

Investigating the Effect of Road Characteristics on Pedestrian and Bike Crash Frequency

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAI'I AT MĀNOA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE
IN
CIVIL ENGINEERING

April 2024

By

Shiva Azimi

Thesis Committee:

Roger B. Chen, Ph.D., Chair

A. Ricardo Archilla, Ph.D.

Jonghyun "Harry" Lee, Ph.D.

Keywords: Traffic Safety, Pedestrian Crashes, Bike Crashes, Road Alignments, Negative
Binomial Regression, Zero-Inflated Negative Binomial Regression

ACKNOWLEDGEMENTS

I would like to express my gratitude to my family and my soulmate, Soroush, for their unwavering support and encouragement throughout every facet of my life, particularly during my educational journey.

Additionally, I am deeply grateful for the invaluable support and guidance provided by Dr. Chen, my committee chair, whose mentorship has been instrumental in both my master's program and the completion of my thesis. I would also like to express my sincere appreciation to Dr. Archilla and Dr. Lee for their service on my committee, as well as their ongoing assistance and counsel during my academic endeavors.

ABSTRACT

The World Health Organization's 2023 report highlights a critical global concern, revealing that vulnerable road users, including pedestrians and cyclists, account for over half of all road traffic fatalities. This alarming statistic underscores the imperative need for enhanced pedestrian and bicycle safety. Road alignment emerges as a pivotal factor influencing the frequency of crashes involving these groups. This thesis examines the relationship between road alignments and features and incidents involving pedestrians and bikes in Oahu, Hawaii, from 2015 to 2022. Structured into two primary sections, the study first develops a methodology for road segmentation based on road alignments, leveraging GPS data and. Subsequently, it employs both Negative Binomial regression and Zero-Inflated Negative Binomial regression analyses to explore the association between segment-level horizontal and vertical road alignments, alongside other road features, and the frequency of crashes involving pedestrians and cyclists. This approach aims to shed light on how specific road configurations contribute to the road safety challenges faced by vulnerable user groups. Key findings indicate that both extremely sharp curves and straight segments elevate crash risks due to navigation difficulty and decreased vigilance respectively. While steeper grades initially decrease crashes, further increases in steepness can escalate crash frequency due to heightened navigational challenges, potentially overriding the cautionary effect observed with moderate steepness. These findings highlight the complex interplay between road geometry and crash incidents. Moreover, higher traffic volumes and more lane numbers emphasizes the heightened risk to pedestrians and bicyclists on roads probably due to increased interactions and challenges of crossing and maneuvering through multi-lane roads. Road segments with higher speed limits tend to have a lower incidence of crashes, likely due to inherent safety features designed for higher speeds and lower speeds in areas frequented by pedestrians and bicyclists. The findings of this study can aid planners and policymakers in enhancing pedestrian and cyclist safety measures.

Table of Contents

ACKNOWLEDGEMENTS	ii
ABSTRACT	iii
LIST OF TABLES	vi
LIST OF FIGURES	vii
Chapter 1: INTRODUCTION AND LITERATURE REVIEW	1
1.1 Factors affecting severity	2
1.2 Factors affecting frequency	2
1.2.1 Factors affecting bike crash frequency	2
1.2.2 Factors affecting pedestrian crash frequency	3
1.3 Extracting Road Alignments	4
1.4 Main Contributions	6
Chapter 2: DATASET	8
2.1 Introduction to the Dataset	8
2.2 Data Sources	8
2.2.1 Crash Dataset	10
2.2.2 Road Features Dataset	11
2.2.3 GPS Dataset	12
2.3 Preprocessing GPS Points	14
2.3.1 Savitzky-Golay Filter	14
2.4 Processing GPS Points	17
2.4.1 Horizontal Alignment	18
2.4.2 Vertical Alignment	21
2.4.3 Homogeneous segments	22
2.5 Descriptive Analysis of the Aggregated Dataset	31
Chapter 3: METHODOLOGY	38
3.1 Crash Frequency Count Models	39
3.1.1 Generalized Linear Models of Count Data	39
Chapter 4: RESULTS	45
4.1 Model Results	47

4.1.1 Negative Binomial Regression Model Results	48
4.1.2 Zero-Inflated Negative Binomial Regression Model Results.....	60
4.1.3 Negative Binomial and Zero-Inflated Negative Binomial Comparison.....	67
Chapter 5: CONCLUSIONS AND RECOMMENDATIONS	70
5.1 Conclusions	70
5.2 Practical Implementation.....	71
5.3 Limitation and Future Direction.....	72
Chapter 6: REFERENCES.....	74

LIST OF TABLES

Table 1: Overview of Dataset Sources and Key Variables	9
Table 2: Descriptive statistics of the continuous variables	32
Table 3: Estimation Result of Horizontal Alignment Parameters in Negative Binomial Model 1	49
Table 4: Estimation Result of Horizontal and Vertical Alignment Parameters in Negative Binomial Model 2	50
Table 5: Estimation Result of Road Features along with Horizontal and Vertical Alignment Parameters in Negative Binomial Model 3.....	51
Table 6: Statistical Evaluation of Negative Binomial Model Enhancements through Likelihood Ratio Tests	60
Table 7: Estimation Result of Horizontal Alignment Parameters in Zero-Inflated Negative Binomial Model 1	61
Table 8: Estimation Result of Horizontal and Vertical Alignment Parameters in Zero-Inflated Negative Binomial Model 2.....	62
Table 9: Estimation Result of Road Features along with Horizontal and Vertical Alignment Parameters in Zero Inflated Negative Binomial Model 3	63
Table 10: Statistical Evaluation of Zero-Inflated Negative Binomial Model Enhancements through Likelihood Ratio Tests	67
Table 11: Model AIC and BIC Statistics Comparison.....	69

LIST OF FIGURES

Figure 1: Flowchart of Crash Frequency Data Analysis and Modeling	7
Figure 2: Comprehensive Map of Pedestrian and Bicycle Crash Incidents on Oahu (2015-2022)	11
Figure 3: Coverage of Road Network with Gathered GPS Data	13
Figure 4: Original vs. Smoothed Road Points Using the Savitzky-Golay Filter on Atkinson Drive	17
Figure 5: Bearing between two consecutive GPS points. lat and lon represents the latitude and longitude of a GPS point.....	19
Figure 6: Road Segmentation Visualization: a) Original GPS Points, b) Classified Points before Smoothing, c) Classified Points after Smoothing	24
Figure 7: Analyzed Road Segments Visualization	25
Figure 8: A sketch illustrating the importance of considering the standard deviations. a) segment with a lower standard deviation of absolute average curvature, and b) same absolute average curvature – higher standard deviation (plan view)	26
Figure 9: Proportion of Road Segments by Curve Class	33
Figure 10: Proportion of Total Road Length by Curve Class	33
Figure 11: Proportion of Crash Counts by Curve Class.....	34
Figure 12: Proportion of Road Segments by Grade Class	34
Figure 13: Proportion of Total Road Length by Grade Class	35
Figure 14: Proportion of Crash Counts by Grade Class.....	35
Figure 15: Distribution of Segment Combinations across Curve and Grade Classes.....	36
Figure 16: Aggregate Length of Segments by Curve and Grade Class Combination	36
Figure 17: Crash Count Distribution by Curve and Grade Class Combinations	37
Figure 18: Proportion of road segments by number of crashes	39
Figure 19: Correlation Matrix Heatmap of Road Segment Features	46
Figure 20: Crash Coefficients (NB Model 3) Across Road Function Classes.....	58

Chapter 1: INTRODUCTION AND LITERATURE REVIEW

Road safety remains a paramount public concern globally and in the United States, necessitating sustained research and policy initiatives aimed at mitigating risks and protecting all road users. The World Health Organization's Global Status Report on Road Safety 2023 [1] starkly highlights the global challenge, reporting approximately 1.19 million fatalities annually due to road traffic crashes. Notably, the WHO emphasizes that over half of all road traffic deaths involve vulnerable road users, including pedestrians, cyclists, and motorcyclists, further accentuating the urgency for targeted safety intervention.

Transitioning to the national perspective, the United States mirrors this urgent public health challenge. According to the National Highway Traffic Safety Administration (NHTSA) [2], the year 2021 witnessed significant traffic-related fatalities, with 39,508 fatal crashes reported, including 7,388 pedestrian and 966 bicyclist deaths, marking the highest number since 2005 and the largest annual percentage increase in the Fatality Analysis Reporting System (FARS) history. This data represents a distressing trend, particularly for non-motorized road users who face heightened risks in traffic environments. The pedestrian fatalities in 2021 marked a 13% increase from the previous year, accompanied by over 60,000 pedestrian injuries nationwide, pointing to a critical need for enhanced protective measures and safety awareness. Similarly, the slight rise in bicyclist fatalities from 938 in 2020 to 966 in 2021 indicates persistent safety challenges for cyclists sharing the road with motorized vehicles.

Research on pedestrian and bicycle crash analysis predominantly divides into two main groups: one focusing on the factors influencing the severity of these incidents and the other on the factors affecting their frequency.

1.1 Factors affecting severity

A comprehensive body of research delves into identifying critical determinants that influence the severity of crashes involving pedestrians and cyclists. Prominent among these factors are the age and gender of both the driver and the pedestrian or cyclist [3], [4], meteorological conditions [3], [5], [6], illumination [7], [8], the built environment [9], [10], infrastructure for bicycles and pedestrians [11], [12], and road alignment [13], [14]. These elements have received considerable attention in scholarly investigations.

1.2 Factors affecting frequency

1.2.1 Factors affecting bike crash frequency

Extensive research has been conducted on the factors influencing the frequency of crashes between bicycles and motorized vehicles. A significant portion of this literature examines the characteristics of the areas surrounding crash sites at various levels, such as Traffic Analysis Zones (TAZ) and census blocks, to understand their impact on crash frequency. Some studies have explored how the proximity and mix of different land uses correlate with the number of bicycle crashes [15], [16]. In terms of socioeconomic characteristics, Siddiqui et al. [17] analyzed pedestrian and bicycle crashes at the TAZ level, finding that household composition variably influences crash frequencies. A consensus emerges from research focusing on vehicular and bicycle traffic volumes, indicating a positive relationship between these volumes and the frequency of bicycle crashes [18], [19].

The role of bicycle infrastructure in influencing crash rates has also been scrutinized. The absence of dedicated bicycle lanes or paths has been consistently associated with an increase in collisions [11]. Furthermore, intersections have been pinpointed as high-risk zones for cyclists [20]. Prati et al.'s systematic literature review [21] emphasized that 57.6% of the studies reviewed identified infrastructure characteristics as crucial factors.

Road features and alignments receive attention in another strand of research. For instance, Robartes et al. [13] examined variables including the roads' horizontal and vertical alignments, finding that vertical grades and horizontal curves positively correlate with the frequency and severity of cyclist accidents. Specific road features, such as surface hazards (e.g., sand), road grades, and curves, have also been recognized for their role in heightening accident risks [22].

1.2.2 Factors affecting pedestrian crash frequency

Investigations into pedestrian crashes have revealed critical factors influencing their frequency. A portion of this research has concentrated on personal attributes of pedestrians and drivers, including age, gender, and health issues, to pinpoint individuals at heightened risk of involvement in road accidents [23], [24]. Another focus has been on demographic characteristics surrounding pedestrian crash hotspots, such as the age distribution of the population, household size, and land use patterns [25], [26].

Further studies have explored the influence of road and traffic features on pedestrian accidents. Haleem et al. [12] observed a linkage between elevated pedestrian crash rates and several variables, including average annual daily traffic (AADT), the proportion of truck traffic, and the existence of crosswalks. Chen et al. [9] highlighted that steep terrains and higher speed limits are associated with a rise in pedestrian crash frequencies.

Rahman et al. [27] delved into the factors affecting both the occurrence and severity of pedestrian crashes, noting the significant role of roadway design. They discovered that an increase in the number of lanes tends to raise crash rates, whereas higher speed limits and larger medians and shoulder widths contribute to a reduction in incidents. Obinguar et al.'s macroscopic analysis in the Philippines [24] identified primary roads and sections with poor surfaces as more susceptible to pedestrian accidents. Riccardi et al., utilizing Italian crash datasets, [14] underscored road

alignment as a critical factor, with straight alignments (tangents) being more prone to accidents compared to curves. Jima et al. [28] further examined the impact of road geometry on crash rates, finding that pedestrian accidents are most likely on straight segments, with pedestrians being the at-fault party in approximately 45% of collisions. Baireddy et al. [29] assessed various factors, including road characteristics, and determined that roadway functional class and lane count are pivotal in influencing pedestrian crash occurrences.

The critical role of road alignment in influencing the frequency of pedestrian and bicycle crashes has been underscored by different studies. However, a predominant reliance on preprocessed data sources for information on road alignments constitutes a notable methodological limitation. This reliance inherently restricts researchers' ability to assess the impact of road alignments in areas where such preprocessed data are unavailable. Consequently, this approach potentially omits significant insights into how road alignments contribute to crash frequencies in underrepresented regions. A more robust methodology would involve the development of a framework capable of extracting both vertical and horizontal road alignment data directly from raw sources, such as Geographic Information Systems (GIS) or Global Positioning Systems (GPS). This advancement would enable a more comprehensive analysis, extending the investigation to previously inaccessible areas and ensuring a more thorough understanding of the relationship between road alignment and crash frequencies.

1.3 Extracting Road Alignments

Identifying the horizontal and vertical alignments of roads from geospatial data is crucial for safety assessments. Many U.S. states maintain databases of such information, particularly for major roads. The development of an automated framework to extract these alignments is highly beneficial, offering significant savings in both cost and time. Common data sources for road

alignment include GPS data, satellite imagery, and laser-scanned data. However, satellite imagery falls short in extracting vertical alignment.

Several studies have focused on deriving road features directly from GPS data, which includes latitude, longitude, and altitude of road points. Bartin et al. [30] developed a web-based tool, CurvS, to extract and analyze road horizontal alignment information, such as straight segments (tangents), curve lengths, and curve radii. They tested this tool to rural roads in New Jersey and freeways in Nevada. Bíl et al. [31] introduced a software and an ArcGIS toolbox, ROCA, for segmenting roads into tangents and curves horizontally, calculating attributes like curve radius, detour ratio, and tangent azimuth. Their tests on the Czech road network revealed that curves with radii of 50 meters pose significantly higher dangers than those with radii of 1000 meters. Xu et al. [32] proposed a GIS and GPS data-based method for computing horizontal curves using regression analysis of road vertex direction-location profiles to determine curve radii, calculating direction changes at each point by considering adjacent points.

The challenges associated with the tools designed for extracting road alignments are multifaceted. Firstly, there is a pronounced emphasis on delineating only the horizontal aspects of road alignment, with vertical alignment often overlooked. This focus limits the scope of safety assessments and may neglect critical factors influencing accident rates. Secondly, the depth of horizontal alignment data extracted tends to be restricted, primarily involving the categorization of road segments into tangents and curves and the determination of curve radii. Such an approach may not capture the full complexity of road geometry and its potential impact on safety. Lastly, a prevalent issue with these tools is their lack of ongoing maintenance, rendering them obsolete and unusable for contemporary analyses. This absence of support and updates significantly undermines the potential for these tools to contribute to current and future road safety evaluation efforts.

1.4 Main Contributions

This study sets itself apart by shifting the analytical focus from a regional to a road segment level. Predominant research in the domain of pedestrian and bicycle crash frequency typically adopts a macroscopic view, analyzing data at broader geographic levels such as Traffic Analysis Zones (TAZ) or census blocks. Such a perspective, while useful, tends to average the data across all roads within a region, thereby diluting the distinct impact of individual road features and alignments. By concentrating on the specific characteristics of each road segment, this research endeavors to isolate and understand the pivotal factors influencing pedestrian and bike crashes. The detailed examination targets the road segments on Oahu Island, Hawaii, as the primary study area.

Secondly, this research introduces an automated framework for the computation of both horizontal and vertical road alignments using merely the latitude, longitude, and altitude data of road points. This method offers a streamlined, cost-efficient approach, compared to methods that require more manual intervention or the use of more advanced equipment for acquiring essential road alignment data, with potential for application worldwide, given the availability of the requisite geocoded data points.

Thirdly, by investigating the association between the frequency of pedestrian and bicycle crashes and the horizontal and vertical alignments of roads—alongside other road attributes—this study seeks to identify significant patterns and contributing factors. The aim is to uncover insights that could guide the development of more focused and effective road safety measures.

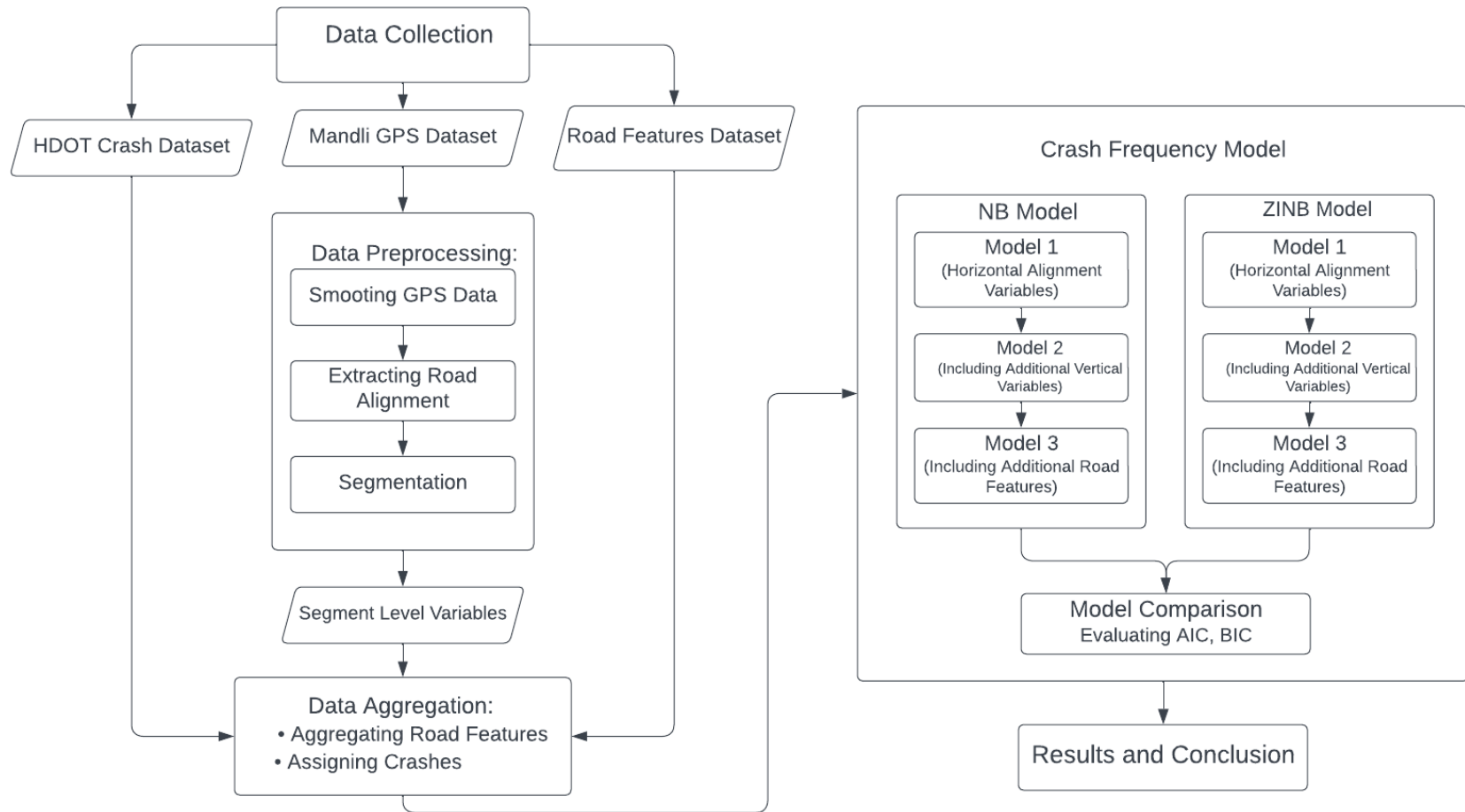


Figure 1: Flowchart of Crash Frequency Data Analysis and Modeling

Chapter 2: DATASET

2.1 Introduction to the Dataset

This thesis aims to explore the correlation between crashes involving vulnerable road users (pedestrians and bicyclists) and road alignment characteristics, using crash data spanning from 2015 to 2022, a comprehensive dataset provided jointly by the State of Hawaii, Department of Transportation, and ArcGIS Online Database. Additionally, detailed roadway alignment (horizontal and vertical) derived from GPS data for Hawaii roads, obtained from Mandli Communications Positional System. This section describes the different data sources and outlines the processes involved in preparing and processing the data to establish a foundation for the subsequent analyses aimed at understanding the relationship between roadway designs on crash occurrences. Additionally, it sets the stage for a descriptive analysis of the dataset, aiming to preliminarily explore the influence of road alignment on vulnerable road user crashes.

2.2 Data Sources

This research draws upon data from diverse sources to facilitate a comprehensive analysis. The subsections below detail each data sources utilized in the study. Table 1 provides a summarized view of these sources along with the key variables each contains.

Table 1: Overview of Dataset Sources and Key Variables

Data Source	Variable	Description
Crash Dataset	Crash Location	Location data of incidents, used for mapping crash sites.
	Incident Type	Specifies if the crash involved pedestrian or bicyclist.
Mandli Communications GPS Data	GPS Points	Latitude, Longitude, and Altitude data used to determine road alignments.
Road Features Dataset (HPMS)	AADT	Annual Average Daily Traffic, indicates road usage volume.
	Speed Limit	Maximum legal speed allowed on road segments.
	Lane Width	Width of individual lanes, impacts road capacity and safety.
	Number of Lanes	Total lanes available, indicates road capacity.
	Road Functional Classification	Categorizes roads based on their primary function within the transportation network, including: <ol style="list-style-type: none"> 1. Interstate Roads, 2. Principal Arterial-Other Freeways & Expressways, 3. Principal Arterial-Other, 4. Minor Arterial, 5. Major Collector, 6. Minor Collector, 7. Local

2.2.1 Crash Dataset

This dataset details the vehicle accidents recorded from 2015 to 2022, obtained through the Hawaii Department of Transportation (HDOT). For the years 2015 to 2018, data was sourced from the ArcGIS Online HDOT Database Crash Data. For the period from 2019 to 2022, data was sourced from the Hawaii Department of Transportation Crash Records. While both datasets come from HDOT, their data collection methods may be identical, focusing on detailed geographical information. This dataset is particularly focused on incidents involving pedestrians or bicyclists on the island of Oahu and provides geocoded data. While the geographical specifics of this dataset are detailed, it's notable that the scope of contextual and temporal specifics is not covered. The dataset's geographical precision offers a robust foundation for spatial analysis but lacks temporal and contextual details like the time of day, weather conditions, and crash severity. Such information could enrich our understanding of crash patterns and contributing factors, such as visibility, road conditions, and behavioral factors. Despite this, the dataset's focus on geographical specifics aligns well with examining road alignment characteristics, making it suitable for analyzing how road design influences safety.

These coordinates were then imported into ArcGIS for analysis alongside the road network data from Mandli Communications. . Due to the absence of crash direction data, the analysis could not account for the road segment behind each crash relative to the travel direction. Therefore, we assigned crashes to the nearest road segment, focusing only on the road alignment within that segment.

Figure 2 shows a comprehensive map of pedestrian and bicycle crash incidents on Oahu from 2015 to 2022.

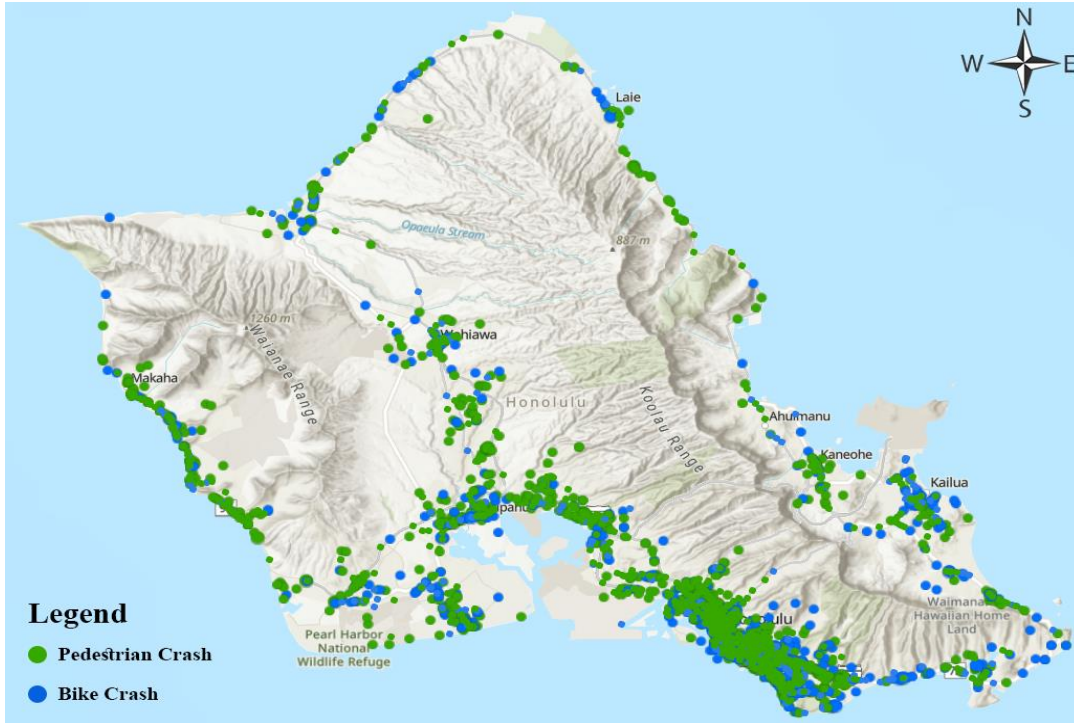


Figure 2: Comprehensive Map of Pedestrian and Bicycle Crash Incidents on Oahu (2015-2022)

2.2.2 Road Features Dataset

To enrich segmented analysis, we incorporated additional road features, obtained annually from the ArcGIS Online Database, sourced from the Hawaii Department of Transportation (HDOT). This dataset details roadway segments, each characterized by attributes derived from the Highway Performance Monitoring System (HPMS). These attributes include the Annual Average Daily Traffic (AADT), Speed Limit, Lane Width, Number of Lanes, and the Road's Functional Classification.

To integrate these variables into the road segments, we utilize the Route Beginning Milepost (BMP), and Ending Milepost (EMP) indicators in the dataset. This integration method aligns with the data GPS point's collection trajectory, which adheres to a positive milepost direction (+MP). The cumulative distance between consecutive GPS points is calculated to span the BMP to EMP

range for each Route, ensuring that the variables are mapped accurately within the predefined segments. By attributing specific environmental and infrastructural factors to distinct road segments, the analysis gains depth. This variable set underscores the importance of a multi-faceted approach in transportation safety research, recognizing the complex interplay between road design, traffic behavior, and environmental conditions.

2.2.3 GPS Dataset

To collect road alignment data across Oahu, a script was employed using Selenium Library in Python [34], an automation tool that facilitates interaction with web browsers. Selenium automates browsers to perform tasks such as navigating through web pages, and extracting data, reducing the time and effort required for manually collecting large datasets. In this case, Selenium was programmed to navigate the Mandli Communications website [35] and extract GPS data points for each road, systematically capturing latitude, longitude, and altitude information for data points at intervals of every 0.002 mile. **Figure 3** presents the extent of the road network for which GPS data was gathered.



Figure 3: Coverage of Road Network with Gathered GPS Data

2.2.3.1 Mandli POS

Mandli Communications utilizes a positional system to collect an array of data from a moving vehicle, including vehicle position, altitude, velocity, track, speed, and dynamics [36]. Central to this system is the Position and Orientation System (POS) unit, which is essential for obtaining GPS data points such as latitude, longitude, and altitude—key for road alignment analysis. This unit, using GPS technology, accurately determines the vehicle's global position. The system also includes an Inertial Measurement Unit (IMU) for enhanced positional accuracy and a wheel encoder Distance Measurement Instrument (DMI) for precise distance tracking. Our study

specifically leverages the GPS data acquired by the POS unit to analyze Oahu's road alignment, focusing on the roads' geometric characteristics.

2.3 Preprocessing GPS Points

Before progressing with the extraction of road geometry, it is important to address the preprocessing step due to the close placement of GPS data points, at 0.002-mile intervals. For this purpose, the Savitzky-Golay filter was applied for smoothing [37]. While smoothing can be beneficial in less dense datasets to reveal overarching patterns, its necessity grows in densely collected datasets where minor variations could be misinterpreted as significant geometric features. Unlike basic smoothing methods, such as the weighted moving average, that might blur significant directional or elevational changes, the Savitzky-Golay filter reduces noise while preserving essential characteristics of the road's alignment, such as curvature and slope. Furthermore, it avoids the challenges of setting a simplification threshold required by algorithms like Douglas-Peucker, which can risk losing crucial details. This capability makes it suitable for processing GPS data for road geometry analysis.

2.3.1 Savitzky-Golay Filter

The Savitzky-Golay filter works by fitting a low-degree polynomial to successive subsets of adjacent data points, using a least squares method. The filter calculates the smoothed value for each data point by convolving the original data points with a set of polynomial coefficients. These coefficients are determined in a way that the resulting polynomial curve closely matches the original data's shape within a specified window. This process is repeated across the dataset, ensuring each point is smoothed in context with its neighbors. The Savitzky-Golay filter we

applied to smooth the GPS data, fits a cubic polynomial to the data within a moving window of 9 points.

In practical terms, if we consider a dataset consisting of points $\{x_j, y_j\}$ where x_j represents an independent variable and y_j is the observed value, the smoothing process treats these points with a predefined number of convolution coefficients C_i , according to the formula [37]:

$$Y_j = \sum_{i=\frac{1-m}{2}}^{\frac{m-1}{2}} C_i (y_{j+i}), \quad \frac{m+1}{2} \leq j \leq n - \frac{m-1}{2} \quad (1)$$

Here, j represents the position of a data point within the series, and m reflects the total number of points used in the smoothing window, extending from $i = \frac{1-m}{2}$ to $i = \frac{m-1}{2}$ around each j . n is the total number of points in the specific road segment under analysis. This equation calculates the smoothed data point Y_j by applying the convolution coefficients to the observed values. A simple version of the Savitzky-Golay filter utilizes a quadratic polynomial coupled with a window size of 5 points. This elementary form is often employed for basic data smoothing needs.

To illustrate, for a dataset smoothed with a 5-point quadratic Savitzky-Golay filter, the j^{th} smoothed data point, Y_j can be calculated as:

$$Y_j = \frac{1}{35} (-3y_{j-2} + 12y_{j-1} + 17y_j + 12y_{j+1} - 3y_{j+2}) \quad (2)$$

In this equation, $m = 5$ reflects the total number of points used in the smoothing window, extending from $i = -2$ to $i = 2$ around each j .

To better capture the intricacies of road alignments in our analysis, we selected a cubic polynomial with a larger window size of 9. This enhanced configuration allows for a more nuanced smoothing

of our high-resolution GPS data. The equation is detailed below, taking into account a 9-point span centered around each data point (with $m = 9$ and i ranging from -4 to 4):

$$Y_j = \frac{1}{231} (-21y_{j-4} + 14y_{j-3} + 39y_{j-2} + 54y_{j-1} + 59y_j + 54y_{j+1} + 39y_{j+2} + 14y_{j+3} - 21y_{j+4}) \quad (3)$$

The convolution coefficients are derived from established tables that specify the coefficients based on the polynomial degree and the chosen window size.

Figure 4 is a visual representation of a road's original horizontal alignment with its smoothed version after applying the Savitzky-Golay filter. The original road's trajectory is marked in red, showing every minor undulation, while the blue line illustrates the smoothed alignment, where inconsequential fluctuations are minimized. This smoothing process preserves the road's essential characteristics, ensuring that important aspects of its curvature are captured without the noise of trivial variations.

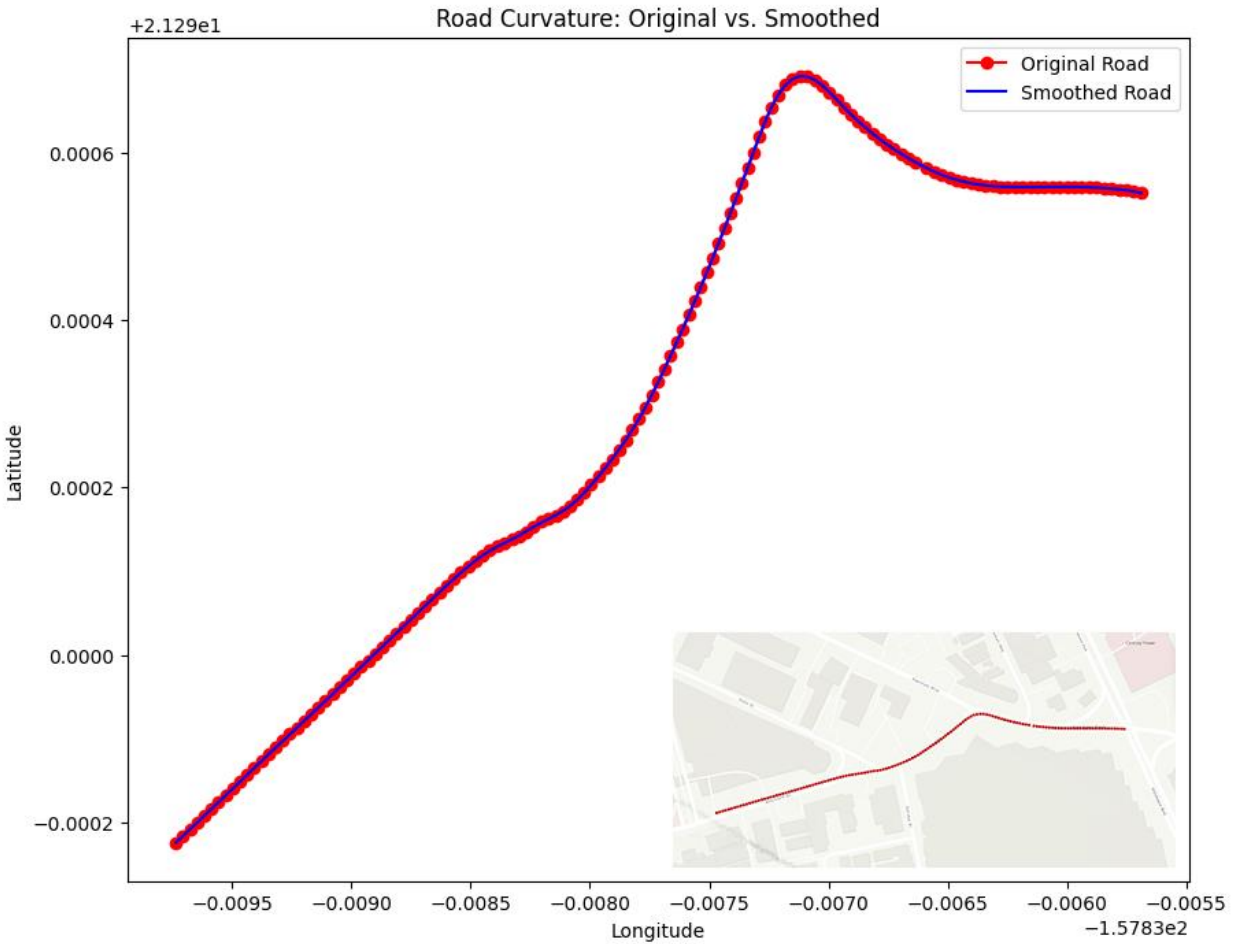


Figure 4: Original vs. Smoothed Road Points Using the Savitzky-Golay Filter on Atkinson Drive

2.4 Processing GPS Points

In this section, we discuss processing of road alignments, which is composed of horizontal and vertical elements that collectively define the geometry of the roadway. We examine each alignment type and the specific geometric considerations it entails.

2.4.1 Horizontal Alignment

2.4.1.1 Bearing Angle

In the analysis of horizontal alignments, an initial step involves the calculation of the bearing between consecutive geospatial points to ascertain the road's directional changes. This bearing change offers a quantitative measure of how the road deviates or curves away from a linear path which impacts road safety.

The bearing between two points, P_1 and P_2 is computed based on their geographic coordinates, $P_1(lat_1, lon_1)$ and $P_2(lat_2, lon_2)$, respectively. The calculation proceeds as follows [38]:

The longitude difference, $\Delta lon = lon_2 - lon_1$, and the latitudes and longitudes converted to radians, serve as inputs to derive the bearing (θ) using the formula:

$$\theta = \arctan2(\sin(\Delta lon) \cdot \cos(lat_2), \cos(lat_1) \cdot \sin(lat_2) - \sin(lat_1) \cdot \cos(lat_2) \cdot \cos(\Delta lon)) \quad (4)$$

Where $\arctan2$ yields the arctangent of the quotient of its arguments, considering the sign of both to determine the correct quadrant.

The initial bearing calculated in radians is converted to degrees and normalized to a 0° to 360° range to obtain the final compass bearing.

This bearing calculation is repeated for each pair of consecutive points along the road. The change in bearing between these pairs indicates the road's deviation from a straight path.

To capture the concept of road deviation through bearing changes between three consecutive points P_1 , P_2 and P_3 , we calculate the bearing from P_1 and P_2 and from P_2 to P_3 , and then analyze the difference between these two bearings. This difference provides insight into the nature of the turn.

Let B_{12} represent the bearing from point P_1 to P_2 , and B_{23} represent bearing from P_2 to P_3 . The change in bearing, which we can denote as ΔB is calculated as follows:

$$\Delta B = B_{23} - B_{12} \quad (5)$$

However, to account for the cyclic nature of bearings (where a bearing of 360° is equivalent to 0°), the difference should be normalized to the range of -180° to 180° to represent the most direct change in direction. Thus, the normalized change in bearing, ΔB_{norm} , can be calculated as:

$$\Delta B_{norm} = ((\Delta B + 180) \text{ mode } 360) - 180 \quad (6)$$

The sign of the normalized bearing change ΔB_{norm} indicates the direction of the turn. A positive value of ΔB_{norm} signifies a right turn, while a negative value indicates a left turn. The distinction between left and right turns, when combined with crash direction data, provides a nuanced understanding of how turning movements may contribute to accidents. The bearing angle concept is depicted in **Figure 5**, as sourced from [38].

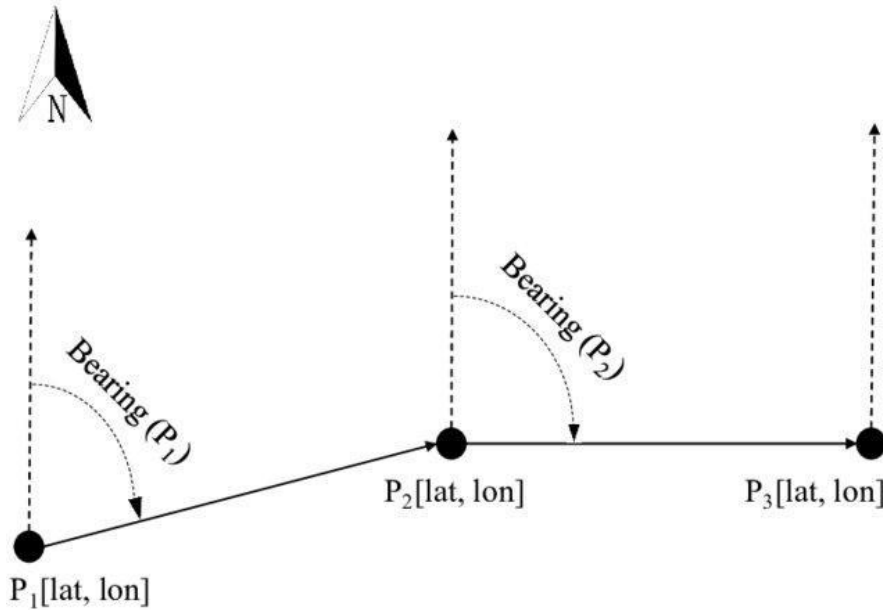


Figure 5: Bearing between two consecutive GPS points. lat and lon represents the latitude and longitude of a GPS point

2.4.1.2 Radius

In assessing the horizontal geometry of roadways, estimating the curvature and corresponding radius at discrete points along the road's trajectory is essential. The radius of curvature is measured in meters. This analytical approach is aligned with the methodology outlined by Andra's̃ik and Bı'1 [39], which provides for the estimation of curvature in the absence of continuous derivatives, a scenario with GPS data representing road alignments as a series of discrete points.

For each point X_i on a road segment, the curvature is calculated using the change in direction over the path defined by the points X_{i-1} , X_i and X_{i+1} . This is quantified as the angle $A_I(X_i)$ between the vectors $u_i = X_{i-1} - X_i$ and $v_i = X_{i+1} - X_i$. The radius of a circumscribed circle (*RCC*) that approximates the road's bend at X_i is computed as:

$$RCC(X_i) = \frac{\|X_{i+1} - X_{i-1}\|}{2\sin(180 - A_I(X_i))} \quad (7)$$

The radius of osculation (*ROC*), which provides a localized measure of curvature at X_i , is determined by the formula:

$$ROC(X_i) = \frac{\|X_{i+1} - X_{i-1}\|}{\|T(X_{i+1}) - T(X_{i-1})\|} \quad (8)$$

In this context, $T(X_i)$ is the tangent unit vector at point X_i calculated by the difference of the position vectors normalized:

$$T(X_i) = \frac{(X_{i+1} - X_{i-1})}{\|X_{i+1} - X_{i-1}\|} \quad (9)$$

The Euclidean distance between two consecutive points X_i and X_{i+1} , is computed using their position vectors within a Cartesian coordinate system. This distance is represented by the norm of the vector difference between the two points, expressed as:

$$\|X_{i+1} - X_i\| = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \quad (10)$$

Where (x_i, y_i) and (x_{i+1}, y_{i+1}) correspond to the Cartesian coordinates of points X_i and X_{i+1} .

The curvature ($K(X_i)$) of a road at a specific point using the osculating circle (also known as the radius of curvature) can be calculated as the reciprocal of the radius of the osculating circle $ROC(X_i)$. The formula for curvature at a point is given by:

$$K(X_i) = \frac{1}{ROC(X_i)} \quad (11)$$

where a larger value of $K(X_i)$ indicates a sharper curve, and a smaller value corresponds to a gentler curve.

2.4.2 Vertical Alignment

Following our examination of horizontal alignment characteristics, our analysis extends into the domain of vertical alignment. Vertical alignment encompasses the elevation changes and gradients of road segments, which are crucial for understanding the roadway's design and its implications for vehicular dynamics and safety. In this context, we assess several key variables that articulate vertical profile:

2.4.2.1 Slope (degree)

Slope Measures the steepness of a road segment by calculating the angle formed by the elevation change between two points relative to their horizontal distance. It's given by the formula:

$$Slope(degrees) = \arctan\left(\frac{Elevation\ Change}{Horizontal\ Distance}\right) \quad (19)$$

2.4.2.2 Grade

Represents the slope as a percentage, providing an intuitive measure of how steep a segment is.

It's calculated as:

$$Grade(\%) = \left(\frac{Elevation\ Change}{Horizontal\ Distance}\right) \times 100 \quad (20)$$

The Grade (%) Change between two consecutive points can be calculated using the difference in their grade percentages. This calculation highlights the rate at which the road's incline or decline changes over a short distance. The formula to calculate Grade (%) Change (G_{change}) is given by:

$$G_{change} = G_{i+1} - G_i \quad (21)$$

Where G_{i+1} is the grade at point $i+1$, calculated as $\left(\frac{Elevation_{i+1}-Elevation_i}{Horizontal\ Distance_{i+1,i}}\right) \times 100$ and G_i is the grade at point i , calculated as $\left(\frac{Elevation_i-Elevation_{i-1}}{Horizontal\ Distance_{i,i-1}}\right) \times 100$.

This equation allows us to assess how abruptly the road slope changes between two points.

2.4.3 Homogeneous segments

In our approach to classifying road segments as either Tangents (straight paths) or Curves, we draw upon a methodology that involves setting a threshold based on the radius of osculation (ROC), as inspired by the findings in [39]. The determination of an optimal ROC threshold is essential for the accurate differentiation between Curves and Tangents, ensuring that each segment is classified in a manner that reflects its true geometric character. According to the analysis presented by Andra's'ik and B'ıl [39], a ROC threshold of 741 meters has been identified as the most effective criterion for this classification process. This specific threshold was determined through a classification tree approach.

Figure 6 provides an illustrative example of classified road segments with: **a)** the original GPS road points, **b)** the classified GPS road points before applying the Savitzky-Golay smoothing filter, and **c)** the classified GPS road points after smoothing with the Savitzky-Golay filter. Here, the classification into 'left turn' and 'right turn' is derived from the sign of the bearing change, and is based solely on the direction of the GPS data collection, denoted as the positive milepost direction (+MP). This classification does not reflect the actual travel direction of vehicles or crash direction data.

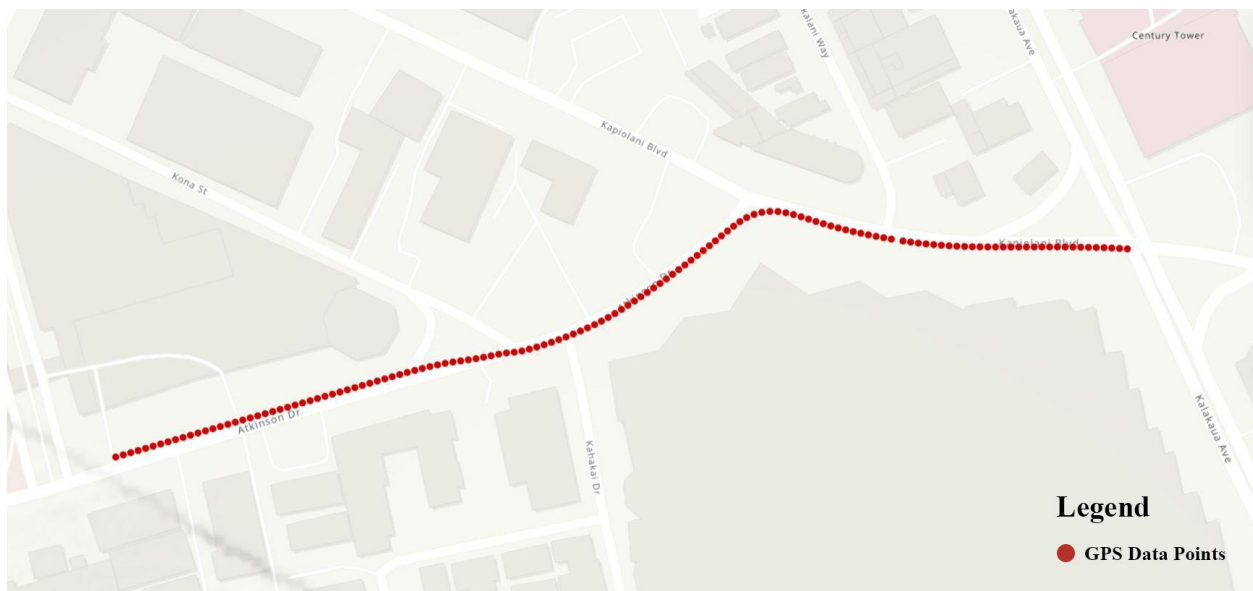


Figure 6a: Road Segmentation Visualization using Original GPS Points

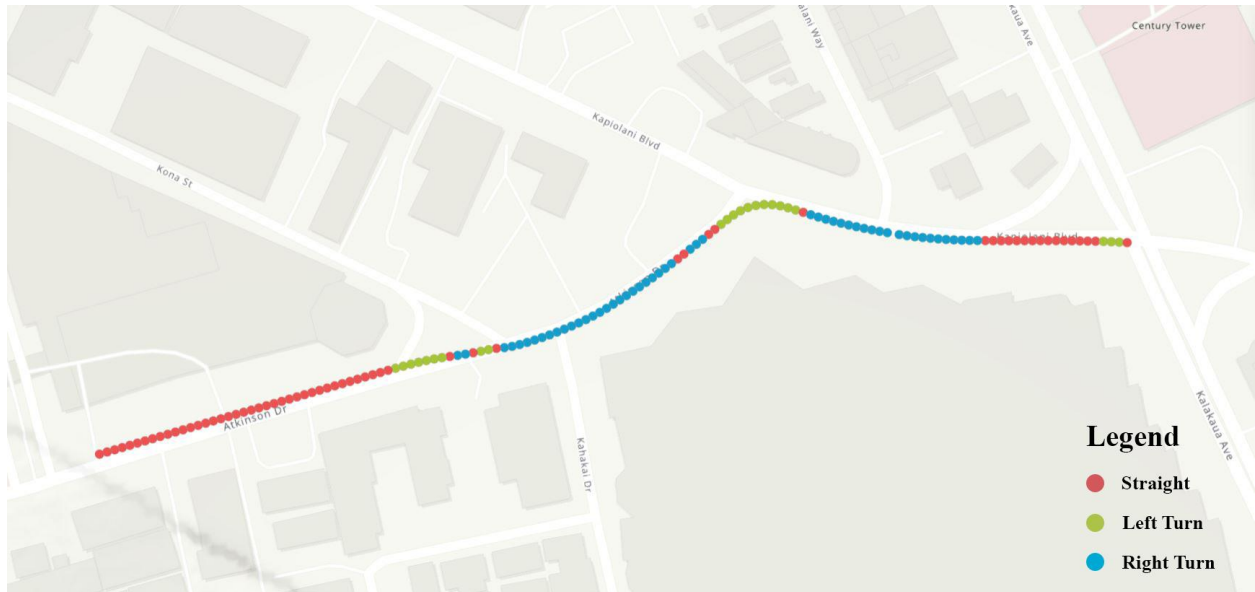


Figure 6b: Road Segmentation Visualization using Classified Points before Smoothing

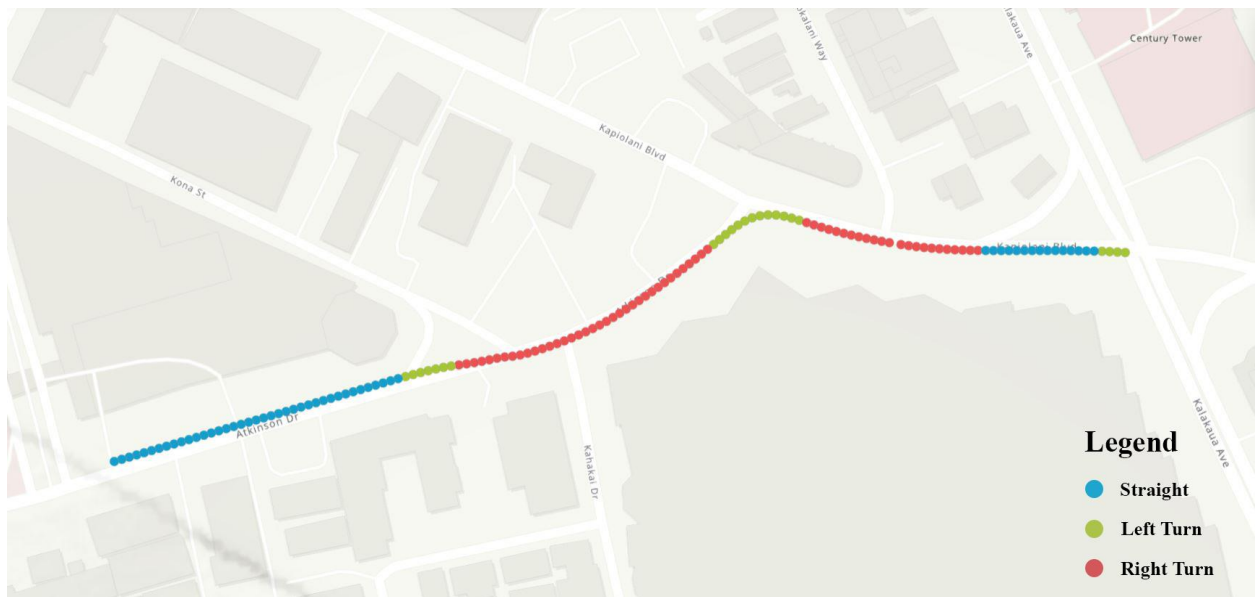


Figure 6c: Road Segmentation Visualization using Classified Points after Smoothing

Figure 6: Road Segmentation Visualization: a) Original GPS Points, b) Classified Points before Smoothing, c) Classified Points after Smoothing

With the current crash direction data limitations, the road segmentation is better visualized in **Figure 7**, with each segment differentiated by color purely to distinguish adjacent segments and

the colors themselves do not represent specific features of the road. This segmentation enables the analysis of road geometries at a segment level, providing a foundation for evaluating the influence of road design on traffic behavior and safety.

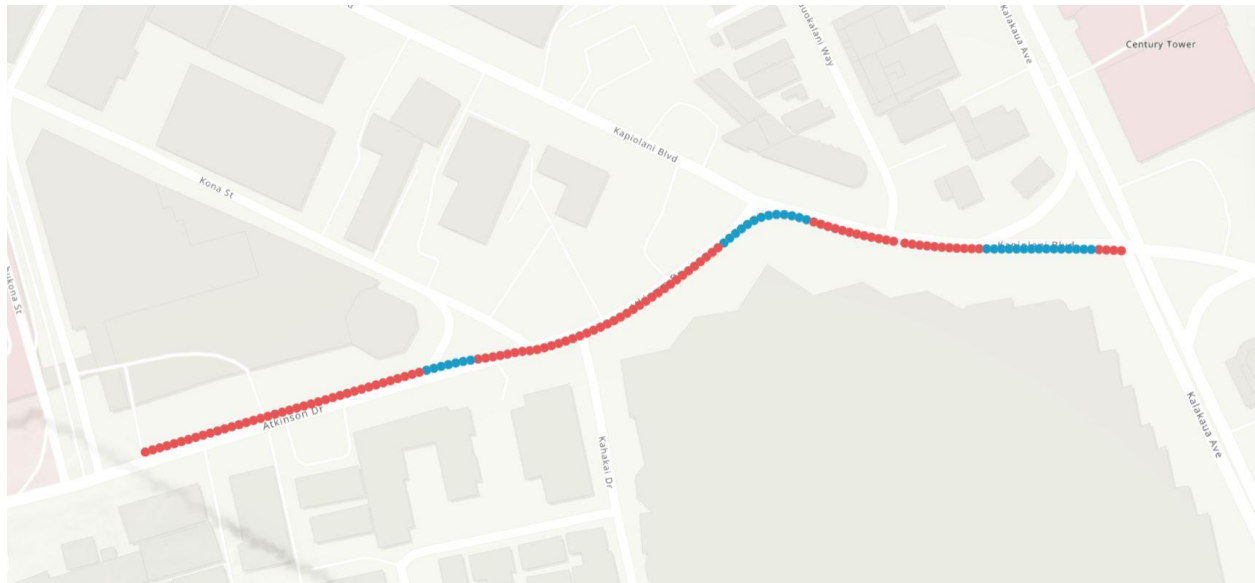


Figure 7: Analyzed Road Segments Visualization

2.4.3.1 Segment-level variables

Before delving into the analysis of segment-level variables, it is essential to discuss the significance of classifying road segments into homogenous categories based on their horizontal alignment. Homogeneous segmentation ensures that each segment analyzed is uniform in terms of its geometric properties, such as curvature or straightness. This methodology contrasts with fixed-size segmentation. While fixed-size segments offer a generalized view, they may inadvertently combine varying geometries—like parts of a curve with a tangent—within a single segment, potentially diluting the specific impacts of these geometric features on road safety.

Figure 8, as illustrated in [40], serves as a compelling example for this understanding. It demonstrates how segments with identical mean curvature can present distinct driving experiences and crash risks due to their internal curvature variability.

Two segments may have the same mean curvature, One segment with two gentle curves and another with two straight segments with a sharp curve in between. This observation underscores the importance of considering not just the mean curvature but also the standard deviation, which captures the variability within the segment.

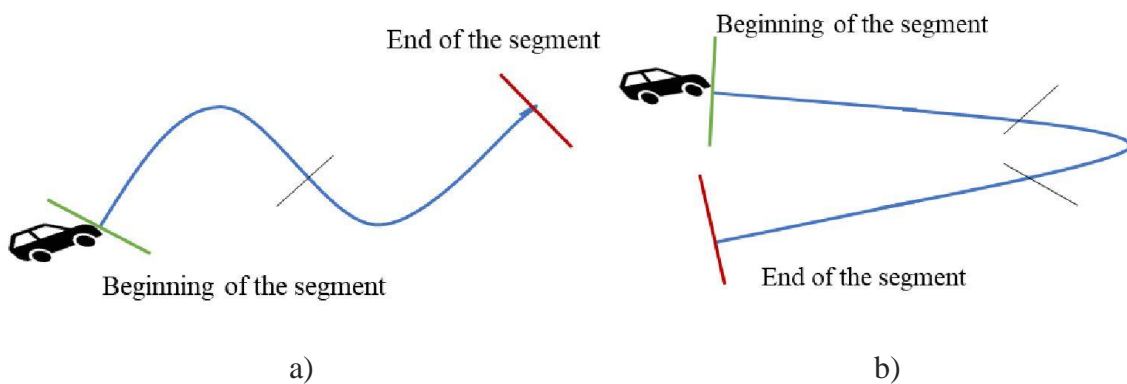


Figure 8: A sketch illustrating the importance of considering the standard deviations. a) segment with a lower standard deviation of absolute average curvature, and b) same absolute average curvature – higher standard deviation (plan view)

While mean curvature and standard deviation offer insights into the overall bend of a road segment, our approach allows us to consider the specific homogeneous segments that drivers actually encounter.

In our model, road segments are classified based on their homogeneous alignment characteristics rather than relying solely on averaged measures. For instance, the segment depicted in **(a)** of **Figure 8**, which showcases a segment composed of two gentle curves, would be segmented as two

individual curves. In cases where the direction of driving is known, these could be identified as sequential curves, such as a right turn followed by a left turn, also known as a 'reverse curve'. In the case of the segment shown in (b) of **Figure 8**, where there are two straight segments with a sharp curve in between, our model would recognize and classify this as a straight, curved, straight segment. This segmentation reflects the actual driving experience more closely than an averaged curvature value could.

Transitioning from the analysis of point-level variables to segment-level metrics, our model incorporates a set of variables that are calculated after classifying points into homogenous road segments. These segment-level variables offer a comprehensive view of the road alignment characteristics, providing insights into their impact on road safety. we proceed to evaluate:

2.4.3.1.1 Total Length

This metric denotes the longitudinal extent of a segment, measured in meters. To compute the Total Length of a road segment within our model, we begin by converting geographic coordinates (latitude and longitude) into Cartesian coordinates (X, Y) . This step enables us to apply Euclidean distance for precision:

$$Distance = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \quad (12)$$

Given X_1, Y_1 and X_2, Y_2 as the coordinates of consecutive points, the Total Length is the sum of these distances across the segment:

$$TotalLength = \sum_{i=1}^{n-1} \sqrt{(X_{i+1} - X_i)^2 + (Y_{i+1} - Y_i)^2} \quad (13)$$

2.4.3.1.2 Cumulative Bearing Angle Change

The Cumulative Bearing Change for a road segment is the sum of the bearing Angle changes at each point, derived from the directional differences between three consecutive points. The formula for this calculation is:

$$\text{Cumulative Bearing Angle Change} = \sum_{i=2}^{n-1} B_{change}(i) \quad (14)$$

Where $B_{change}(i)$ is the bearing change at the point i^{th} , indicating the segment's total directional variation. This variable encapsulates the overall directional evolution of a segment, reflecting its potential complexity for navigation.

2.4.3.1.3 Degree of Curvature

The Degree of Curvature for a segment is quantified by comparing the curvature of the road to a standard circle, expressed in degrees for a given arc length. The equation for calculating the Degree of Curvature (D) can be given as:

$$D = \frac{L}{R} \times \frac{180}{\pi} \quad (15)$$

Where L represents the arc length of the segment, and R is the radius of the curve in meters. This formula calculates the angle in degrees that a curve subtends at the center of a circle of radius R over the length of the arc L . The smaller the radius (R), the higher the degree of curvature (D), indicating a sharper curve.

In practical road design, the Degree of Curvature is determined by the angle subtended by an arc of a specific standard length. This method is widely adopted in the U.S., where the arc length is typically standardized to 100 feet. The formula adjusts to:

$$D = \frac{5729.58}{R} \quad (16)$$

Here, D is in degrees and R is in feet. This specific equation provides a uniform measure of curvature.

2.4.3.1.4 Detour Ratio

The Detour Ratio is calculated as the quotient of the actual path distance of a road segment to the straight-line (geodesic) distance between the segment's endpoints. This measure highlights the directness or sinuosity of the route. Mathematically, it is expressed as:

$$DetourRatio = \frac{Total\ Length}{Geodesic\ Distance} \quad (17)$$

Where Total Length is the sum of distances between consecutive points along the road segment, computed as described previously.

The Geodesic Distance between two points, given their latitudes (ϕ_1, ϕ_2) and longitudes (λ_1, λ_2), can be calculated using the Haversine formula [41]:

$$GeodesicDistance = 2r \cdot \arcsin \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right) \quad (18)$$

Here, r is the Earth's radius (approximately 6,371 kilometers or 3,959 miles), and the latitudes and longitudes are expressed in radians. The Haversine formula accounts for the spherical shape of the Earth, providing an accurate measure of the straight-line distance between two points on the globe. This approach to calculating the Detour Ratio offers a precise measure of how much a road segment deviates from the most direct path between its endpoints, reflecting the segment's sinuosity.

2.4.3.1.5 Curve Classification

In our model, each curve within the road segments is classified according to the Highway Performance Monitoring System (HPMS) Curve Classes [42], based on the Degree of Curvature.

The classification is defined by the following criteria:

- Class A: Degree of Curvature < 3.5
- Class B: $3.5 \leq$ Degree of Curvature < 5.5
- Class C: $5.5 \leq$ Degree of Curvature < 8.5
- Class D: $8.5 \leq$ Degree of Curvature < 14
- Class E: $14 \leq$ Degree of Curvature < 28
- Class F: Degree of Curvature ≥ 28

This categorization allows us to assess the curvature's impact on road safety by delineating segments into discrete classes that reflect the curve's sharpness and navigational challenge. A higher class (towards F) indicates a sharper curve, typically necessitating lower speeds and greater caution, thereby affecting driver behavior and potential safety risks. This curve classification as detailed in the HPMS Field Manual, provides a standardized method for evaluating and comparing road curvature across various segments.

2.4.3.1.6 Grade Classification

In our analysis, the classification based on grade is aligned with guidelines specified by the Highway Performance Monitoring System (HPMS) [42]. Categorizing road grades into these classes, from A indicating less steep slopes to F denoting steeper grades, allows us to evaluate how incline steepness impacts vehicular dynamics, such as acceleration and braking efficiency. The grade classification is defined as follows:

- Class A: 0.0 - 0.4 percent

- Class B: 0.5 - 2.4 percent
- Class C: 2.5 - 4.4 percent
- Class D: 4.5 - 6.4 percent
- Class E: 6.5 - 8.4 percent
- Class F: 8.5 percent or greater

Furthermore, in directional analysis, when the driving direction is known, we also take into account the sign of the grade. A positive grade indicates an uphill slope, while a negative grade denotes a downhill slope. This distinction is important as the same grade value with opposite signs can exert significantly different effects on driving behavior and vehicle control.

The approach of calculating vertical alignment parameters, like slope, is adapted to fit within the confines of previously determined horizontal segments. This adaptation offers more targeted insights into how vertical alignment interacts with horizontal features to affect road safety.

2.5 Descriptive Analysis of the Aggregated Dataset

This section offers a descriptive analysis of an aggregated dataset that includes road segments, crash data, and road features from 2015 to 2022. By integrating diverse data sources, such as detailed roadway alignment from GPS data, crash records, and road attributes from the ArcGIS Online Database provided by the Hawaii Department of Transportation (HDOT), we aim to shed light on the relationship between road characteristics and crash incidents involving vulnerable road users. **Table 2** presents the descriptive statistics of the continuous explanatory variables utilized in our analysis. This table offers key metrics such as mean, standard deviation, minimum, and maximum for each variable.

Table 2: Descriptive statistics of the continuous variables

Variable	Min	Max	Mean	S.D.
Total Length	6.413	3,684.504	239.691	246.336
Horizontal Radius	20.809	25,355.563	3,165.492	2,999.477
Cumulative Bearing Angle Change	0.280	891.667	33.698	48.673
Detour Ratio	0.996	9.116	1.087	0.322
Grade (%)	-5.651	6.491	0.123	1.010
AADT	90	195,650	11,407.56	16,280.39
Speed (mph)	5	60	28.185	12.376
Lane Width (feet)	6	18	11.626	1.067
Lane Number	1	11	3.650	1.540

The following figures illustrate some aspects of the dataset, each provides insights into the potential influence of road design on safety outcomes, setting the stage for more detailed statistical modeling and analysis.

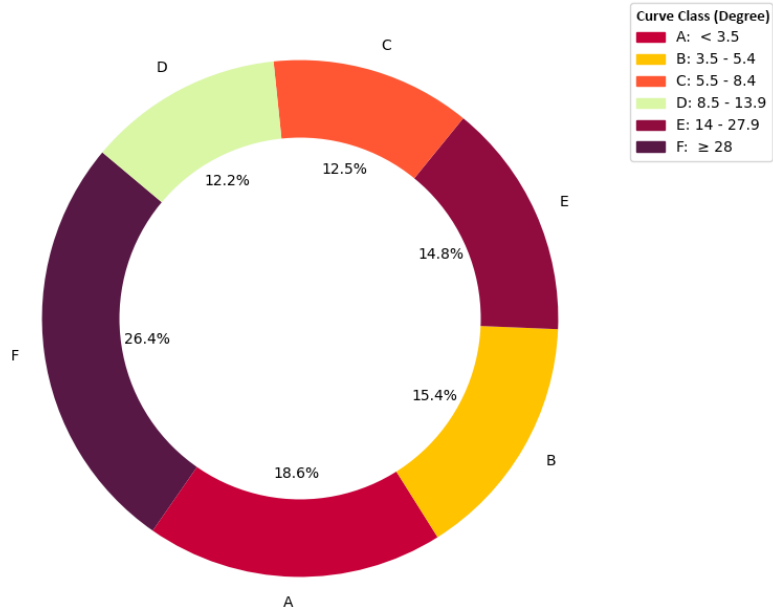


Figure 9: Proportion of Road Segments by Curve Class

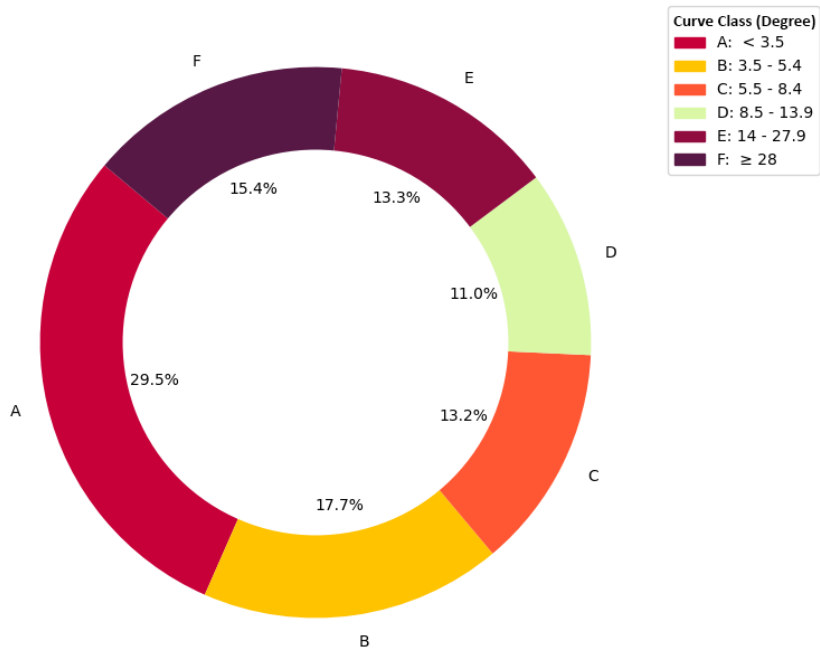


Figure 10: Proportion of Total Road Length by Curve Class

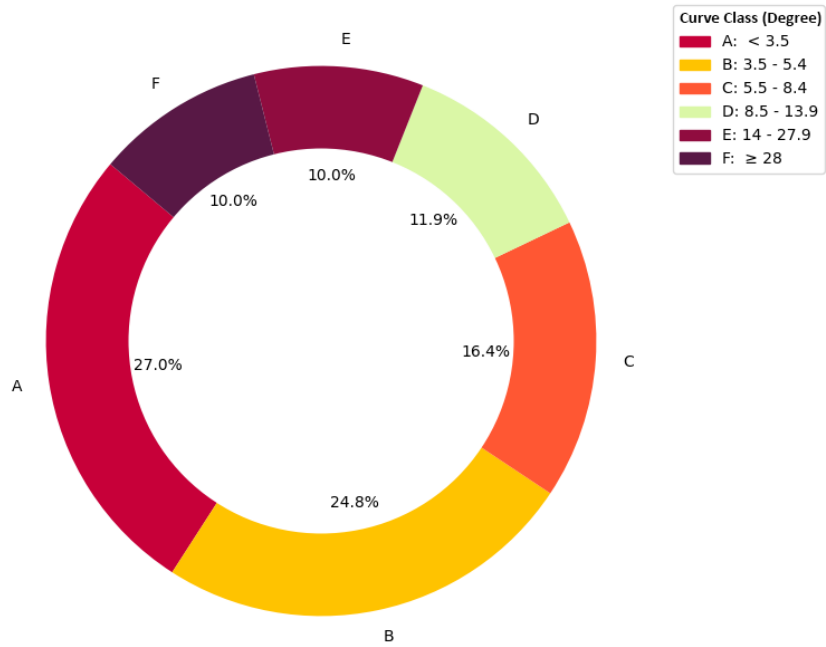


Figure 11: Proportion of Crash Counts by Curve Class

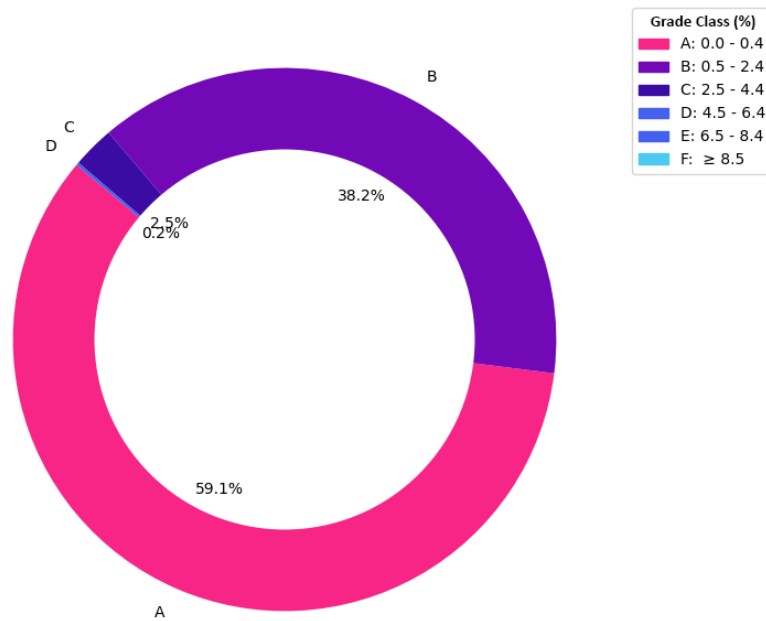


Figure 12: Proportion of Road Segments by Grade Class

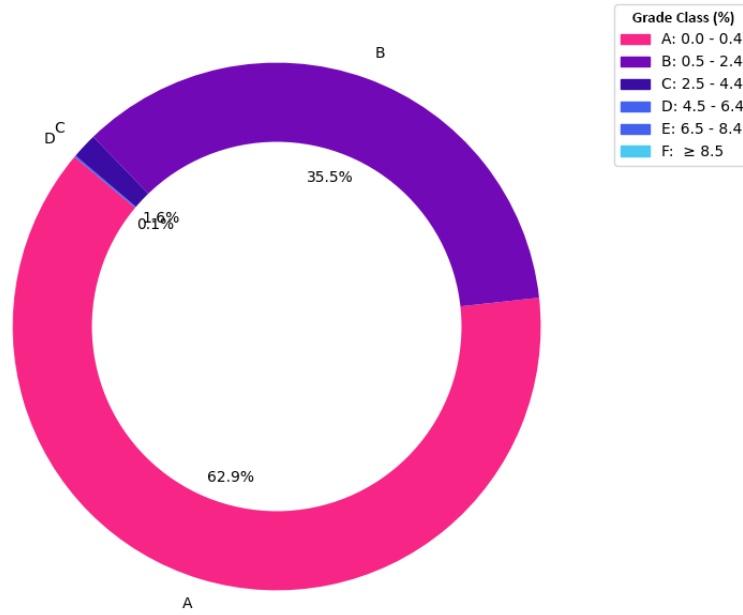


Figure 13: Proportion of Total Road Length by Grade Class

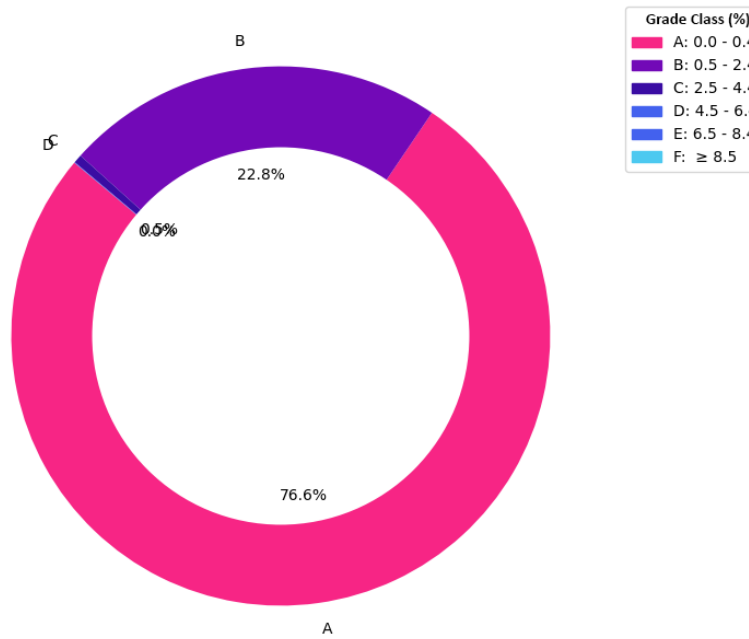


Figure 14: Proportion of Crash Counts by Grade Class

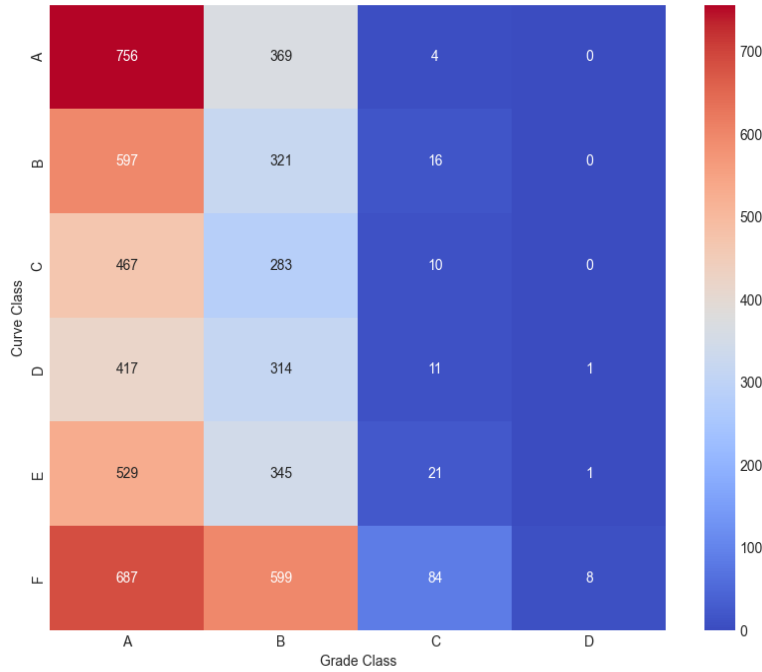


Figure 15: Distribution of Segment Combinations across Curve and Grade Classes

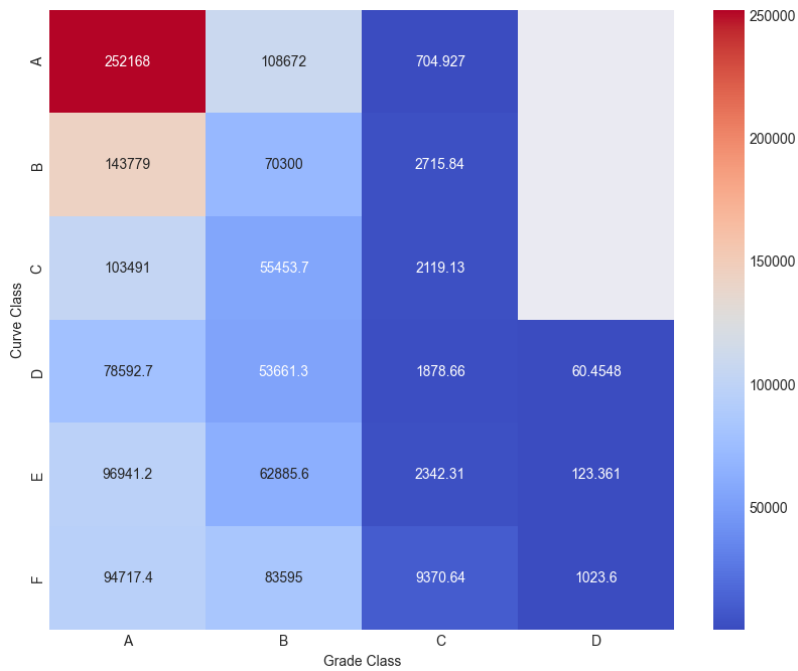


Figure 16: Aggregate Length of Segments by Curve and Grade Class Combination

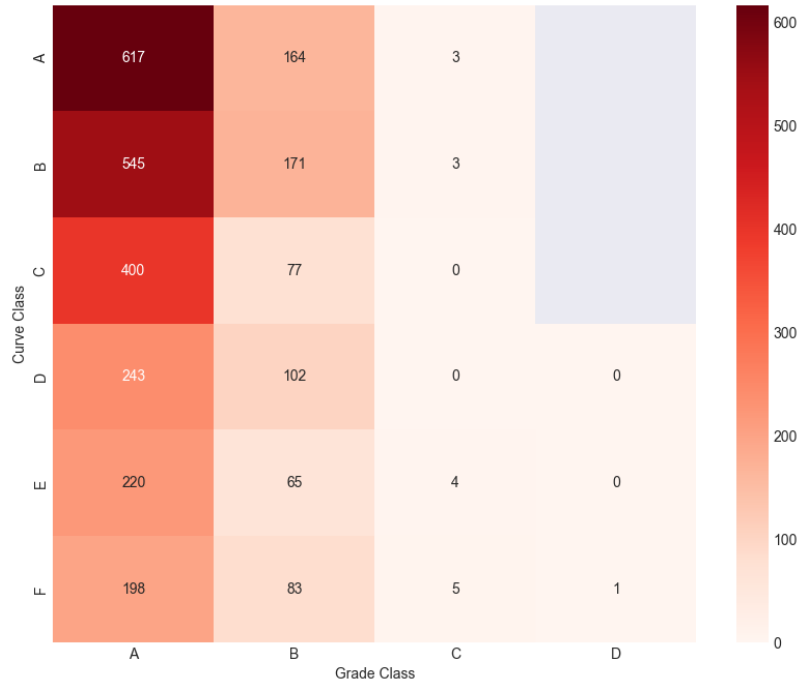


Figure 17: Crash Count Distribution by Curve and Grade Class Combinations

Chapter 3: METHODOLOGY

In this section, the methodology adopted to investigate the relationship between road alignment characteristics and crash occurrences is provided. The analysis leverages Generalized Linear Models to elucidate patterns and relationships within the data.

Distinct from approaches that utilize fixed-size segments for crash frequency analysis [40], this study employs a segmentation strategy based on the homogeneity of road alignment characteristics. This decision bypasses the necessity of selecting an optimal segment length—a balance between overly short segments, which may result in an excessive number of segments with few or no crashes, and overly long segments, which could dilute the discernibility of geometric features' effects. Although fixed-size segmentation offers the flexibility of adjusting segment length to influence crash distribution, our approach allows us to reflect real-world road geometries, capturing nuances that fixed-length segments might overlook, without heavily relying on segment length as a determining factor for analysis.

Upon examining the crash frequency histogram for our homogeneously segmented data, as illustrated in **Figure 18**, it becomes evident that a significant majority of segments, 80.07%, recorded no crashes, with an additional 10.62% documenting a single crash occurrence. This distribution, notably skewed towards lower crash frequencies, can be attributed to several factors inherent to our dataset's nature. Our dataset's extensive coverage, encompassing over 600 roads, leads to a diverse array of road segments. This variety results in a substantial number of segments—specifically 6,067 in our homogeneous segmentation approach compared to 937 segments for a fixed 0.1-mile segment length analysis. Such a disparity highlights the challenges of uniform size segmentation in accurately capturing road geometries.

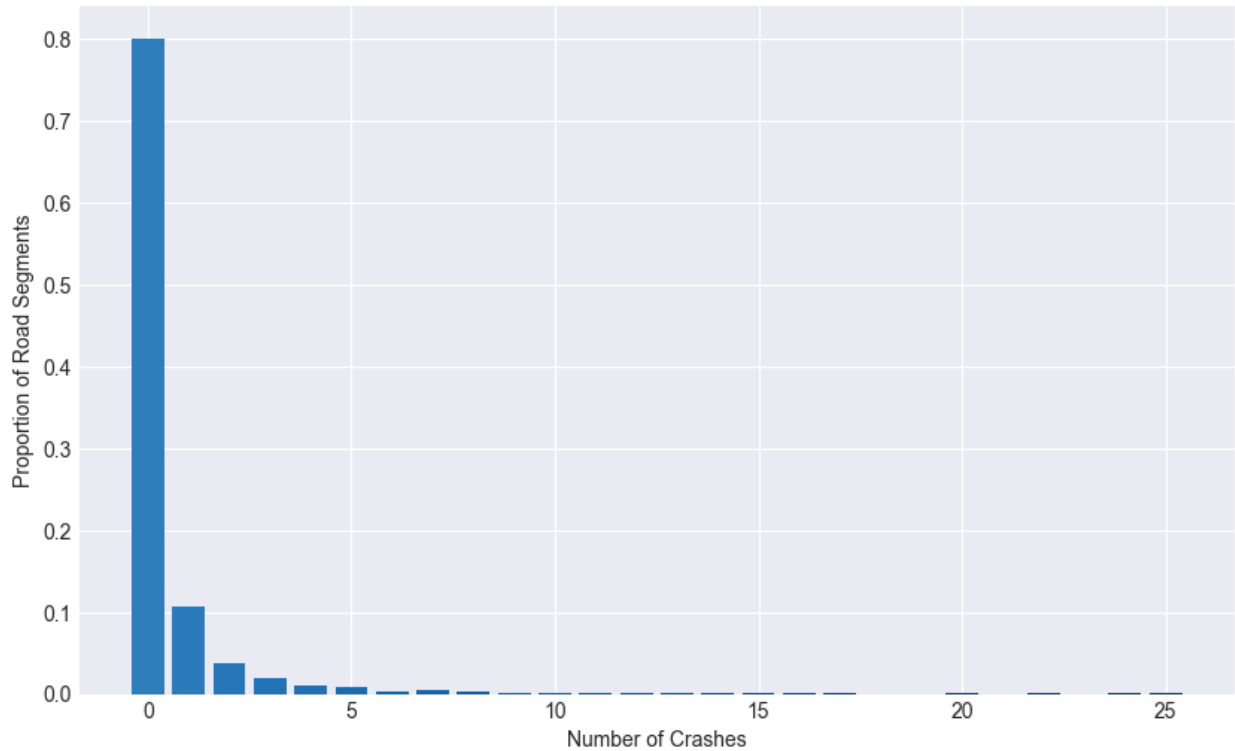


Figure 18: Proportion of road segments by number of crashes

Additionally, the exclusion of crashes occurring significantly away from the available road network, beyond a certain threshold, due to the absence of corresponding road GPS data further influences our crash data size, reducing the total count to 2903 crashes, including 997 bicycle crashes and 1906 pedestrian crashes. This combination of factors contributes to the high proportion of segments with zero crashes observed in our analysis.

3.1 Crash Frequency Count Models

3.1.1 Generalized Linear Models of Count Data

In addressing the complex nature of crash frequency data, the selection of an appropriate statistical modeling approach is paramount. This research employs Generalized Linear Models (GLM) as the foundational methodology [43]. Generalized Linear Models (GLM) are indeed widely used as a

foundational methodology for analyzing crash frequency data in transportation research. Crash frequency data are count-based and often exhibit properties such as overdispersion and non-normal distributions, making GLMs a suitable choice because they can handle such complexities. GLMs extend traditional linear models by allowing for response variables to have error distribution models other than a normal distribution.

Two primary distributions considered within the GLM framework for count data are the Poisson [44] and negative binomial [45]. The initial modeling effort typically begins with the Poisson regression model, predicated on the assumption that the mean and variance of crash counts are equal. While this model offers simplicity and interpretability, its applicability is contingent upon the absence of overdispersion in the crash data. Pearson's chi-squared test for over dispersion yielded a dispersion statistic of 2.791, significantly greater than 1, indicating notable overdispersion in the dataset.

Should empirical evidence indicate the presence of overdispersion, as is often the case with crash frequency data, the analysis progresses to the negative binomial regression model. This model introduces an additional parameter to account for the excess variance, providing a more flexible and realistic modeling of crash counts. The negative binomial model's ability to accommodate overdispersion not only enhances the accuracy of estimates but also ensures the reliability of inferential statistics derived from the model.

The Zero-Inflated Negative Binomial (ZINB) model [46] emerges as a statistical solution tailored to address the unique challenges presented by crash frequency data, notably the prevalent occurrence of zero crash counts. This model handles the dual phenomena often observed in crash data: segments that inherently record no crashes due to intrinsic safety (true zeros) and segments

where crashes are possible but none were observed during the study period (excess zeros). This differentiation is critical in crash data analysis, as it allows for a nuanced understanding that some roadway segments may be inherently safer, exhibiting no crash events, whereas others may simply not have recorded crashes by chance, despite the potential risk.

Having discussed the foundational aspects of Generalized Linear Models (GLM) for analyzing crash frequency data, this study progresses by adopting the Negative Binomial and Zero-Inflated Negative Binomial (ZINB) models. This decision stems from identifying inherent limitations within conventional models, such as the Poisson model, which exhibit inadequacies in fully capturing the dataset's characteristics, especially when confronted with a notable overdispersion and presence of zero counts.

Subsequent sections include discussions on the Negative Binomial and Zero-Inflated Negative Binomial models as outlined in Reference [47].

3.1.1.1 Negative binomial regression

A key limitation of Poisson regression stems from its reliance on the Poisson distribution, necessitating equal mean and variance in count data. When variance exceeds the mean, indicating over-dispersion, the Negative Binomial model becomes a preferable alternative [48]. Negative Binomial regression models the expected crash frequency for a given segment i as a function of various explanatory variables. These variables are aggregated into a vector x_i , allowing for the estimation of crash frequencies based on specific segment characteristics.

$$\lambda_i = \exp(\beta x_i + \varepsilon_i) \tag{22}$$

where ε_i represents the error term associated with segment i . Typically, it is assumed that (ε_i) follows a Gamma distribution with a mean of one and a variance of α . This incorporation of the error term ε_i permits the variance of n_i to deviate from its mean, allowing for greater flexibility in modeling the dispersion of crash frequencies.

$$VAR[n_i] = E[n_i][1 + \alpha E[n_i]] \quad (23)$$

$VAR[n_i]$ represents the variance of the count data n_i for segment i . It indicates how much the count data spread out or vary from the expected count $E[n_i]$. $E[n_i]$ is the expected count (or mean) for segment i , as predicted by the model. α is the dispersion parameter of the Negative Binomial distribution.

The formula for calculating the probability of a specific crash frequency n_i using the Negative Binomial distribution is given by:

$$P(n_i) = \left(\frac{\Gamma\left(n_i + \frac{1}{\alpha}\right)}{n_i!} \right) \times \left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i} \right)^{\frac{1}{\alpha}} \times \left(\frac{\lambda_i}{\frac{1}{\alpha} + \lambda_i} \right)^{n_i} \quad (24)$$

In this equation, $\Gamma(\cdot)$ denotes the Gamma function, and α represents the overdispersion parameter, which adjusts for the variance exceeding the mean in the count data. The model's parameters (β) can be estimated by maximum likelihood.

3.1.1.2 Zero-inflated negative binomial regression

This approach, as outlined in [49], effectively distinguishes between inherent safety and low-probability events within crash data analysis [48], employing a dual-model framework that

incorporates both a binary variable C_i for zero occurrence and a count model n_i^* for crash frequencies.

$$\begin{cases} n_i = 0, & \text{if } C_i = 1 \\ n_i^*, & \text{if } C_i = 0 \end{cases} \quad (25)$$

If we denote w_i as the probability indicating that segment i is completely safe ($C = 1$), then the probability function for n_i can be articulated as:

$$P(n_i) = w_i d_i + (1 - w_i) g(n_i) \quad (26)$$

Here, d_i is calculated as $1 - \min \{n_i, 1\}$ and $g(n_i)$ serves as a function for the probability of count data, employing either Poisson or Negative Binomial distributions [49]. Typically, a binary logit model is employed to represent the binary status inherent in the two components of a zero-inflated model. Consequently, the probability density function of a Zero-Inflated Negative Binomial model is formulated as a blend of these two processes, as follows [50], [51]:

$$P(n_i) = w_i + (1 - w_i) \left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i} \right)^{\frac{1}{\alpha}}, \text{ if } n_i = 0 \quad (27)$$

$$P(n_i) = (1 - w_i) \left(\frac{\Gamma\left(n_i + \frac{1}{\alpha}\right) \times \left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}} \times \left(1 - \frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^n}{\Gamma\left(\frac{1}{\alpha}\right) n_i!} \right), \text{ if } n_i = 1, 2, 3, \dots \quad (28)$$

First, equation calculates the probability of observing no events (zero count) for a given observation in the dataset. It combines the probability of a segment being absolutely safe (w_i) with

the probability of no events occurring due to chance, as per a negative binomial distribution. Second equation calculates the probability of observing a positive number of events (one or more counts) for a given observation. It uses the negative binomial distribution to account for overdispersion, adjusted by the probability of the segment not being perfectly safe ($1-w_i$). Despite consistent progress in the field of crash frequency modeling, the effectiveness of these models is constrained by the databases available for parameter estimation [47]. Furthermore, there is no universally accepted guideline that determines the superiority of one analytical method over another in the analysis of crash data [52].

Chapter 4: RESULTS

In this section, we have utilized Generalized Linear Models (GLMs) to develop models for crash frequency prediction. The focus was placed on employing the Negative Binomial Regression Model and the Zero-Inflated Negative Binomial Regression Model due to their effectiveness in handling overdispersion and excess zeros in count data. Findings include parameter estimates, standard errors, and t-statistic along with model fit indicators such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These results provide the groundwork for a detailed examination of the influence of road features on crash frequency, which will be elaborated on in the following sections.

Prior to the model development, an examination of the correlations among explanatory variables was conducted, utilizing the correlation heatmap depicted in **Figure 19**. This analysis was aimed at addressing potential efficiency issues due to multicollinearity, thereby avoiding the inclusion of highly correlated variables. The heatmap provides a visual representation of the linear relationships between variable pairs, with coefficients nearing +1 or -1 indicating strong linear relationships, either positive or negative, respectively. Coefficients approaching 0 imply a minimal linear relationship.

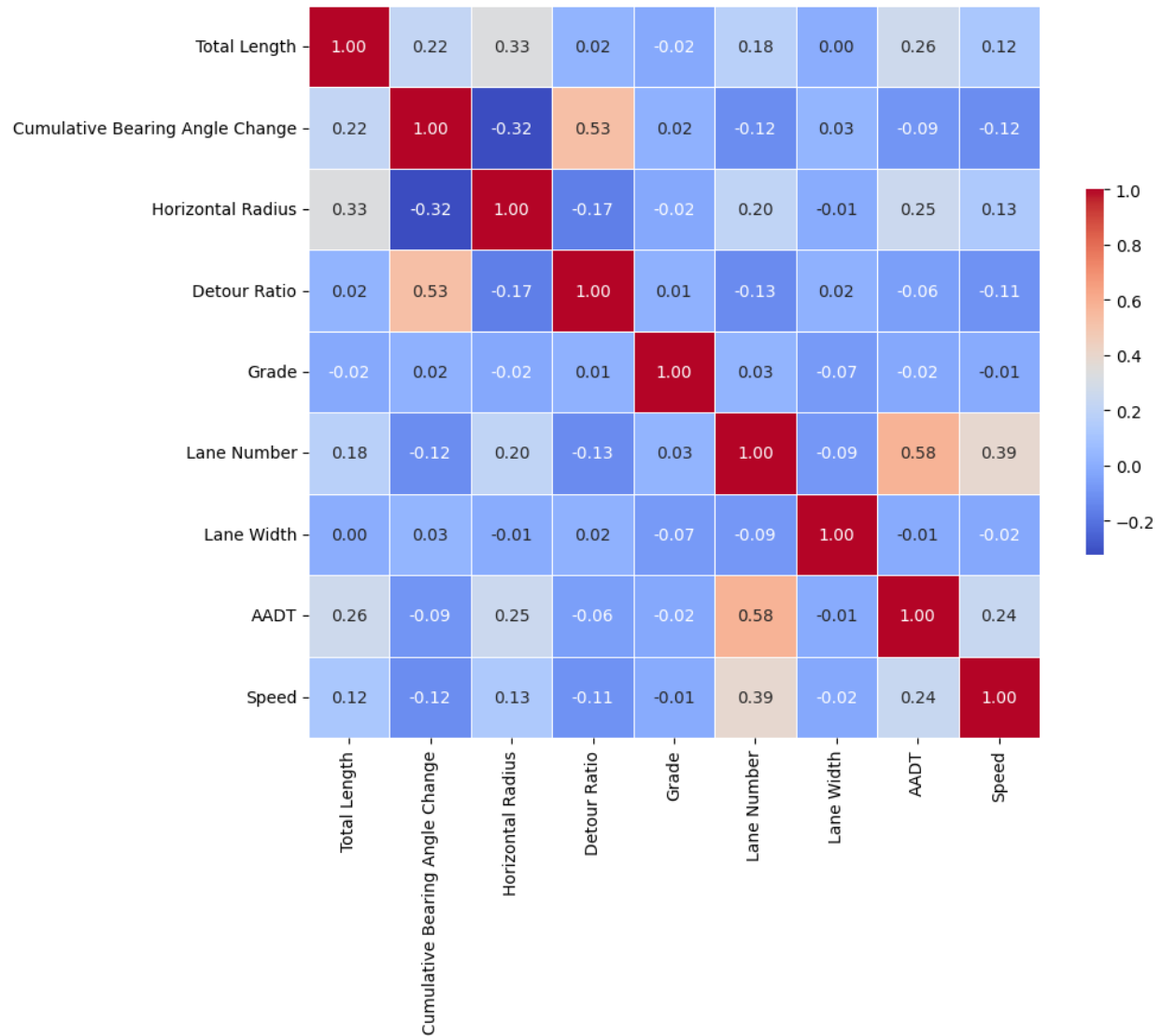


Figure 19: Correlation Matrix Heatmap of Road Segment Features

From this analytical process, several observations were made:

LANENUM and AADT: A moderate correlation of 0.578 between the number of lanes (LANENUM) and Annual Average Daily Traffic (AADT) is expected, as highways with more lanes are typically designed to accommodate higher traffic volumes. This correlation suggests that both variables, while related, offer unique contributions to the model, reflecting road capacity and traffic volume without inducing significant multicollinearity concerns.

LANENUM and SPEED: The moderate correlation (0.386) between lane numbers and speed limits aligns with the understanding that roads with more lanes, such as highways, often allow for higher speed limits. This correlation is logical but not so strong as to diminish the distinct impacts of road size and permitted speeds on crash frequency.

Detour Ratio and Cumulative Bearing Change: The correlation of 0.527 between detour ratio and cumulative bearing change aligns with expectations, reflecting a moderate relationship where segments that deviate more from a straight path tend to require longer routes compared to the direct distance between their endpoints. Despite this relationship, the correlation level is moderate and does not necessitate excluding either variable from the model.

The correlations, ranging from low to moderate, underscore that the explanatory variables possess enough independence for their collective integration into the model, thereby ensuring a robust analysis without the undue influence of multicollinearity.

4.1 Model Results

In this analysis, we explore the results of the Negative Binomial Regression Model and the Zero-Inflated Negative Binomial Regression Model, employing a nested model approach across three progressive stages of model development. Each stage incrementally incorporates a broader set of variables to assess their collective impact on crash frequency.

Initially, we explore the influence of horizontal road alignment features. Subsequently, we enhance our models by integrating vertical alignment factors, offering a more dimensional view of road geometry's effects. Finally, we expand our analysis to include a broader spectrum of road and traffic attributes, aiming to capture different aspects of road characteristics that contribute to crash

occurrences. The approach embodies the principle of nested models by allowing for the evaluation of model fit and complexity incrementally, facilitating a deeper understanding of the dynamics at play in road safety analysis.

4.1.1 Negative Binomial Regression Model Results

4.1.1.1 Model 1: Crash Frequency Analysis with Horizontal Alignment

The initial Model 1 focuses exclusively on horizontal alignment parameters, reflecting the fundamental aspects of road geometry. The estimation results, displayed in **Table 3**, include key parameters: Horizontal Radius, Total Length, Cumulative Bearing Change, Detour Ratio, with selected parameters undergoing logarithmic transformation: $\log(\text{Total Length})$, $\log(\text{Horizontal Radius})$. The logarithmic term illustrates how variations in parameters such as road length or horizontal radius affect crash frequency. They help in handling non-linear relationships, which makes the effect of changes in the predictor on the response more uniform across its range. Log transformations can stabilize the variance of a variable, particularly when data exhibit heteroscedasticity. It scales down large values more than smaller ones, reducing the impact of outliers and extreme values which could disproportionately affect the model's results.

Table 3: Estimation Result of Horizontal Alignment Parameters in Negative Binomial Model 1

Variable	Coefficient	Std. Error	T-Statistic
Constant	-8.291	0.471	-17.603
Log(Total Length)	0.721	0.035	20.597
Horizontal Radius	-0.0002	1.76e-05	-11.364
Log(Horizontal Radius)	0.616	0.056	11.002
Cumulative Bearing Change	-0.0002	0.001	-0.2
Detour Ratio	-0.278	0.180	-1.554
Number of Observations	4097		
Log-Likelihood Constants Only	-4728.620		
Log-Likelihood at Convergence	-4230.291		
AIC	8472.582		
BIC	8510.490		

4.1.1.2 Model 2: Enhanced Model with Horizontal and Vertical Alignments

Building upon the initial model, the Model 2 incorporates vertical alignment parameters to capture the combined effects of road alignment on crash frequency. The estimation results for this enhanced model are illustrated in **Table 4** and include additional parameters. The absolute value of the road grade, indicating steepness and the squared value of the grade, capturing non-linear effects of steepness on crash frequency. This approach indicates that the connection between road

steepness and crash risk does not follow a straightforward linear pattern. Instead, it can vary—either increasing or decreasing—at different levels of steepness, revealing a parabolic relationship. The omission of the grade's sign in our analysis is attributable to the unavailability of crash direction data. This constraint necessitates a more generalized approach to evaluating the impact of road steepness on safety outcomes.

Table 4: Estimation Result of Horizontal and Vertical Alignment Parameters in Negative Binomial Model 2

Variable	Coefficient	Std. Error	T-Statistic
Constant	-6.928	0.477	-14.524
Log(Total Length)	0.682	0.035	19.497
Horizontal Radius	-0.0002	1.8e-05	-11.111
Log(Horizontal Radius)	0.519	0.057	9.110
Cumulative Bearing Change	-0.0002	0.001	-0.2
Detour Ratio	-0.290	0.175	-1.656
Absolute (Grade (%))	-1.155	0.087	-13.272
Square (Grade (%))	0.194	0.031	6.242
Number of Observations	4097		
Log-Likelihood Constants Only	-4728.620		
Log-Likelihood at Convergence	-4080.450		
AIC	8176.9		
BIC	8227.444		

4.1.1.3 Model 3: Extended Analysis Including Road Feature Parameters

The Model 3 extends the analysis by including key road features along with the previously considered horizontal and vertical alignment parameters. The estimation results, showcased in **Table 5**, introduce the following road feature parameters: Log(AADT) Speed, Lane Number, Lane Width. Typically, in crash frequency and severity modeling, the log transformation of AADT is employed to normalize the distribution of traffic volumes and to mitigate the influence of extreme values. This relationship captures the diminishing effect on crash frequency with larger values of this parameter, which is typical for relationships that grow in a decreasingly significant way, such as the case with traffic volume where the addition of a certain number of vehicles has a smaller impact on crash risk in already heavy traffic conditions compared to light traffic conditions.

Table 5: Estimation Result of Road Features along with Horizontal and Vertical Alignment Parameters in Negative Binomial Model 3

Variable	Coefficient	Std. Error	T-Statistic
Constant	-11.453	0.695	-16.479
Log(Total Length)	0.691	0.037	18.687
Horizontal Radius	-0.0001	1.95e-05	-5.128
Log(Horizontal Radius)	0.386	0.059	6.539
Cumulative Bearing Change	0.001	0.001	0.7
Detour Ratio	0.038	0.157	0.24
Absolute (Grade (%))	-0.844	0.094	-8.983
Square (Grade (%))	0.134	0.035	3.837
Log(AADT)	0.369	0.043	8.581
Speed (mph)	-0.029	0.003	-9.733
Lane Width (feet)	-0.007	0.025	-0.26

Lane Number	0.1045	0.021	4.976
Function Class : Local	0.8994	0.375	2.398
Function Class : Minor Collector	2.0371	0.174	11.707
Function Class: Major Collector	2.2240	0.159	13.987
Function Class : Minor Arterial	2.5295	0.152	16.641
Function Class : Principal Arterial-Other	2.1710	0.152	14.2782
Function Class : Principal Arterial-Other Freeways & Expressways	0.8251	0.216	3.820
Number of Observations	4097		
Log-Likelihood Constants Only	-4728.620		
Log-Likelihood at Convergence	-3758.966		
AIC	7553.932		
BIC	7667.656		

4.1.1.4 Estimation Results Interpretation

Log(Total Length): The positive and highly significant coefficient ($p < 2e-16$) for log-transformed total length indicates that longer road segments are associated with an increase in crash frequency, possibly due to a greater exposure to risk over longer distances. The statistical significance highlights the robustness of this relationship across the dataset, suggesting that road length is a key factor in predicting crash occurrences.

Horizontal Radius: Its negative sign and significant result suggest that as the radius of a curve decreases (making the curve sharper), crash frequency tends to increase. This negative relationship

indicates that sharper curves, which are inherently more challenging to navigate, are associated with a higher risk of crashes.

Log (Horizontal Radius): The positive sign and significant result for the log-transformed radius suggest an opposite trend for less sharp curves. As the radius increases, making curves less pronounced, the frequency of crashes also increases. This positive relationship can be seen as reflecting a different set of risk factors associated with straighter roads or gently curving roads, such as higher speeds or less driver attentiveness.

The contrast between the negative sign for Horizontal Radius and the positive sign for Log (Horizontal Radius) highlights the complex relationship between road geometry and crash frequency. It indicates that both extremely sharp curves and straighter segments or gently curving roads have their own unique sets of challenges and risks contributing to crash frequency. Sharp curves may lead to crashes due to the immediate difficulty in navigation, while straighter roads might encourage higher speeds or decreased vigilance, leading to different types of crashes. This dual finding emphasizes the need for tailored safety measures that consider the specific characteristics of road geometry to effectively mitigate crash risks.

The Cumulative Bearing Change: While this parameter exhibits a positive relationship with crash frequency, suggesting that greater directional changes might intuitively correlate with increased crashes, it is not statistically significant in the model (p-value of 0.297). This indicates that the cumulative changes in road direction do not significantly impact crash frequency within the context of this study. It suggests that crash occurrences may be more strongly associated with specific road features or individual curve characteristics, rather than the aggregate change in direction.

The Detour Ratio: Despite showing a positive coefficient, suggesting a potential increase in crash frequency with greater road indirectness, this parameter does not reach statistical significance in the model (p-value of 0.8104).

Absolute (Grade (%)): The negative coefficient indicates that, initially, as roads become steeper, crash frequency tends to decrease. This could be because drivers adopt more cautious behaviors on steeper grades, reducing the likelihood of crashes.

Square (Grade (%)): The positive coefficient for the squared term suggests a non-linear relationship: while steeper grades initially lead to fewer crashes, there's a point where further increases in steepness begin to increase crash frequency. This could be due to the added navigational challenges and potential for loss of vehicle control on very steep roads, which might override the cautionary effect observed with moderate steepness.

This indicates that the relationship between road grade and crash frequency is indeed complex and nonlinear. Initially, steeper grades may deter risky driving behaviors, but as steepness increases beyond a certain threshold, the inherent dangers of navigating such terrain outweigh the cautionary benefits.

Log (AADT): This variable's significant correlation with crash frequency particularly emphasizes the heightened risk to pedestrians and bicyclists on roads with higher traffic volumes. It suggests that increased vehicle flow on these roads elevates the potential for dangerous interactions between vehicles and more vulnerable road users, leading to a higher incidence of crashes. This relationship highlights the critical need for enhanced safety measures and infrastructure on busier roads to protect pedestrians and bicyclists. It is worth mentioning that the higher number of crashes

observed on roads with increased AADT could also be attributed simply to the greater exposure to traffic.

Speed: The observed significant negative correlation between speed limits and crash frequency implies that roads with higher speed limits tend to have a lower incidence of crashes. This finding suggests that the characteristics and safety features inherent in roads designed for higher speeds might be contributing factors to this trend, rather than a direct causative link between higher speeds and reduced crash occurrences. Additionally, considering the focus of this study on pedestrian and bicycle crashes, it's pertinent to acknowledge that areas typically frequented by these road users often have lower speed limits.

Lane Width: Although the negative coefficient might suggest that wider lanes are associated with fewer accidents, the lack of significant correlation (p-value of 0.7953) indicates that within this study's dataset, variations in lane width do not significantly affect crash frequency.

Lane Number: The significant positive correlation between lane number and crash frequency underscores a multifaceted relationship. Roads with more lanes not only suggest busier traffic environments, drawing higher volumes of vehicles, but also introduce complexities in multi-lane navigation that elevate the risk for pedestrians and bicyclists. This increased traffic density and the inherent challenges of crossing and maneuvering through multi-lane roads heighten their exposure to potential accidents.

Function Class: When analyzing the function class of roads with Interstate Roads as the baseline, we can infer how different types of roads correlate with crash frequencies, particularly considering their traffic dynamics and infrastructure characteristics associated with each class.

1. Interstate Roads (Baseline)

Designed for high-speed, long-distance travel with minimal access points, these roads generally report lower crash frequencies due to their streamlined traffic flow and limited interactions with non-vehicular traffic, serving as the baseline for comparison.

2. Principal Arterial-Other Freeways & Expressways

These roads, while similar to interstates in terms of intended use for high-speed travel, have a slightly higher crash frequency, as indicated by the coefficient of 0.8251. This could be due to slightly more frequent access points and interactions with local traffic compared to interstates.

3. Principal Arterial-Other

Serving major traffic movements with more direct access to land compared to freeways, these arterial roads show a higher crash frequency (coefficient of 2.1710). The increased interactions with urban environments and crossing traffic can contribute to this rise.

4. Minor Arterial

These roads facilitate movement within smaller regions and connect to the arterial network, showing even higher crash frequencies (coefficient of 2.5295). Minor arterials balance access and mobility but introduce higher crash risks due to more frequent intersections and a mix of traffic speeds.

5. Major Collector

Connecting local roads to arterials, Major Collectors have a significant positive coefficient (2.2240), indicating higher crash frequencies, likely due to mixing various types of traffic and serving both mobility and access functions.

6. Minor Collector

These roads, with a coefficient of 2.0371, indicate a substantial increase in crash frequencies, potentially due to their function in collecting traffic from local roads and interfacing with a variety of road users, including pedestrians and bicyclists.

7. Local Roads

The coefficient for Local Roads (0.8994) shows an increase in crash frequency compared to Interstate Roads but is less significant than other road types. This could imply that, despite their lower speeds and frequent intersections, local roads might inherently be safer due to less complex traffic dynamics or more cautious driving behavior in more densely populated or residential areas.

The observed pattern of crash coefficients across varying road function classes demonstrates a parabolic trend, as visualized in **Figure 20**.

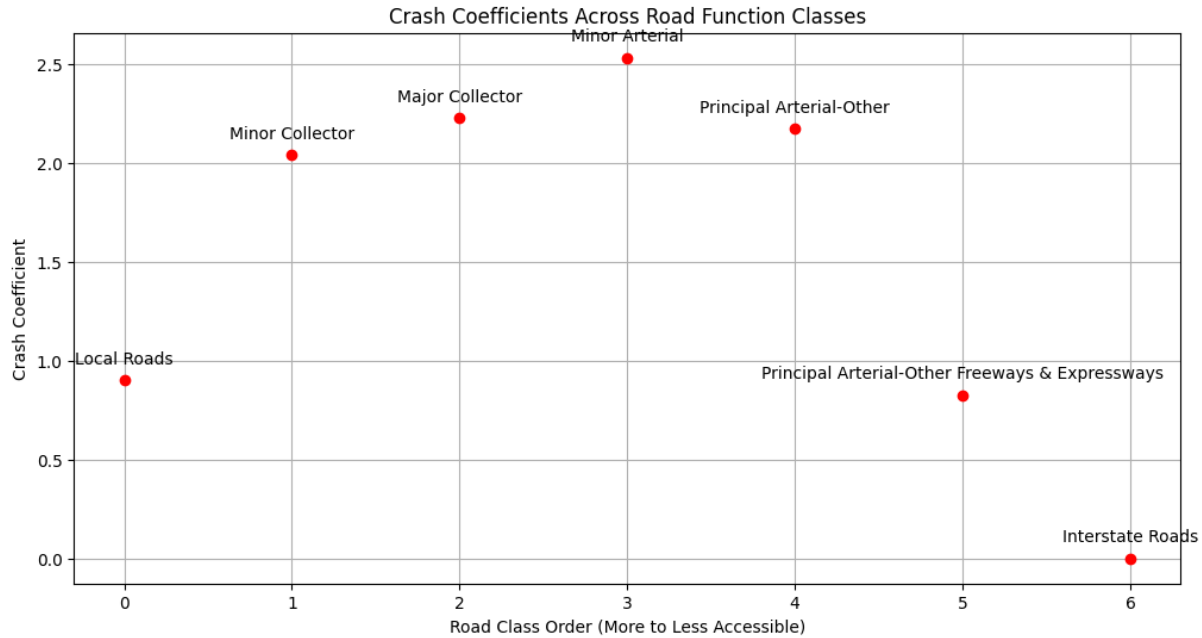


Figure 20: Crash Coefficients (NB Model 3) Across Road Function Classes

Starting from local roads, there is an initial increase in crash coefficients. Local roads often have complex traffic dynamics with frequent intersections and pedestrian crossings, which can elevate the risk of crashes. Where lower speeds and more cautious driving prevail despite higher accessibility, resulting in a reduced, yet still significant, crash frequency compared to interstates.

As we advance to the minor and major collector roads, the crash coefficients continue to rise. These roads typically handle traffic from residential areas and funnel them towards larger arterial roads, increasing the likelihood of crashes due to the blend of local and through traffic.

The trend peaks at the minor arterial roads, where the crash coefficients are highest. This peak reflects the intricate balance these roads strike between providing access to local destinations and

facilitating through traffic, leading to a higher frequency of crashes due to the diversity of traffic flows and the presence of more conflict points and interactions with urban settings.

Moving forward, the curve starts to descend towards the principal arterial-other freeways and expressways. These road classes are designed for higher-speed travel and often have controlled access, which reduces direct conflict points and, consequently, the crash coefficients.

Finally, the curve reaches its lowest point at the interstate roads. As the high-speed, limited-access highways designed primarily for long-distance travel, interstates exhibit the lowest crash coefficients, likely due to their streamlined traffic flow, fewer access points, and absence of intersections, which collectively contribute to reduced crash occurrences compared to other road types.

4.1.1.5 Likelihood Ratio Tests

We evaluated the significance of each successive model enhancement using Likelihood Ratio Tests (LRT). These statistical tests provide a quantitative measure to compare the fit of nested models, allowing us to assess the contribution of added variables to our understanding of crash frequency dynamics. The results are presented in **Table 6**, suggesting significant improvements in model fit with the addition of variables at each stage. The LRT is calculated based on the difference in log-likelihoods between a pair of models, according to the formula:

$$LRT = -2\log\left(\frac{\mathcal{L}_r}{\mathcal{L}_u}\right) = -2(\log(\mathcal{L}_r) - \log(\mathcal{L}_u)) \quad (29)$$

Where \mathcal{L}_r represents the likelihood of the restricted model, and \mathcal{L}_u denotes the likelihood of the unrestricted model.

Table 6: Statistical Evaluation of Negative Binomial Model Enhancements through Likelihood Ratio Tests

Model Configurations Compared	$\log(\mathcal{L}_r)$	$\log(\mathcal{L}_u)$	LRT Statistic	Degrees of Freedom (df)
Null Model and Model 1	-4728.620	-4230.291	996.656	5
Model 1 and Model 2	-4230.291	-4080.450	299.682	2
Model 2 and Model 3	-4080.450	-3758.966	642.967	10

4.1.2 Zero-Inflated Negative Binomial Regression Model Results

We replicate this nested model approach with the Zero-Inflated Negative Binomial (ZINB) Models. In the ZINB model, we include the same set of variables in both count and logit components to ensure an evaluation of how each variable influences not only the frequency of crashes but also the likelihood of zero occurrences.

The estimation results for each developmental stage of the ZINB model are detailed in **Table 7 through 9**. The model incorporating only horizontal alignment parameters is showcased in **Table 7**. The expanded model, which includes both horizontal and vertical alignment parameters, is depicted in **Table 8**. Lastly, the most comprehensive model, integrating key road features alongside the alignment parameters, is illustrated in **Table 9**.

4.1.2.1 Model 1: Crash Frequency Analysis with Horizontal Alignment

Table 7: Estimation Result of Horizontal Alignment Parameters in Zero-Inflated Negative Binomial Model 1

	Count Model (Negative Binomial with log link)			Zero-Inflation Model (Binomial with logit link)		
Variable	Coefficient	Std. Error	T-Statistic	Coefficient	Std. Error	T-Statistic
Constant	-0.531	0.083	-6.393	-2.074	0.707	-2.935
Log(Total Length)	0.319	0.066	4.797	-1.596	0.413	-3.867
Horizontal Radius	-0.419	0.101	-4.136	0.074	0.517	0.143
Log(Horizontal Radius)	0.845	0.128	6.623	0.916	0.950	0.963
Cumulative Bearing Change	0.078	0.0815	0.955	0.470	0.270	1.741
Detour Ratio	-0.092	0.073	-1.265	-1.814	3.657	-0.496
Log(theta)	-1.056	0.095	-11.082	-	-	-
Number of Observations	4097					
Log-Likelihood Constants Only	-3914.744					
Log-Likelihood at Convergence	-3721.033					

AIC	7468.066
BIC	7550.200

4.1.2.2 Model 2: Enhanced Model with Horizontal and Vertical Alignments

Table 8: Estimation Result of Horizontal and Vertical Alignment Parameters in Zero-Inflated Negative Binomial Model 2

Variable	Count Model (Negative Binomial with log link)			Zero-Inflation Model (Binomial with logit link)		
	Coefficient	Std. Error	T-Statistic	Coefficient	Std. Error	T-Statistic
Constant	-0.638	0.096	-6.637	-2.055	0.720	-2.85
Log(Total Length)	0.304	0.067	4.527	-1.581	0.327	-4.84
Horizontal Radius	-0.399	0.101	-3.953	-0.0220	0.599	-0.033
Log(Horizontal Radius)	0.748	0.125	5.987	1.182	1.171	1.008
Cumulative Bearing Change	0.092	0.081	1.139	0.484	0.198	2.420
Detour Ratio	-0.104	0.0740	-1.411	-1.383	2.263	-0.610
Absolute (Grade (%))	-0.778	0.108	-7.205	0.661	0.346	1.91
Square (Grade (%))	0.402	0.111	3.635	-0.263	0.339	-0.776

Log(theta)	-0.94	0.103	-9.185	-	-	-
Number of Observations	4097					
Log-Likelihood Constants Only	-3914.744					
Log-Likelihood at Convergence	-3650.747					
AIC	7335.494					
BIC	7442.900					

4.1.2.3 Model 3: Extended Analysis Including Road Feature Parameters

Table 9: Estimation Result of Road Features along with Horizontal and Vertical Alignment Parameters in Zero Inflated Negative Binomial Model 3

Variable	Count Model (Negative Binomial with log link)			Zero-Inflation Model (Binomial with logit link)		
	Coefficient	Std. Error	T-Statistic	Coefficient	Std. Error	T-Statistic
Constant	-2.788	0.305	-9.148	-6.554	4.469	-1.467

Log(Total Length)	0.543	0.047	11.543	-0.893	0.197	-4.530
Horizontal Radius	-0.344	0.078	-4.420	0.316	0.316	1
Log(Horizontal Radius)	0.414	0.092	4.484	0.523	0.491	1.067
Cumulative Bearing Change	0.091	0.062	1.462	-0.609	0.435	-1.4
Detour Ratio	-0.071	0.211	-0.337	1.985	1.133	1.753
Absolute (Grade (%))	-0.687	0.091	-7.543	1.179	0.793	1.487
Square (Grade (%))	0.331	0.095	3.466	-1.277	1.145	-1.115
Log(AADT)	0.411	0.065	6.329	-0.832	0.398	-2.091
Speed	-0.290	0.047	-6.228	0.371	0.183	2.021
Lane Width	0.019	0.039	0.492	-0.428	0.567	-0.754
Lane Number	0.14185	0.04801	2.955	-0.4205	0.5573	-0.755
Function Class : Local	1.03495	0.4695	2.204	3.2764	6.14	0.534
Function Class : Minor Collector	1.86949	0.2154	8.679	2.9982	3.3827	0.886
Function Class: Major Collector	2.10836	0.19633	10.739	2.49	3.593	0.693
Function Class : Minor Arterial	2.37974	0.18902	12.59	3.2585	3.4412	0.947
Function Class : Principal Arterial-Other	1.98535	0.19004	10.447	2.9649	3.3859	0.876

Function Class : Principal Arterial-Other Freeways & Expressways	0.69212	0.27479	2.519	4.021	3.6749	1.094
Log(theta)	-0.86732	0.06154	-14.094	-	-	-
Number of Observations	4097					
Log-Likelihood Constants Only	-3914.744					
Log-Likelihood at Convergence	-3484.261					
AIC	7042.522					
BIC	7276.288					

4.1.2.4 Estimation Results Interpretation

Reflecting on the count model component of the Zero-Inflated Negative Binomial (ZINB) model, we note its consistency with the Negative Binomial model estimations in terms of the sign and significance of parameters. We shift our focus on the zero-inflation part of the ZINB model. The subsequent analysis will center on interpreting the dynamics influencing zero-crash occurrences. We highlight the factors that significantly influence the likelihood of observing zero-crash counts.

Log (Total Length): The significantly negative coefficient indicates that longer road segments are less likely to have zero-crash counts. This suggests that the increased length of a road correlates with more frequent crash occurrences, reducing the probability of a zero-crash scenario.

Log (AADT): The negative coefficient for Log (AADT) is significant, which suggests that as average annual daily traffic increases, the likelihood of reporting zero crashes decreases. This reinforces the notion that busier roads have a higher chance of reporting crashes, aligning with the count model's finding that higher traffic volumes are associated with an increase in crash frequency.

Speed: The positive and significant coefficient for speed in the zero-inflation part indicates that road segments with higher speed limits are more likely to have zero crashes reported. This finding is consistent with the idea that roads with higher speed limits, which tend to be designed for vehicles rather than pedestrians and bicyclists, have fewer crashes involving vulnerable road users.

4.1.2.5 Likelihood Ratio Tests

LRT were also applied to evaluate enhancements in the Zero-Inflated Negative Binomial (ZINB) model, with results presented in **Table 10**.

Table 10: Statistical Evaluation of Zero-Inflated Negative Binomial Model Enhancements through Likelihood Ratio Tests

Model Configurations Compared	$\log(\mathcal{L}_r)$	$\log(\mathcal{L}_u)$	LRT Statistic	Degrees of Freedom (df)
Null Model and Model 1	-3914.744	-3721.033	387.42	10
Model 1 and Model 2	-3721.033	-3650.747	140.57	4
Model 2 and Model 3	-3650.747	-3484.261	332.97	20

4.1.3 Negative Binomial and Zero-Inflated Negative Binomial Comparison

Both models show significant improvements with the addition of variables at each stage, as evidenced by the LRT statistics. The Negative Binomial model exhibits larger LRT values, suggesting more pronounced improvements in fit with each successive model compared to the ZINB model. The ZINB model, even in its simplest form (Null model), starts with a higher log-likelihood than the Negative Binomial model. This suggests that the ZINB model, by considering zero inflation from the start, may already align more closely with the underlying distribution of the crash data, particularly in datasets where zero-crash segments are significant. It acknowledges that not all road segments have the same risk or conditions leading to crashes, and some may inherently

be more prone to reporting zero crashes, not just because of lower risk but possibly due to factors like low traffic volume.

When evaluating the performance of fitted models, in addition to log-likelihood, it's essential to consider criteria that balance the model's fit with its complexity. Two widely used criteria for this purpose are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), calculated as follows:

$$AIC = -2LL + 2P \quad (30)$$

$$BIC = -2LL + P \cdot \ln(n) \quad (31)$$

In these equations, LL represents the log-likelihood of the fitted model, P denotes the number of parameters within the model, and n stands for the number of observations. Both AIC and BIC introduce a penalty for increasing the number of parameters, thereby mitigating the risk of overfitting. The penalty is more substantial in the case of BIC, especially as the number of observations (n) grows, which can make BIC a stricter criterion for model selection. This is because the penalty term in BIC is $P \cdot \ln(n)$, where P is the number of parameters and n is the number of observations. The logarithmic function of n means that as the dataset gets larger, the penalty for adding more parameters to the model increases more steeply in BIC than in AIC, where the penalty term is simply $2P$.

A model with lower AIC and BIC values is generally preferred, as it suggests an optimal balance between fitting the data well and maintaining a simpler model structure. **Table 11** compares the AIC and BIC statistics for the two models, providing insight into which model achieves a better balance between goodness of fit and complexity.

Table 11: Model AIC and BIC Statistics Comparison

Model Description	Log-Likelihood	AIC	BIC
Final NB Model (Model 3)	-3758.966	7553.932	7667.656
Final ZINB Model (Model 3)	-3484.261	7042.522	7276.288

When comparing the AIC and BIC between the final Negative Binomial (NB) and Zero-Inflated Negative Binomial (ZINB) models, we observe that the ZINB model outperforms the NB model in both metrics. Lower AIC and BIC values suggest a model that provides a better fit to the data while efficiently balancing model complexity. This indicates that the ZINB model, even with its additional parameters, manages to explain the variability in the data more effectively than the NB model.

Chapter 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

This thesis represents an effort to enhance our understanding of road safety, specifically focusing on crashes involving vulnerable road users in the State of Hawaii, Oahu, for the period from 2015 to 2022. It encompasses two main parts: firstly, the development of an automated framework for road segmentation; and secondly, the application of Negative Binomial (NB) and Zero-Inflated Negative Binomial (ZINB) methodologies to assess the relationship between crash occurrences and various road attributes.

The most important conclusions and contributions derived from this research are as follows:

A contribution of this study is its approach to examining the impact of road features on crash occurrences across variably sized segments determined by road alignment characteristics, specifically targeting data that involves vulnerable road users in Hawaii. To the best of the author's knowledge, this angle of inquiry has not been explored previously, making it a notable addition to road safety research.

By developing an automated framework for segmenting roads based on their alignment using GPS data, this approach enables a detailed analysis of crash frequencies at the segment level without the need to access specialized datasets on road alignment, which are often limited in availability and scope. This methodology can be universally applied wherever GPS data of roads are available, offering a repeatable approach for conducting road safety studies. It allows researchers to customize road segmentation according to the unique requirements of their studies.

In terms of model evaluation, the zero-inflated model presented better statistics in comparison with the negative binomial regression. The findings from NB and ZINB models underscored the complex interplay between road alignment and safety, offering insights into how different aspects of road design and traffic dynamics can exacerbate or mitigate crash occurrences.

Our analysis emphasizes the significant impact of both horizontal and vertical road alignments on crash frequencies. Key factors such as the total length of road segments, the radius of curves, and the steepness of roads were identified as having substantial effects on crash occurrences. Additionally, traffic dynamics factors, including speed and average annual daily traffic (AADT), as well as road characteristics such as the number of lanes, emerged as crucial determinants of crash frequency.

Furthermore, the examination of road function classes and their correlation with crash frequencies provides detailed insights into the distinct safety profiles of different road types, from local streets to interstate highways. This analysis underscores the importance of developing customized safety interventions that address the unique challenges and characteristics of each road class, aiming to improve road safety across a variety of traffic environments.

5.2 Practical Implementation

This study's findings contribute to our understanding of road safety offering valuable insights for policymakers and safety practitioners. Identifying key roadway features that affect crash rates allows for the creation of targeted safety interventions. The utilization of island-wide road segmentation based on alignment enables the application of these insights to similar road types, fostering a safer environment for all road users, particularly the most vulnerable. To further enhance road safety, several interventions are suggested:

Traffic Calming in High-Risk Areas: Implementing measures such as speed humps, raised crosswalks and narrowed lanes can significantly lower vehicle speeds, reducing the risk of accidents in areas prone to high crash frequencies.

Enhanced Crossing Facilities for Pedestrians and Bicyclists: Improve crossing facilities at intersections and mid-block locations on roads with higher crash coefficients, particularly where minor arterials transition to local streets. Install pedestrian refuge islands, and dedicated bicycle lanes to protect vulnerable road users.

Infrastructure Upgrades: Upgrade road infrastructure to include protected bicycle lanes and pedestrian pathways, ensuring safe travel for non-motorized road users. Improve lighting along pathways and crossings to enhance visibility at night.

Segment-Specific Interventions: Utilize the segmentation based on road alignment to implement targeted interventions. For segments with sharp curves (as indicated by a smaller horizontal radius), enhance road markings, and install warning signs to improve visibility and awareness.

Data-Driven Road Safety Audits: Conduct road safety audits focusing on segments identified with unique alignment characteristics contributing to higher crash rates. These audits can help identify potential hazards and suggest remedial actions based on empirical data.

5.3 Limitation and Future Direction

To further advance road safety research, the following areas offer promising avenues for future studies and enhancements in data improvements:

Crash Type: Enhancing the crash dataset to focus more on types of crashes that are directly influenced by road alignment, such as roadway departure incidents, could yield more precise insights into how horizontal and vertical alignments impact crash occurrences. Crashes involving pedestrians and bicyclists provide valuable information but may be more concentrated in specific areas, potentially limiting the scope of analysis on road alignment's broader effects.

Crash Direction: Incorporating data on the direction of crashes would allow for a more nuanced analysis, including the potential to consider the impact of road grade direction on crash frequency. This addition would enable the examination of how the orientation of slopes may affect safety, offering a deeper understanding of the interplay between road geometry and crash risks.

Segment Sequence Analysis: Extending the analysis to include not just the segment where a crash occurs but also preceding segments could reveal patterns in how sequences of road features contribute to crash risks. By analyzing the roads, a driver traverses before a crash, research could uncover specific combinations of road characteristics that elevate the likelihood of incidents, guiding more targeted interventions. This approach necessitates incorporating crash direction data.

Comprehensive Road Feature Dataset: Developing a more detailed dataset that encompasses a wider range of road features, including those not directly related to alignment, would support a holistic approach to road safety analysis. This could include factors such as lighting quality, signage, and road surface conditions, each of which plays a role in shaping the safety landscape of road networks.

Chapter 6: REFERENCES

- [1] “Global Status Report on Road Safety 2023.” Accessed: Mar. 09, 2024. [Online]. Available: <https://iris.who.int/bitstream/handle/10665/375016/9789240086517-eng.pdf?sequence=1>
- [2] “National Highway Traffic Safety Administration (NHTSA)- Fatality Analysis Reporting System (FARS) History.” Accessed: Mar. 09, 2024. [Online]. Available: <https://www-fars.nhtsa.dot.gov/Trends/TrendsGeneral.aspx>
- [3] R. Tay, J. Choi, L. Kattan, and A. Khan, “A multinomial logit model of pedestrian-vehicle crash severity,” *Int J Sustain Transp*, vol. 5, no. 4, 2011, doi: 10.1080/15568318.2010.497547.
- [4] A. Behnood and F. Mannering, “Determinants of bicyclist injury severities in bicycle-vehicle crashes: A random parameters approach with heterogeneity in means and variances,” *Anal Methods Accid Res*, vol. 16, 2017, doi: 10.1016/j.amar.2017.08.001.
- [5] X. Zhai, H. Huang, N. N. Sze, Z. Song, and K. K. Hon, “Diagnostic analysis of the effects of weather condition on pedestrian crash severity,” *Accid Anal Prev*, vol. 122, 2019, doi: 10.1016/j.aap.2018.10.017.
- [6] J. K. Kim, S. Kim, G. F. Ulfarsson, and L. A. Porrello, “Bicyclist injury severities in bicycle-motor vehicle accidents,” *Accid Anal Prev*, vol. 39, no. 2, 2007, doi: 10.1016/j.aap.2006.07.002.
- [7] J. K. Kim, G. F. Ulfarsson, V. N. Shankar, and F. L. Mannering, “A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model,” *Accid Anal Prev*, vol. 42, no. 6, 2010, doi: 10.1016/j.aap.2010.04.016.
- [8] S. Boufous, L. De Rome, T. Senserrick, and R. Ivers, “Risk factors for severe injury in cyclists involved in traffic crashes in Victoria, Australia,” *Accid Anal Prev*, vol. 49, 2012, doi: 10.1016/j.aap.2012.03.011.
- [9] P. Chen and J. Zhou, “Effects of the built environment on automobile-involved pedestrian crash frequency and risk,” *J Transp Health*, vol. 3, no. 4, 2016, doi: 10.1016/j.jth.2016.06.008.
- [10] P. Chen and Q. Shen, “Built environment effects on cyclist injury severity in automobile-involved bicycle crashes,” *Accid Anal Prev*, vol. 86, 2016, doi: 10.1016/j.aap.2015.11.002.
- [11] B. E. Hagel *et al.*, “The relationship between visibility aid use and motor vehicle related injuries among bicyclists presenting to emergency departments,” *Accid Anal Prev*, vol. 65, 2014, doi: 10.1016/j.aap.2013.12.014.
- [12] K. Haleem, P. Alluri, and A. Gan, “Analyzing pedestrian crash injury severity at signalized and non-signalized locations,” *Accid Anal Prev*, vol. 81, 2015, doi: 10.1016/j.aap.2015.04.025.
- [13] E. Robartes and T. D. Chen, “The effect of crash characteristics on cyclist injuries: An analysis of Virginia automobile-bicycle crash data,” *Accid Anal Prev*, vol. 104, 2017, doi: 10.1016/j.aap.2017.04.020.

- [14] M. Rella Riccardi, F. Galante, A. Scarano, and A. Montella, “Econometric and Machine Learning Methods to Identify Pedestrian Crash Patterns,” *Sustainability (Switzerland)*, vol. 14, no. 22, 2022, doi: 10.3390/su142215471.
- [15] G. Vandebulcke, I. Thomas, and L. Int Panis, “Predicting cycling accident risk in Brussels: A spatial case-control approach,” *Accid Anal Prev*, vol. 62, 2014, doi: 10.1016/j.aap.2013.07.001.
- [16] S. Narayanamoorthy, R. Paleti, and C. R. Bhat, “On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level,” *Transportation Research Part B: Methodological*, vol. 55, 2013, doi: 10.1016/j.trb.2013.07.004.
- [17] C. Siddiqui, M. Abdel-Aty, and K. Choi, “Macroscopic spatial analysis of pedestrian and bicycle crashes,” *Accid Anal Prev*, vol. 45, 2012, doi: 10.1016/j.aap.2011.08.003.
- [18] C. Hamann and C. Peek-Asa, “On-road bicycle facilities and bicycle crashes in Iowa, 2007-2010,” *Accid Anal Prev*, vol. 56, 2013, doi: 10.1016/j.aap.2012.12.031.
- [19] J. P. Schepers, P. A. Kroeze, W. Sweers, and J. C. Wüst, “Road factors and bicycle-motor vehicle crashes at unsignalized priority intersections,” *Accid Anal Prev*, vol. 43, no. 3, 2011, doi: 10.1016/j.aap.2010.11.005.
- [20] L. Chen, C. Chen, R. Srinivasan, C. E. McKnight, R. Ewing, and M. Roe, “Evaluating the safety effects of bicycle lanes in New York City,” *Am J Public Health*, vol. 102, no. 6, 2012, doi: 10.2105/AJPH.2011.300319.
- [21] G. Prati, V. M. Puchades, M. De Angelis, F. Fraboni, and L. Pietrantonio, “Factors contributing to bicycle–motorised vehicle collisions: A systematic literature review,” *Transport Reviews*, vol. 38, no. 2. 2018. doi: 10.1080/01441647.2017.1314391.
- [22] X. Yan, M. Ma, H. Huang, M. Abdel-Aty, and C. Wu, “Motor vehicle-bicycle crashes in Beijing: Irregular maneuvers, crash patterns, and injury severity,” *Accid Anal Prev*, vol. 43, no. 5, 2011, doi: 10.1016/j.aap.2011.04.006.
- [23] E. Jiménez-Mejías, V. Martínez-Ruiz, C. Amezcua-Prieto, R. Olmedo-Requena, J. D. D. Luna-Del-Castillo, and P. Lardelli-Claret, “Pedestrian- and driver-related factors associated with the risk of causing collisions involving pedestrians in Spain,” *Accid Anal Prev*, vol. 92, 2016, doi: 10.1016/j.aap.2016.03.021.
- [24] D. D. Obinguar and M. Iryo-Asano, “Macroscopic analysis on the frequency and severity of pedestrian crashes on National Roads in Metro Manila, Philippines,” *IATSS Research*, vol. 45, no. 4, 2021, doi: 10.1016/j.iatssr.2021.06.003.
- [25] J. Su, N. N. Sze, and L. Bai, “A joint probability model for pedestrian crashes at macroscopic level: Roles of environment, traffic, and population characteristics,” *Accid Anal Prev*, vol. 150, 2021, doi: 10.1016/j.aap.2020.105898.
- [26] S. Ukkusuri, S. Hasan, and H. M. A. Aziz, “Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency,” *Transp Res Rec*, no. 2237, 2011, doi: 10.3141/2237-11.

- [27] M. Rahman, K. M. Kockelman, and K. A. Perrine, “Investigating risk factors associated with pedestrian crash occurrence and injury severity in Texas,” *Traffic Inj Prev*, vol. 23, no. 5, 2022, doi: 10.1080/15389588.2022.2059474.
- [28] D. Jima and T. Sipos, “The Impact of Road Geometric Formation on Traffic Crash and Its Severity Level,” *Sustainability (Switzerland)*, vol. 14, no. 14, 2022, doi: 10.3390/su14148475.
- [29] R. Baireddy, H. Zhou, and M. Jalayer, “Multiple Correspondence Analysis of Pedestrian Crashes in Rural Illinois,” *Transp Res Rec*, vol. 2672, no. 38, 2018, doi: 10.1177/0361198118777088.
- [30] B. Bartin, S. Demiroglu, K. Ozbay, and M. Jami, “Automatic Identification of Roadway Horizontal Alignment Information Using Geographic Information System Data: CurvS Tool,” in *Transportation Research Record*, vol. 2676, no. 1, 2022. doi: 10.1177/03611981211036364.
- [31] M. Bíl, R. Andrášik, J. Sedoník, and V. Cícha, “ROCA – An ArcGIS toolbox for road alignment identification and horizontal curve radii computation,” *PLoS One*, vol. 13, no. 12, 2018, doi: 10.1371/journal.pone.0208407.
- [32] H. Xu and D. Wei, “Improved identification and calculation of horizontal curves with geographic information system road layers,” *Transp Res Rec*, vol. 2595, pp. 50–58, 2016, doi: 10.3141/2595-06.
- [33] “ArcGIS Online Database.” Accessed: Mar. 28, 2024. [Online]. Available: <https://maps.arcgis.com/index.html>
- [34] “Selenium Python Library.” Accessed: Mar. 08, 2024. [Online]. Available: <https://pypi.org/project/selenium/>
- [35] “Mandli Communications Website”, Accessed: Mar. 28, 2024. [Online]. Available: <https://www.mandli.com/>
- [36] “Mandli Positional System”, Accessed: Mar. 28, 2024. [Online]. Available: <https://www.mandli.com/solutions/positional/>
- [37] A. Savitzky and M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures,” *Anal Chem*, vol. 36, no. 8, 1964, doi: 10.1021/ac60214a047.
- [38] S. Dabiri, C. T. Lu, K. Heaslip, and C. K. Reddy, “Semi-supervised deep learning approach for transportation mode identification using GPS trajectory data,” *IEEE Trans Knowl Data Eng*, vol. 32, no. 5, 2020, doi: 10.1109/TKDE.2019.2896985.
- [39] R. Andrášik and M. Bíl, “Efficient road geometry identification from digital vector data,” *J Geogr Syst*, vol. 18, no. 3, 2016, doi: 10.1007/s10109-016-0230-1.
- [40] Mohammadreza Hashemi, “Influence of Roadway Characteristics in The Modeling of The Frequency of Roadway Departure Crashes on Two-Lane Two-Way State Roads,” University of Hawaii at Manoa, 2019.
- [41] N. Xiao, *GIS Algorithms: Theory and Applications for Geographic Information Science & Technology*. 2016. doi: 10.4135/9781473921498.

- [42] “Highway Performance Monitoring System Field Manual.” Accessed: Mar. 28, 2024. [Online]. Available: <https://gis.transportation.wv.gov/ftp/TMA/Districts/9-23-2021%20Draft%20HPMS%20Field%20Manual%202021.pdf>
- [43] A. Agresti, *Foundations of Linear and Generalized Linear Models*. 2015.
- [44] S. P. Miaou, “The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions,” *Accid Anal Prev*, vol. 26, no. 4, 1994, doi: 10.1016/0001-4575(94)90038-8.
- [45] V. Shankar, F. Mannering, and W. Barfield, “Effect of roadway geometrics and environmental factors on rural freeway accident frequencies,” *Accid Anal Prev*, vol. 27, no. 3, 1995, doi: 10.1016/0001-4575(94)00078-Z.
- [46] J. Lee and F. Mannering, “Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis,” *Accid Anal Prev*, vol. 34, no. 2, 2002, doi: 10.1016/S0001-4575(01)00009-4.
- [47] D. Lord and F. Mannering, “The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives,” *Transp Res Part A Policy Pract*, vol. 44, no. 5, 2010, doi: 10.1016/j.tra.2010.02.001.
- [48] S. P. Washington, M. G. Karlaftis, F. L. Mannering, and P. Anastasopoulos, *Statistical and econometric methods for transportation data analysis, Second edition*. 2010. doi: 10.1201/9781420082869.
- [49] R. Winkelmann, *Econometric analysis of count data*. 2008. doi: 10.1007/978-3-540-78389-3.
- [50] D. Lord, S. P. Washington, and J. N. Ivan, “Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory,” *Accid Anal Prev*, vol. 37, no. 1, 2005, doi: 10.1016/j.aap.2004.02.004.
- [51] D. Lord, S. Washington, and J. N. Ivan, “Further notes on the application of zero-inflated models in highway safety,” *Accid Anal Prev*, vol. 39, no. 1, 2007, doi: 10.1016/j.aap.2006.06.004.
- [52] F. L. Mannering and C. R. Bhat, “Analytic methods in accident research: Methodological frontier and future directions,” *Anal Methods Accid Res*, vol. 1, 2014, doi: 10.1016/j.amar.2013.09.001.