2023

TRI-DIMENSIONAL SEGMENT-BASED URBAN TRAFFIC CRASH MODELS

Farzin Maniei

TRI-DIMENSIONAL SEGMENT-BASED URBAN TRAFFIC CRASH MODELS

By

FARZIN MANIEI

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2023

UNIVERSITY OF
TEXAS
ARLINGTON

ACKNOWLEDGEMENT

I would like to thank my supervising professor, Dr. Stephen P. Mattingly, for his insight, encouragement, and support. I also want to thank my committee members, Dr. Kate Hyun, Dr. Taylor Li, and Dr. Victoria Chen, for their thoughtful ideas and valuable guidance. This work depended on the traffic count data provided by TxDOT Traffic Analysis and System Support. I am grateful to Laura Dabalain and Eric Oeding for supporting this research by providing the traffic count data. I would like to acknowledge partial support from the National Institute for Transportation and Communities (NITC; grant number #1578).

I would also like to express my deep gratitude to all my friends and fellow graduate students at UTA, who always helped and supported me during my research. I thank my family, especially my dear aunt, Dr. Farah Maniei, and my dear uncle, Dr. Ali Abolmaali, for their constant support, love, and encouragement. In particular, I thank my lovely parents, Farrokh Maniei and Sholeh Abolmaali, and my dear brothers Shervin and Ramin for their unconditional love and tremendous support. This dissertation would not have been possible without your help and support.

November 3rd , 2023

ABSTRACT

TRI-DIMENSIONAL SEGMENT-BASED URBAN TRAFFIC CRASH MODELS

Farzin Maniei, Ph.D.

The University of Texas at Arlington, 2023

Supervising Professor: Stephen P. Mattingly, Ph.D.

The continuous expansion of highway and freeway networks further exacerbates the risk of traffic crashes and highlights the critical importance of freeway safety management. To improve overall road safety, many organizations acknowledge the necessity of pinpointing locations experiencing higher-than-expected crashes (known as hotspot identification, HSID), identifying the factors contributing to traffic crashes, and determining the most effective preventive measures. Previous studies highlighted two major drawbacks associated with this approach: (1) HSID based on the total number of crashes can lead to incorrectly identifying hazardous areas; (2) the arbitrary selection of a fragment size (due to the lack of explicit recommendations) used for dividing the highway and freeways into small segments to aggregate data may affect the factors that correlate with crash rates in predictive models. This study addresses the urgent need for investigating the merits of expanding traffic crash analysis from total crashes to traffic crash subsets and providing a standard approach for recommending the fragment size when aggregating crash groups and roadway data based on three crash characteristics (i.e. crash units, manner of collision, and crash severity). The study performs feature selection with a unique approach that harnesses the Laplacian score joined with a distance-

based entropy measure, called LSDBEM. followed by an unsupervised clustering method, K-means clustering, to provide a recommended fragment size (RFS) for data aggregation. The LSDBEM is utilized to satisfy prior to clustering. After the feature selection, the method applies an unsupervised clustering method, K-means clustering, to capture the pattern of traffic crashes on freeways within Dallas County. The investigation considers the LSDBEM/K-means method for fragment sizes ranging from 0.10 mile to 0.25 mile with an increment of 0.01 mile. To evaluate the use of crash features or the total crash rate (TCR) to establish the clustering pattern and the recommended fragment size (RFS), the study compares the LSDBEM/K-means method results for TCR and FCRs. The dissertation assesses the impacts of using higher dimensions of traffic crash characteristics (across four different scenarios) for crash prediction models on model performance, the statistical significance of crash contributing factors, and the identification of crash hot spots. The investigation of higher dimensions of traffic crash characteristics estimates many count data regression models including Poisson, negative binomial (NB), negative binomial type P (NBP), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), zero-inflated negative binomial type P (ZINBP), generalized Poisson type 1 (GP-1), generalized Poisson type 2 (GP-2), and Hurdle regression models.

The dissertation evaluates the performance and suitability of the RFS across the four scenarios of traffic crash characteristic dimensions. This analysis estimates crash prediction models with fragment sizes ranging from 0.10 mile to 0.25 mile with 0.01-mile increments. The evaluation focuses on comparing crash prediction model performance using the root mean square error (RMSE) for the testing dataset. The investigation determines the circumstances that support the adoption of the RFS as the standardized approach for data aggregation.

The study results show that LSDBEM/K-means clustering method provides a standardized approach to determine a recommended fragment size for data aggregation. The clustering results demonstrate that FCR-based clustering creates more cohesive clusters than TCR-based clustering, which promotes the use of three traffic crash dimensions for safety analysis and modeling. The additional crash dimensions indicate substantially different top ten hotspots for each crash group, especially when compared to hotspots identified using the total number of crashes (scenario 1). The crash prediction models for scenario 4, formed by all three dimensions, provides a better understanding of the crash mechanisms, but scenario 4 may not always work for all crash groups due to insufficient observations. The investigation of RFS reveals that it minimizes the multicollinearity among the explanatory variables. The evaluation of the testing RMSE shows that the minimum RMSE ($RMSE_{min}$) occurs at the RFS for some SV-related and MV-related crash groups. Moreover, the crash groups with sufficient non-zero observations generate a RMSE for the RFS ($RMSE_{RFS}$) remains within the proximity (20%) of $RMSE_{min}$, which makes the RFS an acceptable approach for standardized data aggregation when sufficient non-zero observations exist. The future studies need to confirm the benefit of the RFS for data aggregation, its appropriate use cases, and its impact on crash prediction models by examining other highways and freeways.

TABLE OF CONTENTS

LIST OF ILLUSTRATION

LIST OF TABLES

CHAPTER 1. INTRODUCTION AND BACKGROUND

1.1. BACKGROUND

Traffic crashes stand as a major cause of both fatalities and injuries on a global scale (World Health Organization (WHO), 2018). With the rapid expansion of highway and freeway networks, the potential for traffic crashes also rises, underscoring the vital importance of managing freeway safety. Many agencies recognize the need to identify crash contributing factors and locations experiencing higher than expected crashes; they may use this information to enhance overall road safety by locating hazardous areas and identifying and prioritizing effective preventive measures.

These traffic crash hotspots denote places where traffic crashes happen more frequently, or the probability of crashes is notably higher than nearby spots along a specific route or throughout a network. Identifying these hotspots (HSID) enables targeted interventions to reduce the likelihood of crashes and enhance safety in these risky zones by comprehending the factors contributing to such crashes. One of the popular approaches to HSID uses crash prediction models based on the total number of crashes. Previous research indicates two major shortcomings related to this approach: (1) HSID based on the total number of crashes may result in misidentifying the hazardous areas; (2) the predictive models at the aggregate level suffer from the arbitrary selection of the segment length used for dividing the highway and freeways into small segments to aggregate data.

Although many HSID studies concentrate on the total crash count, a handful of recent investigations identify crash hotspots by traffic crash subsets, such as the number of vehicles involved in crashes (crash units), the manner of collision, and the crash severity. The findings from these HSID studies demonstrate that crash hotspots identified using any of these traffic crash

characteristics diverge from those identified based on the total number of crashes. Excluding any of these three characteristics could disrupt the comprehension of traffic crashes, the factors at play, and the identification of hotspots (Wang & Feng, 2019). However, the merits of simultaneously including all three traffic crash characteristics require evaluation due to a lack of quality data on all three characteristics and computation complexity. This study investigates the effect of including all three traffic crash characteristics in traffic safety studies.

As shown in previous studies, the hotspots for single-vehicle (SV) crashes and multi-vehicle crashes may occur in different locations along a corridor with minor overlaps while HSID based on total number of crashes fails to detect some of those hotspots (Wang & Feng, 2019). Moreover, different crash groups likely vary in the crash contributing factors or in the intensity of their impact. For instance, multi-vehicle sideswipe crashes more likely occur when narrower lanes, poor pavement, and low visibility exist, but single-vehicle fixed-object related crashes are more likely when no paved shoulder is provided, or shoulder width is narrow.

The low frequency of traffic crashes forces crash prediction models to rely heavily on the aggregation of diverse datasets, encompassing traffic crash, traffic characteristics, and geometric characteristics data (Wang & Feng, 2019). In most studies, data aggregation requires dividing highways or freeways into small segments using a consistent length known as the "segment length." However, this study employs the term "fragment size" to avoid any potential confusion with the geometric attributes of the highway, like the distance between ramps. Earlier work (Pedregosa, et al., 2011) criticizes the arbitrary selection of fragment size (for aggregating crash data in crash frequency analysis because as Ahmed and Abdel-Aty (2012) notes a change in the fragment size for data aggregation can result in changes in the statistical significance of explanatory variables. This instability of the explanatory variable significance indicates that crash

frequency analysis based on total crashes may capture correlated patterns rather than causal factors. Despite the evident effects of fragment size (segment length) on the aggregation of traffic crash data, no explicit recommendations exist to guide the selection or determination of the fragment size (segment length) for aggregating crash-related data. These concerns demonstrate an urgent need to consider crash subsets and to establish a standard approach for establishing the fragment size when identifying hot spots.

To ensure appropriate data aggregation for urban/suburban highways and freeways, it is advised not to employ a segment length smaller than 0.1 mile, as recommended by the American Association of State Highway and Transportation Officials (2010). Similarly, a spacing interval exceeding 0.25 mile for traffic operational characteristics is discouraged, following the guidance of the Alabama Department of Transportation (2015).

This research seeks to address the major shortcomings associated with HSID using predictive models and other crash frequency modeling: (a) arbitrary selection of fragment size for roadway segmentation; (b) non-comprehensive consideration of some of crucial traffic crash characteristics. The study investigates the effect of fragment size used to segmentize a roadway for aggregating data and clustering roadway segments by including three crucial traffic crash characteristics for crash prediction models and HSID: the number of vehicles involved in the crash (crash units), manner of collision (crash type), and crash severity simultaneously. Also, it introduces a method to define a recommended fragment size (RFS) to overcome the arbitrary selection of fragment size (segment length) because a standardized approach for aggregation may support future metanalyses across crash prediction models to identify contributing factors more explicitly rather than correlated patterns in a particular dataset. In addition, this study investigates the effect of crash dimensions in identifying crash hotspots and contributing factors by defining four scenarios using

traffic crash characteristics. The study compares the prediction model performances across four scenarios using the Akaike Information Criterion (AIC) and investigates the top ten segments hotspots for each crash group to discover the differences or commonalities in HSID results across the four scenarios. Using the four-scenario structure, the research evaluates the impacts of transitions between scenario levels on the intensity and significance of the contributing factors and the merits of using all three crash dimensions when creating crash prediction models. Since the fragment size selection has a ripple effect in the modeling process, the study examines the impact of fragment size on the crash count prediction modeling process by performing prediction models for various fragment sizes ranging from 0.10 mile to 0.25 mile with an increment of 0.01 mile. Finally, the study investigates the potential benefit of the RFS to improve crash count prediction model performance and accuracy, leading to justifiable and reliable fragment size selection for data aggregation.

1.2. LITERATURE REVIEW

Many traffic safety studies focus on traffic crash hotspot identification (HSID) using Geographic Information Systems (GIS) based spatial analysis, statistical models, and machine learning. The GIS-based HSID conventionally provides crash concentration maps using Kernel Density Estimation (KDE) and absolute counts of crashes. The GIS-based HSID results carry two main issues: the accuracy of the concentration maps and only relying on total number of crash (Truong & Somenahalli, 2011). Among these approaches, statistical models (predictive models) hold substantial appeal in practice by offering engineers valuable insights into traffic crash contributing factors through a timely efficient computing method. The statistical model drawbacks include using total crash count for HSID and applying an arbitrary fragment size to divide a corridor into small fragments for aggregating data during the model estimation or application

process. HSID based on total crash count conceals some crucial information needed to properly identify hotspots, understand the crash contributing factors and propose effective countermeasures to mitigate traffic crashes. In fact, previous traffic crash investigations that rely on total crash counts have displayed limitations in identifying certain contributing factors and have shown a tendency for false positive outcomes in hotspot identification (Cheng, et al., 2017).

For understanding traffic crashes, previous research utilized different traffic crash dimensions including crash units, manner of collision, and crash severity. Previous research indicates that whether adopting an aggregated or disaggregated approach, crashes should be analyzed by focusing on crash units (i.e. single-vehicle (SV) and multi-vehicle (MV)) (Yu, et al., 2013). Another study confirms the need to incorporate crash units to distinctively delineate single-vehicle (SV) and multi-vehicle (MV) crash hotspots due to their different spatial distributions and underlying factors (Wang & Feng, 2019). Another essential characteristic of traffic crashes, the manner of collision (crash type), or the first event in a crash, is highlighted as a critical dimension to be included in the analysis of traffic crashes (Pande, et al., 2010) because crash type helps in revealing the underlying contributory factors associated with each specific crash type (Valent, et al., 2002). By incorporating crash types, many traffic safety studies reveal hidden details about traffic crashes (Golob, et al., 2004a; Cheng, et al., 2017). For instance, one study shows that rear-end traffic crashes concentrate in regimes with variable speed and heavily congested flow or free flow while lane-changing crashes concentrate in regimes of variable speed and only free flow (Golob, et al., 2004a). Crash severity introduces another complexity to understanding traffic crashes and previous studies demonstrate that collapsing all crashes into a single analysis may obscure the varying degrees of injury severity (Valent, et al., 2002). Compared to the total crash count, the level of significance or the magnitude of crash contributing factors may vary for

different crash severity (Zeng, et al., 2019). In addition, incorporating traffic crash dimensions helps to address some of the potential unobserved heterogeneity (Mannering & Bhat, 2014). While previous studies identify the importance of including the different crash dimensions in crash studies, this research specifically investigates the implications of including additional dimensions in a crash study through an innovative four scenario investigation of including additional crash dimensions. This dissertation evaluates the impact using modeling performance/HSID.

As aforementioned, the selection of fragment size (segment length) for roadway segmentation and data aggregation (including traffic and crash data) has a direct influence on traffic safety modeling results, potentially influencing the statistical significance of variables. Therefore, research needs to undertake an in-depth exploration into data aggregation using fragment size (segment length) and its impacts on HSID and crash modeling. Thomas (1996) highlights the impact of segment length (fragment size in this study) on a statistical description of crash count data, described as the "size problem," which results from the arbitrary selection of segment lengths for data aggregation. Several studies investigate other alternatives to roadway segmentation using roadway attributes including continuous risk profile (Kwon, et al., 2013), sliding moving window (Qin & Wellner, 2012; Kwon, et al., 2013), peak searching (Kwon, et al., 2013), fixed length and variable length segmentation (Koorey, 2009), clustering methods (Valent, et al., 2002; Depaire, et al., 2008; Lu, et al., 2013). However, no specific approach or guideline finds a recommended fragment size (RFS) for roadway segmentation and data aggregation. As an approach to roadway segmentation, the AASHTO Highway Safety Manual (2010) suggests that creating segments with consistent geometry and Annual Average Daily Traffic (AADT) could mitigate this concern. However, this approach introduces new challenges including producing inconsistent segment length (small segment lengths or very large segment length) and dependency on the universally

available quality data (on geometric characteristics and traffic operational characteristics) across all segments (Ghadi & Torok, 2019). Selecting segment lengths to aggregate crash data holds significance in identifying crash hotspots (Cook, et al., 2011) and maintaining the consistency of hotspot identification (Geyer, et al., 2008). Moreover, the outcomes of safety analyses can be influenced by selecting extraordinarily long or short roadway segments (Lu, et al., 2013). Despite its crucial importance, the need for more comprehensive guidance regarding the optimal fragment length for model performance still exists.

1.3. STUDY OBJECTIVES AND STRUCTURE

This study provides an innovative method to overcome the issue of arbitrary selection of fragment size by introducing a recommended fragment size (RFS) for data aggregation. Also, it evaluates the advantages of incorporating higher dimensions of traffic crash characteristics, including crash units, manner of collisions (crash type), and crash severity, in crash modeling, examining crash prediction models, significance of crash contributing factors, and examining HSID for different crash groups defined under four scenarios. In addition, the research examines the impact of fragment size (ranging from 0.10 mile to 0.25 mile with an increment of 0.01 mile) on crash prediction model performances and accuracy. The investigation evaluates the potential benefit of the RFS to improve crash prediction model results under the four crash modeling dimension scenarios to confirm its suitability for a standardized data aggregation method; this would resolve the shortcoming of traffic safety study results affected by the arbitrary selection of fragment size.

This dissertation consists of five chapters. Chapter 2 establishes a methodology to provide a recommended fragment size (RFS) for crash data aggregation to overcome the arbitrary selection of fragment size. This chapter calculates featured crash rates (FCRs) for different crash groups based on three crash characteristics (i.e. crash units, manner of collision, and crash severity). The

study performs feature selection with a unique approach that harnesses the Laplacian score joined with a distance-based entropy measure, called LSDBEM, followed by an unsupervised clustering method, K-means clustering, to provide a recommended fragment size (RFS) for data aggregation. The LSDBEM is utilized to satisfy prior to clustering. After the feature selection, the method applies an unsupervised clustering method, K-means clustering, to capture the pattern of traffic crashes on freeways within Dallas County. The investigation considers the LSDBEM/K-means method for fragment sizes ranging from 0.10 mile to 0.25 mile with an increment of 0.01 mile. To evaluate the use of crash features or the total crash rate (TCR) to establish the clustering pattern and the recommended fragment size (RFS), the study compares the LSDBEM/K-means method results for TCR and FCRs.

Chapter 3 examines the effect of higher dimensions of traffic crash characteristics on crash prediction models and crash hotspots identified using crash prediction models. To do this, the study defines four scenarios based on the dimensions of traffic crash characteristics involved to form crash groups with the RFS of 0.10 mile. This research estimates several count data regression models including Poisson, negative binomial (NB), negative binomial type P (NBP), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), zero-inflated negative binomial type P (ZINBP), generalized Poisson type 1 (GP-1), generalized Poisson type 2 (GP-2), and Hurdle regression models for all scenarios' crash groups. For each crash group, t the study determines the outperforming models based on AIC and uses them to identify hotspots. To deal with the traffic crash fluctuation, the study applies the empirical bayes method and the potential for safety improvement, known as PSI, to rank the traffic crash hotspots for each crash group. The study evaluates the modeling impact of the higher dimensions by comparing the crash prediction model performance and accuracy. The dissertation investigates crash hotspots under different scenarios

and reveals the effect of the higher dimensions of traffic crash characteristics on changes in the significance and magnitude of contributing factors. The new HSID methodology provides a strategy for identifying hotspots with specific contributing factors that may impact crash mitigation strategies.

Chapter 4 explores the impact of various fragment sizes on crash prediction model performance and accuracy under four scenarios and examines the potential benefit of RFS to the model performance, ultimately affecting the HSID results. The crash data for each scenarios' crash group is aggregated for fragment sizes ranging from 0.10 mile to 0.25 mile with an increment of 0.01. For each fragment size, the study estimates crash prediction models for each crash group and identifies the outperforming models. The chapter evaluates the RFS and identifies the study contexts where it should be adopted. Finally, the discussion, conclusions, and the future extension of work are provided in Chapter 5.

CHAPTER 2. Unsupervised Approach to Investigate Urban Traffic Crashes Based on Crash Unit, Crash Severity, and Manner of Collision

2.1. INTRODUCTION

Traffic crashes represent one type of **"incident,"** defined as an "unplanned randomly occurring traffic event that adversely affects normal traffic operation" (Wang & Feng, 2019). Previous studies arbitrarily select the segment length as a constant value between 0.1 mi. and 1.0 mi (or, in some studies, 100 m to 1.6 km) based on the study's objectives (Texas Department of Transportation (TxDOT) - Traffic Safety Division, 2020). Choosing different segment lengths for aggregation may result in some variables becoming either statistically significant or insignificant (Ahmed & Abdel-Aty, 2012). It is recommended not to use a segmentation length smaller than 0.1 mile (American Association of State Highway and Transportation Officials, 2010) or a spacing interval greater than 0.25 mile to segment and aggregate traffic data for urban/suburban highways and freeways (Alabama Department of Transportation, 2015); however, no specific method currently exists to select segment length. This paper adopts the term fragment size to avoid confusion because the tern "segment length" is used to refer to not only explanatory variable representing the length of roadway section in some studies but also the length selected to divide a roadway to smaller units for data aggregation in some other studies. This study proposes an innovative method to provide a recommended fragment size for data aggregation based on historical crash risk.

Since selecting of fragment size (segment length) for aggregation may cause variables to become statistically significant or insignificant, creating a standard methodology for selecting a suitable fragment size (segment length) for aggregation appears essential for future research. Previous studies argue that the selection of arbitrary fixed-size fragments (segments) for aggregating crash

data generates fundamental problems in crash frequency analysis (Pedregosa, et al., 2011). Previous research fails to provide any standardized guidance or methodology to select the fragment size (segment length) to aggregate crash data. Since the selection of fragment size (segment length) impacts traffic safety research, this study seeks to investigate and propose a method to find a recommended segment length.

Generally, safety studies can investigate traffic crashes based on different crash characteristic dimensions such as number of vehicles involved (Xu, et al., 2018), manner of collision (Cheng, et al., 2017; Bhowmik, et al., 2018; Mahmud & Gayah, 2021), and crash severity (Yu & Abdel-Aty, 2013a; Afghari, et al., 2020). This study also seeks to capture the crash patterns and transitions between crash combinations across highways based on three major traffic crash characteristics: number of vehicles involved in crashes (crash units), manner of collision, and crash severity, simultaneously.

The number of vehicles involved in a crash represents an important crash characteristic dimension that will affect the results of aggregate traffic crash analyses. Previous studies investigate traffic crashes based on number of vehicles involved by grouping the crashes into two categories: single-vehicle (SV) and multi-vehicle (MV) crashes because the crash contributing factors may differ or demonstrate different impacts for SV and MV crashes (Abdel-Aty, et al., 2006; Ivan, et al., 1999; Islam & Pande, 2020b). Yu and Abdel-Aty (Yu & Abdel-Aty, 2013a) show that the selected crash contributing factors have different impacts on SV and MV crashes and recommend that future safety analyses need to consider the number of vehicles involved as a traffic characteristic for both aggregate and disaggregate approaches. This study includes the number of vehicles involved in crashes by creating SV and MV categories for crash features. The manner of collision, which refers to the first event in a crash, represents another important traffic

crash characteristic. Some previous studies refer to the manner of collision as crash type and show that including the manner of collision (crash type) reveals facts about traffic crashes that traffic studies conducted based on total crashes would fail to recognize (Golob, et al., 2008). This and other studies (Islam, et al., 2017; Cheng, et al., 2017) support the importance of including the manner of collision in safety analyses; therefore, the authors integrate the manner of collision as another traffic crash feature dimension. Several studies also investigate the impact of traffic crash contributing factors on crash severity (Abdel-Aty M. A., 2003; Islam & Pande, 2020b). This study combines crash severity as a crash feature with the number of vehicles involved and the manner of collision to create a more refined crash combination than TCR.

This study investigates the effect of segment length for aggregating data and clustering roadway segments using the number of vehicles involved in the crash, manner of collision, and crash severity simultaneously. The clustering approach is selected for roadway segmentation because it can mitigate crash heterogeneity for within-group elements by grouping roadway segments with similar crash distributions into homogeneous groups, according to (Lu, et al., 2013). The focus on the crash characteristics makes grouping the data based on the crash characteristics critical for understanding patterns in the crash data. However, some temporal instability (Islam & Mannering, 2020a) and unobserved heterogeneity associated with environmental characteristics and driver behaviors (Islam, et al., 2020c) may affect the study result. To reduce computation complexities and ease implementation, the study excludes the temporal instability and unobserved heterogeneity associated with environmental characteristics and driver behaviors. The authors also propose a standard method to provide a recommended fragment size (RFS) for aggregating crash data that can be used as a foundation for all future traffic crash analyses requiring data aggregation, which

may reduce the impact of arbitrary selection of fragment size (segment length) on crash frequency analysis (CFA).

2.2. LITERATURE REVIEW

2.2.1. Fragment Size (Segment Length)

As aforementioned, selecting the segment length to aggregate traffic and crash data impacts both CFA and RTCPM since it may affect the variables' statistical significance; therefore, the impact of segment length on safety analyses requires further investigation. Thomas (Thomas, 1996) studies the effect of segment length on crash count and density. Thomas (Thomas, 1996) argues that the arbitrary selection of segment length to aggregate data creates an unaddressed problem called a "size problem". According to AASHTO Highway Safety Manual (2010), creating segments with consistent geometry and Annual Average Daily Traffic (AADT) may address this concern. However, it introduces new issues due to the inconsistent and small segment lengths and the need for universal data availability for all segments (Ghadi & Torok, 2019). Segment length selection to aggregate crash data impacts the identification of crash hotspots (Cook, et al., 2011) and affects the consistency of hotspot identification (Geyer, et al., 2008). Also, the safety analysis outcomes can be affected for both extremely long and short roadway segments (Lu, et al., 2013). Despite the importance of segmentation length, there is minimal guidance on segmentation.

2.2.2. Segmentation Approaches

Various approaches to segmentize a roadway using a subset of sources, including traffic data, roadway characteristics, and traffic crash data exist but a typical approach segments a roadway based on its characteristics to account for unobserved heterogeneity. However, roadway segmentation by roadway characteristics may lead to long segments since many roadways may have little to no variation in roadway attributes over a long stretch (Green, 2018). For example, a

very long segment length may occur because a long stretch of a highway has constant shoulder width, the number of lanes, cross slope, and median width on a straight section (Green, 2018). While a homogeneous long segment can be divided to smaller segments to redistribute traffic crashes into resulting smaller segments, dividing the homogeneous long segments into small segments may lead to an arbitrary selection of break points or selection of a (segment) length with no specific guidelines (Green, 2018). Besides, quality roadway characteristics data may not be available, requiring costly data collection. Other than roadway characteristics, traffic data can be used to develop a homogenous segment when variation in roadway attributes is negligible (Borsos, et al., 2014). Even though traffic data may help to divide long segments into smaller segments, it may not be helpful for roadways with limited access over a long distance due to minor changes in traffic volume (Green, 2018).

Other alternatives to roadway segmentation by roadway attributes exist. These alternatives include continuous risk profile (Kwon, et al., 2013), sliding moving window (Qin & Wellner, 2012; Kwon, et al., 2013), peak searching (Kwon, et al., 2013), fixed length and variable length segmentation (Koorey, 2009), clustering methods (Valent, et al., 2002; Depaire, et al., 2008; Lu, et al., 2013). Among these alternatives, the clustering techniques are beneficial for roadway segmentation using traffic crash data, especially when quality data on traffic and roadway attributes are unavailable because they may reveal undiscovered relationships in traffic crash data (De Luca, et al., 2012; Depaire, et al., 2008; Golob, et al., 2004a; Lu, et al., 2013). Valent et al. (2002) applied a clustering method using a specific crash type to analyze traffic crashes. The clustering method can mask the underlying contributing factors for the specific crash type (Valent, et al., 2002). Depaire et al. (2008) utilized latent class clustering by using the heterogeneity of traffic crash data to segment a roadway. Lu et al. (2013) used Fisher's clustering to create a

segmentation based on sections with similar crash distributions. The segmentation produced by Fisher's clustering improved the predictive model performance. Due to the lack of quality data on roadway attributes, this study performs the clustering method using the heterogeneity of crash data.

An essential aspect of traffic safety studies is unobserved heterogeneity. Studies can only include some information to capture data for all potentially contributing causes of traffic crashes (Chang, Yasmin, Huang, & Chan, 2021; Mannering, Shankar, & Bhat, 2016). A popular approach to address unobserved heterogeneity is to group the traffic crash data into homogeneous groups by different attributes (Mannering & Bhat, 2014). Some traffic crash attributes are crash units (number of vehicles involved in crashes), crash type (manner of collision), and level of crash severity. Generally, previous research classifies crashes based on crash units by grouping crashes into two major classic groups: single-vehicle (SV) and multi-vehicle (MV). Previous traffic crash studies based on total crashes have failed to identify some contributing factors and hotspots with a false positive tendency (Cheng, et al., 2017). Regardless of applying an aggregate or disaggregate approach, a crash analysis should be performed based on the crash units (number of vehicles involved in crashes) (Yu, et al., 2013b). Another typical dimension in traffic safety studies is the manner of collision (crash type), which refers to the first event in a crash; other studies refer to it as crash type. Previous studies document the importance of including the manner of collision (crash type) in traffic crash analysis (Pande, et al., 2010). The traffic crash type can be considered a dimension of group traffic crashes since it helps mask the underlying contributing factors associated with a manner of collision (Valent, et al., 2002). It is also highlighted that the traffic crashes need to be separately investigated by manner of collision since the crash mechanism may potentially vary for different manner of collision (Bhowmik, Yasmin, & Eluru, 2018). The previous

studies confirm that the contributing factors and their statistical significance are different for various manner of collisions (Mahmud & Gayah, 2021). Crash severity represents another dimension to consider in capturing the heterogeneity of traffic crashes. According to Xu et al. (Yang, et al., 2009), crash severity is determined by the most seriously injured individuals in the crash, ranging from low-cost property damage to extremely costly severe injuries or fatalities. The analysis of all crashes together may conceal the injury level of crashes (Valent, et al., 2002). For unobserved heterogeneity, this study considers crash severity alongside the crash unit (number of vehicles involved in the crash) and the crash type (manner of collision).

This study proposes a method to identify a RFS using an unsupervised clustering method on traffic crash data. The study addresses the heterogeneity of traffic crash data by grouping traffic crashes based on crash unit, crash type, and crash severity. A feature for each group of crashes is defined, and its corresponding crash rate is calculated, known as the featured crash rate (FCR). To discover the most critical features for clustering, the Laplacian score with distance-based entropy measure (LSDBEM) is used for K-means clustering feature selection identifies the features providing the most information to capture the similarities between segments. The LSDBEM-selected features significantly improve K-mean clustering results by forming homogeneous clusters (Liu, et al., 2009). Additional dimensions, such as roadway geometry, can be included and investigated in future studies to address unobserved spatial heterogeneity. While roadway geometry attributes may represent a better approach to form homogeneous segments. In the absence of quality geometry attributes, the proposed K-means clustering using crash units, crash type, and crash severity provides another strategy for crash data aggregation.

2.3. DATA DESCRIPTION

This study uses crash data from the network of urban freeways within Dallas County in Texas. The study area includes mainlane segments for both directions of Texas Loop 12, IH-20, IH-30,

IH-35E, IH-45, IH-635, and US-75 (see **Fig. 2.1.**). The data includes crash data, roadway geometric characteristics, and traffic characteristics for the 5-year period of 2015–2019. A statistics summary of crash units with the manner of collision and crash severity is provided in **Table 2.1.** and **Table 2.2.**, respectively.

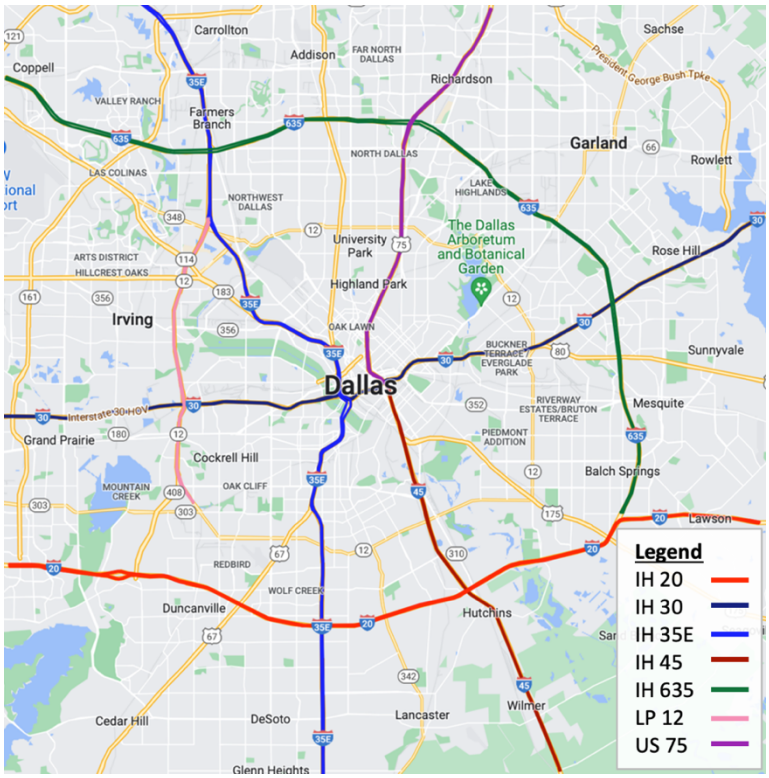


Fig. 2.1. Study area map (produced using Google Maps®).

2.3.1. Crash Data Features

The crash data from the Texas Department of Transportation (TxDOT) C.R.I.S. (Crash Record Information System) includes features from three groups: crash fields, unit fields, and person fields. The crash fields provide information about crashes. These include geospatial data such as latitude, longitude, reference marker, offset distance, highway system, roadway part, highway name, and the roadway geometry at the crash location. The crash fields also include crash characteristics like manner of collision and crash severity. This study only uses the information in

the crash fields. Also, traffic count data for the study area is obtained from TxDOT for the 5-year period of 2015-2019.

Table 2.1. Crash units and manner of collision summary (2015 – 2019)

Highway	Single-Vehicle (SV)			Multi-Vehicle (MV)				Total
	Object Related (OBJ)	Overturned (OVT)	Other (OTH)	Angled (ANGL)	Rear-End (RRND)	Sideswipe (SDSW)	Stopped (STPD)	
IH-20 EB	464	60	41	2	839	838	244	2488
IH-20 WB	475	31	43	5	854	754	215	2377
IH-30 EB	585	22	19	3	804	936	406	2775
IH-30 WB	552	27	23	2	979	842	451	2876
IH-35E NB	1011	71	52	10	2109	1853	1121	6227
IH-35E SB	825	50	34	9	1673	1750	945	5286
IH-45 NB	166	10	6	2	150	156	47	537
IH-45 SB	231	10	7	4	174	174	27	627
IH-635 NB	846	45	7	8	1819	1604	473	4802
IH-635 SB	802	62	4	5	1924	1442	562	4801
LP-12 NB	218	16	6	2	357	340	131	1070
LP-12 SB	235	16	3	4	236	313	62	869
US-75 NB	352	19	1	2	1138	779	373	2664
US-75 SB	370	13	3	1	1321	791	492	2991
Dallas County	7132	452	249	59	14377	12572	5549	40390

2.3.2. Data Preparation

The crash data provides a separate entry for every individual involved in a traffic crash sharing the same crash ID as other individuals but with a different case number. The analysis aggregates the traffic crash entries for each day by crash ID and the total number of vehicles involved in the crashes to form a new crash data set. The new crash IDs include the crash date and time to avoid loss during when fusing five years of data together. To standardize crash location, the analysis calculates the milepost values from the crash location reference marker and offset values provided in the crash data. The analysis only uses crash data for the main segment of each roadway and excludes the crashes involving active work zones, construction areas, pedestrians, or wrong-way

driving. The researchers geovalidated the crash data points by importing crash data points as KMZ files to Google Earth® to ensure the feature values for roadway segments, and vehicle travel directions are consistent with the location of crash data points.

Table 2.2. Crash severity summary

	Crash Severity (2015 - 2019)					
Highway & Travel Direction	Suspected Serious Injuries (A)	Suspected Minor Injuries (B)	Possible Injuries (C)	Fatal (K)	Not Injured (N)	Total
IH-20 EB	54	313	479	17	1625	2488
IH-20 WB	65	288	455	12	1557	2377
IH-30 EB	37	191	463	18	2066	2775
IH-30 WB	45	264	513	7	2047	2876
IH-35E NB	80	509	1116	25	4497	6227
IH-35E SB	88	409	899	23	3867	5286
IH-45 NB	23	54	121	4	335	537
IH-45 SB	14	79	144	8	382	627
IH-635 NB	135	651	1203	21	2792	4802
IH-635 SB	124	659	1108	23	2887	4801
LP-12 NB	21	117	332	3	597	1070
LP-12 SB	21	112	234	5	497	869
US-75 NB	63	322	769	10	1500	2664
US-75 SB	58	337	837	9	1750	2991
Dallas County	828	4305	8673	185	26399	40390

The Instruction to Police for Reporting Crashes (Thomas, 1996) categorizes crash severity levels as A - Suspected Serious Injury, B – Suspected Minor Injury, C – Possible Injury, K – Fatal Injury, N – Not Injured, and 99 – Unknown (see **Table 2.3.** for the definitions). The study area traffic crash data shows that crash severity at levels A, B, C, K, and N are 2.05%, 10.55%, 21.15%, 0.48%, and 64.57% of total crashes for 2015-2019 in Dallas County, respectively. Since fatal crash percentages remain very small, a separate fatal crash characteristic may not be necessary. Therefore, the analysis groups fatal and suspected serious injury crashes together since they are close in terms of severity level and represent a low portion of total crashes. Similarly, the analysis

groups suspected minor and possible injury crashes together because they do not necessarily represent distinct crash severities and likely experience a significant overlap, which would make distinctive clustering more difficult. Non-injury remains a separate crash characteristic and the authors exclude crashes with unknown severity from the study.

Table 2.3. Traffic crash categories.

Number of Vehicle Involved in Crashes	**Description**
Single-Vehicle (SV)	Crashes that only involves one motor vehicle.
Multi-Vehicle (MV)	Crashes that involve two or more motor vehicles.
Manner of Collision	
Fixed Object (OBJ)	Crashes that involve hiding fixed objects as the first harmful event.
Over-turned (OVT)	Crashes that the first harmful event is identified as vehicle overturn.
Angled (ANG)	Crashes that two motor vehicles are collided at an angle caused by at-least one vehicle deviating, turning left/right, or backing.
Rear-End (RRND)	Crashes that a motor vehicle is rear-ended by another motor vehicle.
Sideswipe (SDSW)	Crashes that a motor vehicle is sideswiped by another motor vehicle.
Stopped (STPD)	Crashes that a motor vehicle that is stopped on travel way is collided by a motor vehicle in motion.
Other (OTH)	Crashes that the manner of collision is none of the items above.
Crash Severity	
A - Suspected Serious Injury	Severe injury that prevents continuation of normal activities leading to temporarily or permanent incapacitation.
B - Suspected Minor Injury	Evident injury such as bruises, abrasions, or minor lacerations which do not incapacitate.
C - Possible Injury	Injury claimed, reported, or indicated by behavior but without visible wounds, includes limping or complaint of pain
K - Fatal	If death resulted due to injuries sustained from the crash, at the scene or within 30 days of crash.
N - Not Injured	The person involved in the crash did not sustain as A, B, C, or K injury.
99 - Unknown	Unable to determine whether injuries exist. Some examples may include hit and run, fled scene, fail to stop or render aid.

2.4. Methodology

2.4.1. Introduction

K-means clustering is an unsupervised learning method to group unlabeled objects by similarities (Pedregosa, et al., 2011). Previous traffic crash studies use this technique to cluster traffic data

based on similarities. Using clustering approach, recent research captures congestion-sensitive spots (Bhatia, et al., 2020), groups traffic flow data (Azizi & Hadi, 2021; Xu, et al., 2012; Xu, et al., 2013), classifies the crash risk for urban expressways (Cheng, et al., 2022) or other objectives. In this study, K-means clustering segmentizes urban freeway highways with features defined as crash rates calculated based on jointly considering the number of vehicles involved in the crash, manner of collision, and crash severity (crash combination).

2.4.2. Feature Selection

The K-means clustering results heavily depend on the features selected for grouping the objects into the clusters. The main goal is to compare and group highway segments by crash combination crash rates, which creates 21 features. Before applying the K-means clustering, the methodology implements feature reduction approaches to avoid redundancies and improve clustering results. This study deals with a multivariate problem in which feature values form a sparse matrix for each highway and freeway direction of travel. The methodology requires an appropriate unsupervised feature selection method to address the multivariate nature of the problem, potential redundancy, and existing sparsity in the features. The recent review by Solorio-Fernández et. al. (Solorio-Fernández, et al., 2020) categorizes feature selection candidates for this study under multivariate spectral/sparse learning methods. This study adopts the "Laplacian Score Combined with Distance-Based Entropy Measure" (LSDBEM) (Liu, et al., 2009) because it finds the best subset of features capturing underlying clustering structures before performing clustering methods. Unlike the supervised and semi-supervised feature selection approaches, the unsupervised feature selection methods have no privilege of relying on labeled data to alleviate irrelevant and redundant features. As an unsupervised feature selection, the LSDBEM employs evaluation metrics to eliminate redundant features (He, Cheng, Hu, Zhu, & Wen, 2017). Several studies utilized the

LSDBEM as unsupervised feature selection to capture the relevancy, eliminate the redundancy, and identify the most important features for unsupervised clustering, such as K-means clustering (Barile, et al., 2022; Karim, et al., 2020; Wang, et al., 2022). Karim et al. (2020) extensively implemented the LSDBEM for feature selection. They compared it with two other unsupervised feature selections, Principal Component Analysis (PCA) and Multi-Cluster-based Feature Selection (MCFS). The feature selection results show that 75% of the features selected by LSDBEM are in common with features selected by PCA and MCFS (Karim, et al., 2020). Also, Karim et al. (2020) utilized various clustering methodologies, including Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH), Hierarchical Distance-Based Spatial Clustering of Applications with Noise (DBSCAN), Ordering Points To Identify Cluster Structure (OPTICS), K-modes, Spectral, and K-means. They evaluated the clustering results using the Davies-Bouldin index, Calinski-Harabasz, and silhouette coefficient score. The K-means clustering results showed a significant purity with a very negligible difference (0.1%) compared to the outperforming clustering method OPTICS. As the method name implies, LSDBEM is a combination of the Laplacian score and an entropy measure that are separately explained in separate subsections. Prior to LSDBEM, all-zero and single non-zero features are discussed in the following subsection.

2.4.3. Dropping All-zero Features and Features with single non-zero value

A feature (crash group) that has a zero value (zero crash count) for all the objects (sub-segments) has no impact on the clustering result. Therefore, a zero-value feature can be excluded from the set of selected features for clustering. The single non-zero feature (a crash group with non-zero crash count for only one sub-segment) may be excluded because it will either not affect clustering or form a trivial single object cluster with a single object.

2.4.4. Feature Selection Using Laplacian Score (fsulaplacian):

He et al. (He, et al., 2005) introduce an unsupervised method to rank features based on a Laplacian score calculated using the nearest neighbor similarity graph as a feature selection method called “Laplacian Score”. This method has a proven record of capturing significant features. A detailed Laplacian Score algorithm may be found in a study by Pande et al. (Pande & Abdel-Aty, 2006). The algorithm favors features with large variance because “the algorithm assumes that two data points of an important feature are close if and only if the similarity graph has an edge between the two data points” (Pande & Abdel-Aty, 2006). A feature with a large score s_r represents an important feature. This can be used with the distance-based entropy measure to determine important features.

2.4.5. Distance-based Entropy Measure:

Liu et al. (Liu, et al., 2009) showed that the best subset for clustering can be identified by combining the Laplacian score method with the distance-based entropy measure. The process starts sorting features by their corresponding Laplacian score in ascending order i.e. from the most important feature to the least important feature (note that the lowest the Laplacian score, the highest the importance of the feature (Liu, et al., 2009)). Then, the top two important features are selected as the current subset of features and the distance-based entropy measure is calculated. In the next step, the next subset is formed by adding the next important feature to the current subset and the corresponding distance-based entropy measure is calculated. This process is iterated until all features are in the current subset. Among all the subsets that are investigated in the process, the subset with the highest distance-based entropy measure is the best subset of the features for clustering purposes (Liu, et al., 2009)).

Feature Selection Steps:

The feature selection procedure for this study is as follows:

1. The features are generated for all the possible combinations of traffic crash groups. **Fig. 2.2.** shows traffic crash groups, their abbreviations, and the generated features. The crash rates calculated for each of the generated features are called featured crash rates (FCRs). The naming convention of features is in the format of 'A-B-C' in which A, B, and C are the traffic crash abbreviations for the number of vehicles involved in crashes, manner of collision, and crash severity. For instance, 'SV-OBJ-N' is the feature for single-vehicle object-related crashes with no injuries. Also, 'MV-RRND-B+C' is the feature for multi-vehicle rear-end crashes with suspected minor or possible injuries.
2. All the unknown severity, all-zero, and single non-zero features are dropped.
3. The function "fsulaplacian" is applied to the current set to find all the feature scores.
4. The distance-based entropy measure is applied to the features with their corresponding Laplacian scores. The subset with the highest distance-based entropy measure is selected as the best subset of features for clustering.

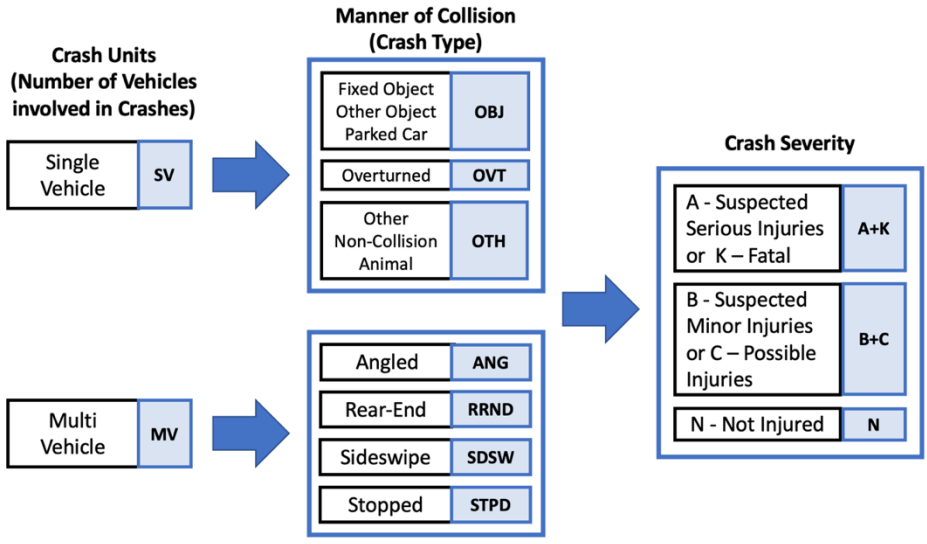


Fig. 2.2. Three dimensions of traffic crashes and the generated features.

2.4.6. K-means Clustering Algorithm

As aforementioned, K-means is an unsupervised learning technique that clusters unlabeled objects by similar features. The K-means algorithm starts with k centroids to group the objects in k clusters (a centroid for each cluster). After assigning all objects to their nearest cluster, the algorithm calculates a new set of centroids by finding the mean values of the objects in each cluster. This process iterates until the associated cost function, the Sum of Squared Error (SSE) within each cluster (also known as cluster inertia), reaches its minimum value and determines the final clusters and their corresponding centroids (Bhatia, et al., 2020). Raschka and Mirjalili (Raschka & Mirjalili, 2017) provide the formal definition of a K-means clustering algorithm as follows:

"Step 1: Randomly pick k centroids from the sample points as initial cluster centers.

Step 2: Assign each sample to the nearest centroid $\mu^j, j \in \{1, \dots, k\}$.

Step 3: Move each centroid to the center of the samples that were assigned to it.

Step 4: Repeat steps 2 and 3 until the cluster assignments do not change or a user-defined tolerance or maximum number of iterations is reached."

In "Step 2", the term "nearest" implies the distance comparison requiring a measure. The distance refers to the differences between values of features for each sample (object) and values of features for the centroids. The shorter the distance to a centroid, the closer the sample (object) to a centroid. For "Step 4", the K-means function (***KMeans***) from Python libraries ***"sklearn.cluster"*** has input variables for a user-define tolerance and maximum number of iteration as stop conditions to terminate the iterative process and report the clustering results. The parameter of the K-means function (***KMeans***) are discussed in the result section under algorithm implementation.

2.4.7. Elbow Curve and Silhouette Coefficient

As described in the previous section, the K-means clustering algorithm starts with randomly selected *k* centroids to group the objects in *k* clusters but selecting a preferred k value represents a challenge (Bhatia, et al., 2020). Running K-means clustering for a range of *k*-values and monitoring the cost function value associated with each *k*-value can overcome this obstacle (Bhatia, et al., 2020). The Elbow method can assist in finding a preferred *k* based on the marginal improvement associated with adding another cluster (Bhatia, et al., 2020). An Elbow Curve, which is a plot of the cost function against the number of clusters, visualizes this process. In an Elbow Curve, a point where the marginal gain drops such that it generates an angular point called an Elbow Point should occur. The number of clusters corresponding to the Elbow Point is the optimal number of clusters, *k** (Bhatia, et al., 2020). Mathematically, the maximum absolute value of the second derivative of the Elbow Curve is the Elbow Point (Bhatia, et al., 2020). **Fig. 2.3.** shows an Elbow Curve and its Elbow Point. Silhouette Analysis evaluates the tightness of objects within the clusters and assesses the clustering quality using the Silhouette Coefficient (Anon., 2011). In fact, the silhouette coefficient measures cluster cohesion and separation simultaneously. Cluster cohesion refers to how objects within a cluster are similar to each other. Cluster separation represents how cluster objects are different from the objects in other clusters. The greater the silhouette coefficient, the stronger the cohesion and the greater the separation. The silhouette coefficient ranges from -1 to 1, and it equals zero when the cluster cohesion and separation are the same; a value that approaches one indicates that separation greatly exceeds the within-cluster distance (Anon., 2011).

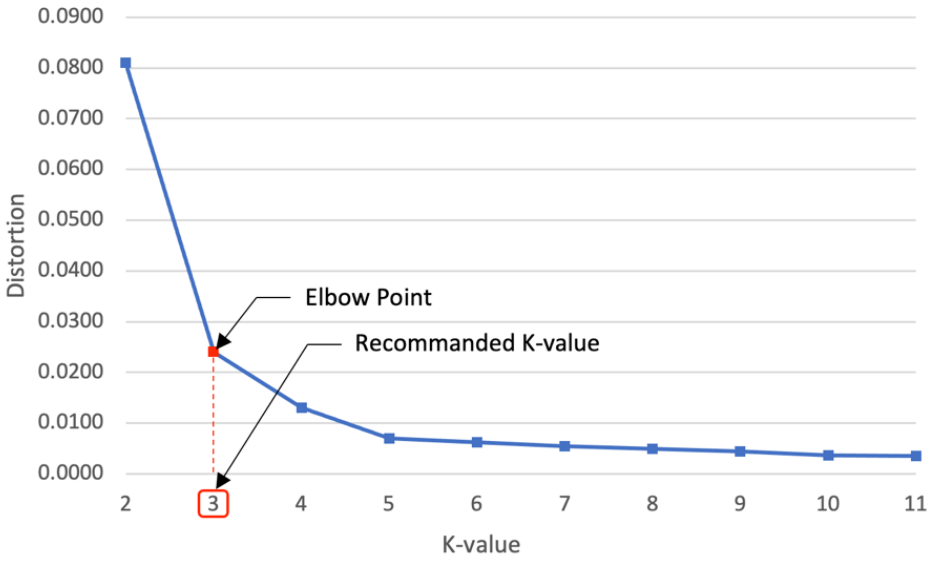


Fig. 2.3. Elbow curve and elbow point.

2.4.8. Search Algorithm

This section describes the algorithm to search for the RFS. This algorithm utilizes the K-means clustering algorithm to cluster the highway segments as the objects with the FCR based on the dimensions described in the data preparation section. The study calculates the featured crash segment length from 0.10 mile to 0.25 mile in the study. The search process starts with the initial value of 0.1 mile to perform K-means clustering to cluster highway segments and continues through the remaining segment lengths using increments of 0.01 mile. The algorithm normalizes the FCRs by dividing each feature value by its corresponding maximum FCR. As described in the feature selection section, the method investigates all the features to select the final features for K-means clustering. The K-means clustering algorithm uses the final features to cluster highway segments. Applying the K-means clustering provides a path to group highway segments by comparing feature similarities of the segments at an aggregate level. After completing the clustering for all segment lengths, the recommended clustering corresponds to the result with the greatest silhouette coefficient since it provides clusters with higher cohesion and better separation.

Also, this will provide a sufficient range of segment length scenarios to investigate the effect of segment length on data aggregation and find the significant crash combinations.

2.5. RESULTS

This section discusses the search algorithm results. The search algorithm is applied to crash data for mainlane segments in both directions of Texas Loop 12, IH-20, IH-30, IH-35E, IH-45, IH-635, and US-75 within Dallas County limits. By applying the search algorithm, the results provide the best set of features for clustering, preferred number of cluster k^*, and silhouette coefficient for each segment length ranging from 0.10 to 0.25 mile. This section also compares the FCR k-means clustering results with the findings from TCR k-means clustering results to evaluate the benefits of using FCR over TCR.

2.5.1. Algorithm Implementation

This study methodology develops a library of functions in Python 3 to perform the entire process, from data cleaning and preparation to feature selection and K-means clustering. The K-means clustering and elbow point detection use the ***"KMeans"*** and ***"Kneelocator"*** functions from Python libraries ***"sklearn.cluster"*** and ***"kneed"***, respectively. The ***KMeans*** function requires values for the attributes ***n_init=50*** and ***max_iter=1000***. ***n_init*** is the number of times that the k-means algorithm will be applied with different centroid seeds. The final k-means clustering result is the best output of ***n_init*** successive runs in terms of inertia. ***max_iter*** sets the maximum number of iterations that the k-means algorithm will be applied in a single run (Raschka & Mirjalili, 2017; Solorio-Fernández, et al., 2020). The ***KMeans*** function is applied with large enough values for the attributes ***n_init=50*** and ***max_iter=1000*** to minimize the impact of random centroids on the final result. For each run, the average computing time is 155 s and 58 s for FCR and TCR (6-Core Intel Core i7, 2.6 GHz CPU, 16 GB memory), respectively.

2.5.2. Clustering Results

The study forms traffic crash clusters by applying K-means to FCRs and TCRs data for each highway mainlane travel direction. As a sample, the clustering results for IH-20 EB (all 16 values) are shown in **Table 2.4**. Compared with TCR, the FCR-based clustering results consistently provide clusters with greater cohesion within the cluster and better separation between clusters based on their silhouette scores. For each highway travel direction, the recommended FCR-based cluster reaches silhouette scores between 0.7415 and 0.9699, which is significantly greater than the recommended TCR-based clustering results with silhouette scores between 0.6056 and 0.7255. To evaluate the significance of FCR over TCR, paired T-test is performed on $d = SC_{FCR} - SC_{TCR,}$ in which SC_{FCR} and SC_{TCR} are the silhouette scores of FCR and TCR-based clustering across all highways. By calculating d for all highways, it is obtained that $\mu_d = 0.2177$ and $S_d = 0.0054$. The hypothesis test is defined as H_0: $\mu \leq 0$ and H_a: $\mu > 0$. Considering the level of significance $\alpha = 0.01$ and $n = 14$, the value of t for the right-tailed test is $t(13, 0.01) = 2.6503$. the value of critical t, t_c is $(\mu_d - \mu)/\left(S_d/\sqrt{n}\right)$. Then $t_c = 151.16$. Thus, $t_c = 151.16 \gg 2.6503$. It yields to reject H_0 and accept H_a, i.e. $SC_{FCR} - SC_{TCR,} > 0$. Therefore, $SC_{FCR} > SC_{TCR,}$with significance level of $\alpha = 0.01$ and $C.I. = (0.2139, 0.2251)$. This shows that FCR-based clusters outperformed TCR-based clusters. For each highway travel direction, the recommended FCR-based cluster reaches silhouette scores between 0.7415 and 0.9699, which is significantly (p-value < 0.0000) greater than the recommended TCR-based clustering results with silhouette scores between 0.6056 and 0.7255.

Table 2.4. Clustering results comparison (IH-20 EB).

	Featured Crash Rate (FCR)			Total Crash Rate (TCR)	
Len. Of Seg.	**Recom'd K-value**	**Silhouette Coefficient**	**Set of Features**	**Recom'd K-value**	**Silhouette Coefficient**
0.10	3	0.9699	[‘SV-OBJ-A+K’, 'SV-OVT-A+K']	4	0.6276
0.11	3	0.9647	[‘SV-OBJ-A+K’, 'SV-OVT-A+K']	4	0.6294
0.12	5	0.3910	[‘SV-OBJ-A+K’, 'SV-OBJ-B+C', 'SV-OTH-N', 'SV-OVT-A+K', 'SV-OVT-N', 'MV-RRND-B+C', 'MV-RRND-N', 'MV-SDSW-B+C', 'MV-STPD-B+C', 'MV-STPD-N']	4	0.5958
0.13	3	0.9573	[‘SV-OBJ-A+K’, 'SV-OVT-A+K']	4	0.6115
0.14	3	0.9540	[‘SV-OBJ-A+K’, 'SV-OVT-A+K']	4	0.6228
0.15	3	0.9041	[‘SV-OBJ-A+K’, 'MV-RRND-A+K']	4	0.6115
0.16	3	0.9451	[‘SV-OBJ-A+K’, 'SV-OVT-A+K']	3	0.6448
0.17	4	0.6872	['SV-OVT-A+K', 'MV-SDSW-B+C']	3	0.6385
0.18	4	0.8219	[‘SV-OBJ-A+K’, 'SV-OTH-N', 'MV-STPD-A+K']	5	0.5840
0.19	3	0.9346	[‘SV-OBJ-A+K’, 'SV-OVT-A+K']	4	0.5881
0.20	4	0.6661	['SV-OBJ-B+C', 'SV-OVT-A+K']	3	0.6103
0.21	3	0.8605	[‘SV-OBJ-A+K’, 'MV-RRND-A+K']	3	0.6396
0.22	2	0.9123	['SV-OVT-A+K', 'MV-STPD-A+K']	5	0.5614
0.23	3	0.9190	[‘SV-OBJ-A+K’, 'SV-OVT-A+K']	3	0.6530
0.24	5	0.4367	['SV-OTH-N', 'SV-OVT-N', 'MV-SDSW-B+C', 'MV-STPD-B+C']	4	0.6632
0.25	3	0.6906	['SV-OTH-N', 'SV-OVT-N', 'MV-STPD-A+K']	4	0.6564

2.5.3. Feature Selection

The FCR-based clustering results provide the sets of significant features associated with the clustering. Also, **Fig. 2.4.** and **Fig. 2.5.** show heatmap representations of feature significance for the urban highway travel directions for the sixteen segment length values ranging from 0.10 to 0.25 mile. Due to the sixteen values, the frequency of features appearing significant varies between 0 and 16. The results demonstrate that the significant features differ depending on the urban highway and travel direction; however, some features appear frequently in most trials generated

by different segment lengths. For IH-20 EB, the methodology selects ‘SV-OBJ-A+K’ and 'SV-OVT-A+K' as the significant features for more trials (segment length), including the RFS, than other features. For IH-20 EB, severe single-vehicle crashes with a clear crash class create the best crash data clusters (see **Fig. 2.2.** for abbreviations). The feature significance appears relatively insensitive to the segment length selected to aggregate the crash data. In most cases, the most frequently significant features (during the sixteen trials) for each highway appear in the cluster with the highest silhouette score. However, a few less frequently selected features also appear in the clusters with the highest silhouette scores, such as features ‘SV-OBJ-A+K’ and ‘SV-OVT-B+C’ for IH-30 EB and feature 'SV-OVT-N' for IH-35E NB. Other less frequently significant features include 'SV-OBJ-N’, 'SV-OVT-N’, and 'SV-OTH-N’, which makes sense because these crashes may be uniformly distributed along a highway since no injuries occur and they only involve a single vehicle. For most freeways, one to three features frequently appear for clustering with the first and second-ranked highest silhouette scores; however, US-75 SB has ten frequently appearing features. **Fig. 2.4.(o)** shows the Dallas County heatmap that summarizes the total frequency of the significant features for the studied highways. The potential range of values in this figure is [0, 224]. Based on **Fig. 2.4.(o)**, the most frequently significant features are ‘SV-OVT-B+C’, 'MV-SDSW-N’, 'MV-STPD-N’, ‘SV-OBJ-A+K’, 'MV-SDSW-B+C’, and 'MV-RRND-A+K’, in descending order.

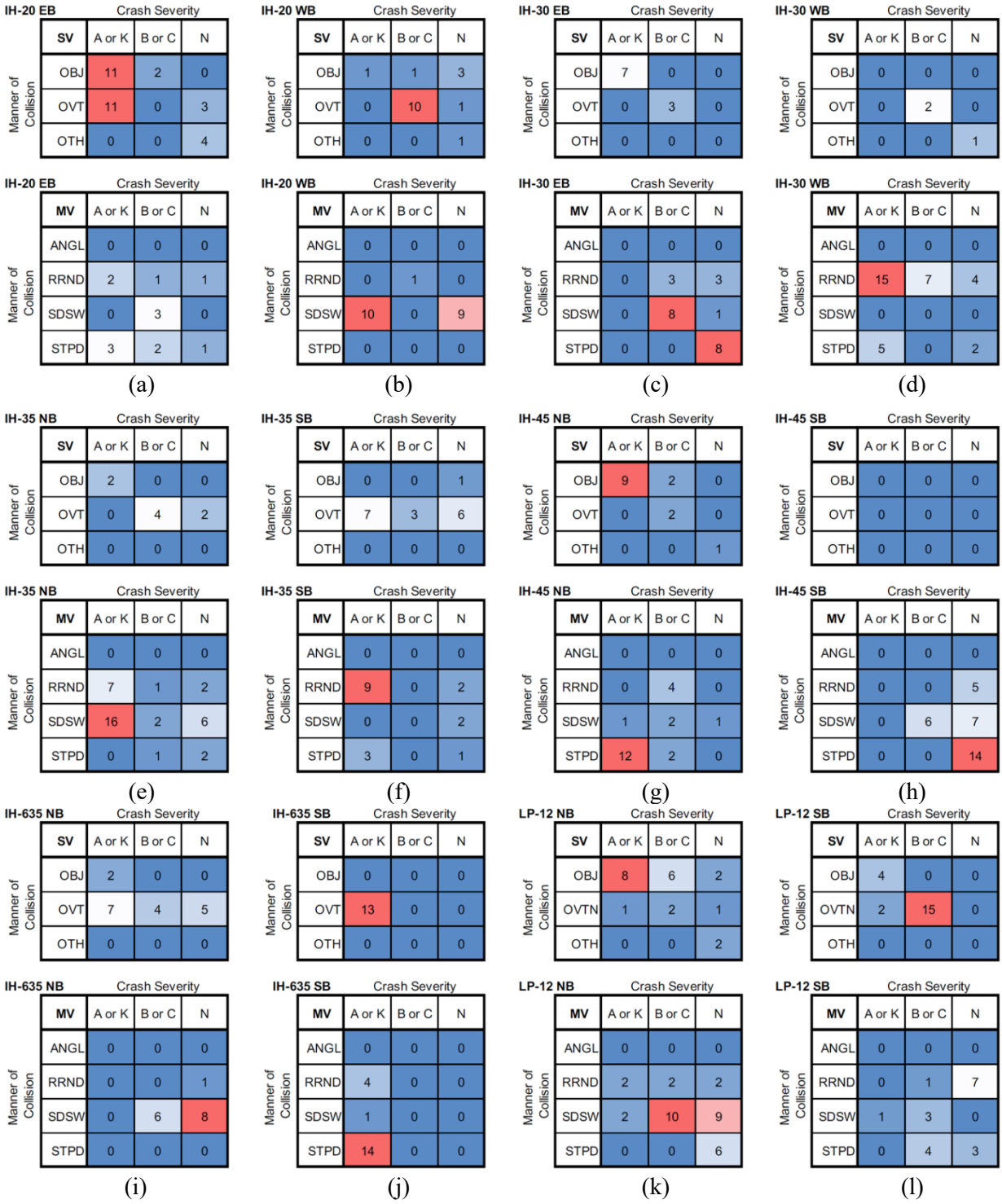
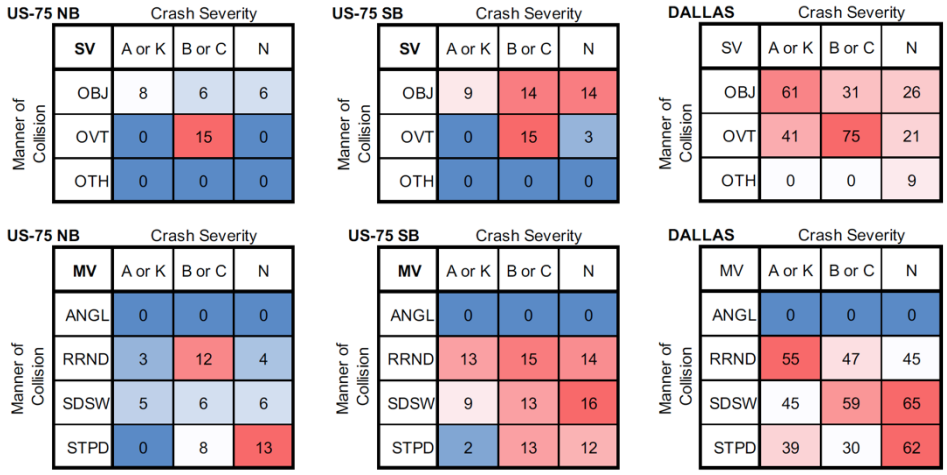


Fig. 2.4. Heatmap of significant features for the highways IH-20, IH-30, IH-35, IH-45, IH-635, and LP-12.



(a) (b) (c)

Fig. 2.5. Heatmap of significant features for the highway US-75 and Dallas County.

2.5.4. Segment Length

The results show that the segment length impacts the clustering results and their corresponding silhouette scores. **Table 2.5** provides a comparison between the top two RFS values for FCR and TCR. The FCR clustering tends to recommend much shorter fragment sizes than the TCR because they also capture trends in specific crash combinations more effectively than the TCR. For almost all the highways, the FCR clustering methodology selects two features, which generate clusters with silhouette scores at least 0.1 larger than the best corresponding TCR result. The additional information provided by the FCR strengthens the clustering and segregates the freeway into segments with different crash risks for the selected features.

2.5.5. Z-score Analysis of FCR-based Clusters

The features' Z-scores for the clusters with highest silhouette score is provided in **Table 2.6**. For each highway travel direction, the features F_1 , F_2, and F_3 correspond to the set of features in **Table 2.5.** for clustering with the highest silhouette scores. In most two-cluster and two- feature cases, the clustering results for $k^* = 2$ (two clusters) show that one feature appears with a large positive Z-score in one cluster while the other feature shows a small value (somewhat close to zero) and the feature values reverse in the other cluster. For instance, the clustering result for IH-20 WB

shows that single-vehicle overturned crashes with minor or possible injuries has a Z-score of 4.32 for cluster #2, meaning, cluster #2 represents single vehicle overturned crashes with minor or possible injuries but not multi-vehicle sideswipe fatal and serious crashes; cluster #1 represents risky locations for multi-vehicle sideswipe fatal and serious crashes but not single-vehicle object crashes with minor or possible injuries. The large Z-score also indicates the intensity of the risk for cluster #2 is much higher than cluster #1. The same pattern for cluster #1 and #2 applies to other highway travel directions with $k^* = 2$ (two clusters) IH-35 NB, IH-35 SB, IH-45 NB, IH-635 SB, and LP-12 SB for their corresponding features. For IH-35 NB, cluster #2 identifies high-risk multi-vehicle sideswipe fatal and serious crash locations. For IH-35 SB, cluster #2 identifies high-risk multi-vehicle rear-end fatal and serious injury crash locations. For IH-45 NB and IH 635 SB, cluster #2 identifies high-risk multi-vehicle stopped fatal and serious injury crash locations. For LP-12 SB, cluster #2 identifies high-risk single-vehicle overturned minor and possible injury crash locations. Another two-cluster case, IH-30 WB, follows a different pattern where cluster #1 represents a low crash risk for both features and cluster #2 represents a high crash risk for fatal and serious multi-vehicle rear-end and stopped crashes. For $k^* = 3$ (three clusters), one cluster indicates a high-risk location for one crash type and another cluster indicates a high-risk location for the other selected crash type; the third cluster indicates low-risk crash locations for both selected crash features. IH-20 EB identifies high-risk locations for single-vehicle object and overturn crashes with fatal and serious injury, IH-30 EB identifies high-risk locations for single-vehicle object fatal and serious injury crashes and single-vehicle overturn crashes with minor and possible injury, and US-75 NB identifies high-risk locations for single-vehicle overturned minor and possible injury crashes and multi-vehicle rear-end minor and possible injury crashes. Another three-cluster case, IH-635 NB, adds a third feature to the clustering results; this case creates a low-

risk crash cluster for single-vehicle overturn crashes. The other clusters separate high-risk single-vehicle overturned fatal and serious crash locations from high-risk single-vehicle overturned minor and possible injury crash locations. Only two freeway corridors (IH-45 SB and LP-12 NB) showed $k^* = 4$ (four clusters). For the IH-45 SB case, one cluster identifies low-risk locations for multi-vehicle sideswipe crashes with minor and possible injuries and multi-vehicle stopped crashes with property damage only. Another cluster identifies locations with high-risk for multi-vehicle sideswipe minor and possible injury crashes and low-risk for multi-vehicle stopped crashes with property damage only. The final two clusters contain moderate risk for multi-vehicle sideswipe minor and possible injury crash locations and high and moderate risk for multi-vehicle stopped crashes with property damage only. The LP-12 NB case identifies clusters with low risk for both features (single-vehicle object fatal and serious injuries and multi-vehicle stopped property damage), high risk for both features, and high risk for one feature/low risk for the other feature. Finally, US-75 SB demonstrated $k^* = 5$ (five clusters), as with all clusters with $k^* > 2$, one cluster represents low crash risk locations for the selected features. Similar to other cluster amounts, one cluster characterizes locations with high risk for single-vehicle overturned minor and probable injury crashes and low risk for multi-vehicle sideswipe property damage only crashes. Two other clusters identify locations with high and moderate risk for multi-vehicle sideswipe property damage only crashes and low risk for single-vehicle overturned minor and probable injury crashes. The final cluster includes locations with moderate risk for single-vehicle overturned minor and possible injury crashes and slightly above average risk for multi-vehicle sideswipe property damage only crashes. Overall, the clustering represents an effective strategy for identifying data patterns for the selected crash features, which can directly identify high and low risk locations for these crash combinations.

2.5.6. Silhouette Scores and Fragment Sizes

A stairs-type stacked plot of silhouette scores for FCR and TCR clusters versus various fragment sizes for all highway travel directions is shown in **Fig. 2.6**. The silhouette scores for the FCR and TCR clustering results for the selected features are illustrated in blue and orange color, respectively. Overall, the silhouette scores of the TCR-based clustering results show greater stability across the various fragment sizes than the silhouette scores of the FCR-based clustering results. While the TCR-based clustering is more resistant to changes in the fragment sizes used for data aggregation, its silhouette scores remain under 0.80 while FCR-based clustering shows silhouette scores greater than 0.80 for some fragment sizes. However, the TCR-based clustering result supersedes the FCR-based clustering for US-75 SB for all fragment sizes but 0.23 mile where FCR-based clustering result reaches the highest silhouette score. For IH-635 SB, the FCR-based clustering show highest silhouette scores for all fragment sizes comparing to TCR-based. These trends can be related to the traffic crash data distribution along US-75 SB and IH-635 SB.

2.6. CONCLUSIONS AND RECOMMENDATIONS

This paper develops a recommended fragment size (segment length) using three dimensions of traffic crashes (i.e., number of vehicles involved in the crash, manner of collision, and crash severity) and clustering methods as an innovative data-driven method to aggregate crash data. This strategy provides a standard approach for future studies to aggregate crashes and resolves the previously identified concern associated with the arbitrary selection of segment length in previous research. The proposed method harnesses the advantages of LSDBEM and K-means clustering algorithm as unsupervised learning applied to highway segments as the objects.

The study defines featured crash rates (FCRs) using three dimensions of traffic crash characteristics: number of vehicles involved in the crash, manner of collision, and crash severity. The FCR-based clustering results show that RFS varies for each highway travel direction.

Table 2.5. RFS values (FCR vs TCR).

	Featured Crash Rate (FCR)					Total Crash Rate (TCR)			
Highway & Travel Direction	**RFS Rank**	**Len. Of Seg.**	**K-value**	**Silh. score**	**Set of Features**	**RFS Rank**	**Len. Of Seg.**	**K-value**	**Silh. score**
IH-20 EB	1st	0.10	3	0.9699	[‘SV-OBJ-A+K’, 'SV-OVT-A+K']	1st	0.24	4	0.6632
IH-20 EB	2nd	0.11	3	0.9647	[‘SV-OBJ-A+K’, 'SV-OVT-A+K']	2nd	0.25	4	0.6564
IH-20 WB	1st	0.10	2	0.9223	[‘SV-OVT-B+C’, 'MV-SDSW-A+K’]	1st	0.20	3	0.6575
IH-20 WB	2nd	0.11	2	0.9153	[‘SV-OVT-B+C’, 'MV-SDSW-A+K’]	2nd	0.13	3	0.6413
IH-30 EB	1st	0.14	3	0.8880	[‘SV-OBJ-A+K’, ‘SV-OVT-B+C’]	1st	0.24	3	0.6704
IH-30 EB	2nd	0.15	3	0.8716	[‘SV-OBJ-A+K’, ‘SV-OVT-B+C’]	2nd	0.25	4	0.6470
IH-30 WB	1st	0.10	2	0.9128	['MV-RRND-A+K', 'MV-STPD-A+K']	1st	0.25	3	0.6852
IH-30 WB	2nd	0.12	2	0.8994	['MV-RRND-A+K', 'MV-STPD-A+K']	2nd	0.16	4	0.6680
IH-35E NB	1st	0.10	2	0.8726	['SV-OVT-N', 'MV-SDSW-A+K’]	1st	0.24	3	0.6478
IH-35E NB	2nd	0.11	2	0.8601	['MV-RRND-A+K', 'MV-SDSW-A+K’]	2nd	0.12	3	0.6435
IH-35E SB	1st	0.11	2	0.9366	['SV-OVT-A+K', 'MV-RRND-A+K']	1st	0.13	2	0.7255
IH-35E SB	2nd	0.12	2	0.9363	['SV-OVT-A+K', 'SV-OVT-N']	2nd	0.14	3	0.6754
IH-45 NB	1st	0.13	2	0.9240	[‘SV-OBJ-A+K’, 'MV-STPD-A+K']	1st	0.16	3	0.6532
IH-45 NB	2nd	0.12	2	0.9216	[‘SV-OBJ-A+K’, 'MV-STPD-A+K']	2nd	0.14	3	0.6473
IH-45 SB	1st	0.16	4	0.8114	['MV-SDSW-B+C', 'MV-STPD-N']	1st	0.10	2	0.6817
IH-45 SB	2nd	0.21	3	0.7530	['MV-SDSW-B+C', 'MV-STPD-N']	2nd	0.18	2	0.6800
IH-635 NB	1st	0.12	3	0.9042	['SV-OVT-A+K', ‘SV-OVT-B+C’, 'SV-OVT-N']	1st	0.17	2	0.6886
IH-635 NB	2nd	0.18	2	0.8915	['SV-OVT-A+K', 'SV-OVT-N']	2nd	0.12	4	0.6579
IH-635 SB	1st	0.11	2	0.9358	['SV-OVT-A+K', 'MV-STPD-A+K']	1st	0.11	3	0.6406
IH-635 SB	2nd	0.12	2	0.9341	['SV-OVT-A+K', 'MV-STPD-A+K']	2nd	0.17	3	0.6215
LP-12 NB	1st	0.11	4	0.8260	[‘SV-OBJ-A+K’, 'MV-STPD-N']	1st	0.19	3	0.6514
LP-12 NB	2nd	0.15	4	0.7981	[‘SV-OBJ-A+K’, 'MV-STPD-N']	2nd	0.14	3	0.6374
LP-12 SB	1st	0.11	2	0.9167	['SV-OVT-A+K', ‘SV-OVT-B+C’]	1st	0.20	4	0.6056
LP-12 SB	2nd	0.18	2	0.8664	['SV-OVT-A+K', ‘SV-OVT-B+C’]	2nd	0.17	4	0.6035
US-75 NB	1st	0.22	3	0.7627	[‘SV-OVT-B+C’, 'MV-RRND-B+C']	1st	0.23	4	0.6147
US-75 NB	2nd	0.25	3	0.7512	[‘SV-OVT-B+C’, 'MV-RRND-B+C']	2nd	0.21	4	0.5945
US-75 SB	1st	0.22	5	0.7415	[‘SV-OVT-B+C’, 'MV-SDSW-N']	1st	0.11	4	0.6905
US-75 SB	2nd	0.10	3	0.6037	[‘SV-OBJ-A+K’, 'SV-OBJ-B+C', 'SV-OBJ-N', ‘SV-OVT-B+C’, 'MV-RRND-A+K', 'MV-RRND-B+C', 'MV-RRND-N', 'MV-SDSW-A+K’, 'MV-SDSW-B+C', 'MV-SDSW-N', 'MV-STPD-B+C', 'MV-STPD-N']	2nd	0.10	4	0.6107

Table 2.6. Z-score values of selected features used in LSDBEM/K-means clustering for FCR.

Highway Travel Direction	Cluster ID	Feature Mean Per Cluster			Feature Total Mean			Feature Variance			Feature Z-Score		
		F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
IH-20 EB	1	0.00	0.00	-	0.07	0.02	-	0.04	0.02	-	-0.35	-0.17	-
	2	0.63	0.00	-	0.07	0.02	-	0.04	0.02	-	2.69	-0.17	-
	3	0.00	0.78	-	0.07	0.02	-	0.04	0.02	-	-0.35	5.74	-
IH-20 WB	1	0.00	0.03	-	0.03	0.03	-	0.02	0.02	-	-0.23	0.01	-
	2	0.68	0.00	-	0.03	0.03	-	0.02	0.02	-	4.32	-0.21	-
IH-30 EB	1	0.00	0.00	-	0.11	0.07	-	0.06	0.04	-	-0.45	-0.32	-
	2	0.60	0.00	-	0.11	0.07	-	0.06	0.04	-	2.05	-0.32	-
	3	0.10	0.66	-	0.11	0.07	-	0.06	0.04	-	-0.01	2.89	-
IH-30 WB	1	0.00	0.04	-	0.04	0.04	-	0.03	0.03	-	-0.22	-0.02	-
	2	0.84	0.09	-	0.04	0.04	-	0.03	0.03	-	4.31	0.31	-
IH-35E NB	1	0.04	0.00	-	0.04	0.04	-	0.02	0.03	-	0.00	-0.27	-
	2	0.03	0.60	-	0.04	0.04	-	0.02	0.03	-	-0.04	3.35	-
IH-35E SB	1	0.02	0.00	-	0.02	0.05	-	0.02	0.04	-	0.01	-0.26	-
	2	0.00	0.77	-	0.02	0.05	-	0.02	0.04	-	-0.16	3.71	-
IH-45 NB	1	0.04	0.00	-	0.04	0.05	-	0.03	0.04	-	0.01	-0.24	-
	2	0.00	0.91	-	0.04	0.05	-	0.03	0.04	-	-0.23	4.14	-
IH-45 SB	1	0.01	0.00	-	0.14	0.12	-	0.05	0.08	-	-0.61	-0.42	-
	2	0.21	0.87	-	0.14	0.12	-	0.05	0.08	-	0.32	2.71	-
	3	0.47	0.00	-	0.14	0.12	-	0.05	0.08	-	1.51	-0.42	-
	4	0.20	0.44	-	0.14	0.12	-	0.05	0.08	-	0.29	1.16	-
IH-635 NB	1	0.00	0.00	0.03	0.03	0.08	0.03	0.03	0.06	0.03	-0.20	-0.32	-0.04
	2	0.00	0.81	0.08	0.03	0.08	0.03	0.03	0.06	0.03	-0.20	3.08	0.26
	3	0.83	0.16	0.07	0.03	0.08	0.03	0.03	0.06	0.03	4.89	0.35	0.22
IH-635 SB	1	0.02	0.00	-	0.02	0.04	-	0.02	0.02	-	0.01	-0.26	-
	2	0.00	0.60	-	0.02	0.04	-	0.02	0.02	-	-0.15	3.69	-
LP-12 NB	1	0.00	0.03	-	0.07	0.15	-	0.04	0.08	-	-0.32	-0.42	-
	2	0.72	0.65	-	0.07	0.15	-	0.04	0.08	-	3.22	1.80	-
	3	0.00	0.73	-	0.07	0.15	-	0.04	0.08	-	-0.32	2.11	-
	4	0.63	0.10	-	0.07	0.15	-	0.04	0.08	-	2.79	-0.18	-
LP-12 SB	1	0.04	0.00	-	0.04	0.07	-	0.04	0.05	-	0.02	-0.28	-
	2	0.00	0.87	-	0.04	0.07	-	0.04	0.05	-	-0.21	3.45	-
US-75 NB	1	0.00	0.94	-	0.14	0.19	-	0.07	0.03	-	-0.52	4.05	-
	2	0.00	0.15	-	0.14	0.19	-	0.07	0.03	-	-0.52	-0.23	-
	3	0.62	0.22	-	0.14	0.19	-	0.07	0.03	-	1.79	0.13	-
US-75 SB	1	0.00	0.07	-	0.08	0.10	-	0.05	0.02	-	-0.37	-0.25	-
	2	0.94	0.08	-	0.08	0.10	-	0.05	0.02	-	3.97	-0.18	-
	3	0.46	0.13	-	0.08	0.10	-	0.05	0.02	-	1.75	0.14	-
	4	0.00	1.00	-	0.08	0.10	-	0.05	0.02	-	-0.37	6.05	-
	5	0.00	0.35	-	0.08	0.10	-	0.05	0.02	-	-0.37	1.63	-

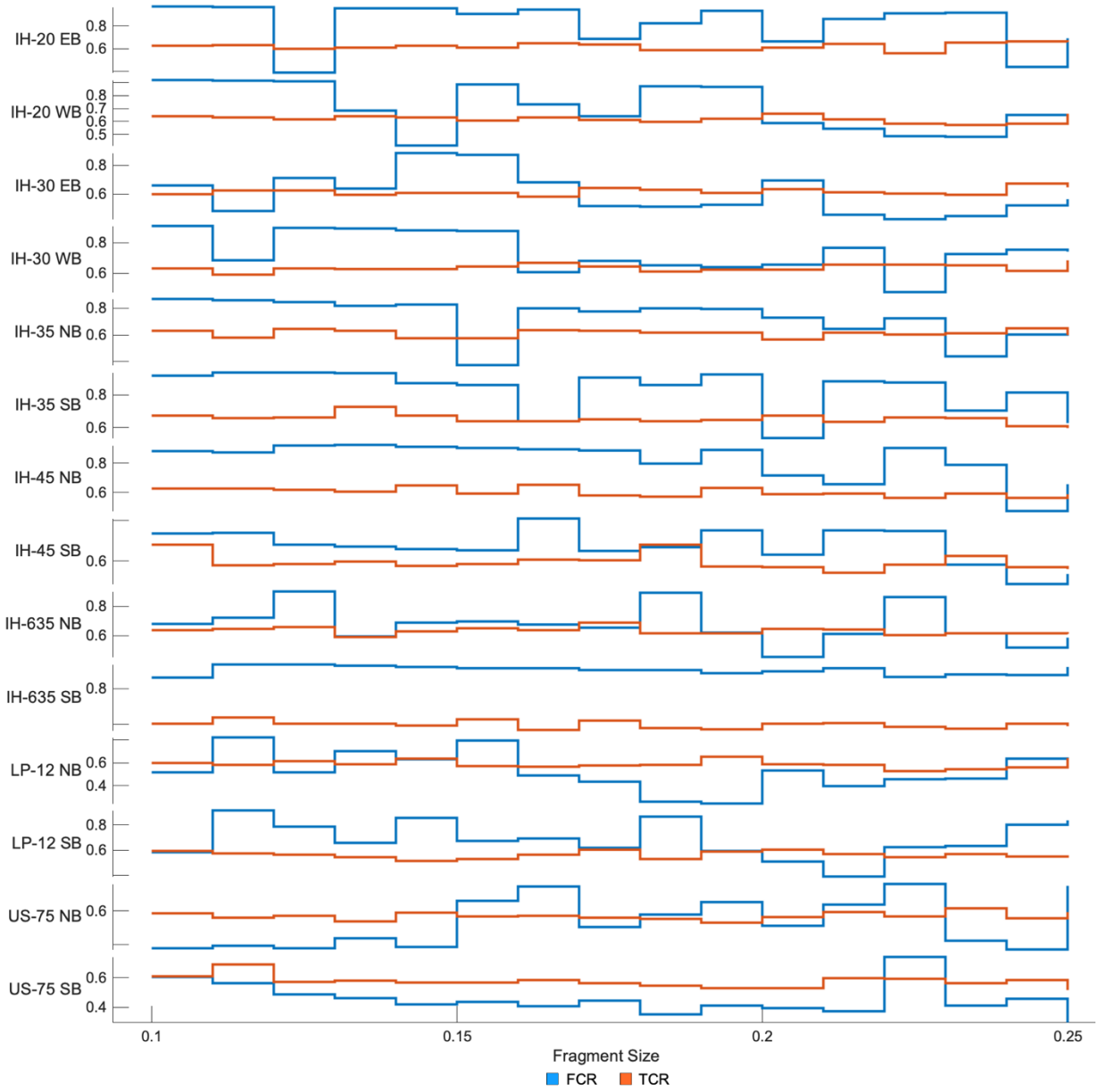


Fig. 2.6. Silhouette scores for FCR and TCR clusters vs fragment size.

The typical segment length of 0.10 mile that has been used in several studies matches the RFS only for IH-20 EB, IH-20 WB, IH-30 WB, IH-35E NB, and US-75 SB that, which is less than forty percent of highway travel directions. The RFS based on FCR clustering varies between 0.1 and 0.22, while the RFS based on TCR clustering cover the entire range from 0.1 to 0.25.

The variation in RFS across the different highways and travel directions indicates that a single "best" segment length does not exist, and the segment length should be selected based on observed

crash data. However, the RFS based on FCR clustering and TCR clustering is the same for US 75 SB (0.10 mile) and IH 635 SB (0.11 mile) (see **Table 2.5**). The FCR-based clustering results not only provide a RFS using three dimensions of traffic crashes characteristics but also identify the significant features for each highway travel direction which is impractical using TCR-based clustering. This paper proposed a data-driven methodology that overcomes the arbitrary selection of segment length using three dimensions of traffic crash characteristics.

The significant improvement in silhouette score between the FCR and TCR clustering methods indicates more cohesive and distinct clusters. This improvement will make the aggregated crash data more valuable and guarantee that the within-cluster segments experience similar crash risk for the selected features. The highest FCR-based silhouette scores range between 0.7415 and 0.9699. The methodology typically chooses two features for the best silhouette scores. However, the methodology evaluated several sets of features before selecting the set of features to represent the data clusters best. While the selected features vary significantly between freeways and travel directions, the features used to select the clusters associated with the RFS typically reflect the most commonly selected features for a particular freeway and travel direction. This study provides a foundation for highway segmentation that benefits future traffic and crash studies and RTCPMs using aggregated data.

Because this study establishes a standardized method for selecting a segment length to aggregate crash data for future safety analyses and RTCPMs, many opportunities for future research exist. The total assessment of this method's impact requires investigating the improvement in crash modeling that results. In addition, this method may eliminate the need for disaggregating locationally specific static crash modification factors for RTCPMs if the clustering can effectively capture aggregate static crash contributing factors. Future research should also examine the RFS's

temporal stability and cluster structure's temporal stability. An extension of this study is to consider the temporal instability and unobserved heterogeneity associated with the environmental characteristics and driver behaviors by introducing featured crash rates (FCRs) for each year, including the environmental characteristics, and applying the LSDBEM/K-means. The study only investigates the clustering and recommended fragment length using all three traffic crash characteristics combined. The LSDBEM/K-means clustering can be applied to crash groups for scenarios including crash units only, and crash units and manner of collision combined to compare with FCR and TCR clustering results. The future study should investigate the value or importance of including additional crash characteristics in predicting crash risk and identifying contributing factors. Future studies need to extend this study by investigating each traffic crash characteristic separately and comparing the results with all three traffic crash characteristics considered. This study considered each highway and travel direction separately and created distinctive clusters for each. The future research can also consider the network wide clustering for a comparison. Future studies should apply this method on other freeway networks and explore applying it (or a variation) for two-lane highways and arterials. While this study includes three crash dimensions in its features, future studies may consider fewer (e.g., number of vehicles and manner of collision) and more crash dimensions (e.g., roadway geometry or AADT). The clustering may also involve other non-crash features and incorporate spatial correlation. A future study may expand the proposed RFS method to segmentize highways with a variable segment length rather than a constant length of the segment. The fragment size (segment length) selected for data aggregation may impact the statistical significance of explanatory variables in crash prediction models; a future study investigates these impacts and investigates the potential advantages of the recommended fragment size (RFS) for crash prediction models .

CHAPTER 3. Traffic Crash Hotspot Identification and Static Contributing Factors by Crash Unit, Manner of Collision, and Crash Severity

3.1. INTRODUCTION

According to the World Health Organization (WHO), traffic crashes are a primary cause of death and injury worldwide, with an estimated 1.35 million deaths annually (World Health Organization (WHO), 2018). As the highway/freeway systems expand rapidly, the risk of crashes increases, making freeway safety management a top priority. In recent years, identifying traffic crash hotspots has become essential for pinpointing hazardous locations, prioritizing effective countermeasures, and improving road safety. Traffic crash hotspots represent locations where the frequency of crash occurrences is higher or the likelihood of crashes is more significant than the neighboring locations along the targeted corridor or across a network. Hotspot identification (HSID) allows for targeted interventions to reduce the likelihood of crashes and improve safety in hazardous areas by understanding the contributing factors in traffic crashes. Most HSID studies focus on the total number of crashes. However, a few recent studies investigated the crash hotspots by considering one or two traffic crash characteristics, such as the number of vehicles involved in crashes (crash units), manner of collision, and crash severity. HSID studies show that crash hotspots based on any of the three traffic crash characteristics differ from crash hotspots based on the total crashes. Leaving out any of the three crash characteristics may disrupt understanding traffic crashes, contributing factors, and identifying the hotspots. This study develops a methodology to include all three crash dimensions simultaneously in HSID and analyzes the effect of adding additional crash dimensions on HSID and contributing factors (Wang & Feng, 2019).

This study investigates the crash dimensions' role in properly characterizing hotspots and contributing factors. The number of vehicles involved in crashes (crash units), which categorizes traffic crashes into single-vehicle (SV) and multi-vehicle (MV), may be used to identify different

crash hotspots for each category. Previous research indicates a crucial need to characterize SV and MV crash hotspots separately because they have different spatial distributions and contributing factors (Wang & Feng, 2019). Other research supports that the factors that lead to crashes may be different or have different impacts on SV and MV crashes (Ivan, et al., 1999; Abdel-Aty, et al., 2006). Because the crash unit plays a significant role in both aggregate and disaggregate traffic safety analysis approaches (Yu & Abdel-Aty, 2013), this study considers crash units in crash HSID. The manner of collision, or the first event in a crash, represents another essential traffic crash characteristic since the contributing factors vary for the various manners of collision; some studies refer to the manner of collision as a crash type. An earlier study indicates that including the manner of collision reveals important information that total crashes may fail to recognize (Golob, et al., 2008). By incorporating the manner of collision (crash type) rather than solely focusing on total crashes, many traffic safety studies have uncovered unknown details about traffic crashes (Golob, et al., 2004a; Cheng, et al., 2017). Cheng et al. (2017) also support the importance of including the manner of collision in safety analyses. Thus, this study includes the manner of collision as another traffic crash feature dimension. In traffic crash studies, crash severity represents another crucial characteristic considered alongside crash units and crash type. Previous research has also shown that contributing factors to traffic crashes may have varying impacts on different levels of crash severity (Abdel-Aty, 2003; Jung, et al., 2010). While crash severity has been incorporated into studies that identify crash hotspots at intersections, it is often overlooked in HSID studies for uninterrupted facilities. All three dimensions exhibit different relationships with contributing factors and spatial distributions.

Many approaches to hotspot identification exist, and they can be categorized into Geographic Information Systems (GIS) based spatial analysis, statistical models, and machine learning. GIS

has become a valuable tool for studying traffic crashes spatially and identifying areas with a high frequency of crashes by mapping and analyzing crash data (Al-Aamri et al., 2021). The conventional approach to identifying high-crash areas involves creating crash concentration maps, which rely on Kernel Density Estimation (KDE) and absolute counts of crashes to determine the density of crashes in a given area (Truong & Somenahalli, 2011). However, some research identifies two potential issues with this method. Firstly, the accuracy of concentration maps may vary depending on the search bandwidth used. Secondly, the absolute crash counts may not accurately indicate safety issues, as they ignore important factors such as crash types and exposure measures like vehicular volumes (Truong & Somenahalli, 2011). Another approach uses regression models developed in previous studies or using historical data to predict the number of crashes, examine factors associated with crashes, and identify traffic crash hotspots. These previous studies use several types of count data regression models: Poisson, Negative Binomial, Poisson Lognormal, Zero-Inflated Poisson/Negative Binomial, Gamma, Generalized Estimating Equation, Negative Multinomial, and Hurdle (Hilbe, 2014). The regression models aim to predict crash frequencies based on roadway geometry features, traffic characteristics, and weather conditions. [Explain the step to identify crash hotspots] (Highway Safety, 2005). Some recent studies apply machine learning methods to conduct traffic safety studies and identify traffic crash hotspots. Machine learning methods used to investigate traffic crash include random forest, decision tree, support vector machine (SVM), Naive Bayes, and neural network (Santos et al., 2022). The statistical models still appeal significantly because they provide engineers insight into features that may contribute to higher crash frequencies.

This study investigates the implications of HSID (crash hotspot identification) based on simultaneously considering three traffic crash characteristics: crash units, manner of collision, and

crash severity. This approach can identify distinct contributing factors for each group of crashes. To carefully examine the benefits of the proposed strategy and compare it with previous findings, the study conducts HSID for three other scenarios: (A) total crashes (with no incorporation of crash characteristics), (B) single-vehicle (SV) and multi-vehicle (MV) crashes (with only crash units considered), and (C) all crash groups categorized by crash units and crash types. The authors examine the spatial differences and changes in contributing factors for the traffic crash hotspots generated using all three crash characteristics and scenarios (A), (B), and (C) to determine the potential advantages of using HSID based on these three traffic crash dimensions.

3.2. LITERATURE REVIEW

3.2.1. Traffic crash hotspots:

Traffic safety represents a significant concern due to the socioeconomic burden associated with road crashes each year across the globe. Many countries worldwide have launched measures to promote traffic and road safety to reduce fatalities as their population grows; however, these countries have failed to meet the WHO's worldwide objective of reducing road traffic deaths by half by 2020 (World Health Organization (WHO), 2018). Traffic safety research investigates the likelihood of traffic crash occurrence, identifies crash hotspot locations (locations with high crash frequency), and characterizes traffic crash contributing factors to improve traffic safety and mitigate the socioeconomic impacts of traffic crashes. In theory, a crash hotspot refers to a place where a greater number of crashes have taken place in comparison to other comparable locations along the targeted corridor or across a network (Sørensen & Elvik, 2007). The process of identifying hotspots and potential safety issues aims to highlight the roadway sections with a high risk of crashes. The HSID process aims to suggest possible countermeasures to reduce the risk of crashes by analyzing various crash patterns and identifying contributing factors (Montella, 2010).

Therefore, HSID is a focal point of traffic safety analyses by identifying contributing factors and mitigation strategies for the hazardous areas.

3.2.2. Hotspot identification approaches and crash prediction models:

Various methods are available for identifying hotspots, including crash frequency (CF), crash rate (CR), quality control (QC), equivalent property damage only (EPDO), empirical Bayes (EB), full Bayes (FB), and potential for safety improvement (PSI). While CF (Oppe, 1991) and CR (Lord & Park, 2008) are easy to implement, CF overlooks the influence of traffic volume, and CR may incorrectly estimate the effect of the exposure variable. Neither method considers random fluctuations in the crash count (Hauer, 1997). The QC method proposed by Norden et al. (1956) considers the discreteness of crash data and assumes that crashes follow the Poisson distribution. Still, setting a comparison threshold reduces the probability of mistakenly identifying low-traffic areas as safe. However, considering the discreteness of crash data using a threshold may not be applicable for a corridor with high traffic volume and traffic crash occurrences across the entire corridor. The EPDO method, proposed by Tamburri and Smith (1970), considers crash severity in determining hotspots, but it tends to exaggerate the risk in a hotspot location where serious crashes only occasionally occur. However, none of these methods considers the problem of regression to the mean (RTM) in the crash count, which results in incorrect hotspot identification due to random fluctuations in crash characteristics (Hauer, 1980).

The EB method has gained widespread use in hotspot identification, as it accounts for the random fluctuations in crash counts and resolves issues with CF and CR (Hauer, et al., 1988; Persaud, et al., 1999; Sørensen & Elvik, 2007). However, creating the safety analysis model requires many samples, and the model must have a simple form for ease of implementation. Schluter et al. (1997) introduce the FB method as an improved version of EB to estimate the posterior distribution of

crashes and use the estimate as an index to identify hotspots. Miranda-Moreno and Fu (2007) compare FB and EB and indicate that both methods produce similar results with an adequate sample size. However, FB appears superior for small sample sizes. Huang et al. (2009) utilize a Bayesian framework and a hierarchical model considering spatial and temporal correlation and demonstrate that FB can accommodate complex model forms and outperform EB in hotspot identification.

In 1999, Persaud et al. introduced the PSI approach, which calculates the disparity between the anticipated crash frequency at a particular location and the average forecasted crash frequency for that location using crash prediction models. In studies, researchers have commonly employed a combination of EB or FB with PSI (Wang & Feng, 2019) or FB and PSI (Dong, et al., 2016) to account for the stochastic nature of crashes.

Crash prediction models have been investigated for several decades. Initially, researchers relied on linear regression models to estimate crashes and establish correlations between crash frequency and explanatory factors (Joshua & Garber, 1990; Okamoto & Koshi, 1989). However, linear regression models had limitations in handling crash data's discrete and non-negative nature (Lord & Mannering, 2010; Miaou & Lum, 1993). As a result, many researchers adopted count data models for crash prediction. The Poisson regression model was the preferred option for researchers because it assumed that the variance of the data was equivalent to its mean. However, crash data was often characterized by over-dispersion, which occurred when the variance of crash data exceeded its mean. To address the over-dispersion issue, researchers employed negative binomial (NB) regression models (Abdel-Aty & Radwan, 2000; Miaou, 1994). As statistical methods advanced and computing power improved, more sophisticated techniques have been developed to model crash data. Lord and Mannering (2010) and Mannering and Bhat (2014) comprehensively

documented the current trends in crash prediction and future directions. Despite the complexity, the conventional NB model remained popular due to its simple implementation.

The negative binomial (Al-Aamri, et al., 2021) model has various parameterizations in the literature. However, the NB-1 and NB-2 (Cameron & Trivedi, 1986) are commonly utilized for count data modeling (Wang et al., 2019; Giuffre et al., 2014; Ismail & Zamani, 2013; Hilbe, 2011; Winkelmann, 2008; Chang & Xiang, 2003; Miaou & Lord, 2003). These two models differ based on the relationship between the variance and mean of the data. The NB-1 assumes a linear relationship between the variance and mean, while the NB-2 assumes a quadratic relationship. Comprehensive estimation procedures for both forms are outlined in Hardin and Hilbe (2018), Lord and Park (2015), and Hilbe (2011). In traffic safety, the NB-2 is frequently utilized to estimate safety performance functions (SPFs), while the NB-1 is employed in a limited number of studies. For instance, Chang and Xiang (2003) employed both NB-1 and NB-2 models to investigate the association between crashes and traffic congestion levels on freeways. The authors discovered that both models exhibited consistent results for the relationship between crashes and traffic volume, number of through lanes, and median. Giuffre et al. (2014) used NB-1 and NB-2 models to develop SPFs for urban unsignalized intersections and found that NB-1 fitted the data better than NB-2. Wang et al. (2019) used the NB-1 model in combination with the standard Poisson, NB-2, and NB-P models to estimate SPFs and select a superior-performing model for rural two-lane intersections.

However, some drawbacks to using the NB-1 and NB-2 models exist. These models restrict the variance structure when estimating SPFs, with the NB-1 and NB-2 models imposing a linear and quadratic link between the mean and variance of crash data, respectively (Park, 2010). This limited variance structure can result in biased model parameter estimates and inaccurate crash predictions (Wang et al., 2019). Additionally, the NB-1 and NB-2 models are non-nested, meaning a statistical

test cannot directly compare them to determine the better model (Wang et al., 2019; Greene, 2008). To address this issue, Greene (2008) introduced a new NB regression functional form called the NB-P, which nests both the NB-1 and NB-2 models. The NB-P model extends traditional NB models to address the restricted variance structure problem and reduces to the NB-1 when P = 1 and NB-2 when P = 2. The parametric nature of the NB-P model allows analysts to test the NB-1 and NB-2 functional forms against the more general NB-P model. It offers a better model fit and estimation accuracy due to its flexible variance structure. The researchers focused on constructing models only based on traffic factors. The previous research showed that the NB-P model outperformed the Poisson, NB-1, and NB-2 models. They concluded that the NB-P model's malleable variance structure notably enhanced estimation accuracy. In a recent study, Wang and colleagues (2020) utilized the NB-P model to investigate different intersection safety performance functions in urban and suburban settings.

According to the literature review, the NB-P model has yet to be widely adopted in traffic safety studies and crash prediction, despite its potential to improve estimation accuracy compared to traditional NB models. To the best of the authors' knowledge, the NB-P model has been applied to estimate SPFs for urban roads only by Wang et al. (2020) and the NB-P model has not been used for estimating multivariate SPFs. Given the benefits of the NB-P model in terms of offering a flexible variance structure and the ability to statistically test the NB-1 and NB-2 models against an available alternative, the authors see an opportunity to apply it in this study.

The traditional Poisson and negative binomial models used in crash count analysis are not well-equipped to handle excessive zeros (Dong et al., 2014). To address this issue, zero-inflated models have been widely employed (Carson & Mannering, 2001; Qin et al., 2005). These models assume that the extra zeros in the dataset come from two states: a true-zero state where the roadway

segment is inherently safe and a nonzero state where no crashes occur during the observation period (Shankar et al., 1997). As zero-inflated models, zero-inflated Poisson and zero-inflated negative binomial models have been commonly used in the literature to address the issue of excessive zeros in crash frequency analysis (Lee & Mannering, 2002; Chin & Quddus, 2003). These models have been shown to provide a statistically better fit to the data in various studies (Malyshkina & Mannering, 2010). However, it is highly unlikely that roadway segments are intrinsically safe. For instance, crashes can occur due to the unsafe behavior of drivers, even on well-designed roadway segments. Therefore, the fundamental assumption of the zero-inflated model is flawed (Lord et al., 2005, 2007). As an alternative approach, the hurdle model, also known as the two-part model, has been employed to handle excessive zeros in the dataset (Ma et al., 2016). The hurdle model, a two-part model, first determines whether the count value is zero or positive and then, if positive, uses a truncated count distribution for analysis (Cragg, 1971). The hurdle model assumes that roadway segments with zero crashes observed during the study period are only safe during that period, not inherently.

This study selects statistical crash models with Potential for Safety Improvement (PSI) for HSID. Among three main HSID methods (Geographic Information Systems (GIS) based spatial analysis, statistical models, and machine learning), statistical crash models with Potential for Safety Improvement (PSI) represent the prevailing method to detect crash hotspots since PSI accounts for random fluctuations in the crash characteristics. By leveraging the statistical models, studies can identify traffic crash hotspots (Thakali et al., 2015). Traditionally, crash prediction models have been used at the micro-levels for intersections, segments, or corridors. Nevertheless, some researchers have applied crash prediction models at the macro-level by integrating safety into transportation planning zones (Lee, 2014; Park et al., 2015); however, the micro-level models

outperform macro-level models (Huang et al., 2016). Thus, this study develops its tri-dimensional statistical crash models with PSI methodology to identify hotspots at a micro-level.

The study aims to identify the best model for estimating segment potential for safety improvement (PSI) for an urban highway. To achieve this, various count regression models are applied including Poisson (P), negative binomial (NB), negative binomial type p (NBP), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), zero-inflated negative binomial type p (ZINBP), Consul's Generalized Poisson (GP-1), Famoye's Generalized Poison (GP-2), and hurdle regression. Since the crash groups considered in this study may have negligible or significant dispersion, the statistical model development needs to include models capable of addressing a variety of different data structures. listed to handle crash groups with either. Also, as the study conders more traffic crash dimensions to form crash groups, crash groups with an excessive number of zeros appear likely to occur, which requires including zero-inflated count regression models as possible solutions. The study selects a comprehensive set of statistical models to cope with the differences in data structures across the crash groups.

When estimating crash prediction models, previous research includes many features representing traffic operational characteristics, roadway geometry characteristics, and ambient conditions as the explanatory variables in the predictive models. In previous studies, the most popular traffic operational characteristics are the average daily traffic (ADT) (Daniels, et al., 2010; Yu, et al., 2014; Wang, et al., 2017), average daily lane occupancy, average daily speed (Wang, et al., 2017), annual average daily traffic (AADT) (Montella, 2010; Yu, et al., 2014; Eustace, et al., 2015; Yang, et al., 2021), annual average daily traffic for passenger car/trucks ($AADT_{PC}/AADT_{T}$) (Yu, et al., 2014), truck or heavy vehicle percentage (Abdel-Aty & Pande, 2007) (Montella, 2010), speed limit (Abdel-Aty & Pande, 2007; Chiou & Fu, 2013), 85th percentile of speed reduction (ΔV_{85th})

(Montella, 2010), v/c ratio (Abdel-Aty & Pande, 2007), and vehicle miles traveled (Yang, et al., 2021). Some of the road geometry characteristics in the previous studies include lane configuration (Daniels, et al., 2010), curvature (Abdel-Aty & Pande, 2007; Montella, 2010; Chiou & Fu, 2013), deflection angle, tangent length preceding a curve, vertical alignment (Montella, 2010), maximum upward/downward grade (Chiou & Fu, 2013), ramp presence (Abdel-Aty & Pande, 2007; Yu, et al., 2014), segment on bridge (Montella, 2010; Eustace, et al., 2015), median type (Montella, 2010; Yu, et al., 2014), median width, shoulder width (Yu, et al., 2014), number of lanes (Abdel-Aty & Pande, 2007; Chiou & Fu, 2013; Yu, et al., 2014; Wang, et al., 2017; Yang, et al., 2021), inside/outside shoulders width (Wang, et al., 2017), the difference between the assumed and demanded side friction factors (Δf_r^c) (Montella, 2010), road surface condition (Yang, et al., 2021), and road surface type (Eustace, et al., 2015; Yang, et al., 2021). For model estimation, this study evaluates a wide range traffic operational characteristics and roadway geometry characteristics, but only validates and retains a limited number of features based on their data quality.

3.2.3. Traffic Crash Dimensions:

Previous research typically considers three vital characteristics to describe traffic crashes; these are the number of vehicle involved in a traffic crash (crash units), manner of collision (crash type), and crash severity.

The number of vehicles involved in a traffic crash is a critical characteristic, also known as "crash units." This characteristic involves classifying crashes as either single-vehicle (SV) or multi-vehicle (MV). For traffic crash hotspots based on site locations, crash units can be a distinguishing factor in identifying high-risk areas (Wang & Feng, 2019) or contributing factors (Ivan, et al., 1999; Abdel-Aty, et al., 2006), (Yu, et al., 2013) and their related impacts (Yu & Abdel-Aty, 2013; Dong, et al., 2018). The study reveals that separate HSID analyses for single-vehicle (SV) and

multi-vehicle (MV) crashes are necessary due to the dissimilarities in their spatial distribution and associated contributing factors (Wang & Feng, 2019). This holds true for both aggregate and disaggregate approaches in the study of traffic crashes. (Yu & Abdel-Aty, 2013). Wang and Feng (2019) conducted a study on HSID, focusing on single-vehicle (SV) and multi-vehicle (MV) crashes. They used separate crash prediction models that were proposed in earlier studies. (Ivan, et al., 1999; Lord, et al., 2005; Geedipally & Lord, 2010; Ma, et al., 2016). According to Wang and Feng (2019), notable differences in both the significant crash contributing factors and the identified hotspots occur when comparing the results from total crashes to the results from single-vehicle (SV), and multi-vehicle (MV) crashes.

Studying the manner of collision is another essential aspect of traffic crash analysis. This is also known as "crash type" in some studies and refers to the initial event that occurs during a traffic crash or incident when a collision or unexpected event occurs. Including crash type in traffic crash analysis is crucial, particularly for real-time crash risk assessment (Pande & Abdel-Aty, 2006). Numerous empirical studies have demonstrated that traffic crashes are specific to their type, regardless of the level of aggregation (Golob, Recker, & Pavlis, 2008). Cheng et al. (2017) demonstrate the presence of spatial correlations between the crash types of neighboring intersections. Despite the importance of considering crash types, few studies have focused on conducting HSID based on crash types. This study includes crash type as a crucial traffic crash characteristic to identify crash hotspots.

One important aspect of traffic crashes is crash severity, which is determined by the extent of damage caused by the crash, ranging from minor property damage to extremely costly severe injuries or fatality. This can be determined by the level of injuries sustained by those involved or the amount of damage caused (Xu, et al., 2013). Developing crash prediction models that account

for different crash severity levels has provided valuable insights into decreasing the probability of severe crashes (Xu, et al., 2013). Crash severity is essential for studying single-vehicle (SV) crashes and their contributing factors (Jung, et al., 2010) because crash hotspots vary by severity (Dezman, et al., 2016). Several studies have explored crash hotspots with crash severity by applying negative binomial and Bayesian spatial statistical methods (Mitra, 2009), multivariate crash count models, equivalent property damage only (EPDO), and two-stage models (Afghari, et al., 2020). Afghari et al. (Afghari, et al., 2020) declare that the traditional approaches do not consider the unobserved heterogeneity associated with the correlations between crash counts for each severity level. To account for this, the current study considers the severity of crashes in addition to the number of vehicles involved (crash units) and the type of collision (crash types).

3.3. DATA DESCRIPTION

This study aims to analyze mainlane segments in both directions of IH-20 within Dallas County. The Texas Department of Transportation's C.R.I.S. (Crash Record Information System) is the source of the traffic crash data for 2015-2019, including information on crashes, roadway geometry, and traffic characteristics.

3.3.1. Crash Data Features

Features from three groups - crash fields, unit fields, and person fields - are included in the crash data obtained from the Texas Department of Transportation (TxDOT) C.R.I.S. (Crash Record Information System). The crash fields provide the information required for this study, including latitude, longitude, reference marker, offset distance, highway system, roadway part, highway name, manner of collision, crash severity, and other geometric design features such as curve type, curve degree (curvature), curve length, curve delta degree, left shoulder type, left shoulder use, left shoulder width, right shoulder type, right shoulder use, right shoulder width, median type, median, number of lanes, roadbed width, surface condition, surface type, and surface width.

3.3.2. Traffic Characteristics features

The crash data is organized into three categories: crash fields, unit fields, and person fields. For this study, only the crash fields are relevant and include details such as latitude, longitude, highway name, crash severity, and geometric design features. Traffic characteristics such as adjusted average daily traffic amounts, single-unit truck percentages, combo truck percentages, adjusted percentage of average daily traffic for trucks, and speed limits are also included.

3.3.3. Data Preparation

To prepare the data, individual entries for each person involved in a crash are combined into a single entry using the newly assigned crash ID, a combination of crash date, crash time, and existing crash ID. To maintain the crash locations, the crash location reference markers and offset values provided in the crash data sets are utilized to calculate milepost values. This new dataset is then grouped by roadway travel direction for analysis. Only crashes on main roadway segments are considered, and those involving work zones, pedestrians, or wrong-way driving are excluded. This study utilizes feature engineering techniques to filter out crash data pertaining only to the main segment of each roadway while excluding data involving pedestrians, active work zones, construction areas, and wrong-way driving. The crash data points are geovalidated using KMZ files imported to Google Earth® to ensure the consistency of feature values for roadway segments and vehicle travel directions. Geometric design features are also validated using these files, revealing that feature values for left and right shoulder width, median width, number of lanes, surface type, and surface width are inconsistent with true measurements and are thus excluded from this study.

The Texas Department of Transportation (TxDOT) - Traffic Safety Division (2020) provides categories for crash severity levels, with A representing suspected serious injury, B for suspected

minor injury, C for possible injury, K for fatal injury, N for not injured, and 99 for unknown. **Table 3.1.** contains the definitions for these categories. **Table 3.2.** presents the summary statistics of crash data, including the percentage range of traffic crash severity for corridors. The data shows that fatal crashes occur at a considerably low percentage, indicating that they may not be a key feature in differentiating roadway segments and forming clusters. To address this, fatal and suspected serious injury crashes are grouped together, while suspected minor and possible injury crashes are also combined. Non-injury crashes remain a separate characteristic, and crashes with unknown severity are excluded from the study.

Traffic crash groups and their abbreviations are shown in **Fig. 3.1**. In this study, four scenarios are considered to form crash groups as shown (orange boxes) in **Fig 3.1**. The crash count calculated for each of the generated crash group. Depending on the scenario, the naming convention of crash group is in a format of 'A', 'A-B', or 'A-B-C' in which A, B, and C are the traffic crash abbreviations for the number of vehicles involved in crashes, manner of collision, and crash severity, respectively. For instance, 'SV-OBJ-N' is the crash group for single-vehicle object-related crashes with no injuries. Also, 'MV-RRND-B+C' is the feature for multi-vehicle rear-end crashes with suspected minor or possible injuries. In scenario 1, 'TNC' is the crash group including all crashes occurring in each segment. The statistics summary of crash counts for each crash group of each scenario are provided in **Tables 3.3. – 3.12.** According to the statistics summaries, insufficient observations for MV-ANG related crash groups in scenarios 3 and 4 for both IH-20 EB and IH-20 WB exist to perform statistical modeling. Additionally, the crash groups 'SV-OTH-A+K' and 'SV-OVT-A+K' for IIH-20 EB and 'SV-OTH-A+K' and 'SV-OTH-B+C' for IIH-20 WB show insufficient observations for modeling.

Table 3.1. Traffic Crash Categories and Definitions.

Traffic Crash Data Categories	
Number of Vehicle Involved in Crashes	
Single-Vehicle (SV)	Crashes that only involves one motor vehicle.
Multi-Vehicle (MV)	Crashes that involve two or more motor vehicles.
Manner of Collision	
Fixed Object (OBJ)	Crashes that involve hiding fixed objects as the first harmful event.
Over-turned (OVT)	Crashes that the first harmful event is identified as vehicle overturn.
In-Transport (TRNSP)	
Angled (ANG)	Crashes that two motor vehicles are collided at an angle.
Rear-End (RE)	Crashes that a motor vehicle is rear-ended by another motor vehicle.
Sideswipe (SDSP)	Crashes that a motor vehicle is sideswiped by another motor vehicle.
Other (OTH)	Crashes that the manner of collision is none of the items above.
Crash Severity	
A - Suspected Serious Injury	Severe injury that prevents continuation of normal activities leading to temporarily or permanent incapacitation.
B - Suspected Minor Injury	Evident injury such as bruises, abrasions, or minor lacerations which do not incapacitate.
C - Possible Injury	Injury claimed, reported, or indicated by behavior but without visible wounds, includes limping or complaint of pain
K - Fatal	If death resulted due to injuries sustained from the crash, at the scene or within 30 days of crash.
N - Not Injured	The person involved in the crash did not sustain as A, B, C, or K injury.
99 - Unknown	Unable to determine whether injuries exist. Some examples may include hit and run, fled scene, fail to stop and render aid.

Table 3.2. IH 20 (EB/WB) Traffic Crash Statistics by Severity.

Crash Data (2015-2019)	**% Range Across Corridors**		**% Total Crashes in Dallas County**
Crash Severity	Min	Mx	
99 - UNKNOWN	0.63%	3.06%	1.19%
A - SUSPECTED SERIOUS INJURY	1.27%	4.14%	2.05%
B - SUSPECTED MINOR INJURY	6.82%	13.52%	10.55%
C - POSSIBLE INJURY	16.45%	30.57%	21.15%
K - FATAL INJURY	0.24%	1.39%	0.48%
N - NOT INJURED	54.97%	73.46%	64.57%

Table 3.3. Statistical Summary of TNC & SV Crashes in Scenario 2, 3, & 4 for IH-20 EB

IH-20 EB	Scenario 1	Scenario 2 (SV)	Scenario 3 (SV)			Scenario 4 (SV)	
	TNC	SV	SV-OBJ	SV-OTH	SV-OVT	SV-OBJ-A+K	SV-OBJ-B+C
Total	2471	565	464	41	60	30	140
Min	0	0	0	0	0	0	0
Mod	0	0	0	0	0	0	0
Median	6	1	1	0	0	0	0
Max	99	17	14	3	3	2	4
Mean	10.21	2.33	1.92	0.17	0.25	0.12	0.58
Variance	149.32	8.66	6.53	0.22	0.26	0.13	0.85
# of Non-Zero	215	157	143	34	52	28	84
N	242	242	242	242	242	242	242

Table 3.4. Statistical Summary of SV Crashes in Scenario 4 for IH-20 EB

IH-20 EB	Scenario 4 (SV)						
	SV-OBJ-N	SV-OTH-A+K	SV-OTH-B+C	SV-OTH-N	SV-OVT-A+K	SV-OVT-B+C	SV-OVT-N
Total	294	0	6	35	7	35	18
Min	0	0	0	0	0	0	0
Mod	0	0	0	0	0	0	0
Median	1	0	0	0	0	0	0
Max	10	0	1	3	1	3	2
Mean	1.21	0.00	0.02	0.14	0.03	0.14	0.07
Variance	3.38	0.00	0.02	0.17	0.03	0.16	0.08
# of Non-Zero	122	0	6	30	7	32	17
N	242	242	242	242	242	242	242

Table 3.5. Statistical Summary of TNC & MV Crashes in Scenario 2 & 3 for IH-20 EB

IH-20 EB	Scenario 1	Scenario 2 (MV)	Scenario 3 (MV)			
	TNC	MV	MV-ANG	MV-RRND	MV-SDSW	MV-STPD
Total	2471	1906	2	837	823	244
Min	0	0	0	0	0	0
Mod	0	0	0	0	0	0
Median	6	5	0	2	2	0
Max	99	90	1	61	22	14
Mean	10.21	7.88	0.01	3.46	3.40	1.01
Variance	149.32	99.76	0.01	30.17	16.52	2.89
# of Non-Zero	215	207	2	180	174	114
N	242	242	242	242	242	242

Table 3.6. Statistical Summary of MV Crashes in Scenario 4 for IH-20 EB (cont'd)

IH-20 EB	Scenario 4 (MV)					
	MV-ANG-A+K	MV-ANG-B+C	MV-ANG-N	MV-RRND-A+K	MV-RRND-B+C	MV-RRND-N
Total	1	1	0	19	283	535
Min	0	0	0	0	0	0
Mod	0	0	0	0	0	0
Median	0	0	0	0	1	1
Max	1	1	0	2	14	46
Mean	0.00	0.00	0.00	0.08	1.17	2.21
Variance	0.00	0.00	0.00	0.10	3.11	16.09
# of Non-Zero	1	1	0	16	123	156
N	242	242	242	242	242	242

Table 3.7. Statistical Summary of MV Crashes in Scenario 4 for IH-20 EB

IH-20 EB	Scenario 4 (MV)					
	MV-SDSW-A+K	MV-SDSW-B+C	MV-SDSW-N	MV-STPD-A+K	MV-STPD-B+C	MV-STPD-N
Total	9	214	600	5	109	130
Min	0	0	0	0	0	0
Mod	0	0	0	0	0	0
Median	0	0	1	0	0	0
Max	2	7	17	1	8	6
Mean	0.04	0.88	2.48	0.02	0.45	0.54
Variance	0.04	1.93	9.56	0.02	0.95	0.94
# of Non-Zero	8	105	155	5	64	81
N	242	242	242	242	242	242

Table 3.8. Statistical Summary of TNC & SV Crashes in Scenario 2, 3, & 4 for IH-20 WB

IH-20 WB	Scenario 1	Scenario 2 (SV)	Scenario 3 (SV)			Scenario 4 (SV)	
	TNC	SV	SV-OBJ	SV-OTH	SV-OVT	SV-OBJ-A+K	SV-OBJ-B+C
Total	2367	548	474	43	31	27	160
Min	0	0	0	0	0	0	0
Mod	0	0	0	0	0	0	0
Median	7	1	1	0	0	0	0
Max	62	14	12	3	2	2	7
Mean	9.78	2.26	1.96	0.18	0.13	0.11	0.66
Variance	107.74	6.83	5.26	0.22	0.14	0.13	1.02
# of Non-Zero	217	172	167	35	28	23	98
N	242	242	242	242	242	242	242

Table 3.9. Statistical Summary of SV Crashes in Scenario 4 for IH-20 WB

IH-20 WB	Scenario 4 (SV)						
	SV-OBJ-N	SV-OTH-A+K	SV-OTH-B+C	SV-OTH-N	SV-OVT-A+K	SV-OVT-B+C	SV-OVT-N
Total	287	2	3	38	4	12	15
Min	0	0	0	0	0	0	0
Mod	0	0	0	0	0	0	0
Median	1	0	0	0	0	0	0
Max	7	1	1	3	1	1	1
Mean	1.19	0.01	0.01	0.16	0.02	0.05	0.06
Variance	2.28	0.01	0.01	0.19	0.02	0.05	0.06
# of Non-Zero	131	2	3	32	4	12	15
N	242	242	242	242	242	242	242

Table 3.10. Statistical Summary of TNC & MV Crashes in Scenario 2 & 3 for IH-20 WB

IH-20 WB	Scenario 1	Scenario 2 (MV)	Scenario 3 (MV)			
	TNC	MV	MV-ANG	MV-RRND	MV-SDSW	MV-STPD
Total	2367	1819	5	852	749	213
Min	0	0	0	0	0	0
Mod	0	0	0	0	0	0
Median	7	5	0	2	2	0
Max	62	49	2	31	21	11
Mean	9.78	7.52	0.02	3.52	3.10	0.88
Variance	107.74	71.07	0.03	19.89	12.99	2.11
# of Non-Zero	217	210	4	182	182	107
N	242	242	242	242	242	242

Table 3.11. Statistical Summary of MV Crashes in Scenario 4 for IH-20 WB (cont'd)

IH-20 WB	Scenario 4 (MV)					
	MV-ANG-A+K	MV-ANG-B+C	MV-ANG-N	MV-RRND-A+K	MV-RRND-B+C	MV-RRND-N
Total	0	2	3	28	286	538
Min	0	0	0	0	0	0
Mod	0	0	0	0	0	0
Median	0	0	0	0	1	1
Max	0	1	1	2	11	20
Mean	0.00	0.01	0.01	0.12	1.18	2.22
Variance	0.00	0.01	0.01	0.11	2.60	9.68
# of Non-Zero	0	2	3	27	134	160
N	242	242	242	242	242	242

Table 3.12. Statistical Summary of MV Crashes in Scenario 4 for IH-20 WB

IH-20 WB	Scenario 4 (MV)					
	MV-SDSW-A+K	MV-SDSW-B+C	MV-SDSW-N	MV-STPD-A+K	MV-STPD-B+C	MV-STPD-N
Total	12	195	542	4	83	126
Min	0	0	0	0	0	0
Mod	0	0	0	0	0	0
Median	0	0	1	0	0	0
Max	2	8	18	1	6	5
Mean	0.05	0.81	2.24	0.02	0.34	0.52
Variance	0.06	1.55	7.70	0.02	0.59	0.88
# of Non-Zero	11	101	168	4	56	78
N	242	242	242	242	242	242

3.4. METHODOLOGY

3.4.1. Introduction

This study uses count regression models to predict the crash frequencies and investigates roadway geometric features and traffic characteristics as explanatory variables. Based on the expected crash frequencies from the regression models, the potential for safety improvement (PSI) is calculated by accounting for dispersions and used to identify traffic crash hotspots. Identifying traffic crash hotspots involves the examination of various traffic crash characteristics, such as the number of vehicles involved (crash units) (Wang & Feng, 2019), the manner of collision (crash type) (Golob, et al., 2004a; Cheng, et al., 2017), and the severity of the crash (crash severity) (Abdel-Aty, 2003; Jung, et al., 2010).

3.4.2. Poisson and negative binomial regression model

In traffic crash analysis, a typical prediction model is Poisson regression model which ignores the over-dispersion in the crash data (Lord & Mannering, 2010). To address the over-dispersion problem, negative binomial regression model (Al-Aamri, Hornby et al.) has been applied in previous studies (Anastasopoulos & Mannering, 2009; Geedipally & Lord, 2010; Ma, et al., 2017; Ma, et al., 2017). According to (Chow & Steenhard, 2009), the standard negative binomial probability mass function and the negative binomial regression model are defined as:

$$f(Y_i = y_i \mid \lambda_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\,\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\lambda_i + \alpha^{-1}}\right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\lambda_i + \alpha^{-1}}\right)^{y_i} \quad (1)$$

$$\log(\lambda_i) = \beta_0 + \beta\, X_i \quad (2)$$

where y_i is the number of traffic crashes on segment i during a period of time, and λ_i is the mean predicted crash frequency for segment i. The assumption is that λ_i is a function of exploratory vector X_i. β_0 and β are regression coefficients and α is the over-dispersion parameter. Noteably, the NB model contracts to a Poisson regression model when there is no over-dispersion (i.e. $\alpha = 0$). The traditional NB model is also known as the NB-2 model (Khattak, et al., 2021). Replacing α^{-1} in the NB-2 model (Eq. 1) with $\alpha^{-1}\lambda_i$ yields a re-parameterization of the variance structure the NB model, transforming it to a different functional form known as the NB-1 (Khattak, et al., 2021). Another functional form of the negative binomial regression model called negative binomial model-type P exists. The NB-P model is introduced by Greene (2008), and the parameter "P" represents the relationship between mean and variance: $E(Y_i) = \lambda_i, Var(Y_i) = \lambda_i + \alpha\,\lambda_i^P$.

$$f(Y_i = y_i \mid \lambda_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1}\lambda_i^{2-P})}{\Gamma(y_i + 1)\,\Gamma(\alpha^{-1}\lambda_i^{2-P})} \left(\frac{\alpha^{-1}\lambda_i^{2-P}}{\lambda_i + \alpha^{-1}\lambda_i^{2-P}}\right)^{\alpha^{-1}\lambda_i^{2-P}} \left(\frac{\lambda_i}{\lambda_i + \alpha^{-1}\lambda_i^{2-P}}\right)^{y_i} \quad (3)$$

3.4.3. Zero-inflated regression model

The zero-inflated models include zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models to estimate a dual-state process. ZIP and ZINB models assume that two distinct processes are responsible for zero-count and non-zero-count observations. Zero-inflated models have two components: one to model zero count observations and another for non-zero-count observations. The ZINB probability mass function is defined as:

$$f(Y_i = y_i) = \begin{cases} q_i + (1 - q_i)\, g(y_i = 0) & \text{if } y_i = 0 \\ (1 - q_i)\ g(y_i) & \text{if } y_i > 0 \end{cases} \quad (4)$$

where q_i is the logistic link function to estimate the probability of zero-count and non-zero-count observations, respectively. In equation Eq. 1, $g(y_i)$ is the standard negative binomial probability density function shown in equation Eq. 4. In the absence of over-dispersion (α=0), the ZINB model reduces to a ZIP model.

3.4.4. Generalized Poisson regression model

The Generalized Poisson (GP) regression frequently represents count data. In contrast to the Negative Binomial (NB) regression, the GP models can account for both over-dispersed and under-dispersed data. Like the NB regression, the GP model requires an additional parameter, a scale or dispersion parameter. A unique aspect of the GP dispersion parameter is that it can assume positive or negative values for over-dispersed and under-dispersed data, respectively. The GP distribution's probability mass function (p.m.f.) is defined as:

$$\Pr(Y_i = y_i) = \frac{\theta(\theta + \alpha y_i)^{y_i - 1} \exp(-\theta - \alpha y_i)}{y_i!}, \quad y_i = 0, 1, 2, \cdots, \tag{5}$$

where $\theta > 0$, and $0 \leq \alpha < 1$. For the GP regression, the mean and the variance are $E(Y_i) = \lambda_i = (1-\alpha)^{-1}\theta, Var(Y_i) = (1-\alpha)^{-3}\theta = (1-\alpha)^{-2}\lambda_i = \phi\lambda_i$. The dispersion factor is denoted as ϕ.

3.4.5. Hurdle regression model

The hurdle regression model is another statistical technique to model data containing excess zero. The model is a two-part regression model that first models the probability of a zero count and then models the count distribution for non-zero values. The hurdle regression is defined as follows:

$$f(Y_i = y_i | \pi, \theta) = \begin{cases} 1 - \pi, & \text{if } y_i = 0 \\ \pi f(y_i|\theta), & \text{if } y_i > 0 \end{cases} \tag{6}$$

where $0 \leq \pi \leq 1$ is the probability that the dependent variable is positive, $f(y_i|\theta)$ is the probability density of the dependent variable given that it is positive, and θ is the parameters of the model (Mullahy, 1986). The first part estimates the probability of a zero-count using a binary logistic

regression model. The second part models the non-zero counts using a standard regression model, such as Poisson or negative binomial regression. Hurdle regression is useful in situations where the traditional regression models for count data, such as Poisson or negative binomial regression, are not appropriate due to the presence of excess zeros. Examples of such data include health care utilization, insurance claims, and environmental counts (Afghari, et al., 2021).

3.4.6. Model comparison and model selection

This study estimates nine different types of count data models and investigates what model type provides the best performance. According to Hilbe (2014), three primary statistical tests to compare zero-inflated models exist: boundary likelihood ratio test, Vuong test, and AIC/BIC tests. This study applies the Vuong test, and AIC/BIC tests to compare models.

3.4.6.a. *Vuong test:*

The second test is the Vuong test (Vuong, 1989), which compares a ZI model to a non-inflated model that is not nested within it, such as ZINB to NB. The test statistic follows a normal distribution N(0,1), where large positive values favor ZINB, and significant negative values select NB. The Vuong test statistical (V) is introduced as:

$$V = \frac{\overline{m} * \sqrt{N}}{S_m} \tag{7}$$

in which $m_i = log\left[\frac{f_1(y_i)}{f_2(y_i)}\right]$, N is the number of observation; $\overline{m}$ and S_m are the mean and the standard deviation of m_i, respectively; f_1, f_2 are two competing models.

There are three potential outcomes for V:

(1) If $|V| < 1.96$ for a confidence level of 0.95 then neither model is favored by the test results;

(2) If $V \gg 0$, then model 1 is favored;

(3) If $V \ll 0$, then model 2 is favored.

If the p-value is insignificant, the Vuong test cannot differentiate between ZINB and NB. However, the Vuong test is biased toward ZI models, and correction factors are available to account for this.

3.4.6.b. *AIC/BIC tests:*

Comparing different models and selecting the best model using a consistent methodology is crucial. The Akaike information criterion (AIC), proposed by Akaike (1974), is a widely used method for model selection (Akaike, 1974). AIC provides an estimate for the amount of relative information lost when a model represents the data-generating process. AIC is defined as:

$$AIC = -2Ln(L) + 2K \quad (8)$$

where K represents the number of the independent variables and L is the maximum value of the likelihood function. The Bayesian Information Criterion (BIC) is another statistical criterion used to evaluate the relative fit of different statistical models to a given dataset (Schwarz, 1978). BIC is a measure of the goodness of fit of a model, adjusted for the number of parameters in the model, and can be applied to compare the different models. The BIC statistic is defined as:

$$BIC = -2Ln(L) + Ln(n) * K \quad (9)$$

where L represents the likelihood of the data given the model, K is the number of independent variables in the model, and n is the number of observations. The BIC penalizes models with more parameters, as represented by the second term of the equation, and aims to balance the trade-off between model complexity and goodness of fit. The BIC can be used to compare models and select the one that provides the best balance between fit and parsimony. The model with the lowest BIC is favored, and a difference in BIC of more than ten between the two models is considered strong evidence for selecting the model with the lower BIC. The BIC is a popular alternative to the Akaike Information Criterion (AIC) and is often used in the context of maximum likelihood estimation.

While the AIC tends to favor more complex models than the BIC, the BIC tends to be more effective at selecting the correct model when the sample size is small, or the number of parameters is large. Finally, AIC/BIC tests should be used to determine if a standard non-inflated model might fit the data better than a ZI model. A negative binomial or NB-P model may be more appropriate than a zero-inflated Poisson or negative binomial model.

3.4.7. Hotspots identification (HSID)

This study identifies hotspots using the Potential for Safety Improvement (PSI) index, which determines locations on the roadway that can become safer by implementing safety measures. Since traffic crashes are random occurrences, the number of observed traffic crashes in a given segment may fluctuate naturally over time. Therefore, the observed frequency of traffic crashes over a brief period cannot be used as a reliable indicator of the expected frequency of traffic crashes under the same circumstances over an extended period (2010). Instead, the Bayesian expected number of traffic crashes is used to address the regression-to-the-mean issue for a more reliable estimate of the expected frequency of traffic crashes. The PSI can be obtained by subtracting the predicted crash count from the expected crash frequency estimated by a crash prediction model (Montella, 2010). Sites with higher PSI values are more likely to benefit from safety improvements, and several studies (Persaud, et al., 1999) (El-Basyouny & Sayed, 2006) use it to rank hotspots. The study calculates the PSI using the empirical bayes (EB) method. Using the crash prediction model, the EB and the PSI are as follows (2010):

$$EB_i = w_i \times N_{predicted,\ i} + (1 - w_i) \times N_{observed,\ i} \quad (10)$$

$$PSI_i = EB_i - N_{predicted,\ i} \quad (11)$$

where

EB_i : the expected average crashes frequency for the segment i;

w_i: the weighted adjustment to be applied on the regression prediction;

$N_{predicted,\ i}$: the predicted average crashes frequency for the segment i, predicted using the regression model for the segment i ;

$N_{observed,\ i}$: the observed average crashes frequency for the segment i;

PSI_i : the potential for safety improvement for the segment i.

The weighted adjustment factor, w_i, can be calculated using the overdispersion parameter, k, associated with the regression model:

$$w_i = \frac{1}{1 + k \times N_{predicted,\ i}} \tag{12}$$

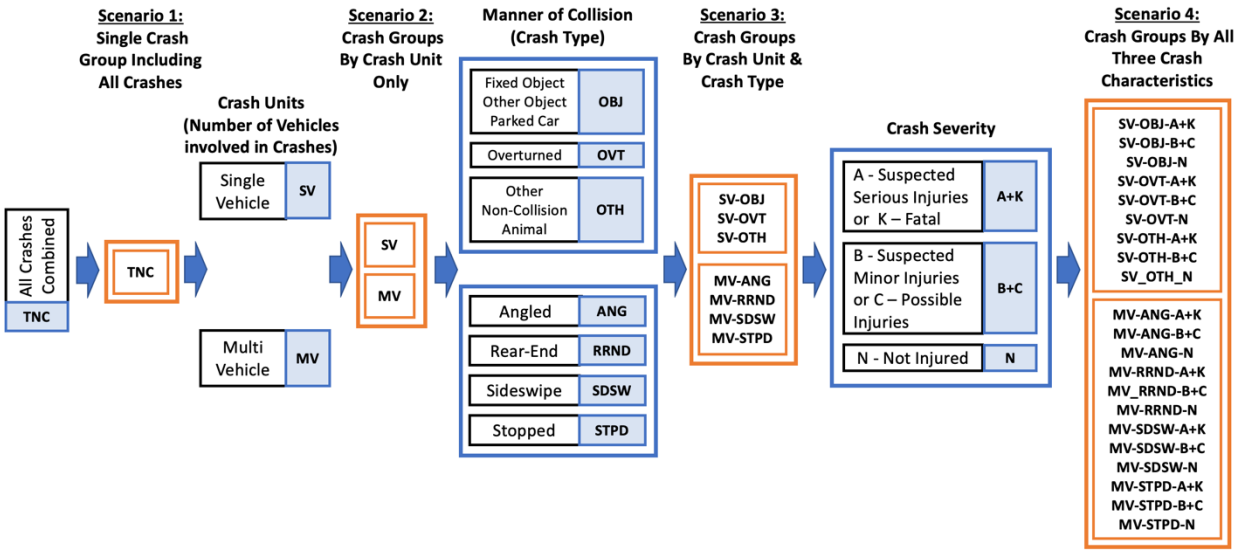


Fig. 3.1. Three Dimensions of Traffic Crashes and Four Scenarios.

Typically, traffic safety research must address unobserved heterogeneity. Inherently, the scope of the study allows some information to account for all potential causal factors contributing to traffic crashes (Chang, Yasmin, Huang, & Chan, 2021; Mannering, Shankar, & Bhat, 2016). A prevalent strategy to tackle unobserved heterogeneity involves categorizing traffic crash data into homogeneous groups using traffic crash attributes (Mannering & Bhat, 2014). This study attempts to address the

unobserved heterogeneity by considering higher dimensions of traffic crash characteristics, including crash units, manner of collision, and crash severity in scenarios 2, 3, and 4 as shown in **Fig. 3.1.**

3.4.8. Modeling Process

The study estimates various count data regression models to characterize different crash groups using four scenarios. The crash groups for each scenario are shown in **Fig. 3.1**. A flowchart summarizing the methodology used in the study is presented in **Fig. 3.2**. The explanatory variables are shown in **Table 3.3**. The investigation checks the explanatory variables for multicollinearity using the variance inflation factor (VIF) and determines no serious multicollinearity between the explanatory variables since the $\overline{VIF}$ is 4.14 and 4.51 and VIF_{MAX} is 8.98 and 9.44 for IH 20 EB and IH 20 WB, respectively. The author evaluates a total of nine regression models, including Poisson, NB, NBP, ZIP, ZINB, ZINBP, GP-1, GP-2, and Hurdle regression models, to identify the most suitable regression model for each group of traffic crashes. A brute force approach is employed to determine the recommended parameter P for the NBP and ZINBP models by analyzing model performance for parameter values ranging from 1 to 2, with an increment of 0.01. The NBP and ZINBP models with the best performance are compared with other models. The study considers the crash count for each crash group as the dependent variable. For each crash group, the dataset for regression consists of the crash count data and explanatory variables data. The analysis splits the dataset into train and test sets. The process estimates count data regression models using the training set and applies the estimated model to the test set to generate the model RMSE.

A model selection process is adopted to determine the outperforming model for each group of traffic crashes. Vuong's test compares the base models (Poisson, NB, and NBP) with their corresponding zero-inflated models (ZIP, ZINB, and ZINBP). The process compares the models selected by Vuong's test with the remaining models (GP-1, GP-2, and Hurdle) based on AIC and

BIC values to identify the final model for each group of traffic crashes. The EB and PSI methods identify traffic crash hotspots for each group of traffic crashes, using the selected model for each group of crashes.

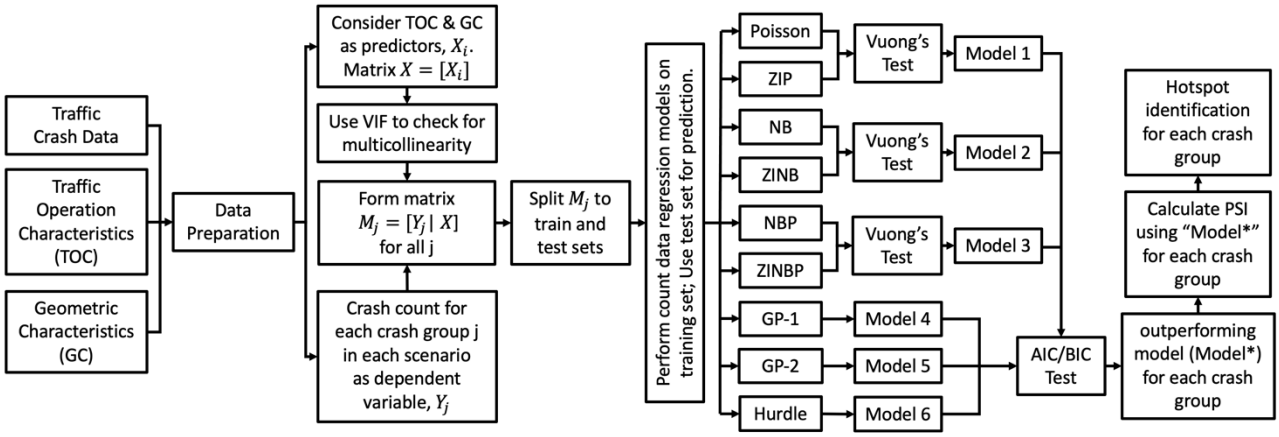


Fig. 3.2. Modeling Process Flow Chart.

3.4.9. Modeling Implementation:

This study methodology develops a library of functions in Python 3 to perform the entire process, from data cleaning and preparation to feature selection, developing regression models, and model selection. To avoid overfitting, the traffic crash data for each travel direction is split to train and test sets with ratio of 70% and 30%, respectively. For each run, the average computing time is 1151.21s and 801.52s for EB and WB (6-Core Intel Core i7, 2.6 GHz CPU, 16 GB memory), respectively.

3.5. RESULTS

This section provides the modeling results of the crash prediction models for four scenarios. Also, hotspots identification results for each groups of traffic crashes are shown.

3.5.1. Modeling results

The regression models are performed on crash groups for all four scenarios. The independent variables are the variables that are shown in **Table 3.13**. **Table 3.14** illustrates the model selection

for crash group 'MV-RRND-N'**.** As described in Section 4.6, the analysis applies Vuong's test to each pair of non-nested models (the basic model and its corresponding zero-inflated model) to select the outperforming models for each pair of non-nested models. As shown in **Table 3.14**, the Vuong's test statistical V values are -2.1653, 0.0002, and 1.5808 for the non-nested models [Poisson, ZIP], [NB, ZINB], and [NBP, ZINBP], respectively. Since $V < 0$ for [Poisson, ZIP], the ZIP model is selected as the outperforming model. Because $V > 0$ for non-nested models [NB, ZINB] and [NBP, ZINBP], the NB and NBP models outperform their zero-inflated versions. The model selection process compares the models emerging from these pairwise comparisons with the GP1, GP2, and Hurdle models using AIC values. For this example, the NB model has the minimum AIC value of 632.82 and represents the best model for crash group 'MV-RRND-N'. For each crash group, the dependent variables are the traffic crash count for the crash group. **Tables 3.15. -3.22.** and **3.23. -3.31.** show the outperforming model parameter estimation results for IH 20 EB and IH 20 WB, respectively.

As shown in the modeling results, the type of outperforming model may vary for each group of crashes. Also, the dispersion values differ for each group of crashes. Subsections 5.1.1. and 5.1.2 investigate the modeling results of each scenario for IH 20 EB and IH 20 WB.

Table 3.13. Description of Explanatory Variables

Explanatory Variables	Description
$AADT_T$	Annual average daily traffic for single-unit/combo trucks (1000 vehicles per day, 1000 vpd)
$AADT_{NT}$	Annual average daily traffic for non-truck vehicle (1000 vehicles per day, 1000 vpd)
CDD	Horizontal curve delta angle a.k.a central angle (degree)
R	Horizontal curve radius (1000 ft)
TR-Seg	Binary variable with value of 1 if traffic crash site is on a segment of roadway transitioning from straight (tangent) segment to a curved segment or a curved-to-right segment to a curved-to-left segment or vice versa. Otherwise, 0.
LT-Seg	Binary variable with value of 1 if traffic crash site is on a curved-to-left segment. Otherwise, 0.
RT-Seg	Binary variable with value of 1 if traffic crash site is on a curved-to-right segment. Otherwise, 0.

Table 3.14. An example of model selection process

	MV-RRND-N								
Model	Poisson	ZIP	NB	ZINB	NBP	ZINBP	GP-1	GP-2	Hurdle
Vuong's Test Statistical V	V = -2.1653		V = 0.0002		V = 1.5808		-	-	-
p-value	0.0152		0.4999		0.0570		-	-	-
Vuong's Test	V < 0		V > 0		V >0		-	-	-
Selected Model by Vuong's Test	ZIP		NB		NBP		-	-	-
AIC	799.07		632.82*		637.25		637.94	9066.31	806.50

* The model with smallest AIC as selected model for the crash group.

3.5.2. IH 20 EB modeling results

In scenario 1 (**Table 3.15.**), for TNC, the explanatory variables $AADT_T$ and $AADT_{NT}$ are significant at the 99% confidence interval (C.I.) with GP-2 representing the outperforming model. Looking at the scenario 2 (**Table 3.15.**), $AADT_T$ and $AADT_{NT}$ remain significant at the 99% C.I. for both SV and MV crashes. Also, *CDD* and *RT-Seg* become significant at the 95% and 90% C.I. for only SV crashes. Navigating to scenario 3, the modeling results for traffic crashes categorized under SV show that for SV-OBJ crashes, *CDD* level of significance improves to 99% C.I. while the other explanatory variables for SV crashes maintain the same level of significance for SV-OBJ crashes. For SV-OTH crashes, the $AADT_T$ level of significance drops to 90% C.I. and the curve radius *R* becomes significant at the 95% C.I. Meanwhile, the results for SV-OVT show $AADT_T$ is the only variable remaining significant with a level of significance at the 99% C.I. In scenario 3, the modeling results for traffic crashes under MV category show that $AADT_T$ level of significance drops to the 95% C.I. for MV-RRND crashes but $AADT_{NT}$ maintains its significance at the 99% C.I (**Table 3.17.**). For MV-SDSW, variables $AADT_T$, $AADT_{NT}$, and *CDD* become significant at 95%, 99%, and 99% C.I., respectively. However, only $AADT_{NT}$ appears significant at the 99% C.I.

In scenario 4 (**Table 3.18.**), the modeling results show $AADT_T$ is significant at the 99% C.I. only for SV-OBJ-N crashes and is not fount significant for SV-OBJ-A+K and SV-OBJ-B+C. $AADT_{NT}$ becomes significant at the 90%, 99%, and 99% C.I. for SV-OBJ-A+K, SV-OBJ-B+C, and SV-OBJ-N crashes, respectively. This shows that the $AADT_{NT}$ level of significance decreases for SV-OBJ-A+K crashes compared to the $AADT_{NT}$ level of significance for SV-OBJ crashes. *CDD* appears significant at the 95%, 90%, and 99% C.I. for SV-OBJ-A+K, SV-OBJ-B+C, and SV-OBJ-N crashes, respectively. In contrast with SV-OBJ crashes, the *CDD* level of significance increases for SV-OBJ-N crashes while it decreases for SV-OBJ-A+K crashes. *R* becomes significant at the 90% C.I for SV-OBJ-N crashes but it is not significant for SV-OBJ. In scenario 4 for MV crashes (**Table 3.19.**), *R* is significant at the 90% and 95% C.I. for SV-OTH-B+C and SV-OTH-N, respectively. For SV-OTH-B+C and SV-OTH-N crashes, only the $AADT_{NT}$ becomes significant at the 95% and 90% C.I.

Table 3.15. Modeling results for IH 20 EB (Scenario 1 and 2)

Variables	Scenario 1		Scenario 2			
	TNC		SV		MV	
	Mean	St.	Mean	St.	Mean	St.
Intercept	-1.677[a]	0.508	-1.874[a]	0.525	-1.775[a]	0.524
$AADT_T$	0.119[a]	0.033	0.102[a]	0.03	0.1[a]	0.034
$AADT_{NT}$	0.044[a]	0.006	0.027[a]	0.006	0.044[a]	0.006
CDD	0.057	0.039	0.071[b]	0.032	0.043	0.033
R	0.006	0.033	0.033	0.029	-0.001	0.033
TR-Seg	-0.625	0.477	-0.644	0.45	-0.584	0.467
LT-Seg	-0.611	0.658	-0.595	0.583	-0.436	0.612
RT-Seg	-0.93	0.82	-1.426[c]	0.813	-0.67	0.742
Dispersion	0.222[a]	0.021	0.687[a]	0.132	0.251[a]	0.026
Model	GP-2		NB		GP-2	
RMSE	102.63		33.452		78.975	
AIC	1057.66		676.983		975.289	
BIC	1085.829		705.152		1003.458	
Mean	10.331		2.438		7.893	
Variance	152.264		9.268		100.171	

Table 3.16. Modeling results for IH 20 EB (Scenario 3 (SV))

Variables	Scenario 3 (SV)					
	SV-OBJ		SV-OTH		SV-OVT	
	Mean	St.	Mean	St.	Mean	St.
Intercept	-1.663[a]	0.534	-3.436[a]	1.072	-4.831[a]	1.323
$AADT_T$	0.082[a]	0.029	0.11[c]	0.067	0.198[a]	0.059
$AADT_{NT}$	0.023[a]	0.007	0.008	0.014	0.022	0.015
CDD	0.057[a]	0.015	-0.008	0.068	-0.01	0.056
R	0.034	0.027	0.102[b]	0.043	0.018	0.06
TR-Seg	-0.348	0.366	0.255	0.923	-0.873	1.103
LT-Seg	-0.41	0.417	0.604	1.198	0.641	0.971
RT-Seg	-1.169[c]	0.62	-9.554	135.37	0.219	1.486
Dispersion	0.724[a]	0.147	0.252	0.344		
Model	GP-1		GP-2		Poisson	
RMSE	23.495		4		4.472	
AIC	615.823		179.039		211.569	
BIC	643.992		207.208		236.609	
Mean	1.917		0.169		0.248	
Variance	6.557		0.216		0.262	

Table 3.17. Modeling results for IH 20 EB (Scenario 3 (MV))

Variables	Scenario 3 (MV)					
	MV-RRND		MV-SDSW		MV-STPD	
	Mean	St.	Mean	St.	Mean	St.
Intercept	-1.914[a]	0.566	-1.26[b]	0.503	-2.84[a]	0.791
$AADT_T$	0.068[b]	0.034	0.066[b]	0.027	0.045	0.042
$AADT_{NT}$	0.039[a]	0.007	0.027[a]	0.006	0.037[a]	0.009
CDD	0.027	0.032	0.045[a]	0.017	-0.032	0.056
R	-0.012	0.035	0.015	0.029	-0.043	0.05
TR-Seg	-0.389	0.49	-0.534	0.383	-0.317	0.718
LT-Seg	-0.382	0.624	-0.281	0.398	0.511	0.887
RT-Seg	-0.809	0.794	-0.373	0.515	0.33	1.21
Dispersion	1.042[a]	0.157	3.238[a]	0.6	0.426[a]	0.101
Model	NB		NBP		GP-2	
RMSE	37.108		34.641		15.492	
AIC	760.024		745.51		441.529	
BIC	788.194		773.679		469.698	
Mean	3.459		3.401		1.008	
Variance	30.299		16.59		2.904	

Table 3.18. Modeling results for IH 20 EB - Scenario 4 (SV)

Variables	Scenario 4 (SV)					
	SV-OBJ-A+K		SV-OBJ-B+C		SV-OBJ-N	
	Mean	St.	Mean	St.	Mean	St.
Intercept	-3.306[a]	1.052	-2.478[a]	0.744	-2.332[a]	0.668
$AADT_T$	-0.032	0.07	0.041	0.038	0.098[a]	0.034
$AADT_{NT}$	0.027[c]	0.014	0.029[a]	0.009	0.023[a]	0.008
CDD	0.101[b]	0.043	0.034[c]	0.02	0.06[a]	0.017
R	-0.22	0.401	0.011	0.033	0.047[c]	0.028
TR-Seg	-0.756	1.182	0.005	0.431	-0.54	0.437
LT-Seg	-2.095	1.689	0.143	0.491	-0.383	0.467
RT-Seg	-1.034	1.565	-0.409	0.736	-1.325[c]	0.742
Dispersion					0.537[a]	0.132
Model	Poisson		ZIP		GP-1	
RMSE	2.646		7.874		17.635	
AIC	149.532		363.079		491.308	
BIC	174.571		391.248		519.477	
Mean	0.124		0.579		1.215	
Variance	0.126		0.851		3.389	

Table 3.19. Modeling results for IH 20 EB - Scenario 4 (SV)

Variables	Scenario 4 (SV)							
	SV-OTH-B+C		SV-OTH-N		SV-OVT-B+C		SV-OVT-N	
	Mean	St.	Mean	St.	Mean	St.	Mean	St.
Intercept	-4.608[b]	1.972	-3.523[a]	1.177	-5.745[a]	1.851	-6.22[b]	2.481
$AADT_T$	0.136	0.146	0.1	0.071	0.196[b]	0.077	0.221[c]	0.116
$AADT_{NT}$	-0.005	0.029	0.011	0.015	0.029	0.021	0.016	0.028
CDD	-0.02	0.13	-0.004	0.071	0.02	0.066	-0.052	0.084
R	0.147[b]	0.069	0.084[c]	0.044	0.045	0.059	-3.204	3.732
TR-Seg	-47.471	3.98E+10	0.436	0.92	-19.952	1.53E+04	4.613	4.839
LT-Seg	1.102	2.264	0.466	1.268	-0.092	1.262	5.671	5.168
RT-Seg	-47.131	4.76E+10	-92.935	1.88E+20	-21.96	6.29E+04	6.957	5.44
Dispersion								
Model	ZIP		ZIP		Poisson		Poisson	
RMSE	0		4		3		2.45	
AIC	64.876		158.209		153.575		98.774	
BIC	93.045		186.379		178.614		123.813	
Mean	0.025		0.145		0.145		0.074	
Variance	0.024		0.174		0.157		0.077	

The modeling results for scenario 4 MV is shown in **Table 3.20.-3.22.**. No explanatory variables appear statistically significant for the MV-RRND-A+K, MV-SDSW-A+K, and MV-STPD-A+K crashes. $AADT_T$ and $AADT_{NT}$ are significant at the 95% and 99% C.I. for MV-RRND-B+C crashes as they were for MV-RRND crashes in scenario 3 (**Table 3.17.**). for MV-RRND-N crashes, $AADT_{NT}$ remains significant at the same level of 99% C.I. as MV-RRND crashes but the $AADT_T$ is not statistically significant.

For MV-SDSW-B+C crashes, the explanatory variables $AADT_T$, $AADT_{NT}$, *CDD, and R* become significant at the 90%, 99%, 90%, and 90% C.I., respectively. **Table 3.21.** shows that $AADT_T$, $AADT_{NT}$, and *CDD* appear significant at the 90%, 99%, and 99% C.I. for MV-SDSW-N, respectively. $AADT_{NT}$ is the only explanatory variable found significant for MV-STPD-B+C and MV-STPD-N with the level of significance at the 99% C.I. (**Table 3.22.**).

Table 3.20. Modeling results for IH 20 EB - Scenario 4 (MV)

	Scenario 4 (MV)					
Variables	**MV-RRND-A+K**		**MV-RRND-B+C**		**MV-RRND-N**	
	Mean	**St.**	**Mean**	**St.**	**Mean**	**St.**
Intercept	-3.13[b]	1.532	-2.51[a]	0.728	-2.255[a]	0.643
$AADT_T$	0.015	0.101	0.092[b]	0.041	0.039	0.037
$AADT_{NT}$	0.022	0.02	0.028[a]	0.009	0.042[a]	0.008
CDD	-0.228	0.394	0.013	0.036	0.039	0.038
R	-0.62	4.872	0.014	0.04	-0.028	0.04
TR-Seg	2.562	5.775	0.044	0.597	-0.777	0.566
LT-Seg	-89.992	1.72E+20	-0.008	0.726	-0.649	0.725
RT-Seg	3.947	6.166	-0.003	0.887	-2.066[c]	1.09
Dispersion			0.965	0.67	1.123[a]	0.192
Model	ZIP		ZINB		NB	
RMSE	3		13.602		27.477	
AIC	104.659		494.913		632.817	
BIC	132.829		526.212		660.986	
Mean	0.079		1.169		2.211	
Variance	0.098		3.121		16.159	

Table 3.21. Modeling results for IH 20 EB - Scenario 4 (MV)

Scenario 4 (MV)						
Variables	**MV-SDSW-A+K**		**MV-SDSW-B+C**		**MV-SDSW-N**	
	Mean	**St.**	**Mean**	**St.**	**Mean**	**St.**
Intercept	-3.423	2.667	-2.854[a]	0.826	-1.506[a]	0.55
$AADT_T$	0.041	0.142	0.076[c]	0.04	0.055[c]	0.031
$AADT_{NT}$	0.027	0.033	0.029[a]	0.01	0.028[a]	0.008
CDD	0.054		0.042[c]	0.025	0.048[a]	0.018
R	-0.254	1	0.055[c]	0.029	0.005	0.033
TR-Seg	-2.942	6.494	-0.792	0.691	-0.505	0.502
LT-Seg	-0.195		-0.414	0.586	-0.238	0.485
RT-Seg	-2.709	6.446	-0.018	0.663	-0.606	0.596
Dispersion			0.476[a]	0.134	2.371[b]	0.956
Model	ZIP		GP-1		ZINBP	
RMSE	1.414		11.705		27.24	
AIC	72.743		416.737		664.89	
BIC	100.912		444.906		696.189	
Mean	0.037		0.884		2.479	
Variance	0.044		1.937		9.595	

Table 3.22. Modeling results for IH 20 EB - Scenario 4 (MV)

Scenario 4 (MV)						
Variables	**MV-STPD-A+K**		**MV-STPD-B+C**		**MV-STPD-N**	
	Mean	**St.**	**Mean**	**St.**	**Mean**	**St.**
Intercept	-5.13[c]	2.727	-3.296[a]	1.007	-3.587[a]	1.043
$AADT_T$	-0.342	0.316	0.055	0.057	0.039	0.05
$AADT_{NT}$	0.064	0.041	0.031[a]	0.012	0.038[a]	0.012
CDD	-0.983	1.28	-0.103	0.11	0.006	0.053
R	-0.515	5.62	-1.577	4.598	-0.003	0.049
TR-Seg	-11.277	1.86E+04	2.2	5.395	-0.895	0.885
LT-Seg	11.106	13.83	2.98	5.579	0.058	0.908
RT-Seg	-150.253	1.40E+35	2.752	6.049	-0.025	1.263
Dispersion			0.712[a]	0.236	0.367[b]	0.164
Model	Poisson		GP-2		GP-2	
RMSE	1		8.888		10.44	
AIC	46.645		296.954		312.847	
BIC	71.684		325.123		341.016	
Mean	0.021		0.45		0.537	
Variance	0.02		0.954		0.947	

3.5.3. IH 20 WB modeling results

The modeling results for IH 20 WB are shown in **Table 3.23. -3.31**. In scenario 1 (**Table 3.23.**), both $AADT_T$, and $AADT_{NT}$ become significant at the 99% C.I. for TNC. Also, $AADT_T$, and $AADT_{NT}$ appear significant at the 99% C.I. for MV crashes in scenario 2. For SV crashes, $AADT_T$, $AADT_{NT}$, and LT-Seg are significant at the 99%, 90% and 90% C.I. showing a decrease in the $AADT_{NT}$ level of significance in comparison to TNC in scenario 1. In scenario 3 for SV crashes (**Table 3.24.**), $AADT_T$, $AADT_{NT}$, and LT-Seg are significant at the 99%, 95%, and 90% C.I. for SV-OBJ crashes. In comparison to scenario 2 SV crashes, the $AADT_{NT}$ level of significance increases by including the crash severity. For SV-OTH, $AADT_T$ appears significant at the 99% C.I. and no explanatory variables are statistically significant for SV-OVT crashes.

The modeling results for scenario 3 MV crashes (**Table 3.25.**), $AADT_T$, and $AADT_{NT}$ become significant at the 99% C.I. for MV-RRND crashes. For MV-SDSW crashes, $AADT_T$, $AADT_{NT}$, and RT-Seg are significant at the 99%, 99% and 95% C.I., which shows an improvement in RT-Seg level of significance in comparison with MV crashes. For MV-STPD crashes, $AADT_T$, and $AADT_{NT}$ become significant at the 95% and 99% C.I. Studying scenario 4 for SV crashes, **Table 3.26.** shows that $AADT_{NT}$ is significant at the 95% C.I. for SV-OBJ-A+K crashes. $AADT_T$, $AADT_{NT}$, CDD, and LT-Seg are statistically significant at the 99%, 90%, 90%, and 95% C.I. for SV-OBJ-B+C crashes. The modeling result for SV-OTH-N crashes shows that $AADT_T$ becomes significant at the 99% C.I. (**Table 3.27.**). For MV-RRND-B+C crashes, $AADT_T$, and $AADT_{NT}$ are significant at the 95% and 99% C.I. (**Table 3.29.**). Both $AADT_T$, and $AADT_{NT}$ are significant at the 99% C.I for MV-RRND-N crashes. $AADT_T$ appears significant at the 95%, 95%, and 99% for MV-SDSW-A+K, MV-SDSW-B+C, and MV-SDSW-N crashes (**Table 3.30.**). $AADT_{NT}$ and RT-Seg are significant at the 99% and 90% C.I. for MV-SDSW-B+C crashes and at the 99% and 95%

for MV-SDSW-N crashes, respectively. No explanatory variables become significant for SV-OBJ-N (**Table 3.26.**), SV-OTH-A+K, SV-OTH-B+C (**Table 3.27.**), SV-OVT-A+K, SV-OVT-B+C, SV-OVT-N (**Table 3.28.**), MV-RRND-A+K (**Table 3.29.**), and MV-STPD-A+K (**Table 3.31.**).

Table 3.23. Modeling results for IH 20 WB - Scenario 1 & 2.

Variables	Scenario 1		Scenario 2			
	TNC		SV		MV	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-1.586[a]	0.486	-0.917[b]	0.397	-1.66[a]	0.442
$AADT_T$	0.21[a]	0.041	0.131[a]	0.029	0.153[a]	0.032
$AADT_{NT}$	0.032[a]	0.005	0.01[c]	0.005	0.036[a]	0.005
CDD	-0.01	0.031	-0.012	0.018	-0.0	0.021
R	0.03	0.035	0.015	0.027	0.033	0.027
TR-Seg	0.183	0.371	0.339	0.304	0.139	0.321
LT-Seg	0.656	0.53	0.902[c]	0.461	0.364	0.416
RT-Seg	-0.655	0.772	-0.171	0.568	-1.08[c]	0.599
Dispersion	0.194[a]	0.02	1.477[a]	0.385	0.669[a]	0.091
Model	GP2		ZINBP		NB	
RMSE	92.244		21.726		63.119	
AIC	1054.879		674.864		967.783	
BIC	1083.048		706.163		995.952	
Mean	9.975		2.339		7.636	
Variance	112.273		7.254		72.73	

Table 3.24. Modeling results for IH 20 WB - Scenario 3 (SV).

Variables	Scenario 3 (SV)					
	SV-OBJ		SV-OTH		SV-OVT	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-1.448[a]	0.449	-2.89[a]	0.898	-3.937[a]	1.425
$AADT_T$	0.141[a]	0.035	0.236[a]	0.076	0.079	0.087
$AADT_{NT}$	0.013[b]	0.005	-0.013	0.013	0.026	0.016
CDD	-0.015	0.024	0.076	0.07	-0.058	0.095
R	0.012	0.034	0.053	0.061	-0.452	2.608
TR-Seg	0.243	0.377	-0.211	0.791	1.618	3.379
LT-Seg	0.81[c]	0.476	-3.651	3.006	1.737	3.56
RT-Seg	0.012	0.683	-95.165	1.40E+20	-91.748	2.18E+20
Dispersion	0.675[a]	0.142				
Model	NB		ZIP		ZIP	
RMSE	18.358		4.899		2.828	
AIC	631.143		171.279		148.565	
BIC	659.312		199.448		176.734	
Mean	1.959		0.178		0.128	
Variance	5.285		0.221		0.137	

Table 3.25. Modeling results for IH 20 WB - Scenario 3 (MV).

Variables	Scenario 3 (MV)					
	MV-RRND		MV-SDSW		MV-STPD	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-2.851[a]	0.601	-1.184[a]	0.433	-3.917[a]	0.922
$AADT_T$	0.145[a]	0.037	0.104[a]	0.028	0.117[b]	0.05
$AADT_{NT}$	0.044[a]	0.006	0.022[a]	0.005	0.043[a]	0.01
CDD	-0.013	0.022	0.02	0.016	0.021	0.028
R	0.043	0.029	0.024	0.023	0.034	0.036
TR-Seg	0.266	0.353	-0.128	0.315	0.033	0.464
LT-Seg	0.547	0.465	-0.095	0.373	0.246	0.587
RT-Seg	-0.752	0.674	-1.548[b]	0.659	-16.097	1169.715
Dispersion	0.698[a]	0.182	0.86[a]	0.147	0.72[a]	0.217
Model	ZINB		GP1		NB	
RMSE	30.919		29.933		10.677	
AIC	753.647		715.175		418.995	
BIC	784.946		743.344		447.164	
Mean	3.521		3.095		0.88	
Variance	19.977		13.041		2.114	

Table 3.26. Modeling results for IH 20 WB - Scenario 4 (SV).

Variables	Scenario 4 (SV)					
	SV-OBJ-A+K		SV-OBJ-B+C		SV-OBJ-N	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-4.165[a]	1.583	-1.762[a]	0.565	1.234[c]	0.68
$AADT_T$	-0.065	0.108	0.118[a]	0.04	-0.038	0.043
$AADT_{NT}$	0.042[b]	0.019	0.012[c]	0.007	-0.002	0.006
CDD	-0.032	0.12	-0.055[c]	0.032	0.007	0.017
R	-0.521	3.847	-0.045	0.058	-0.006	0.026
TR-Seg	1.352	4.878	0.555	0.405	-0.12	0.366
LT-Seg	1.182	5.124	1.302[b]	0.554	0.44	0.361
RT-Seg	-15.121	4591.077	0.507	0.827	-0.086	0.616
Dispersion	0.116	0.107				
Model	GP1		ZIP		Hurdle	
RMSE	3		7.616		13.191	
AIC	134.564		392.757		494.651	
BIC	162.733		420.926		544.73	
Mean	0.112		0.661		1.186	
Variance	0.133		1.022		2.293	

Table 3.27. Modeling results for IH 20 WB - Scenario 4 (SV).

Variables	Scenario 4 (SV)					
	SV-OTH-A+K		SV-OTH-B+C		SV-OTH-N	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-7.452	6.405	-8.612	8.795	-3.295[a]	0.926
$AADT_T$	1.201	1.61	-0.216	0.471	0.259[a]	0.081
$AADT_{NT}$	-0.187	0.261	0.094	0.103	-0.018	0.014
CDD	-0.602	2.85E+05	-2.476	1.70E+07	0.072	0.063
R	-3.295	5.25E+06	-2.799	6.48E+08	0.071	0.06
TR-Seg	-17.834	5.86E+06	-16.908	8.41E+08	-0.061	0.784
LT-Seg	-16.232	5.21E+06	-3.733	8.91E+08	-3.149	2.628
RT-Seg	-6.89	5.18E+06	-3.96	9.02E+08	-45.756	2.86E+09
Dispersion						
Model	Poisson		Poisson		Poisson	
RMSE	0		1		4.899	
AIC	33.307		34.233		152.841	
BIC	58.346		59.272		177.88	
Mean	0.008		0.012		0.157	
Variance	0.008		0.012		0.191	

Table 3.28. Modeling results for IH 20 WB - Scenario 4 (SV).

Variables	Scenario 4 (SV)					
	SV-OVT-A+K		SV-OVT-B+C		SV-OVT-N	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-4.999	4.401	-5.254[b]	2.357	-5.31[b]	2.114
$AADT_T$	0.157	0.258	0.028	0.147	0.101	0.12
$AADT_{NT}$	0.021	0.051	0.031	0.027	0.027	0.024
CDD	0.083	0.34	-0.002	0.1	-0.237	0.204
R	-0.8	5.218	-0.259	1.323	-0.219	0.834
TR-Seg	2.058	6.795	1.027	2.124	1.921	1.765
LT-Seg	-3.855	15.243	0.617	2.59	3.817	2.541
RT-Seg	-1.924	14.321	-113.635	1.13E+25	-30.387	3.58E+07
Dispersion						
Model	ZIP		Poisson		Poisson	
RMSE	1		2		1.732	
AIC	45.162		77.053		96.265	
BIC	73.331		102.092		121.304	
Mean	0.017		0.05		0.062	
Variance	0.016		0.047		0.058	

Table 3.29. Modeling results for IH 20 WB - Scenario 4 (MV).

Variables	Scenario 4 (MV)					
	MV-RRND-A+K		MV-RRND-B+C		MV-RRND-N	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-3.603[a]	1.186	-3.14[a]	0.718	-3.727[a]	0.739
$AADT_T$	-0.005	0.091	0.11[b]	0.043	0.157[a]	0.042
$AADT_{NT}$	0.023	0.015	0.036[a]	0.008	0.049[a]	0.008
CDD	-0.105	0.181	-0.008	0.026	-0.012	0.023
R	-2.263	7.315	0.046	0.031	0.044	0.031
TR-Seg	4.259	8.949	0.352	0.411	0.147	0.389
LT-Seg	3.415	9.361	0.623	0.517	0.432	0.495
RT-Seg	-12.285	5491.828	-0.985	0.94	-0.707	0.728
Dispersion			0.614[a]	0.177	0.742[a]	0.257
Model	Poisson		NB		ZINB	
RMSE	3.464		11.874		20.494	
AIC	123.708		481.481		639.202	
BIC	148.747		509.65		670.501	
Mean	0.116		1.182		2.223	
Variance	0.111		2.614		9.718	

Table 3.30. Modeling results for IH 20 WB - Scenario 4 (MV).

Variables	Scenario 4 (MV)					
	MV-SDSW-A+K		MV-SDSW-B+C		MV-SDSW-N	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-7.396[b]	2.985	-2.315[a]	0.663	-1.397[a]	0.463
$AADT_T$	0.328[b]	0.155	0.093[b]	0.047	0.089[a]	0.03
$AADT_{NT}$	0.02	0.035	0.023[a]	0.008	0.022[a]	0.006
CDD	-0.111	1.76E+07	0.032	0.023	0.018	0.017
R	-2.034	2.14E+07	-0.045	0.055	0.034	0.023
TR-Seg	-33.936	1.75E+08	0.112	0.415	-0.349	0.379
LT-Seg	-33.428	2.95E+08	-0.702	0.615	0.019	0.399
RT-Seg	-32.226	2.93E+08	-2.035[c]	1.117	-1.341[b]	0.676
Dispersion			0.622	0.394	1.623[a]	0.348
Model	ZIP		ZINBP		NBP	
RMSE	2.45		10.392		22.158	
AIC	72.301		420.76		630.482	
BIC	100.47		452.059		658.652	
Mean	0.05		0.806		2.24	
Variance	0.056		1.56		7.735	

Table 3.31. Modeling results for IH 20 WB - Scenario 4 (MV).

Variables	Scenario 4 (MV)					
	MV-STPD-A+K		MV-STPD-B+C		MV-STPD-N	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-6.116[c]	3.413	-6.53[a]	1.809	-3.073[a]	0.878
$AADT_T$	0.09	0.212	0.192[b]	0.086	0.069	0.05
$AADT_{NT}$	0.025	0.038	0.056[a]	0.019	0.036[a]	0.01
CDD	0.323	0.6	0.017	0.04	0.015	0.02
R	-3.127	7.466	0.054	0.051	0.041	0.044
TR-Seg	2.865	8.971	0.094	0.678	-0.218	0.488
LT-Seg	-31.023	2.40E+04	0.15	0.896	0.196	0.508
RT-Seg	-26.124	8.53E+05	-21.21	2.80E+04	-34.632	1.57E+07
Dispersion			0.583[b]	0.249	N/A	
Model	Poisson		GP2		ZIP	
RMSE	410.992		5.745		7.483	
AIC	49.993		249.596		331.697	
BIC	75.033		277.765		359.866	
Mean	0.017		0.343		0.521	
Variance	0.016		0.591		0.881	

3.5.4. Model performance comparison

As discussed in section 4.8, the modeling process includes steps to identify the best performing model for each crash group. Vuong's test is used to compare non-nested models and investigates base models versus zero-inflated models. Afterwards, the study compares the models selected by Vuong's test with the GP-1, GP-2, and Hurdle regression models using AIC to identify the outperforming model for each crash group. **Fig. 3.3.** shows the AIC values for the outperforming model for each crash group. As shown in **Fig. 3.3.**, the AIC values for tri-dimensional crash groups (using all three traffic crash characteristics) are smaller than the AIC values for their parent crash group using two traffic crash characteristics (crash units and crash type), which are smaller than the AIC values for their corresponding grandparent crash group using only one traffic characteristic (crash unit) that are smaller than AIC value for TNC crash group. This implies that the tri-dimensional analysis improves the model performance.

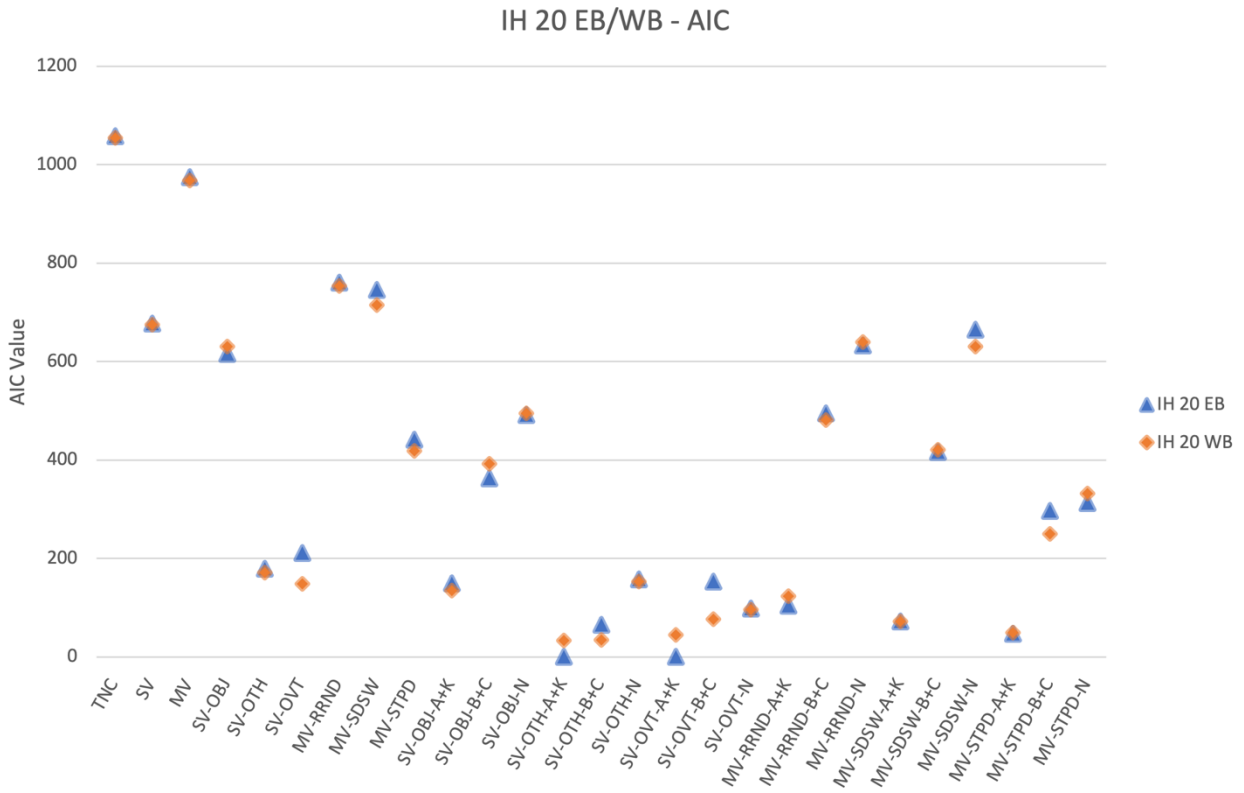


Fig. 3.3. Model performance by AIC for all crash groups.

3.5.5. Hotspot identification results

PSI identifies hotspots using the selected crash prediction models (**Table 3.15. -3.22.** for IH 20 EB and **Table 3.23.-3.31.** for IH 20 WB) for each crash group within each scenario. The study compares the hotspot identification results by looking at the top ten most hazardous hotspots (top ten largest PSI values). The segment ID for the top ten hotspots for each group of crashes are shown in **Tables 3.32.-3.34.** and **Tables 3.35.-3.37.** for IH 20 EB and IH 20 WB, respectively. Subsections 5.2.1. and 4.5.2.2 present the hotspot identification results of each scenario for IH 20 EB and IH 20 WB.

3.5.6. IH 20 EB hotspot results

The segment ID for the top 10 hotspots along IH 20 EB are shown in **Tables 3.32.-3.34**. No hotspot can be identified for SV-OTH, SV-OVT, SV-OBJ-A+K, SV-OBJ-B+C, SV-OTH-B+C, SV-

OTH-N, SV-OVT-B+C, and SV-OVT-N crashes because their corresponding prediction models are not significant. As shown in **Tables 3.32.**, there are some hotspot segments that are common between scenario 1 and scenario 2. Comparing the top ten hotspots for TNC with SV and MV crashes, there are 70% and 80% of the hotspot segments for SV and MV crashes in common with TNC hotspot segments, respectively. Also, 50% of the hotspot segments for SV and MV crashes are in common. Comparing MV crashes with TNC, the top six hotspot segment are the same with a slight difference for segments #164 and #120, switching places in the rankings for MV crashes. Studying scenario 3 for SV crashes (**Tables 3.33.**), There are only two segments, #111 and #176, that are hotspot segments for SV-OBJ crashes and not for SV crashes in scenario 2 but segment # 111 is a hotspot for scenario 1, TNC. In scenario 3 (SV), **Table 3.33.** shows that the first ten segments of IH 20 EB as hotspots for SV-OTH and SV-OVT crashes. As shown in **Table 3.33.,** segment #175 and #22 for MV-RRND crashes are not appeared as hotspot segments in MV crashes in scenario 2. For MV-SDSW and MV-STPD crashes show 50% and 40% hotspot segments in common with hotspot segments for MV crashes in segment 2. Segments #0 and #130 are hotspots segments for SV-OBJ-N crashes that are not identified as hotspots for TNC, SV, and SV-OBJ crashes. In scenario 4 (SV), segments #44 and #69 for MV-RRND-B+C crashes and segment #22 and #117 for MV-RRND-N crashes are hotspots that do not appear in the top ten hotspots for scenario 1, 2, and 3. As shown in **Table 3.34.,** 60% and 50% of the hotspot segments for MV-SDSW-B+C and MV-SDSW-N crashes are identified in the top ten only for scenario 4. Similarly, for MV-STPD-B+C and MV-STPD-N crashes, 60% and 40% hotspot segments have their first appearance as the top ten hotspot segments for scenario 4.

Table 3.32. Top 10 IH 20 EB hotspots - Scenario 1 & 2.

	Scenario 1	Scenario 2	
Hotspots Rank	**TNC**	**SV**	**MV**
1	110	175	110
2	164	164	120
3	120	179	164
4	88	84	88
5	7	120	7
6	111	7	111
7	175	88	109
8	179	110	99
9	109	30	112
10	99	69	119
PSI_{1st}	65	10	59
PSI_{10th}	17	4	12

Table 3.33. Top 10 IH 20 EB hotspots - Scenario 3 (SV) & (MV).

	Scenario 3 (SV)	Scenario 3 (MV)		
Hotspots Rank	**SV-OBJ**	**MV-RRND**	**MV-SDSW**	**MV-STPD**
1	175	110	164	110
2	84	120	88	164
3	88	111	99	120
4	120	7	110	3
5	164	88	7	7
6	179	109	44	17
7	7	112	84	19
8	30	164	79	31
9	111	175	176	69
10	176	22	179	72
PSI_{1st}	7	47	17	4
PSI_{10th}	4	5	8	1

Table 3.34. Top 10 IH 20 EB hotspots - Scenario 4 (SV) & (MV).

	Scenario 4 (SV)	Scenario 4 (MV)					
Hotspots Rank	**SV-OBJ-N**	**MV-RRND-B+C**	**MV-RRND-N**	**MV-SDSW-B+C**	**MV-SDSW-N**	**MV-STPD-B+C**	**MV-STPD-N**
1	175	110	110	88	164	110	17
2	84	120	120	96	88	120	19
3	120	175	164	99	44	88	69
4	164	7	7	110	84	106	99
5	179	44	88	156	99	111	110
6	0	69	109	165	179	115	119
7	7	88	111	7	7	164	120
8	88	109	112	42	79	0	0
9	130	111	22	45	120	1	1
10	176	112	117	57	4	2	2
PSI_{1st}	4	6	34	2	12	3	1
PSI_{10th}	2	2	5	1	5	0	0

3.5.7. IH 20 WB hotspot results

Tables 3.35.-3.37. show the hotspot results for IH 20 WB for the crash groups in each scenario. Since the prediction model is found significant for SV-OTH, SV-OVT, SV-OBJ-A+K, SV-OBJ-B+C, SV-OTH-A+K, SV-OTH-B+C, SV-OTH-N, SV-OVT-B+C, SV-OVT-N, MV-RRND-A+K, MV-SDSW-A+K, MV-STPD-A+K, and MV-STPD-N crashes, no hotspot can be identified using the empirical bayes methods. As shown in **Table 3.35.**, 40% and 80% of the hotspot segments for SV and MV crashes are in common with the TNC hotspot segments. Except for segments #110, #130, and #164, the rank of hotspot segments for TNC and MV crashes are the same. In scenario 3 (SV) (**Table 3.36.**), segments #176 and #84 are identified as top ten hotspots for SV-OBJ crashes that do not appear as the top ten hotspots for TNC and SV crashes. In **Table 3.36.** for scenario 3 (MV) crash groups, 70%, 60%, and 40% of the top ten hotspot segments are previously appeared in scenario 1 or 2. Similarly, in scenario 4 (MV), there are 50%, 40%, 50%, and 30% of the top ten hotspot segments for MV-RRND-B+C, MV-RRND-N, MV-SDSW-B+C, and MV-SDSW-N that are not identified as the top ten hotspots for scenarios 1, 2, and 3. For MV-STPD-B+C crashes, only segments #130, #88, and # 129 are in common with MV or MV-STPD hotspot segments (**Table 3.37.**).

Table 3.35. Top 10 IH 20 WB hotspots - Scenario 1 & 2.

	Scenario 1	Scenario 2	
Hotspot Rank	**TNC**	**SV**	**MV**
1	88	110	88
2	110	88	130
3	130	146	164
4	164	234	110
5	31	84	31
6	7	164	7
7	0	3	0
8	44	7	44
9	45	30	156
10	2	99	211
PSI_{1st}	36	10	34
PSI_{10th}	11	4	13

Table 3.36. Top 10 IH 20 WB hotspots - Scenario 3 (SV) & (MV).

	Scenario 3 (SV)	Scenario 3 (MV)		
Hotspot Rank	**SV-OBJ**	**MV-RRND**	**MV-SDSW**	**MV-STPD**
1	88	130	88	130
2	110	88	164	30
3	146	110	211	129
4	176	31	7	0
5	84	164	110	2
6	164	19	120	88
7	234	44	155	4
8	3	131	156	19
9	7	7	31	20
10	31	21	33	22
PSI_{1st}	6	20	13	4
PSI_{10th}	3	6	6	1

Table 3.37. Top 10 IH 20 WB hotspots - Scenario 4 (MV).

	Scenario 4 (MV)				
Hotspot Rank	**MV-RRND-B+C**	**MV-RRND-N**	**MV-SDSW-B+C**	**MV-SDSW-N**	**MV-STPD-B+C**
1	130	130	211	88	130
2	63	110	92	164	88
3	88	88	120	31	129
4	128	31	159	7	135
5	3	129	0	110	238
6	19	145	2	155	0
7	21	0	7	156	1
8	98	131	14	33	2
9	164	22	33	45	3
10	223	44	88	176	4
PSI_{1st}	5	12	3	12	2
PSI_{10th}	2	4	1	5	0

3.6. DISCUSSION

The findings of this study show consistency with previous research, which indicated significant variations in the contributing factors of crash groups of each scenario. The modeling results indicate that traffic operation characteristics, $AADT_T$ and $AADT_{NT}$, significantly impact TNC (scenario 1), SV, and MV crash occurrences (scenario 2) for both IH 20 EB & WB. This implies that higher traffic volume negatively impacts both SV and MV crashes. This finding is consistent with the results from previous research on SV crashes (Persaud & Mucsi, 1995; Geedipally & Lord, 2010) and MV crashes (Persaud & Mucsi, 1995; Geedipally & Lord, 2010; Yu & Abdel-Aty, 2013). However, some previous studies conclude that traffic volume has no significant impact on SV crash occurrences, which this study confirms for some cases (Yu & Abdel-Aty, 2013; Wang & Feng, 2019). For example, in scenario 3, $AADT_{NT}$ shows no significant impact on SV-OTH and SV-OVT crashes for IH 20 EB/IH 20 WB and $AADT_T$ reveals no significant impact on MV-STPD and SV-OVT crashes for IH 20 EB and IH 20 WB, respectively. As shown in scenario 4, $AADT_T$ and $AADT_{NT}$ ae not significant contributing factors for some crash groups with certain crash

severity, despite their significant impact on the corresponding crash supergroup in scenario 3, or vice versa. $AADT_T$ has a greater magnitude effect on SV-OVT-N and SV-OVT-B+C crashes for IH 20 EB, in a descending order. Meanwhile for IH 20 WB, MV-SDSW-A+K, SV-OTH-N, and MV-STPD-B+C crashes are impacted the most by traffic volume of trucks, $AADT_T$.

The modeling results proclaim that some of the geometric characteristics, the horizontal curve delta angle, CDD, appear as a significant contributing factor for SV, SV-OBJ, SV-OBJ-A+K, SV-OBJ-B+C, SV-OBJ-N, MV-SDSW, MV-SDSW-B+C and MV-SDSW-N for IH 20 EB. It implies that the greater the horizontal curve delta angle, the greater the likelihood of crashes for the listed crash groups. While for IH 20 WB, the horizontal curve delta angle, CDD, become a significant contributing factor for SV-OBJ-B+C crashes but with an opposite effect; the greater the horizontal curve delta angle, the lower the likelihood of crashes. This noticeable difference between IH 20 EB and WB might be justified if another potential explanatory variables such as sunlight glare, and pavement marking visibility appear as significant contributing factors for the crash groups of interest.

Curve radius, R, is another geometric characteristic that is a significant contributing factor for SV-OTH, SV-OBJ-N, SV-OTH-B+C, SV-OTH-N, and MV-SDSW-B+C crashes for IH 20 EB. However, the curve radius is not significantly contributing to any crash groups for IH 20 WB. These results suggest that a larger curve radius contributed to a higher number of the named crash groups for IH 20 EB. This finding is inconsistent with using a larger radius to provide smoother transitions between tangent segments.

As a part of geometric characteristics, the geometric type of the segment is included by introducing binary variables *TR-Seg*, *LT-Seg*, and *RT-Seg* as defined in **Table 3.13**. The results show that the likelihood of crash occurrence on curved-to-right segments is smaller for SV, SV-

OBJ, and SV-OBJ-N crashes across IH 20 EB, and MV, MV-SDSW-B+C, and MV-SDSW-N crashes along IH 20 WB. The curved-to-left segment, *LT-Seg,* negatively impacts on SV, SV-OBJ, and SV-OBJ-B+C crashes for IH 20 WB increasing the crash likelihood for these crash groups.

Considering the HSID results, the hotspots for each crash group in each scenario show different locations, nearly half of the top ten hotspots. Therefore, hotspot segments among the top ten hotspots are mutually shared for traffic crash groups within and across scenarios. The study result indicates that TNC and MV crashes demonstrate higher consistency in hotspot segments than SV crashes. This finding is consistent with previous research and confirm their finding (Wang & Feng, 2019). The HSID results show that the top ten hotspot segments are substantially different when all three crash characteristics are included. This study excluded spatial correlation between the segments that may impact the hotspot segments mutually shared between various crash groups. Investigating the spatial correlation combined with the regression models used in this study is for future research.

3.7. CONCLUSIONS

This study aims to investigate the traffic crash contributing factors, including traffic operation characteristics with geometric characteristics and their impact on different crash groups. Four scenarios are defined to form crash groups: Scenario 1 includes all crashes combined; For scenario 2, crash unit dimension is considered to form crash groups, SV and MV crashes; In scenario 3, crash unit and crash type dimensions are used to group the crashes; Finally, all three crash characteristics, crash unit, crash type, and crash severity, are considered to classify crashes. The traffic crash data, traffic operation, and geometric characteristics data are collected for both directions of a 24.26 mile of IH 20 within the Dallas County limit. An extensive group of count data regression models, including Poisson, NB, NBP, ZIP, ZINB, ZINBP, GP-1, GP-2, and Hurdle regression models, are utilized to perform prediction models for all the crash groups defined. As

expected, the modeling results show that the outperforming models differ for each group of crashes same as the dispersion magnitudes. According to the results, the contributing factors, such as truck AADT, non-truck AADT, horizontal curve delta degree, horizontal curve radius, curved-to-right segment, and curved-to-left segment, became significant contributing factors for some crash groups. At the same time, they are not statistically significant for other crash groups, especially compared to total crashes, SV, and MV crashes. As shown in previous research, it is also shown that the top ten hotspot segments for each crash group vary for nearly half segments. However, it is confirmed that MV crashes hotspot segments are more representative and aligned with total crash hotspot segments. This study concludes that scenario 4, which includes all three traffic crash characteristics, crash unit, crash type, and crash severity, provides a better understanding and a clearer vision of contributing factors and hotspot segments. It will assist in proposing appropriate measures to mitigate specific crash groups at a specific location on a roadway. However, some disadvantages associated with scenario 4 exist. As shown, by including all three dimensions, some crash groups disappear from the modeling due to an insufficient number of occurrences, making regression model estimation infeasible. Therefore, the crash groups with very limited observations must be excluded from the analysis. The crash groups with very limited observations may represent outliers in their parent crash groups from scenarios 1, 2, and 3, but low crash counts do not reduce the importance of a crash group in all cases. Also, traffic safety analysis using all three dimensions may suffer from the availability of quality data and the workload associated with adding more dimensions.

This study came with limitations due to the availability of quality data on traffic operation and geometric characteristics. Future research needs to be conducted using more accurate data, including potential explanatory variables such as operating speed, pavement marking visibility,

sunlight glare, cross-slope, shoulder width, and lane width. As noted, the spatial correlation was excluded from this study, and it is suggested that future studies investigate the spatial correlation combined with, but not limited to, the selected regression models in this study.

CHAPTER 4. INVESTIGATING THE IMPACT OF RECOMMENDED FRAGMENT SIZE TO IMPROVE CRASH COUNT PREDICTION MODELS

4.1. INTRODUCTION

In recent years, identifying traffic crash hotspots has become crucial for specifying hazardous locations, prioritizing effective countermeasures, and enhancing road safety. Traffic crash hotspots refer to areas where the frequency or likelihood of crashes is significantly higher than neighboring locations along a targeted corridor or across a network. Hotspot identification (HSID) supports providing focused interventions to mitigate traffic crash likelihood and ameliorate safety in hazardous areas by understanding the contributing factors to traffic crashes. Often, traffic hotpot studies use crash frequency analysis (CFA), which divides a highway into small fragments (segments) with constant length for data aggregation. Previous studies highlight that the selection of the fragment size (segment length) may affect the crash frequency model estimation, the accuracy of CFA-based HSID, and eventually the effectiveness of interventions to mitigate traffic crash likelihood. Therefore, finding the appropriate fragment size for data aggregation appears vital due to its ripple effect throughout traffic safety studies.

Hotspot identification encompasses many approaches that can be classified into three main categories: Geographic Information Systems (GIS)-based spatial analysis, statistical models, and machine learning. Traditionally, GIS-based HSID involves the creation of crash concentration maps, which rely on Kernel Density Estimation (KDE) and absolute crash counts to determine the density of crashes in specific regions (Truong & Somenahalli, 2011). However, the accuracy of concentration maps can be challenged due to the selected search bandwidth for KDE and reliance on absolute crash counts (Truong & Somenahalli, 2011). Other studies apply regression models to predict crash frequencies, investigate factors associated with crashes, and identify traffic crash

hotspots‘ these count data regression models include Poisson, Negative Binomial, Poisson Lognormal, Zero-Inflated Poisson/Negative Binomial, Gamma, Generalized Estimating Equation, Negative Multinomial, and Hurdle models (Hilbe, 2014). These regression models aim to forecast crash frequencies based on features such as roadway geometry, traffic characteristics, and weather conditions (Highway Safety, 2005) and identify hotspots based on the estimated crash frequencies. Recent studies h employ machine learning methods, such as random forest, decision tree, support vector machine (SVM), Naive Bayes, and neural network algorithms, to investigate traffic safety and identify traffic crash hotspots (Santos et al., 2022). This study develops a crash frequency analysis by estimating crash count regression models including Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), generalized Poisson type 1 (GP-1), generalized Poisson type 2 (GP-2), Hurdle regression.

In crash prediction models, the data aggregation, including traffic crash data, traffic characteristics data, and geometric characteristics data, plays a vital role (Wang & Feng, 2019). Traditionally, a constant length referred to as the "segment length" divides highways or freeways into smaller sections. However, to prevent confusion with the geometric features of the road, this study chooses to utilize the term "fragment size" instead. Previous research often employs arbitrary values for segment length and typically range from 0.1 mile to 1.0 mile or sometimes 100 meters to 1.6 kilometers, based on their specific research objectives (Texas Department of Transportation (TxDOT) - Traffic Safety Division, 2020). Nevertheless, using different segment lengths for data aggregation can result in certain variables being either statistically significant or insignificant (Ahmed & Abdel-Aty, 2012). Earlier studies highlight the inherent problems in crash frequency analysis when utilizing fixed-size fragments (segments) to aggregate crash data (Pedregosa, et al., 2011). However, until recently, no researchers identify a specific methodology to determine the

suitable fragment size. Given the potential impact on the statistical significance of variables, the selection of an appropriate fragment size (segment length) for data aggregation appears essential.

The fragment size selection is vital to understanding the significance of changes in the crash prediction model results, the resulting hotspots, and the effectiveness of the remedies to improve traffic safety. Due to the ripple effect of the fragment size in traffic safety studies, this study investigates the impact of fragment size on crash prediction models and inspects the potential advantages of the RFS to produce valid models and improve crash prediction model performance and accuracy.

4.2. LITERATURE REVIEW

An important part of traffic safety studies focuses on traffic crash hotspot identification (HSID) using crash count prediction models performed by dividing a highway into small fragments (segments) for data aggregation. The arbitrary selection of fragment size (segment length) may adversely impact the crash prediction models results, but no previous research clearly illustrates the magnitude of these impacts or a strategy for choosing a fragment size. Effective hotspot identification and selection of remedies to reduce traffic safety hazards rely on crash prediction model validity. Therefore, researchers face a crucial challenge when identifying the appropriate fragment size for data aggregation due to its chain impact throughout a traffic safety study. Recent research (Maniei & Mattingly, 2023a) provide an innovative approach to find a recommended fragment size (RFS) for data aggregation, but the modeling performance using the RFS requires evaluation.

4.2.1. Crash prediction models

Many studies adopted count data models for crash prediction. The Poisson regression model became popular among count data models assuming the mean and the variance of the data are

equal. However, crash data often exhibited over-dispersion, where the variance exceeded the mean. To address this issue, researchers turned to negative binomial (NB) regression models (Abdel-Aty & Radwan, 2000; Miaou, 1994). The commonly employed models for count data modeling are NB-1 and NB-2 with different assumptions for variance structure. Since these restricted variance structures could lead to biased model parameter estimates and inaccurate crash predictions, Greene (2008) introduced a new NB regression model called the NB-P encompassing to the NB-1 and NB-2 when $P = 1$ and $P = 2$, respectively. This model provided better fit and estimation accuracy due to its flexible variance structure. The researchers concluded that the NB-P model's flexible variance structure significantly improved estimation accuracy (Wang, et al., 2020). Due to excessive zeros for no-crash areas, crash count analysis needs to handle the excessive zeros that traditional Poisson and NB models cannot handle. To address this issue, many investigators employ zero-inflated models (Carson & Mannering, 2001; Qin, et al., 2005). Previous crash frequency analysis studies frequently utilized the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models to deal with the problem of excessive zeros (Lee & Mannering, 2002; Chin & Quddus, 2003) and show a statistically better fit to the data (Malyshkina & Mannering, 2010). However, the assumption that roadway segments were intrinsically safe seems doubtful even for well-designed roadway segments due to unsafe driver behavior. Therefore, the fundamental assumption of the zero-inflated model appeared flawed (Lord, et al., 2005). An alternative approach was the Hurdle model, also known as the two-part model, to handle excessive zeros in the dataset (Ma, et al., 2016). Unlike the zero-inflated models, the Hurdle model did not assume that roadway segments with zero crashes observed during the study period were inherently safe, but rather that they were safe only during that specific period.

This study aims to identify the impact of the fragment size on crash prediction models by investigating many count regression models including Poisson, NB, NBP, ZIP, ZINB, ZINBP, Consul's Generalized Poisson (GP-1), Famoye's Generalized Poison (GP-2), and Hurdle regression. Since the study considers various crash groups, these models must represent crash groups with either negligible dispersion or significant dispersion. Due to the study's structure, more crash groups with an excessive number of zeros appear as the study adds more traffic crash dimensions when forming the crash groups, which makes the need to include zero-inflated count regression models inevitable.

4.2.2. Crash prediction models and traffic crash dimensions

Three essential characteristics of traffic crashes exist: number of vehicle involved in a traffic crash (crash units), manner of collision (crash type), and crash severity. The number of vehicles involved in a traffic crash, also known as "crash units", is considered a critical characteristic, typically classified as either single-vehicle (SV) or multi-vehicle (MV). Previous research shows that crash units play a determining role in predicting traffic crashes, identifying significant contributing factors (Ivan, et al., 1999; Abdel-Aty, et al., 2006), (Yu, et al., 2013) and their associated impacts (Yu & Abdel-Aty, 2013; Dong, et al., 2018). The previous work demonstrates the need to separate the analyses of single-vehicle (SV) and multi-vehicle (MV) crashes due to their distinct spatial distribution and contributing factors (Wang & Feng, 2019). This distinction holds when adopting aggregate and disaggregate approaches to studying traffic crashes (Yu & Abdel-Aty, 2013). Earlier studies propose utilizing separate crash prediction models (Ivan, et al., 1999; Lord, et al., 2005; Geedipally & Lord, 2010; Ma, et al., 2016). Wang and Feng (2019) observed significant differences in both the traffic crash contributing factors when comparing the results of total crashes with those of single-vehicle (SV) and multi-vehicle (MV) crashes.

The manner of collision or crash type represents another crucial traffic crash characteristic for traffic crash analysis. Crash type pertains to the initial event or collision during a traffic crash, and the differences in the types holds significant importance in traffic crash analysis (Pande & Abdel-Aty, 2006). Several empirical studies show that traffic crashes exhibit distinct characteristics based on their crash type, regardless of the level of aggregation (Golob, et al., 2008). Despite the importance of considering crash types in traffic crash studies, limited research comprehensively investigates traffic crash contributing factors based on crash types. This study addresses this gap by incorporating crash type as a critical characteristic for predicting traffic crashes.

The crash severity, which specifies the magnitude of damage and injuries caused, represents another vital traffic crash characteristic (Xu, et al., 2013). Crash severity ranges from minor property damage to severe injuries or fatalities. The development of crash prediction models that consider different levels of crash severity yields valuable insights into reducing the likelihood of severe crashes (Xu, et al., 2013). Examining crash severity is vital when studying single-vehicle (SV) crashes and their contributing factors (Jung, et al., 2010). Several studies investigate crash severity crash hotspots using negative binomial and Bayesian spatial statistical methods (Mitra, 2009), multivariate crash count models, equivalent property damage only (EPDO), and two-stage models (Afghari, et al., 2020). Afghari et al. (2020) argue that traditional approaches fail to consider the unobserved heterogeneity associated with the correlations between crash counts for each severity level. In fact, a critical part of traffic safety studies is unobserved heterogeneity. Studies can only include some information to capture data for all potentially contributing causes of traffic crashes (Chang, Yasmin, Huang, & Chan, 2021; Mannering, Shankar, & Bhat, 2016). To address unobserved heterogeneity, investigators often categorize the traffic crash data into homogeneous crash groups using different attributes (Mannering & Bhat, 2014). To overcome this

issue, the present study considers crash severity, the number of vehicles involved (crash units), and the type of collision (crash types) by creating four different scenarios for crash dimensionality.

4.2.3. Crash prediction models and data aggregation

Crash prediction models typically divide corridors into small segments using a fixed segment length and aggregating the crash data and other geometric and operational data.

According to Green (2018), many different approaches to segmenting a roadway using a subset of data sources, such as traffic data, roadway characteristics, and traffic crash data, exist. One standard method involves segmenting a roadway based on its geometric characteristics to account for unobserved heterogeneity; however, this approach may result in long segments with little variation in roadway attributes over a considerable distance. For instance, Green (2018) points out that a highway segment can become very long if it features a straight section with constant shoulder width, number of lanes, cross slope, and median width. Furthermore, the limited availability of quality roadway characteristics data may necessitate expensive data collection efforts. In situations with insignificant variations in roadway attributes, Borsos et al. (2014) suggest using traffic data to create homogeneous segments. While this approach can help divide long segments into smaller ones, Green (2018) notes that it may not be effective for roadways with limited access over a long distance, where minor changes in traffic volume occur.

Other alternatives to roadway segmentation by roadway attributes include continuous risk profile (Kwon, et al., 2013), sliding moving window (Qin & Wellner, 2012; Kwon, et al., 2013), peak searching (Kwon, et al., 2013), fixed length and variable length segmentation (Koorey, 2009), and clustering methods (Valent, et al., 2002; Depaire, et al., 2008; Lu, et al., 2013). The use of clustering techniques has proven beneficial for segmenting roadways based on traffic crash data, particularly in cases where high-quality data on traffic and roadway attributes is unavailable. These

techniques have the potential to reveal previously unknown relationships within the crash data (Golob, et al., 2004a; Depaire, et al., 2008; De Luca, et al., 2012; Lu, et al., 2013). Valent et al. (2002) employ a clustering method focused on a specific crash type to analyze traffic crashes but caution that this approach could obscure the underlying contributing factors associated with that particular crash type. Depaire et al. (2008) utilize latent class clustering to segment roadways leveraging the heterogeneity of traffic crash data. Similarly, Lu et al. (2013) employ Fisher's clustering technique to create a segmentation based on sections exhibiting similar crash distributions. This application of Fisher's clustering result in an improvement in the performance of predictive models. Maniei and Mattingly (2023a) apply the Laplacian score with distance-based entropy measure (LSDBEM) and K-meaning clustering to recommend a segment length for roadway segmentation, called recommended fragment size (RFS).

Given the lack of high-quality data on roadway attributes, the current study examines if the crash prediction model results improve using RFS for data aggregation as proposed by Maniei and Mattingly (2023a). For the appropriate aggregation of data on urban/suburban highways and freeways, previous studies recommend not using a segment length smaller than 0.1 mile (American Association of State Highway and Transportation Officials, 2010) or a spacing interval larger than 0.25 mile for traffic operational characteristics (Alabama Department of Transportation, 2015). The study estimates and evaluates crash prediction models using data aggregation fragment sizes ranging from 0.10 mile to 0.25 mile with an increment of 0.01 mile as recommended in the earlier studies. To address unobserved heterogeneity, the investigation develops crash prediction models for crash groups formed by four scenarios using crash units, crash types, and crash severity.

4.3. DATA DESCRIPTION

This study evaluates the impact of fragment size using both directions of mainlane IH-20 segments located in Dallas. The Texas Department of Transportation (TxDOT) Crash Record Information System (C.R.I.S.) data used for this analysis covers the period from 2015 to 2019. This dataset encompasses a wide range of information and includes details on crashes, roadway geometry, and traffic characteristics.

4.3.1. Crash Data Features

The TxDOT C.R.I.S. data consists of features categorized into three groups: crash fields, unit fields, and person fields. The crash fields provide data such as latitude, longitude, reference marker, offset distance, highway system, roadway part, highway name, manner of collision, crash severity, and geometric design features like curve type, curve degree (curvature), curve length, curve delta degree, left shoulder type, left shoulder use, left shoulder width, right shoulder type, right shoulder use, right shoulder width, median type, median, number of lanes, roadbed width, surface condition, surface type, and surface width. Additionally, the data also includes traffic characteristics such as adjusted average daily traffic levels, percentages of single-unit trucks and combo trucks, adjusted percentage of average daily traffic attributed to trucks, and speed limits.

4.3.2. Data Preparation

For this study, only crashes occurring on main roadway segments are considered. Crashes involving work zones, pedestrians, or wrong-way driving are excluded. Feature engineering techniques are applied to filter the crash data specifically associated with the main segments of each roadway while disregarding data related to pedestrians, active work zones, construction areas, and wrong-way driving. To ensure the reliability and consistency of the data, the crash data points are subjected to geovalidation using KMZ files imported to Google Earth®, verifying the accuracy of feature values for roadway segments, vehicle travel directions, and the geometric design

features. All feature exhibiting inconsistencies with actual measurements are excluded from the study such as left and right shoulder width, median width, number of lanes, surface type, and surface width.

The Traffic Safety Division of the Texas Department of Transportation (TxDOT) has established categories for different levels of crash severity. These categories are denoted by letters, with A representing suspected serious injury, B for suspected minor injury, C for possible injury, K for fatal injury, N for not injured, and 99 for unknown (2020). The definitions for these severity categories can be found in **Table 4.1**. The crash data summary is provided in **Table 4.2**, including the percentage range of traffic crash severity for various corridors. The data reveals that fatal crashes occur at a relatively low percentage, suggesting that they may not serve as a significant factor in differentiating roadway segments or forming clusters. To address this, the study combines fatal and suspected serious injury crashes into one group, while suspected minor and possible injury crashes are also merged. Non-injury crashes are treated as a separate characteristic, and crashes with unknown severity are excluded from the analysis.

Traffic crash groups and their abbreviations are shown in **Fig. 4.1**. In this study, four scenarios are considered to form crash groups as shown (orange boxes) in **Fig. 4.1**. The crash count calculated for each of the generated crash group. Depending on the scenario, the naming convention of crash group is in a format of ‘A’, ‘A-B’, or ‘A-B-C’ in which A, B, and C are the traffic crash abbreviations for the number of vehicles involved in crashes, manner of collision, and crash severity, respectively. For instance, ‘SV-OBJ-N’ is the crash group for single-vehicle object-related crashes with no injuries. Also, ‘MV-RRND-B+C’ is the feature for multi-vehicle rear-end crashes with suspected minor or possible injuries. In scenario 1, ‘TNC’ is the crash group including all crashes occurred in each segment.

Table 4.1. Traffic Crash Categories.

Traffic Crash Data Categories	
Number of Vehicle Involved in Crashes	
Single-Vehicle (SV)	Crashes that only involves one motor vehicle.
Multi-Vehicle (MV)	Crashes that involve two or more motor vehicles.
Manner of Collision	
Fixed Object (OBJ)	Crashes that involve hiding fixed objects as the first harmful event.
Over-turned (OVT)	Crashes that the first harmful event is identified as vehicle overturn.
In-Transport (TRNSP)	
Angled (ANG)	Crashes that two motor vehicles are collided at an angle.
Rear-End (RE)	Crashes that a motor vehicle is rear-ended by another motor vehicle.
Sideswipe (SDSP)	Crashes that a motor vehicle is sideswiped by another motor vehicle.
Other (OTH)	Crashes that the manner of collision is none of the items above.
Crash Severity	
A - Suspected Serious Injury	Severe injury that prevents continuation of normal activities leading to temporarily or permanent incapacitation.
B - Suspected Minor Injury	Evident injury such as bruises, abrasions, or minor lacerations which do not incapacitate.
C - Possible Injury	Injury claimed, reported, or indicated by behavior but without visible wounds, includes limping or complaint of pain
K - Fatal	If death resulted due to injuries sustained from the crash, at the scene or within 30 days of crash.
N - Not Injured	The person involved in the crash did not sustain as A, B, C, or K injury.
99 - Unknown	Unable to determine whether injuries exist. Some examples may include hit and run, fled scene, fail to stop and render aid.

Table 4.2. IH 20 (EB/WB) Traffic Crash Statistics by Severity.

Crash Data (2015-2019)	**% Range Across Corridors**		**% Total Crashes in Dallas County**
Crash Severity	Min	Mx	
99 - UNKNOWN	0.63%	3.06%	1.19%
A - SUSPECTED SERIOUS INJURY	1.27%	4.14%	2.05%
B - SUSPECTED MINOR INJURY	6.82%	13.52%	10.55%
C - POSSIBLE INJURY	16.45%	30.57%	21.15%
K - FATAL INJURY	0.24%	1.39%	0.48%
N - NOT INJURED	54.97%	73.46%	64.57%

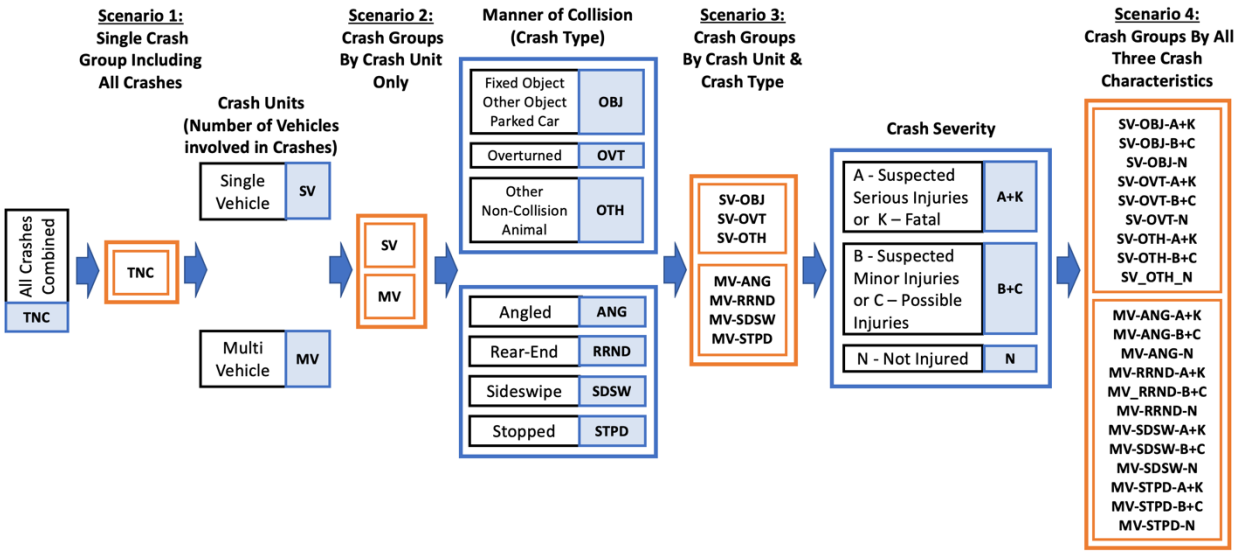


Fig. 4.1. Three Dimensions of Traffic Crashes and Four Scenarios.

4.4. METHODOLOGY

4.4.1. Introduction

This study methodology consists of three stages, as shown in **Fig. 4.2**.: (1) identifying a recommended fragment size (RFS) to aggregate the crash data; (2) performing count regression models for four crash group scenarios with fragment sizes ranging between 0.10 and 0.25 in 0.01-mile increments; (3) comparing the model performance including Akaike Information Criterion (AIC) and root mean square error (RMSE) for all iterations and investigating the potential benefits of the RFS for crash prediction model performance. Stage 1 requires calculating the feature crash rates (FCRs) for the scenario 4 crash groups (**Fig. 4.1**). The approach selects the RFS using the Laplacian score accompanied with a distance-based entropy measure, LSDBEM, (Liu, et al., 2009) to determine the best subset of features for K-means clustering. The methodology identifies the RFS as the fragment size associated with the K-means clusters with the highest silhouette score (Maniei & Mattingly, 2023a). The previous study shows that clustering roadway segments using FCRs outperforms the clustering results based on total crash rates, TCRs (Maniei & Mattingly, 2023a). Stage 2 estimates crash count data regression models including Poisson, NB, NBP, ZIP,

ZINB, ZINBP, GP-1, GP-2, and Hurdle regression models using traffic operational and geometric characteristics (**Table 4.3**) as explanatory variables for the four crash group scenarios with fragment sizes, ranging from 0.10 to 0.25 miles in 0.01 mile increments. For each fragment size increment the methodology selects the best model for further analysis. Finally, stage 3 compares the selected models for each crash group and fragment size to investigate the magnitude of improvement in the model performance measures when the RFS is used for data aggregation rather than other fragment sizes for the same crash group. Before making any comparisons, stage 3 eliminates all models from fragment sizes that exhibit excessive multicollinearity among independent variables.

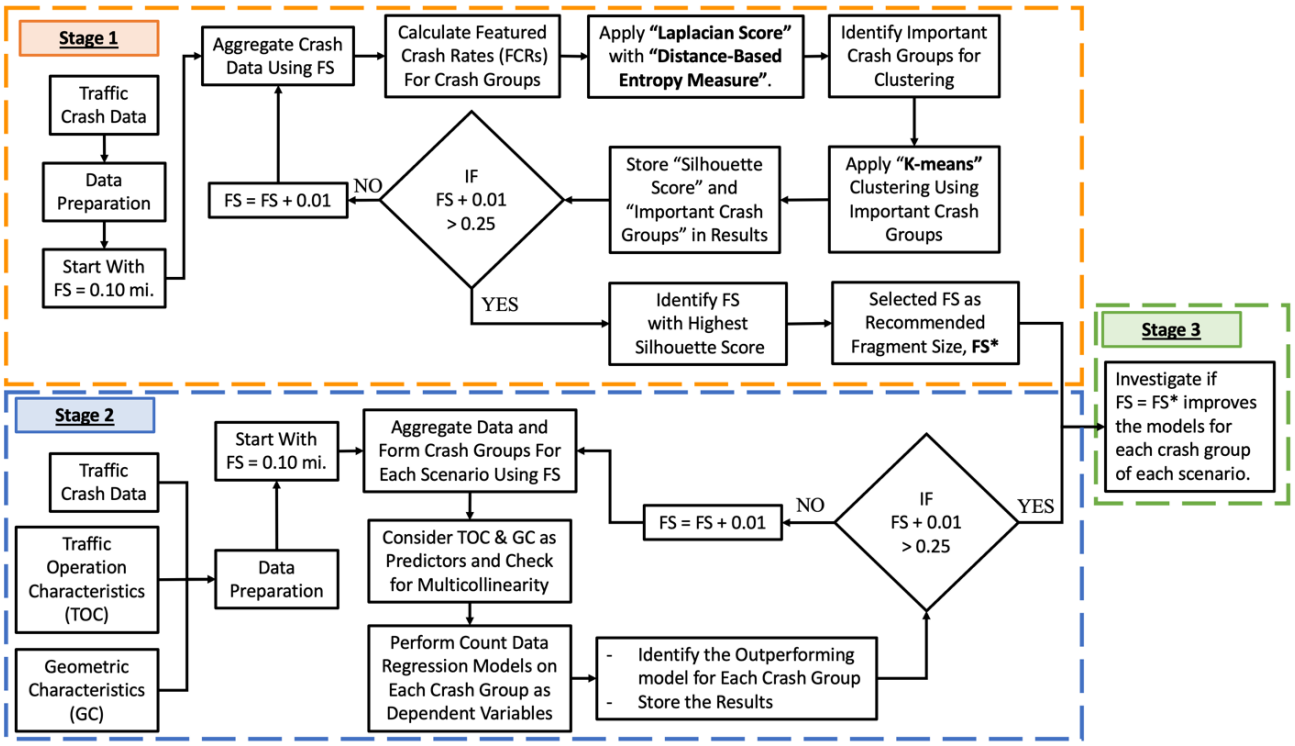


Fig. 4.2. Methodology Flow Chart.

Table 4.3. Description of Explanatory Variables (Maniei & Mattingly, 2023b).

Explanatory Variables	Description
$AADT_T$	Annual average daily traffic for single-unit/combo trucks (1000 vehicles per day, 1000 vpd)
$AADT_{NT}$	Annual average daily traffic for non-truck vehicle (1000 vehicles per day, 1000 vpd)
CDD	Horizontal curve delta angle a.k.a central angle (degree)
R	Horizontal curve radius (1000 ft)
TR-Seg	Binary variable with value of 1 if traffic crash site is on a segment of roadway transitioning from straight (tangent) segment to a curved segment or a curved-to-right segment to a curved-to-left segment or vice versa. Otherwise, 0.
LT-Seg	Binary variable with value of 1 if traffic crash site is on a curved-to-left segment. Otherwise, 0.
RT-Seg	Binary variable with value of 1 if traffic crash site is on a curved-to-right segment. Otherwise, 0.

4.4.2. Identifying recommended fragment size (RFS)

To identify the most important featured crash rates (FCRs) for clustering, Maniei and Mattingly (2023a) evaluated the Laplacian score with distance-based entropy measure (LSDBEM) for the FCRs calculated for each crash group. The method performed the K-mean clustering using the selected FCRs to segmentize the roadway based on the crash data. The process considered each fragment size between 0.10 and 0.25 mile in 0.01-mile increments. Maniei and Mattingly (2023a) proposed that the fragment size corresponding to the outperforming clustering result based on its silhouette score can be used for the roadway segmentation as a recommended fragment size (RFS). See Maniei and Mattingly (2023a) for a complete discussion of the methodology. This study utilized the LSCBEM followed by the K-means clustering to find the RFS.

4.4.3. Developing crash prediction models

Traffic safety studies use several count regression models to predict crash counts. And they most commonly use Poisson regression. However, Poisson regression overlooks the over-dispersion present in the crash data (Lord & Mannering, 2010). To handle the over-dispersion, other studies utilize the negative binomial regression model (Anastasopoulos & Mannering, 2009; Geedipally & Lord, 2010; Ma, et al., 2017; Ma, et al., 2017). The definition of standard negative binomial

probability mass function and the negative binomial regression model can be found in previous studies (Chow & Steenhard, 2009). Notably, the NB model contracts to a Poisson regression model when there is no over-dispersion. Furthermore, an alternative variation of the negative binomial regression model exists (Greene, 2008), referred to as the negative binomial model-type P (NBP) and the parameter P providing a general form to describe the relationship between mean and variance: $E(Y_i) = \lambda_i, Var(Y_i) = \lambda_i + \alpha \lambda_i^P$ where y_i is the number of traffic crashes happened on segment i during a period of time, and λ_i is the mean predicted traffic crash frequency for segment i. The presumption is that λ_i is a function of exploratory vector X_i. However, the NBP model is still unable to deal with the potential under-dispersion in count data. Another model called Generalized Poisson (GP) regression is also used to analyze the count data handling both over-dispersed and under-dispersed data. Depending on the mean and variance relationship, the GP model can be implemented in two forms: Consul's Generalized Poisson (GP-1) and Famoye's Generalized Poison (GP-2).

Besides the over-dispersion, according to Dong et al. (2014), there is a challenge to handle excessive zeros in crash count data that the commonly used Poisson and negative binomial are unsuitable for dealing with it. To tackle this issue, researchers widely employ zero-inflated models, as Carson and Mannering (2001) and Qin et al. (2005) noted. The zero-inflated models, including Poisson (ZIP) and zero-inflated negative binomial (ZINB) models, operate under the assumption that the additional zeros in the dataset arise from two distinct states: a true-zero state, indicating an inherently safe roadway segment, and a nonzero state, where no crashes occur during the observation period (Shankar, et al., 1997). Numerous studies demonstrate that these models offer a statistically superior fit to the data (Malyshkina & Mannering, 2010). Similar to the NB model, the ZINB model reduces to a ZIP model when over-dispersion is zero. Arguably, the assumption

that roadway segments are intrinsically safe, which forms the basis of the zero-inflated models, is highly improbable (Lord & Park, 2008); (Lord, et al., 2005). Even on well-designed roadway segments, crashes can occur due to unsafe driver behavior. Consequently, the fundamental assumption of the zero-inflated model is flawed. In response, researchers have turned to an alternative approach known as the Hurdle or two-part model, as Ma et al. (2016) employed, to handle excessive zeros in the dataset. The Hurdle model operates in two stages: firstly, it determines whether the count value is zero or positive, and secondly, if positive, it employs a truncated count distribution for analysis, following the methodology outlined by Cragg (1971). The Hurdle model assumes that roadway segments with zero observed crashes during the study period are only safe for that specific duration rather than being inherently safe. As aforementioned, the hurdle regression is a two-part model. The first part handles the probability of a zero count and the second part deals with the count distribution for non-zero values (Mullahy, 1986). In previous research, Maniei and Mattingly (2023b) implemented the count regression models including Poisson, NB, NBP, ZIP, ZINB, ZINBP, GP-1, GP-2, and Hurdle regression models to predict crash counts for various crash groups and identify the hotspots using the potential for safety improvement (PSI). In this study, all the models mentioned above will be developed to investigate the impact of various fragment sizes and recommended fragment size (RFS) on the model performance measures. This process is depicted as stage 2 in **Fig. 4.2**.

4.4.4. Modeling Process and Model Selection

The study estimates many count data regression models to analyze different crash groups using the chosen independent variables (**Table 4.3.**). The modeling process assesses the multicollinearity among the explanatory variables using the variance inflation factor (VIF). The multicollinearity is discussed in the results section. The selection process considers nine different count regression

models for each crash group. The model selection process identifies the superior model for each crash group. Vuong's test compares the base models (Poisson, NB, and NBP) with their corresponding zero-inflated models (ZIP, ZINB, and ZINBP). The process evaluates the models selected by Vuong's test with the remaining models (GP1, GP2, and Hurdle) based on their AIC values to determine the final model for each crash group.

4.4.5. Fragment Sizes and RFS Evaluation

This study determines the outperforming models based on their AIC values and evaluates these models across the fragment sizes in stage 3. For all scenario's crash groups, the values of dispersion parameter detected by crash prediction models are illustrated and explored for various fragment sizes via bubble charts. This reveals the effect of the fragment size on the amount of dispersion in the aggregated data, which leads to improving the performance of some model types and may make them the outperforming model for a particular fragment size and scenario. After identifying the outperforming model for each fragment size, the evaluation eliminates fragment sizes that exhibit serious multicollinearity among the explanatory variables using variance inflation factor (VIF); The explanatory variables with VIF greater than 10 show a serious multicollinearity. Also, the modeling result carries out a serious multicollinearity if the average VIF values for all explanatory variables, $\overline{VIF}$, exceeds 5. The study excludes fragment sizes resulting in serious multicollinearity are excluded from statistical modeling. To emphasize the predictive capabilities of the models, the evaluation compares the remaining models using the testing data set RMSE. The minimum RMSE, $\mathrm{RMSE}_{\mathrm{min}}$, represents the smallest RMSE of the remaining fragment size models; the study evaluates the performance of the RFS model by comparing its $\mathrm{RMSE}_{\mathrm{RFS}}$ value with $\mathrm{RMSE}_{\mathrm{min}}$ using a percentage difference measure. This evaluation occurs across all four

scenarios and the inquiry identifies the scenarios and circumstances within a scenario where the RFS returns satisfactory performance.

4.4.6. Modeling Implementation

In this study, a library of functions is scripted using Python programming language (Python 3) to implement the process, encompassing data cleaning, preparation, feature selection, regression model development, and model selection. To mitigate overfitting, the traffic crash data for each travel direction is divided into training and testing sets, with a ratio of 70% for training and 30% for testing. Each run's average computation time is 1151.21s and 801.52s for IH 20 EB and WB (utilizing a 6-Core Intel Core i7, 2.6 GHz CPU, and 16 GB memory), respectively.

4.5. RESULTS

This section discusses the study result. Stage 1 determines the RFS. Stage 2 estimates and selects count data regression models using crash data, traffic operational characteristics, and geometric characteristics that are aggregated using fragment sizes ranging from 0.10 mile to 0.25 mile with an increment of 0.01 mile. Prior to performing statistical models, the dependent and independent variables are checked for multicollinearity using the variance inflation factor (VIF). Finally, the stage 3 analysis investigates the suitability of the RFS using testing data.

4.5.1. Recommended fragment size (RFS)

The study methodology starts with finding the recommended fragment size (RFS) proposed by Maniei and Mattingly (2023a). Their study suggested applying the LSDBEM followed by K-means clustering to the featured crash rates calculated for the tridimensional crash groups. The tridimensional crash groups are shown as scenario 4 crash groups in **Fig. 4.1**. This process is denoted as stage 1 of the study methodology, iterating over the fragment sizes ranging between 0.10 mile to 0.25 mile with an increment of 0.01 mile. **Table 4.4.** presents the stage 1 results for IH 20 EB and WB. The previous study concluded that the fragment size for the clustering with the

highest silhouette score is the recommended fragment size (RFS). The two highest silhouette scores are shown with asterisks in **Table 4.4**. In descending order, the two highest silhouette scores for IH 20 EB are 0.9699 and 0.9647. For IH 20 WB, the two highest silhouette scores are 0.9223 and 0.9153, in descending order. According to Maniei and Mattingly (2023a), the RFS for both IH 20 EB and WB is 0.10 mile. The RFS is used in stage 3 to compare the statistical model performance for the RFS against other fragment size values used to aggregate the data to develop the statistical models.

Table 4.4. Stage 1 results using LSDBEM and K-mean clustering for IH 20 EB/WB.

	Silhouette Scores for Various Fragment Sizes															
FS (mile)	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20	0.21	0.22	0.23	0.24	0.25
IH 20 EB	0.9699*	0.9647**	0.391	0.957	0.954	0.904	0.945	0.687	0.822	0.935	0.666	0.861	0.912	0.919	0.437	0.691
IH 20 WB	0.9223*	0.9153**	0.908	0.682	0.416	0.885	0.730	0.638	0.873	0.868	0.589	0.544	0.484	0.482	0.649	0.598

4.5.2. Multicollinearity and modeling results

In stage 2, statistical models are developed to predict the crash count for each crash group in each scenario using explanatory variables in **Table 4.3**. Under a normal modeling process, the explanatory variables need to be investigated for multicollinearity since it may adversely impact the statistical model results and validity. This study utilizes the variance inflation factor (VIF) to investigate the multicollinearity among the explanatory variables. The VIF estimates the inflation in the variance of statistical model coefficients due to multicollinearity. The values of explanatory variables differ for each fragment size which causes changes in the corresponding VIF values; therefore, the analysis investigates the multicollinearity at the beginning of each iteration in stage 2 (**Fig. 4.2.**). Serious multicollinearity among the explanatory variables exists if $VIF_{max} > 10$ or $VIF_{mean} > 5$. The multicollinearity analysis results for IH 20 EB and WB are shown in **Table 4.5.** and **Table 4.6.**, respectively. For IH 20 EB, a serious multicollinearity among explanatory variables occurs for the fragment sizes of 0.18, 0.19, 0.21, 0.22, 0.23, and 0.25 mile because $VIF_{mean} > 5$.

For IH 20 WB, a serious multicollinearity exists among the explanatory variables for the fragment sizes of 0.18, 0.19, 0.21, 0.22, 0.23, 0.24, and 0.25 mile because $VIF_{mean} > 5$. The fragment size of 0.10 mile shows the best VIF results for both IH 20 EB and WB.

Table 4.5. Multicollinearity analysis of explanatory variables for IH 20 EB.

IH 20 EB - Variance Inflation Factor (VIF)																
Fragment Size (mile)	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20	0.21	0.22	0.23	0.24	0.25
$AADT_T$	8.883	9.096	9.144	9.046	9.129	9.124	9.131	9.195	9.254	9.298	9.131	9.190	9.305	9.323	9.507	9.460
$AADT_{NT}$	8.983	9.231	9.238	9.168	9.100	9.193	9.103	9.166	9.448	9.404	9.074	9.370	9.706	9.508	9.629	9.288
CDD	3.926	5.334	4.893	4.432	4.500	4.951	4.296	4.550	7.129	7.920	4.198	7.051	6.959	7.007	6.068	7.010
R	1.085	1.106	1.112	1.106	1.142	1.103	1.141	1.135	1.143	1.137	1.128	1.321	1.142	1.149	1.142	1.146
TR-Seg	1.410	1.640	1.655	1.410	1.726	1.686	1.808	1.746	2.549	2.455	1.809	2.630	2.933	2.858	3.358	2.947
LT-Seg	2.702	2.966	2.504	3.017	2.436	2.722	2.219	2.362	2.151	2.797	2.224	1.879	1.956	2.168	1.602	2.028
RT-Seg	2.016	2.939	2.948	2.253	2.609	2.761	2.523	2.703	4.689	4.866	2.465	5.011	4.407	4.283	3.403	4.387
VIF_{Max}	8.983	9.231	9.238	9.168	9.129	9.193	9.131	9.195	9.448	9.404	9.131	9.370	9.706	9.508	9.629	9.460
VIF_{Mean}	4.144	4.616	4.499	4.348	4.378	4.506	4.317	4.408	5.195	5.411	4.290	5.207	5.201	5.185	4.958	5.181

Table 4.6. Multicollinearity analysis of explanatory variables for IH 20 WB.

IH 20 WB - Variance Inflation Factor (VIF)																
Fragment Size (mile)	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20	0.21	0.22	0.23	0.24	0.25
$AADT_T$	9.439	9.339	9.643	9.345	9.260	9.502	9.566	9.663	9.666	9.581	9.582	9.547	9.597	9.861	9.913	9.700
$AADT_{NT}$	9.234	9.204	9.476	9.235	9.155	9.397	9.353	9.362	9.573	9.494	9.251	9.530	9.505	9.921	9.714	9.372
CDD	4.821	5.645	6.323	5.026	5.580	5.396	5.993	5.115	7.931	7.461	5.510	6.988	7.407	7.335	7.340	6.465
R	1.108	1.117	1.127	1.113	1.117	1.117	1.160	1.147	1.152	1.124	1.158	1.152	1.150	1.151	1.146	1.139
TR-Seg	1.418	1.639	1.804	1.286	1.704	1.818	2.082	1.930	2.765	2.260	2.124	3.060	3.120	3.038	3.193	3.236
LT-Seg	2.831	2.979	3.007	3.289	2.823	2.726	2.729	2.502	2.526	3.068	2.685	1.951	2.209	2.161	1.887	1.897
RT-Seg	2.688	3.140	3.577	2.566	3.216	2.998	3.340	2.793	4.708	4.180	2.849	4.142	4.230	4.253	4.332	3.514
VIF_{Max}	9.439	9.339	9.643	9.345	9.260	9.502	9.566	9.663	9.666	9.581	9.582	9.547	9.597	9.921	9.913	9.700
VIF_{Mean}	4.506	4.723	4.994	4.551	4.694	4.708	4.889	4.645	5.474	5.310	4.737	5.196	5.317	5.389	5.361	5.046

4.5.3. Model performance measure

The outperforming model for each crash group scenario is the model with the lowest AIC value. The outperforming model's AIC values are investigated by scatter diagrams color-coded by the fragment sizes without serious multicollinearity for SV-related and MV-related crash groups of IH 20 EB and WB. As an example, a scatter diagram of MV-related crash groups for IH 20 WB is shown in **Fig. 4.3**. Generally, all the scatter plots show a declining trend in AIC values by increasing the fragment size with few minor exceptions. In addition, the lower bound and the upper

bound of AIC values decrease by moving from each scenario to the next scenario. Also, the range of AIC values becomes tighter by moving from each scenario to the next scenario.

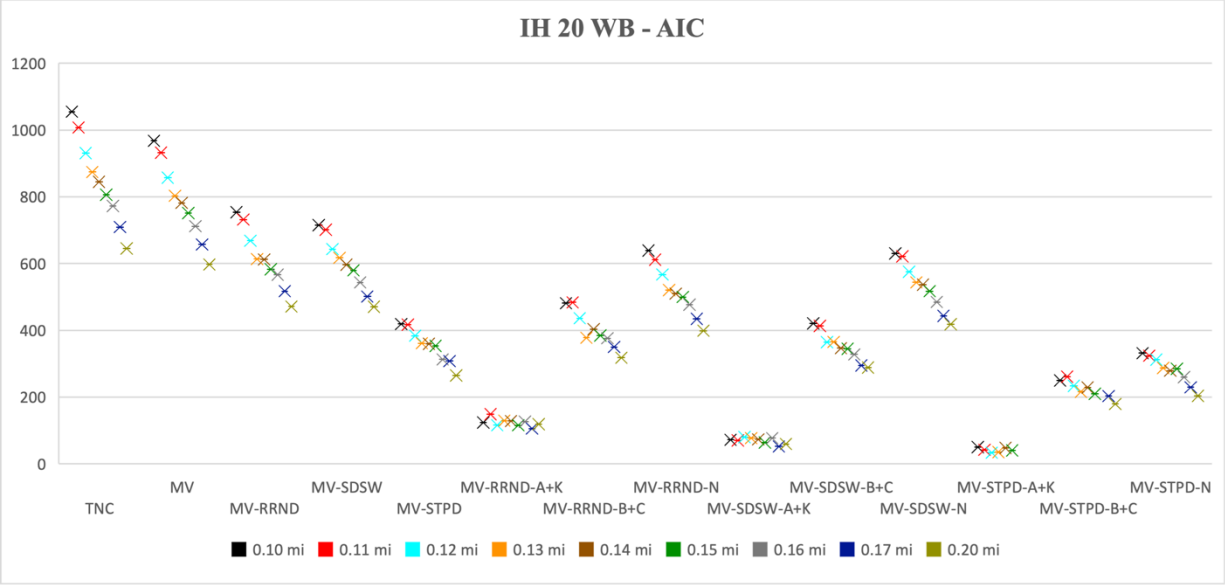


Fig. 4.3. IH 20 WB Outperforming Models for TNC and All MV Crash Group.

The root mean square error (RMSE) measures the model accuracy using the predicted and actual values for test set observations. The RMSE values associated with selected models of SV-related and MV-related crashes for IH 20 EB and WB are investigated by bar plots. **Fig. 4.4.** shows a bar plot for IH 20 WB depicting RMSE for SV-related crash groups. Mathematically, the statistical models with smaller RMSE are preferable. Unlike the AIC values, the RMSE values fluctuate for various fragment sizes for each crash group. Generally, the RMSE values show a declining trend by moving from each scenario to the next, similar to AIC values.

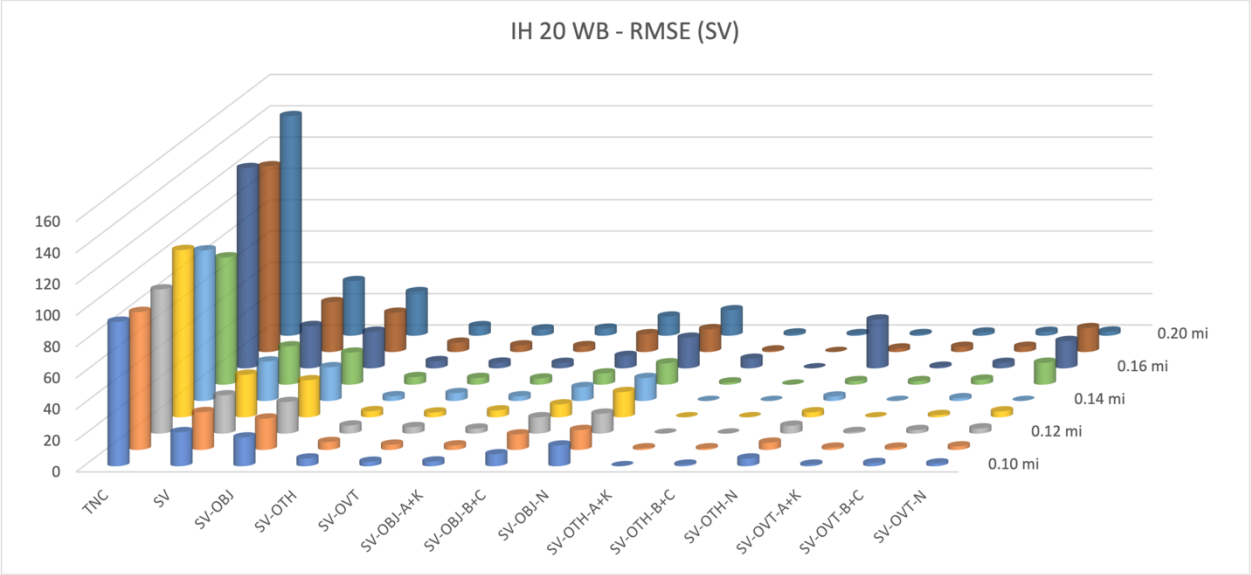


Fig. 4.4. IH 20 WB Model RMSE Values for TNC and All SV Crash Group.

Moreover, selection of fragment size impact data distribution after data aggregation. As a result, the amount of the dispersion in aggregated data may differ for various fragment sizes, if dispersion exist. Among the count regression model, there is a model effectively capturing the dispersion (variability in the aggregated data) as a good-fit model. On the other hand, the outperforming model is the model with the lowest AIC meaning the model effectively captures the pattern of aggregated data. . In this study, the count regression models account for the potential dispersion in the crash groups except Poisson and ZIP models. The amount of the dispersion in crash groups for various fragment sizes is investigated and no specific trend is detected. However, the amount of dispersion for each crash group is impacted by various fragment sizes.

4.5.4. RFS impact on model performance

Stage 3 investigates the impact of fragment size on data aggregation and model performance. Specifically, this stage evaluates if the RFS produces the best modeling results or improves the prediction model results. As previously noted, the multicollinearity between the explanatory variables shown in **Tables 4.5.** and **4.6.** led to the elimination of some fragment sizes from the RFS

evaluation because the fragment sizes produced flawed models. The fragment size of 0.10 mile provides the smallest multicollinearity between the explanatory variables.

The study investigates the model performance and accuracy of the prediction models for all four scenarios using their AIC and RMSE values. The minimum AIC, AIC_{min}, for SV-related crashes in scenarios 1, 2, and 3 occur with a fragment size (FS) of 0.24 mile except for 'SV-OTH' crash group with AIC_{min}, at FS = 0.17 mile for IH 20 EB (**Table 4.7.**). In scenario 4, the AIC_{min} fluctuates over various fragment sizes (**Table 4.7.-4.8.**). For IH 20 WB, AIC_{min} occurs at FS = 0.20 mile except for some crash groups in scenario 4 (**Table 4.7.-4.8.**). The analysis for this freeway does not estimate models for 'SV-OTH-A+K' and 'SV-OVT-A+K' due insufficient observations. Considering MV-related crashes, AIC_{min} values mostly appear at FS =0 0.24 mile for scenarios 1, 2, and 3 for IH 20 EB. In scenario 4, the AIC_{min} for 'MV-related' crashes appear at FS = 0.17, 0.20, or 0.24 mile or IH 20 EB (**Table 4.9.-4.10.**). For IH 20 WB, MV-related crash groups show their AIC_{min} at FS = 0.20 mile for all scenarios except for crash groups 'MV-RRND-A+K', 'MV-SDSW-A+K', and 'MV-STPD-A+K' at fragment size of 0.17, 0.17, and 0.12 mile, respectively (**Table 4.9.-4.10.**). The minimum AIC, AIC_{min}, does not occur at the RFS (0.10 mile) for either IH 20 EB or IH 20 WB.

Table 4.7. Minimum AIC Values and Corresponding Fragment Size for TNC & SV-related Crashes.

		Scenario 1	**Scenario 2 (SV)**	**Scenario 3 (SV)**			**Scenario 4 (SV)**	
		TNC	SV	SV-OBJ	SV-OTH	SV-OVT	SV-OBJ-A+K	SV-OBJ-B+C
IH 20 EB	AIC_{min}	599.89	410.01	379.32	121.91	150.29	101.96	233.56
	FS	0.24	0.24	0.24	0.17	0.24	0.17	0.24
IH 20 WB	AIC_{min}	645.119	434.033	410.032	128.496	98.739	90.421	257.464
	FS	0.20	0.20	0.20	0.20	0.20	0.20	0.20

Table 4.8. Minimum AIC Values and Corresponding Fragment Size for TNC & SV-related Crashes.

		Scenario 4 (SV)						
		SV-OBJ-N	SV-OTH-A+K	SV-OTH-B+C	SV-OTH-N	SV-OVT-A+K	SV-OVT-B+C	SV-OVT-N
IH 20 EB	AIC_{min}	327.02	-	28.372	112.59	-	112.18	61.13
	FS	0.24	-	0.12	0.15	-	0.17	0.24
IH 20 WB	AIC_{min}	344.681	18.00	24.175	137.682	22	49.937	58.917
	FS	0.20	0.20	0.15	0.12	0.15	0.16	0.17

Table 4.9. Minimum AIC Values and Corresponding Fragment Size for TNC & MV-related Crashes.

		Scenario 1	Scenario 2 (MV)	Scenario 3 (MV)			Scenario 4 (MV)	
		TNC	MV	MV-RRND	MV-SDSW	MV-STPD	MV-RRND-A+K	MV-RRND-B+C
IH 20 EB	AIC_{min}	599.89	564.65	448.29	452.83	299.49	81.16	61.79
	FS	0.24	0.24	0.24	0.24	0.24	0.16	0.20
IH 20 WB	AIC_{min}	645.119	597.328	471.249	470.524	264.57	105.69	317.86
	FS	0.20	0.20	0.20	0.20	0.20	0.17	0.20

Table 4.10. Minimum AIC Values and Corresponding Fragment Size for TNC & MV-related Crashes.

		Scenario 4 (MV)						
		MV-RRND-N	MV-SDSW-A+K	MV-SDSW-B+C	MV-SDSW-N	MV-STPD-A+K	MV-STPD-B+C	MV-STPD-N
IH 20 EB	AIC_{min}	433.45	49.11	292.24	414.95	20	204.98	229.05
	FS	0.17	0.17	0.24	0.24	0.20	0.17	0.24
IH 20 WB	AIC_{min}	398.811	52.433	288.76	418.419	33.527	179.67	203.65
	FS	0.20	0.17	0.20	0.20	0.12	0.20	0.20

For assessing the predictive quality of the crash models the evaluation prefers the lowest RMSE for the testing data. The $RMSE_{min}$ with its corresponding fragment size and $RMSE_{RFS}$ values for each scenario crash group for SV-related and MV-related crash groups are shown in **Table 4.11.-4.12.** and **4.13.-4.14.** for both IH 20 EB and WB, respectively. No prediction model was developed for 'SV-OTH-A+K' and 'SV-OVT-A+K' since the size of observations is insufficient. Comparing the $RMSE_{min}$ and $RMSE_{RFS}$ values for the SV-related crash groups for IH 20 EB, the percentage difference of $RMSE_{RFS}$ values are within 20% of the $RMSE_{min}$ for all scenario crash groups except for 'SV-OTH-N' and 'SV-OVT-N'. Considering SV-related crash groups for IH 20 WB, the percentage differences are within 13% of the $RMSE_{min}$ for scenarios 1, 2, and 3 except for the 'SV-

OTH' crash group. The RMSE comparison of scenario 4 SV-related crash groups show the percentage difference of five crash groups is less than 7%. However, the crash groups 'SV-OTH-B+C', 'SV-OTH-N, 'SV-OVT-A+K', and 'SV-OVT-B+C' have large percentage differences ranging from 30% to 200% due to a very limited number of non-zero observation after data aggregation. The RMSE percentage difference values of MV-related crash groups for scenarios 1, 2, and 3 remain within 17% and 18% for IH 20 EB and IH 20 WB, respectively. In scenario 4 for both IH 20 EB and IH 20 WB, MV-related crash groups show percentage differences within 20% except for some crash groups (**Table 4.13.-4.14.**).

Table 4.11. Minimum RMSE Values and Corresponding Fragment Size for TNC & SV-related Crashes.

		Scenario 1	**Scenario 2 (SV)**	**Scenario 3 (SV)**			**Scenario 4 (SV)**	
		TNC	SV	SV-OBJ	SV-OTH	SV-OVT	SV-OBJ-A+K	SV-OBJ-B+C
IH 20 EB	$RMSE_{min}$	102.63	27.586	22.271	3.606	3.742	2.646	7.746
	FS	0.10	0.13	0.15	0.12	0.11	0.10	0.15
	$RMSE_{RFS}$	102.6	33.5	23.5	4.0	4.5	2.6	7.9
	%-Diff	0.00%	19.22%	5.35%	10.36%	17.77%	0.00%	1.64%
IH 20 WB	$RMSE_{min}$	81.284	21.726	18.358	3	2.828	2.828	7.141
	FS	0.15	0.10	0.10	0.14	0.10	0.11	0.15
	$RMSE_{RFS}$	92.24	21.73	18.36	4.90	2.83	3.00	7.62
	%-Diff	12.63%	0.00%	0.00%	48.08%	0.00%	5.90%	6.44%

Table 4.12. Minimum RMSE Values and Corresponding Fragment Size for TNC & SV-related Crashes.

		Scenario 4 (SV)						
		SV-OBJ-N	SV-OTH-A+K	SV-OTH-B+C	SV-OTH-N	SV-OVT-A+K	SV-OVT-B+C	SV-OVT-N
IH 20 EB	$RMSE_{min}$	15.395	-	0	2.236	-	3	1.414
	FS	0.15	-	0.10	0.16	-	0.10	0.20
	$RMSE_{RFS}$	17.64	-	0.00	4.00	-	3.00	2.45
	%-Diff	13.56%	-	0.00%	56.57%	-	0.00%	53.62%
IH 20 WB	$RMSE_{min}$	12.689	0	0	1	0	1.414	1.732
	FS	0.11	0.10	0.12	0.20	0.13	0.11	0.10
	$RMSE_{RFS}$	13.19	0.00	1.00	4.90	1.00	2.00	1.73
	%-Diff	3.88%	0.00%	200.00%	132.19%	200.00%	34.33%	0.00%

Table 4.13. Minimum RMSE Values and Corresponding Fragment Size for TNC & MV-related Crashes.

		Scenario 1	Scenario 2 (MV)	Scenario 3 (MV)			Scenario 4 (MV)	
		TNC	MV	MV-RRND	MV-SDSW	MV-STPD	MV-RRND-A+K	MV-RRND-B+C
IH 20 EB	$RMSE_{min}$	102.63	78.975	37.108	34.598	13.153	2.45	1.732
	FS	0.10	0.10	0.10	0.24	0.16	0.15	0.12
	$RMSE_{RFS}$	102.63	78.975	37.108	34.641	15.492	3	13.602
	%-Diff	0.00%	0.00%	0.00%	0.12%	16.33%	20.18%	154.82%
IH 20 WB	$RMSE_{min}$	81.284	63.119	30.919	27.037	9	3.162	1.732
	FS	0.15	0.10	0.10	0.15	0.15	0.13	0.16
	$RMSE_{RFS}$	92.24	63.12	30.92	29.93	10.68	3.46	11.87
	%-Diff	12.63%	0.00%	0.00%	10.17%	17.05%	9.12%	149.08%

Table 4.14. Minimum RMSE Values and Corresponding Fragment Size for TNC & MV-related Crashes.

		Scenario 4 (MV)						
		MV-RRND-N	MV-SDSW-A+K	MV-SDSW-B+C	MV-SDSW-N	MV-STPD-A+K	MV-STPD-B+C	MV-STPD-N
IH 20 EB	$RMSE_{min}$	1.414	1.414	1	6.782	1	8.888	9.747
	FS	0.24	0.16	0.11	0.15	0.10	0.10	0.17
	$RMSE_{RFS}$	27.477	1.414	11.705	27.24	1	8.888	10.44
	%-Diff	180.42%	0.00%	168.52%	120.26%	0.00%	0.00%	6.87%
IH 20 WB	$RMSE_{min}$	2.45	2	1.414	0	0	5.745	7.483
	FS	0.20	0.13	0.15	0.20	0.14	0.10	0.10
	$RMSE_{RFS}$	20.494	2.45	10.392	22.158	410.992	5.745	7.483
	%-Diff	157.29%	20.22%	152.09%	200.00%	200.00%	0.00%	0.00%

4.6. DISCUSSION

This study objective is to examine the impact of the fragment size on the data aggregation and the traffic crash prediction results under different conditions by defining four scenarios. In addition, it is explored if the recommended fragment size (RFS) makes some improvement in crash prediction results as suggested by Maniei and Mattingly (2023a).

The research investigates the impact of the fragment size on multicollinearity. Prior to developing statistical models, the modelling process must investigate the multicollinearity among the explanatory variables. The study result shows the fragment size affects the explanatory variables VIFs (**Table 4.5.** and **4.6.**), and results in invalid models for some fragment sizes. The minimum VIF_{max} and VIF_{mean} occur at the RFS of 0.10 mile, because the RFS suggested by

Maniei and Mattingly (2023a) is based on the LSDBEM that accounts for multicollinearity (Liu, Yang, Ding, & Ma, 2009).

The AIC values for the selected models are compared over the various fragment sizes for all scenarios for IH 20 EB and WB. The AIC_{min} values for all scenario crash groups and their corresponding fragment size (FS) are shown in **Table 4.7.-4.8.** and **Table 4.9.-4.10.** for SV-related and MV-related crash groups. The result shows that the AIC_{min} values for scenario 1, SV-related, and MV-related crash groups in scenarios 2 and 3 occur at FS of 0.24 and 0.20 mile for IH 20 EB and IH 20 WB, respectively, except for 'SV-OVT' with AIC_{min} at FS=0.17 mile for IH 20 EB. The AIC_{min} values for scenario 4 fluctuate over the various fragment sizes but there are crash groups with AIC_{min} values at FS of 0.24 and 20 mile for IH 20 EB and IH 20 WB, respectively. No crash group shows AIC_{min} at the fragment size of 0.10 mile, which is the RFS suggested by Maniei and Mattingly (2023a). Therefore, the RFS fails to produce the AIC_{min} values for the selected models for all scenarios crash groups. Besides the fragment size, the impact of traffic crash dimensions on the AIC values are investigated. The result shows that the AIC values are constantly improved by including higher dimensions of traffic crashes. The selected model AIC values decline for each crash group in scenarios 2, 3, and 4 from their parent crash groups in scenarios 1, 2, and 3 as shown in **Fig. 4.3.** However, the improvement in AIC values is negligible for crash groups including 'MV', 'MV-SDSW', and 'MV-SDSW-N' for both IH 20 EB and WB. Because these crash groups show minor improvement in AIC values, it can be argued that the RFS of 0.10 mile might be as effective as the FS corresponding to AIC_{min} values for these crash groups.

Additionally, the various fragment sizes for IH 20 EB and IH 20 WB impacted the dispersion for each crash group. The changes in dispersion of the crash groups resulting from various fragment sizes led to different types of count regression models appearing suitable as the selected model for

a crash group because the selected model type better captured the dispersion, or variation in the data. The AIC value for each model type indirectly related to the dispersion derived by the model assumptions on mean and variance relationship. Among all the models performed for each crash group for each fragment size, the outperforming model effectively captured the magnitude of the dispersion (or variability in the aggregated data). Therefore, the outperforming model, the model with the lowest AIC, effectively captured the dispersion in the aggregated data. **Fig. 4.5.A. – 4.5.C.** showed the top two suitable models as GP2 and GP1 for the 'TNC' crash group, Poisson and GP1 for the SV-related crash groups, and ZIP and GP2 for the MV-related crash groups.

Fig. 4.5.A.

Fig. 4.5.B.

Fig. 4.5.C.

Fig. 4.5.A. – 4.5.C. Word Cloud Diagram of Outperforming Model Type for 'TNC', 'SV'-related, and 'MV'-related Crashes, respectively.

The $\mathrm{RMSE}_{\mathrm{min}}$ values for all scenario crash groups for IH 20 EB and WB with their corresponding FS and $\mathrm{RMSE}_{\mathrm{RFS}}$ are shown in **Table 4.11.-4.12.** and **Table 4.13.-4.14.** for SV-related and MV-related crash groups. The results show that the $\mathrm{RMSE}_{\mathrm{min}}$ values for the selected models fluctuate over the various fragment sizes. Unlike AIC, $\mathrm{RMSE}_{\mathrm{min}}$ values differ for various FS in scenarios 1, 2, and 3 as they do in scenario 4. However, the FS of 0.10 mile coincides with $\mathrm{RMSE}_{\mathrm{min}}$ for some crash groups including 'SV-OBJ-A+K', 'SV-OTH-B+C', and 'SV-OVT-B+C' for IH 20 EB and 'SV', 'SV-OBJ', 'SV-OVT', 'SV-OTH-A+K', 'MV', 'MV-RRND', 'MV-STPD-B+C', and 'MV-STPD-N'. Moreover, the percentage difference values of the RMSE values for the recommended

fragment size, $RMSE_{RFS}$, and $RMSE_{min}$ values are within at most 20% for all crash groups in scenarios 1, 2, and 3 for both IH 20 EB and IH 20. WB, except the 'SV-OTH' crash group for IH 20 WB. This implies that the crash prediction models for RFS of 0.10 mile provide sufficient goodness of fit for scenarios 1, 2, and 3 compared to the crash prediction models for the fragment sizes with $RMSE_{min}$. However, a similar conclusion only applies for some of crash groups in scenario 4 for both IH 20 EB and IH 20 WB. This shows that crash groups in scenario 1, 2, and 3 benefited from the RFS for data aggregation by reaching the $RMSE_{min}$ or its proximity. The RFS only performed acceptably for some crash groups in scenario 4. In particular, the RFS provides unsatisfactory performance for crash groups with a very limited number of non-zero observations after data aggregation. This shows that the RFS represents a reasonable strategy to simplify fragment selection and modeling for all scenarios and crash groups with sufficient non-zero observations.

4.7. CONCLUSIONS

This study explores the impact of various fragment sizes and traffic crash dimensions on multicollinearity and statistical model performance and accuracy. Also, it examines the potential benefits of the RFS obtained by LSDBEM/K-means for the statistical models. Stage 1 recommends a RFS value of 0.10 mile for both IH 20 EB and WB.

This study considers the variance inflation factor (VIF), AIC, and RMSE for all predictive models with the dispersion values to study the impact of fragment sizes. The VIF results illustrate that the multicollinearity between the explanatory variables differ for various fragment sizes such that serious multicollinearity appears for all FS of 0.18 mile and greater for both IH 20 EB and WB, except for the FS of 0.20 and 0.24 mile for IH 20 EB and the FS of 0.24 mile for IH 20 WB. This indicates that after data aggregation begins to exhibit multicollinearity additional increases in FS may result in specific data aggregations that do not show serious multicollinearity, but

predicting these patterns a priori appears difficult. The result shows that the minimal multicollinearity occurs for a FS of 0.10 mile for both IH 20 EB and WB, which is the RFS obtained in stage 1. The AIC values for outperforming models fluctuate over the various fragment sizes and scenario's crash groups. The AIC_{min} values for all crash groups (SV and MV related) in scenarios 1, 2, and 3 occur at FS of 0.24 mile except for 'SV-OVT' with AIC_{min} at FS of 0.17 mile. Therefore, the RFS (0.10 mile) does not achieve the lowest AIC values for scenarios 1, 2, and 3. Since the difference between the AIC values for the various fragment sizes are negligible for some of crash groups, it the RFS (0.10 mile) appears to roughly correspond to AIC_{min} for those crash croups. Similarly in scenario 4, crash groups with negligible differences between AIC values exist while a few crash groups reach AIC_{min} at FS of 0.24 and 0.20 mile for both IH 20 EB and WB. Therefore, the minimum AIC, AIC_{min}, does not occur at FS = 0.10 mile for all four scenarios. Another measure, RMSE, represents the accuracy of the crash prediction model when estimating the testing data set values. The RMSE values for each crash group differ across the fragment sizes. For some SV-related and MV-related crash groups, the $RMSE_{min}$ corresponds with the RFS of 0.10 mile. For other crash groups, the $RMSE_{RFS}$ remains within proximity (20%) to the $RMSE_{min}$, meaning that the RFS performance is approximately as good as using the fragment size associated with $RMSE_{min}$. For scenario 4, the RMSE values do not perform as well as those in scenarios 1, 2, and 3. This implies that the RMSE result fluctuates when crash severity is included for grouping crashes. Mostly, the RMSE values for the outperforming models for crash groups in scenarios 2, 3, and 4 decrease from their corresponding parent crash groups in scenarios 1, 2, and 3, respectively. Various fragment sizes impact the dispersion detected for each crash group, which impacts the types of count regression models emerging as the outperforming models. The model selection process identifies the top two 'TNC' crash group models as GP2 and GP1, the top two

SV-related crash group models as Poisson and GP1, and the top two MV-related crash group models as ZIP and GP2. The General Poisson regression models (GP1 and GP2) appear to perform well regardless of the crash group.

An expanded analysis and improved data quality may address many of the limitations associated with this study. The study findings are limited to the available traffic operational and roadway geometry data for IH 20 in Dallas County. Expanding the crash investigation to more sites appears necessary to confirm the benefit of the recommended fragment size (RFS) for different highways and freeways and determine the contexts where applying the RFS procedure seems most appropriate. The study currently excludes any spatial correlation effect for the traffic crash data. Future research should examine the potential contributing factors absent from this study such as operating speed, pavement marking visibility, sunlight glare, ambient lighting, cross-slope, roadway profile grade, shoulder width, and lane width. The scope of the study does not include the discussion of significant contributing factors for crash groups and various fragment sizes due to the numerous models developed in this study. An innovative method to illustrate the models' coefficients and standard errors would provide insights about the contributing factors becoming statistically significant across the various fragment sizes for different crash groups. Although the RFS improves the multicollinearity among the explanatory variables, the outperforming models with minimum AIC typically occur at fragment sizes other than the RFS. Therefore, future research needs to evaluate the trade-off between selecting a fragment size to reach minimal multicollinearity versus choosing the fragment size with minimum AIC or RMSE as the outperforming model. Also, future studies must confirm the benefit of the RFS for data aggregation and denoting the restrictions on its application by examining other highways and freeways, and perhaps using other model types.

CHAPTER 5. CONCLUSION

This dissertation develops an innovative data-driven methodology to aggregate crash data and recommends a fragment size for aggregating crash, roadway, and traffic data. The RFS represents a solution for the arbitrary selection of a fragment size (segment lengths) that prior research deems a concern. The new method utilizes LSDBEM for crash feature selection and the K-means clustering algorithm as unsupervised learning tools to categorize highway segments based on traffic crash patterns. Throughout this analysis, the crash analyses use featured crash rates (FCRs) based on three traffic crash characteristics (i.e. crash units, type, and severity). Also, the study investigates the effect of higher dimensions using four scenarios of traffic crash characteristics on crash prediction models, the statistical significance of explanatory variables, and traffic crash hotspots identified using crash prediction models. The dissertation examines the impact of fragment size on the multicollinearity among explanatory variables, and crash prediction model under the four scenarios. Finally, the dissertation evaluates the suitability and use cases of the RFS (obtained by LSDBEM/K-means) for data aggregation in crash prediction modeling.

The dissertation evaluates the performance of the new clustering method by comparing the results based on FCR and TCR for the mainlane segments of urban highways' (Texas Loop 12, IH-20, IH-30, IH-35E, IH-45, IH-635, and US-75) within Dallas County for fragment sizes ranging from 0.10 mile to 0.25 mile in 0.01 mile increments. The clustering using FCR outperforms the TCR-based clustering, which indicates that FCRs represents a better strategy for choosing an aggregation pattern for the crash data. This suggests using the FCR-based clustering result for each highway travel direction as the RFS for data aggregation.

The dissertation limits the crash modeling investigation to EB and WB IH-20 with a 0.1-mile fragment size for data aggregation; however, the RFS evaluation expands the analysis to consider

fragment sizes ranging from 0.10 mile to 0.25 mile in 0.01-mile increments. When the dissertation compares the modeling results for the four scenarios using the three crash feature dimensions, the modeling outcomes demonstrate the anticipated results where the outperforming models differ for each crash group. According to the findings, specific contributing factors, such as truck AADT, non-truck AADT, horizontal curve delta degree, horizontal curve radius, curved-to-right segment, and curved-to-left segment, emerge as significant contributors within specific crash groups. However, the statistical significance of these factors changes for overall crashes, SV, and MV crashes. Consistent with prior research, the top ten hotspot segments vary for nearly half of the segments when considering different crash groups. The dissertation also confirms that the segments identified as MV crash hotspots more closely align with the total crash hotspots than those identified as hotspots for SV crashes. The study results conclude that scenario 4, which encompasses all three dimensions of traffic crash characteristics (i.e. crash unit, crash type, and crash severity), furnishes a more comprehensive insight into contributing factors and hotspot segments; however, including all three dimensions of traffic crash characteristics may not always produce meaningful results when a scenario 4 crash groups has insufficient observations.

Also confirming previous studies, the dissertation shows that the fragment size for data aggregation directly impacts the modeling process and results. The data aggregation impacts the multicollinearity among the explanatory variables, the dispersion values in the dependent variables (crash count for each crash group), the type of count data regression model becoming the outperforming model, and the significant explanatory variables. The data aggregation pattern also affects the performance and accuracy of the outperforming models. The result illustrates that the multicollinearity among explanatory variables differ for various fragment sizes, and the minimum multicollinearity among the explanatory variables occurs at the RFS values obtained using the

LSDBEM/K-means method, meaning the data aggregation recommended using the RFS limits the multicollinearity among the explanatory variables. Considering the crash prediction models, the AIC values for outperforming models fluctuate over various fragment sizes. The root mean square error (RMSE) values for the testing data estimated by the crash prediction models also fluctuate across the fragment sizes. The minimum RMSE, $RMSE_{min}$, occurs at the RFS for some crash groups, and the RMSE at RFS, $RMSE_{RFS}$, for all crash groups with sufficient non-zero observations is similar (within 20%) to the $RMSE_{min}$. This indicates that the RFS represents an acceptable strategy for standardizing data aggregation.

This study encountered some limitations. The investigation suffered from insufficient traffic operational and roadway geometric characteristic data. There is a need for quality data to expand the consideration of contributing factors or introducing real-time contributing factors. Ideally, a future study should include more precise quality data, and introduce traffic operational and roadway explanatory variables like operating speed, visibility of pavement markings, sunlight glare, cross-slope, shoulder width, and lane width. The application of the methods described within this dissertation were limited to Dallas County urban freeways and the modeling only considered a single urban highway (IH 20 within Dallas County). Previous studies noted the potential for spatial correlation among crash data, but the limited traffic operational and roadway geometric made including the spatial correlation less important. Future studies with better data should also add spatial correlation within the crash modeling methodology. Due to their infrequent occurrence and the complexity of their contributing factors, crash data patterns may change over time and introduce temporal instability in RFS and cluster structures. Future research must investigate the RFS and cluster structure temporal stability. Expanding on this research likely involves considering the temporal instability and unobserved variations related to environmental

characteristics and driver behaviors. The study solely focuses on clustering and recommended fragment length by incorporating all three traffic crash characteristics, but the LSDBEM/K-means clustering technique can be employed to analyze crash groups, considering scenarios such as crash units only or crash units combined with the manner of collision. This allows for a comparison with clustering results based on FCR and TCR. Subsequent research should explore the significance of incorporating additional crash characteristics in predicting crash risk and identifying contributing factors. Further investigations should delve into each traffic crash characteristic independently, comparing the outcomes with those obtained when considering all three characteristics simultaneously. This study created distinct clusters for each highway and travel direction separately, but future research could expand its scope by exploring network-wide clustering for comparative analysis.

This dissertation provides the foundation for many additional research studies. Future research should investigate using the clustering approach to develop aggregate modeling techniques that resemble principal components analysis by using the aggregate data from the clusters, and another study should evaluate the impacts of estimating crash models for each cluster. While this study includes three crash dimensions in its features, future studies may investigate fewer (e.g., number of vehicles and manner of collision, similar to scenario 2 and 3) and more crash dimensions (e.g., roadway geometry or AADT). The clustering may also include other non-crash features and incorporate spatial correlation. This study may expand the new RFS method to segmentize highways with a variable segment length rather than a constant length of the segment. The new clustering method can also form clusters based on traffic operational characteristics data, and find the RFS based on the traffic characteristics, which may be critical for real-time crash prediction. A subsequent inquiry could also modify the RFS approach to segmentize highways with a variable

fragment size (segment length) rather than adhering to a constant fragment size (segment length). Also, additional research needs to investigate the modeling balance between a fragment size that minimizes multicollinearity, the AIC testing data RMSE to identify an ideal model. Finally, the study methodology needs to be applied to other urban (or rural) highways/freeways in different locations to confirm the benefit of the RFS for data aggregation, its appropriate use cases, and its impact on crash prediction models.

REFERENCES

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

Abdel-Aty, M. & Pande, A., 2007. Crash data analysis: Collective vs. individual crash level approach. *Journal of safety research,* 38(5), pp. 581-587.

Abdel-Aty, M., Pemmanaboina, R. & Hsia, L., 2006. Assessing Crash Occurrence on Urban Freeways by Applying a System of Interrelated Equations. *Journal of the Transportation Research Board,* Volume 1953, p. 1–9.

Afghari, A. P., Haque, M. M. & Washington, S., 2020. Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes. *Accident Analysis and Prevention,* Volume 144, pp. 1-11.

Afghari, A. P. et al., 2021. How much should a pedestrian be fined for intentionally blocking a fully automated vehicle? A random parameters beta hurdle model with heterogeneity in the variance of the beta distribution. *Analytic Methods in Accident Research,* 01 12, Volume 32, p. 100186.

Aguero-Valverde, J., 2013. Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. *Accident Analysis & Prevention,* Volume 50, pp. 365-373.

Aguero-Valverde, J. & Jovanis, . P. P., 2008. Analysis of Road Crash Frequency with Spatial Models. *Transportation Research Record,* 1 January, 2061(1), pp. 55-63.

Aguero-Valverde, J. & Jovanis, P. P., 2010a. Spatial Correlation in Multilevel Crash Frequency Models. Volume 2165, pp. 21-32.

Aguero-Valverde, J. & Jovanis, P. P., 2010b. Spatial Correlation in Multilevel Crash Frequency Models - Effects of Different Neighboring Structures. *Transportation Research Record,* Volume 2165, p. 21–32.

Ahmed, M. M. & Abdel-Aty, M. A., 2012. The Viability of Using Automatic Vehicle Identification Data for Real-Time Crash Prediction. *IEEE Transactions on Intelligent Transportation Systems,* June, 13,(2), pp. 459-468.

Akaike, H. A., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19(6), pp. 716-723.

Al-Aamri, A. K. et al., 2021. Mapping road traffic crash hotspots using GIS-based methods: A case study of Muscat Governorate in the Sultanate of Oman. *Spatial Statistics,* 01 04, Volume 42, p. 100458.

Alabama Department of Transportation, 2015. *Alabama Speed Management Manual.* [Online] Available at: https://www.dot.state.al.us/publications

American Association of State Highway and Transportation Officials, 2010. *Highway Safety Manual.* Washington DC: s.n.

American Association of State Highway and Transportation Officials, 2010. *Highway Safety Manual.* Washington DC: s.n.

Anastasopoulos, P. C. & Mannering, F. L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention,* 41(1), pp. 153-159.

Anon., 2011. *sklearn.cluster.KMeans.* [Online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Anon., 2019. Investigation of the consequences of the modifiable areal unit problem in macroscopic traffic safety analysis: A case study accounting for scale and zoning. *Accident Analysis and Prevention,* Volume 132, pp. 1-17.

Anon., 2020. *Traffic Management Data Dictionary (TMDD).* [Online] Available at: http://www.ite.org/tmdd

Anon., n.d.

Azizi, L. & Hadi, M., 2021. Using Traffic Disturbance Metrics to Estimate and Predict Freeway Traffic Breakdown and Safety Events. *Transportation Research Record,* 2675(10), p. 723–733.

Barile, C., Casavola, C., Pappalettera, G. & Paramsamy Kannan, V., 2022. Laplacian score and K-means data clustering for damage characterization of adhesively bonded CFRP composites by means of acoustic emission technique. *Applied Acoustics,* Volume 185, p. 108425.

Barua, S., El-Basyouny, K. & Islam, M. T., 2014. A Full Bayesian multivariate count data model of collision severity with spatial correlation. *Analytic Methods in Accident Research,* Volume 3, p. 28–43.

Barua, S., El-Basyouny, K. & Islam, M. T., 2016. Multivariaterandomparameterscollisioncountdatamodels with spatialheterogeneity. *AnalyticMethodsinAccidentResearch,* Volume 9, p. 1–15.

Barua, S., El-Basyouny, K. & Islam, M. T., 2016. Multivariaterandomparameterscollisioncountdatamodels with spatialheterogeneity. *AnalyticMethodsinAccidentResearch,* Volume 9, pp. 1-15.

Bhatia, J. et al., 2020. SDN-based real-time urban traffic analysis in VANET environment. *Computer Communications,* Volume 149, p. 162–175.

Bhowmik, T., Yasmin, S. & Eluru, N., 2018. A joint econometric approach for modeling crash counts by collision type. *Analytic Methods in Accident Research,* Volume 19, pp. 16-32.

Bonneson, A. . J., National Research, C., American Association of State, H. T. O. & National Cooperative Highway Research, P., 2010. *Highway Safety Manual.* Washington DC: American Association of State Highway and Transportation Officials.

Borsos, A., Ivan, J. N. & Orosz, G., 2014. Development of safety performance functions for two-lane rural first-class main roads in Hungary. *Transport Research Arena (TRA) 5th Conference: Transport Solutions from Research to Deployment.*

Briz-Redón, Á., Martínez-Ruiz, F. & Montes, F., 2019. Investigation of the consequences of the modifiable areal unit problem in macroscopic traffic safety analysis: A case study accounting for scale and zoning. *Accident Analysis and Prevention,* Volume 132, pp. 1-17.

Cai, Z. et al., 2021. Modeling of Low Visibility-Related Rural Single-Vehicle Crashes Considering Unobserved Heterogeneity and Spatial Correlation. *Sustainability,* Volume 13, pp. 1-24.

Cameron, A. C. & Trivedi, P. K., 1986. Econometric models based on count data. Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics,* 01, 1(1), pp. 29-53.

Carson, J. & Mannering, F., 2001. The effect of ice warning signs on ice-accident frequencies and severities. *Accident Analysis & Prevention,* 01 01, 33(1), pp. 99-109.

Chang, F., Yasmin, S., Huang, H. & Chan, A. H., 2021. Injury severity analysis of motorcycle crashes: A comparison of latent class clustering and latent segmentation based models with unobserved heterogeneity. *Analytic Methods in Accident Research,* Volume 32, pp. 1-28.

Chang, G.-L. & Xiang, H., 2003. *The relationship between congestion levels and accidents,* s.l.: s.n.

Cheng, W. et al., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. *Accident Analysis and Prevention,* Volume 99, p. 330–341.

Cheng, Z. et al., 2022. Crash Risks Evaluation of Urban Expressways: A Case Study in Shanghai. *IEEE Transactions on Intelligent Transportation Systems,* pp. 1-11.

Chin, H. C. & Quddus, M. A., 2003. Modeling Count Data with Excess Zeroes: An Empirical Application to Traffic Accidents. *Sociological Methods & Research,* 01 08.pp. 90-116.

Chiou, Y.-C. & Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. *Accident analysis and prevention,* Volume 50, pp. 73-82.

Chow, N. & Steenhard, D., 2009. *A Flexible Count Data Regression Model Using SAS.* s.l., s.n., pp. 1-14.

Cook, D., Souleyrette, R. & Jackson, J., 2011. Effect of Road Segmentation on Highway Safety Analysis. *Transportation Research Board 90th Annual Meeting,* Volume 11-1995.

Cragg, J. G., 1971. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica,* 09, 39(5), pp. 829-844.

Daniels, S., Brijs, T., Nuyts, E. & Wets, G., 2010. Explaining variation in safety performance of roundabouts. *Accident Analysis & Prevention,* 42(2), pp. 393-402.

De Luca, M., Mauro, R., Lamberti, R. & Dell'Acqua, G., 2012. Road safety management using Bayesian and cluster analysis. *Procedia - Social and Behavioral Sciences,* p. 1260–1269.

Depaire, B., Wets, G. & Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. *Accident Analysis and Prevention,* Volume 40, p. 1257–1266.

Depaire, B., Wets, G. & Vanhoof, K., 2008. Traffic Accident Segmentation by Means of Latent Class Clustering. *Accident Analysis and Prevention,* 40(4), pp. 1257-1266.

Dezman, Z. et al., 2016. Hotspots and causes of motor vehicle crashes in Baltimore, Maryland: A geospatial analysis of five years of police crash and census data. *Injury,* November, 47(11), pp. 2450-2458.

Dong, B., Ma, X., Chen, F. & Chen, S., 2018. Investigating the Differences of Single-Vehicle and Multivehicle Accident Probability Using Mixed Logit Model. *Journal of Advanced Transportation,* pp. 2-9.

Dong, C. et al., 2014. Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression. *Safety Science,* 01 12, Volume 70, pp. 63-69.

Dong, N. et al., 2016. Macroscopic hotspots identification: A Bayesian spatio-temporal interaction approach. *Accident Analysis & Prevention,* 01 07, Volume 92, pp. 256-264.

El-Basyouny, K. & Sayed, T., 2006. Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. *Transportation Research Record,* 01 01, 1950(1), pp. 9-16.

El-Basyouny, K. & Sayed, T., 2009. Urban Arterial Accident Prediction Models with Spatial Effects. *Transportation Research Record,* 2102(1), p. 27–33.

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

Fitzpatrick, K., Schneider IV, W. & Carvell, J., 2006. Using the Rural Two-Lane Highway Draft Prototype Chapter. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 1950, p. 44–54.

Gaweesh, S. M., Ahmed, M. M. & Piccorelli, P. V., 2019. Developing crash prediction models using parametric and nonparametric approaches for rural mountainous freeways: A case study on Wyoming Interstate 80. *Accident Analysis and Prevention,* Volume 123, p. 176–189.

Geedipally, S. R. & Lord, D., 2010. Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson–gamma models. *Accident Analysis & Prevention,* July, 42(4), pp. 1273-1282.

Geyer, J. et al., 2008. *Methods for identifying high collision concentration locations for potential safety improvements,* s.l.: University of California at Berkeley.

Ghadi, M. & Torok, A., 2019. A comparative analysis of black spot identification methods and road accident segmentation methods. *Accident Analysis and Prevention,* p. 1–7.

Ghandour, A. J., Hammoud, H. & Telesca, L., 2019. Transportation hazard spatial analysis using crowd-sourced social network data. *Physica A,* Volume 520, p. 309–316.

Giuffrè, O. et al., 2014. Estimating the Safety Performance Function for Urban Unsignalized Four-Legged One-Way Intersections in Palermo, Italy. *Archives of Civil Engineering,* March, 60(1), pp. 41-54.

Golob, T. F., Recker, W. & Pavlis, Y., 2008. Probabilistic models of freeway safety performance using traffic flow data as predictors. *Safety Science,* Volume 46, p. 1306–1333.

Golob, T. F., Recker, W. W. & Alvarez, V. M., 2004a. Freeway safety as a function of traffic flow. *Accident Analysis and Prevention,* Issue 36, p. 933–946.

Golob, T. F., Recker, W. W. & Alvarez, V. M., 2004a. Freeway safety as a function of traffic flow. *Accident Analysis and Prevention,* Volume 36, p. 933–946.

Green, E. R., 2018. Segmentation Strategies for Road Safety Analysis. *UKnowledge,* pp. 1-200.

Greene, W., 2008. Functional forms for the negative binomial model for count data. *Economics Letters,* 01 06, 99(3), pp. 585-590.

Guo, F., Wang, X. & Abdel-Aty, M. A., 2010. Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis and Prevention,* 42(1), p. 84–92.

Hardin, J. W. & Hilbe, J. M., 2018. *Generalized linear models and extensions.* Fourth ed. s.l.:Stata press.

Hauer, E., 1980. Bias-by-selection: Overestimation of the effectiveness of safety countermeasures caused by the process of selection for treatment. *Accident Analysis & Prevention,* 12(2), pp. 113-117.

Hauer, E., 1997. *Observational before/after studies in road safety. estimating the effect of highway and traffic engineering measures on road safety.* Bingley: Emerald Group Publishing Limited.

Hauer, E., Ng, J. C. N. & Lovell, J., 1988. Estimation of safety at signalized intersections. *Transportation research record,* Volume 1185, pp. 48-61.

He, X., Cai, D. & Niyogi, P., 2005. Laplacian Score for Feature Selection. *Advances in neural information processing systems,* Volume 18.

Hilbe, J. M., 2011. *Negative binomial regression.* s.l.:Cambridge University Press.

Hilbe, J. M., 2014. *Modeling count data.* New York(NY): Cambridge University Press.

Huang, H. & Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. *Accident Analysis and Prevention,* Volume 42, p. 1556–1565.

Huang, H., Chin, H. C. & Haque, M. M., 2009. Empirical evaluation of alternative approaches in identifying crash hot spots: Naive Ranking, Empirical Bayes, Full Bayes Methods. *Transportation Research Record,* 2103(1), pp. 32-41.

Huang, H. et al., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. *Journal of Transport Geography,* 01 06, Volume 54, pp. 248-256.

Huang, H. et al., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. *Analytic Methods in Accident Research,* Volume 14, pp. 10-21.

Islam, M., Alnawmasi, N. & Mannering, F., 2020c. Unobserved heterogeneity and temporal instability in the analysis of work-zone crash-injury severities. *Analytic Methods in Accident Research,* Volume 28, p. 100130.

Islam, M. & Mannering, F., 2020a. A temporal analysis of driver-injury severities in crashes involving aggressive and non-aggressive driving. *Analytic Methods in Accident Research,* 27.p. 100128.

Islam, M. & Pande, A., 2020b. Analysis of Single-Vehicle Roadway Departure Crashes on Rural Curved Segments Accounting for Unobserved Heterogeneity. *Transportation Research Record,* 2674(10), pp. 146-157.

Islam, M., Perez-Bravo, D. & Silverman, K. K., 2017. *Performance-based Assessment to Transportation Safety Planning for Metropolitan Travel Improvement Study.* Washington, DC, Transportation Research Board.

Islam, M. T., El-Basyouny, K., Ibrahim, S. E. & Sayed, T., 2016. Before–After Safety Evaluation Using Full Bayesian Macroscopic Multivariate and Spatial Models. *Transportation Research Record: Journal of the Transportation Research Board,* 2601(1), p. 128–137.

Ismail, N. & Zamani, H., 2013. *Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models.* s.l., Citeseer, pp. 1-28.

Ivan, J. N., Pasupathy, R. K. & Ossenbruggen, P. J., 1999. Differences in causality factors for single and multi-vehicle crashes on two-lane roads. *Accident Analysis and Prevention,* Issue 31, p. 695–704.

Jung, S., Qin, X. & Noyce, D. A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis and Prevention,* Volume 42, p. 213–224.

Karim, A., Azam, S., Shanmugam, B. & Kannoorpatti, K., 2020. Efficient Clustering of Emails Into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework. *IEEE access,* Volume 8, pp. 154759-154788.

Khattak, M. W. et al., 2021. Estimation of safety performance functions for urban intersections using various functional forms of the negative binomial regression model and a generalized Poisson regression model. *Accident Analysis & Prevention,* 01 03, Volume 151, p. 105964.

Koorey, G., 2009. Road Data Aggregation and Sectioning Considerations for Crash Analysis. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 2103, p. 61–68.

Kwon, O. H., Park, M. J., Yeo, H. & Chung, K., 2013. Evaluating the performance of network screening methods for detecting high collision concentration locations on highways. *Accident Analysis and Prevention,* Volume 51, pp. 141-149.

Lee, J., 2014. *Development of Traffic Safety Zones and Integrating Macroscopic and Microscopic Safety Data Analytics for Novel Hot Zone Identification,* Orlando, Florida: University of Central Florida.

Lee, J. & Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. *Accident Analysis & Prevention,* 34(2), pp. 149-161.

Liu, R., Yang, N., Ding, X. & Ma, L., 2009. An Unsupervised Feature Selection Algorithm: Laplacian Score Combined with Distance-Based Entropy Measure. *2009 Third International Symposium on Intelligent Information Technology Application,* pp. 65-68.

Lord, D., Manar, A. & Vizioli, A., 2005. Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. *Accident Analysis & Prevention,* January, 37(1), pp. 185-199.

Lord, D. & Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice,* 01 06, 44(5), pp. 291-305.

Lord, D. & Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice,* June, 44(5), pp. 291-305.

Lord, D. & Park, B., 2015. Appendix D: Negative Binomial Regression Models and Estimation Methods. *Crime Stat Version,* Volume 3.

Lord, D. & Park, P. Y.-J., 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. *Accident Analysis & Prevention,* 01 07, 40(4), pp. 1441-1457.

Lu, J., Gan, A., Haleem, K. & Wu, W., 2013. Clustering-based roadway segment division for the identification of high-crash locations. *Journal of Transportation Safety & Security,* 5(3), pp. 224-239.

Mahmud, A. & Gayah, V. V., 2021. Estimation of crash type frequencies on individual collector roadway segments. *Accident Analysis & Prevention,* Volume 161, p. 106345.

Ma, L., Yan, X., Wei, C. & Wang, J., 2016. Modeling the equivalent property damage only crash rate for road segments using the hurdle regression framework. *Analytic Methods in Accident Research,* 01 09, Volume 11, pp. 48-61.

Malyshkina, N. V. & Mannering, F. L., 2010. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. *Accident Analysis & Prevention,* 01 01, 42(1), pp. 131-139.

Maniei, F. & Mattingly, S. P., 2023a. Unsupervised Approach to Investigate Urban Traffic Crashes Based on Crash Unit, Crash Severity, and Manner of Collision. *Working Paper.*

Maniei, F. & Mattingly, S. P., 2023b. Traffic Crash Hotspot Identification and Static Contributing Factors by Crash unit, Manner of Collision, and Crash Severity. *Working Paper.*

Mannering, F. L. & Bhat, C. . R., 2014. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research,* Volume 1, pp. 1-22.

Mannering, F. L., Shankar, V. & Bhat, C. . R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research,* Volume 11, pp. 1-16.

Mathworks, 2021. *fsulaplacian.* [Online]
Available at: https://www.mathworks.com/help/stats/fsulaplacian.html#mw_16dfb112-dac6-4cb6-84d4-6314df56c122
[Accessed 2022].

Ma, X., Chen, S. & Chen, F., 2016. Correlated Random-Effects Bivariate Poisson Lognormal Model to Study Single-Vehicle and Multivehicle Crashes. *Journal of Transportation Engineering,* November.142(11).

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

Ma, Z. et al., 2017. Predicting expressway crash frequency using a random effect negative binomial model: A case study in China. *Accident Analysis & Prevention,* Volume 98, pp. 214-222.

Miaou, S.-P. & Lord, D., 2003. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods. *Transportation Research Record,* 01 01, 1840(1), pp. 31-40.

Miaou, S.-P. & Song, J. J., 2005. Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis and Prevention,* Volume 37, p. 699–720.

Miranda-Moreno, L. F. & Fu, L., 2007. *Traffic safety study: Empirical Bayes or full Bayes?,* s.l.: s.n.

Mitra, S., 2009. *Spatial Autocorrelation and Bayesian Spatial Statistical Method for Analyzing Intersections Prone to Injury Crashes.* Washington, D.C.: Transportation Research Record: Journal of the Transportation Research Board.

Montella, A., 2010. A comparative analysis of hotspot identification methods. *Accident Analysis & Prevention,* 01 03, 42(2), pp. 571-581.

Montella, A., 2010. A comparative analysis of hotspot identification methods. *Accident Analysis & Prevention,* 42(2), pp. 571-581.

Mullahy, J., 1986. Specification and testing of some modified count data models. *Journal of Econometrics,* 01 12, 33(3), pp. 341-365.

Norden, M., Orlansky, J. & Jacobs, H., 1956. *Application of statistical quality-control techniques to analysis of highway-accident data.* Washington, DC, Highway Research Board, pp. 17-31.

Oppe, S., 1991. Development of traffic and traffic safety: Global trends and incidental fluctuations. *Accident Analysis & Prevention,* 01 10, 23(5), pp. 413-422.

Pande, A. & Abdel-Aty, M., 2006. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. Transportation Research Record:. *Journal of the Transportation Research Board,* Volume 1953, pp. 31-40.

Pande, A., Abdel-Aty, M. & Das, A., 2010. A classification tree based modeling approach for segment related crashes on multilane highways. *Journal of Safety Research,* p. 391–397.

Park, J., Abdel-Aty, M., Lee, J. & Lee, C., 2015. Developing crash modification functions to assess safety effects of adding bike lanes for urban arterials with different roadway and socio-economic characteristics. *Accident Analysis & Prevention,* 01 01, 74(0001-4575), p. https://doi.org/10.1016/j.aap.2014.10.024.

Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research ,* pp. 2825-2830.

Persaud, B., Lyon, C. & Nguyen, T., 1999. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. *Transportation Research Record,* 01 01, 1665(1), pp. 7-12.

Persaud, B., Lyon, C. & Nguyen, T., 1999. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. *Transportation Research Record,* 1665(1), pp. 7-12.

Persaud, B. & Mucsi, K., 1995. Microscopic accident potential models for two-lane rural roads. *Transportation Research Record,* pp. 134-139.

Qin, X., Ivan, J. N., Ravishanker, N. & Liu, J., 2005. Hierarchical Bayesian estimation of safety performance functions for two-lane highways using Markov chain Monte Carlo modeling. *Journal of Transportation Engineering,* 131(5), pp. 345-351.

Qin, X. & Wellner, A., 2012. Segment length impact on highway safety screening analysis. *Transportation Research Record: Journal of the Transportation Research Board,* Volume 12-0644.

Raschka, S. & Mirjalili, V., 2017. *Python Machine Learning.* Second ed. Birmingham: Packt Publishing.

Sørensen, M. & Elvik, R., 2007. *Black spot management and safety analysis of road network.* Oslo: Institute of transport economics.

Schlüter, P. J., Deely, J. J. & Nicholson, A. J., 1997. Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. *Journal of the Royal Statistical Society: Series D (The Statistician),* 01 09, 46(3), pp. 293-316.

Schwarz, G., 1978. Estimating the dimension of a model Annals of Statistics. *The Annals of Statistics,* 03, 6(2), pp. 461-464.

Shankar, V., Milton, J. & Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis & Prevention,* 01 11, 29(6), pp. 829-837.

Solorio-Fernández, S., Carrasco-Ochoa, J. A. & Martínez-Trinidad, J. F., 2020. A review of unsupervised feature selection methods. *Artificial Intelligence Review,* pp. 907-948.

Song, J., Ghosh, M., Miaou, S. & Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis,* 97(1), pp. 246-273.

Son, H. '., Kweon, Y.-J. & Park, B. '., 2011. Development of crash prediction models with individual vehicular data. *Transportation Research Part C,* Volume 19, p. 1353–1363.

Tamburri, T. N. & Smith, R. N., 1970. The safety index: A method of evaluating and rating safety benefits. *Highway Research Record,* Issue 332, pp. 28-43.

Texas Department of Transportation (TxDOT) - Traffic Safety Division, 2020. *Instruction to Police for Reporting Crashes,* Austin: s.n.

Thakali, L., Kwon, T. J. & Fu, L., 2015. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation,* 01 06, 23(2), pp. 93-106.

Thomas, I., 1996. Spatial data aggregation: exploratory analysis of road accidens. *Accident Analysis and Prevention,* 28(2), pp. 251-264.

Valent, F. et al., 2002. Risk factors for fatal road traffic accidents in Udine, Italy. *Accident Analysis and Prevention,* Volume 34, p. 71–84.

Vuong, Q. H., 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica,* 03, 57(2), pp. 307-333.

Wang, D. et al., 2022. Assessing dynamic metabolic heterogeneity in non-small cell lung cancer patients via ultra-high sensitivity total-body [18F]FDG PET/CT imaging: quantitative analysis of [18F]FDG uptake in primary tumors and metastatic lymph nodes. *European Journal of Nuclear Medicine and Molecular Imaging,* 49(13), pp. 4692-4704.

Wang, K., Zhao, S. & Jackson, E., 2019. Functional forms of the negative binomial models in safety performance functions for rural two-lane intersections. *Accident Analysis & Prevention,* 01 03, Volume 124, pp. 193-201.

Wang, K., Zhao, S. & Jackson, E., 2020. Investigating exposure measures and functional forms in urban and suburban intersection safety performance functions using generalized negative binomial - P model. *Accident Analysis & Prevention,* 01 12, Volume 148, p. 105838.

Wang, L., Abdel-Aty, M. & Lee, J., 2017. Safety analytics for integrating crash frequency and real-time risk modeling for expressways. *Accident analysis and prevention,* Volume 104, pp. 58-64.

Wang, X. & Feng, M., 2019. Freeway single and multi-vehicle crash safety analysis: Influencing factors and hotspots. *Accident Analysis and Prevention,* Volume 132, pp. 1-12.

Wang, Y. & Kockelman, K. M., 2013. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. *Accident Analysis & Prevention,* Volume 60, pp. 71-84.

Wen, H. et al., 2018. The Effects of Traffic Composition on Freeway Crash Frequency by Injury Severity: A Bayesian Multivariate Spatial Modeling Approach. *Journal of Advanced Transportation,* p. Article ID 6964828..

Winkelmann, R., 2008. *Econometric analysis of count data.* s.l.:Springer Science & Business Media.

World Health Organization (WHO), 2018. *Road safety,* Geneva: WHO.

Xie, K., Wang, X., Ozbay, K. & Yang, H., 2014. Crash frequency modeling for signalized intersections in a high-density urban road network. *Analytic Methods in Accident Research,* Volume 2, pp. 39-51.

Xu, C. et al., 2018. Utilizing Structural Equation Modeling and Segmentation Analysis in Real-time Crash Risk Assessment on Freeways. *KSCE Journal of Civil Engineering,* Volume 22, p. 2569–2577.

Xu, C., Liu, P., Wang, W. & Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. *Accident Analysis & Prevention,* Volume 47, pp. 162-171.

Xu, C., Tarko, A. P., Wang, W. & Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis and Prevention ,* Volume 57, p. 30– 39.

Yang, D., Xie, K., Ozbay, K. & Yang, H., 2021. Fusing crash data and surrogate safety measures for safety assessment: Development of a structural equation model with conditional autoregressive spatial effect and random parameters. *Accident analysis and prevention,* Volume 152, pp. 105971-105971.

Yang, Y., Ding, X. & Ma, L., 2009. *An Unsupervised Feature Selection Algorithm: Laplacian Score Combined with Distance-based Entropy Measure.* NanChang, China, s.n., pp. 65-68.

Yasmin, S. & Eluru, N., 2018. A joint econometric framework for modeling crash counts by severity. *Transportmetrica A: Transport Science,* 14(3), pp. 230-255.

Yu, H., Liu, P., Chen, J. & Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. *Accident analysis and prevention,* Volume 66, pp. 80-88.

Yu, R. & Abdel-Aty, M., 2013a. Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes. *Accident Analysis and Prevention,* Volume 58, p. 97– 105.

Yu, R. & Abdel-Aty, M., 2013. Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes. *Accident Analysis and Prevention,* Volume 58, p. 97– 105.

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

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

Zeng, Q. et al., 2019. Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors. *Accident Analysis & Prevention,* Volume 127, pp. 87-95.