otherwise) | 0.356 | 0.48 | 0.218 | 0.41 | 0.356 | 0.48 | 0.346 | 0.48 | 0.243 | 0.43 |
| No passing zone (1 if the no passing zone is present, 0 otherwise) | 0.094 | 0.29 | 0.309 | 0.46 | 0.123 | 0.33 | 0.014 | 0.12 | 0.173 | 0.38 |
| No traffic control devices (1 if no traffic control is present, 0 otherwise) | 0.238 | 0.43 | 0.183 | 0.39 | 0.166 | 0.37 | 0.102 | 0.30 | 0.238 | 0.43 |
| Stop or yield or warning sign (1 if stop/yield/warning signs are present, 0 otherwise) | 0.079 | 0.27 | 0.096 | 0.30 | 0.108 | 0.31 | 0.253 | 0.44 | 0.015 | 0.12 |
| Rider age (1 if rider's age is greater than 65 years old, 0 otherwise) | 0.085 | 0.28 | 0.096 | 0.29 | 0.088 | 0.28 | 0.076 | 0.27 | 0.082 | 0.27 |
| Curve grade (1 if the roadway alignment is curve and graded, 0 otherwise) | 0.039 | 0.19 | 0.293 | 0.46 | 0.036 | 0.19 | 0.005 | 0.07 | 0.056 | 0.23 |
| Curve level (1 if the roadway alignment is curve and leveled, 0 otherwise) | 0.096 | 0.29 | 0.621 | 0.49 | 0.087 | 0.28 | 0.014 | 0.12 | 0.086 | 0.28 |
| Straight level (1 if the roadway alignment is straight and leveled, 0 otherwise) | 0.699 | 0.46 | 0.018 | 0.13 | 0.709 | 0.45 | 0.903 | 0.30 | 0.661 | 0.47 |
| Afternoon (1 if crash happened in the afternoon, 0 otherwise) | 0.255 | 0.44 | 0.265 | 0.44 | 0.252 | 0.43 | 0.235 | 0.42 | 0.220 | 0.41 |
| Morning (1 if crash happened in the morning, 0 otherwise) | 0.244 | 0.43 | 0.249 | 0.43 | 0.245 | 0.43 | 0.264 | 0.44 | 0.255 | 0.44 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dark, not lighted (1 if crash happened in dark, not lighted condition, 0 otherwise) | 0.245 | 0.43 | 0.218 | 0.41 | 0.184 | 0.39 | 0.118 | 0.32 | 0.519 | 0.50 |
| Daylight (1 if crash happened in daylight, 0 otherwise) | 0.674 | 0.47 | 0.745 | 0.44 | 0.693 | 0.46 | 0.668 | 0.47 | 0.413 | 0.49 |
| Fall (1 if crash happened in fall, 0 otherwise) | 0.267 | 0.44 | 0.258 | 0.44 | 0.279 | 0.45 | 0.293 | 0.45 | 0.307 | 0.46 |
| Winter (1 if crash happened in winter, 0 otherwise) | 0.143 | 0.35 | 0.144 | 0.35 | 0.161 | 0.37 | 0.194 | 0.40 | 0.142 | 0.35 |
| Rear-end (1 if the crash type is rear-end, 0 otherwise) | 0.029 | 0.17 | 0.001 | 0.03 | 0.284 | 0.45 | 0.221 | 0.42 | 0.003 | 0.05 |
| Run-off-road (ROR) (1 if the crash type is ROR, 0 otherwise) | 0.861 | 0.35 | 0.987 | 0.11 | 0.149 | 0.36 | 0.002 | 0.05 | 0.995 | 0.07 |
| Turning (1 if the crash happened involving turning maneuver, 0 otherwise) | 0.077 | 0.27 | 0.008 | 0.09 | 0.213 | 0.41 | 0.048 | 0.21 | 0.001 | 0.03 |
| Hit fixed object (1 if the crash happened involving hitting fixed object, 0 otherwise) | 0.142 | 0.35 | 0.270 | 0.44 | 0.043 | 0.20 | 0.004 | 0.06 | 0.049 | 0.22 |
| Overturned (1 if crash happened involving overturning, 0 otherwise) | 0.696 | 0.46 | 0.586 | 0.49 | 0.203 | 0.40 | 0.016 | 0.13 | 0.439 | 0.50 |

For each crash cluster, model estimation followed a systematic procedure beginning with a MNL specification, followed by an RPL model, and extended to an RPLHM. All random-parameter models were estimated using simulated maximum likelihood with 1,000 Halton draws to ensure stable and efficient estimation. Model selection was based on comparative improvements in log-likelihood (LL) values and reductions in the AIC/N. Using this framework, Clusters 1 (LL: -3111.53308, AIC/N: 1.852 ), Cluster 2 (LL:-2892.93482, AIC/N: 1.735 ) and Cluster 3 (LL: -2538.99162, AIC/N: 1.985) were best represented by the RPLHM model, as the inclusion of heterogeneity in means yielded meaningful improvements in both LL and AIC/N relative to the MNL and RPL models. In contrast, for Cluster 4 (LL: -2474.44496, AIC/N: 2.137), the MNL model was retained because extensions to random-parameter structures did not provide sufficient improvement in model fit. Finally, Cluster 5 (LL: -1047.8101, AIC/N: 1.892) was optimally modeled using an RPL specification, indicating the presence of unobserved heterogeneity without statistically significant heterogeneity in means.

## Model Estimation Results

*Rider Behavior*

The results indicate that helmet not worn consistently shows strong positive coefficients for KA outcomes across multiple clusters, with the highest value in Cluster 1 (1.583), highlighting the elevated risk for KA injuries among motorcyclists not wearing helmets on rural undivided roads[65]. Helmet worn but damaged generally shows negative coefficients in Clusters 1, 2, and 3, reflecting a protective effect against higher severity injuries, while Cluster 4 shows a positive coefficient (0.233, Table 4).

Unsafe speed has negative coefficients in Clusters 1 and 2 but shifts to a positive coefficient in Cluster 4 (0.641, Table 4), which indicates that the effect of unsafe speed on crash severity varies depending on the context and may either increase or decrease risk[55].

The marginal effect analysis further clarifies these relationships. Helmet not worn results in the greatest increase in the probability of KA in Cluster 2 (0.0710, Table 5), followed by Cluster 1 (0.0625, Table 5), emphasizing its strong role in severe injury risk[65]. Helmet worn but damaged generally reduces the probability of BC across clusters, while Cluster 4 is notable for an increased probability of BC. Unsafe speed, while only slightly increasing the likelihood of KA in Clusters 1 and 2, leads to a substantial increase in the probability of O in Cluster 4 (0.1350, Figure 15), in line with its positive coefficient in this cluster[55].

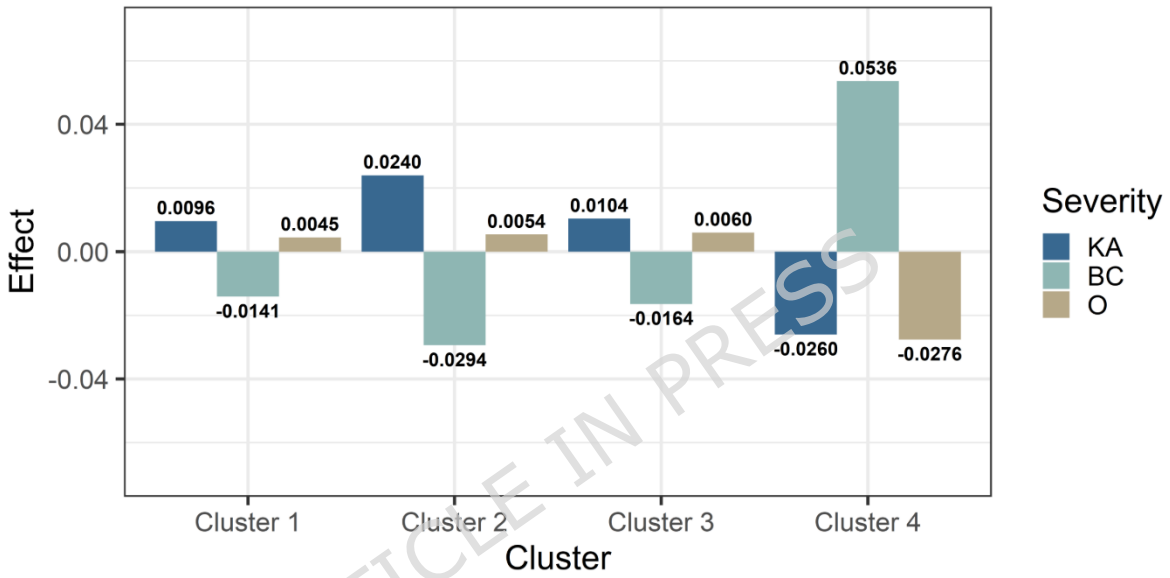



**Figure 15.** Marginal effects of the variable helmet worn but damaged

*Rider Demographic*

Rider age greater than 65 years old shows positive coefficients in Cluster 1 (0.780), Cluster 3 (0.965, Table 4), and Cluster 4 (0.696), indicating that older motorcyclists are generally associated with higher crash severity[54]. Marginal effect analysis reveals that while probabilities for KA and BC decrease slightly in Clusters 1 and 3, the probability of O increases (0.0090 in Cluster 1; 0.0131 in Cluster 3, Table 5). Cluster 4 shows a pronounced rise in O (0.1465) with substantial declines in KA and BC. These results suggest older riders may be more likely to sustain minor rather than severe injuries, possibly reflecting cautious behavior but heightened physical vulnerability.

*Crash Characteristics*

The relationship between crash speed and injury severity was observed to differ across clusters. For crash speed 30–45 mph, negative coefficients for KA were identified in Cluster 1 (–0.330, Table 4), Cluster 3 (–0.329), and Cluster 5 (–0.538), indicating a lower likelihood of KA and BC at moderate speeds within these groups. In Cluster 4, a positive coefficient

(0.279) was noted, corresponding to a higher probability of O, which suggests that moderate speeds in this cluster are more frequently associated with minor outcomes. In contrast, crash speed greater than 65 mph was consistently associated with strong positive coefficients for KA across all clusters, with values of 0.718 in Cluster 1, 0.347 in Cluster 2, 1.035 in Cluster 3, 0.933 in Cluster 4, and 0.693 in Cluster 5, providing clear evidence of a strong association with an elevated risk of KA[54].

These findings from the model estimation are further supported by the marginal effect analysis. For crash speed 30–45 mph, Clusters 1 and 3 exhibited small increases in KA and BC and a decrease in O, while Cluster 4 showed a notable increase in O (0.0588, Table 5) alongside reductions in KA and BC severities, reinforcing the observation that moderate speeds are generally associated with less severe outcomes in this context. For crash speed greater than 65 mph, the marginal effects (see Figure 16) revealed substantial increases in the probability of KA, particularly in Cluster 4 (0.1899), with corresponding declines in BC and O, highlighting the significantly greater risk of KA severity at higher speeds (see Figure 16).

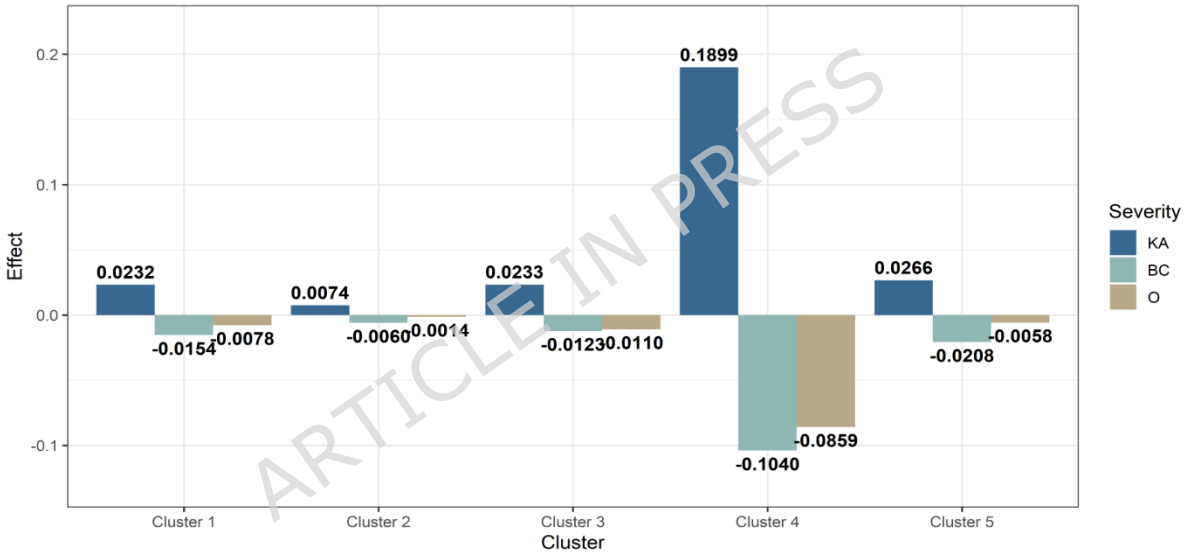



**Figure 16.** Marginal effects of the variable crash speed greater than 65 mph

*Traffic Condition*

The model estimation results indicate that the presence of a center stripe or divider is associated with injury severity only in Cluster 5, where a positive coefficient for BC (0.283, Table 4) was found. Marked lane present is linked to reduced severity across multiple clusters, as reflected by negative coefficients for O severity in Cluster 1 (–0.643), Cluster 2 (–0.509), and Cluster 3 (–0.455suggesting a protective effect for motorcyclists[55]. The presence of stop or yield or warning signs is associated with increased severity in certain clusters, as shown by positive coefficients for KA in Cluster 2 (0.249) and Cluster 4 (0.279).

The marginal effect analysis further clarifies these findings. For center stripe or divider in Cluster 5, the probability of KA (–0.0120) and O (–0.0062) decreases, while BC increases (0.0182, Table 5). For marked

lane present, marginal effects (Figure 17) in Cluster 2 show a slight increase in KA (0.0049) and reductions in BC (–0.0038) and O (–0.0011), while Cluster 4 shows a notable increase in KA (0.0569) and declines in BC (–0.0312) and O (–0.0257). For stop or yield or warning sign presence, Cluster 2 presents a slight increase in KA and reductions in BC and O, while Cluster 4 again demonstrates a pronounced increase in KA (0.0569) with decreases in BC and O.

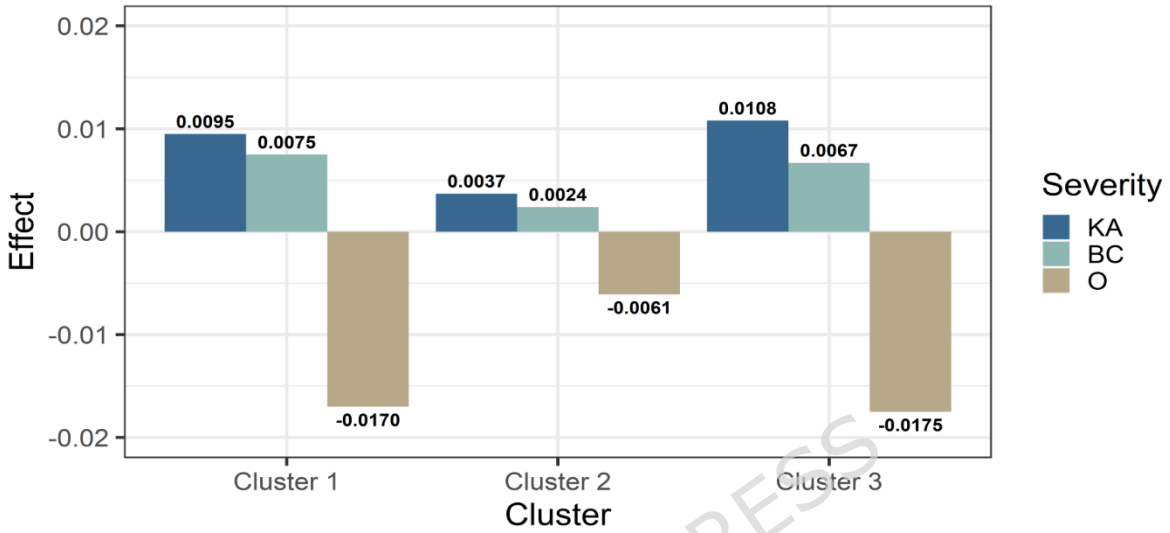



**Figure 17.** Marginal effects of the variable marked lane

*Roadway Characteristics*

The association between roadway alignment and injury severity was observed to vary across clusters. For curve grade, model estimation results indicated a positive coefficient for O severity in Cluster 1 (0.582, Table 4) and a negative coefficient for O severity in Cluster 2 (–0.584), suggesting that curve-graded segments are linked to an increased likelihood of O severity in Cluster 1 and a reduced likelihood of O severity in Cluster 2[66]. For straight level, a positive coefficient for KA severity was found in Cluster 1 (0.549), indicating that straight and level roadways are associated with a higher risk of KA severity in this cluster[54].

Marginal effect analysis provided further insight into these patterns. For curve grade, Cluster 1 exhibited small decreases in the probability of KA severity (–0.0015) and BC severity (–0.0017), along with an increase in the probability of O severity (0.0032, Table 5). In contrast, Cluster 2 showed increases in the probability of KA severity (0.0054) and BC severity (0.0036), with a decrease in the probability of O severity (–0.0090, Figure 18). For straight level in Cluster 1, the marginal effects demonstrated a substantial increase in the probability of KA severity (0.0559), accompanied by reductions in the probability of BC severity (–0.0383) and O severity (–0.0176).

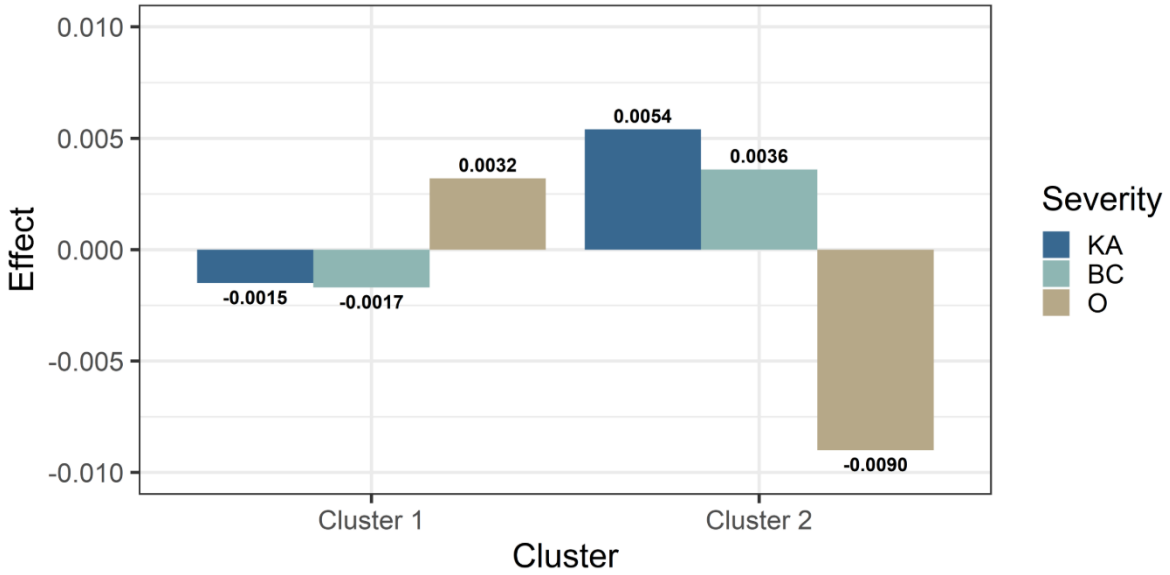



**Figure 18.** Marginal effects of the variable curve-graded roadway

*Temporal Factors*

The relationship between crash injury severity was found to vary across temporal conditions, particularly by time of day and season. For afternoon, model estimation results indicated negative coefficients for O severity in Cluster 1 (–0.373, Table 4), Cluster 2 (–0.370), and Cluster 3 (–0.393), suggesting a reduced likelihood of O severity and generally less severe outcomes for motorcycle crashes occurring in the afternoon (Jafari et al., 2025a). For morning, a positive coefficient for KA severity was identified in Cluster 3 (0.573), indicating that crashes in the morning are more likely to result in KA severity. Similarly, fall was associated with a positive coefficient for KA severity in Cluster 3 (0.427), highlighting greater crash severity in this season.

Marginal effect analysis supported these findings. For afternoon, all clusters showed small increases in the probability of KA severity (0.0040 to 0.0067) and BC severity (0.0023 to 0.0041), but a more notable decrease in the probability of O severity (–0.0057 to –0.0108, Table 5; Figure 19). For morning in Cluster 3, the probability of KA severity increased (0.0203), with decreases for BC severity (–0.0099) and O severity (–0.0104). For fall in Cluster 3, the probability of KA severity also increased (0.0175), with corresponding declines for BC severity (–0.0088) and O severity (–0.0087).

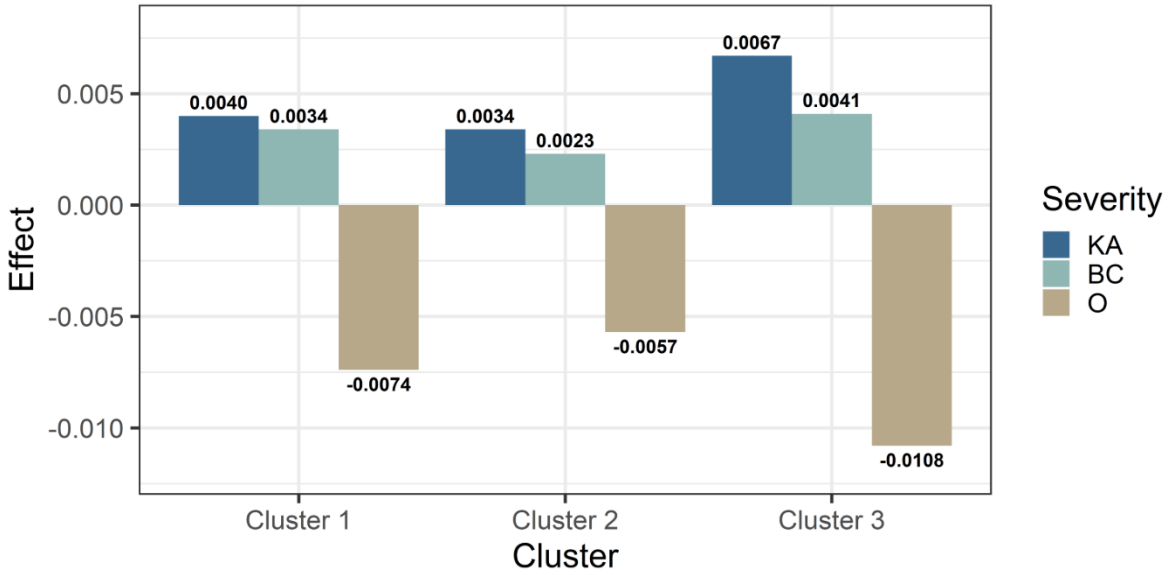



**Figure 19.** Marginal effects of the variable afternoon

*Lighting Condition*

The association between lighting condition and injury severity was observed to be distinct across crash settings. For dark, not lighted conditions, model estimation results indicated negative coefficients for O severity in Cluster 1 (–0.237, Table 4), Cluster 2 (–0.498), Cluster 3 (–0.415), Cluster 4 (–0.545), and Cluster 5 (-1.298), suggesting that crashes in unlit dark conditions are linked to a reduced likelihood of O severity for motorcyclists[66]. In contrast, daylight was associated with positive coefficients for BC severity across all clusters Cluster 1 (0.951), Cluster 2 (1.320), Cluster 3 (0.454), Cluster 4 (0.505), and Cluster 5 (0.671) indicating a greater risk of BC severity during daylight hours[66].

Marginal effect analysis further clarifies these patterns. For dark, not lighted conditions, Cluster 4 exhibited an increase in the probability of KA severity (0.0502) and BC severity (0.0646), but a substantial decrease in the probability of O severity (–0.1148, Table 5). In other clusters, decreases in O severity were also consistent, though of smaller magnitude. For daylight, the marginal effects (Figure 20) showed that in Cluster 2, the probability of KA severity decreased (–0.0632) and in Cluster 4, O severity decreased (–0.0598), while BC severity increased (0.1160). Across other clusters, similar trends were observed, with decreases in KA severity (–0.0313 to –0.0632) and O severity (–0.0034 to –0.0598), and increases in BC severity (0.0358 to 0.1160), confirming that daylight hours are generally associated with a higher probability of BC severity and lower probabilities of KA and O severities.

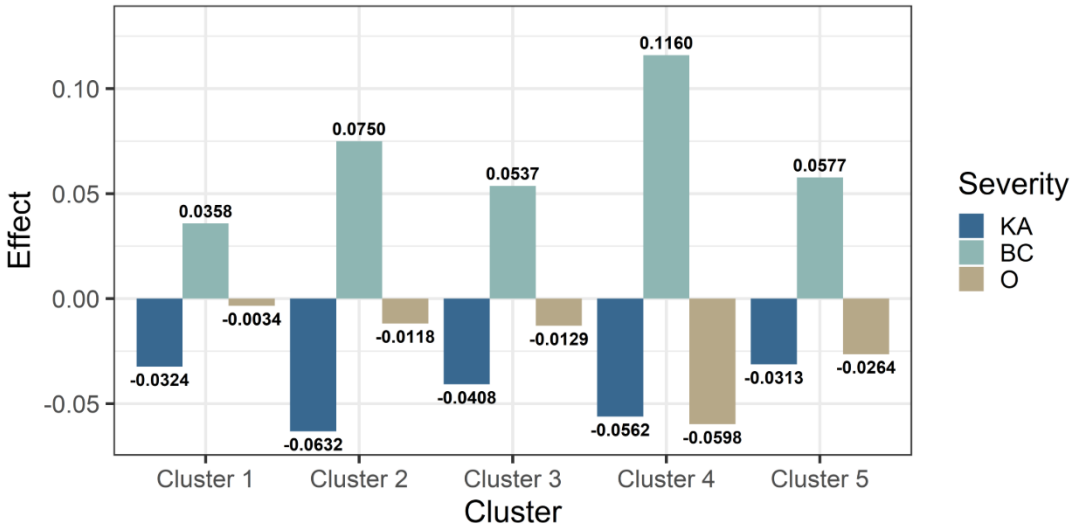



**Figure 20.** Marginal effects of the variable daylight

*First Harmful Event*

The association between first harmful event and injury severity varied by crash type and cluster. For rear-end crashes, a positive coefficient for BC severity was observed in Cluster 3 (0.642, Table 4), indicating a higher likelihood of moderate injuries in these cases. Run-off-road events displayed negative coefficients for KA severity in Cluster 1 (–0.392) and Cluster 3 (–1.868), but a positive coefficient for KA severity in Cluster 2 (0.818). For crashes involving hitting a fixed object, a negative coefficient for BC severity was found in Cluster 3 (–0.885). Overturned crashes exhibited strong positive coefficients for BC severity in Cluster 1 (1.078), Cluster 2 (1.454), and Cluster 5 (0.481), while Cluster 4 had a negative coefficient (–0.956), highlighting variability across clusters[55].

Marginal effect analysis provided further insights. In Cluster 3, rear-end crashes increased the probability of BC severity (0.0236) and reduced KA severity (–0.0181) and O severity (–0.0055, Table 5). For run-off-road, Cluster 1 showed a reduction in KA severity (–0.0481) with increases in BC severity (0.0319) and O severity (0.0162), while Cluster 2 demonstrated an increase in KA severity (0.1660), with declines in BC (–0.1317) and O (–0.0343). Crashes involving fixed objects in Cluster 3 showed a small increase in KA severity (0.0022), a decrease in BC severity (–0.0042), and a slight increase in O severity (0.0020). For overturned crashes, Clusters 1 and 2 revealed notable increases in BC severity (0.1018 and 0.1750) and decreases in KA severity (–0.0711 and –0.1489) and O severity (–0.0307 and –0.0261). In Cluster 4, the effect was reversed with an increase in KA severity (0.1065), a decrease in BC severity (–0.2197), and an increase in O severity (0.1132, Figure 21).

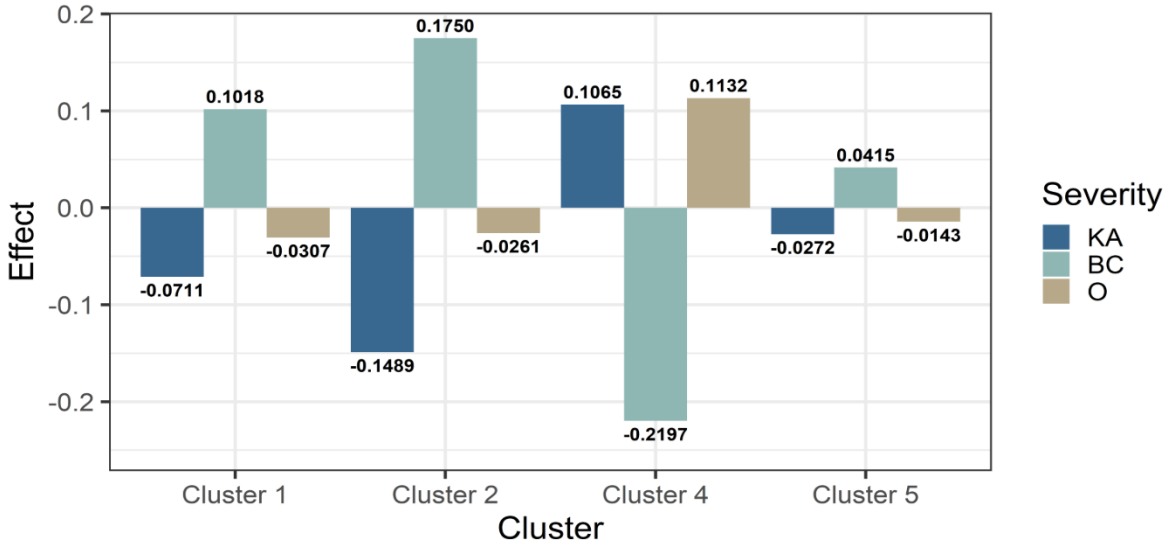



**Figure 21.** Marginal effects of the variable overturned

*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 roads 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 4). 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 4), 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 4.** 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** | | | | | | | | | | |

| 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** |
| 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 | - | - | - | - | - | - | - | - |

| 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 |
| 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 helmet not worn | - | - | - | - | 1.173 | 2.69 | - | - | - | - |
| Effect of no no-passing zone on the mean of random parameter daylight | - | - | - | - | -0.924 | -1.97 | - | - | - | - |
| **Statistics** | | | | | | | | | | |
| Number of observations | 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 5.** 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 | KA | 0.0062 | 0.0066 | - | -0.0590 | - |

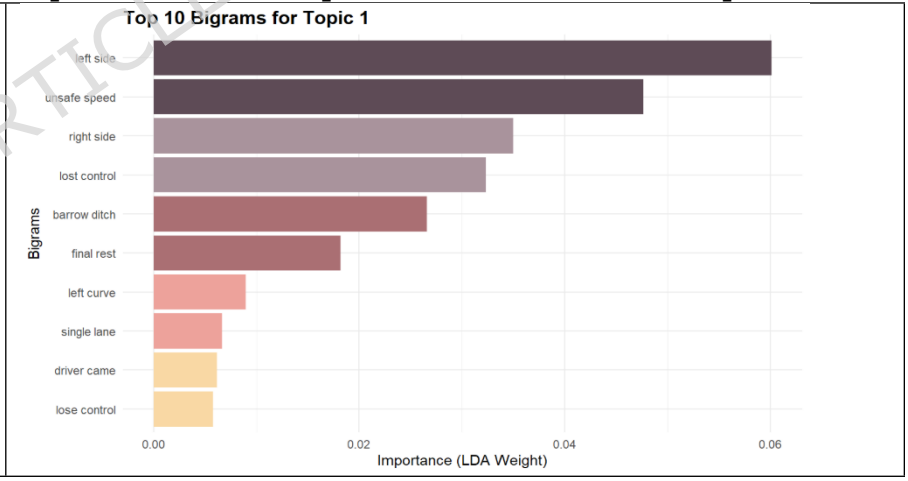
| | | | | | | |
|---|---|---|---|---|---|---|
| otherwise) [O] | BC | 0.0052 | 0.0045 | - | -0.0760 | - |
| | O | -0.0114 | -0.0111 | - | 0.1350 | - |
| **Rider Demographic** | | | | | | |
| Rider age (1 if rider's age is greater than 65 years old, 0 otherwise) [O] | KA | -0.0048 | - | -0.0079 | -0.0641 | - |
| | 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 otherwise) [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 curved 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 | - | - |

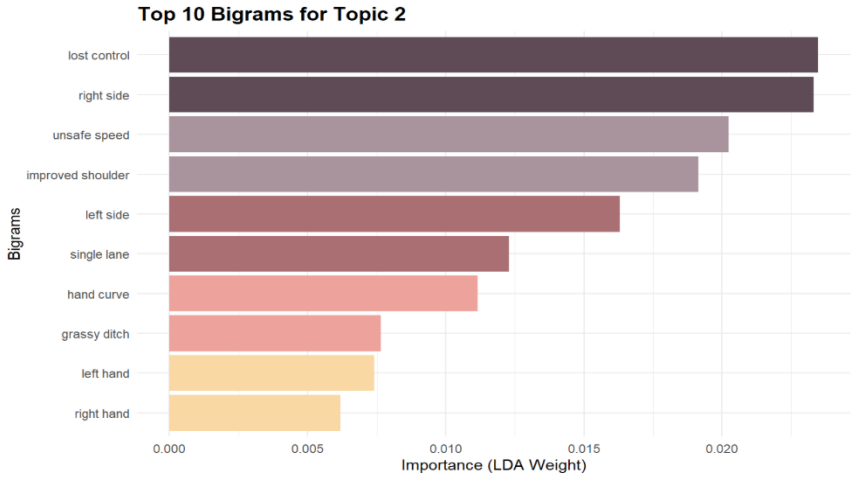
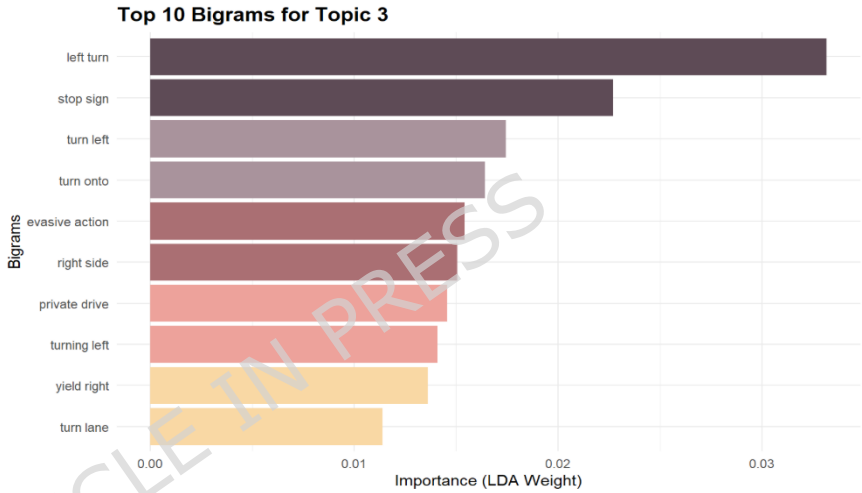
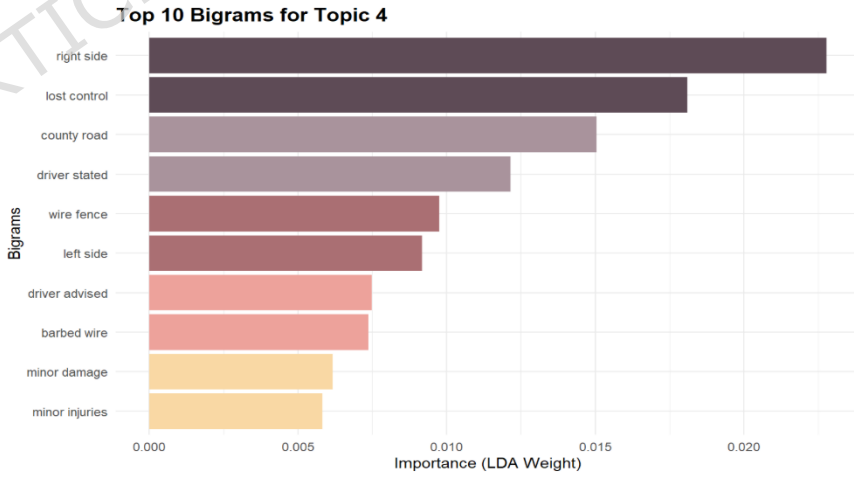
| | | | | | | |
|---|---|---|---|---|---|---|
| Fall (1 if crash happened in fall, 0 otherwise) [KA] | KA | - | - | 0.0175 | - | - |
| | BC | - | - | -0.0088 | - | - |
| | O | - | - | -0.0087 | - | - |
| **Lighting Condition** | | | | | | |
| Dark, not lighted (1 if crash happened in dark, not lighted 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 |

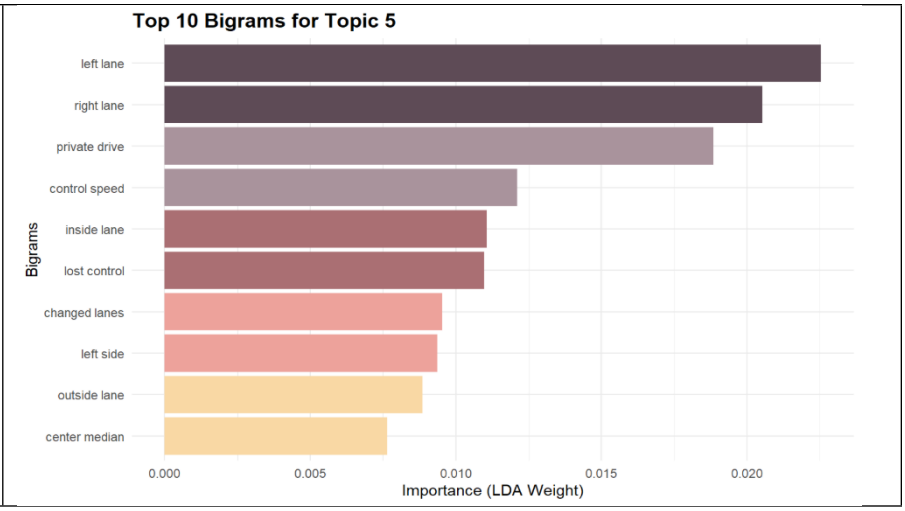
**Advanced Topic Modeling on Crash Narratives**

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 as illustrated in Table 6. Topic 1 highlights incidents where excessive speed and subsequent loss of control often on curves led vehicles to leave the roadway, emphasizing the dangers of poor speed management. Building on this, Topic 2 focuses on single-vehicle crashes, especially on narrow or winding roads, where drivers' inability to adapt to challenging geometry or surface conditions frequently results in run-off-road events, reinforcing the role of roadway design and individual driver response. Topic 3 shifts the focus to intersection areas, where turning maneuvers, failures to yield, and improper entries or exits are common causes of crashes, underlining persistent intersection safety challenges. Moving to less urbanized environments, Topic 4 captures crashes occurring on rural county roads, where loss of control often leads to contact with roadside objects such as fences, typically causing property damage and minor injuries. Finally, Topic 5 brings attention to multi-lane settings and driveways, where crashes are frequently linked to lane-change maneuvers and inadequate speed control, illustrating the risks created by complex traffic movements and higher speed.

**Table 6.** Five bigram topic models from police crash narrative report

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

| | |
|---|---|
| Topic 2: Single-Vehicle Roadway Departure |   |
| Topic 3: Turn-Related Crash at Intersection |   |
| Topic 4: Run-Off-Road Crashes on County Roads |   |

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

### Convergent Validity and Policy Implications

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 roads. 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 SSA. Table 7 is 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 7.** Comparative Table: Clusters, Model Results, and Narrative Themes

| Cluster Findings | Marginal Effects | LDA Topics |
|---|---|---|
| **C1: Overturns and run-off-road crashes at high speeds**<br>Overturned vehicles, run-off-road events, high speeds, other contributing factors, straight level roads. | **Helmet not worn [KA]:** 0.0625 ↑<br>**Crash speed >65 mph [KA]:** 0.0232 ↑<br>**Run-off-road [KA]:** -0.0481 ↓<br>**Straight, level [KA]:** 0.0559 ↑<br>**Overturned [BC]:** 0.1018 ↑<br>**Marked lane [O]:** -0.0170 ↓<br>**Afternoon [O]:** -0.0074 ↓<br>**Daylight [BC]:** 0.0358 ↑ | **Topic 1:** Speeding-related roadway departure<br>**Topic 2:** Single-vehicle roadway departure |
| **C2: Crashes with fixed objects on curves at unsafe speeds**<br>Curve, grade alignments, high speed, run-off-road into fixed objects, unsafe speed major factor. | **Helmet not worn [KA]:** 0.0710 ↑<br>**Crash speed >65 mph [KA]:** 0.0074 ↑<br>**Run-off-road [KA]:** 0.1660 ↑<br>**Curve, grade [KA]:** 0.0054 ↑<br>**Marked lane [O]:** -0.0061 ↓<br>**Overturned [BC]:** 0.1750 ↑<br>**Afternoon [O]:** -0.0057 ↓<br>**Daylight [BC]:** 0.0750 ↑ | **Topic 2:** Single-Vehicle Roadway Departure<br>**Topic 1:** Speeding-related roadway departure |
| **C3: Severe straight road crashes with riding control issues**<br>Straight road segments, fixed object, trees or obstacles, mix of crash types related to turning issues. | **Helmet not worn [KA]:** 0.0373 ↑<br>**Crash speed >65 mph [KA]:** 0.0233 ↑<br>**Run-off-road [KA]:** -0.0292 ↓<br>**Morning [KA]:** 0.0203 ↑<br>**Rear-end [BC]:** 0.0236 ↑<br>**Daylight [BC]:** 0.0537 ↑ | **Topic 3:** Turn-Related Crash at intersection<br>**Topic 5:** Lane-change and speed control conflicts |

| | | |
|---|---|---|
| | **Dark, not lighted [KA]:** 0.0050 ↑ | |
| **C4: Crashes at intersections under dark, lighted conditions**<br>Unsafe speed, signalized intersections, angle/head-on, complex geometry, non-passing zones, fixed objects. | **Helmet worn, damaged [KA]:** -0.0260 ↓<br>**Crash speed >65 mph [KA]:** 0.1899 ↑<br>**Unsafe speed [KA]:** -0.0590 ↓<br>**Crash speed 30–45 mph [KA]:** -0.0257 ↓<br>**Rider age >65 [KA]:** -0.0641 ↓<br>**Overturned [KA]:** 0.1065 ↑<br>**Daylight [BC]:** 0.1160 ↑<br>**Dark, not lighted [KA]:** 0.0502 ↑ | **Topic 3:** Turn-related crash at intersection |
| **C5: Run-off-road crashes at night with animals present**<br>Small cluster, dark roads, animals, run-off-road, unsafe speed, failed yielding/signaling. | **Helmet not worn [KA]:** 0.0261 ↑<br>**Crash speed >65 mph [KA]:** 0.0266 ↑<br>**Daylight [BC]:** 0.0577 ↑<br>**Dark, not lighted [KA]:** 0.0191 ↑<br>**Center divider [BC]:** 0.0182 ↑ | **Topic 4:** Run-off-road crashes on county roads |

The combined findings from the cluster analysis, marginal effects modeling, and LDA topic modeling consistently show that excessive speed, loss of control, poor roadway geometry, low lighting, and helmet non-use are central to severe motorcycle crashes on rural undivided roads. High speeds and run-off-road events, often involving overturns and fixed object collisions, are particularly prominent in the most severe crash clusters, as highlighted by both the statistical models and narrative themes [67,68]. Intersection-related crashes, especially under dark or poorly lit conditions, further increase risk due to inadequate lighting, insufficient channelization, and improper yielding or turning behaviors. Animal-related crashes at night present additional dangers, with poor visibility and frequent wildlife crossings significantly raising crash severity for motorcyclists[69].

Addressing these risks requires a Safe System Approach (SSA) tailored to rural settings, as recommended by USDOT [70]. Key interventions include setting context-sensitive speed limits, targeted enforcement, and geometric safety improvements like rumble strips and dynamic warning signs[68,71]. Road upgrades, such as expanding clear zones, installing energy-absorbing barriers, and improving pavement markings and intersection design should be prioritized[72,73]. Education campaigns focused on helmet use, hazard perception, and animal avoidance, along with helmet rebate programs and roadside safety checks, can further reduce risks[74]. Enhancing post-crash response and partnering with wildlife agencies to implement fencing and animal detection systems will help address the unique challenges found in

rural motorcycle crashes[69].

## Conclusions

The persistent challenge of motorcycle crash severity on rural undivided roads 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 roads.

The integrated analysis revealed that motorcycle crashes resulting in severe or fatal injuries on rural undivided roads are shaped by a combination of speed-related behaviors, roadway conditions, and rider actions. Excessive speeds, particularly those exceeding 65 mph, run-off-road incidents, and overturns emerged as leading contributors to crash severity, often occurring on straight stretches or in areas with fixed roadside objects. The risk of serious injury was further heightened when riders were not wearing helmets, traveled at night or in poorly lit conditions, or encountered hazardous roadway features such as sharp curves or steep grades. Crashes at intersections and during turning maneuvers also proved especially dangerous, particularly when drivers failed to yield, traffic signals were inadequate, or visibility was low. Analysis of narrative data added further depth, showing that animal crossings, sudden loss of control, and common speeding behaviors are frequent and significant threats in rural environments.

This analytical framework was structured to explicitly account for the fact that motorcycle crash injury severity on rural undivided roads is not governed by a single, uniform risk process. The cluster analysis revealed that crashes occur within distinct clusters characterized by different combinations of speed, roadway geometry, lighting, and conflict type. Afterwards, sequential estimation using MNL, RPL, and RPLHM formulations allowed the analysis to systematically reduce assumptions of parameter homogeneity and to empirically assess whether additional model flexibility was justified. Performance comparisons based on log-likelihood and AIC showed that clusters associated with complex road geometry (e.g., curved roads), high-speed, and run-off road/loss-of-control scenarios required models that explicitly captured both unobserved heterogeneity in risk effects, whereas clusters reflecting more constrained environments were adequately represented by simpler structures. Importantly, the consistency between cluster-specific econometric results and independently derived LDA topics from police narratives provides external validation of the structured data findings, reinforcing that the identified performance differences reflect

substantive behavioral and environmental distinctions rather than specification-driven effects.

This study has several limitations that should be considered when interpreting the findings. First, the analysis is based exclusively on police-reported crash data from the Texas CRIS database, and the results may not be directly transferable to other states or regions with different roadway designs, traffic compositions, enforcement practices, or reporting standards, introducing potential sources of contextual bias. Second, the mixed logit models rely on variables selected through cluster correspondence analysis and machine-learning–based importance ranking; while this enhances model interpretability, some relevant but unobserved factors may remain excluded. Third, the study does not explicitly model temporal dynamics or post-crash processes, limiting insights into how injury severity relationships may evolve over time. Additionally, detailed roadside safety features such as specific barrier and guardrail types could not be examined due to data constraints, and the focus on rural undivided roads further limits generalizability to urban or divided roadway environments.

**Acknowledgments**

This study utilized crash data obtained from the Texas Crash Records Information System (CRIS)[75], which provided the necessary information to analyze and interpret injury severity outcomes.

**Data availability statement**

The data supporting the findings of this study are available from the corresponding author upon reasonable request.



**Declaration of conflicting interests**

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

No external funding was received to support this research.

**References**

1. NHTSA. Traffic Safety Facts, 2022 Data | Motocycles. *NHTSA, Washington D.C.* (2024).
2. IIHS-HLDI. Fatality Facts 2023: Motorcycles and ATVs. *IIHS-HLDI crash testing and highway safety* (2025).
3. TxDOT. C.R.I.S. Query: Crash Records Information System. *TxDOT Crash Query Tool* (2024).
4. Islam, M. The effect of motorcyclists' age on injury severities in single-motorcycle crashes with unobserved heterogeneity. *Journal of Safety Research* **77**, 125–138 (2021).

5. Goodwin, A. H., Wang, Y. C., Foss, R. D. & Kirley, B. The role of inexperience in motorcycle crashes: Crash trends of novice and returning motorcycle riders. *Journal of Safety Research* **82**, pp 371-375 (2022).
6. Phillips, J., Walford, N., Hockey, A., Foreman, N. & Lewis, M. Older people and outdoor environments: Pedestrian anxieties and barriers in the use of familiar and unfamiliar spaces. *Geoforum* **47**, 113–124 (2013).
7. Christian, J. M., Thomas, R. F. & Scarbecz, M. The Incidence and Pattern of Maxillofacial Injuries in Helmeted Versus Non-Helmeted Motorcycle Accident Patients. *Journal of Oral and Maxillofacial Surgery* **72**, 2503–2506 (2014).
8. De Rome, L. *et al.* Effectiveness of motorcycle protective clothing: Riders' health outcomes in the six months following a crash. *Injury* **43**, 2035–2045 (2012).
9. Carvalho, H. B. D. *et al.* 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 (2016).
10. Sarmiento, J. M. *et al.* Alcohol/Illicit Substance Use in Fatal Motorcycle Crashes. *Journal of Surgical Research* **256**, 243–250 (2020).
11. Maistros, A., Schneider, W. H. & Savolainen, P. T. A comparison of contributing factors between alcohol related single vehicle motorcycle and car crashes. *Journal of Safety Research* **49**, 129.e1-135 (2014).
12. Islam, S. & Brown, J. A comparative injury severity analysis of motorcycle at-fault crashes on rural and urban roadways in Alabama. *Accident Analysis & Prevention* **108**, pp 163-171 (2017).
13. Haque, M. M., Chin, H. C. & Debnath, A. K. An investigation on multi-vehicle motorcycle crashes using log-linear models. *Safety science* https://www.sciencedirect.com/science/article/pii/S0925753511002608?casa_token=3r75a1Vi0g4AAAAA:CbrLCRce1ewT_vBQqmd7AeWdEp_nSuCMCdWKH6G3iC_yvTIukD-Qnv357IH2_7bKzze0p_QkuA (2012).
14. Schneider, W. H. & Savolainen, P. T. Comparison of Severity of Motorcyclist Injury by Crash Types. *Transportation Research Record: Journal of the Transportation Research Board* **2265**, 70–80 (2011).
15. Das, S., Dzinyela, R., Liu, J., Dadashova, B. & Silvestri-Dobrovolny, C. Understanding patterns of factor influences in motorcycle crashes with fixed objects. *Journal of Transportation Safety & Security* 1–27 (2024) doi:10.1080/19439962.2024.2429077.
16. Abdulrazaq, M. A. & Fan, W. (David). A priority based multi-level heterogeneity modelling framework for vulnerable road users. *Transportmetrica A: Transport Science* **0**, 1–34 (2025).
17. Abdulrazaq, M. A. & Fan, W. (David). Seasonal instability in the determinants of vulnerable road user crashes: a partially temporally constrained modeling approach. *Accident Analysis & Prevention* **224**, 108277 (2026).
18. Abdulrazaq, M. A. & Fan, W. D. Temporal dynamics of pedestrian injury severity: a seasonally constrained random parameters approach.

*International Journal of Transportation Science and Technology* **20**, 237–257 (2025).
19. Khan, G., Bill, A. R., Chitturi, M. V. & Noyce, D. A. Safety Evaluation of Horizontal Curves on Rural Undivided Roads. *Transportation Research Record: Journal of the Transportation Research Board* **2386**, 147–157 (2013).
20. Xin, C., Wang, Z., Lin, P.-S., Lee, C. & Guo, R. 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 (2017).
21. Das, S., Mousavi, S. M. & Shirinzad, M. Pattern recognition in speeding related motorcycle crashes. *Journal of Transportation Safety & Security* **14**, 1121–1138 (2022).
22. Lemonakis, P., Eliou, N. & Karakasidis, T. 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 (2021).
23. Chakraborty, R., Das, S., Mimi, M. S. & Kutela, B. 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 (2024) doi:10.1177/03611981241297976.
24. Das, S. Identifying key patterns in motorcycle crashes: findings from taxicab correspondence analysis. *Transportmetrica A: Transport Science* **17**, 593–614 (2021).
25. Rahman, M. A., Das, S. & Sun, X. Understanding the Drowsy Driving Crash Patterns from Correspondence Regression Analysis. *Journal of Safety Research* **84**, 167–181 (2023).
26. Das, S., Avelar, R., Dixon, K. & Sun, X. Investigation on the wrong way driving crash patterns using multiple correspondence analysis. *Accident Analysis & Prevention* **111**, 43–55 (2018).
27. Liu, S., Lin, Z. & Fan, W. (David). 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 (2022).
28. Das, S. & Sun, X. Exploring Clusters of Contributing Factors for Single-Vehicle Fatal Crashes Through Multiple Correspondence Analysis. in (2014).
29. Das, S., Dutta, A. & Tsapakis, I. Topic Models from Crash Narrative Reports of Motorcycle Crash Causation Study. *Transportation Research Record: Journal of the Transportation Research Board* **2675**, pp-449-462 (2021).
30. Kutela, B., Shita, H., Mbuya, C. & Chimba, D. Prediction of sensor damage in automated vehicle involved collisions using parametric and non-parametric approaches. *International Journal of Crashworthiness* **30**, 337–349 (2025).

31. Levy, K. E. C. & Franklin, M. Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry. *Social Science Computer Review* **32**, 182–194 (2014).
32. Zhao, X., Zhan, M. & Jie, C. Examining multiplicity and dynamics of publics' crisis narratives with large-scale Twitter data. *Public Relations Review* **44**, 619–632 (2018).
33. Anastasopoulos, P. Ch. & Mannering, F. L. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. *Accident Analysis & Prevention* **43**, 1140–1147 (2011).
34. Eluru, N., Bhat, C. R. & Hensher, D. A. *A Mixed Generalized Ordered Response Model for Examining Pedestrian and Bicyclist Injury Severity Level in Traffic Crashes*. (University of Sydney, 2007).
35. Xin, C., Wang, Z., Lee, C. & Lin, P.-S. 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 (2017) doi:10.3141/2637-05.
36. Das, S., Sakib, N., Geedapally, S. & Wei, Z. Understanding Pedestrian Hit and Run Crash Patterns using Louisiana Data. *Transportation Safety and Environment* https://doi.org/10.1093/tse/tdaf039 (2025) doi:10.1093/tse/tdaf039.
37. Islam, M. The effect of motorcyclists' age on injury severities in single-motorcycle crashes with unobserved heterogeneity. *Journal of Safety Research* **77**, 125–138 (2021).
38. Das, S., Jafari, M., Hossain, A., Chakraborty, R. & Mimi, M. S. Toll road crash severity using mixed logit model incorporating heterogeneous mean structures. *Transportmetrica A: Transport Science* 1–21 (2024) doi:10.1080/23249935.2024.2343755.
39. Ukkusuri, S., Miranda-Moreno, L. F., Ramadurai, G. & Isa-Tavarez, J. The Role of Built Environment on Pedestrian Crash Frequency. *Safety Science* **50**, 1141–1151 (2012).
40. Hossain, A., Barua, S., Das, S. & Starewich, M. 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 (2025) doi:10.1080/23249935.2025.2481578.
41. Wang, M.-H. Investigating the Difference of Factors Contributing to Motorcyclist Fatalities in Single Motorcycle and Multiple Vehicle Crashes in Taiwan. in 17p (2022).
42. Kim, Y., Yeo, H., Lim, L. & Noh, B. Integrating visual and community environments in a motorcycle crash and casualty estimation. *Accident Analysis & Prevention* **208**, (2024).
43. Markos, A., D'Enza, A. I. & Van De Velden, M. Beyond Tandem Analysis: Joint Dimension Reduction and Clustering in R. *J. Stat. Soft.* **91**, (2019).

44. Chakraborty, R., Mills, D. & Das, S. Children on wheels: Identifying crash determinants using cluster correspondence analysis. *Accident Analysis & Prevention* **216**, (2025).
45. Chakraborty, R., Das, S., Mimi, M. S. & Kutela, B. Investigating Factor Associations in Barrier Crashes through Cluster Correspondence Analysis. *Transportation Research Record* **2679**, 860–879 (2025).
46. Rahman, M. A. *et al.* Exploring Attribute Associations in Pedestrian-Involved Hit-and-Run Crashes through Cluster Correspondence Analysis. *Transportation Research Record* **2678**, 1079–1098 (2024).
47. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at https://doi.org/10.48550/arXiv.1705.07874 (2017).
48. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research* (2003).
49. Jafari, M. *et al.* Assessing motorcyclist injury severity on curved road segments with temporal dynamics and unobserved heterogeneity. *Sci Rep* **15**, 13110 (2025).
50. Behnood, A. & Mannering, F. 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 (2017) doi:10.1016/j.amar.2017.04.001.
51. Das, S., Jafari, M., Dzinyela, R. & Khan, M. N. Applying hybrid dimension reduction and econometric model to investigate rider behaviors in roadway departure motorcycle crashes. *Transportation Letters* 1–21 (2025) doi:10.1080/19427867.2025.2516422.
52. Das, S., Tamakloe, R., Zubaidi, H., Obaid, I. & Ashifur Rahman, M. Bicyclist injury severity classification using a random parameter logit model. *International Journal of Transportation Science and Technology* **12**, 1093–1108 (2023).
53. Hossain, A., Sun, X., Das, S., Jafari, M. & Codjoe, J. 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 (2024).
54. Jafari, M., Das, S., Barua, S., Mimi, M. S. & Starewich, M. Crash outcomes of yellow school buses: Random parameter and correlated random parameter logit models with heterogeneity in means. *Accident Analysis & Prevention* **219**, 108109 (2025).
55. Jafari, M. *et al.* Assessing motorcyclist injury severity on curved road segments with temporal dynamics and unobserved heterogeneity. *Scientific Reports* **15**, 13110 (2025).
56. Chen, T. *et al.* xgboost: Extreme Gradient Boosting. (2024).
57. Ashifur Rahman, M., Das, S. & Sun, X. Using Cluster Correspondence Analysis to Explore Rainy Weather Crashes in Louisiana. *Transportation Research Record* **2676**, 159–173 (2022).

58. Das, S., Tabesh, M., Dadashova, B. & Dobrovolny, C. Diagnosis of Encroachment-Related Work-Zone Crashes by Applying Pattern Recognition. *Transportation Research Record: Journal of the Transportation Research Board* **2677**, 222–236 (2023).
59. The R Foundation. R: The R Project for Statistical Computing. https://www.r-project.org/ (2024).
60. Effati, M. & Ramezanpoor, A. 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 (2025).
61. Arnadottir, A. T., Ulfarsson, G. F. & Kim, S. 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 (2025).
62. Das, S., Dzinyela, R., Liu, J., Dadashova, B. & Silvestri-Dobrovolny, C. Understanding patterns of factor influences in motorcycle crashes with fixed objects. *Journal of Transportation Safety & Security* **17**, 483–509 (2025).
63. Kutela, B., Mihayo, M. P., Khalaf, H. M. M. B., Kidando, E. & Kitali, A. E. Investigating the influence of lighting conditions on pre-crash vulnerable road users' visibility. *Advances in Transportation Studies* **65**, 103–118 (2025).
64. Song, L. *et al.* 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 (2025).
65. Kang, L., Vij, A., Hubbard, A. & Shaw, D. The unintended impact of helmet use on bicyclists' risk-taking behaviors. *Journal of Safety Research* **79**, 135–147 (2021).
66. Ye, Y., He, J., Yan, X., Wang, C. & Qin, P. 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 (2025).
67. Gross, F., Jovanis, P. P., Eccles, K. & Chen, K.-Y. *Safety Evaluation of Lane and Shoulder Width Combinations on Rural, Two-Lane, Undivided Roads - FHWA-HRT 09-031*. https://www.fhwa.dot.gov/publications/research/safety/09031/index.cfm (2009).
68. Liu, C. Design Directional Raised Rumble Aggregates and Strips for Awakening Wrong-Way Drivers. *International Journal of Transportation Science and Technology* **4**, 151–156 (2015).
69. Grace, M. K., Smith, D. J. & Noss, R. F. Reducing the threat of wildlife-vehicle collisions during peak tourism periods using a Roadside Animal Detection System. *Accident Analysis & Prevention* **109**, 55–61 (2017).

70. USDOT. What Is a Safe System Approach? | US Department of Transportation. https://www.transportation.gov/safe-system-approach (2023).
71. NHTSA. Speed Safety Camera Enforcement | NHTSA. https://www.nhtsa.gov/book/countermeasures-that-work/speeding-and-speed-management/countermeasures/enforcement/speed-safety-camera-enforcement (2024).
72. Dobrovolny, C. S. *et al. Addressing the Motorcyclist Advisory Council Recommendations: Synthesis on Barrier Design for Motorcyclists Safety*. https://rosap.ntl.bts.gov/view/dot/56065 (2021).
73. FHWA. Pavement marking safety study | FHWA. https://highways.dot.gov/safety/other/visibility/benefits-pavement-markings-renewed-perspective-based-recent-and-ongoing-5 (2024).
74. NHTSA. Promote Bicycle Helmet Use with Education | NHTSA. https://www.nhtsa.gov/book/countermeasures-that-work/bicycle-safety/countermeasures/other-strategies-behavior-change (2024).
75. TXDOT. CRIS Query. https://cris.dot.state.tx.us/public/Query/app/home (2025).