




2022

## INCORPORATING SPEED INTO CRASH MODELING FOR RURAL TWO-LANE HIGHWAYS

Fahmida Rahman

University of Kentucky, fra228@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0001-8143-6540>

Digital Object Identifier: <https://doi.org/10.13023/etd.2022.322>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

### Recommended Citation

Rahman, Fahmida, "INCORPORATING SPEED INTO CRASH MODELING FOR RURAL TWO-LANE HIGHWAYS" (2022). *Theses and Dissertations--Civil Engineering*. 121.  
[https://uknowledge.uky.edu/ce\\_etds/121](https://uknowledge.uky.edu/ce_etds/121)

This Doctoral Dissertation is brought to you for free and open access by the Civil Engineering at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Civil Engineering by an authorized administrator of UKnowledge. For more information, please contact [UKnowledge@lsv.uky.edu](mailto:UKnowledge@lsv.uky.edu).

## **STUDENT AGREEMENT:**

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

## **REVIEW, APPROVAL AND ACCEPTANCE**

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Fahmida Rahman, Student

Dr. Mei Chen, Major Professor

Dr. Mei Chen, Director of Graduate Studies

INCORPORATING SPEED INTO CRASH MODELING FOR RURAL TWO-LANE  
HIGHWAYS

---

DISSERTATION

---

A dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy in the  
College of Engineering  
at the University of Kentucky

By  
Fahmida Rahman  
Lexington, Kentucky  
Director: Dr. Mei Chen, Professor of Civil Engineering  
Lexington, Kentucky  
2022

Copyright © Fahmida Rahman 2022  
<https://orcid.org/0000-0001-8143-6540>

## ABSTRACT OF DISSERTATION

### INCORPORATING SPEED INTO CRASH MODELING FOR RURAL TWO-LANE HIGHWAYS

Rural two-lane highways account for 76% in mileages of the total paved roads in the US. In Kentucky, these roads represent 85 % of the state-maintained mileages. Crashes on these roads account for 40% of all crashes, 47% of injury crashes, and 66% of fatal crashes on state-maintained roads. These statistics draw attention to the need to investigate the crashes on these roads. Several factors such as road geometries, traffic volume, human behavior, etc. contribute to crashes on a road. Recently, studies have identified speed as one of the key factors of crashes as well as the severity associated with them and indicated the need to incorporate speed into predicting crashes and severity. Such studies are limited for rural two-lane highways due to the lack of measured speed data in the past. This study fills this gap by utilizing widely available measured speed data on these roads and investigates the relationship between speed and crashes on rural two-lane highways.

This study collected crash, speed, traffic, and road geometric data for rural two-lane highways in Kentucky. Particularly for the speed, this study utilized GPS-based probe data. The speed data was integrated with the crash data and road attributes for the rural two-lane highways. This study utilized the speed measures directly calculated from the measured speed data and evaluated the effect of speed on the crashes of these roads. At first, this study investigated the effect of speed by incorporating average speed along with traffic volume and length in the crash prediction model for total number of crashes. A zero-inflated negative binomial model was utilized to account for the overdispersion from excess zero crashes in the dataset. From the model, a negative relationship was identified between average speed and number of crashes. One possible explanation is that rural two-lane roads with higher speeds tend to be those main corridors with better geometric conditions. Furthermore, the significance of speed in the model varies with the operating speed on these roads. This suggested considering speed as a categorizer to develop separate models for different speed ranges. Separating models based on speed provided improved prediction performance compared to an overall model.

Operating speed often reflects geometric conditions. Therefore, this study also evaluated how the change in the 85<sup>th</sup> percentile speed from one section to another road section affects the crashes of a road. The analysis showed that more crashes tend to occur when the 85<sup>th</sup> percentile speed differential between consecutive segments increases. However, further investigation showed that speed differential may not be a suitable indicator of identifying the locations with a high risk of crashes, rather it can be applied for design improvement of the roads.

Later, this study investigated spatial heterogeneity of the effect of speed in addition to other factors utilizing a geographically weighted regression model. The model

accounted for the geographical location of the data and helped to investigate the spatially varying effect of speed. The results from this model showed that the significance of speed can vary at different locations, which is not observed in the global model. In some regions, speed actually reflects the local geometric conditions of the roads. On the road with poor geometric conditions, crashes tend to be higher. The safety improvement strategies for these roads can focus on improving the geometric conditions such as providing shoulders, realigning the sharp curves, etc. Furthermore, speed seemed to increase crashes in some locations with good geometric conditions and low traffic volume. Speed was indeed a critical factor for these locations and safety countermeasures should be recommended considering the operating condition.

Utilizing measured speed data, this study also explored the effect of speed separately on KABC and PDO crashes for these roads. Separate models were developed for KABC and PDO crashes using a zero-inflated Poisson model form. Results from the models showed that speed had a positive relationship with KABC crashes, but a negative relationship with PDO crashes. For the KABC crashes, more KABC crashes tend to occur on high-speed roads. In contrast, PDO crashes tend to be higher on low-speed roads with poor geometric conditions. Furthermore, this study separated the models for each severity level using speed as a categorizer. The models developed at individual speed ranges revealed a varying effect of speed over the different speed ranges of these roads. For example, speed had a positive effect on KABC crashes of low and medium-speed roads, whereas it had a negative influence on crashes of high-speed roads. Further investigation of the study data showed that most of the low and medium-speed roads had poor geometric conditions (narrow shoulder and lane widths with the presence of sharp curves), whereas, high-speed roads had standard geometric conditions. Especially on low-speed roads, it is understandable that a crash can be severe when speed goes up under such restrictive geometric conditions of the roads. In contrast, on high-speed roads, the number of severe crashes tends to be low under standard geometric conditions. Additionally, separating models considering speed ranges provided 19% and 6.5% improvement respectively for KABC and PDO crashes compared to the overall models. Such models can help the agencies to adopt strategies for minimizing crashes at different severity levels based on the speed condition of the road.

This study further looked at the effect of speed using Random Forest model since it can deal with multicollinearity between explanatory variables and requires no assumptions on the functional form. After including all the traffic and geometric variables in the model, speed showed 11.5% importance. Compared to the traditional count model, the model provided a better fit with an improved performance of 13%. For better predictability, planning level safety analysis can utilize such machine learning model.

**KEYWORDS:** Rural Two-Lane Highways, Highway Safety Manual, Probe Speed, Zero Inflated Model, Data Mining, Geographically Weighted Regression

---

Fahmida Rahman

---

08/04/2022

---

INCORPORATING SPEED INTO CRASH MODELING FOR RURAL TWO-LANE  
HIGHWAYS

By  
Fahmida Rahman

Dr. Mei Chen

---

Director of Dissertation

Dr. Mei Chen

---

Director of Graduate Studies

08/04/2022

---

Date

DEDICATION

*To my parents, brother, teachers, and friends*

## ACKNOWLEDGMENTS

First, I would like to express my gratitude to the Almighty who has given me the wisdom, patience, and strength and has made me successful in the completion of this dissertation. His consistent kindness over me keeps me focused and confident to do this research.

I am immensely grateful to my research advisor, Dr. Mei Chen, who helped me with her constant guidelines and supported me throughout the course of this research. Her constructive feedback at various phases of this research assisted me in shaping up the deliverables and gaining insights.

I would like to thank my committee members Dr. Reginald Souleyrette, Dr. Nikiforos Stamatiadis, Dr. Cidambi Srinivasan, and Dr. Mei Chen for their evaluation and instructive reviews on my document. Their feedback helped me to improve the document substantially.

I am thankful to Dr. Eric Green, William Staats, and Dr. Chris Blackden for their assistance in data management and pre-processing works.

Sincere thanks go to my co-workers, especially Xu Zhang who supported me from the very beginning of my graduate program here at the University of Kentucky. He helped me thoroughly with the conceptual understanding of my research. Furthermore, Eugene Antwi Boasiako, Vedant Goyal, Riana Tanzin, Shraddha Sagar, Jawad Hoque were there to amend me with the challenging working environment here. Their perseverance and dedication to work encouraged me to make every possible attempt to achieve the goal. In addition, I am thankful to the departmental faculties and staffs for their time and effort in any situation of assistance.



I also would like to thank my friends Shaowli Kabir, Shafika Showkat Moni, Ayesha Siddiqua Dina, Monika Islam Khan, Sayma Afrin, Shaikh Bony, here in Lexington, who kept me supporting and have been always there when I needed them.

Lastly, I am grateful to my mom, dad, and younger brother for believing in me to pursue my goals. I would like to also thank all of my friends in Bangladesh for their love and inspiration.

## TABLE OF CONTENTS

ACKNOWLEDGMENTS .....	iii
TABLE OF CONTENTS.....	v
LIST OF TABLES.....	ix
LIST OF FIGURES .....	xi
CHAPTER 1. INTRODUCTION .....	1
1.1 <i>Background</i> .....	1
1.2 <i>Research Objectives</i> .....	3
CHAPTER 2. LITERATURE REVIEW .....	5
2.1 <i>Speed in Safety Performance</i> .....	5
2.1.1 Relationship between Crashes and Speed.....	5
2.1.1.1 Individual Driver-based Studies .....	5
2.1.1.2 Segment-based Studies .....	8
2.1.2 Summary of the Crash-Speed Relation Studies at Different Facility Types	16
2.1.3 Existing Gaps .....	27
2.2 <i>Additional Relevant Background Studies</i> .....	27
2.2.1 Speed Measures in Analyzing Crashes of Rural Two-Lane Highways .....	28
2.2.2 Analytical Methods for Crash Prediction .....	28
2.2.2.1 Statistical Models.....	29
2.2.2.1.1 Traditional Models.....	29
2.2.2.1.2 Spatial Models .....	30
2.2.2.2 Machine Learning Models .....	30
2.2.2.3 Summary .....	32
2.2.3 Studies Incorporating Speed Measures in Analyzing Crash Severity .....	32
2.3 <i>Literature Review Summary</i> .....	34
CHAPTER 3. DATA COLLECTION PROCESSING.....	37
3.1 <i>Data Sources</i> .....	37
3.2 <i>Data Pre-Processing</i> .....	37
3.3 <i>Summary</i> .....	40
CHAPTER 4. METHODOLOGY .....	41

4.1	<i>Potential Factors of Crashes</i> .....	41
4.2	<i>Methods for Variable Selection</i> .....	42
4.2.1	Pearson Correlation Coefficient.....	42
4.2.2	Spearman’s Correlation Coefficient .....	43
4.3	<i>Spatial Dependency Test</i> .....	44
4.4	<i>Modeling Approach</i> .....	45
4.4.1	HSM Method .....	45
4.4.2	Statistical Models.....	47
4.4.2.1	Traditional Count Models .....	47
4.4.2.1.1	Poisson Model.....	47
4.4.2.1.2	Negative Binomial Model.....	48
4.4.2.1.3	Zero Inflated Models.....	49
4.4.2.2	Spatial Count Models.....	53
4.4.2.2.1	Geographically Weighted Poisson Regression Model.....	53
4.4.2.2.2	Geographically Weighted Zero Inflated Poisson Regression .....	56
4.4.3	Machine Learning Model.....	58
4.4.3.1	Random Forest Model Calibration Process .....	59
4.4.3.2	Variable Importance from Random Forest Model.....	61
4.5	<i>Evaluation of Model Performance</i> .....	62
4.6	<i>Summary</i> .....	65
CHAPTER 5. INVESTIGATING SIGNIFICANCE OF SPEED .....		67
5.1	<i>Influence of Speed from Operational Context</i> .....	67
5.1.1	Objective.....	67
5.1.2	Dataset and Speed Variable Selection .....	67
5.1.3	Incorporating Speed for Better Performance .....	70
5.1.3.1	Speed Categorizer .....	72
5.1.3.1.1	Low-Speed Roads .....	72
5.1.3.1.2	Medium-Speed Roads.....	73
5.1.3.1.3	High-Speed Roads .....	77
5.1.3.2	Overall Performance .....	79
5.1.4	Findings and Significance of the Analysis.....	80

5.2	<i>Influence of Speed from Consistency Context</i> .....	81
5.2.1	Objective.....	82
5.2.2	Dataset and Variable Selection.....	82
5.2.3	Analysis and Results.....	87
5.2.3.1	Comparison with Models based on Speed Metric.....	92
5.2.4	Application and Limitations.....	96
5.2.5	Major Findings and Significance of the Analysis.....	98
5.3	<i>Summary</i> .....	99
CHAPTER 6. SPATIAL VARYING EFFECT OF THE FACTORS ON CRASHES ..		101
6.1	<i>Objectives</i> .....	101
6.2	<i>Dataset and Variable Selection</i> .....	101
6.3	<i>Spatial Autocorrelation Check</i> .....	106
6.4	<i>Analysis and Results</i> .....	106
6.4.1	Spatial Variation Analysis.....	109
6.4.1.1	AADT.....	110
6.4.1.2	Segment Length.....	111
6.4.1.3	Average Speed.....	112
6.4.1.4	Degree of Curvature.....	115
6.4.1.4.1	Data Analysis for Degree of Curvature:.....	117
6.5	<i>Major Findings and Significance of the Analysis</i> .....	124
CHAPTER 7.EFFECT OF SPEED AT DIFFERENT LEVELS OF CRASH SEVERITY		126
7.1	<i>Objective</i> .....	126
7.2	<i>Dataset and Variables</i> .....	126
7.3	<i>Analysis based on Traditional Count Models</i> .....	129
7.3.1	Model Development for KABC and PDO crashes.....	130
7.3.2	Severity Analysis at Different Speed Ranges.....	134
7.4	<i>Spatial Analysis</i> .....	146
7.4.1	Spatial Modeling and Results.....	146
7.4.1.1	AADT.....	150
7.4.1.2	Average Speed.....	152

7.4.1.3	Degree of Curvature.....	152
7.5	<i>Major Findings and Significance of the Analysis</i> .....	158
CHAPTER 8.	MACHINE LEARNING MODEL-BASED ANALYSIS.....	162
8.1	<i>Objectives</i> .....	162
8.2	<i>Dataset and Variables</i> .....	162
8.3	<i>Analysis and Results</i> .....	163
8.3.1	Model Comparison.....	168
8.4	<i>Findings and Significance of the Analysis</i> .....	170
CHAPTER 9.	CONCLUSION.....	172
9.1	<i>Summary</i> .....	172
9.2	<i>Study Limitations and Future Work</i> .....	177
APPENDIX.....		179
REFERENCES .....		180
VITA.....		189

## LIST OF TABLES

Table 1	Summary of the Segment-based Studies	18
Table 2	Default distribution for Crash Severity Level on Rural Two-Lane Highways (Source: HSM (7))	33
Table 3	Attributes Associated with the Segments	39
Table 4	Base Condition for Rural Two-Lane, Two-Way Highways (Source: HSM)	46
Table 5	Hyperparameters for RF Model Calibration	60
Table 6	Descriptive Statistics of the Explanatory Variables	68
Table 7	Model Parameters and Goodness-of-Fit	70
Table 8	Model for Low-Speed Category	73
Table 9	Model comparison for Medium-Speed Category	74
Table 10	AADT Categorizer-based Models and Comparison	76
Table 11	Model comparison for High-Speed Category	78
Table 12	Performance Comparisons	80
Table 13	Summary Statistics of the Variables	85
Table 14	Spearman's Correlation Test	87
Table 15	Parameter Estimates and Performance Measures	88
Table 16	Likelihood Ratio Test	90
Table 17	Mean Crash Rates for Different Design Categories	92
Table 18	Comparison between Speed Differential and Speed Metric-based Models	94
Table 19	Percentage of CURE within $\pm 2\sigma$ Boundaries	96
Table 20	Summary Statistics of the Variables	105
Table 21	Spatial Dependency of the Variables	106
Table 22	Variable Coefficients and Model Performance	108
Table 23	Class-wise Distribution of Degree of Curvature	118
Table 24	Summary Statistics for Degree of Curvature	118
Table 25	Distribution of Curve Class in More Homogenous Dataset	121
Table 26	Comparison of ZIP Models based on Degree of Curvature	122
Table 27	Statistics of Crash Rates for different curve classes	123
Table 28	Summary Statistics of the Road Attributes	129
Table 30	Multicollinearity Check	131
Table 29	Parameter Estimates KABC and PDO Models and Goodness-of-Fit	131
Table 31	Multicollinearity Check for Low, Medium, and High-Speed Roads	136
Table 32	ZIP Models for Low, Medium, and High-Speed Roads	137
Table 33	Performance Comparisons	142
Table 34	Spatial Dependency of the Variables	146
Table 35	Variable Coefficients and Model Performance	148
Table 36	Summary Statistics of the Road Attributes	163
Table 37	Best Combination of Hyperparameters	164

Table 38 Ranking of the Variables 165  
Table 39 Performance of the RF Model 168  
Table 40 Comparison of Model Performance 168  
Table 41 Variable Importance from RF and ZINB model 170

## LIST OF FIGURES

- Figure 1 Crash Involvement Rate vs Variation from the Average Speed [Source: Solomon (4)] 7
- Figure 2 Segment Aggregation 39
- Figure 3 Rural Two-Lane Segments in Kentucky 39
- Figure 4 Illustration of Monotonic and Non-Monotonic Relationships 44
- Figure 5 Demonstration of Geographically Weighted Regression Modeling Process 54
- Figure 6 Demonstration of 5-fold Cross-Validation 61
- Figure 7 Marginal Model Plots for Model (2) 71
- Figure 8 CURE Plots for Model (2) 72
- Figure 9 Marginal Model Plots for Medium Speed Roads 75
- Figure 10 CURE Plots with  $\pm 2\sigma$  for the Explanatory Variables in Medium Speed Model 76
- Figure 11 CURE Plots with  $\pm 2\sigma$  for the Models of Medium-Speed Roads 77
- Figure 12 Marginal Model Plots for High-Speed Roads 79
- Figure 13 HERE Link Issue 84
- Figure 14 Aggregation based on Degree of Curvature 85
- Figure 15 Consecutive Rural Two-Lane Segments in the Direction of Increasing Mile Points 86
- Figure 16 CURE Plots for Speed Differential -based Models 90
- Figure 17 CURE plots for Speed Differential and Speed Metric-based Models 95
- Figure 18 Application of the Analysis 98
- Figure 19 Spatial Distribution of Variables 104
- Figure 20 Correlation Analysis 105
- Figure 21 Kentucky (a) Regions, (b) Terrain, and (c) Area Type 110
- Figure 22 Spatial Distribution of the Coefficients for AADT 111
- Figure 23 Spatial Distribution of the Coefficients for Length 112
- Figure 24 Spatial Distribution of the Coefficients and Significance for Speed 114
- Figure 25 Variable Ranking from GWZIP Model 115
- Figure 26 Spatial Distribution of the Coefficients and Significance for Degree of Curvature 116
- Figure 27 Distribution of Degree of Curvature 117
- Figure 28 Crash Rate vs Degree of Curvature 119
- Figure 29 Before Aggregation 120
- Figure 30 After Aggregation 120
- Figure 31 Percentage of Zero and Non- zero crashes for K, ABC, and PDO 127
- Figure 32 Spatial Distribution of KABC and PDO Crashes 128
- Figure 33 CURE Plots for Average Speed Model 134
- Figure 34 Distribution of Shoulder Width, Lane Width, and Degree of Curvature over Low, Medium, and High-Speed Roads 141



Figure 35 CURE Plots for KABC and PDO Models	145
Figure 36 Spatial Distribution of the Coefficients for AADT and Effect of AADT Changes	151
Figure 37 Spatial Distribution of the Coefficients for Average Speed and Effect of Speed Changes	155
Figure 38 Spatial Distribution of the Coefficients for Degree of Curvature and Effect of Curvature Changes	157
Figure 39 Partial Dependence Plots	167
Figure 40 Comparison between predicted and actual number of crashes	169
Figure 41 Comparison of CURE Plots	170

## CHAPTER 1. INTRODUCTION

### 1.1 Background

Safety on rural roads is a serious concern in the United States (U.S.). A recent analysis based on the data released by Fatality Analysis Reporting System (FARS) showed that fatal crash rate (per 100 vehicle miles traveled) in rural areas was almost 1.7 times higher than for urban areas in 2020, although only approximately 19% of the U.S. population lives in rural areas (1; 2). These statistics draw attention to the crashes in rural areas.

A large portion of the rural areas in U.S. includes rural two-lane highways. Nationwide, the total length of paved roads is 4,000,000 miles, of which 80% are rural roads, and 85% of these rural roads are rural two-lane highways (3). Past analysis for Kentucky undertaken by Kentucky Transportation Cabinet (KYTC) identified that “rural two-lane highways account for about 85 % of the state-maintained mileage, however, only 34 percent of the vehicle miles are traveled. These roads account for 40% of all the crashes on state-maintained roads, 47% of injury crashes, and 66% of fatal crashes. Moreover, the fatal crash rate on rural two-lane highways is approximately twice the overall fatal crashes on all state-maintained roads ”(4). Furthermore, crashes on rural two-lane highways are recorded three times higher on horizontal curves than on tangent sections (5). Overall, all these records show that both the frequency and severity of crashes on rural two-lane highways need serious attention.

Many factors contribute to crashes on a roadway including road geometries, traffic volume, environment, speed characteristics, human behavior, etc. Among them, speed is often considered a major factor (6). The first edition of the Highway Safety Manual (HSM) includes safety performance functions (SPFs) to estimate annual crash frequency for multiple facility types including rural two-lane/two-way roads (7). The SPF equations incorporate traffic volume and length. These equations were developed for the base conditions. If there is a deviation from the base condition, crash modification factors (CMFs) are estimated and included in SPFs. This requires a detailed inspection of a large base condition list, for example, shoulder width and type, lane width, curve, grade,

driveway density, lighting, etc. Neither the SPF equations nor the CMFs consider speed as one of the factors.

Existing studies, which investigated the role of speed on crash prediction, confirmed the correlation between speed and crashes (6; 8-12) and suggested including speed as a variable in the model (12-18). However, most of these are performed for heavily traveled corridors such as arterials, interstates, state highways, multilane, etc.(12; 15-20). The relationship between speed and crashes on rural two-lane highways has been mostly explored in the context of geometric design consistency (21-25). Speed, in particular the 85<sup>th</sup> percentile speed, is used as an indicator of design consistency from one segment to another. Due to the lack of measured data, speed is often estimated using models (21-25).

In recent years, measured speed data have become widely available and especially abundant on higher functional class roads. Many studies used these datasets (e.g., probe vehicle data, GPS taxi data, loop detector data, etc.) directly to examine the relationship between crashes and measured speed (19; 20; 26-30). Some noted that including speed would enhance the performance of crash prediction models compared to the traditional method (27). In particular, a recent study by Dutta and Fontaine used measured speeds from two states to develop safety performance functions for rural interstates, multilane highways, and two-lane highways (31). They found that speed is significant at different severity levels as well as for the total number of crashes.

Regardless, in case of assessing the effect of speed on different levels of crash severity, other facility types received much attention (19; 27; 32-36), whereas, limited works were found, especially for rural two-lane highways (31; 37-39). As speed parameters, mainly speed limit and design speed have been explored for rural two-lane highways (37-39). For these roads, speed limit may not always capture the actual operating condition. All these indicate a research concern for looking into the different levels of crash severity on rural two-lane highways utilizing measured speeds.

In summary, crash occurrence and severity of rural two-lane highways require research attention to identify the role of speed, especially based on the measured speed dataset. This study attempts to address this research need by utilizing the availability of speed data from GPS-based probes.

## 1.2 Research Objectives

Previous section indicates that research on rural two-lane highways still requires attention to carefully investigate the effect of speed on crash occurrence and severity. With the advancements in GPS technologies, speed data availability has become better than before on these roads. In this study, the author utilizes the GPS-based probe speed data to estimate different speed measures and incorporate them into the crash prediction model for rural two-lane highways. The goal is to investigate the role of speed on the crashes of these roads utilizing measured data. The primary objectives of this research can be listed as follows.

- Investigate the effect of speed on the crashes of rural two-lane highways.
- Develop crash prediction models for the rural two-lane highways by integrating speed measures along with geometric and traffic factors. This includes exploring the effect of these variables on crashes utilizing both statistical (traditional and spatial modeling) and machine learning (ML) techniques.
- Explore whether speed influences crashes at different levels of severity. If speed is significant, incorporate it in predicting crashes at different levels of severity with other factors.

To develop the prediction models for the total number of crashes and number of crashes at different levels of severity, this study adopted both statistical models and data mining tools. These approaches evaluate the significance of speed along with other explanatory variables for crashes of rural two-lane highways. Furthermore, the crash prediction models at different severity levels provide an idea of whether the influence of speed varies over the different levels of severity. Additionally, the performance of the developed models is compared with the traditional model form that does not include speed factors.

This document is organized into nine chapters. Below are the contents of the chapters in brief.

- Chapter 1: An overview of the research statement and the objectives.
- Chapter 2: Review of existing literature focusing on the incorporation of speed measures in crash prediction model and related to the major objectives of this research.

- Chapter 3: Description of data sources and pre-processing.
- Chapter 4: Details on methodological approaches.
- Chapter 5 to Chapter 8: Model development and analysis of results in evaluating the effect of speed along with other factors.
- Chapter 9: Summary of the major findings and recommendations for future works

## CHAPTER 2. LITERATURE REVIEW

This chapter documents the existing studies that considered speed in developing crash prediction models and examines the influence of speed on crashes in addition to other factors. The review is separated into two major sections, followed by a summary.

The first section discusses the past efforts incorporating the speed in analyzing crashes of different facility types including rural two-lane highways. It provides an idea of the several speed measures used in crash predictions as well as how those speed measures were estimated. The findings about the relationship between speed and crashes are also documented. This section helped to identify the gaps from a broader perspective and set up the major goals of this study.

Once the broader research need of considering actual speed in crash prediction of rural two-lane highways is identified, the second section discusses the more relevant studies to further provide a background for the individual objectives of this study. The review includes identifying the speed measures investigated for rural two-lane highways in the existing studies, reviewing different analytical methods used for crash prediction, and identifying the studies that incorporated speed into severity analysis.

Lastly, a summary of the review concludes the gaps in the existing literature and provides insights into the importance of this research in addressing those gaps.

### 2.1 Speed in Safety Performance

#### 2.1.1 Relationship between Crashes and Speed

Existing literature that examined the effect of speed in crash analysis can be classified into two major categories. These are as follows:

- Individual driver-based studies
- Segment-based studies

##### 2.1.1.1 Individual Driver-based Studies

The driver-based studies mainly relate the difference between the pre-crash speed of a vehicle and the aggregated speed over a segment to the crashes. The pre-crash speed is mainly associated with the crash occurrence, whereas, the segment speed is derived

from a non-crash condition of the road. (11). Generally, the data source for pre-crash speed can be police reports.

Solomon first observed the relationship between the individual vehicle's pre-crash speed and crash risk based on 35 segments from rural areas (6). He noticed that the pre-crash speeds of the vehicles were either below or above the average speed of the segment during a non-crash situation. This can be shown as the U-shaped relationship between the speed deviation and crash involvement rate (Figure 1). However, the pre-crash speed records by the police report may not be always accurate and the average speed was assumed to be constant over a segment despite the changes in road geometries. All these limitations may result in an inaccurate estimation of the speed and crash relationship.

Similar to Solomon's study, Kloeden et al. experimented on pre-crash speeds of the vehicles involved in crash events to quantify the crash involvement rate (11). Unlike the police report data in Solomon's study, they determined the pre-crash speeds using computer-aided crash reconstruction techniques. The pre-crash speeds of the vehicles were compared with the other vehicles traversing at an average speed without being involved in crashes. It was observed that those vehicles were traveling faster than the average speed. The study further concluded that slow-moving vehicles were less prone to high crash risks. Therefore, the U-shape relationship is not supported by this study.

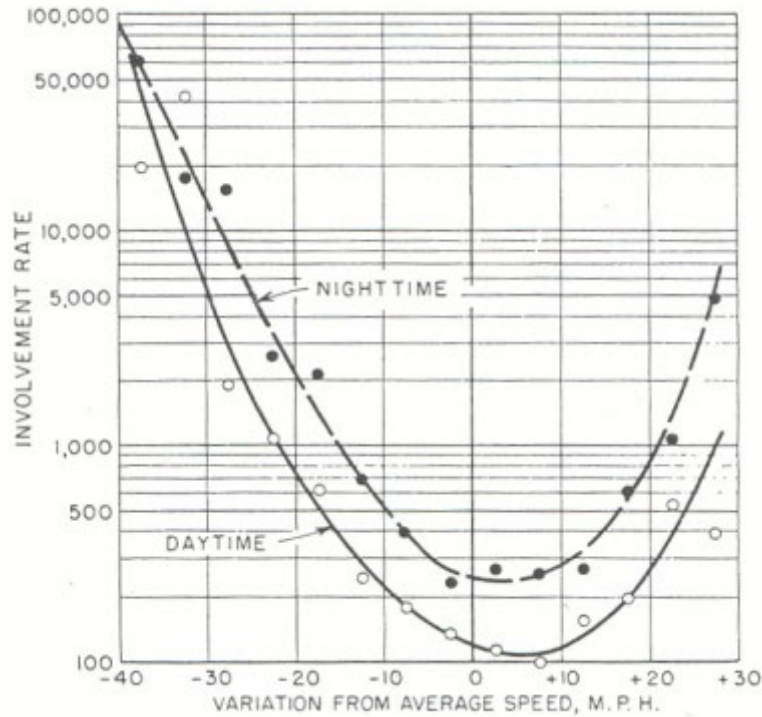


Figure 1 Crash Involvement Rate vs Variation from the Average Speed [Source: Solomon (4)]

Another individual driver-based study by Fildes et al. tested the relationship between vehicle's speed and crash risk for urban arterials and rural undivided highways in Australia (9). The speed data were collected by interviewing the drivers about their accident history during the past five years. The analysis based on the accident history reveals that drivers traveling above the 85<sup>th</sup> percentile speed had a higher crash risk, whereas, drivers traveling below the 15<sup>th</sup> percentile speed were less likely to be involved in a crash. Nonetheless, the speed data only belonged to the group of victims who survived the crashes (property damage crashes). Therefore, it was not possible to get the data for the fatal ones, and the study may not reflect the results for fatal crashes. Specific weak points of the study were (a) a small number of locations (two per road type) and (b) a small number of days for speed measurement (4–6 days per location).

West et al. collected self-reports on the driving behavior completed by the drivers traveling urban and motorways routes (40). The study verified the reports based on an observer who accompanied the drivers. The authors performed a multiple logistic



regression, which showed a positive relationship between the observed speed and self-reported crash involvement.

Maycock et al. utilized a case-control study to find a relationship between crash rates and speed measures (41). The experiment included all the UK roads. Individual driver's speeds were collected using a radar gun, and the crash history of each driver was collected based on a questionnaire survey. They found a positive association between crash rates and the individual's relative speed with respect to the average speed of the control vehicles on a road. It means that vehicles, driving at a speed higher than the average speed of a road, tend to have more crashes. However, they only got a 46% response from the survey. This sample may not be well-representative and may produce biased results.

Richards et al. obtained a relationship between crash risk and the individual vehicle's speed change during the crash occurrence (42). The study was performed in London. Individual vehicle's speed change during crashes (delta-v) was calculated by measuring the vehicles' residual crash and utilizing a forensic investigation of the damage. Findings from this study showed a positive association between delta-v and the risk of fatality (%).

In summary, the individual driver-based studies compared the speed of a vehicle involved in a crash with the prevailing speed of a segment. The studies, in general, implied that a deviation of speed from the average speed of a segment leads to high crash risk. However, data sources such as police reports may not be always reliable for the accurate representation of such speed and crash relationship.

#### 2.1.1.2 Segment-based Studies

It is not always easy to obtain the individual driver's speed right before the crash occurrence. Therefore, studies also started utilizing the aggregated speed measures of a segment to investigate the relationship between speed and crash occurrence. The aggregated speed measure reflects the operating condition of a road and can be determined based on traffic volume and geometric conditions of the road. Such speed measure partly reflects the geometric condition of a road.

Garber and Gadiraju explored speed-crash relationship for freeways and arterials in both rural and urban settings (10). As the speed variables, they considered average speed and speed variance. They collected 24 hours of speed data and aggregated it to calculate the average speed. The speed variance was estimated as the difference between design speed (from the highway log sheet) and posted speed limit. According to the ANOVA test, they found significant effects of average speed and the speed variance on the crash rate. Regardless, they concluded that the crash rates increased with increasing speed variance for all functional classes. Contrarily, the crash rate decreased with an increase in the average speed from a lower functional class to a higher functional class (such as interstate) road. The higher average speeds were due to the better geometric conditions of the higher functional class roads.

Anderson et al. investigated how the geometric design consistency parameter influences the number of crashes on the horizontal curves of rural two-lane highways (43). The reduction of the 85<sup>th</sup> percentile speed between adjacent tangent and curve or curve and tangent was used as the consistency parameter. The authors used a speed prediction model to calculate the 85<sup>th</sup> percentile speed of a segment. After fitting a Poisson regression model for the number of crashes as a function of traffic, geometric, and speed reduction parameters, the study found a strong positive relationship between the number of crashes and the speed reduction on horizontal curves.

Another study by De Oña et al. used a speed prediction model for calculating the 85<sup>th</sup> percentile speed of the two-way rural two-lane horizontal curves in Spain (23). The study data included 3 years of crash data (2006-2008). From the analysis, it was concluded that the reduction in 85<sup>th</sup> percentile speed between the consecutive elements of horizontal curves significantly affects the crash frequency of this type of road. Ng and Wai utilized an operating speed prediction model to calculate the 85<sup>th</sup> percentile speed for rural two-lane (horizontal curves) in British Columbia, Canada (25). The finding was that the larger difference between the operating and design speed, the more collisions are expected to occur.

Cafiso et al. estimated the 85<sup>th</sup> percentile speed, average speed, and standard deviation based on the regression models for two-lane rural roads located in Italy (22). For accident data, 5 years of data were collected associated with homogenous sections.

The authors showed that an increase in the standard deviation of speed can increase accidents. The increase in standard deviation is likely to occur at the transition of curves, for example, transitioning from a long tangent section to a sharp horizontal curve. While driving in this transition, drivers usually become more cautious. Moreover, a value of speed differentials, i.e., difference between the 85<sup>th</sup> percentile speeds of the two horizontal elements of rural two-lanes, higher than 10 km/hr. is associated with an increase in the number of accidents.

Using a speed prediction equation, the study by Wu et al. found a positive association between design inconsistency (the difference between the 85<sup>th</sup> percentile and design speed) and number of crashes per year for horizontal curves (44). Kononov et al. developed a Neural Network model-based SPF for urban freeways and multilane highways (16). The SPF showed a good fit from the cumulative residuals (CURE) plots (crash rates vs. annual average daily traffic) with a sigmoidal shape. The authors related the changes in flow, speed, and density with the changes in the sigmoidal shape of the SPF using the Highway Capacity Model. Critical density point and supercritical density point were identified based on the changes in the slope of the SPF's shape. They measured the speed related to those two critical points. The analysis showed that the number of crashes moderately increased when the freeway segments were operating at a free-flow condition. However, after reaching a critical density point, the slope of the SPF got steeper indicating a sharp increase in crashes with increasing annual average daily traffic (AADT). As the density increases with increasing traffic, a supercritical point was found in the SPF reflecting a high level of congestion with decreasing operating speed. After reaching the super-critical point, the crash rates tended to be lower than in the critical zone.

A recent study by Llopis-Castelló et al. was performed on 71 homogenous segments of rural two-lane highways in Italy (24). They also used the speed prediction models to calculate the 85<sup>th</sup> percentile speed. They concluded that inconsistency among the 85<sup>th</sup> percentile speed profiles causes an increase in crash frequency.

Another recent work by Dutta and Fontaine was based on rural four-lane roads (45). In their study, loop detectors were used to collect hourly speed data and hourly volume. Hourly crash records were also obtained for their analysis. The results indicated

that speed is negatively correlated with crashes, i.e. lower average speed (congestion) results in higher crash frequencies. However, datasets for this study were mostly from those locations that could not capture a broader variation in the traffic conditions.

Kweon and Kockelman experimented on major highways located in Washington State (17). Speed data were collected for 5-minute intervals using loop detectors. The authors calculated speed measures such as average speed and variance for five different time periods: the whole day, morning peak, morning off-peak, afternoon peak, and afternoon off-peak. Speed limit was also included in the analysis. For each time period, they calibrated average speed model and variance model utilizing measured speed data. The best models were the afternoon average speed and afternoon speed variance. Hence, they used the afternoon average speed and variance generated by these models in the crash prediction model along with the speed limit. The crash prediction model was developed based on 4-years of crash data associated with homogenous segments. The study found that an increase in speed limit causes a decrease in non-fatal crashes. Nevertheless, the speed limit was not significant for fatal crashes, due to the lack of sufficient variation in data as well as 99% of the data not including fatal crashes.

Taylor et al. used fixed sensor speed data for urban single carriageways in the UK (18). These roads were linked to a 1590 injury crash record. The study developed a non-linear regression model and showed that crash frequency increases with average speed. Initially, the relationship between average speed and crash frequency was negative. It was due to masking, which means other unaccounted variables (flow, pedestrian activity) were strongly correlated with the crash frequency. After taking those variables into account, the relationship became positive for urban roads. However, it could not be solved for rural roads even after accounting for those unobserved variables.

Kockelman and Ma collected loop detector-based speed data for the freeways of Southern California (12). They aggregated the speed data to obtain speed measures such as average speeds and speed variances for within the lane, across the lane, and total segment. The crash data were from 1 month period with a total record of 744. The authors developed multiple least-squared regression models and binomial models considering all the speed measures separately. The analysis showed no evidence of the relationship between the speed measures and crash occurrence. The limitations of this

study involved data accuracy issues, for example, errors in the crash reports, speed data aggregation along and across the segment, etc.

Based on the hypothesis that one or more speed-related measures are the determinant of crash risk for highways (such as interstates, expressways, and two-lane highways), Stout developed logistic regression models for estimating the probability of crashes (46). He tested different speed measures such as mean speed, speed variance, speed dispersion (the difference between the case hour the 85<sup>th</sup> percentile and mean speeds), and speed departure (the difference between the case hour the 85<sup>th</sup> percentile speed and the speed limit) in the model. The speed measures were estimated using an aggregated speed dataset collected by the automatic traffic recorders (ATR) from 1998 to 2003. The results from the model showed that speed is not the main factor for the crash risk. Even though speed variation can be indicated as one of the factors for interstates and expressways, it was not specified for two-lane highways. The research further mentioned that the aggregated speed data from ATR may not be suitable for estimating speed variance.

Finch et al. utilized the main roads from rural areas in Finland, Denmark, Switzerland, and the United States to conduct a meta-analysis (47). In a meta-analysis, data from different studies are combined to observe the common effects of the variables. Using this analysis, the authors found that a reduction in speed limits causes a reduction in the average speed of a road segment. Later, a linear regression model was fitted to see how changes in average speed affect the crash rate. They observed that crash rates are positively correlated with average speed, and a 1 mph increase in average speed causes a 5% increase in crash rates.

Baruya developed a crash prediction model with speed, flow, and geometric parameters for the European roads (48). He investigated the effect of average speed on the crash frequency, where the average speed was estimated from a regression model. According to the study's findings, higher crash frequency is associated with a lower average speed. Congestion and road environment would be the reasons for the lower average speed. However, this study performed a cross-sectional study, which does not allow for assessing the effect of an individual variable. Moreover, only 3 to 4 countries

were included in this analysis assuming that a similar relationship should exist in other countries of the UK. Therefore, the relationship remained unverified for other countries.

Pei et al. conducted a cross-sectional study on the freeway segments of Hong Kong (28). They collected 4 hours of speed data (30-sec epoch) for 3 months using GPS taxis. Based on these data, they calculated mean speed and standard deviation of the speeds. Crash data were also collected for the same 3-month period. The analysis concluded that higher mean speed leads to lower crash frequencies in terms of distance exposure. In contrast, crash severity was positively related to the mean speed. However, standard deviation of speed did not show any relationship with crash frequencies or crash severity. One possible reason was that the standard deviation parameter was not representative of the speed variability for a mixed traffic condition in their study. Regardless, the study had data limitations. Only 4 hours of speed dataset was used to calculate the speed measure, which may cause a sampling bias.

Wang et al. also did a cross-sectional study focusing on urban arterials (49). They concluded that mean speed is positively correlated with crash frequency. After quantifying the relationship, an increase of 10 kmph in mean speed would cause a 3% increase in crash frequency. Najjar and Mandavilli studied rural and urban state roads in Kansas by incorporating average speed limit as an operational condition in their crash prediction model (50). The study used an Artificial Neural Network (ANN) based data mining approach to observe the contribution of speed limit on the crash rates. They found that rural two-lane highways and urban expressway networks had the highest crash rates in rural and urban categories, respectively. However, some of the results for rural two-lane highways were not consistent with the existing literature or engineering judgment. For example, they developed SPF for similar shoulder widths (99% of the segments with 10 ft. width) with different pavement types, where the results were different even though the shoulder widths were the same. There was no explanation for this result. Due to these limitations, the Kansas Department of Transportation (DOT) did not apply the ANN model for practical purposes.

Banihashemi et al. conflated 2011-13 crash data with 2013 GPS probe speed data for urban interstates and arterials to explore the influence of speed on crash severity (19). Their assumption was that crash severity is affected by speed differentials (the difference

between the 85<sup>th</sup> percentile speed during off-peak and the speed limit). The study identified higher speed differentials as a reason for lower severity, which was counter-intuitive. This contradiction could be due to aggregating crashes from both directions. In addition, they linked three years of crash data with one year of speed data where the attributes provided by Roadway Information Data or National Performance Management Research Data Set (NPMRDS) travel times may not be consistent over the period. This may result in contradictory findings. Another reason for the counterintuitive result could be not considering additional geometric factors for developing the crash prediction model.

Wang et al. utilized GPS taxi data to calculate mean speed and speed variation (20). They integrated these speed measures along with other traffic characteristics in developing a hierarchical Poisson log-normal model for predicting crash frequency on the urban arterials in Shanghai, China. After analyzing the effect of these measures on crash frequency, they concluded that the crash frequency of a segment increased with higher mean speed and speed variation. This finding helped them in policy-making for speed management in Shanghai. They further quantified the contribution of speed to crashes. They showed that a 1% increase in mean speed is related to a 0.7% increase in crash frequency, and a 1% increase in speed variation is associated with a 0.74% increase in crash frequency. However, it was concluded that the impact of speed is less for the arterials than the freeways and rural roads as confirmed, which can also be confirmed by Elvik's meta-analysis (51).

Stipancic et al. experimented on the relationship between macroscopic traffic flow surrogate safety measure (SSM) and crash frequency as well as severity for different road types in Canada (e.g., Motorway, primary, secondary, tertiary, and residential) (52). To calculate different measures like congestion index (CI), average speed (V), and the coefficient of variation of speed (CVS), they used GPS data and large usage-based insurance data. Results showed that CI and CVS were both positively correlated with crash frequency. In terms of severity, an increase in CVS was related to a higher number of fatal and injury crashes. Unlike CI and CVS, the average speed was negatively correlated with the crash frequency.

A pilot project for observing the speed-safety relationship on rural roads was conducted by Das et al. (31). One of the research questions was to identify whether different speed parameters contribute to crash occurrence or not. The speed (estimated from NPMRDS travel time data) and crash datasets were obtained for Washington state and Ohio state for the year 2015. Their study focused on the crashes of rural roads consisting of interstates, multilane and two-lane highways. Interestingly, their developed crash prediction models for these roads were based on bi-directional crashes, speed, and segment attributes. The models were for annual –level crashes prediction and daily-level crash prediction, and each of these levels was separated by different crash severity level i.e., total KABCO (K= Fatal, A= Incapacitating Injury, B=Non-incapacitating Injury, C=Minor Injury, O= Property Damage Only) crashes, KABC crashes and Property Damage Only (PDO) crashes. The reason to separate the models by severity level was to explore how the effect of variables differs based on the severity level. The major findings after analyzing the annual-level crash prediction models for rural two-lane highways showed that average operating speed difference on weekdays and weekends was positively correlated with PDO crashes for both states, whereas, standard deviation in hourly operating speed was positively associated with the KABC crashes for both states. The daily-level crash prediction model further depicted that daily average operating speed was positively related to only KABCO crashes, and standard deviation of daily average operating speed was positively associated with both KABC and KABCO crashes.

A case study undertaken by Ederer et al. explored the relationship between percentile speeds and crashes (29). For the estimation of percentile speeds, they used probe vehicle speed data collected for the arterials in Atlanta. Based on the analysis, the authors suggested using the difference between the 85<sup>th</sup> percentile and the median speed as the safety performance metric since it showed a strong positive relationship with the expected number of crashes per segment.

Hutton et al. utilized SHRP2 Naturalistic Driving Study Data and Roadway Inventory Database to investigate the association between speed and crash frequency for urban and sub-urban arterial segments (14). Among the different speed measures they experimented with, higher speed variance caused higher crash frequency, and average speed depicted a negative correlation with crash frequency.



Llopis-Castelló et al. evaluated the effectiveness of the jurisdiction-based SPFs and SPFs including consistency parameters compared to the HSM method for rural two-lane highways (53). One of the consistency parameters was the difference between inertial operating speed and the 85<sup>th</sup> percentile speed. The 85<sup>th</sup> percentile speed profile was determined using the speed model by Ottesen and Krammes for curves and Polus et al for tangents (54; 55). After analyzing CURE plots, SPF based on consistency parameters provided the most accurate results. They concluded that SPF with consistency parameters is more accurate, includes interaction between infrastructure and human behavior, is not entirely dependent on field data collection, is easier to apply, and is more practical in terms of highway engineering.

Igene and Ogirigbo utilized speed differential model proposed by Abdelwahab et al. (56; 57). They included speed differential in the crash prediction model and found that higher speed differentials cause higher crashes. In their case, design consideration based on design speed had proved to be inadequate as some segments showed poor design from the 85<sup>th</sup> percentile speed differential, which was not identified by the difference between the 85<sup>th</sup> percentile speed and design speed. They recommended using the driver's operating speed instead of design speed for road design.

Gemechu and Tulu evaluated whether design consistency measures can also help in crash prediction in addition to identifying geometric inconsistencies (58). To calculate the operating speed-based design consistency measures, they determined the 85<sup>th</sup> percentile speed based on spot speeds observed at the center of curves and midpoint of tangents. They developed crash frequency models separately for each design consistency measure. They found the design consistency measures as significant in each model, and the speed differential was positively related to the number of crashes. Based on their study, the highest crashes in the poor design category indicate safety is related to design consistency.

### 2.1.2 Summary of the Crash-Speed Relation Studies at Different Facility Types

Since this research focuses on integrating segment-level speed measures for crash prediction of rural two-lane highways, the purpose of this subsection is to summarize the

segment-based studies, especially in terms of facility types and sources of speed. Previous subsection discussed such studies in detail focusing on how the segment-level speed measures interact with crash occurrence. This subsection briefly summarizes those studies with their specific research focus, facility type, study area, speed data source, and speed measures. Table 1 presents the summary. This will help to understand the gaps in the literature and the need for this particular study, which is discussed in the next subsection.

Table 1 Summary of the Segment-based Studies

<b>Year</b>	<b>Study</b>	<b>Research Focus</b>	<b>Facility Type</b>	<b>Study Area</b>	<b>Speed Data Source /Estimation</b>	<b>Speed Measure(s)</b>
1994	Finch et al. (47)	Effect of average speed change on the change in crash rates after decreasing the speed limit	Rural Roads, Main Lanes	Finland, Denmark, Switzerland, and the U.S.	Meta-analysis database	Change in Average Speed
1998	Baruya (48)	Effect of average speed on crash rates along with cross-sectional attributes	Rural Single Carriageway	UK	-	Average Speed*
1999	Anderson et al. (43)	Effect of design consistency on crash frequency	Rural Two-Lane Highways	Minnesota, New York, Oregon, Pennsylvania, Texas, and Washington in the U.S.	Fitzpatrick's Speed Model	The 85th percentile speed difference between two successive segments
2000	Garber and	Effect of average speed, the standard deviation of	Urban and Rural Highways	Virginia, U.S.	Sporadic monitoring	Average Speed and Standard Deviation

	Ehrhart (59)	speed, and the flow parameters on crash rates				
2000	Taylor et al. (18)	The combined effect of average speed and speed variation on crashes per year	Urban Single Carriageway	UK	Spot Speed at a fixed location	Average Speed, Coefficient of Variation, and difference between Average Speed and Speed Limit
2000	Anderson (21)	Effect of operating speed on crash rates of horizontal curves	Rural Two-Lane Highways	New York, Texas, and Washington in the U.S.	Ottesen and Krammes operating speed model	The 85th percentile speed difference between two successive segments
2002	Taylor et al. (60)	Effect of average speed on crash frequency using homogenous segmentation	Rural Single Carriageway	UK	Automatic equipment at each location	Average Speed
2002	Ng (25)	Effect of geometric design consistency on crash frequency per year	Rural Two-Lane Highways	British Columbia, Canada	Operating Speed Model	The 85th percentile speed difference between two successive

						segments, and the difference between the 85th percentile speed and the design speed*
2005	Kweon and Kockelman (17)	Effect of speed limit changes on average speed, hence, on crash severity	Interstate and State Highway	Washington, U.S.	Loop Detector	Speed Limit, Average Speed, and Standard Deviation
2007	Kockelman and Ma (12)	Effect of average speed and speed variation on crash probability	Interstate and State Highway	California, U.S.	Loop Detector	Average Speed and Standard Deviation
2009	Najjar and Mandavilli (50)	Effect of the speed limit on crash rates	Rural and Urban	Kansas, U.S.	Posted Speed Limit	Speed Limit
2010	Council et al. (61)	Comparing speed-related crashes with respect to total crashes	All type	North Carolina and Ohio in the U.S.	-	Speed Limit

2010	Cafiso et al. (22)	The combined effect of design consistency parameters, roadway features, and exposure on the number of crashes	Rural Two-Lane Highways	Italy	Equation of the 85th percentile speed, average speed, and standard deviation	The 85th percentile speed difference between two successive segments and Standard Deviation of the 85th percentile speeds
2011	Bornheimer (62)	Effect of the speed limit on total crashes	Rural Two-Lane, Two-Way Roads	Kansas, U.S.	Posted Speed Limit	Speed Limit
2011	Dell'acqua and Russo (63)	Effect of average speed on crash frequency	Rural Two-Lane Highways	Italy	-	Average Speed
2011	Kononov et al. (16)	Effect of operating speed on SPF	Urban Freeways, Multilane Highways	California and Colorado in the U.S.	Operating Speed Model	The 85th percentile speed

2012	Pei et al. (28)	Effect of average speed and speed variation on crash risk based on time exposure	Urban Roads	Hong Kong, China	GPS Taxis	Average Speed and Standard Deviation
2012	De one and Garach (23)	Effect of geometric design consistency on crash severity	Rural Two-Lane, Two-Way Roads	Spain	Operating Speed Model	The 85th percentile speed difference between two successive segments
2013	Quddus (33)	Effect of average speed and speed variation on crash frequency	Major Arterials	London, UK	Highways Agency (HA)	Average Speed and Standard Deviation
2013	Wu et al. (44)	Effect of design consistency on crash frequency	Highways	Pennsylvania, U.S.	Operating Speed Model	The 85th percentile speed difference between two successive segments
2018	Llopis-Castelló	Effect of geometric design consistency on total crashes	Rural Roads	Italy	Operating Speed Model	Difference between the 85th

	et al. (24)					percentile speed profiles
2018	Wang et al. (20).	Effect of mean speed and speed variation on crash frequency	Urban Arterials	Shanghai, China	GPS Taxi data	Mean Speed and Speed Variation
2019	Banihas hemi et al. (19)	Effect of operating speed and speed limit on crash severity	Urban Interstates and Major Arterials	Washington, Florida, New York, Pennsylvania, Indiana, and North Carolina in the U.S.	HERE GPS data	Difference between the 85th percentile speed at off-peak period and the Speed Limit
2019	Dutta and Fontaine (45)	Effect of average speed on total crashes	Rural Four-Lane Highways	Virginia, U.S.	Loop Detector	Average Speed
2020	Das et al. (31).	Effect of several speed measures on number of crashes	Rural Roads	Washington and Ohio in the U.S.	NPMRDS travel time data	Average hourly speed, Average hourly speed during non-peak and non-event



						<p>periods, Standard deviation of hourly operating speeds, Standard deviation of monthly operating speeds, Differences in the operating speeds during weekdays and weekends</p>
2020	Ederer et al. (29)	Effect of percentile speeds on the number of crashes per segment	Arterial	Atlanta, U.S.	Probe speed data	<p>The 15<sup>th</sup> percentile speed, Median speed, the 85<sup>th</sup> percentile speed, Difference between Median speed and the 15<sup>th</sup> percentile speed, Difference between the 85<sup>th</sup></p>

						percentile speed and Median speed
2021	Llopis-Castelló et al. (53)	Effectiveness of jurisdiction-based SPFs and SPFs including consistency parameters compared to the HSM method	Rural Two-Lane Highways	North Carolina, U.S.	Ottesen and Krammes and Polus et al Operating Speed Model	Difference between inertial operating speed and the 85 <sup>th</sup> percentile speed
2021	Igene and Ogirigbo (57)	Evaluation of the geometric design Consistency and road safety utilizing operating speed	Rural two-way single carriageway	Nigeria	Abdelwahab et al. Speed Differential Model	Absolute difference in the 85th percentile speed between two successive segments
2021	Gemechu and Tulu (58)	Evaluation of design consistency measures for crash prediction	Rural Two-Lane Highways	Ethiopia	Spot Speed at mid-segment	The 85th percentile speed difference between two successive segments, and the difference between

						the 85th percentile speed and the design speed
--	--	--	--	--	--	--

**\*Note:**

**Average Speed:** The summation of the instantaneous or spot-measured speeds at a specific location of vehicles divided by the number of vehicles observed. (MUTCD 2010)

**Design Speed:** The design speed is a selected speed used to determine the various geometric design features of the roadway (AASHTO Green Book)

**Operating Speed:** Operating speed is the speed at which drivers are observed operating their vehicles. The 85th percentile of the distribution of observed speeds is the most frequently used descriptive statistic for the operating speed associated

### 2.1.3 Existing Gaps

From Table 1, several studies used segment-based speed measures for different facility types like interstates, multilane highways, urban roads, arterials, etc. (12; 15-20). A variety of speed measures were experimented into the crash prediction models for these facilities. With the advancement in modern technology, both loop detector and GPS-based data have been utilized to investigate the relationship between crashes and speed measures. These facilities got attention over time by using good quality data and including speed measures along with other geometric attributes in the model.

However, rural two-lane highways still require serious attention. While a few studies used actual speed data (31), most of the studies primarily utilized prediction models to estimate average speed, standard deviation, and the 85<sup>th</sup> percentile speed. (21-24; 43; 44; 62; 63). These estimated speed measures may not reflect the actual operating condition on these roads. As a result, there is a necessity to investigate the speed and crash relationship for this facility based on the speed measures calculated from measured data (e.g., high-frequency GPS data) rather than from models. Moreover, the existing literature body still lacks significant work for rural two-lane highways in terms of relating crash severity with speed. Therefore, this research attempts to address these gaps by investigating the effect of speed measures on the number of total crashes and crashes at different severity levels for rural two-lane highways.

## 2.2 Additional Relevant Background Studies

The review in the previous section provides an idea of the existing research gap from a broader perspective. It helped to identify that rural two-lane highways still require attention to evaluate the effect of measured speed on crashes. This study set the major research objectives (Section 1.2) based on that. This section briefly discusses the most relevant studies that can provide additional background for each of these research objectives. In this way, the significance of the specific research objectives under this study can be understood further.

### 2.2.1 Speed Measures in Analyzing Crashes of Rural Two-Lane Highways

For rural two-lane highways, the relationship between speed measures and crashes has been mostly explored in the context of geometric design consistency (21-25). Speed, particularly the 85<sup>th</sup> percentile speed, was used as an indicator of design consistency between consecutive segments. The 85<sup>th</sup> percentile speed was estimated mainly using previously developed speed models (22; 23; 25; 53; 57; 64; 65) and required calibration using speed data (24; 44). The speed data used for developing these models are primarily spot speed collected with a radar gun or laser gun (54; 55; 64). Using spot speed collected at mid-point (assuming constant speed over the segment) of the section for speed estimation may be questionable (66). In reality, speed fluctuates over the section. Spot speed may fail to capture this and may not result in an accurate estimation of the speed measures. It may further affect analysis related to evaluating the role of speed on crashes of rural two-lane highways. Limited studies were found to utilize a complete set of measured speed data in developing speed models or in directly estimating speed measures before incorporating speed in crash prediction models of rural two-lane highways (67).

In summary, existing works mainly looked at the 85<sup>th</sup> percentile speed, especially as a design consistency measure, and incorporated it into the crash prediction model for rural two-lane highways. In addition, the measure was estimated either based on a model or spot speeds. This particular study tries to explore other speed measures (for example, average speed) for crashes of rural two-lane highways while utilizing a ubiquitous source of speed dataset. Using the complete set of measured speed in estimating different speed measures (including speed differentials) can offer a more complete picture in analyzing the crashes of these roads.

### 2.2.2 Analytical Methods for Crash Prediction

Earlier research assumed a linear relationship between traffic volume and crash frequency. Later, it was observed that crash frequencies are non-linearly correlated with the traffic volume and segment length. With time, other geometric and traffic conditions were explored for crash prediction. Research also started to include speed characteristics in the crash prediction model for different facility types. To explore the relationships

between the explanatory variables and crashes, different statistical and machine learning models have been utilized. This subsection summarizes the different modeling techniques adopted in existing studies related to crash prediction.

### 2.2.2.1 Statistical Models

In analyzing crashes, statistical models are fitted based on historical data to capture the relationship between crashes and other factors. Selection of the model may depend on how well the data fits with the functional form of the model. Existing literature shows the use of several statistical models for crash prediction. These can be separated into types as below:

- Traditional Models
- Spatial Models

#### 2.2.2.1.1 TRADITIONAL MODELS

According to Hauer (68), the following additive and multiplicative are generally used for predicting crashes of a road segment.

$$\text{Additive Form: } Y = L \times (\beta_0 x_0 + \beta_1 x_1 + \dots + \beta_n x_n) \quad (1)$$

$$\text{Multiplicative Form: } Y = L \times (\beta_0 x_1^{\beta_1} x_2^{\beta_2} \dots \dots) \quad (2)$$

$$\text{Exponential base Multiplicative Form: } Y = L \times (e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2} \dots) \quad (3)$$

Where, Y is the expected number of crashes, L is the segment length,  $x_n$  is the  $n^{th}$  explanatory variable, and  $\beta_n$  is the  $n^{th}$  regression coefficient. The additive model is best suited for evaluating the effect of point attributes such as the presence of driveways or narrow bridges, whereas, multiplicative models are appropriate for assessing the effect of segment-related attributes like lane width or shoulder width on crash occurrences (68). In addition, studies used these additive and multiplicative model forms to observe how speed affects crashes on a road segment (23; 43; 60).

Generalized Linear Modeling (GLM) approach is also used in quantifying the relationship between crash occurrence and road attributes (16; 22). This approach assumes a distribution from the exponential family for the crash occurrence. The popular

GLM technique includes Poisson, Negative Binomial (NB), Zero Inflated Poisson (ZIP), Zero Inflated Negative Binomial (ZINB), etc. These are widely used in predicting number of crashes and analyzing the contributing factors of crashes (17; 22-24; 43; 45; 62; 69-73). Other modeling techniques such as least-square linear regression, multivariate models, and random effect parameter models are also utilized for crash modeling (12; 37; 39; 44; 74; 75).

In general, all the above modeling approaches assume a stationary pattern of the crash data as well as constant effect of these variables over the spatial domain. These models estimate an average coefficient value for each explanatory variable of crashes.

#### 2.2.2.1.2 SPATIAL MODELS

While the traditional models assume a constant effect of the explanatory variables, in reality, the effect may show spatial heterogeneity considering the spatial dependency of crashes and the road attributes (76-80). To capture the spatial heterogeneity, studies utilized spatial modeling techniques such as geographically weighted regression (GWR) models (81). GWR models received significant attention in traffic safety analysis as a diagnostic tool (77; 78; 82-94).

The diagnostic power of this tool has helped to understand the spatial effect of different factors (for example, geometrics, traffic condition, land use, socio-demographics, etc.) on crashes particularly the macro-level crashes analyzed at the spatial units like traffic analysis zone (TAZ) or county (78; 82-87; 92). In those studies, GWR approach revealed significant varying relationships existing at different locations and provided a better understanding of the critical parameters of crashes for different regions. They utilized such findings in developing localized safety improvement policies and recommendations.

#### 2.2.2.2 Machine Learning Models

ML models have become popular in addressing multicollinearity issues and providing better performance in predicting crashes (16; 95; 96). These models are also applied to identify the important variables for crashes. For rural two-lane highways, Wei et al. used eXtreme Gradient Boosting to classify the short-duration crash occurrence

(97). To further investigate the relationship between the explanatory variables and the predicted crashes, they applied an artificial intelligence technique (SHapley Additive explanation) and found length, AADT, average visibility, daily precipitation, speed variation, etc. as the important variables for the crash occurrence. Wen et al. utilized different ML models in predicting run-off-road (ROR) crashes for highways (98). They revealed some important factors such as length, AADT, number of lanes, degree of curvature, etc. for ROR crashes. According to their observations, a complex ML model, for example, a random forest (RF) model, works better in capturing the associations and providing accurate predictions than a simple ML model, such as a classification and regression tree (CART). Zhang et al. applied an ensemble machine learning technique to improve the predictive performance of crash frequency model (99). They also identified the most significant factors for crash frequency, which included AADT, number of lanes, segment length, shoulder width, lane width, etc.

Studies also applied the ML models to prioritize the variables, which can ease data collection efforts. For example, Saha et al. utilized RF model to prioritize HSM variables for urban and suburban arterial roads since the detailed data required by HSM may not always be available (100). After investigating the variable importance, they found traffic volume, roadside object density, and minor commercial driveway density as the top-ranked variables. They also observed that some variables which HSM considers as the less important variables, such as roadside object density can fall into the list of top variables. Due to the same reason of data unavailability for HSM variables, another study by Saha et al. utilized a Boosted Regression Tree (BRT) approach to prioritize the HSM variable list based on their importance in the prediction (101).

Some studies applied tree-based models such as CART or RF to screen out the important variables for crash prediction and included the variables in the black box models like support vector machines (SVM) (102; 103). Overall, literature shows a growing application of different ML models, for example, decision jungle, nearest neighbor classification, decision tree, neural network, RF, DVM, BRT, Cubist, etc. in developing crash prediction models as classification or regression and identifies the robustness of these models considering the predictive capability (16; 95; 96; 104-106).



### 2.2.2.3 Summary

Overall, count models such as Poisson, NB, ZIP, ZINB, etc. are widely used in crash analysis of a segment. These modeling approaches can offer interpretability of the effect of a factor and transferability of the models. However, they may not capture spatial heterogeneity in the effect of the factors if spatial dependency is present in the dataset. Such issues can be further addressed by utilizing spatial models such as GWR method. These techniques were mostly applied in zonal or county level crash analysis rather than segment level crashes. This study utilizes the GWR tools to analyze the spatially varying effects of the factors on the crashes of rural two-lane highways in addition to investigating the stationary effect based on the traditional count models.

Even though the statistical models have advantages like better interpretability and transformability, they can be prone to multicollinearity issues when a large number of independent variables are considered. A fixed functional form is also required before applying these models. To overcome the issues in these models, studies were found to use ML tools in predicting crashes (16; 95; 96). Such studies are rather limited for crash prediction of rural two-lane highways especially incorporating speed as one of the factors (97). This study attempts to fill the gap by developing a crash prediction model based on an ML technique while considering speed along with other geometric and traffic factors.

### 2.2.3 Studies Incorporating Speed Measures in Analyzing Crash Severity

According to the HSM, severity of a crash can be classified as Fatality (K), Injury (A/B/C), and Property Damage (O), where injury can be further divided into Incapacitating Injury (A), Non-incapacitating Injury (B), Possible Injury (C). A segment with more fatalities or injuries is of more concern compared to a segment with higher property damage crashes. As per HSM, the costs of fatality and injury are 48 times and 11 times higher than the property damages (7). Therefore, it is important to take into consideration the different severity levels while estimating the crash frequency and investigate the causes that result in the crashes at individual severity levels.

The general approach to crash prediction described in HSM is that it applies CMFs and calibration factors to the base SPF for a road segment. The purpose of this approach is to predict how the design and operational changes can influence the safety of

a specific road. Based on this approach, the total number of crashes can be estimated. In terms of severity, HSM applies fixed proportions to the total number of crashes for estimating the crash counts at different crash severity levels (K, A, B, C, O). For a rural two-lane highway, HSM suggests using the percentages shown in Table 2 if data is not available for a particular jurisdiction (7). When data are available, the fixed proportions can be calculated by dividing the observed count under each severity with the total observed crashes from the dataset. These proportions are average values for a typical condition and may not be sensitive to varying geometric and traffic conditions. In practice, the distribution of severity might vary for different segments depending on the characteristics of the road in addition to other factors. Using a default proportion may lead to a biased estimation of crashes for different severity levels.

Table 2 Default distribution for Crash Severity Level on Rural Two-Lane Highways  
(Source: HSM (7))

<b>Crash Severity Level</b>	<b>Percentage of Total Roadway Segment Crashes</b>
Fatal	1.3
Incapacitating Injury	5.4
Non-incapacitating Injury	10.9
Possible Injury	14.5
Property Damage	67.9

Besides HSM, there are other existing research that looked at the different crash severity levels. Some studies only investigated a particular level of severity, especially the severe levels of crashes, which include fatal and injury crashes (19; 33; 36; 39; 71; 107). They primarily explored the factors specific to fatality or injury crashes. Another type of study considered both severe and non-severe crashes and investigated how the effects of road attributes and traffic characteristics vary over the different severity levels (27; 31; 32; 34; 35; 37; 38; 108-113). The modeling strategy of these studies depends on

whether the crash data are aggregated to a road segment level or not. The strategies can be summarized below:

- Developing crash severity prediction model where severity level is predicted either as the discrete variable or proportion (32; 110-113). This type of modeling approach is adopted when crash data are available in a disaggregated format, i.e., individual crash records.
- Estimating the proportion of crashes at each severity level for a specific segment and applying the proportions to the SPF to predict the number of crashes for each severity level (34; 109). Crash data are also required to be in a disaggregated format for this type of approach.
- Developing count models for crash prediction by crash severity levels (27; 31; 35; 37; 38; 108). This approach is suitable when the crash data is available in an aggregated format for each road segment.

While most of the above studies analyzed the crash severity mainly on the higher functional class of roads (27; 33-35; 107; 111; 113), few are focused on the crash severity for rural two-lane highways (31; 37-39; 109). Especially in terms of assessing the effect of speed characteristics on different levels of crash severity, other facility types received much attention (19; 27; 32-36), however, limited works were found for rural two-lane highways (31; 37-39). As speed measures, mainly speed limit and design speeds have been explored for rural two-lane highways (37-39). For these roads, speed limit may not always capture the actual operating condition. Therefore, this study tries to incorporate other speed measures (for example average speed, standard deviation of speed, etc.) to explore the effect of operating conditions on the crash severity of these roads. In addition, it will be worth investigating how the effect of speed differs for different severity levels while controlling for road geometric and traffic factors.

### 2.3 Literature Review Summary

With the advancement in data collection techniques, several studies estimated speed measures from the actual speed data and incorporated the measures in analyzing the role of speed on crashes. These studies were primarily undertaken for the higher

functional class such as interstates, multilane highways, urban arterials, etc. In contrast, measured speed data on rural two-lane highways were sparse in the past. Therefore, investigating the effect of speed on the crashes using the measured data was rather limited for rural two-lane highways. With the proliferation of GPS probe speed data, speed data has been available on this road. Utilizing the availability of such datasets, this study will investigate the effect of speed on the crashes of rural two-lane highways.

In terms of the speed measures, the existing studies mainly looked at the 85<sup>th</sup> percentile speed as a design consistency indicator between consecutive segments and incorporated it into the crash prediction model for rural two-lane highways. The measure was calculated either from a model or spot speeds. This study uses measured speed data to estimate different speed measures such as average speed, standard deviation of speed, 85<sup>th</sup> percentile speed, etc. and investigates the influence of speed on the crashes of these roads from both planning and design consistency aspects.

Count models such as Poisson, NB, ZIP, ZINB, etc. are widely used in analyzing crashes of a segment. These approaches can offer interpretability on the average effect of a factor on crashes. However, they may not capture spatial heterogeneity in the effect of the factors if spatial dependency is present in the dataset. Such issues can be further addressed by utilizing spatial models such as GWR method. The spatial techniques were mostly applied in macro-level crash analysis rather than segment-level crashes. This study utilizes the GWR tools to analyze the spatially varying effects of the factors on the crashes of rural two-lane highways in addition to investigating the stationary effect based on the traditional count models.

Despite having advantages like better interpretability and transformability, statistical models can be susceptible to multicollinearity among the independent variables. Furthermore, they require a presumption on the functional form of the model. To overcome these issues, studies were found to apply ML tools in predicting crashes (16; 95; 96). Such studies are rather limited for crash prediction of rural two-lane highways, especially incorporating speed as one of the factors (97). This study will also explore ML modeling techniques to develop crash prediction model for these roads by incorporating speed along with other geometric and traffic factors.

In case of assessing the effect of speed on crash severity, the higher functional class roads mainly received the research attention (19; 27; 32-36). Limited works were identified for rural two-lane highways (31; 37-39). As speed measures, speed limit and design speed have been explored for rural two-lane highways (37-39). For these roads, speed limit may not always capture the actual operating condition. Therefore, this study will explore the effect of different speed measures (for example average speed, standard deviation of speed) on the crashes at different severity levels for these roads. In addition, this study will investigate how the effect of speed differs for different severity levels while controlling for geometric and traffic factors.

## CHAPTER 3. DATA COLLECTION PROCESSING

This chapter documents the sources of datasets used for this study. The databases for road attributes, crashes, and speed were obtained. After collecting the data, preprocessing was done to link up information from each database together. ArcGIS and Python were used as tools to perform all the preprocessing.

### 3.1 Data Sources

Crash data were collected from the Kentucky State Police database between 2013 and 2017 for rural two-lane highways. The crash dataset came into an aggregated format regardless of the direction of the roads. Roadway attributes such as AADT, degree of curvature, lane width, shoulder width, grades, functional class, etc. were extracted from the Highway Information System (HIS) of KYTC. Third-party GPS-based probe speed data were obtained from HERE Technologies for 2015 to 2017 at 5-minute increments (114). These data were from both directions of the road.

### 3.2 Data Pre-Processing

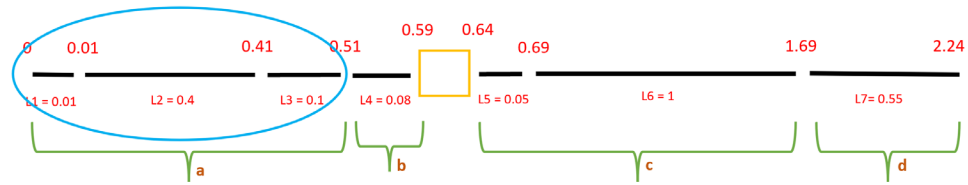
After the data collection, homogenous segmentation was done utilizing the segments from HIS. The ‘Overlay Route Events’ tool in ArcGIS was applied to obtain the homogenous segments based on functional classes, traffic counts, shoulders, grades, horizontal curves, and speed limit. The overall process breaks down a HIS segment any time one of the road attributes is changed.

The crash dataset was spatially joined with the homogenous segments using ArcGIS. Furthermore, crashes that occurred at intersections, identified as areas within 100 feet of intersections, were excluded from the dataset because they are more likely to be associated with a different set of factors. After that, the homogenous segments were linked to the HERE network to obtain speed data for individual segments. A data adequacy screening was performed to ensure only the segments with adequate data (i.e. segments meeting the required minimum data availability rate of 10% ) are used in this study (115). The daytime data consisted of 60% speed data compared to 24 hours of data. Therefore, this study decided to use daytime data to obtain credible speed measures.

Later, average speed, standard deviation of the speed, and the 85<sup>th</sup> percentile speed were calculated from the HERE speed. All these measures were initially corresponded to the HERE links for each direction. To convert them into the homogenous segment level, Space Mean Speed (SMS) was followed. Finally, all speed metrics were averaged from both directions of a segment.

All the above preprocessing resulted in 2,78,187 homogeneous segments. For developing crash prediction models, a check on the minimum segment length is required. According to Hauer and Bamfo, the minimum segment length should be considered as 0.1 mile (116). However, 78% of the homogenous segments were shorter than the required minimum. This study performed additional processing of the homogeneous segments through aggregation so that a sufficient amount of data can be utilized even after applying the constraint for minimum segment length. The aggregation process is described below.

Using Python scripting, this study performed an aggregation process on the 2,78,187 segments to add consecutive segments up to half a mile when there is no intersection between the segments. Figure 2 shows a demonstration of this process. It shows some homogenous segments with increasing mile points (0 to 2.24 mile points) from left to right. As the process allows the aggregation until the summation of the length reaches a maximum of 0.5 miles, Section **a** consisting of L1, L2 and L3 becomes the first aggregated segment. For the next aggregation, the process can only include L4 in Section **b** since an intersection is present after that (yellow rectangle). Similarly, Section **c** and Section **d** were obtained. The road attributes related to the segments were also aggregated as length weighted average and crashes were summed. In this way, the resulting sections can be still homogeneous segments. Such aggregation finally resulted in 44,008 segments with a total of 93,820 crashes summed up from both directions of the road. They represent a total of 21,240 miles of rural two-lane highways in Kentucky, as shown in Figure 3.



Note:  $L$  = Length of the homogenous segments

Figure 2 Segment Aggregation

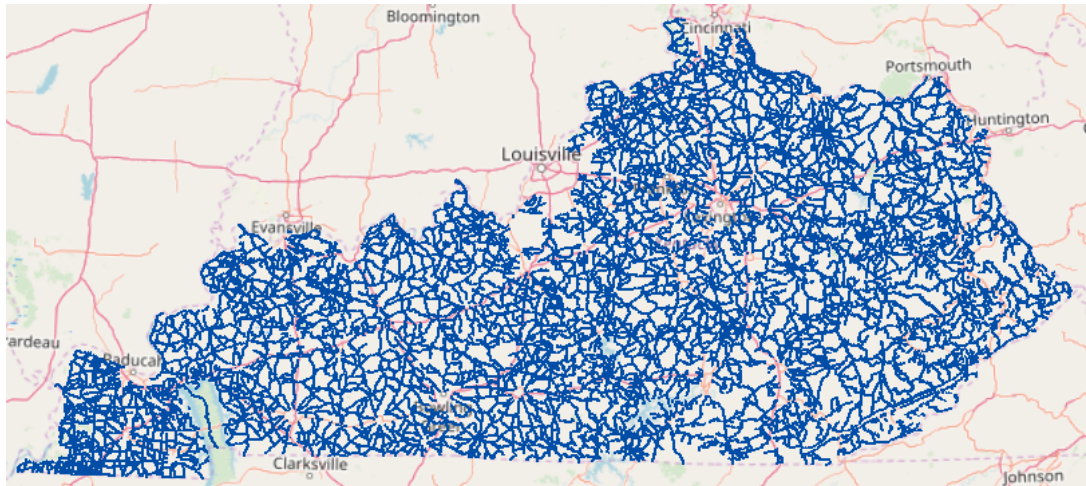


Figure 3 Rural Two-Lane Segments in Kentucky

Overall, the dataset after the aggregation process meets the minimum requirement of 100-200 miles for SPF development following HSM and *Safety Performance Function Decision Guide* (7; 117). Table 3 lists the attributes attached to the segments in the dataset. The list contains road geometries, traffic conditions, speed measures, and crash-related information.

Table 3 Attributes Associated with the Segments

Explanatory Variables	Fields
Geometric Conditions	Section Length, Shoulder Width, Degree of Curvature, Lane Width, Grade.
Traffic Condition	AADT



Speed	Average Speed, Speed Limit, Standard Deviation of Speed, The 85 <sup>th</sup> percentile Speed, Speed Differentials, etc.
<b>Response Variable</b>	
Crashes	Number of total crashes, K, A, B, C, O Crashes, in 5 years.

### 3.3 Summary

Once the necessary data collections were done, this study prepared the dataset by linking HIS segments with the crash and speed databases. This results in a set of homogeneous segments, which were further aggregated based on segment length and presence of intersection to minimize the exclusion of shorter segments. All the pre-processing resulted in 44,008 segments. The attributes associated with these segments will be utilized as the potential factors of crashes during the analysis in this study.

## CHAPTER 4. METHODOLOGY

This chapter discusses the potential factors that will be utilized in the analysis, especially looking into the effect of speed on the crashes of rural two-lane highways. Methods related to the selection of ultimate variables for the model development are also detailed here. In addition, the modeling approaches experimented on in this study are discussed. The chapter concludes with an overview of the evaluation criteria followed by a summary.

### 4.1 Potential Factors of Crashes

For analyzing crashes, the widely used exposure variables in the existing literature include AADT and segment length (20; 22; 27; 29; 43; 62). Following the existing practices, this study also considered AADT and length in developing crash prediction models. In addition, studies also identified significant relationships between crashes and different geometric attributes especially degree of curvature, shoulder width, and lane width (27; 35; 37; 38; 109). This study also included these variables during model development and analyzing the results.

Since the general focus of this study is to evaluate the influence of speed on crashes of rural two-lane highways, several speed measures were tested in individual analysis. Existing studies investigated average speed, the 85<sup>th</sup> percentile speed, std of speed, speed limit, difference between average speed and speed limit, difference between the 85<sup>th</sup> percentile speed and speed limit, etc. as the speed metrics to analyze the effect of speed on crashes for different facility types(16; 18; 19; 35; 48-50; 59; 62; 63). In case of rural two-lane highways, current studies explored the effect of speed mainly based on the 85<sup>th</sup> percentile and speed limit (21-25; 43; 62). This study experimented with the average speed, the 85<sup>th</sup> percentile speed, std of speed, speed limit, difference between average speed and speed limit, and difference between the 85<sup>th</sup> percentile speed and speed limit depending on the focus of the individual analysis.

As the design consistency metric, the 85<sup>th</sup> percentile speed is one of the commonly used candidate measures because it reflects the behavior of most of the drivers, especially in the curve segments (118). For a particular section of the road,

difference between the 85<sup>th</sup> percentile speed and design speed is generally used to identify inconsistency in the design of that section, whereas, the difference of the 85<sup>th</sup> percentile speed between consecutive sections can identify the inconsistency that a driver may experience when traversing from one section to another. The latter is a proper measure to understand the crashes that vary with the changes in the degree of curvature on the horizontal curves of rural two-lane highways (21). Moreover, it is one of the safety criteria suggested by Lamm et al (118). To further evaluate the effect of speed on crashes of rural two-lane highways from design consistency perspective, this study considered speed differential i.e. the difference of the 85<sup>th</sup> percentile speed between consecutive segments (119). This measure reflects the fact that crash occurrence may not only depend on the local design conditions of a particular road segment but also depends on the condition of adjacent segments.

Overall, the possible geometric and traffic factors to analyze the crashes of rural two-lane highways were identified through the existing practices. Regarding speed, this study will evaluate different measures such as average speed, the 85<sup>th</sup> percentile speed, Std of speed, etc. for crash predictions of rural two-lane highways. Incorporation of the speed measures in the crash prediction model will help to understand the role of speed on the crashes of these roads. Such analysis can add further insights into the current state of art practice.

## 4.2 Methods for Variable Selection

### 4.2.1 Pearson Correlation Coefficient

Preliminarily, a correlation check between the explanatory variables and response variables was done using Pearson correlation coefficient estimated by Equation (4). The purpose was to understand the linear association between them and do the primary selection of the explanatory and response variables for the respective analysis. A stronger association is indicated by a value of  $r$  closer to 1, whereas, a value closer to 0 means a complete lack of linear association.

$$r = \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum(Y_i - \bar{Y})^2 \sum(X_i - \bar{X})^2}} \quad (4)$$

Where,

$r$  = Pearson correlation coefficient (range -1 to 1)

$Y_i = i^{th}$  observed value of the variable  $Y$

$X = i^{th}$  observed value of the variable  $X$

$\bar{Y}$  = mean of the observations of the variable  $Y$

$\bar{X}$  = mean of the observations of the variable  $X$

Additionally, this study utilized the Pearson correlation coefficient to further check the multicollinearity between each pair of the explanatory variables before including them in the statistical models. A correlation coefficient of higher than 0.6 was used as an indication for significant multicollinearity following Ji et. al. (86). If the correlation coefficient between two explanatory variables is higher than 0.6, one of the variables was not included in the model development.

#### 4.2.2 Spearman's Correlation Coefficient

Depending on the requirement of the analysis, Spearman's correlation coefficient ( $\rho$ ) was also used to check the correlations between variables. This method does not assume a linear relationship between the variables but rather considers a monotonic relationship i.e. one variable increases with another variable or decreases with another variable, but not necessarily as a straight line, as shown in Figure 4 (120). The null hypothesis for Spearman's correlation test is that the relationship between the variables is not monotonic. A p-value of less than 0.05 for the correlation suggests that the null hypothesis is rejected. Spearman  $\rho$  value indicates the strength and direction of the relationship. Spearman  $\rho$  close to -1 or +1 implies the strongest correlation. Based on the p-value, we can determine whether the correlation is significant or not.

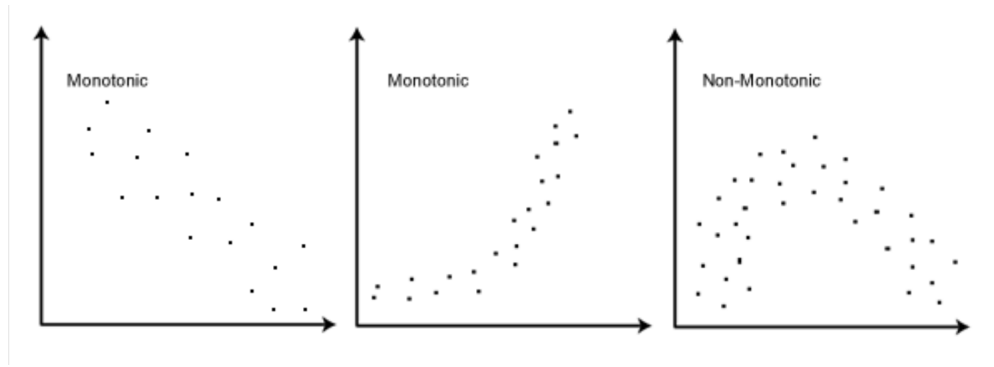


Figure 4 Illustration of Monotonic and Non-Monotonic Relationships

### 4.3 Spatial Dependency Test

To check the appropriateness of developing spatial models, this study had to confirm the spatial dependency of the explanatory and response variables. For this, a spatial autocorrelation test was performed using Moran Global Index (I) calculated based on Equation (5) (121). The null hypothesis for Moran's I is spatial randomness i.e., no spatial dependency. A p-value of less than 0.05 for the correlation suggests that the null hypothesis is rejected. Moran's I can range between -1 and 1 (82). A value of zero means no spatial autocorrelation. The closer the Moran's I value to 1, the stronger the spatial correlation and the higher the similarities between the adjacent neighbors. Conversely, a Moran's I value closer to -1 means a perfect dispersion in the data with lower similarities between the neighbors.

$$I = \frac{n}{W} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (5)$$

Where,

$n$  = number of observations

$w_{ij}$  = spatial weight for a pair of objects

$W$  = sum of spatial weights

$X_i, X_j$  = values of a variable for location  $i$  and  $j$

$\bar{X}$  = mean value of a variable

#### 4.4 Modeling Approach

This study utilized HSM method to compare with the prediction models specific to this research. To fulfill the goals of this study, count models, spatial models, and ML models were experimented. The subsections below discuss the models utilized for this study.

##### 4.4.1 HSM Method

Part C of the HSM presents the traditional approach to predict crash frequency at individual sites on different roadway facilities including rural two-lane highways (7). The general form of the predictive models in the HSM can be expressed as follows:

$$N_{predicted,i} = N_{spf,i} \times (CMF_{1,i} \times CMF_{2,i} \times \dots \times CMF_{n,i}) \times C_i \quad (6)$$

Where,

$N_{predicted,i}$  = predicted number of crashes for a specific year for segment  $i$

$N_{spf,i}$  = predicted number of crashes for a specific year for a segment  $i$  for base conditions

$CMF_{1,i}, CMF_{2,i}, \dots, CMF_{n,i}$  = crash modification factors for  $n$  geometric conditions or traffic control features for segment  $i$

$C_i$  = calibration factor to adjust SPF for local conditions for segment  $i$ .

As shown in Equation (6), there are three components of the HSM models: base SPFs, CMFs, and calibration factors. Base SPFs are generally the statistical models that are used to predict crash frequency for a facility type with definite base conditions. The base SPF (HSM Equation 10-6) introduced by HSM is presented in Equation (7).

$$N_{spf,i} = AADT \times L \times 365 \times 10^{-6} \times e^{(-0.312)} \quad (7)$$

Where,

$N_{spf,i}$  = predicted number of crashes for a specific year for a segment  $i$  for base conditions

$AADT$  = average annual daily traffic (vehicles per day)

$L$  = segment length (miles)

CMFs are used to account for the effects of non-base conditions on predicted crashes. When a segment does not meet any of the base conditions listed in Table 4, a CMF is multiplied by the base SPF shown in Equation (7).

Table 4 Base Condition for Rural Two-Lane, Two-Way Highways (Source: HSM)

Lane Width	12 feet
Shoulder Width	6 feet
Shoulder Type	Paved
Roadside Hazard Rating	3
Driveway Density	5 driveways per mile
Horizontal Curvature	None
Vertical Curvature	None
Central Rumble Strips	None
Passing Lanes	None
Two-way left-turn lanes	None
Lighting	None
Automated speed enforcement	None
Grade Level	0%

Calibration factors are also required “to account for differences between the jurisdiction and time period for which the predictive models were developed and the jurisdiction and time period to which they are applied by HSM users” (122). Calibration factor is estimated as the ratio of the total number of observed crashes to the total number of predicted crashes calculated using the SPFs and CMFs provided in the HSM.

## 4.4.2 Statistical Models

### 4.4.2.1 Traditional Count Models

This study utilized traditional count models for analyzing the crashes of the rural two-lane segments. The specific models that were considered can be listed as:

1. Poisson Model
2. NB Model
3. ZIP Model and
4. ZINB Model

#### 4.4.2.1.1 POISSON MODEL

Poisson regression model is one of the count models, which is widely used to predict crashes assuming that the number of crashes follows Poisson distribution (23; 43; 123). In case of Poisson model, the probability mass function for a given value of  $Y_i = y_i$  can be written as below.

$$Pr(Y_i = y_i | \mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \quad y_i = 0, 1, 2, 3, \dots \quad (8)$$

Where,

$Y_i$  = number of crashes

$\mu_i$  = expected number of crashes and can be estimated from Equation (9)

$$\mu_i = e^{(\beta_0 + \beta_1 AADT + \beta_2 L + \beta_3 V + \sum \beta_n X_n)} \quad (9)$$

Where,

$\mu_i$  = expected number of crashes

$\beta_0$  = random intercept term

$\beta_1, \beta_2, \dots, \beta_n$  = estimated regression coefficients

$AADT$  = average annual daily traffic (vehicles per day)

$L$  = segment length (miles)

$V$  = speed measure (mph)



$X_n$  = other geometric variables (if considered in the specific analysis)

One of the key assumptions of Poisson regression model is that the mean and variance from the observed crash data are equal as presented in Equation (10).

$$\mu_i = Var(Y_i) \quad (10)$$

Where,

$\mu_i$  = expected number of crashes

$Y_i$  = number of crashes

$Var(Y_i)$  = variance of  $Y_i$

#### 4.4.2.1.2 NEGATIVE BINOMIAL MODEL

When the mean and variance of the observed crash data are not equal, the crash data are overdispersed, and NB model is recommended instead of Poisson model (123). The NB Model uses a Gamma Probability Distribution of observed crashes. Let's assume  $Y_i$  represents number of crashes and its values are  $y_i \in 0, 1, 2, 3, \dots$ . The probability of  $Y_i$ , can be written as the distribution of NB below.

$$Pr(Y_i = y_i | \mu_i, \alpha) = \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(1 + y_i)! \Gamma(\alpha^{-1})} \left(\frac{\alpha \mu_i}{1 + \alpha \mu_i}\right)^{y_i} \left(\frac{1}{1 + \alpha \mu_i}\right)^{\alpha^{-1}} \quad (11)$$

Where,

$Y_i$  = number of crashes

$\mu_i$  = expected number of crashes and can be estimated from Equation (12)

$\Gamma$  = gamma function

$\alpha$  = over-dispersion parameter that can be calculated from Equation (13)

$$\mu_i = e^{(\varepsilon_i + \beta_1 AADT + \beta_2 L + \beta_3 V + \sum \beta_n X_n)} \quad (12)$$

Where,

$\mu_i$  = expected number of crashes

$\varepsilon_i$  = gamma-distributed error

$\beta_1, \beta_2, \dots, \beta_n$  = estimated regression coefficients

$AADT$  = average annual daily traffic (vehicles per day)

$L$  = segment length (miles)

$V$  = speed measure (mph)

$X_n$  = other geometric variables (if considered in the specific analysis)

$$\alpha = \frac{Var(Y_i) - \mu_i}{\mu_i^2} \quad (13)$$

Where,

$\mu_i$  = expected number of crashes

$Y_i$  = number of crashes

$Var(Y_i)$  = variance of  $Y_i$  and can be estimated from Equation (14)

$$Var(Y_i) = \mu_i(1 + \alpha\mu_i) \quad (14)$$

#### 4.4.2.1.3 ZERO INFLATED MODELS

Since crashes are rare, the dataset may contain a significant amount of zero crashes. In this study, zero crashes were observed on more than 50% of the rural two-lane segments. As a result, the crash dataset can be significantly overdispersed relative to its mean. To handle the excess zero crashes in the dataset, Poisson-based and Negative Binomial-based zero-inflated models were introduced by Lambert and Greene, respectively (124; 125). This study also adopted ZIP and ZINB approaches to the modeling of crashes on rural two-lane highways. The underlying methodology of these models is discussed as follows.

#### **Zero Inflated Poisson Model**

ZIP is a combination of two models: a binary model and a Poisson model (125). The binary model is used to produce the excess zero crashes and, Poisson model

produces the number of crashes of a segment including zero crashes following a Poisson distribution. If the probability of the data point (i.e., number of crashes) produced by the binary model is  $p_i$ , the probability of the data point generated by the Poisson model will be  $(1-p_i)$ . In ZIP,  $p_i$  is generally fitted using a logistic regression model as a function of the explanatory variables shown in Equation (15) (125).

$$\ln\left(\frac{p_i}{1-p_i}\right) = \gamma_0 + \gamma_1 AADT + \gamma_2 L + \gamma_3 V + \sum \gamma_n X_n \quad (15)$$

Where,

$\frac{p_i}{1-p_i}$  = odds ratio of the probability for binary process to the probability for Poisson

model

$\gamma_0$  = intercept

$\gamma_1, \gamma_2, \dots, \gamma_n$  = estimated regression coefficients

$AADT$  = average annual daily traffic (vehicles per day)

$L$  = segment length (miles)

$V$  = speed measure (mph)

$X_n$  = other geometric variables (if considered in the specific analysis)

Equation 15 can be transformed below to estimate the probability of zero crashes from the binary process. A  $p_i$  value close to 1 implies that segment  $i$  is more likely to have no crashes and therefore safe.

$$p_i = \frac{e^{(\gamma_0 + \gamma_1 AADT + \gamma_2 L + \gamma_3 V + \sum \gamma_n X_n)}}{1 + e^{(\gamma_0 + \gamma_1 AADT + \gamma_2 L + \gamma_3 V + \sum \gamma_n X_n)}} \quad (16)$$

Now, the probability distribution of the number of crashes for segment  $i$  can be expressed as:

$$Pr(Y_i = 0) = p_i + (1 - p_i) \exp(-\mu_i) \quad (17)$$

$$Pr(Y_i = y_i) = (1 - p_i) \frac{\exp(-\mu_i) (\mu_i)^{y_i}}{y_i!}, \quad y_i > 0 \quad (18)$$

Where,

$Y_i$  = number of crashes

$p_i$  = probability of crashes produced by binary model

$\mu_i$  = expected number of crashes and can be estimated from Equation (19)

$$\mu_i = (1 - p_i) e^{(\beta_0 + \beta_1 AADT + \beta_2 L + \beta_3 V + \sum \beta_n X_n)} \quad (19)$$

Where,

$\mu_i$  = expected number of crashes

$\beta_0$  = random intercept term

$\beta_1, \beta_2, \dots, \beta_n$  = estimated regression coefficients

$AADT$  = average annual daily traffic (vehicles per day)

$L$  = segment length (miles)

$V$  = speed measure (mph)

$X_n$  = other geometric variables (if considered in the specific analysis)

### **Zero Inflated Negative Binomial Model**

Similar to ZIP, ZINB is a combination of two models (126): a binary model and a negative binomial (NB) model. The binary model is used to produce the excess zero crashes and, NB model produces the number of crashes of a segment including zero crashes following a binomial process. If the probability of the data point produced by the binary model is  $p_i$ , the probability of the data point generated by the NB model will be  $(1-p_i)$ . In ZINB,  $p_i$  is fitted using a logistic regression model as a function of the explanatory variables shown in the Equation (20) (72).

$$\ln\left(\frac{p_i}{1-p_i}\right) = \gamma_0 + \gamma_1 AADT + \gamma_2 L + \gamma_3 V + \sum \gamma_n X_n \quad (20)$$

Where,

$\frac{p_i}{1-p_i}$  = odds ratio of the probability for binary process to the probability for NB process

$\gamma_0$  = intercept

$\gamma_1, \gamma_2, \dots, \gamma_n$  = estimated regression coefficients

$AADT$  = average annual daily traffic (vehicles per day)

$L$  = segment length (miles)

$V$  = speed measure (mph)

$X_n$  = other geometric variables (if considered in the specific analysis)

Equation 20 can be transformed below.

$$p_i = \frac{e^{(\gamma_0 + \gamma_1 AADT + \gamma_2 L + \gamma_3 V + \sum \gamma_n X_n)}}{1 + e^{(\gamma_0 + \gamma_1 AADT + \gamma_2 L + \gamma_3 V + \sum \gamma_n X_n)}} \quad (21)$$

Now, the probability of the number of crashes on segment  $i$ ,  $Y_i$ , can be written as the distribution of ZINB below (127).

$$Pr(Y_i = 0) = p_i + (1 - p_i) \left( \frac{1}{1 + \alpha \mu_i} \right)^{\alpha - 1} \quad (22)$$

$$Pr(Y_i = y_i) = (1 - p_i) \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(1 + y_i) \Gamma(\alpha^{-1})} \left( \frac{\alpha \mu_i}{1 + \alpha \mu_i} \right)^{y_i} \left( \frac{1}{1 + \alpha \mu_i} \right)^{\alpha - 1} \quad y_i = 1, 2, 3, \dots \quad (23)$$

Where,

$Y_i$  = number of crashes

$p_i$  = probability of crashes produced by binary process

$\Gamma$  = gamma function

$\alpha$  = over-dispersion parameter that can be calculated from Equation (13)

$\mu_i$  = expected number of crashes and can be estimated from Equation (19)

$$\mu_i = (1 - p_i) e^{(\varepsilon_i + \beta_1 AADT + \beta_2 L + \beta_3 V + \sum \beta_n X_n)} \quad (24)$$

Where,

$\mu_i$  = expected number of crashes

$\varepsilon_i$  = gamma-distributed error

$\beta_1, \beta_2, \dots, \beta_n$  = estimated regression coefficients

$AADT$  = average annual daily traffic (vehicles per day)

$L$  = segment length (miles)

$V$  = speed measure (mph)

$X_n$  = other geometric variables (if considered in the specific analysis)

#### 4.4.2.2 Spatial Count Models

Traditional regression models assume that the coefficients of parameters are constant over space. However, crash data may contain spatial heterogeneity, and traditional models may not capture the spatial heterogeneity of the crash data which might lead to a biased estimation of the results (93). To address this, spatial modeling techniques like geographically weighted regression models have been introduced (81). These models take into account the spatial context when establishing the relationship between crashes and the explanatory variables.

Since this study particularly investigates the crashes as a count variable, the Geographically Weighted Poisson Regression (GWP) and Geographically Weighted Zero Inflated Poisson Regression (GWZIP) are suitable choices. The model assumes that the coefficients of the independent variables vary across the space. For each data point, they fit a local model with the closest neighbors and provide a set of estimated coefficients. Analysis based on the local models can help in identifying more appropriate countermeasures (78; 82-87; 92). The underlying methodology of these models is discussed as follows.

##### 4.4.2.2.1 GEOGRAPHICALLY WEIGHTED POISSON REGRESSION MODEL

If the crash dataset contains  $i$  segments, the GWP modeling approach develops a total of  $i$  models. It means that there will be  $i$  local models. Suppose, in case of segment 1 (coordinates of the midpoint of the segment is  $(u_1, v_1)$ ) shown in Figure 5, it has  $l$  closest neighboring segments. For example, if the value of  $l$  is 3, the neighbors can be shown within the red box in Figure 5. Now, the local model is developed with these  $l$  neighbors using Poisson model, which follows the probability mass function written below.

$$Pr(Y_1 = y_1 | \mu_1) = \frac{\mu_1^{y_1} e^{-\mu_1}}{y_1!} \quad y_1 = 0,1,2,3, \dots \quad (25)$$

Where,

$Y_1$  = number of crashes on target segment 1

$\mu_1$  = expected number of crashes for segment 1 that can be estimated from Equation (26)

$$\mu_1 = e^{(\beta_0(u_1, v_1) + \beta_1(u_1, v_1)AADT_1 + \beta_2(u_1, v_1)L_1 + \beta_3(u_1, v_1)V_1 + \sum \beta_n(u_1, v_1)X_{n,1})} \quad (26)$$

Where,

$\mu_1$  = expected number of crashes for segment 1

$\beta_0$  = random intercept term

$\beta_1(u_1, v_1), \beta_2(u_1, v_1), \dots, \beta_n(u_1, v_1)$  = estimated regression coefficients for segment 1

$AADT_1$  = average annual daily traffic (vehicles per day) for segment 1

$L_1$  = length (miles) for segment 1

$V_1$  = speed measure (mph) for segment 1

$X_{n,1}$  = other geometric variables for segment 1 (if considered in the specific analysis)

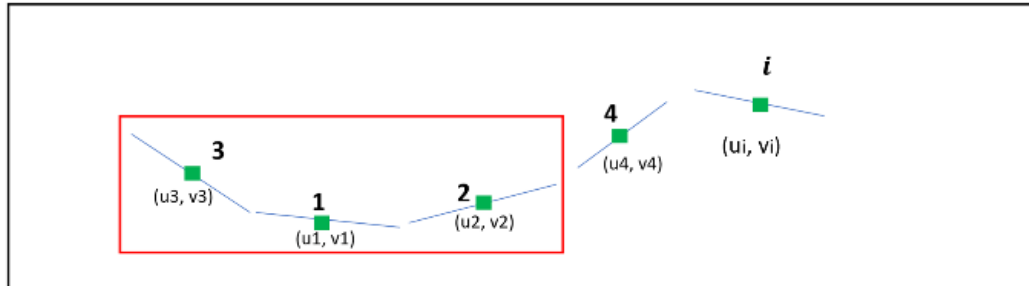


Figure 5 Demonstration of Geographically Weighted Regression Modeling Process

For segment 1, the coefficients in Equation (26) are estimated using the Maximum Likelihood Estimation (MLE) method as below (1):

$$LL(\beta(u_1, v_1)) = \sum_{l=1}^l \left( (y_l X_{m,l}^T \beta(u_1, v_1)) - e^{X_{m,l}^T \beta(u_1, v_1)} - \ln(y_l!) \right) w_{1l}(u_1, v_1) \quad (27)$$

Where,

LL = log-likelihood

$l$  = index of the segments considered for the local model and in this case, the segments are segment 1 to  $l$

$y_l$  = number of crashes on  $l^{th}$  segment

$X_{m,l}^T$  = column vector of the variables AADT, length, speed measure, etc. for  $l$  segments

$w_{1l}(u_1, v_1)$  = weight function to describe the influence of the neighbor segments around the target segment 1. In this study case, adaptive kernel bi square function is used as the weight function and can be calculated as,

$$w_{1l} = \begin{cases} \left[ 1 - \left( \frac{d_{1l}}{h} \right)^2 \right]^2, & \text{if } d_{1l} \leq h \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

Here,

$d_{1l}$  = Euclidian distance between target segment 1 and the  $l^{th}$  neighbor segment

$h$  = bandwidth distance (since this study is considering  $l$  neighboring segments in local model development, it is the distance between target segment 1 and the neighbor, which is farthest from the target segment.)

The  $h$  is determined through the minimization of the cross-validation (CV) score while iterating over a different number of neighbors. CV score is calculated as the sum of the squares of local model residuals (actual number of crashes - predicted number of crashes). In this case, for each of the  $i$  segments, the local model is fitted based on  $l$  nearest neighbors and CV score is calculated. After that, the process is repeated by increasing the number of neighbors, and the number of neighbors associated with the



lowest CV score is utilized to calculate the optimum bandwidth,  $h$ . For example, if the whole process obtains the lowest CV score for 500 closest neighboring segments,  $h$  is estimated as the distance to the 500<sup>th</sup> segment from the target segment 1, and the final local models are fitted using 500 neighboring segments. Finally, the general form of the local model for each of the  $i$  segments can be written as,

$$\mu_i = e^{\beta_0(u_i, v_i) + \beta_1(u_i, v_i)AADT_i + \beta_2(u_i, v_i)L_i + \beta_3(u_i, v_i)V_i + \sum \beta_n(u_i, v_i)X_{n,i}} \quad (29)$$

#### 4.4.2.2.2 GEOGRAPHICALLY WEIGHTED ZERO INFLATED POISSON REGRESSION

Similar to GWP method, GWZIP modeling approach develops a total of  $i$  models, if the crash dataset contains  $i$  segments. Considering the same example in Figure 5, the local model is developed with the  $l$  neighbors using ZIP model, which is a combination of binary model and Poisson model. At target segment 1, the binary part models the excess zero crashes of the neighboring segments, and the Poisson models the number of crashes including the zero crashes following Poisson distribution. For this segment, if the probability of the number of crashes produced by the binary model is  $p_1$ , the probability of the number of crashes produced by the Poisson model is  $(1 - p_1)$ .

Now,  $p_1$  is fitted using a logistic regression model as a function of the explanatory variables as below.

$$\ln\left(\frac{p_1}{1-p_1}\right) = \gamma_0(u_1, v_1) + \gamma_1(u_1, v_1)AADT_1 + \gamma_2(u_1, v_1)L_1 + \gamma_3(u_1, v_1)V_1 + \sum \gamma_n(u_1, v_1)X_{n,1} \quad (30)$$

Where,

$\gamma_0(u_1, v_1)$  = intercept

$\gamma_1(u_1, v_1), \gamma_2(u_1, v_1), \dots, \gamma_n(u_1, v_1)$  = estimated regression coefficients for segment 1,

which is estimated using MLE method presented in Equation (35)

$AADT_1$  = average annual daily traffic (vehicles per day) for segment 1

$L_1$  = length (miles) for segment 1

$V_1$  = speed measure (mph) for segment 1

$X_{n,1}$  = other geometric variables for segment 1 (if considered in the specific analysis)

Equation (30) can be rewritten as,

$$p_1 = \frac{e^{\gamma_0(u_1, v_1) + \gamma_1(u_1, v_1)AADT_1 + \gamma_2(u_1, v_1)L_1 + \gamma_3(u_1, v_1)V_1 + \sum \gamma_n(u_1, v_1)X_{n,1}}}{1 + e^{\gamma_0(u_1, v_1) + \gamma_1(u_1, v_1)AADT_1 + \gamma_2(u_1, v_1)L_1 + \gamma_3(u_1, v_1)V_1 + \sum \gamma_n(u_1, v_1)X_{n,1}}} \quad (31)$$

and the probability distribution of crashes for segment 1 can be expressed as:

$$\Pr(Y_1 = 0) = p_1 + (1 - p_1) \exp(-\mu_1) \quad (32)$$

$$\Pr(Y_1 = y_1) = (1 - p_1) \frac{\exp(-\mu_1) (\mu_1)^{y_1}}{y_1!}, \quad y_1 > 0 \quad (33)$$

Where,

$Y_1$  = number of crashes on target segment 1

$\mu_1$  = expected number of crashes for segment 1 that can be estimated from Equation (34)

$$\mu_1 = (1 - p_1) e^{(\beta_o(u_1, v_1) + \beta_1(u_1, v_1)AADT_1 + \beta_2(u_1, v_1)L_1 + \beta_3(u_1, v_1)V_1 + \sum \beta_n(u_1, v_1)X_{n,1})} \quad (34)$$

Where,

$\mu_1$  = expected number of crashes for segment 1

$\beta_o$  = random intercept term

$\beta_1(u_1, v_1), \beta_2(u_1, v_1), \dots, \beta_n(u_1, v_1)$  = estimated regression coefficients for segment 1

$AADT_1$  = average annual daily traffic (vehicles per day) for segment 1

$L_1$  = length (miles) for segment 1

$V_1$  = speed measure (mph) for segment 1

$X_{n,1}$  = other geometric variables for segment 1 (if considered in the specific analysis)

For segment 1, the coefficients in Equation (30) and Equation (34) are estimated using the MLE method as below (I):

$$\begin{aligned}
& \text{LL}(\gamma(u_1, v_1), \beta(u_1, v_1)) \\
& = \begin{cases} \sum_{l=1}^l \left( \ln \left( e^{X_{m,l}^T \gamma(u_1, v_1)} + e^{-e^{X_{m,l}^T \beta(u_1, v_1)}} \right) - \ln \left( 1 + e^{X_{m,l}^T \gamma(u_1, v_1)} \right) \right) w_{1l}(u_1, v_1), & y_l = 0 \\ \sum_{l=1}^l \left( \left( y_l X_{m,l}^T \beta(u_1, v_1) - e^{X_{m,l}^T \beta(u_1, v_1)} \right) - \ln(y_l!) \right) w_{1l}(u_1, v_1), & y_l > 0 \end{cases} \quad (35)
\end{aligned}$$

Where,

LL = log-likelihood

$l$  = index of the segments considered for the local model and in this case, the segments are segment 1 to  $l$

$y_l$  = number of crashes on  $l^{th}$  segment

$X_{m,l}^T$  = column vector of the variables AADT, length, speed measure, etc. for  $l$  segments

$w_{1l}(u_1, v_1)$  = weight function to describe the influence of the neighbor segments around the target segment 1 and is estimated using Equation (28).

As previously described in Section 4.4.2.2.1, the optimum bandwidth,  $h$ , is determined through the minimization of CV score. The final local models are fitted using the neighboring segments corresponding to the optimum bandwidth. Finally, the general form of the local model for each of the  $i$  segments can be written as,

$$\begin{aligned}
& \mu_i \\
& = (1 - p_1) e^{\beta_0(u_i, v_i) + \beta_1(u_i, v_i) AADT_i + \beta_2(u_i, v_i) L_i + \beta_3(u_i, v_i) V_i + \sum \beta_n(u_i, v_i) X_{n,i}} \quad (36)
\end{aligned}$$

#### 4.4.3 Machine Learning Model

As one of the ML modeling approaches, this study adopted RF regression model to predict the number of crashes. RF model is a supervised learning algorithm that utilizes a decision tree-based ensemble approach. It is a non-parametric model which can capture the non-linear effect of the explanatory variables on the model output. The model is made up of several decision trees. Each tree in the ensemble is built from a number of bootstrap training samples which are randomly drawn from the population data with replacement. Each tree provides prediction results using the testing data. The prediction results from the trees are then averaged. To avoid correlation between individual trees,

RF model uses a subgroup of the explanatory variables for splitting each node under each decision tree. The best split point for each node is determined by applying a splitting algorithm on the subgroup of the selected explanatory variables. The splitting algorithm produces a maximum homogeneity to the successive node at a particular value of a selected variable.

Some of the advantages of RF model are that it does not require any predefined functional form, it can address multicollinearity among the explanatory variables and provides the importance of the variables according to their contribution to model predictions (128; 129).

#### 4.4.3.1 Random Forest Model Calibration Process

The calibration process of RF model involves tuning a set of hyperparameters to obtain good prediction accuracy while reducing the overfitting or underfitting in the trained model. This study controls for the five hyperparameters as presented in Table 5 to calibrate RF model. The author chose these candidate hyperparameters to tune by following Probst et al., Han et al., and Parmar et al. (130-132).

Each of the hyperparameters has its own importance in model predictions. A larger value of `n_estimators` can increase the accuracy; however, the accuracy may no longer be affected by an increasing `n_estimators` after a certain level. Following Saha et al., this study adopts 500, 1000, 5000, and 10,000 for `n_estimators` (100). In case of `max_features`, this study tries  $\sqrt{p}$  in addition to  $p$  as suggested by Genuer et al. for low dimensional regression problems (133). For `max_depth`, the study applies the values as shown in Table 5. This is a critical hyperparameter in RF model as increasing the `max_depth` continuously may cause the overfitting issue in the trained model. To further prevent overfitting, this study also considers `min_samples_leaf` and `min_sample_split`. These hyperparameters can control the growth of the trees, therefore, reducing overfitting with the training data. The smallest value associated with these hyperparameters can end up with the largest tree. Therefore, this study considers other values as shown in Table 5 in addition to the default values of 1 and 2 respectively for `min_samples_leaf` and `min_sample_split`.

Table 5 Hyperparameters for RF Model Calibration

Hyperparameters	Description	Values Tried
n_estimators	Number of trees in the forest	500, 1000, 5000, and 10,000
max_features	Number of explanatory variables in each split	$p, \sqrt{p}$ *
max_depth	Maximum depth of tress	5, 10, 20
min_samples_leaf	Minimum number of samples in a terminal node	1, 2, 4
min_sample_split	Number of samples required to split a node.	2, 5, 10

\* $p$  = total number of explanatory variables

The best combination of the hyperparameters is obtained through CV process, which builds a number of models utilizing different combinations of hyperparameters. All the models are evaluated by CV. This study uses a 5-fold CV to evaluate each model and control overfitting in the models. The 5-fold CV splits the data into 5 stratified parts as illustrated in Figure 6. Each part successively is used as testing data for estimating prediction performance. The remaining data are used as a training set. For each fold, Mean Squared Error ( $MSE$ ) is calculated using Equation (37) and is averaged over the 5 folds. This 5-fold CV is performed for the models with different combinations of hyperparameters, and average  $MSE$  is obtained for individual models. Finally, the best combination of hyperparameters is reported from the model that estimates the lowest  $MSE$ .

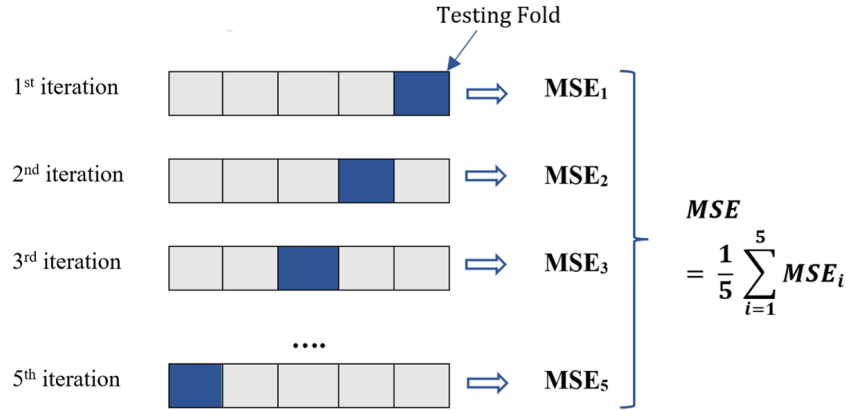


Figure 6 Demonstration of 5-fold Cross-Validation

$$MSE = \frac{1}{n} \sum_{\substack{i \in testing \\ data}}^n (y_i - \hat{y}_i)^2 \quad (37)$$

Where,

$MSE$  = mean squared error using the testing data

$y_i$  = observed value of the  $i^{th}$  observation in the testing data

$\hat{y}_i$  = predicted value of the  $i^{th}$  observation in the testing data

$n$  = number of observations in the testing data

#### 4.4.3.2 Variable Importance from Random Forest Model

After calibrating the RF model and developing the model with the best combination of hyperparameters, variable importance (VI) is measured. The purpose is to rank the explanatory variables. VI indicates the contribution of a variable to the output prediction when all other variables are present in the model. This study particularly used Mean Decrease in Accuracy (MDA) method to measure the VI. MDA measures how much the model accuracy decreases when the testing data of each variable are permuted. If the variable is important, the model accuracy will be highly altered and decreases significantly after permutation. Then, the variables can be ranked according to the mean accuracy decrease. As the accuracy measure,  $MSE$  is calculated for testing data using

Equation (37). For each explanatory variable,  $MSE$  is calculated before and after permutation. The differences between before and after permutation  $MSE$  are averaged over all the trees. Equation (38) shows the VI calculation of an explanatory variable based on the  $MSE$  for testing data (134).

$$VI = \frac{1}{n_{tree}} \sum_{k=1}^{n_{tree}} (EP_k - E_k) \quad (38)$$

Where,

$n_{tree}$  = number of trees in the forest

$E_t$  =  $MSE$  on  $k^{th}$  tree before permuting the values of the variable

$EP_t$  =  $MSE$  on tree  $k^{th}$  tree after permuting the values of the variable

$VI$  = variable importance

#### 4.5 Evaluation of Model Performance

To choose the best models or to evaluate the performance of the models under individual analysis, this study utilized several goodness of fit (GOF) measures. These measures are discussed below.

To determine the best model among a set of statistical models, Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC) are generally used. These can be calculated using Equation (39) and Equation (40) respectively. The smaller the value of AIC and BIC, the better is the model (135).

$$AIC = -2\log Q + 2K \quad (39)$$

$$BIC = -2\log Q + (\ln(i))K \quad (40)$$

Where,

$\log Q$  = log-likelihood of the residual sum of squares

$K$  = number of estimated parameters

$i$  = total number of observations.

To compare the performance of the NB and Poisson models against each other, a Likelihood Ratio (LR) test is performed (136). The null hypothesis is that the Poisson model is better than the NB model. The test statistic follows a chi-square distribution with a degree of freedom (Df) equal to the difference between the number of parameters in the NB and Poisson model, and it can be calculated from Equation (41). If the p-value for the test statistic is less than 0.5, then the null hypothesis is rejected.

$$\lambda = -2 [\text{Log}(\text{Poisson}) - \text{Log}(\text{NB})] \quad (41)$$

Where,

$\lambda$  = test statistic

Log(Poisson) = log-likelihood of Poisson model

Log(NB) = log-likelihood of NB model

To perform an objective assessment of the predictive performance of the statistical and ML models, additional measures are evaluated using data “unseen” by the models. These include Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), Mean Absolute Deviation (MAD), Generalized  $R^2$  and traditional  $R^2$ .

The MAPE estimates the absolute value of the error term as a percentage of the actual number of crashes that excludes the segments with actual zero crashes. Equation (42) shows the mathematical formula. The RMSE is calculated as the square root of the MSE term, which is an average of the square of the prediction error at each segment. Equation (43) shows the calculation. The MAD is the average of the absolute deviation of predicted crashes by the model from the actual crashes as expressed by Equation (44). A lower value for each of these measures implies better accuracy in model prediction.

$$MAPE = \frac{100}{n} \sum \left| \frac{(y_i - \hat{y}_i)}{y_i} \right| \quad (42)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (43)$$



$$MAD = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (44)$$

Where,

$y_i$  = actual number of crashes of  $i^{th}$  segment

$\hat{y}_i$  = predicted number of crashes  $i^{th}$  segment

$\bar{y}$  = mean of actual number of crashes

$n$  = number of segments

Traditional  $R^2$  measures the variance in the response variable that can be described by the explanatory variables in a regression model. It can be calculated from Equation (45). A higher value of  $R^2$  indicates a better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (45)$$

Where,

$y_i$  = actual number of crashes of  $i^{th}$  segment

$\hat{y}_i$  = predicted number of crashes  $i^{th}$  segment

$\bar{y}$  = mean of actual number of crashes

$n$  = number of segments

Generalized  $R^2$  value is calculated from the likelihood function  $Q$  by setting a scale of 1 as the maximum value. It simplifies the traditional  $R^2$  without requiring any specific distribution (e.g., normal distribution) of the response variable. It is calculated with Equation (46).

$$R^2 = 1 - \exp\left[-\frac{2}{i} \{ \text{Log } Q(\hat{\beta}) - \text{Log } Q(0) \} \right] \quad (46)$$

Where,

$\text{Log } Q(\hat{\beta})$  = log-likelihoods of the fitted model

$\text{Log } Q(0) = \text{log-likelihood of the null model with only the intercept}$

To further assess the performance of the models developed in this study, CURE plots are utilized. The goal is to graphically observe how well the models fit the dataset. Following the procedure in Hauer and Bamfo, CURE plots are developed by showing the cumulative residual (i.e. difference between the actual number of crashes and the predicted number of crashes from model) as the increasing order of each explanatory variable (116). The CUREs are treated as a random walk within a 95% confidence interval (CI) based on Equation (47). A cumulative residual curve that stays within 2 standard deviations ( $\pm 2\sigma$ ) 95% of time is considered to be satisfactory (21).

$$\sigma_{CURE} = \sigma_s(n) \sqrt{1 - \frac{\sigma_s^2(n)}{\sigma_s^2(i)}} \quad (47)$$

Where,

$\sigma_{CURE}$  = estimated variance of the random walk

$\sigma_s^2(n)$  = sum of the squared residual from 1 to n

$\sigma_s^2(i)$  = sum of the cumulative residuals over the total observations  $i$

#### 4.6 Summary

The identification of the factors affecting the number of crashes in the existing literature helped to select the potential variables for this study. Especially for the speed measures, current practice is mainly focused on the 85<sup>th</sup> percentile speed and speed limit for rural two-lane highways. These measures may not always represent the actual operating condition of these roads. It seems that additional speed measures should be investigated to properly link the operational conditions of the rural two-lane highways with their crashes. To do the investigation, different statistical and ML models discussed in this chapter can be utilized. After deciding on the final model forms based on the evaluation matrices, an idea can be obtained about the more representative speed measures. Analyzing the results from the model with the speed measures can provide further insights into the relationship between speed and crashes of these roads.

Furthermore, the findings from the analysis can be utilized to identify the appropriate countermeasures for minimizing crashes.

## CHAPTER 5. INVESTIGATING SIGNIFICANCE OF SPEED

This chapter documents the preliminary analysis of incorporating speed into the crash prediction model for rural two-lane highways and investigates the effect of speed on crashes utilizing measured data. The analysis can be separated into two sections based on the speed measures tested. The first section looks at several speed metrics associated with the segment from an operational perspective and identifies the representative speed metric for the crashes on these roads. This analysis will provide insights into how speed corresponding to a segment influences the crash occurrence in that specific segment. The latter section looked at the influence of speed on crashes from a design consistency perspective. This analysis will explore the idea that crash occurrence on a road may not only depend on the local design conditions of that particular road segment but also on the operating condition of the adjacent segment.

### 5.1 Influence of Speed from Operational Context

#### 5.1.1 Objective

The operating speed on rural two-lane highways may vary significantly from one location to another due to a wide range of factors. This section focuses on the role of speed in crash prediction models for these roads by linking measured speeds with volume and geometric information. Initially, to investigate the effect of speed on the crashes of these roads from an operational perspective, this section set up the research goal as:

- Incorporate speed measures into the crash prediction model and investigate whether speed is a significant factor for crashes on rural two-lane highways.

#### 5.1.2 Dataset and Speed Variable Selection

Dataset processed in Section 3.2 was utilized for this preliminary analysis. The dataset contains 44,008 segments with a total of 93,820 crashes aggregated from both directions of the road. As speed measures, Average Speed, the 85<sup>th</sup> Percentile Speed, Difference between Average Speed and Speed Limit, and Difference between the 85<sup>th</sup> Percentile Speed and Speed Limit were calculated for each direction of the road segments. These metrics were averaged from both directions of a segment. Each of these

speed variables was experimented in a ZINB (see Section 4.4.2.1.3) model together with AADT and L presented in Equation (48) to predict the expected number of crashes in 5 years.

$$\mu = e^{\varepsilon} \cdot AADT^{\beta_1} \cdot L^{\beta_2} \cdot e^{\beta_3 V} \quad (48)$$

In the model, AADT and L were natural log-transformed due to skewness in the distribution of these variables. No transformation of the speed measures was deemed necessary because of the normal distribution of the data associated with these variables.

Table 6 presents the descriptive statistics of the explanatory variables i.e., AADT, L, and speed measures considered for the crash prediction model development process in this study. In the study dataset, there were some places with low average speed. Further investigation of these places revealed that they are mostly lower functional class roads with narrow lanes and low speed can be possible. Further, the study data contains some segments from the lowest speed limit such as 10 mph. The database shows these segments as rural two-lane highways and such records can be rare.

Table 6 Descriptive Statistics of the Explanatory Variables

<b>Variables</b>	<b>Unit</b>	<b>Min.</b>	<b>Max.</b>	<b>Mean</b>	<b>Standard Deviation</b>
AADT	vehicle	2	19619	1456	1895
Segment Length (L)	mile	0.10	2.97	0.48	0.30
Average Speed ( $V_a$ )	mph	5.36	69.67	38.94	10.37
Speed Limit ( $V_{sp}$ )	mph	10	55		
The 85 <sup>th</sup> Percentile Speed ( $V_{85}$ )	mph	9.10	70	47.90	8.77
Difference between Average Speed and Speed Limit ( $V_a - V_{sp}$ )	mph	-	20.66	-	10.98
		49.64		14.07	

Difference between the 85 <sup>th</sup> Percentile Speed and Speed Limit ( $V_{85} - V_{sp}$ )	mph	-	32.78	-5.11	9.41
			45.87		

The following five models were evaluated in this study with the rural two-lane segments. The traditional form, with only AADT and L as explanatory variables, was included to provide a baseline for other models. This is to compare if the inclusion of speed as a factor in the crash prediction model helps to improve the prediction performance.

- (1) Model using AADT and L only
- (2) Model using AADT, L and  $V_a$
- (3) Model using AADT, L and  $V_{85}$
- (4) Model using AADT, L and ( $V_a - V_{sp}$ )
- (5) Model using AADT, L and ( $V_{85} - V_{sp}$ )

During the model development process, 75% of the dataset was used to train the model and 25% as the testing dataset. Table 7 summarizes all the experimented models with parameter estimates, AIC, BIC, Generalized  $R^2$ , RMSE, MAPE, and MAD values. Compared to the traditional model, which only includes AADT and L, models with speed measures seem to fit the data better based on AIC and BIC values. Further, speed measures are significant at a 5% significance level in each model. Model (3) with the 85<sup>th</sup> Percentile Speed measure seems to show the lowest error, with Model (2) with Average Speed as the close second. The 85<sup>th</sup> percentile speed is commonly used in safety assessment in highway design considerations (65), thus Model (3) may fit better for design applications. However, a large amount of data is needed to ensure a reliable value of the 85<sup>th</sup> percentile speed. Considering the fact that Average Speed is a better representation of the realistic operating condition for rural two-lane highway facility type, this analysis chose Model (2) as shown in Equation (49) to proceed with the subsequent analyses.

Table 7 Model Parameters and Goodness-of-Fit

Mode I	Parameter Estimate				AIC	BIC	R <sup>2</sup>	RMS E	MAP E (%)	MA D
	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$						
(1)	-		0.9		10696	10700	0.44	3.71	68.48	1.61
	4.1	0.8	8		0	2	6			
	8	1								
(2)	-		1.0	-	10661	10666	0.45	3.66	66.86	1.59
	4.0	0.8	2	0.0	7	8	2			
	9	9		1						
(3)	-		1.0	-	10648	10653	0.45	3.65	66.78	1.58
	3.6	0.8	3	0.0	7	8	4			
	9	8		2						
(4)	-		0.9	-	10685	10690	0.44	3.70	67.98	1.61
	4.6	0.8	9	0.0	0	1	8			
	2	6		1						
(5)	-		0.9	-	10680	10685	0.44	3.70	68.08	1.61
	4.5	0.8	9	0.0	3	3	9			
	2	5		1						

\*Note:  $p$ -value  $< 0.0001$  for all the variables.

$$\mu = e^{-4.09} \cdot AADT^{0.89} \cdot L^{1.02} \cdot e^{-0.01V_a} \quad (49)$$

### 5.1.3 Incorporating Speed for Better Performance

In Equation (49), AADT and L are significant and positively correlated with the number of crashes, as expected. Average Speed factor in the model is negatively associated with the total number of crashes. This finding is consistent with a recent study by Dutta and Fontaine (27). The negative association can also be observed from the marginal model plots that show the direction of responses with respect to an explanatory variable where all other variables are set to their mean value. Based on the marginal

model plots, Figure 7 shows that the number of crashes is decreasing with an increase in Average Speed. Conversely, the number of crashes is increasing with AADT and L.

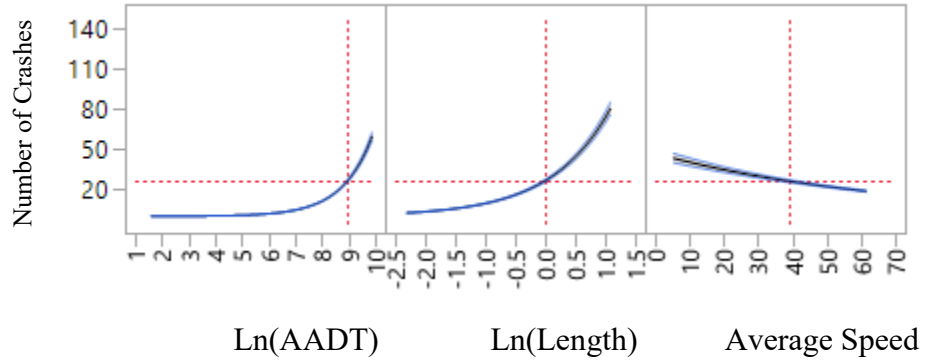


Figure 7 Marginal Model Plots for Model (2)

The observed negative association was further confirmed by normalizing the crash data in terms of vehicle miles traveled (VMT) using AADT and L. The normalized number of crashes showed a decreasing trend with increasing Average Speed. Especially for the crashes at the higher average speeds, even though the total number of crashes can be higher due to the presence of high volume, the crashes are actually low while the other factors, i.e., AADT and L remain constant.

The performance of Model (2) was further assessed using cumulative residual (CURE) plots. Figure 8 shows the CURE plots for Model (2). The appropriateness of the functional form of the model was assessed through the CURE plots for the explanatory variables i.e. AADT, L, and Average Speed. Clearly, a significant portion of the cumulative residual is outside the boundary of  $\pm 2\sigma$ , indicating that the model does not fit the data very well for all the explanatory variables. It seems that the model is highly over-predicting or under-predicting, especially in the higher speed and higher AADT ranges. Hence, the model is not fitting well for the data that vary widely, especially in terms of Average Speed of the segments. These observations prompt to consider a different approach using speed as a categorizer, discussed in the next section.



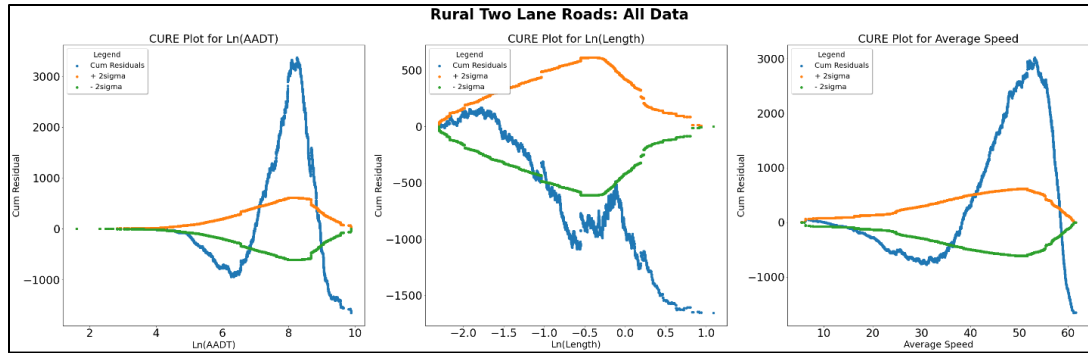


Figure 8 CURE Plots for Model (2)

#### 5.1.3.1 Speed Categorizer

This section explores how to best incorporate speed in developing crash prediction models. From Figure 8, it has already appeared that the model is gradually overpredicting the number of crashes up to an average speed of 30 mph. After 30 mph, the model starts underpredicting, which continues to 50 mph. From the transitions of the CURE plot with respect to Average Speed, there are three evident regions of speed for which the dataset can be grouped to develop three separate models. Therefore, the dataset was split into these three-speed ranges labeled as low, medium, and high speed, respectively. They represent about 21%, 61%, and 18% of the total segments, respectively.

ZINB-based crash prediction model was developed for individual speed ranges. The effect of speed was evaluated at each speed level. The following sub-sections discuss the significance of speed variable in the model and how speed affects the number of crashes differently for various speed ranges.

##### 5.1.3.1.1 LOW-SPEED ROADS

The low-speed roads are comprised of segments where the average speed is below 30 mph. There are 9,371 segments in this category. Seventy-five percent of these segments were used as training samples to develop model for these roads. Among the three explanatory variables, AADT and L are significant, while Average Speed is not in the model. Table 8 shows the final model specification.

Table 8 Model for Low-Speed Category

	<b>Estimate</b>	<b>Std. Error</b>	<b>95% CIs</b>
<b>Intercept (<math>\epsilon</math>)</b>	-4.95	0.107	(-5.16, -4.74)
<b>Ln(AADT)</b>	0.93	0.019	(0.89, 0.97)
<b>Ln(L)</b>	0.92	0.033	(0.85., 0.98)
<b>Model Form</b>	$\mu = e^{-4.95} \cdot AADT^{0.93} \cdot L^{0.92}$		
<b>MAPE</b>	59.79%		
<b>RMSE</b>	1.64		
<b>MAD</b>	0.86		

\*Note: *p*-value < 0.0001 for all the variables.

One way to quantify the contribution of each variable in the model is to measure the importance of the variables. Equation (50) presents a means of estimating the importance of an explanatory variable.

$$\text{Variable Importance} = \frac{\text{Var}(E(\frac{y}{X}))}{\text{Var}(y)} \quad (50)$$

Here,  $\text{Var}(E(\frac{y}{X}))$  is calculated from the expected number of crashes,  $y$ , with respect to the conditional distribution of all variables considered, and the variance is taken over the distribution of variable  $X$ . In the model for low-speed roads, the importance of AADT and L are 68% and 32% respectively.

#### 5.1.3.1.2 MEDIUM-SPEED ROADS

The medium-speed category contains segments with an average speed ranging between 30 mph and 50 mph. The number of segments under this category is 27,075. Two different models were fitted for this category. One is traditional AADT and L only, and the other includes Average Speed along with AADT and L. Table 9 presents the

specifications and performances of these two models. AADT and L are significant at a 5% significance level in both models. Moreover, Average Speed is statistically significant according to the model with speed showing a p-value of less than 0.0001.

Table 9 Model comparison for Medium-Speed Category

	Model Without Speed			Model With Speed		
	Estimate	Std. Error	95% CIs	Estimate	Std. Error	95% CIs
<b>Intercept (<math>\epsilon</math>)</b>	-4.58	0.057	(-4.69, -4.47)	-4.32	0.067	(-4.46, -4.19)
<b>Ln(AADT)</b>	0.88	0.008	(0.87, 0.90)	0.91	0.009	(0.89, 0.92)
<b>Ln(L)</b>	1.06	0.013	(1.03, 1.08)	1.07	0.013	(1.05, 1.09)
<b>Average Speed</b>	–			-0.01	0.001	(-0.01, -0.007)
<b>Model Form</b>	$\mu = e^{-4.58} \cdot AADT^{0.88} \cdot L^{1.06}$			$\mu = e^{-4.32} \cdot AADT^{0.91} \cdot L^{1.07} \cdot e^{-0.01V_a}$		
<b>MAPE</b>	61.04%			60.89%		
<b>RMSE</b>	2.85			2.84		
<b>MAD</b>	1.52			1.52		

\*Note: p-value < 0.0001 for all the variables.

Although Average Speed is significant in the model with speed, its importance is quite low. Its importance factor is only 1%, while AADT and L have 59% and 40%, respectively. It seems that the effect of speed is trivial, which is corroborated by the marginal model plots in Figure 9, where the plots are ordered according to the importance of the variables in the model. The figure shows that the number of crashes is not changing considerably with the predictor Average Speed, whereas, other variables are showing a noticeable influence on the changes in the number of crashes.

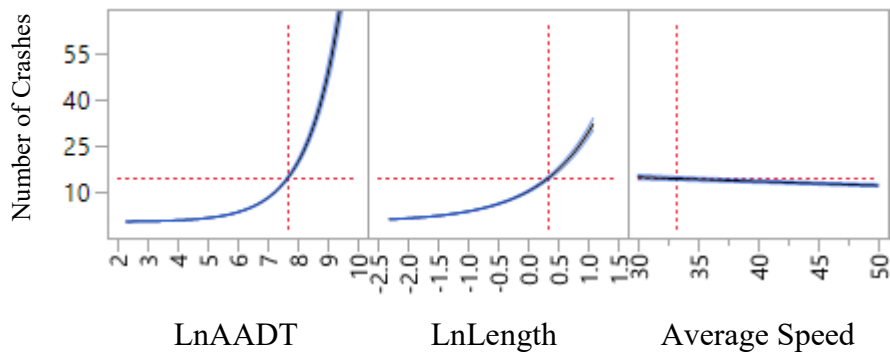


Figure 9 Marginal Model Plots for Medium Speed Roads

This observation suggests that excluding Average Speed from the model may not degrade the model’s performance in any significant way, as corroborated by the performance indicators in Table 9. Nonetheless, having one less variable can reduce model complexity.

As this analysis moves forward with the model without speed for medium-speed roads, CURE plots, shown in Figure 10, were constructed with respect to AADT and L. The plots suggest that the data perhaps should be further divided to improve the model fit. Based on the observation, the consistent under-prediction turns into a consistent over-prediction when Ln(AADT) is roughly 8, which corresponds to an AADT value of approximately 3000. Using this value as a threshold, this dataset with medium-speed range was further split into the low-volume and high-volume sets.

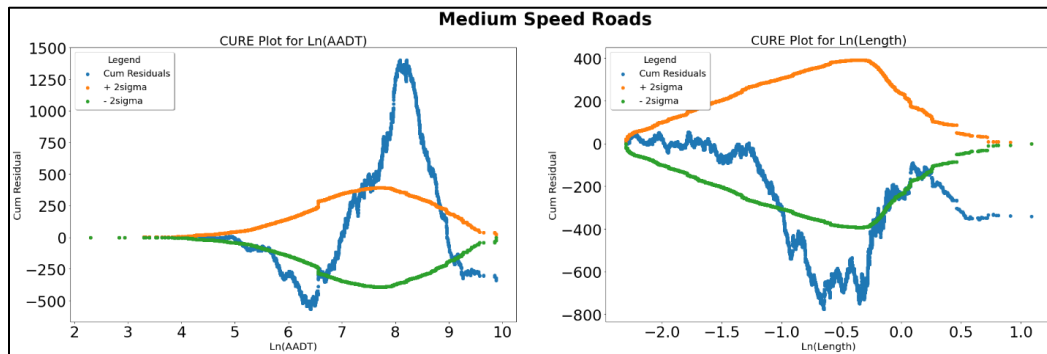


Figure 10 CURE Plots with  $\pm 2\sigma$  for the Explanatory Variables in Medium Speed Model

To understand whether considering AADT as another level of categorizer can improve the performance of the models, this study tested another two sub-models separately for low-volume and high-volume roads. These are Low Volume sub-model and High Volume sub-model. The models were built using the same ZINB formulation incorporating AADT and L in the models. The specifications of these models are presented in Equation (51) and Equation (52).

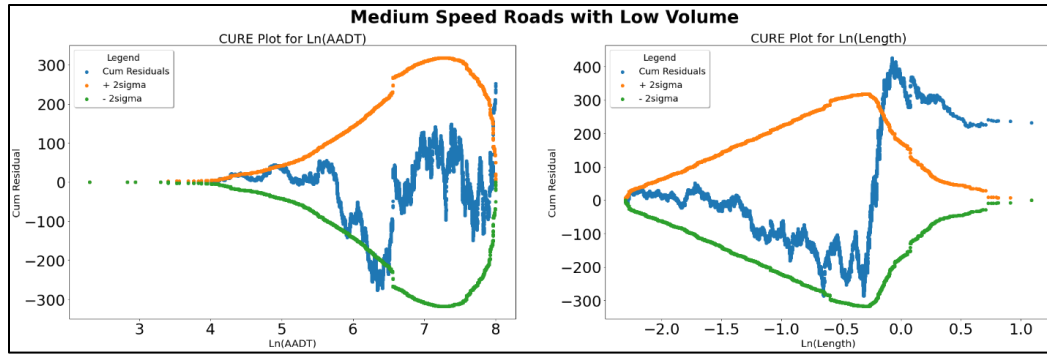
$$\text{Low Volume sub-model: } \mu = e^{-4.99} \cdot AADT^{0.95} \cdot L^{1.07} \quad (51)$$

$$\text{High Volume sub-model: } \mu = e^{-3.31} \cdot AADT^{0.72} \cdot L^{0.99} \quad (52)$$

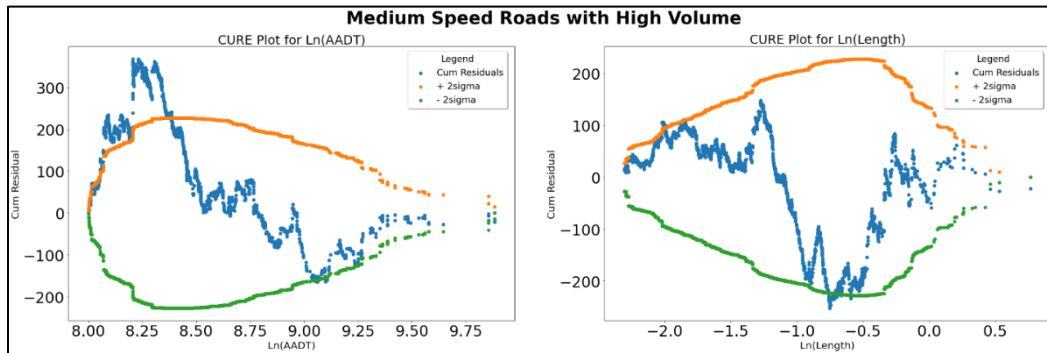
Table 10 shows the prediction performance of these models using the testing datasets. After splitting the segments of medium speed ranges in terms of volume, the combined performance of the models is slightly better than the single model. Moreover, CURE plots fit better after considering AADT categorizer-based separate models for medium-speed roads as shown in Figure 11.

Table 10 AADT Categorizer-based Models and Comparison

<b>Models tested for Medium Speed Roads</b>	<b>No of Segments for Training</b>	<b>No of Segments for Testing</b>	<b>MAPE</b>	<b>RMSE</b>	<b>MAD</b>
Low Volume sub-model	18,342	6,114	60.81%	2.29	1.31
High Volume sub-model	1,964	654	80.26%	5.00	3.18
Combination of sub-models	20,307	6,768	63.03%	2.73	1.50
Single model	20,307	6,768	61.04%	2.85	1.52



(a) Low Volume Roads



(b) High Volume Roads

Figure 11 CURE Plots with  $\pm 2\sigma$  for the Models of Medium-Speed Roads

### 5.1.3.1.3 HIGH-SPEED ROADS

Roads with an average speed of 50 mph are referred to the high-speed roads in this analysis. The number of segments under this category is 7,561. Two models were developed separately for these segments. One followed the traditional form, and the other included AADT, L, and Average Speed. Table 11 shows all the significant variables for each model. Evidently, Average Speed becomes statistically significant for the crashes of high-speed roads at a 5% significance level in the model with speed. The association between Average Speed and number of crashes was found as negative, which is also evident from the marginal model plots in Figure 12. After analyzing the dataset, it was observed that roads in the high-speed category are the ones with better geometric conditions, for example, wider lanes, presence of shoulders, etc. Furthermore, the importance of Average Speed is 8% in the model while AADT and L are of 52% and

40% importance, respectively. It implies that the influence of speed on the crash predictions for high-speed roads is more profound than that for other roads.

Including speed in the crash prediction model shows improved performance over the traditional model without speed. All performance measures shown in Table 11 are better when including Average Speed in the model.

Table 11 Model comparison for High-Speed Category

	Without Speed			With Speed		
	Estimate	Std. Error	95% CIs	Estimate	Std. Error	95% CIs
<b>Intercept (<math>\epsilon</math>)</b>	-2.96	0.136	(-3.23, -2.69)	1.12	0.255	(0.62, 1.62)
<b>Ln(AADT)</b>	0.62	0.017	(0.59, 0.65)	0.73	0.018	(0.69, 0.77)
<b>Ln(L)</b>	0.96	0.019	(0.92, 0.99)	0.98	0.019	(0.95, 1.03)
<b>Average Speed</b>	-			-0.09	0.005	(-0.10, -0.08)
<b>Model Form</b>	$\mu = e^{-2.96} \cdot AADT^{0.62} \cdot L^{0.96}$			$\mu = e^{1.12} \cdot AADT^{0.73} \cdot L^{0.98} \cdot e^{-0.09V_a}$		
<b>MAPE</b>	75.23%			71.06%		
<b>RMSE</b>	7.23			7.16		
<b>MAD</b>	2.51			2.39		

\*Note:  $p$ -value < 0.0001 for all the variables.

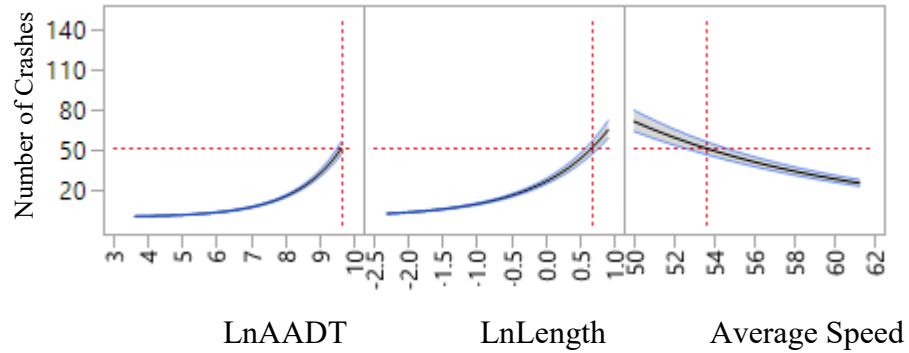


Figure 12 Marginal Model Plots for High-Speed Roads

Although the CURE plots for the high-speed roads show overprediction after an AADT of approximately 5,000, this analysis did not further categorize the dataset based on AADT because the number of segments in the high-speed range is rather limited. As more data become available over time, this analysis can be revisited in the future.

#### 5.1.3.2 Overall Performance

The performance of the speed and AADT categorizer-based models was compared with initially developed Model (2) to demonstrate how the separate models better predict the number of crashes for rural two-lane highways than Model (2). For this purpose, the predictions from all the speed and AADT categorizers-based models (i.e., low speed, medium speed, and high-speed road models) were combined, and error measures were estimated. Table 12 shows a comparison of the combined errors with Model (1) and Model (2). The comparison shows that consideration of speed as a categorizer and further breaking down the model based on AADT improves the overall model performance by reducing the error up to a maximum of 11.3%. Utilizing the actual dataset for calculating speed measures as well as considering speed and AADT as the categorizers, this study demonstrated improvement in the performance of the crash prediction model for rural two-lane highways.



Table 12 Performance Comparisons

	MAPE	RMSE	MAD
<b>Model (1)</b>	68.48%	3.71	1.61
<b>Model (2)</b>	66.86%	3.66	1.59
<b>Combined Models with Speed and AADT as Categorizers</b>	64.49%	3.29	1.50

#### 5.1.4 Findings and Significance of the Analysis

This analysis investigated the effect of measured speed on the crashes of rural two-lane highways by directly using the actual speed dataset. The investigation found that speed is a significant factor considering all the rural two-lane segments in this study. However, the role of speed differs for highways with different speed ranges. The speed is insignificant for low-speed roads, whereas, it is statistically significant but negligible on medium-speed roads, and more profound on high-speed roads. This indicates that speed becomes a significant factor for the crashes of rural two-lane highways from lower to higher speed ranges, and the effect of speed is more evident for the crashes occurring in the high-speed range.

This study also revealed that adding another categorizer level i.e., AADT along with speed and separating the model into AADT sub-groups under each speed category yields better results than one model. If data are adequate for separate models, both speed and AADT can be considered as the categorizers when developing crash prediction models for rural two-lane highways.

Another finding from this study was that speed is negatively correlated with crashes of rural two-lane highways. This negative association is generally consistent with existing studies that found higher average speeds are associated with a lower number of crashes (10; 27; 28; 30; 48). A possible explanation is that rural two-lane highways with higher speeds tend to be those main corridors in the region that often have better geometric conditions (28).

Overall, the findings indicate that the influence of speed on crashes may vary depending on the speed category of rural two-lane segments. This result can be utilized

by DOTs and agencies. For the safety assessment of these roads, they can adopt the approach of separating the crash prediction model for different speed ranges.

This analysis had limitations in terms of the dataset. Even though the 85<sup>th</sup> percentile speed-based model in Table 7 was the best model considering the predictive performance, this analysis did not select it as calculating the 85<sup>th</sup> percentile speed requires a large amount of dataset. This study will further look at the 85<sup>th</sup> percentile speed model when more speed data become available in the future. Moreover, the dataset contained some low functional class roads with lower average speeds although the speed limits from HIS database were 55 mph. It requires further verification of the HIS database and revisiting the models. In addition, some of the average speeds of the roads seemed to be affected by conflation issue of the speed network. In future, this type of issue will be further investigated to see how it can affect the accuracy of the crash prediction models.

## 5.2 Influence of Speed from Consistency Context

Design consistency indicates the conformance of highway geometry to driver's expectation (23). Sudden changes in operating speed over the adjacent road elements can be avoided with a consistent design. Inconsistency in the design may violate driver's expectation, and a driver might choose an inappropriate speed that may lead to an accident. Therefore, design consistency is an important factor for road safety (23; 25).

The 85<sup>th</sup> percentile speed is one of the commonly used candidate measures of design consistency because it reflects the behavior of most of the drivers, especially in the curve segments (118). For a particular section of the road, difference between the 85<sup>th</sup> percentile speed and design speed is generally used to identify inconsistencies in the design of that section, whereas, the difference of the 85<sup>th</sup> percentile speed between consecutive sections can identify the inconsistencies that a driver may experience when traversing from one section to another. The latter is a good measure to understand the crashes that vary with the changes in the degree of curvature on the horizontal curves of rural two-lane highways (21). Moreover, it is one of the safety criteria suggested by Lamm et al. (118).

The crash risk involved with the design inconsistency of a highway can be assessed through crash prediction models by considering the design consistency variables, which allows the incorporation of human factors in assessing safety (53). For rural two-lane highways, several studies incorporated geometric design consistency measures in the crash prediction model and have found a significant influence of the consistency measures on the crashes (21-25; 44; 53; 57; 58; 64; 67; 119). According to Lamm et al., 50% of crashes on rural two-lane highways result from inconsistency in speed, which further implies the importance of evaluating the relationship between crashes and design consistency of these roads (118).

In the first section of this chapter, speed metric-based (such as average speed of a segment) models were developed to explore how the speed of a rural two-lane segment affects the total number of crashes specific to that road. This section investigates the effect of speed differential, i.e. the difference between the 85th percentile speeds of two consecutive segments, on the number of crashes on rural two-lane highways. The measure reflects the fact that crash occurrence may not only depend on the local design conditions of a particular road segment but also depend on the condition of adjacent segments (53).

### 5.2.1 Objective

This analysis investigates the relationship between speed differential and the total number of crashes on rural two-lane highways. The main objectives are:

- Investigate the effect of speed differential in predicting crashes of the rural two-lane segments in this study.
- Compare the crash prediction model incorporating speed differential with the ones considering average speed, the 85<sup>th</sup> percentile speed.

### 5.2.2 Dataset and Variable Selection

Before utilizing the dataset processed in Section 3.2, this specific analysis performed an additional investigation on the dataset. The goal was to check that each HERE link is at least associated with the homogenous segments from the same curve class (see Appendix 1). Generally, the difference in the 85<sup>th</sup> percentile speed can be

observed when traveling from tangent to curve, curve to curve, or curve to tangent. If a HERE link consists of multiple homogeneous segments i.e., the HERE link does not break at least during the change in curvature class, the difference between the 85<sup>th</sup> percentile speed may not be captured for two consecutive homogeneous segments with different degrees of curvature. This study observed such an issue in the dataset. Figure 13 shows some examples related to this issue. These are discussed below. Note that, the red lines represent HERE links and the black lines represent the homogeneous segments.

- In Figure 13(a), HERE Link ID 1225796539 consists of homogeneous segments with curve classes ranging between A and F.
- In Figure 13(b), HERE Link ID 879826964 includes homogeneous segments with curve classes ranging between A and D
- Figure 13(c), HERE Link ID 773386625 is associated with curve Class A and Class D.

The examples clearly indicate that HERE links consist of multiple homogeneous segments with varying curve classes. This study excluded those segments and only included the ones which were not affected by such issues associated with HERE links for experimenting with the effect of speed differential on the number of crashes. Furthermore, if any portion of the unique route was affected by the HERE link issue, the whole unique route is excluded from the dataset. The reason is that the unique routes should include continuous segments for the speed differential analysis. The filtering process resulted in 303 unique routes out of a total 3,700 unique routes. The 303 unique routes correspond to 7,909 homogeneous segments. Later, the segments were aggregated based on the same degree of curvature. It means that if the degree of the curvature for the consecutive segments is the same and no intersection is present between them, the segments are merged into a single segment as shown in Figure 14. The red box shows the segments that are aggregated into one segment based on the same degree of curvature. In addition, length weighted average was used to aggregate the associated roadway attributes and crashes were summed up regardless of the travel direction. The overall aggregation process resulted in a total of 5,182 segments with a total of 8,279 crashes aggregated from both directions of the road.

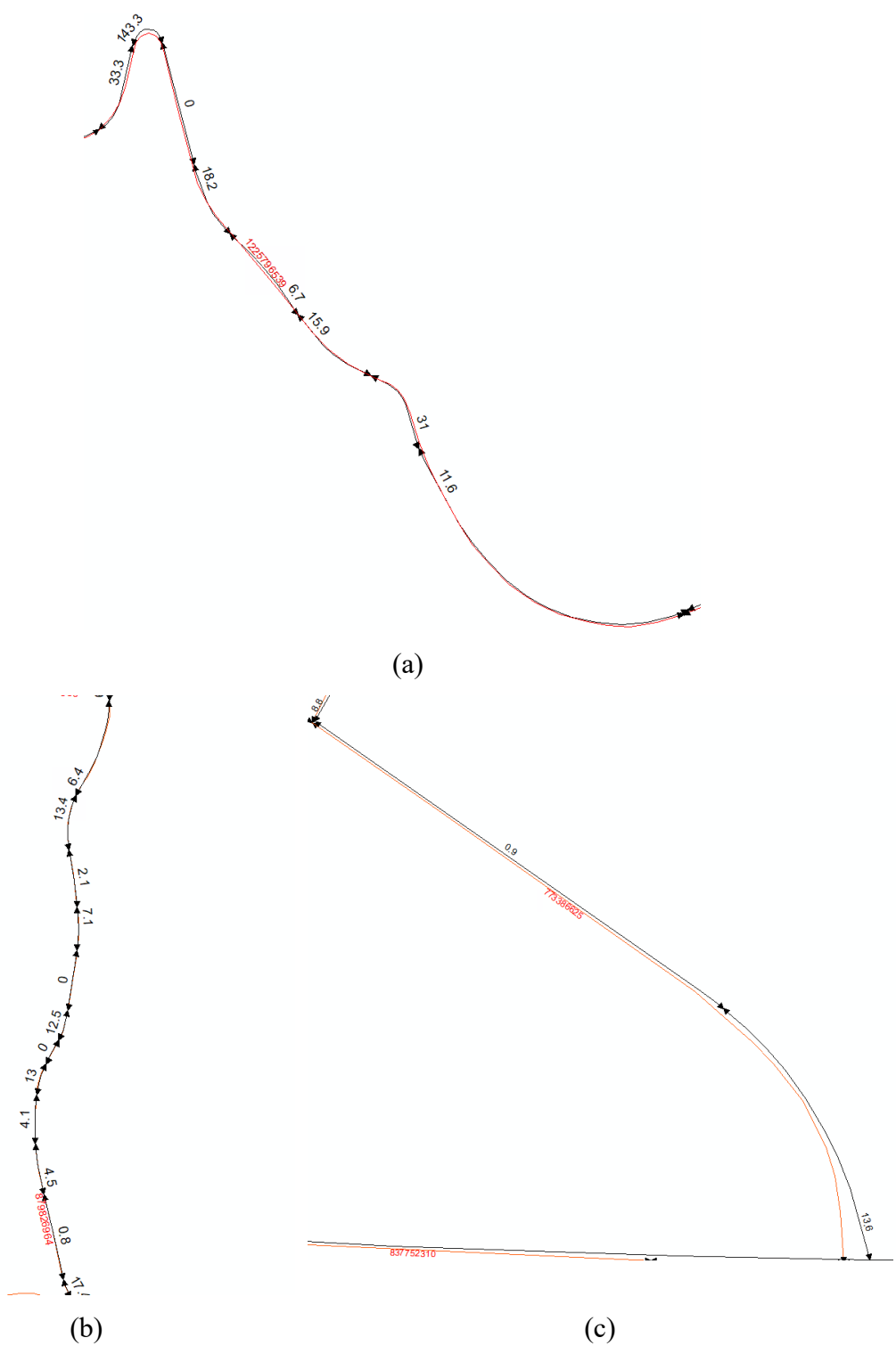


Figure 13 HERE Link Issue

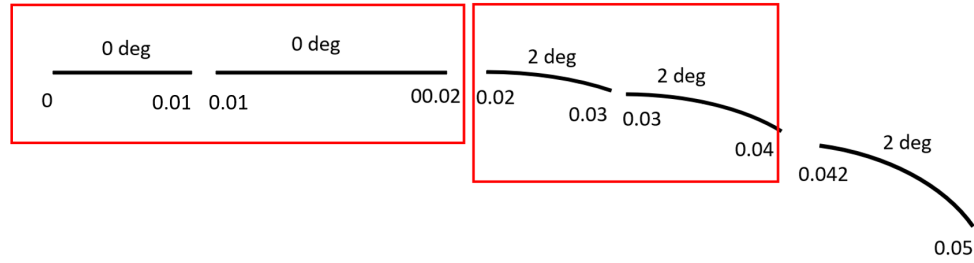


Figure 14 Aggregation based on Degree of Curvature

Explanatory variables considered for this analysis are presented in Table 13. These are AADT, L, Speed Differential, Degree of Curvature, Average Speed, and the 85<sup>th</sup> Percentile Speed of the study segments. Speed Differential ( $\Delta V_{85}$ ) was calculated as the absolute difference between the 85<sup>th</sup> Percentile Speed of consecutive segments. Figure 15 shows an example of consecutive segments where the mile point of the segments increases from left to right. In the direction of increasing mile point,  $\Delta V_{85}$  for segment 2 can be calculated as  $|V_{85,2} - V_{85,1}|$ . Note that, the 85<sup>th</sup> Percentile Speed was calculated for each direction of the road from the probe speed dataset. The average for the 85<sup>th</sup> Percentile Speed from both directions was used in determining  $\Delta V_{85}$ . As the response variable, number of crashes in 5 years was used.

Table 13 Summary Statistics of the Variables

Variables	Unit	Statistics			
		Min.	Max.	Mean	Standard Deviation
AADT	vehicle	14	19619	3854	3156
Segment Length (L)	mile	0.001	1.95	0.20	0.21
Speed Differential ( $\Delta V_{85}$ )	mph	0	40.73	0.98	2.55
Degree of Curvature ( $Cu$ )	degrees	0	79.1	0.98	2.13

Average Speed ( $V_a$ )	mph	4.68	60.95	49.04	10.87
The 85 <sup>th</sup> Percentile Speed ( $V_{85}$ )	mph	11.72	66.93	55.98	8.91
Number of Crashes in 5 years		0	81	1.60	3.67

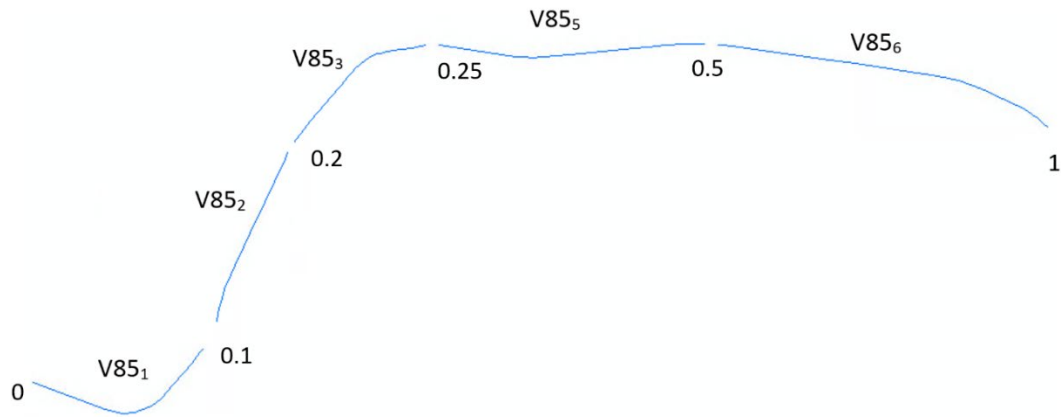


Figure 15 Consecutive Rural Two-Lane Segments in the Direction of Increasing Mile Points

To check the correlations among the variables, this analysis adopted Spearman's correlation method discussed in Section 4.2.2. Table 14 shows the results from Spearman's correlation test. The correlations between number of crashes and the explanatory variables i.e., AADT, L, Speed Differential, and Degree of Curvature were found significant at a 5% level. These explanatory variables were included in the model development considering speed differential.

Table 14 Spearman's Correlation Test

Variable	by Variable	Spearman $\rho$	p-value
AADT	L	-0.0174	0.2106
Number of Crashes	L	0.4976	<.0001*
Number of Crashes	AADT	0.3334	<.0001*
$Cu$	L	-0.2121	<.0001*
$Cu$	AADT	-0.0489	0.0004*
$Cu$	Number of Crashes	-0.1295	<.0001*
$\Delta V_{85}$	L	-0.0122	0.3816
$\Delta V_{85}$	AADT	-0.0302	0.0296*
$\Delta V_{85}$	Number of Crashes	0.0364	0.0088*
$\Delta V_{85}$	$Cu$	-0.1196	<.0001*

Note: (\*) indicates significance at a 5% level

### 5.2.3 Analysis and Results

To develop speed differential-based crash prediction model, four count models were explored. These are Poisson, NB, ZIP, and ZINB (see Section 4.4.2.1). To choose the best-fitted model, AIC, BIC, RMSE, MAD, and MAPE were utilized. An 80% of the 5,182 rural two-lane segments was used to train the models, and 20% for testing. Initially, AADT, L, Speed Differential, and Degree of Curvature were included in the models. For each model form, the Degree of Curvature was identified as statistically insignificant. Therefore, it was excluded from the model. Table 15 shows all the experimented model forms, parameter estimates significant at a 5% level, and performance measures. The table shows that AIC and BIC values are the lowest for NB model. The rest of the performance measures are similar among these four models tested.

CURE plots were utilized to further evaluate the models, as shown in Figure 16. The CURE is changing with respect to AADT, L, and Speed Differential. For unbiased estimation of crashes, the CURE should be within the boundaries of two standard deviations,  $\pm 2\sigma$ . From Figure 19, ZIP shows residuals outside of the boundary after around an AADT of 5,500, a length of 0.4 miles, and a speed differential of 5 mph. The

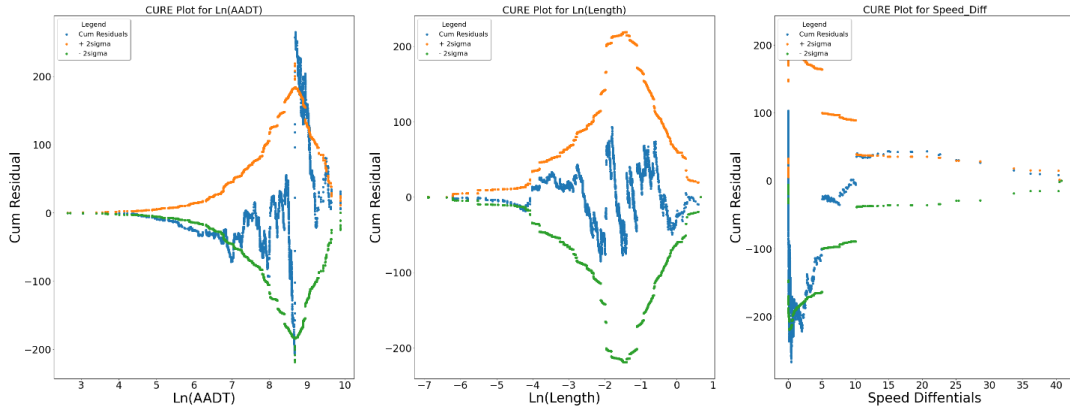


overall magnitude of the residual corresponding to each variable is higher than in the other models. The CURE plots for NB and ZINB models seem to be comparable implying a similar model fit. These models have a comparatively smaller magnitude of the residuals compared to ZIP model considering all the explanatory variables. In case of Poisson model, the lowest magnitude of residuals is observed for each variable in the model.

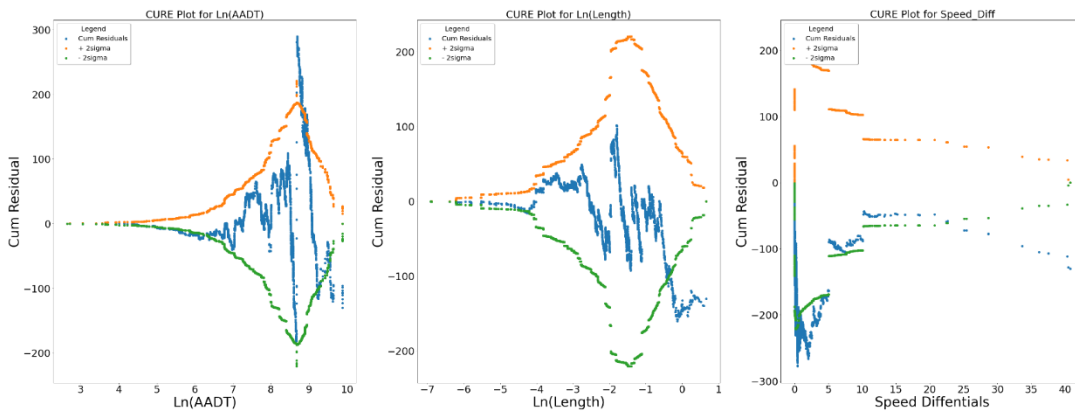
Table 15 Parameter Estimates and Performance Measures

<b>Crash Prediction Models based on Speed Differential</b>								
<b>Variables</b>	<b>Poisson</b>		<b>NB</b>		<b>ZIP</b>		<b>ZINB</b>	
	<b>Estimat e</b>	<b>Std. Erro r</b>	<b>Estimat e</b>	<b>Std. Erro r</b>	<b>Estimat e</b>	<b>Std. Erro r</b>	<b>Estimat e</b>	<b>Std. Erro r</b>
<b>Intercept, <math>\epsilon</math></b>	-3.901	0.131	-4.306	0.199	-3.640	0.139	-4.306	0.199
<b>Ln (AADT)</b>	0.696	0.015	0.747	0.024	0.673	0.016	0.747	0.024
<b>Ln (L)</b>	0.744	0.014	0.762	0.022	0.675	0.015	0.762	0.022
<b><math>\Delta V_{85}</math></b>	0.027	0.004	0.042	0.009	0.031	0.004	0.042	0.009
<b>AIC</b>	15248.02		11916.75		14664.86		11918.76	
<b>BIC</b>	15273.33		11948.39		14696.49		11956.71	
<b>RMSE</b>	2.67		2.67		2.68		2.67	
<b>MAPE (%)</b>	62.76		64.83		59.60		64.83	
<b>MAD</b>	1.21		1.21		1.20		1.21	

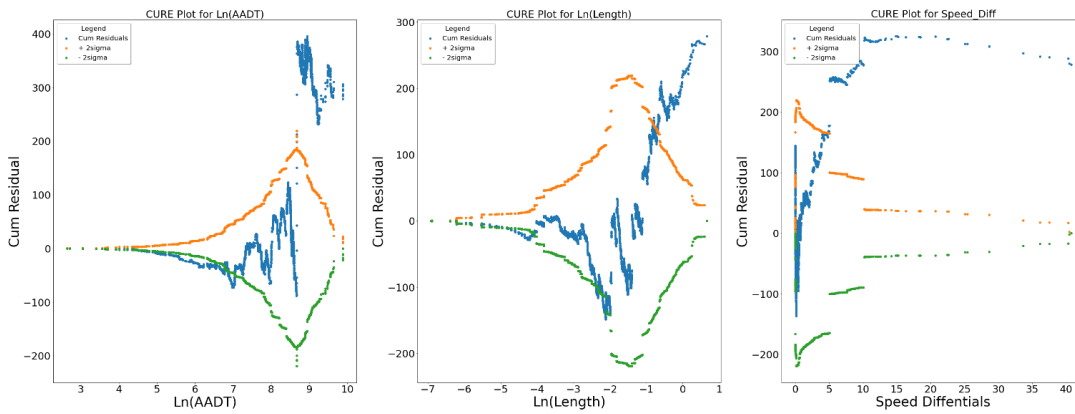
Note: all the co-efficient significant at a 5% level



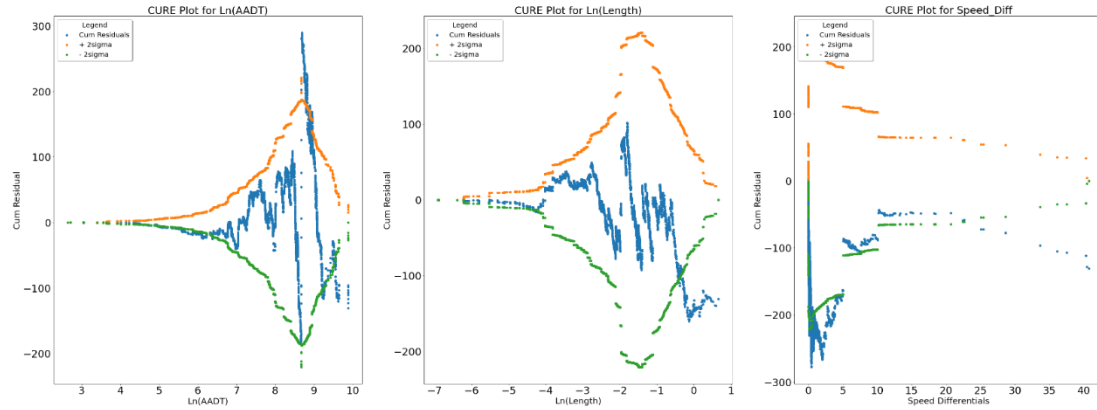
(a) Poisson Model



(b) NB Model



(c) ZIP Model



(d) ZINB Model

Figure 16 CURE Plots for Speed Differential -based Models

Considering the performance measures in Table 15, NB seems to be the best model, whereas, assessment of the CURE plots in Figure 16 shows Poisson model as the best one. To compare the performance of the NB and Poisson models against each other, an LR test was performed, shown in Table 16. According to the test, the NB model is significantly better than the Poisson model. NB model is selected as the final model for further analysis. This is consistent with existing research by Dhahir and Hassan where the author had NB model as the best performing model after all the assessments (67). Equation (53) shows the model form of NB model.

Table 16 Likelihood Ratio Test

Model	Df	Log-Likelihood	$\lambda$	Pr(> $\lambda$ )
Poisson	4	-7620.0		
NB	5	-5953.4	3333.3	<2.2e-16 ***

$$\mu = e^{-4.306+0.747\text{Ln}(\text{AADT})+0.762 \text{Ln}(L)+0.042 \Delta V_{85}} \quad (53)$$

Where,

$\mu$  = predicted mean number of crashes

$AADT$  = average daily traffic

$L$  = segment length

$\Delta V_{85}$  = speed differential

It can be noted that there are studies that treated each direction of the road as a separate site to develop the crash prediction model while incorporating the design consistency measures (21; 23; 43; 67). It was possible as they had crash data available for each direction in addition to the speed profiles. This study is limited in terms of predicting crashes by direction since crash dataset in this study was not separated by directions of the road, although speed data were available for each direction. This analysis averaged the 85<sup>th</sup> percentile speed from both directions and determined the Speed Differential towards the increasing mile points. It was then included in the crash prediction model while using number of crashes aggregated from both directions. The analysis also tested Speed Differential by direction in the crash prediction models as the speed dataset allowed to compute the directional 85<sup>th</sup> percentile speed. However, it did not offer better performance than the models already presented in Table 15. In addition, the coefficients were also found to be similar. Therefore, this analysis stick to the models shown in Table 15.

In the crash model based on design consistency (Equation (53)), Speed Differential is found statistically significant at a 5% level along with the  $AADT$  and  $L$ . It is positively related to the number of crashes, which is in line with existing practices (21-23; 43; 65; 67). A one mph difference in the 85<sup>th</sup> percentile speed results in a 4.3% increase in the number of crashes. This finding is similar to some of the existing studies. For example, Anderson et al. and Dhahir and Hassan found a 6.8% and 6.3% increase in crash frequency, respectively for one mph Speed Differential (43; 67).

Furthermore, the relative importance of Speed Differential was found to be 12.74% in Equation (53). From the CURE plot in Figure 16(b), for the locations where the Speed Differential is 5 mph or less, the CURE is outside the preferable range. The model has higher overpredictions of the number of crashes for these locations. These are actually the locations with good design ( $\Delta V_{85} < 6$  mph) according to the design safety

levels proposed by Lamm et al. (137). Moreover, the number of crashes on 84% of these locations ranges between 0 and 2. It appears that the performance of the model in Equation (53) is not good enough for these locations. Further investigation of those locations showed that the segments with higher over-predictions are from high volume roads (AADT  $\geq 5000$ ) and medium to high speeds (average speed  $> 30$  mph). It seems that the performance of the model is questionable for the high volume and medium to high-speed roads. It would have been interesting to see if developing separate models for these segments could provide more accurate predictions. This can be a future scope of this analysis when more data becomes available.

Since a majority of the segments are of good design consistency as shown in Table 17, this analysis performed an ANOVA test by creating balanced datasets to find statistical evidence of crashes varying significantly over the three design safety levels. While keeping all 35 segments for the poor category, the process generated 500 samples by randomly selecting 140 segments for both good and fair categories. Using the balanced samples, the pairwise student's t-test showed that there is a significant difference in crash rates for the poor category compared to the good and fair categories for each of these 500 samples.

Table 17 Mean Crash Rates for Different Design Categories

Design Safety Level	Ranges (source: Lamm et. al (137))	Number of Segments	Mean Crash Rates
Good	$\Delta V_{85} < 6$ mph	5004	1.963
Fair	$6 \text{ mph} < \Delta V_{85} < 12$ mph	143	2.63
Poor	$\Delta V_{85} > 12$ mph	35	146.223

#### 5.2.3.1 Comparison with Models based on Speed Metric

The performance of the selected crash prediction model considering speed differential (Equation (53)) was further compared with speed metric-based models such as Average Speed-based model and the 85<sup>th</sup> Percentile Speed-based model. The goal is to

find whether Equation (53) is better than the crash prediction models incorporating speed metrics. The analysis used the same NB model form and developed two other models where average speed and the 85<sup>th</sup> percentile speed were considered separately in addition to AADT and L of the segments. Table 18 presents these two models as well as the model based on speed differential.

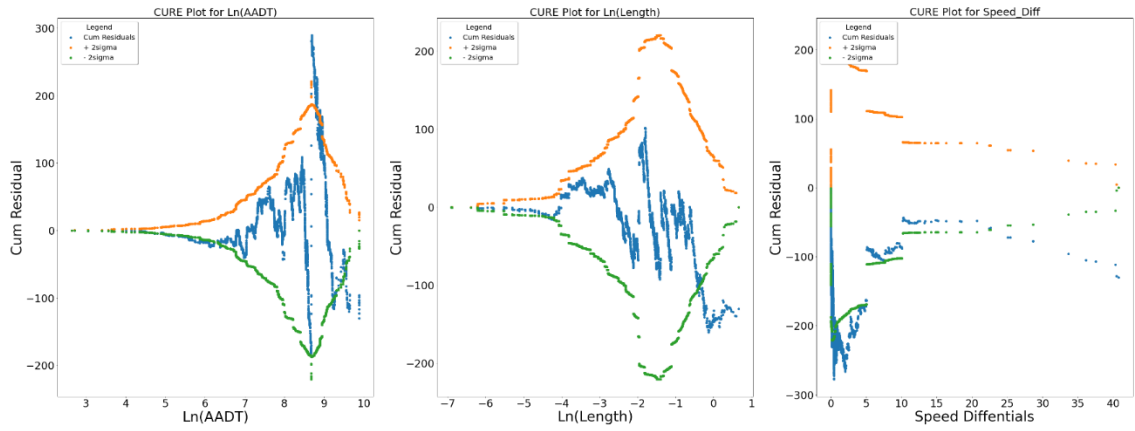
In the models based on speed metrics, the relationship between the speed measures and number of crashes shows that crashes on the higher speed roads tend to be less because of the good geometric conditions. In contrast, the speed differential-based model indicates that a higher inconsistency in the speed from the preceding segment may cause more crashes. In terms of performance, the 85<sup>th</sup> percentile speed-based model seems to be the best model. However, the performance is not substantially better (0.61% decrease in AIC and BIC values) than the model based on speed differential. To choose the best fit model from these three models, further investigation on CURE plots was done. The observations based on CURE plots from Figure 17 are as follows:

- For the model incorporating speed differential, the residuals tend to be outside of the boundaries, especially at around 6000 AADT and 0-5 mph speed differentials (Figure 17 (a)). From Table 19, the residuals remain within the  $\pm 2\sigma$  boundaries for 82% and 65% of the times corresponding to AADT and speed differential.
- The model with average speed shown in Figure 17 (b) tends to highly overestimate and underestimate, especially after an AADT of around 8000 and an average speed of around 50 mph. Table 19 shows that the residuals remain within the  $\pm 2\sigma$  boundaries for 64.5% and 50% of the times corresponding to AADT and average speed.
- The model with the 85<sup>th</sup> percentile speed shown in Figure 17 (c) significantly underestimates and overestimates after an AADT of around 8000 and an 85<sup>th</sup> percentile speed of around 55 mph. In addition, Table 19 shows that the residuals remain within the  $\pm 2\sigma$  boundaries for 64.5% and 41% of the times corresponding to AADT and the 85<sup>th</sup> percentile speed.

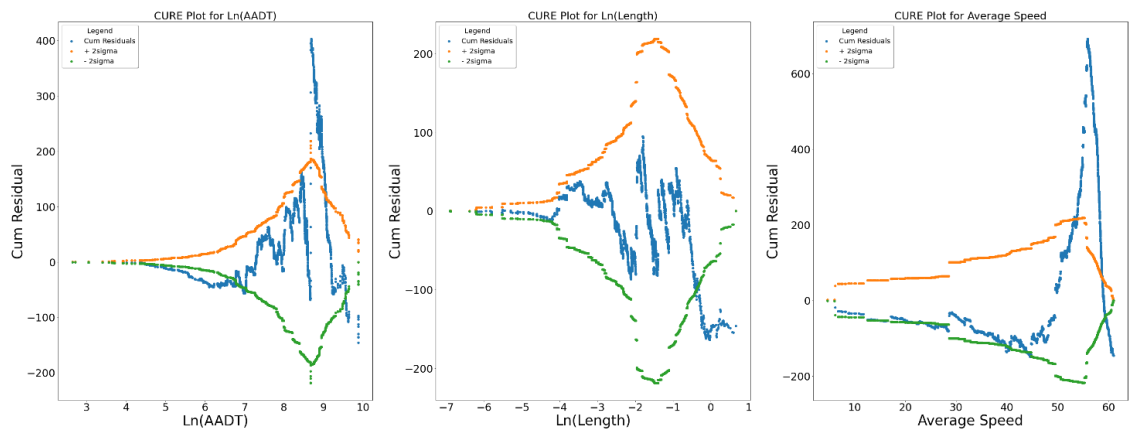
The above observations from CURE plots and the percentage of residual within the  $\pm 2\sigma$  boundaries for each model reveal that the model incorporating Speed Differential provides a better fit compared to the speed metric-based models.

Table 18 Comparison between Speed Differential and Speed Metric-based Models

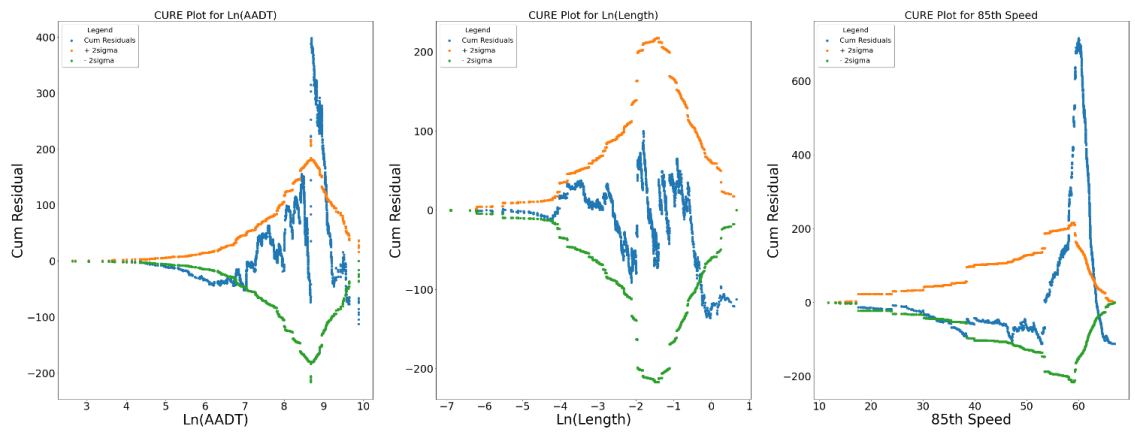
Variables	Model based on Speed Differential		Models based on Speed Metric			
	Estimate	Std. Error	With Average Speed		With The 85 <sup>th</sup> Percentile Speed	
			Estimate	Std. Error	Estimate	Std. Error
<b>Intercept, <math>\epsilon</math></b>	-4.306	0.199	-3.424	0.208	-2.565	0.251
<b>Ln (AADT)</b>	0.747	0.024	0.824	0.026	0.780	0.024
<b>Ln (L)</b>	0.762	0.022	0.826	0.023	0.830	0.023
<b><math>\Delta V_{85}</math></b>	0.042	0.009	-	-	-	-
<b><math>V_a</math></b>	-	-	-0.027	0.003	-	-
<b><math>V_{85}</math></b>	-	-	-	-	-0.032	0.003
<b>AIC</b>	11916.75		11849.11		11843.46	
<b>BIC</b>	11948.39		11880.75		11875.10	
<b>RMSE</b>	2.67		2.65		2.64	
<b>MAPE (%)</b>	64.83		63.70		63.88	
<b>MAD</b>	1.21		1.21		1.21	



(a) Speed Differential Based Model



(b) Average Speed Based Model



(c) The 85<sup>th</sup> Percentile Speed Based Model

Figure 17 CURE plots for Speed Differential and Speed Metric-based Models



Table 19 Percentage of CURE within  $\pm 2\sigma$  Boundaries

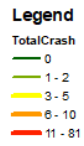
Models	% CURE within $\pm 2\sigma$		
	AADT	Length	Speed Measure
Speed Differential based Model	82	95	65
Average Speed based Model	64.5	95	50
The 85 <sup>th</sup> Percentile Speed-based Model	64.5	96	41

#### 5.2.4 Application and Limitations

The crash prediction model developed based on Speed Differential in this analysis was applied to find out whether it can identify the hot spots with inconsistent speed. In other words, the aim was to investigate if the inconsistency in speed can actually be used in crash prediction. For example, Figure 18(a) and Figure 18(b) present some of the locations with speed inconsistency indicated by speed differentials and high crash locations, respectively. For segments with higher crashes, the speed may not be always inconsistent. Further, segments with the lowest or no speed differential may have high crashes from Figure 18(b). Crashes predicted by the model (Figure 18(c)) may not always capture those high observed crashes. Overall, it looks like high crashes may not necessarily be involved with high-speed differentials based on the study data. Instead of using the speed differential measure for identifying crashes, it can be rather used for design improvements when deemed necessary.



(a) Locations with Speed Inconsistency



(b) Distribution of Observed Total Crashes



(c) Distribution of Predicted Total Crashes

Figure 18 Application of the Analysis

### 5.2.5 Major Findings and Significance of the Analysis

Past research incorporated the Speed Differential as a design consistency measure in the crash prediction model to relate design consistency with road safety. Most of this research mainly used speed prediction models to estimate the 85<sup>th</sup> percentile speed before calculating speed differential. The speed prediction models were mainly developed using spot speed data, which may fail to capture the speed variation over a segment. This may lead to an inaccurate estimation of the Speed Differential and may further affect the accuracy of the design consistency analysis for crashes. This analysis tried to address this issue by utilizing measured speed data in determining Speed Differential and developed crash prediction model based on that. Key observations of the analysis can be listed below:

- Speed Differential was found as a significant predictor of rural two highways crashes. It is positively related to the number of crashes. It implies that crashes are higher when the design inconsistency is higher as indicated by the speed differential.

- It was also observed that a one mph difference in the 85<sup>th</sup> percentile speed results in a 4.3% increase in the number of crashes. This finding is similar to some of the existing studies, such as those by Anderson et al. and Dhahir and Hassan, where the authors found a 6.8% and 6.3% increase in crash frequency, respectively for a one mph Speed Differential (43; 67).
- Crash prediction model incorporating Speed Differential as a consistency measure outperformed the model with speed metric (average speed, the 85<sup>th</sup> percentile speed) as shown in Figure 17 and Table 19.

The above finding implies that crash occurrence on rural two-lane highways is not only dependent on local attributes of that segment but also on the global geometric behavior, i.e., effect of adjacent elements on that segment. Incorporation of that behavior into the model provided further accuracy in crash predictions. These findings can be supported by a recent study by Llopis-Castelló et al. (53). However, the application of the speed differential-based model in identifying hot spots revealed that the higher crash location in this study may not be always involved with speed inconsistency. Therefore, speed differential may not be a suitable factor for predicting crashes in this study, rather it can be useful to take measures for further design improvement of the roads.

The analysis has multiple limitations in terms of the dataset. It could not explore the effect of Speed Differential for each direction of the road since the crash dataset came into an aggregated format regardless of the directions. If directional crash data can be collected, the analysis can be revisited further. Moreover, 92% of the data was from curve Class A and majority of the segments had a good design. This requires further looking into the analysis if more data for other curve classes are available.

### 5.3 Summary

This chapter explored the effect of speed from both an operational perspective and a design perspective. In both cases, speed measure was found significant for the crashes of the rural two-lane highways. For the individual segment-based analysis, Average Speed was the better representation of the operating condition of these roads. This analysis showed a varying effect of speed on crashes from low-speed to high-speed roads.

It implies that speed has a subgroup effect on the crashes of rural two-lane highways. Therefore, it is recommended to consider developing separate models based on the speed of these roads. Although including the speed variable in the model may not always add a dramatic change in the prediction performance, considering the speed during splitting the data for developing separate models can improve the overall performance. This analysis can be applied during planning level safety improvement of these roads. For the speed consistency-based analysis, speed differential from the prior segments showed significant influence on the crashes of a segment. However, further investigation of the dataset and model predictions showed that speed differential can be a good indicator for design rather than potential crash locations.

Until now, this study explored the effect of speed without considering the spatial heterogeneity in the dataset. The next chapter tries to incorporate spatial heterogeneity while investigating the effect of speed in addition to other factors on crashes of rural two-lane segments in this study.

## CHAPTER 6. SPATIAL VARYING EFFECT OF THE FACTORS ON CRASHES

Traditional count models (such as ZINB), used in the previous chapter to investigate the effect of speed on crashes, assume a stationary pattern of the crash data as well as the constant effect of the variables over the spatial domain. These models estimate a single coefficient value as the average effect of a variable on crashes. However, crash data and road attributes can show a similar pattern with the neighboring segments. The pattern can vary within the same jurisdiction based on the geographical locations. Considering this spatial dependency, the relationship between crashes and road attributes may show spatial heterogeneity (76-80). To incorporate such spatial dependency, this chapter adopts spatial modeling techniques. These are GWP and GWZIP models (Section 4.4.2.2 ). These models account for the spatial location by developing local models utilizing the nearest segments when establishing the relationship between crashes and the explanatory variables. The results from the local models are used to diagnose the spatially varying effects of different factors on the crashes in this study.

### 6.1 Objectives

This chapter analyses the spatial pattern of the effect of traffic, geometric and speed conditions on crashes of rural two-lane highways. Here are the objectives:

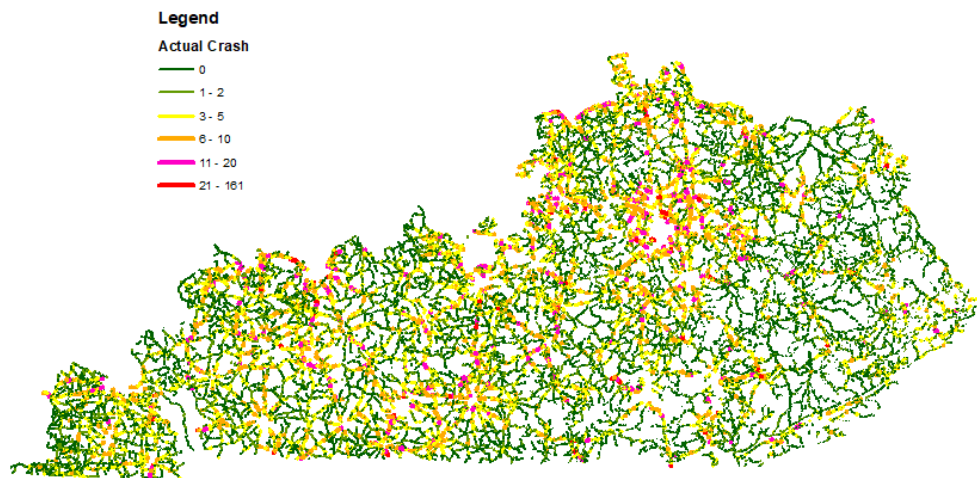
- Investigate whether there exists spatial heterogeneity in the effect of the geometric attributes, traffic volume, and speed on the total number of crashes for rural two-lane highways.
- Compare the performance of GWPR and GWZIP models with the traditional count models.

### 6.2 Dataset and Variable Selection

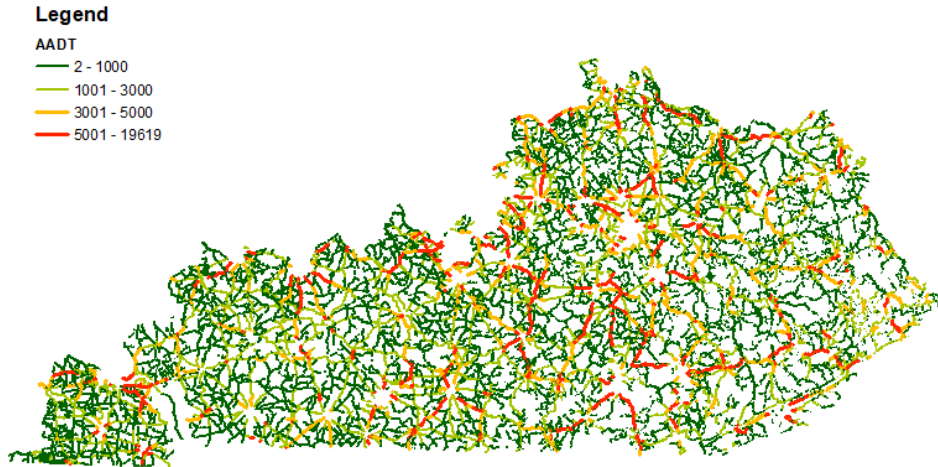
For this analysis, the author wanted to incorporate additional geometric variables including Degree of Curvature. The dataset described in Section 3.2 had issues considering Degree of Curvature. During the aggregation process, the information related to Degree of Curvature got diluted by changing the curve class (see [Appendix](#)). The analysis further processed the dataset following the same aggregation approach in Section

3.2 while considering an additional condition for curve class. It means the aggregation of the segments was done up to half a mile if there is no intersection and the curve class of the segments is the same. In this way, a balanced dataset in terms of Degree of Curvature can be obtained. After the processing, the final dataset contains 53,208 segments with a total of 65,091 crashes aggregated from both directions of the road.

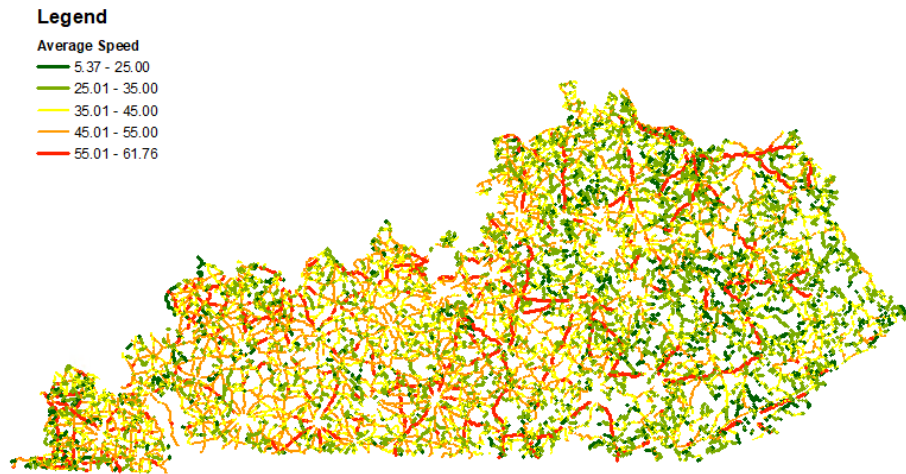
Figure 19(a) presents the distribution of observed number of crashes on these segments over the state. Table 13 presents the statistics of geometric, traffic, and speed attributes on these segments, and Figure 19(b)-(f) shows the spatial distributions of these variables. To select the explanatory variables for developing the spatial models, this analysis initially tested the correlations between each pair of the variables. Figure 20 shows the Pearson correlation coefficients for each pair. Based on the coefficients, lane width and shoulder width showed a higher correlation with AADT. Therefore, they were excluded from the model. The final list of explanatory variables for the model development included AADT, L, Average Speed, and Degree of Curvature. As the response variable, the total number of crashes in 5 years was used.



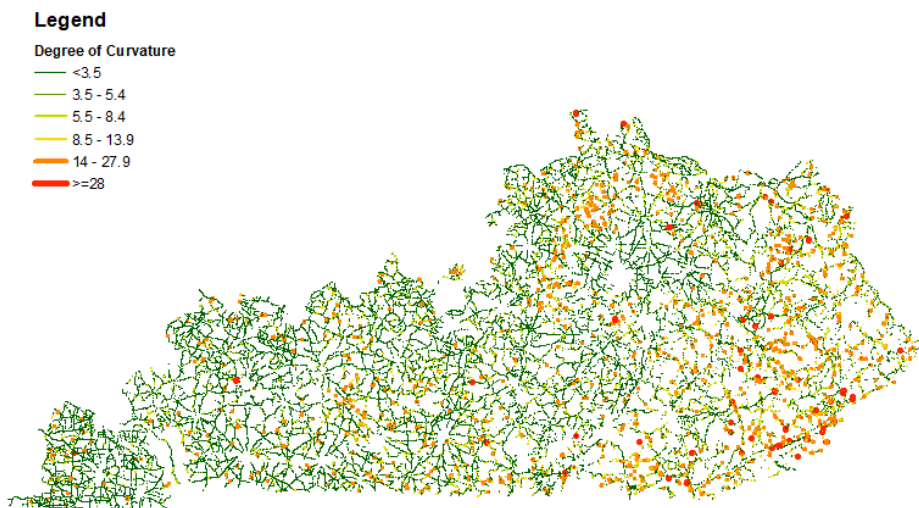
(a) Spatial Distribution of Observed Number of Crashes



(b) Spatial Distribution of AADT

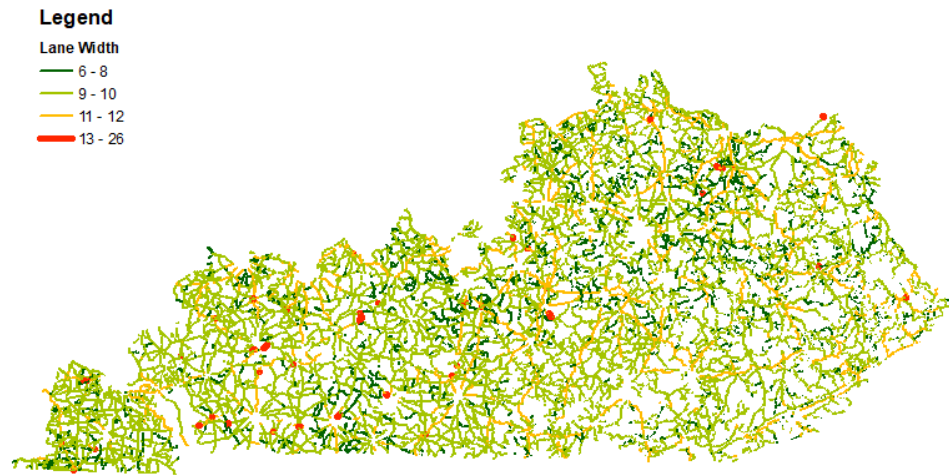


(c) Spatial Distribution of Average Speed

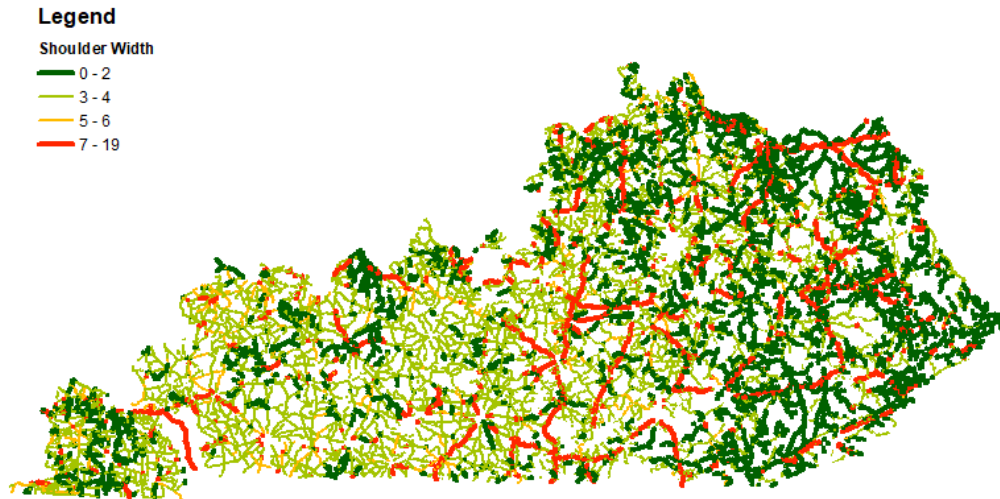




(d) Spatial Distribution of Degree of Curvature



(e) Spatial Distribution of Lane Width



(f) Spatial Distribution of Shoulder Width

Figure 19 Spatial Distribution of Variables

Table 20 Summary Statistics of the Variables

Variables	Unit	Statistics			
		Min.	Max.	Mean	Standard Deviation
AADT	vehicle	2	19619	1355	1772
Segment Length (L)	mile	0.10	2.97	0.26	0.21
Average Speed ( $V_a$ )	mph	5.37	61.76	39.92	9.87
Degree of Curvature ( $C_u$ )	degrees	0	63.81	2.42	3.80
Lane Width (LW)	ft	6	18	9.42	1.14
Shoulder Width (SW)	ft	0	14	3.51	2.06
Number of Crashes in 5 years		0	161	1.22	2.85

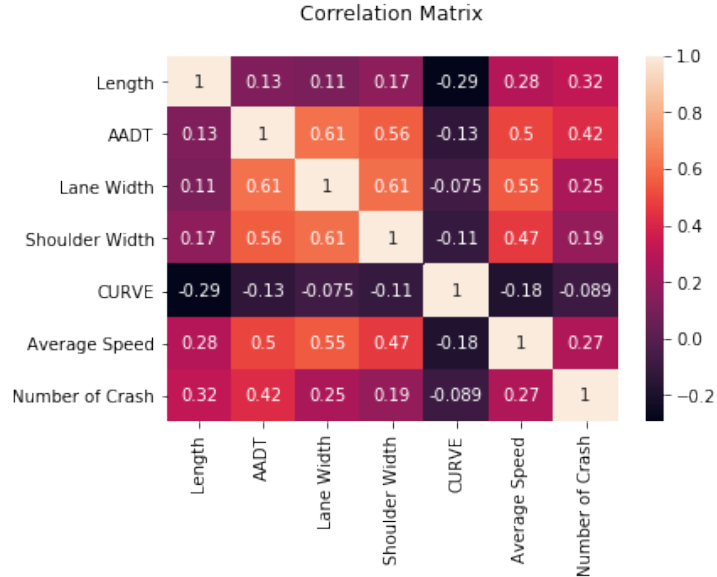


Figure 20 Correlation Analysis

### 6.3 Spatial Autocorrelation Check

To check the spatial dependency of the explanatory and response variables before fitting the spatial models, a spatial autocorrelation test was performed using Moran's I. Table 21 shows the Moran's I value for the variables. For all the explanatory variables in addition to the response variable, the values are positive and significant at a 5% confidence level. It indicates the variables are spatially autocorrelated significantly. The proof of spatial autocorrelation supports the idea of testing the spatial models for the analysis.

Table 21 Spatial Dependency of the Variables

Variables	Moran's I	P-value	Clustered/Spatial Autocorrelation
AADT	0.4778	0	Yes
L	0.1213	0	Yes
$V_a$	0.3984	0	Yes
$Cu$	0.0988	0	Yes
Number of Crash	0.0135	0	Yes

### 6.4 Analysis and Results

Four models, including both global and local models, were evaluated for this analysis. As the global models, Poisson model and ZIP model were developed with AADT, L, Average Speed, and Degree of Curvature utilizing all the segments. For the local models, GWP and GWZIP models were fitted using the same set of variables. The optimum bandwidths were estimated as the farthest neighbor distance associated with 1,360 and 1,050 nearest neighbors, respectively for GWP and GWZIP models. The number of neighbors related to the optimum bandwidth can vary based on the model type (91). Furthermore, the number of neighbors used to estimate the optimum bandwidths for both models meets the sample size requirement by HSM and *Safety Performance Function Decision Guide*. Performance of the models was evaluated using  $R^2$  and RMSE.

Table 22 presents the coefficients of the variables estimated from each model. The global models in Table 22 (i) provide the coefficient values for each variable, assuming their influence on the number of crashes remains constant over the spatial domains. The effect of all variables was found to be statistically significant at a 5% level. In both Poisson and ZIP models, the estimated coefficients are reasonable, for example, number of crashes increases with AADT, L, and Degree of Curvature, which makes sense and is in line with existing literature (27; 31). In addition, Average Speed is negatively related to number of crashes based on the dataset used and it is also consistent with existing research findings (27).

Table 22 (ii) also provides the descriptive statistics of coefficient values for each variable from the local models (i.e., GWP and GWZIP). Like the global models, both GWP and GWZIP models show a positive influence of AADT and L on crashes. The minimum coefficient value for Degree of Curvature suggests that there are locations where the local models determined a negative relationship between number of crashes and Degree of Curvature. This relationship seems to be counterintuitive. After investigating the negative coefficients, this analysis observed that all these negative coefficients depicted no statistical significance in both GWP and GWZIP models. Other existing research observed similar cases of negative relationships for Degree of Curvature from geographically weighted regression models (93), and one of the reasons for estimating such relationships by these models can be that some variables may not be significant in certain road segments (83; 138). In case of Average Speed, both positive and negative influences on crashes can be observed. Existing literature supports both types of findings for this variable in the crash prediction model (18; 27). Further investigation results related to the effect of Average Speed based on the local crash prediction model are discussed in the later subsection (Section 6.4.1.3) of this chapter.

The performance measures in Table 22 show better fits for the local models compared to their corresponding global models. Between the local models, GWZIP seems to perform slightly better. This analysis chose GWZIP model to proceed with the further discussion on the spatial variation of the coefficients in the subsections below.

Table 22 Variable Coefficients and Model Performance

(i) Global Models

Model	Poisson Model			ZIP		
	Coefficient	Std. Error	z-value	Coefficient	Std. Error	z-value
Intercept	-3.9580	0.0320	-123.69	-2.7658	0.0430	-64.27
Ln(AADT)	0.8334	0.0045	186.56	0.7141	0.0056	127.94
Ln(L)	0.8950	0.0065	137.75	0.7728	0.0078	99.24
$V_a$	-0.0149	0.0005	-27.44	-0.0196	0.0006	-31.36
$Cu$	0.0361	0.0013	27.34	0.0403	0.0015	26.2
$R^2$	0.2998			0.3113		
RMSE	2.38			2.38		

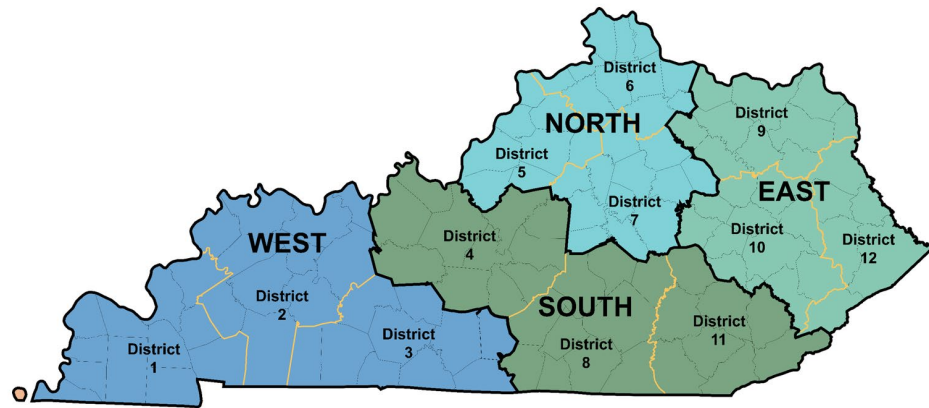
\*Note: all variables showed p-value < 2e-16

(ii) Local Models

Model	GWP				GWZIP			
	Min	Max	Mean	SD	Min	Max	Mean	SD
Intercept	-7.1040	-1.6279	-4.4005	0.9854	-6.8246	-0.3467	-3.4056	1.2214
Ln(AADT)	0.2625	1.2709	0.8615	0.1472	0.0117	1.2094	0.7586	0.1711
Ln(L)	0.5027	1.3820	0.8626	0.1064	0.1275	1.4742	0.7746	0.1565
$V_a$	-0.0661	0.0946	-0.0125	0.0167	-0.0880	0.1135	-0.0160	0.0201
$Cu$	-0.0699	0.1572	0.0524	0.0328	-0.1528	0.1957	0.0580	0.0448
$R^2$	0.4074				0.4109			
RMSE	2.19				2.17			

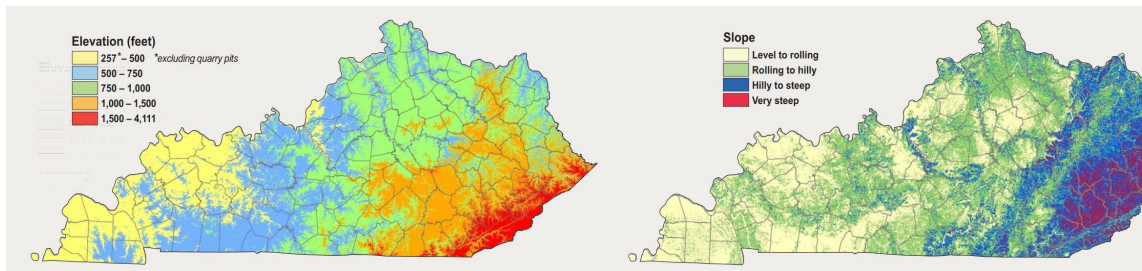
### 6.4.1 Spatial Variation Analysis

Kentucky is divided into four geographical regions with varying terrain and area, as shown in Figure 21(a) -Figure 21(c). In terms of terrain, from East to West, it changes from very steep and hilly to rolling and level. In terms of area type, Eastern Kentucky seems to be more rural whereas Northern Kentucky seems to be largely urbanized. In this study, this knowledge of different regions, terrain types, area types, etc. will be utilized to explain the spatial pattern of coefficients.



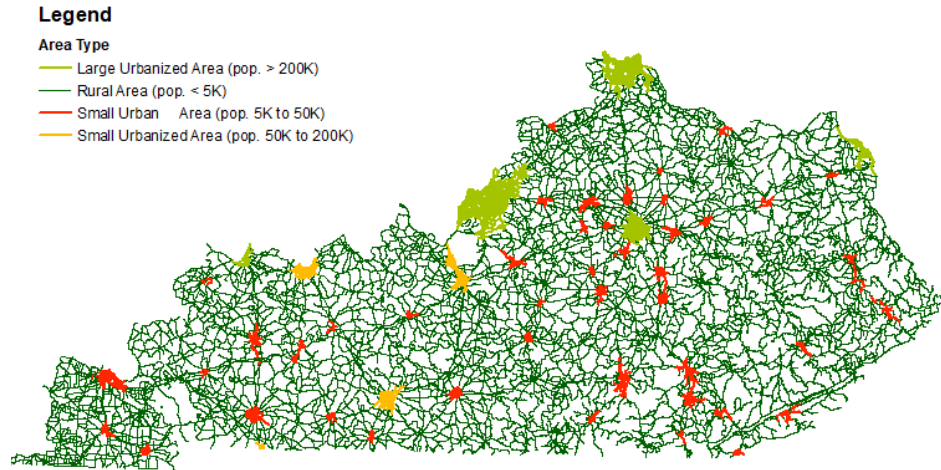
*Source: Kentucky Transportation Cabinet (139)*

(a) Kentucky Regions



*Source: Kentucky Geological Survey (140)*

(b) Terrain



(c) Area Type

Figure 21 Kentucky (a) Regions, (b) Terrain, and (c) Area Type

Spatial variation of the coefficients from the local models is discussed below:

#### 6.4.1.1 AADT

From GWZIP model, AADT was found to be significant for 99.76% of the rural two-lane segments. Such a high percentage is expected due to the predominant influence of exposure variables in crash prediction. Figure 22 shows a distinct spatial pattern of the effect of AADT. For example, Eastern Kentucky and Western Kentucky show a comparatively higher effect of AADT on crashes from Figure 22. These regions are mostly rural and less urbanized with less population density (Figure 21(c)), therefore, low traffic volume is usually observed. An unexpected increase in traffic in these regions may cause random fluctuation in the traffic pattern, which may affect the crashes in these regions. In contrast, the impact of traffic volume transitions from average (coefficients 0.6 -0.8) to lower in most of the Southern and Northern regions. These regions are more urbanized with higher population density. In other words, the usual traffic can be heavy with obvious patterns.

Some road segments in Western Kentucky (close to the Indiana Border) and Southern Kentucky region show the lowest or insignificant influence of AADT (in dark green or cyan color). These segments are mainly close to the urbanized areas. Other

factors such as Average Speed is more significant for these roads, as depicted by the ranking of the variables in Figure 25.

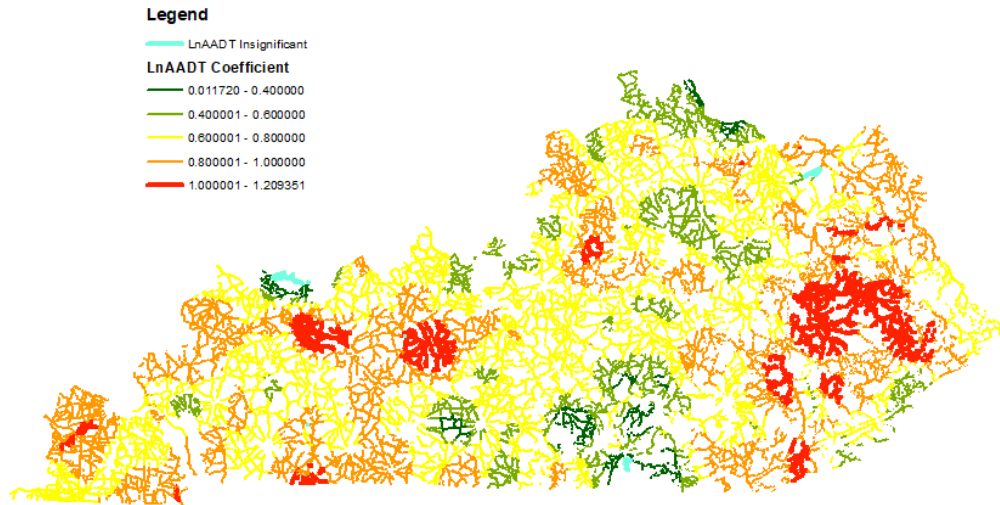


Figure 22 Spatial Distribution of the Coefficients for AADT

#### 6.4.1.2 Segment Length

Length was found significant for 99.1% of the segments from GWZIP model, which is not surprising for an exposure variable. From Figure 23, the spatial distribution of the coefficients seems random, and it is hard to find any distinct pattern for the effect of segment length over different regions. From the segmentation process, around 87.5% of the segments have a length of 0.5 miles or less. Less variation in the lengths for most of the segments may result in such random pattern of the coefficients.



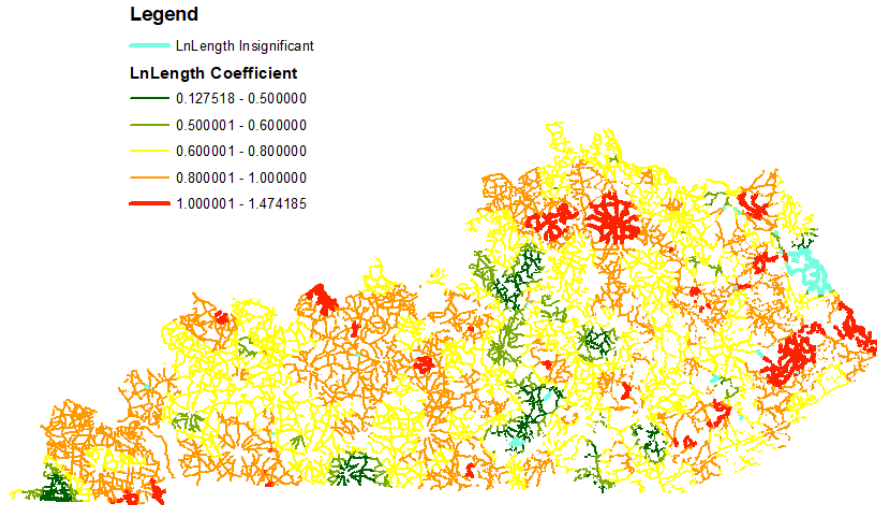


Figure 23 Spatial Distribution of the Coefficients for Length

#### 6.4.1.3 Average Speed

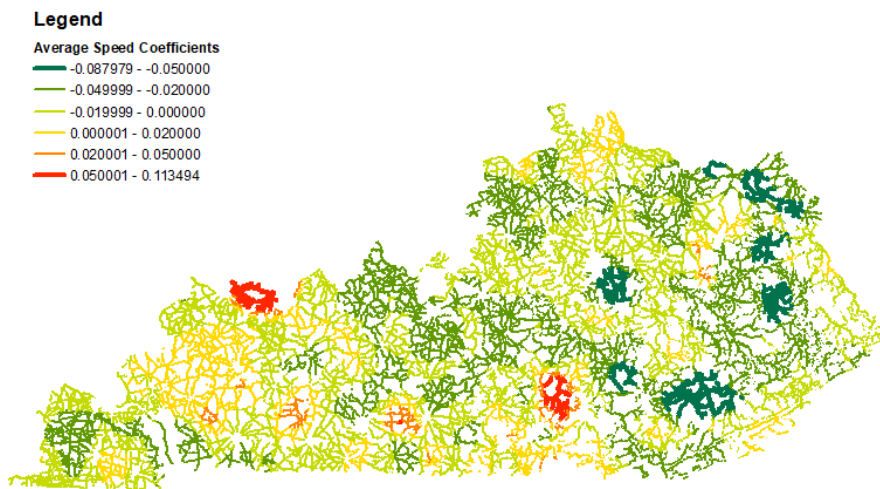
Effect of Average Speed was found significant for 50.1% of the total rural two-lane segments from GWZIP model, shown in Figure 24(b). Around half of those segments with significance have at least one crash record. Among the 50.1% segments, 5.8% showed a positive influence of speed on crashes (global model did not capture this positive effect), and rest of the segments showed a negative influence of speed on crashes. The coefficients of speed for these segments can be shown in Figure 24(a).

After looking closely at the spatial pattern of the significance of speed in Figure 24(b), it seems that speed is a significant factor for most of Eastern and Northern Kentucky. These regions mainly show the negative influence of speed on crashes (Figure 24(a)). Further investigating the features of the roads in these regions, poor geometric conditions were observed. The roads had narrow shoulders (Figure 19(f)) and sharp curvatures (Figure 19(d)). A low to medium average speed was also observed on these roads (see Figure 19(c)). Such results from the local models indicate giving further attention to the poor geometric standards for improving safety on those roads.

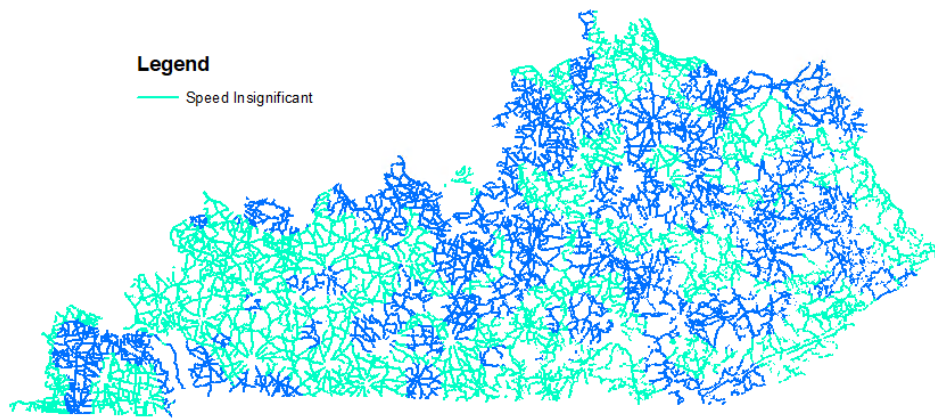
In Western and Southern Kentucky, speed is mainly insignificant except for some places in red color shown in Figure 24(a). These places show a higher positive effect of speed on crashes. There are 488 such segments. These segments are mostly the ones

where a lower or insignificant effect of AADT on crashes was observed (Figure 22). Further looking into the importance of speed in these segments, it turned out that speed is the top or second important variable for most of these segments (Figure 25). This analysis investigated these segments and found that these segments are having better geometric conditions over the flat terrain (Figure 21(b)) with wider shoulders (Figure 19(f)) and straight sections (Figure 19(d)). Moreover, the volume (<1000) is lower on these roads. Therefore, if a crash occurs, speed is clearly the reason and the main factor. This finding is different from the global models. When proposing safety improvement plans for these locations, speed should be given priority.

In addition to the significant positive effect of speed, Western Kentucky showed locations with significant negative effects of speed close to Missouri and Tennessee borders. These locations seem to have mostly narrow shoulders (Figure 19(f)) while operating at medium to high speeds (Figure 19(c)).

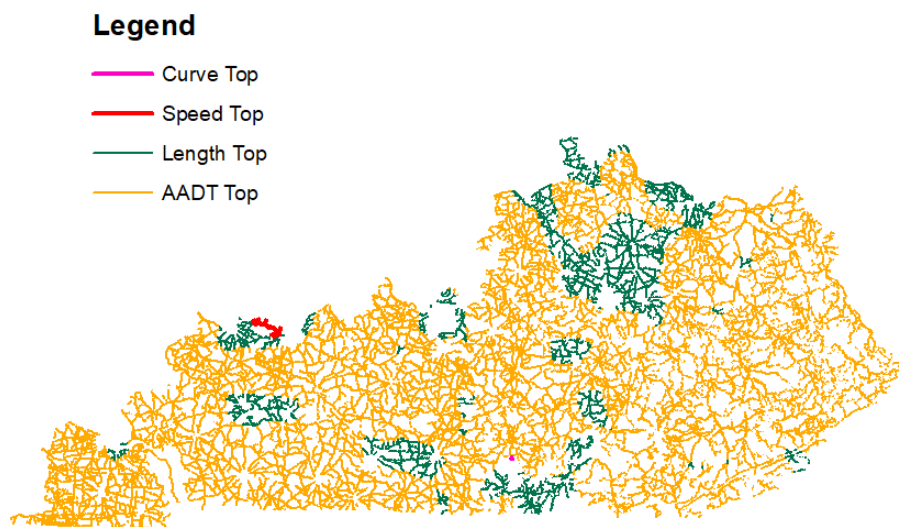


(a) Speed Coefficients from GWZIP Model

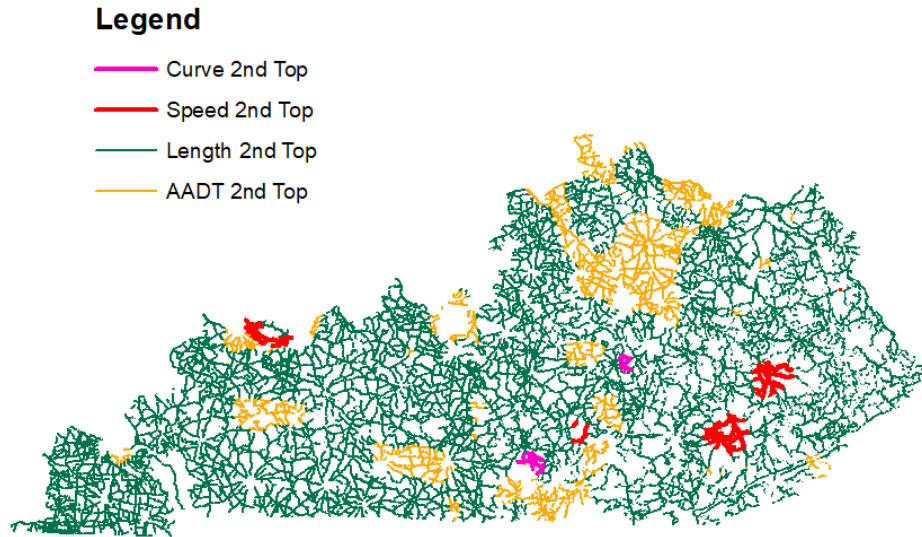


(b) Segments with Speed as an Insignificant Factor from GWZIP Model

Figure 24 Spatial Distribution of the Coefficients and Significance for Speed



(a) Top Ranked Variables from GWZIP Model



(b) Second Ranked Variables from GWZIP Model

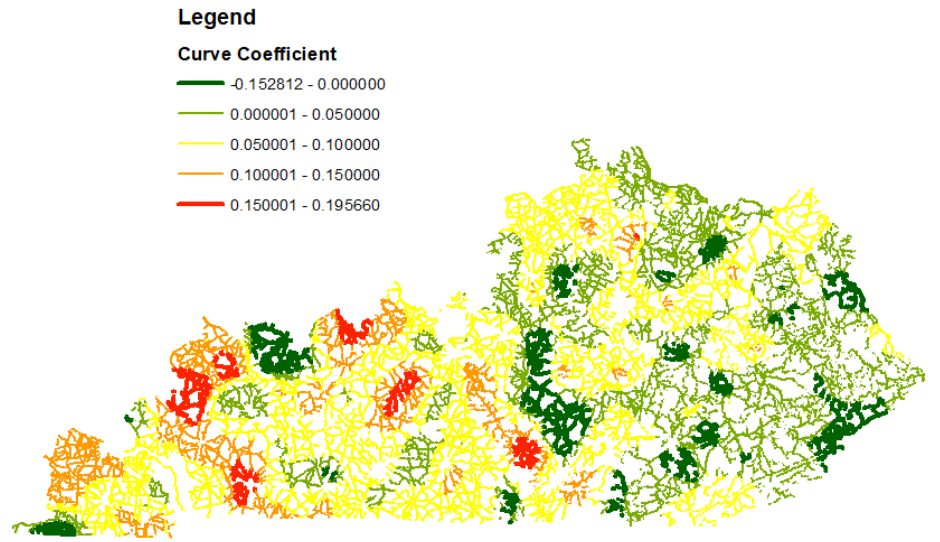
Figure 25 Variable Ranking from GWZIP Model

#### 6.4.1.4 Degree of Curvature

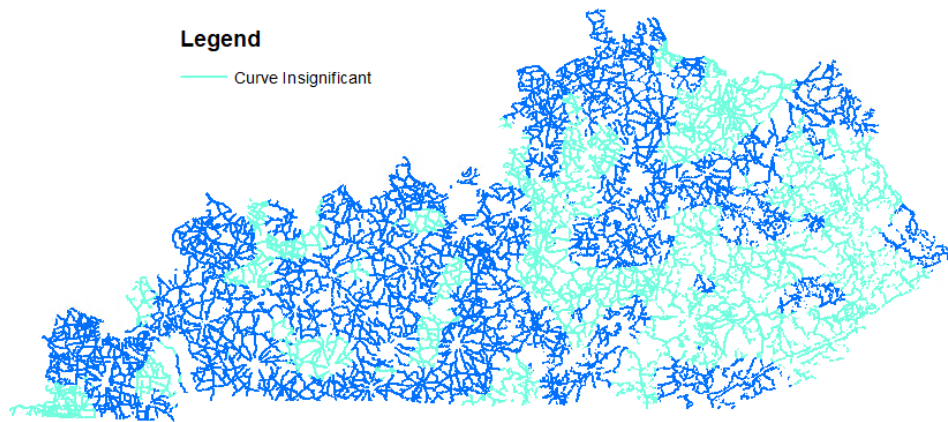
From the GWZIP model, Degree of Curvature was found to be the significant factor on 61.4% of the rural two-lane segments, respectively. As shown in Figure 26(b), most of these significant results are seen in Western and Southern Kentucky. Figure 26(a) shows a higher influence of the Degree of Curvature on Western Kentucky. Those areas are mainly the higher speed roads (Figure 19(c)) with more standard geometric conditions (Figure 19(d) and Figure 19(f)) and flat terrain (Figure 21(b)). An increase in Degree of Curvature can be more critical for the safety of these roads with higher speed conditions compared to low-speed roads.

It is further noticeable from Figure 26(b) that Degree of Curvature is not a significant variable for a large number of segments in Eastern Kentucky. This appears to contradict the assumption that the Degree of Curvature should be a significant factor for crashes in this area of Kentucky due to the presence of sharp curvature (Figure 19(d)). To evaluate the assumption, this analysis further investigated how the Degree of Curvature is being affected by the segmentation process and whether there is strong evidence of the influence of Degree of Curvature on the number of crashes based on the dataset. The

evaluation can help decide whether Degree of Curvature should be considered in the global models, therefore, in the local models.



(a) Degree of Curvature Coefficients from GWZIP Model



(b) Segments with Degree of Curvature as an Insignificant Factor from GWZIP Model

Figure 26 Spatial Distribution of the Coefficients and Significance for Degree of Curvature

The analysis below assesses the influence of curvature globally and decides on whether Degree of Curvature should be included in the crash prediction model.

6.4.1.4.1 DATA ANALYSIS FOR DEGREE OF CURVATURE:

At first, this analysis looked at the distribution and descriptive statistics of Degree of Curvature for the 53,208 segments shown in Figure 26, Table 23, and Table 24. The distribution shows that around 74% of the segments are Class A curves. In addition, the standard deviation of the Degree of Curvature data shows 3.6 degrees of variation from the mean Degree of Curvature. Furthermore, the analysis looked at the scatterplot based on the Degree of Curvature and crash rate per VMT (Figure 28). The scatterplot shows that higher crash rates are mainly in the range of the Class A curve.

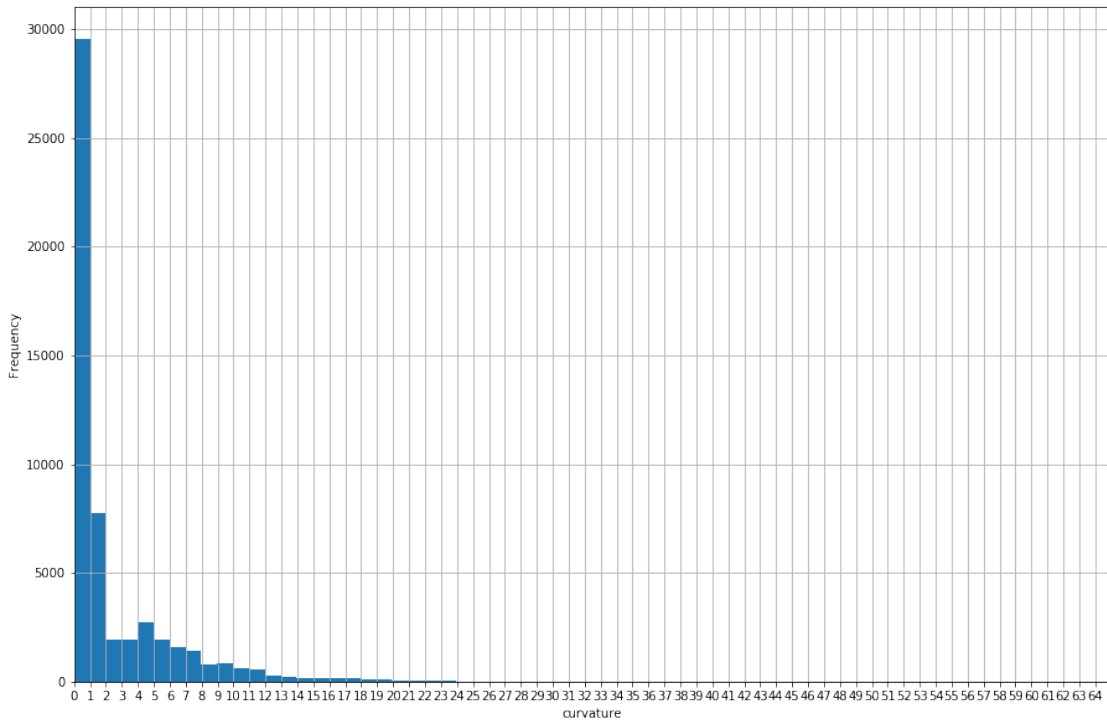


Figure 27 Distribution of Degree of Curvature

Table 23 Class-wise Distribution of Degree of Curvature

Curve Class	Degree of Curvature Range	No of Segments
A	<3.5	39456
B	3.5 – 5.4	5362
C	5.4 – 8.4	4426
D	8.5 – 13.9	2888
E	14 – 27.9	1020
F	>=28	56

Table 24 Summary Statistics for Degree of Curvature

Total Segments	53,208
mean	2.424602
std	3.802812
min	0
25%	0.183304
50%	0.183304
75%	3.600000
max	63.81

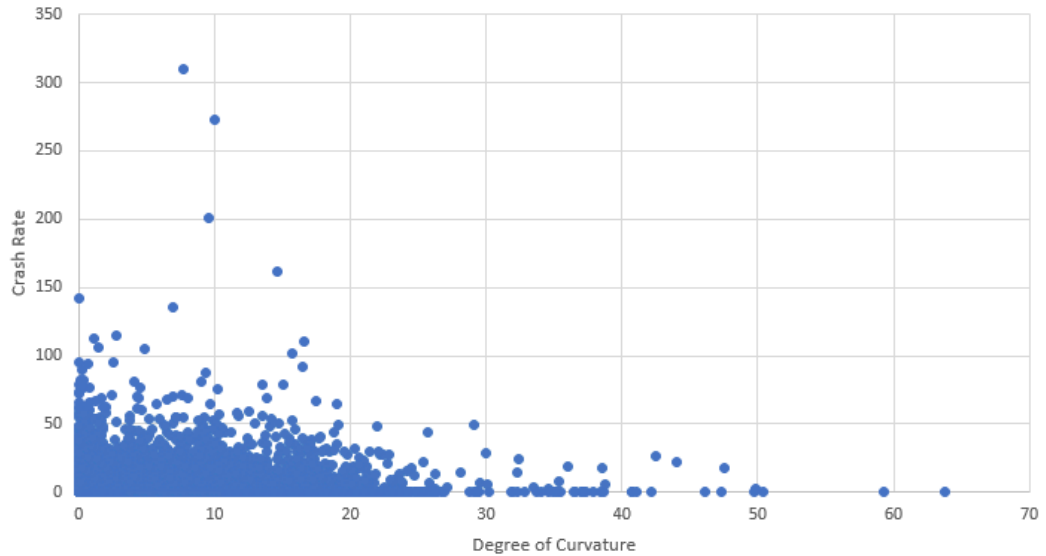


Figure 28 Crash Rate vs Degree of Curvature

Later, the analysis checked the spatial distribution of Degree of Curvature over Kentucky. A comparison was made between the segments before they were aggregated for up to 0.5-mile segments (Figure 29) and the segments after the aggregation process (Figure 30). In Figure 29, the higher classes of curvature (Class D to Class F) seem to be mostly in Eastern Kentucky and Northern Kentucky regions. After the aggregation process of making at least 0.5-mile segments, Figure 30 seems to show a similar pattern, especially for Eastern Kentucky, but some areas turned into lower curvature classes after calculating the length weighted average of curvature.



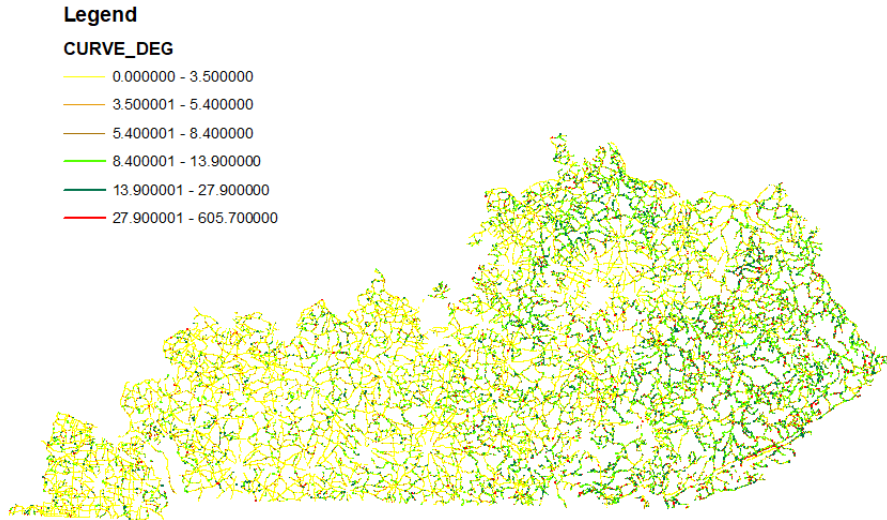


Figure 29 Before Aggregation

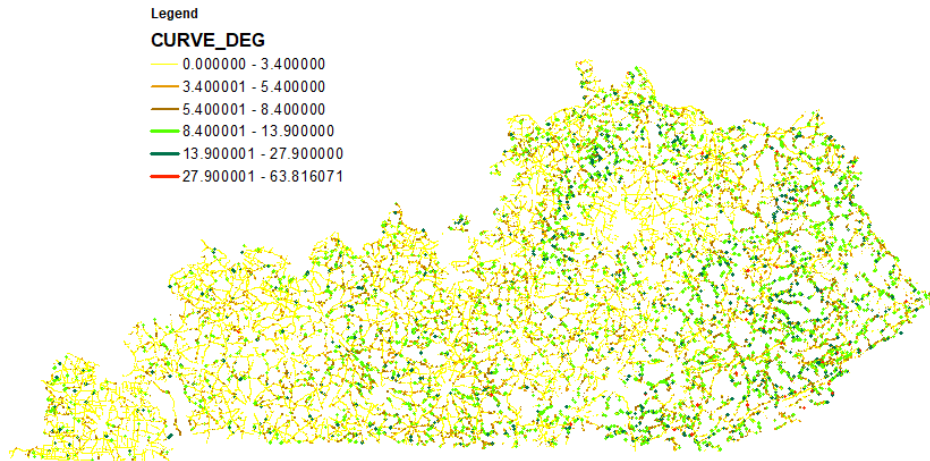


Figure 30 After Aggregation

It is apparent from the comparison of Figure 29 and Figure 30 that there are areas where the data related to Degree of Curvature is being diluted due to the aggregation process. It raises the question of how number of crashes will be affected if the analysis could use an actual Degree of Curvature or a more homogenous Degree of Curvature (where the curvature information wasn't substantially affected by the aggregation process) instead of the diluted Degree of Curvature. Is there a significant influence of

actual Degree of Curvature on the crashes based on the dataset used for the analysis? To answer this question, additional experiments were performed as below:

- Another aggregation process was done to obtain at least 0.1-mile segments. In this way, there will be more homogenous segments in terms of Degree of Curvature. The 0.1-mile aggregation process resulted in 80,221 segments after filtering out the segments shorter than 0.1 miles.
- For the 80,211 segments, several statistics were determined. For example, max and min Degree of Curvature, classes for max and min Degree of Curvature, difference between the curve classes of max and min curve, and percentage of maximum curvature class while aggregating the segments to at least 0.1 miles, etc.
- From 80,221 segments, the analysis filtered out the more homogenous segments using difference between the maximum and minimum curve class and percentage of maximum curvature class. It only chose the segments where difference between the maximum and minimum curve classes (during aggregating the segments) was maximum of one class or the percentage of maximum curvature class in the aggregated segment is at least 70% (this is a subjective value considering it is enough to capture the sharp curves). All these screenings resulted in 39,215 segments in total. Table 25 shows the distribution of Curve classes for these 39,215 segments. While Class A consists of around 88% of the dataset, Class F only has 37 observations.

Table 25 Distribution of Curve Class in More Homogenous Dataset

Curve Class	No of segments
A	34708
B	2024
C	1363
D	827
E	255

F	37
---	----

The 39,215 segments were utilized in developing one of the global models, for example, ZIP model in this case. Two models were tested: one without Degree of Curvature and the other with Degree of Curvature as shown in Table 26. Including Degree of Curvature in the model shows a similar performance from the models as indicated by the  $R^2$  and AIC values. Furthermore, the coefficient for the Degree of Curvature in the model with curve indicates a 2.8% increase in the number of crashes with a unit increase in Degree of Curvature.

Table 26 Comparison of ZIP Models based on Degree of Curvature

	Without Curve			With Curve		
	Estimate	Std. Error	z value	Estimate	Std. Error	z value
(Intercept)	-2.4838567	0.0500012	-49.68	-2.56632	0.050624	-50.69
Ln(AADT)	0.7014621	0.006192	113.29	0.708338	0.006237	113.58
Ln(L)	0.7595176	0.0095035	79.92	0.783896	0.009785	80.11
$V_a$	-0.0236654	0.0006601	-35.85	-0.02317	0.000663	-34.97
$C_u$				0.02801	0.002454	11.41
$R^2$	0.2641			0.2640		
AIC	113888.6			113756.2		

\*  $P$ -value  $< 2e-16$  for all variables in the models

Even though Degree of Curvature does not seem to contribute significantly to the model performance, the analysis further checked how crashes vary over different curve classes based on 39,215 segments. For this, the analysis considered crash rate per VMT. Table 27 presents the mean, minimum and maximum crash rates under each curve class. Except for Class F, it shows an increasing mean crash rate from Class A to Class E. For Class F, it shows the lowest crash rate. This may not necessarily be the case as the sample size is substantially small for this class to provide a consistent result.

Table 27 Statistics of Crash Rates for different curve classes

<b>class</b>	<b>mean</b>	<b>max</b>	<b>min</b>
A	1.941332	94.731389	0.0
B	2.154299	63.818449	0.0
C	2.428271	58.855554	0.0
D	2.983842	90.734427	0.0
E	3.162197	50.714604	0.0
F	1.931905	26.107547	0.0

To see how the mean crash rates significantly vary over the curve classes, an ANOVA test was performed. Since the dataset with 39,215 segments is not balanced in terms of curvature, the analysis prepared 500 random samples where Class A through Class E contained 255 data points in each sample, and all the 37 segments under Class F were included. From the ANOVA test for each of these 500 samples, 81.8% of the samples provided statistically significant evidence of the differences in mean crash rates over the different classes of curvature.

Overall, there is statistical evidence that Degree of Curvature significantly influences the crashes even though it does not add much to the model improvement. Considering its significance, including Degree of Curvature in the GWZIP model can be justified. However, the previous analysis observed those segments in Eastern Kentucky where Degree of Curvature is not significant. Further investigating the dataset, it was found that the dataset is dominated by Class A curves, which may affect the significance of curvature in the spatial models for those regions.

## 6.5 Major Findings and Significance of the Analysis

This chapter investigated the spatial effects of the explanatory variables (AADT, L, Average Speed, and Degree of Curvature) on the number of crashes for rural two-lane segments. For this, GWP and GWZIP models were utilized. Based on the performance, this study chose GWZIP model to analyze the results. In addition, it showed a maximum of 32% improvement over the global ZIP model. The GWZIP model provided evidence of the varying effects of the explanatory variables over the spatial domains. The results from these models helped to diagnose the localized influence of the predictor variables. These can be summarized below:

- After analyzing the spatial distribution of the coefficients of AADT from Figure 22, AADT shows higher coefficient values mostly in the Western and Eastern parts. These are mainly rural and less populated areas as shown in Figure 21(c). For these areas, AADT should be considered a more critical factor to analyze crashes.
- Spatial analysis of Average Speed revealed the regions in Northern and Eastern Kentucky, where speed is significant and negatively associated with crashes (see Figure 24(a)-(b)). These roads are mostly with low geometric standards (Figure 19(f) and Figure 19(d) showing narrow shoulders and sharp curves in those areas) and the speed varies between low and medium (Figure 19(c)). To further enhance safety in these areas, measures should be taken in improving road geometrics. Some areas (Figure 24(a)) in Western Kentucky showed that speed affects the crashes positively and speed was the top-ranked factor. This makes sense for these locations considering the standard geometric conditions and low traffic.
- Many segments in Eastern Kentucky showed the Degree of Curvature as the insignificant variable for predicting number of crashes in those locations as shown in Figure 26(b). After analyzing the data for Degree of Curvature, 74% of the study segments are from curve Class A (Figure 19(d)). Later, based on a balanced dataset with respect to curvature class, this analysis observed an increasing average crash rate with increasing curve class from A through F. The increasing relationship was found significant for the balanced dataset from the ANOVA test.

Possible reasons for the insignificance of Degree of Curvature in Eastern Kentucky from spatial models can be due to the imbalanced data of curvature.

While traditional models identify the same factors as significant over the state, the above analysis results based on the spatial models provide an idea of the local factor that can be significant for one region but may not be in another region in the same jurisdiction. Such insights can be applied to prioritize the important local factors of crashes for a road in a certain area. The most important variable in that area can be utilized to plan an efficient improvement strategy. Furthermore, this analysis provided local models for each road. The model of a certain road can be utilized for analyzing the safety performance of a new road within its close proximity. However, this analysis can be limited due to the aggregation process of the segments especially using the curvature class. This may affect the findings related to the curvature. For future analysis, this study will include a more precise measurement of the curvature before developing models.

So far, this study has investigated the effect of speed on the total number of crashes regardless of the severity level. The next chapter focuses on the number of crashes in terms of severity for exploring the effect of speed.

## CHAPTER 7. EFFECT OF SPEED AT DIFFERENT LEVELS OF CRASH SEVERITY

This chapter incorporates speed with the crashes at different levels of severity and investigates the effect of speed. The analysis can be separated into two parts. The first part adopts the traditional count models and identifies the significance of speed at different severity levels. A comparison is made between speed-based models and the models without considering speed as well as the HSM method. The second part investigates the spatially varying effect of speed in addition to other factors on the crashes at different severity levels.

### 7.1 Objective

As previously mentioned in the literature review (Section 2.2.3), limited work has been done to incorporate the effect of speed on crashes at different severity levels for rural two-lane highways. Therefore, this analysis sets the objectives as below:

- Investigate the effect of speed along with geometric and traffic variables on KABC and PDO crashes.
- Explore the spatial effect of the geometric, traffic, and speed variables on KABC and PDO crashes by utilizing the features of nearest neighbors.

### 7.2 Dataset and Variables

The same dataset (Section 6.2) used for the analysis in Chapter 6 was utilized for this analysis. The dataset contains 53,208 segments with a total of 65,091 crashes aggregated from both directions of the road. As shown in Figure 31, these segments consist of 98.3%, 79.9%, and 61.8% zero crashes correspondingly for K crash, Injury crash (A, B, C), and PDO crash. To develop separate models for each severity level, the dataset at least needs a total of 300 crashes per year (117). In case of K crashes, the data contain only 182.8 crashes per year. To avoid the rarity of more severe crashes, there is a practice of combining two or more severities for developing models (141). This analysis combined the K and ABC crashes due to insufficient crash counts under K crash. Figure 32 (a)-(b) presents the distribution of observed number of KABC crashes and PDO

crashes in the study segments across the state. The figures show the hotspots for these crashes. The crash count at these two types of severity levels seems to be higher in Northern and Western Kentucky.

Table 28 presents the statistics of geometric, traffic, and speed attributes on these segments. L, Degree of Curvature, Lane Width, and Shoulder Width represents the geometrics of the roads, whereas, Average Speed, Speed Limit, and Standard Deviation (Std) of speed represent the speed condition on these roads. In addition, the table shows the summary statistics for the crashes of each severity level.

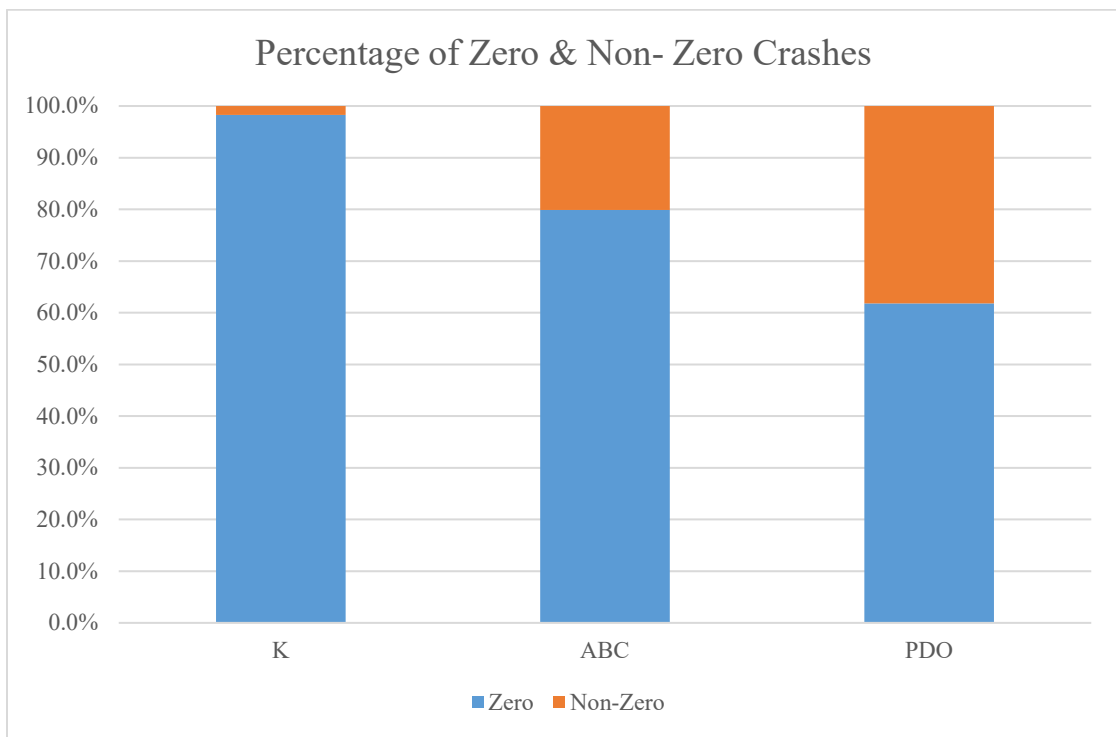
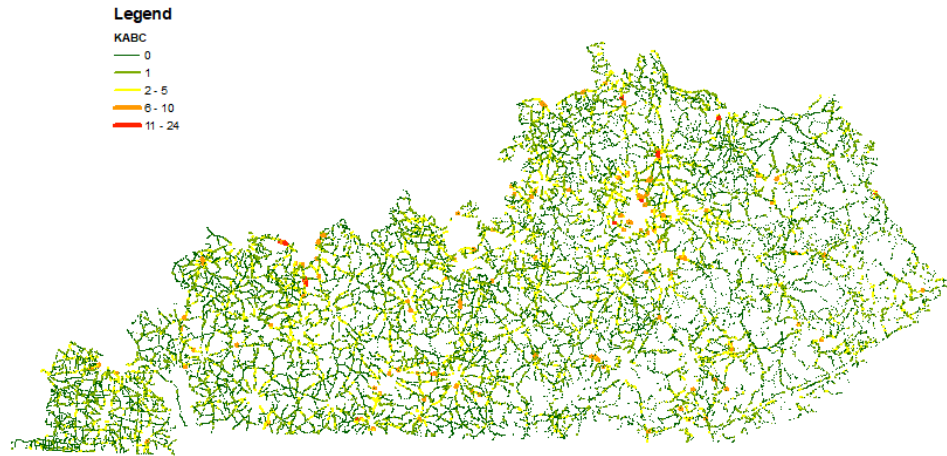
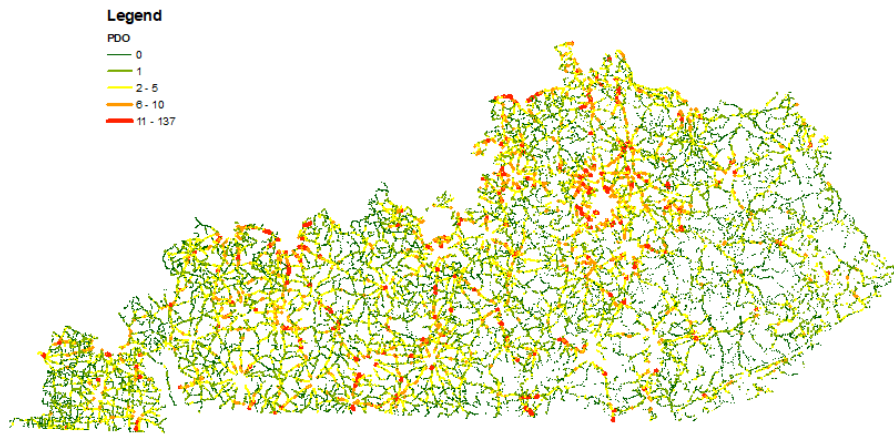


Figure 31 Percentage of Zero and Non- zero crashes for K, ABC, and PDO





(a) Spatial Distribution of Observed Number of KABC Crashes



(b) Spatial Distribution of Observed Number of PDO Crashes

Figure 32 Spatial Distribution of KABC and PDO Crashes

Table 28 Summary Statistics of the Road Attributes

Variables	Unit	Statistics			
		Min.	Max.	Mean	Standard Deviation
AADT	vehicle	2	19619	1355	1772
Segment Length (L)	mile	0.10	2.97	0.26	0.21
Average Speed ( $V_a$ )	mph	5.37	61.76	39.92	9.87
Speed Limit	mph	15	55	53.89	4.22
Standard Deviance (Std) of Speed	mph	6.20	36.74	16.42	4.30
Degree of Curvature ( $Cu$ )	degrees	0	63.81	2.42	3.80
Lane Width (LW)	ft	6	18	9.42	1.14
Shoulder Width (SW)	ft	0	14	3.51	2.06
Number of K Crashes in 5 years		0	2	0.02	0.13
Number of ABC Crashes in 5 years		0	24	0.30	0.78
Number of KABC Crashes in 5 years		0	24	0.32	0.80
Number of PDO Crashes in 5 years		0	137	0.91	2.31

### 7.3 Analysis based on Traditional Count Models

Following the existing practice with crashes aggregated at segment level, this study developed separate count models for KABC and PDO crashes (27; 31; 35; 37; 38; 108). This section tests Average Speed and Std of speed separately for each severity level and finds out their significance in KABC and PDO crashes. The performances of the speed-based models were checked with the traditional model without speed. The best

performing model was selected for further analysis. Additional analysis investigated the effect of the explanatory variables at different speed regions for different severity levels. Separate models were developed for each speed region. Finally, the overall performance of these models was compared with HSM model (Section 4.4.1) in addition to the single model.

### 7.3.1 Model Development for KABC and PDO crashes

Following the existing studies that looked at specifically crash severity, this analysis considered AADT, L, Degree of Curvature, Shoulder Width, and Lane Width as the explanatory variables in developing models for KABC and PDO crashes (27; 35; 37; 38; 109). In addition, the analysis considered Average Speed and Std of speed as the speed factors to find whether speed plays a significant role in crashes of different severity levels. As the response variables, the number of KABC crashes and number of PDO crashes were used.

A multicollinearity check was performed before finalizing the variables for model development. Pearson correlation coefficient showed shoulder width and land width were highly correlated with AADT (Table 30). Therefore, these two variables were not included in the model. As for the speed variables, average speed and Std of speed were experimented. This analysis utilized ZIP model form (Section 4.4.2.1.3) for both KABC and PDO crashes. The following 3 models were evaluated separately for KABC and PDO crashes with the rural two-lane segments. The traditional form without speed was included to provide a baseline for other models. This is to compare if and how the inclusion of speed as a factor in the KABC and PDO crash prediction models helps to improve the prediction performance. To compare the performance of the models, AIC, BIC,  $R^2$ , MAPE, RMSE, and MAD were utilized.

- (1) Model using AADT, L,  $Cu$ ,
- (2) Model using AADT, L,  $Cu$ , and  $V_a$
- (3) Model using AADT, L  $Cu$ , and Std of speed

Table 29 Multicollinearity Check

	AADT	Length	Lane Width	Shoulder Width	Degree of Curvature	Average Speed	Std	KABC	PDO
AADT	1.000000	0.125531	0.613581	0.564023	-0.130119	0.496005	-0.476800	0.347804	0.400887
Length	0.125531	1.000000	0.110590	0.168297	-0.293377	0.280669	0.031417	0.271152	0.301465
Lane Width	0.613581	0.110590	1.000000	0.608630	-0.075209	0.550611	-0.389483	0.205235	0.242013
Shoulder Width	0.564023	0.168297	0.608630	1.000000	-0.106528	0.469152	-0.272384	0.135600	0.185950
Degree of Curvature	-0.130119	-0.293377	-0.075209	-0.106528	1.000000	-0.178460	0.003373	-0.052838	-0.091217
Average Speed	0.496005	0.280669	0.550611	0.469152	-0.178460	1.000000	-0.495122	0.244995	0.246305
Std	-0.476800	0.031417	-0.389483	-0.272384	0.003373	-0.495122	1.000000	-0.191741	-0.201884
KABC	0.347804	0.271152	0.205235	0.135600	-0.052838	0.244995	-0.191741	1.000000	0.575039
PDO	0.400887	0.301465	0.242013	0.185950	-0.091217	0.246305	-0.201884	0.575039	1.000000

For model development, 80% of the segments were used, and the remaining segments were utilized as the testing set. The parameter estimates and the performance of the models are presented in Table 29. All the variables in the KABC and PDO models were found significant except for Std of speed for KABC crashes. All three models under each severity level seem to perform similarly. Considering Average Speed better reflects the operating condition of the rural two-lane highways, the average speed-based models presented in Equation (54) and Equation (55) were chosen to proceed with subsequent analysis.

Table 30 Parameter Estimates KABC and PDO Models and Goodness-of-Fit

(i) Models for KABC Crashes

Variables	Traditional Model		Average Speed Model		Std of Speed Model	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Intercept, $\epsilon$	-5.073	0.078	-4.996	0.081	-5.239	0.131
Ln (AADT)	0.729	0.010	0.752	0.012	0.741	0.012
Ln (L)	0.783	0.016	0.801	0.016	0.781	0.016
$V_a$	-	-	0.005	0.001		
$Cu$	0.053	0.003	0.053	0.003	0.054	0.003

<b>Std</b>			<i>0.005</i>	<i>0.003</i>
<b>AIC</b>	46001	45990	46000	
<b>BIC</b>	46043	46041	46051	
<b>R<sup>2</sup></b>	0.21	0.21	0.21	
<b>RMSE</b>	0.71	0.71	0.71	
<b>MAPE (%)</b>	63.66	63.65	63.59	
<b>MAD</b>	0.4	0.4	0.4	

*\*\*Note: parameter estimates in red italic are insignificant at a 5% significance level*

(ii) Models for PDO Crashes

<b>Variables</b>	<b>Traditional Model</b>		<b>Average Speed Model</b>		<b>Std of Speed Model</b>	
	<b>Estimate</b>	<b>Std. Error</b>	<b>Estimate</b>	<b>Std. Error</b>	<b>Estimate</b>	<b>Std. Error</b>
<b>Intercept, <math>\varepsilon</math></b>	-4.318	0.047	-4.040	0.049	-4.646	0.078
<b>Ln (AADT)</b>	0.756	0.006	0.832	0.007	0.780	0.007
<b>Ln (L)</b>	0.801	0.009	0.867	0.010	0.799	0.009
<b><math>V_a</math></b>			-0.017	0.001		
<b><math>Cu</math></b>	0.030	0.002	0.029	0.002	0.031	0.002
<b>Std</b>					0.010	0.002
<b>AIC</b>	83943		83523		83917	
<b>BIC</b>	83986		83574		83968	
<b>R<sup>2</sup></b>	0.30		0.31		0.30	
<b>RMSE</b>	1.81		1.79		1.80	
<b>MAPE (%)</b>	57.75		57.26		57.68	
<b>MAD</b>	0.81		0.80		0.81	

*\*\*Note: parameter estimates in red italic are insignificant at a 5% significance level*

*Expected Number of KABC Crash*

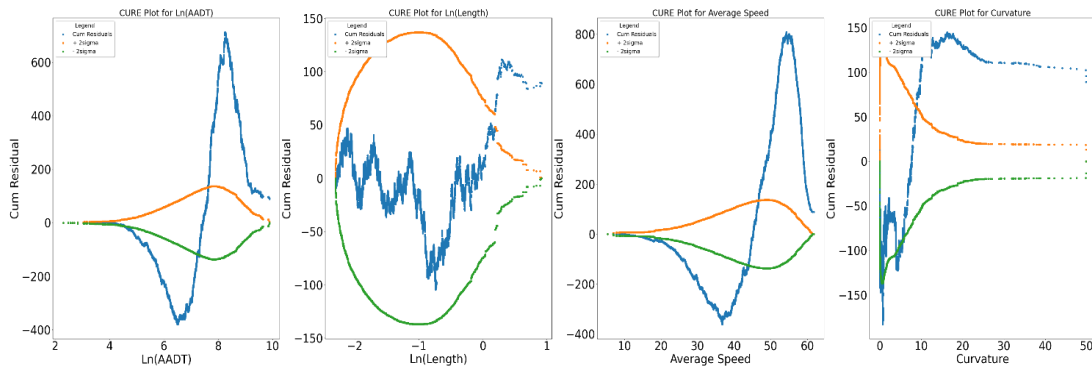
$$= e^{(-4.996+0.752 \ln(AADT)+0.801\ln(L)+0.005V_a+0.053Cu)} \quad (54)$$

*Expected Number of PDO Crash*

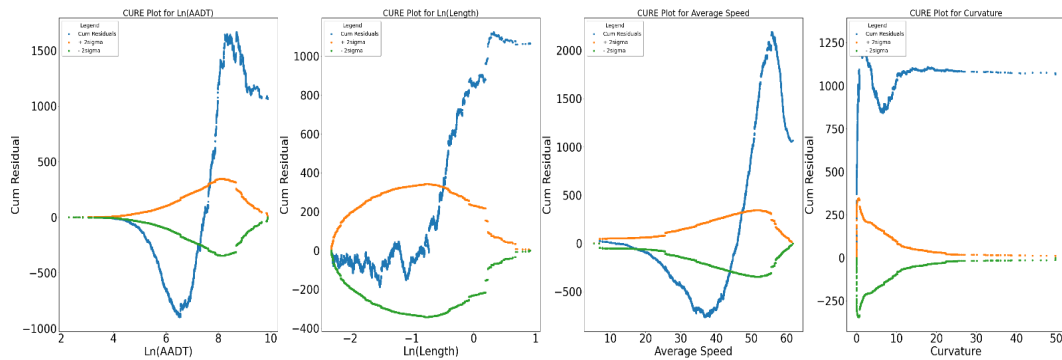
$$= e^{(-4.040+0.832 \ln(AADT)+0.867\ln(L)-0.017V_a+0.029Cu)} \quad (55)$$

In Equation (54) and Equation (55), AADT and L are positively affecting both types of crashes, which is consistent with the existing literature (27; 35; 37; 38; 109). Average Speed shows different relationships for KABC and PDO crashes. The association between Average Speed and KABC crashes is positive. More KABC crashes tend to occur at a higher speed of the study segments. This is in line with the study by Wang et. al. where the authors found that number of severe crashes (especially fatal crashes) is positively affected by the average speed (71). In contrast, Average Speed is negatively related to PDO crashes. After analyzing the data, such negative association for PDO crashes actually reflects those facilities where the geometric condition is better. This inverse relationship between Average Speed and PDO crashes is also consistent with an existing study by Dutta and Fontaine (27). In case of Degree of Curvature, the effect is positive in both types of crashes and aligns with some of the existing findings (27; 38).

To investigate how the functional form of the models in Equation (56) and Equation (57) fit the data, this analysis looked at CURE plots. Figure 33 (a)-(b) shows the CURE plots for all the explanatory variables in the models. Clearly, the models are not fitting the data very well since a significant portion of the CURE is outside the boundary of  $\pm 2\sigma$  for most of the variables. Moreover, it seems that both KABC and PDO models are constantly overpredicting or underpredicting crashes in the higher speed and higher AADT ranges. Further looking into the plot for Average Speed, it seems there are three distinctive speed ranges where the model is consistently overpredicting or underpredicting outside of the preferable ranges. This observation is similar to what was found in case of modeling total number of crashes using all rural two-lane segments. This analysis adopted the same approach of using speed as a categorizer to further investigate the crashes at different severity levels considering geometric, traffic, and speed factors. The related analysis is discussed in the next subsection.



(a) KABC Crash



(b) PDO Crash

Figure 33 CURE Plots for Average Speed Model

### 7.3.2 Severity Analysis at Different Speed Ranges

This subsection explores how the effect of different explanatory variables varies over the KABC and PDO crash severity levels while disaggregating the models by speed ranges of the rural two-lane segments. From Figure 33(a)-(b), it appears that the models are gradually overpredicting the number of both KABC and PDO crashes up to an average speed of 35 mph. After 35 mph, each model starts underpredicting, which continues until Average Speed is approximately 50 mph, after which the overprediction begins. Based on these transitions of CURE plot for Average Speed, the study segments were grouped into three-speed categories to develop separate models for both KABC and PDO. The three-speed ranges are labeled as low (below 35 mph), medium (between 35

mph and 50 mph), and high speed (above 50 mph), respectively. They represent about 32%, 50.6%, and 17.4% of total segments, correspondingly. The general approach for analyzing severity at each speed range has been laid out below:

- Select all the road attributes, i.e., L, Degree of Curvature, Shoulder Width, Lane Width, AADT along with Average Speed as the initial explanatory variable set.
- For each speed range, check the multicollinearity and finalize the explanatory variables.
- Utilizing the final variables, develop ZIP models for KABC and PDO crashes separately for low, medium, and high-speed roads utilizing 80% of the segments under each speed range.
- Investigate how different the influence of each factor is on KABC and PDO crashes at the three categories of speed. This is done by looking at the variable coefficients from each model.
- Look at the performances of KABC and PDO models in the three-speed categories. Later, the performances of the models from different speed categories were compared with the previously developed average speed-based models, i.e., Equation (56) and Equation (57) and HSM approach.

For low-speed and medium-speed roads, there was no significant high multicollinearity for the explanatory variables (Table 31). However, the high-speed roads showed high multicollinearity for the Shoulder Width and Lane Width as depicted in Table 31(c). Based on these observations, this analysis considered all the geometric variables for the low and medium-speed roads, while excluding Shoulder Width and Lane Width from the models of high-speed roads. Table 32 shows the parameter estimates and model performances for KABC and PDO.



Table 31 Multicollinearity Check for Low, Medium, and High-Speed Roads

(a) Low-Speed Roads									
	AADT	Length	Lane Width	Shoulder Width	Degree of Curvature	Average Speed	Std	KABC	PDO
AADT	1.000000	-0.043618	0.389584	0.260542	-0.052088	0.125417	-0.254936	0.319823	0.488201
Length	-0.043618	1.000000	-0.045786	0.043190	-0.251690	0.048549	0.247247	0.104853	0.115301
Lane Width	0.389584	-0.045786	1.000000	0.300626	0.003820	0.231757	-0.063933	0.139655	0.185044
Shoulder Width	0.260542	0.043190	0.300626	1.000000	-0.043054	0.070790	0.036970	0.082938	0.165371
Degree of Curvature	-0.052088	-0.251690	0.003820	-0.043054	1.000000	-0.052840	-0.092806	0.001693	-0.049338
Average Speed	0.125417	0.048549	0.231757	0.070790	-0.052840	1.000000	-0.031077	0.076276	0.056958
Std	-0.254936	0.247247	-0.063933	0.036970	-0.092806	-0.031077	1.000000	-0.066938	-0.094546
KABC	0.319823	0.104853	0.139655	0.082938	0.001693	0.076276	-0.066938	1.000000	0.426987
PDO	0.488201	0.115301	0.185044	0.165371	-0.049338	0.056958	-0.094546	0.426987	1.000000

(b) Medium-Speed Roads									
	AADT	Length	Lane Width	Shoulder Width	Degree of Curvature	Average Speed	Std	KABC	PDO
AADT	1.000000	-0.017487	0.488348	0.358632	-0.071673	0.249636	-0.341124	0.337341	0.463781
Length	-0.017487	1.000000	-0.056486	0.057376	-0.313390	0.126010	0.223508	0.213786	0.291580
Lane Width	0.488348	-0.056486	1.000000	0.412153	0.030939	0.215162	-0.235749	0.140637	0.200139
Shoulder Width	0.358632	0.057376	0.412153	1.000000	-0.046948	0.174461	0.007415	0.073260	0.147402
Degree of Curvature	-0.071673	-0.313390	0.030939	-0.046948	1.000000	-0.081876	-0.098022	-0.021988	-0.078328
Average Speed	0.249636	0.126010	0.215162	0.174461	-0.081876	1.000000	-0.259633	0.139888	0.161876
Std	-0.341124	0.223508	-0.235749	0.007415	-0.098022	-0.259633	1.000000	-0.120253	-0.142992
KABC	0.337341	0.213786	0.140637	0.073260	-0.021988	0.139888	-0.120253	1.000000	0.496731
PDO	0.463781	0.291580	0.200139	0.147402	-0.078328	0.161876	-0.142992	0.496731	1.000000

(c) High-Speed Roads									
	AADT	Length	Lane Width	Shoulder Width	Degree of Curvature	Average Speed	Std	KABC	PDO
AADT	1.000000	-0.015114	0.621673	0.522043	-0.052060	0.361547	-0.474958	0.187620	0.202401
Length	-0.015114	1.000000	-0.015713	0.002389	-0.291914	0.154472	0.094770	0.273286	0.275804
Lane Width	0.621673	-0.015713	1.000000	0.778637	0.010457	0.561911	-0.389600	0.034792	0.090549
Shoulder Width	0.522043	0.002389	0.778637	1.000000	0.016408	0.597192	-0.312582	-0.035224	0.015315
Degree of Curvature	-0.052060	-0.291914	0.010457	0.016408	1.000000	-0.116106	-0.054317	-0.016876	-0.069678
Average Speed	0.361547	0.154472	0.561911	0.597192	-0.116106	1.000000	-0.483281	-0.002460	0.046750
Std	-0.474958	0.094770	-0.389600	-0.312582	-0.054317	-0.483281	1.000000	-0.080129	-0.094768
KABC	0.187620	0.273286	0.034792	-0.035224	-0.016876	-0.002460	-0.080129	1.000000	0.624337
PDO	0.202401	0.275804	0.090549	0.015315	-0.069678	0.046750	-0.094768	0.624337	1.000000

Table 32 ZIP Models for Low, Medium, and High-Speed Roads

Variable	Low-Speed Roads 16,964 Segments				Medium-Speed Roads 26,906 Segments				High-Speed Roads 9,338 Segments							
	KABC		PDO		KABC		PDO		KABC		PDO					
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error				
Intercept, $\epsilon$	-	6.041	-	0.304	-	4.999	-	0.172	5.573	0.206	-4.255	0.124	1.012	0.345	<i>0.266</i>	<i>0.189</i>
<b>Ln (AADT)</b>	0.869	0.033	0.974	0.019	0.898	0.019	0.931	0.011	0.616	0.023	0.686	0.014				
<b>Ln (L)</b>	0.850	0.054	0.836	0.030	0.845	0.023	0.901	0.014	0.876	0.027	0.919	0.016				
<b><math>V_a</math></b>	0.014	0.007	-	<i>0.002</i>	<i>0.004</i>	0.008	0.003	-0.005	0.002	0.096	0.007	0.075	0.004			
<b><math>C_u</math></b>	0.041	0.006	0.013	0.004	0.075	0.004	0.047	0.003	0.045	0.005	0.027	0.004				
<b>LW</b>	-	<i>0.014</i>	-	<i>0.032</i>	0.060	0.018	0.078	0.020	-0.084	0.012						

<hr/>							
<b>SW</b>	<i>-</i>				<i>-</i>		
	<i>0.025</i>	<i>0.021</i>	0.031	0.010	0.064	0.009	-0.029
							0.005
<hr/>							
<b>Performance Measures</b>	<b>KABC</b>	<b>PDO</b>	<b>KABC</b>	<b>PDO</b>	<b>KABC</b>	<b>PDO</b>	
<hr/>							
<b>AIC</b>	7263	14663	24128	42138	13945	25409	
<hr/>							
<b>BIC</b>	7322	14722	24191	42201	13986	25368	
<hr/>							
<b>R<sup>2</sup></b>	0.15	0.30	0.23	0.39	0.17	0.21	
<hr/>							
<b>RMSE</b>	0.37	0.94	0.67	1.39	1.09	3.22	
<hr/>							
<b>MAPE (%)</b>	81.01	67.38	63.47	53.93	49.99	56.47	
<hr/>							
<b>MAD</b>	0.17	0.41	0.41	0.79	0.70	1.42	
<hr/>							

*\*\*Note: parameter estimates in red italic are insignificant at a 5% significance level*

Based on the KABC and PDO models presented in Table 32, below is the discussion on how each variable is affecting the number of crashes at each severity level while disaggregating the models by speed level of the rural two-lane segments.

- **AADT:** Apparently, AADT is significant and positively affects both KABC and PDO crashes on low, medium, and high-speed roads.
- **Length:** At each speed level, L is significant and positively related to the KABC and PDO crashes.
- **Average Speed:** In case of Average Speed, the results show a varying effect over the low, medium, and high-speed roads while considering different levels of severity for crashes. These are:
  - **Low-Speed Roads:** Speed is only significant for severe crashes, i.e., KABC crashes. These roads mostly have poor geometric conditions. From Figure 34, the roads clearly show the presence of narrow shoulders and lanes as well as more curves. On average, the shoulder width is around 2 ft and lane width is around 8.7 ft for these roads. It is understandable that a crash can be severe when speed goes up under such restrictive geometric conditions of the roads.
  - **Medium-Speed Roads:** Speed is statistically significant for both KABC and PDO crashes. It is positively related to the KABC crashes and negatively related to PDO crashes, which is consistent with the initially developed Average Speed Model. However, the effect of speed on KABC crashes is comparatively lower than the low-speed roads as indicated by the coefficient value for these roads. Further, the geometric condition seems to be moderate for these roads from Figure 34. According to the study data, 59% of the medium-speed roads have lanes of less than 10 ft and 94% of these roads have shoulders of less than 6 ft.
  - **High-Speed Roads:** Speed is statistically significant for both KABC and PDO crashes, and it is negatively correlated with each severity level. It is different from what was observed on low and medium-speed roads. These roads are actually the high geometric standard roads indicated in Figure 34. Compared to low and medium-speed roads, the average lane width of these

roads is higher than 10 ft with the presence of more shoulders and more straight sections. Drivers tend to travel at a high speed due to these features. For this group of roads, if the speed of a segment is higher than another segment, that road may be safer conditioned on its geometric condition being better than the other road.

- **Degree of Curvature:** It is significant and positively related to the KABC and PDO crashes at each speed level.
- **Lane Width:** Lane Width was only considered for low and medium-speed roads. For low-speed roads, lane width is only significant for PDO crashes. A wider lane can significantly reduce the number of PDO crashes on these roads. For medium-speed roads, Lane Width had a significant negative relationship with the crashes of each severity level. Wider lanes can reduce crashes at each severity level for these roads.
- **Shoulder Width:** Shoulder Width was only considered for low and medium-speed roads. For low-speed roads, Shoulder Width is only significant for PDO crashes, and it is positively related. Such relationship reflects the narrow shoulder width on these roads. The Average Speed Model did not capture this type of effect of Shoulder Width. On the other hand, Shoulder Width is significant and negatively correlated with both KABC and PDO crashes of medium speed roads. A 1 ft increase in Shoulder Width can reduce more KABC crashes (6.20%) than PDO crashes (2.86%) implying that a wider shoulder has a higher influence in minimizing the severity of a crash on these roads.

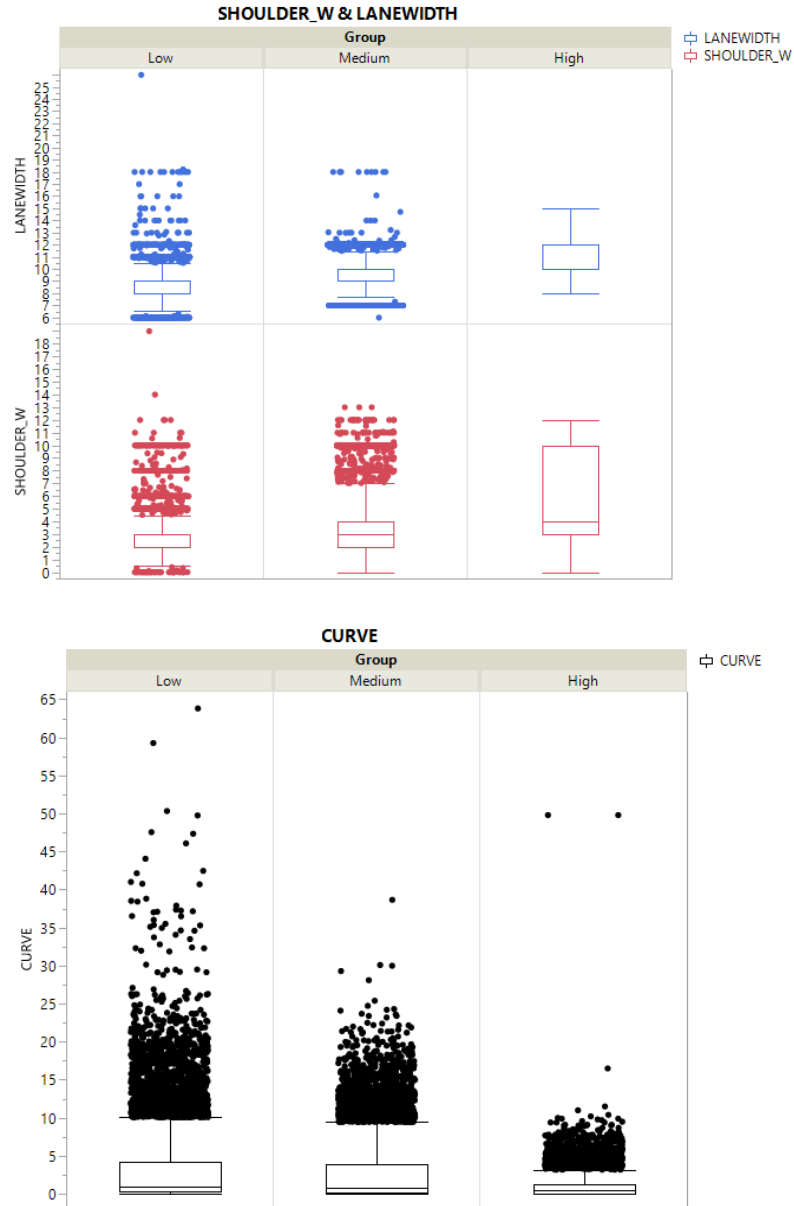


Figure 34 Distribution of Shoulder Width, Lane Width, and Degree of Curvature over Low, Medium, and High-Speed Roads

From the above discussion, there exists a varying effect of the speed and geometric variables on KABC and PDO crashes when speed is considered to divide the rural two-lane highway dataset. This was not captured by the initially developed Average Speed Model, which implies developing separate models for different speed levels. To analyze crashes at different severity levels, speed tends to be a better surrogate for geometric

conditions of low-speed roads compared to medium-speed roads. The geometric condition should be given priority while providing countermeasures, especially for KABC crashes. For high-speed roads, the number of severe and PDO crashes tends to be low under standard geometric conditions.

This analysis further investigated how separating the KABC and PDO models for low, medium, and high-speed ranges can improve the model performance. For both KABC and PDO, the predicted number of crashes from the low, medium, and high-speed models were combined so that their overall performance could be compared to the single model, i.e., Average Speed Model. In addition, the combined performance was compared with No Speed models and HSM-based models. Table 33(i)-(ii) presents the comparisons for the performances. While HSM model performs the worst, the combination of low, medium, and high-speed models performs best, and, Average Speed Model is the second best model for both KABC and PDO crashes. For KABC crashes, the improvement was a maximum of 47% compared to the HSM model, and, for PDO crashes, there was a maximum of 22% improvement with respect to HSM. The CURE plots in Figure 35(i)-(ii) show further evidence of improvement. For each severity level, the combination of low, medium, and high-speed models fits the data best.

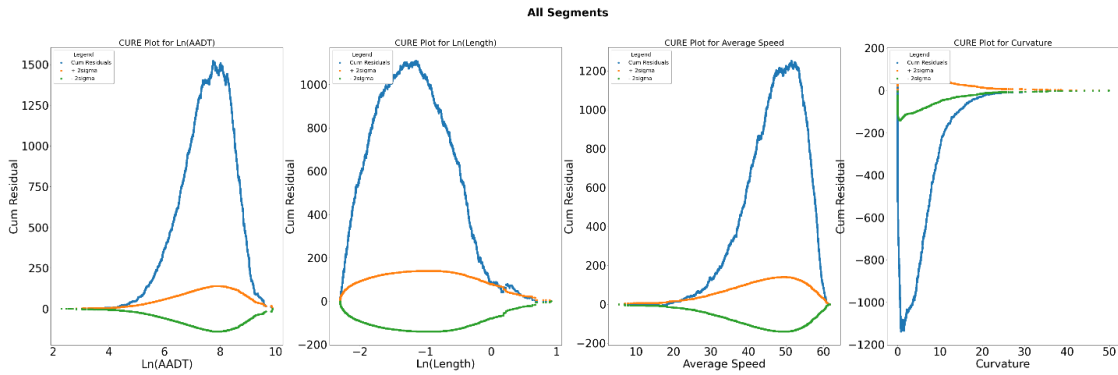
Table 33 Performance Comparisons

(i) Models Tested for KABC Crash

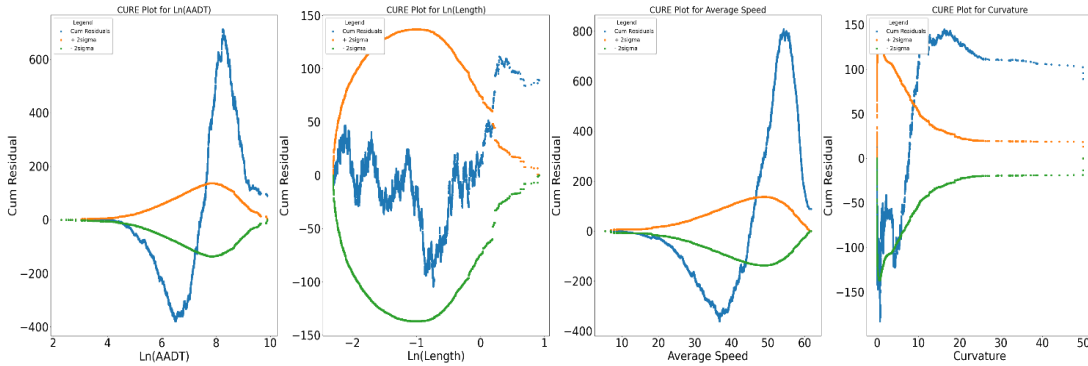
KABC Models	$R^2$	RMSE	MAPE (%)	MAD
No Speed Model	0.21	0.71	63.66	0.40
Average Speed Model	0.21	0.71	63.65	0.40
HSM Model	0.17	0.73	68.03	0.39
Low, Medium and High-Speed Models	0.25	0.69	61.3	0.38

(ii) Models Tested for PDO Crash

PDO Models	$R^2$	RMSE	MAPE (%)	MAD
No Speed Model	0.30	1.81	57.75	0.81
Average Speed Model	0.31	1.79	57.26	0.80
HSM Model	0.27	1.83	68.51	0.80
Low, Medium and High-Speed Models	0.33	1.76	56.90	0.78

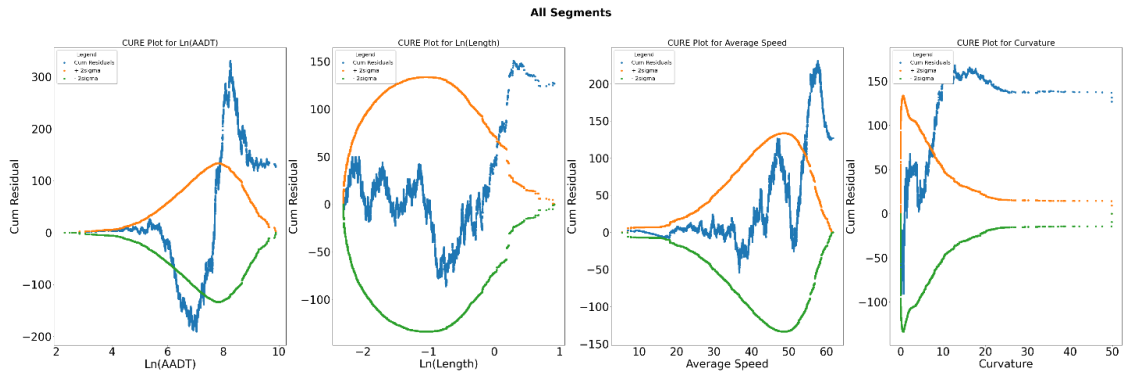


(a) HSM Model



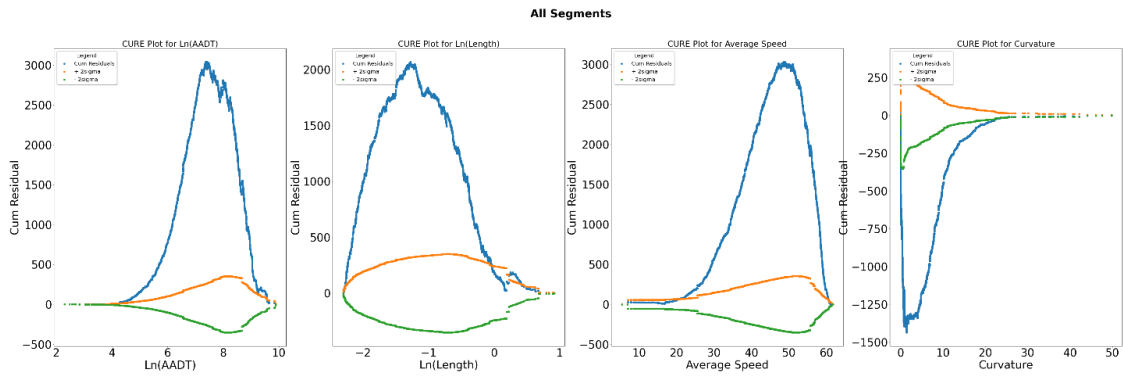
(b) Average Speed Model



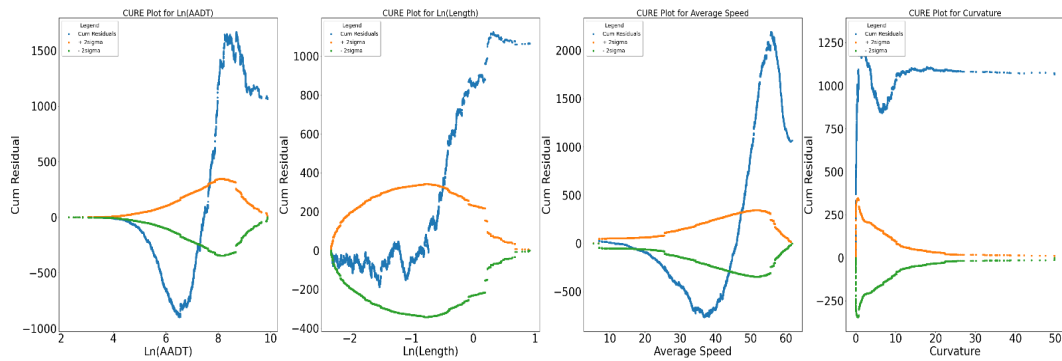


(c) Combined Models

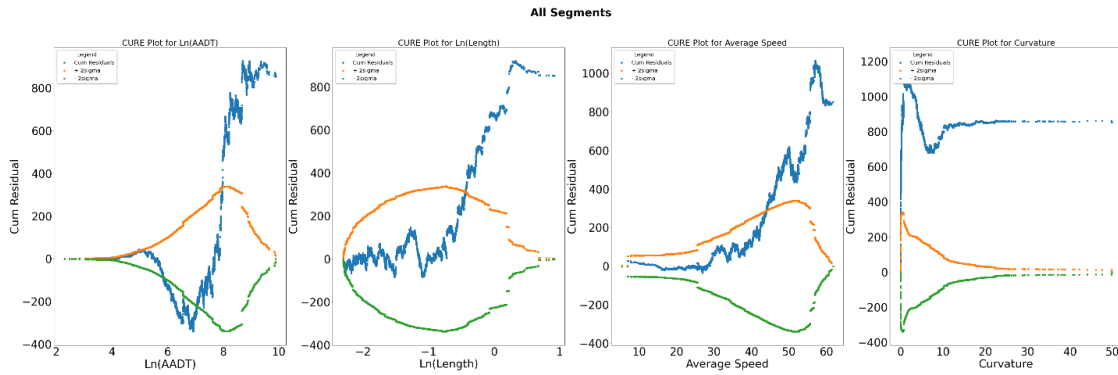
(i) CURE Plots for KABC



(a) HSM Model



(b) Average Speed Model



**(c) Combined Models**

**(i) CURE Plots for PDO**

Figure 35 CURE Plots for KABC and PDO Models

Overall, the analysis in this section showed that speed is indeed a significant factor for each severity level while controlling for geometric and traffic attributes. It also revealed variability in speed's effect for KABC and PDO crashes. It is positively associated with the KABC crashes and negatively related to the PDO crashes. Speed was further used as the categorizer to develop separate KABC and PDO models at low, medium, and high-speed ranges. This provided a different picture of the effect of exploratory variables which was not identified based on the single model, i.e., Average Speed Model. Speed was only found significant for KABC crashes at low speed and seems to be a better surrogate for the geometric condition compared to medium-speed roads. For high-speed roads, the number of severe crashes tends to be low under standard geometric conditions. In addition to these findings, the combined performance of low, medium, and high-speed models outperformed the Average Speed Models as well as the fixed proportion-based HSM models. This suggests developing separate count models for KABC and PDO instead of applying the fixed proportions of different severity to the total number of crashes predicted by the HSM model. In addition, speed should be considered as a categorizer variable to achieve further improvement by developing models at different speed ranges.

## 7.4 Spatial Analysis

Previous section explored the effect of geometric, traffic, and speed variables at different severity levels using ZIP model. The model mainly provided average estimates of the effect. While using a statewide dataset for crash and severity analysis, it is possible that the explanatory variables are spatially distributed showing spatial autocorrelation. This section utilizes GWR models to incorporate the spatial autocorrelation and investigates the locally varying effect of the factors on KABC and PDO crashes.

### 7.4.1 Spatial Modeling and Results

To perform the spatial analysis, all 53,208 homogenous segments were utilized. AADT, L, Average Speed, and Degree of Curvature were selected as the explanatory variables, whereas number of KABC crashes in 5 years and number of PDO crashes in 5 years as the response variables. Due to high multicollinearity, Shoulder Width and Lane Width were not included in the spatial models.

Before fitting spatial models for KABC and PDO crashes, this analysis checked the spatial auto collinearity for the selected explanatory variables using Moran's I. Table 34 shows Moran's I values for the variables. For all explanatory variables and response variables, the values are positive and significant at a 5% confidence level. It means the variables are showing significant spatial autocorrelation. The proof of spatial autocorrelation supports the idea of testing the spatial models for this analysis.

Table 34 Spatial Dependency of the Variables

Variables	Moran's I	P-value	Clustered/Spatial Autocorrelation
AADT	0.4778	0	Yes
L	0.1213	0	Yes
$V_a$	0.3984	0	Yes
$Cu$	0.0988	0	Yes

Number of KABC Crash	0.0078	0	Yes
Number of PDO Crash	0.0133	0	Yes

As the spatial models, GWP and GWZIP models were adopted for both KABC and PDO crashes. For GWP model, the optimum nearest number of neighbors was estimated as 1,360 and 1,800, respectively for KABC and PDO crashes. For GWZIP model, 2,500 and 2,600 were determined as the number of neighbors correspondingly for KABC and PDO crashes. The number of neighbors used for each model meets the sample size requirement by HSM and *Safety Performance Function Decision Guide*. The performances of the GWP and GWZIP models were compared with the global models, i.e., Poisson model and ZIP model. To evaluate the model performance,  $R^2$  and RMSE were used.

Table 35 presents the coefficients of the variables estimated from each model. The global models in Table 35(i) and Table 35(iii) provide the coefficient values for each variable assuming their influence on the number of KABC and PDO crashes remains constant regardless of spatial variation. The effect of all variables was found to be statistically significant at a 5% level.

Table 35(ii) and Table 35(iv) provide the descriptive statistics of coefficient values for each variable from the local models (i.e., GWP and GWZIP). Like the global models, both GWP and GWZIP models show a positive influence of AADT and L on both KABC and PDO crashes. The minimum coefficient value for Degree of Curvature suggests that there are locations where the local models determined a negative relationship between number of KABC crashes and Degree of Curvature and between number of PDO crashes and Degree of Curvature. Such relationship seems to be counterintuitive. After investigating the negative coefficients, this analysis observed that all these negative coefficients depicted no statistical significance in each model. Other existing research observed similar cases of negative relationships for Degree of Curvature from geographically weighted regression models (93), and one of the reasons for these counterintuitive signs can be that some variables may not be significant in certain road segments, therefore, it is possible that the local models estimate counterintuitive coefficients for those variables (83; 138). In case of Average Speed, both positive and

negative influences on KABC and PDO crashes were observed. Further investigation of the results related to the effect of Average Speed from the local models is discussed later in this section.

The performances in Table 35 show better fits for the local models compared to their corresponding global models. Between the spatial models, GWP seems to perform better for each severity level. This analysis chose the GWP model to proceed with the further discussion on the spatial variation of the coefficients (mainly AADT, Average Speed, and Degree of Curvature as Length didn't show any specific patterns) in the subsections below.

Table 35 Variable Coefficients and Model Performance

(i) Global Models for KABC Crash

Model	Poisson Model			ZIP		
	Coefficient	Std. Error	z-value	Coefficient	Std. Error	z-value
Intercept	-5.1735	0.0622	-83.2010	-3.3587	0.1274	-26.3580
Ln(AADT)	0.7541	0.0088	85.2200	0.6117	0.0149	41.1230
Ln(L)	0.8327	0.0127	65.7190	0.6744	0.0204	32.9950
$V_a$	-0.0069	0.0011	-6.3030	-0.0175	0.0018	-9.4810
$C_u$	0.0542	0.0022	25.1310	0.0501	0.0032	15.5570
$R^2$	0.2037			0.2138		
RMSE	0.7173			0.7128		

\*Note: all variables showed p-value < 2e-16

(ii) Local Models for KABC Crash

Model	GWP				GWZIP			
	Min	Max	Mean	SD	Min	Max	Mean	SD
Intercept	-	-	-	-	-	-	-	-
	9.2805	-2.2286	-5.5545	1.0212	-8.1652	-0.4752	3.9789	1.3765
Ln(AADT)	0.1509	1.1516	0.7595	0.1452	0.1126	1.1591	0.6604	0.1638

Ln(L)	0.2034	1.3165	0.8297	0.1552	0.0122	1.2907	0.7129	0.1871
$V_a$	-						-	
	0.0591	0.0920	-0.0013	0.0185	-0.0870	0.0850	0.0146	0.0264
$C_u$	-							
	0.0598	0.1751	0.0726	0.0389	-0.0373	0.2249	0.0793	0.0490
Bandwidth		1360				2500		
$R^2$	0.2851				0.2212			
RMSE	0.6796				0.7094			

(iii) Global Models for PDO Crash

Model	Poisson Model			ZIP		
	Coefficient	Std. Error	z-value	Coefficient	Std. Error	z-value
Intercept	-4.3067	0.0373	115.3800	-2.9444	0.0533	-55.2100
Ln(AADT)	0.8603	0.0052	166.2300	0.7253	0.0068	107.1200
Ln(L)	0.9148	0.0076	120.8600	0.7743	0.0095	81.6400
$V_a$	-0.0176	0.0006	-28.1100	-0.0228	0.0007	-30.4300
$C_u$	0.0276	0.0017	16.7000	0.0346	0.0020	17.4100
$R^2$	0.2691			0.2795		
RMSE	1.9717			1.9576		

\*Note: all variables showed p-value < 2e-16

(iv) Local Models for PDO Crash

Model	GWP				GWZIP			
	Min	Max	Mean	SD	Min	Max	Mean	SD
Intercept	-							
	7.3649	-2.0086	-4.7999	1.0720	-7.2215	-1.1539	3.5831	1.1718
Ln(AADT)	0.4557	1.3345	0.9018	0.1548	0.3312	1.2474	0.7828	0.1620
Ln(L)	0.5818	1.3324	0.8731	0.1018	0.3810	1.1709	0.7581	0.1198

$V_a$	-						-	
	0.0689	0.0723	-0.0167	0.0160	-0.0743	0.0579	0.0216	0.0173
$C_u$	-							
	0.0489	0.1453	0.0433	0.0307	-0.0615	0.1299	0.0483	0.0312
Bandwidth		1800				2600		
$R^2$	0.3600				0.3322			
RMSE	1.8449				1.8846			

#### 7.4.1.1 AADT

Figure 36 shows the distribution of the AADT coefficients and the percent changes in KABC and PDO crashes for a 10% increase in AADT based on GWP. AADT was found significant for all the segments in terms of PDO crashes. In contrast, 99.75% of segments showed AADT as the significant variable for KABC crashes. The insignificant areas (Southern Kentucky) for KABC crashes are mostly low-volume areas Figure 19(b) with medium to high speeds Figure 19(c). Furthermore, sharp curves are present in these areas.

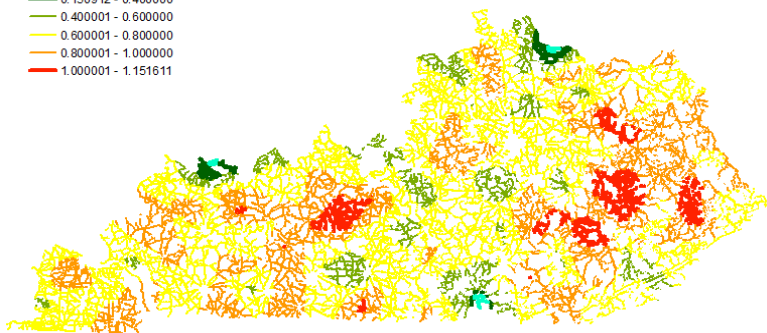
In terms of the effect of AADT changes on KABC and PDO crashes, Eastern Kentucky has the highest impact (in red) of AADT on PDO crashes. In case of KABC, AADT is comparatively less impactful in this region. The roads within this region are mostly below standards, and the average speed is low to medium (Figure 19(c)). People are used to driving in this area with narrow or no shoulders as well as narrow lanes (Figure 19(e)-(f)). Western Kentucky also shows a similar picture for KABC and PDO crashes. These are high-speed roads (Figure 19(c)) with flat terrain, wider lanes, and shoulders shown in Figure 21(b) and Figure 19(e)-(f).

GWP Model

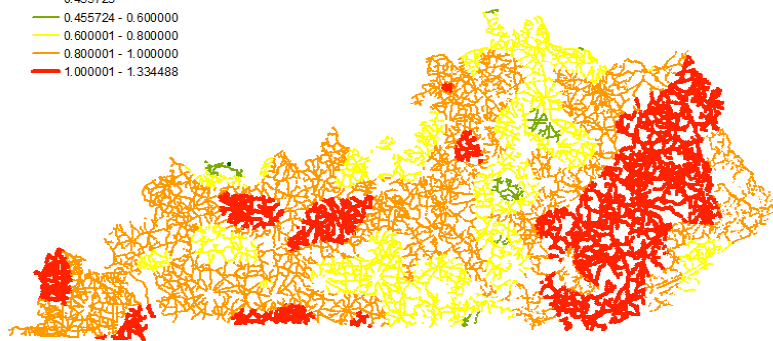
KABC

PDO

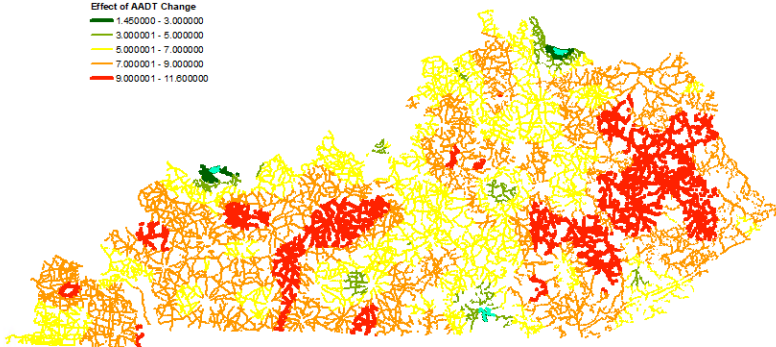
**Legend**  
 LnAADT Insignificant  
**LnAADT Coefficient**  
 0.150912 - 0.400000  
 0.400001 - 0.600000  
 0.600001 - 0.800000  
 0.800001 - 1.000000  
 1.000001 - 1.151611



**Legend**  
**LnAADT Coefficient**  
 0.455723  
 0.455724 - 0.600000  
 0.600001 - 0.800000  
 0.800001 - 1.000000  
 1.000001 - 1.334488



**Legend**  
 LnAADT Insignificant  
**Effect of AADT Change**  
 1.450000 - 3.000000  
 3.000001 - 5.000000  
 5.000001 - 7.000000  
 7.000001 - 9.000000  
 9.000001 - 11.600000



**Legend**  
**Effect of AADT Change**  
 4.440000 - 5.000000  
 5.000001 - 7.000000  
 7.000001 - 9.000000  
 9.000001 - 13.580000

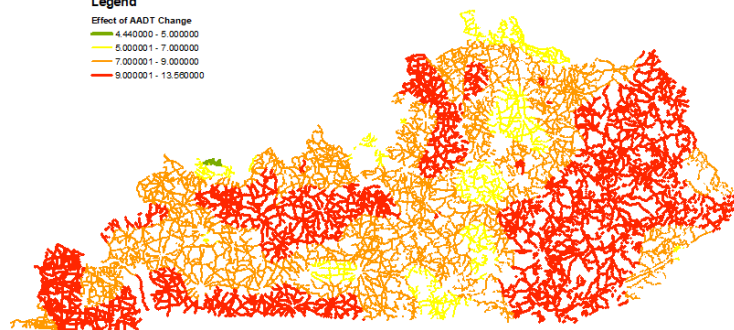


Figure 36 Spatial Distribution of the Coefficients for AADT and Effect of AADT Changes



#### 7.4.1.2 Average Speed

Figure 37 shows the distribution of the Average Speed coefficients and the percent changes in KABC and PDO crashes for a 5 mph increase in Average Speed based on GWP. Figure 37 shows that average speed is significant mainly in Eastern Kentucky in terms of KABC crashes. For PDO crashes, most of the regions are showing speed as significant, except for Western Kentucky.

In Eastern Kentucky, the Average Speed is mainly negatively associated with both KABC and PDO crashes. These are the places with poor geometric conditions. The negative effect draws attention to the geometric conditions. For example, the shoulders are mostly narrow (0-2 ft), and sharp curves are present in this area (Figure 19(f) and Figure 19(d)). The improvement measures for these areas should consider these geometric conditions to minimize both KABC and PDO crashes.

In other regions, Average Speed seems to be positively associated with KABC, but negatively with PDO crashes. However, majority of the positive effect on KABC crashes are insignificant from both GWP and GWZIP models. Segments close to the Indiana border show the positive effect of average speed on KABC and PDO as significant. The highest positive effect of the average speed on both KABC and PDO crashes indicates considering speed as an important safety measure for these segments to minimize crashes at both severity levels.

#### 7.4.1.3 Degree of Curvature

Figure 38 shows the distribution of Degree of Curvature coefficients and the percent changes in KABC and PDO crashes for a 1-degree increase in curvature based on GWP model. In Figure 38, the darker green shows the lowest effect of the degree of curvature whereas the red color shows the highest effect of the degree of curvature. For both KABC and PDO crashes, the effect is increasing from East to West. In Eastern Kentucky, Degree of Curvature has the lowest effect and is mostly insignificant for both severity levels. These unexpected results can be due to the imbalanced data of curvature (discussed in Section 6.4.1.4.1).

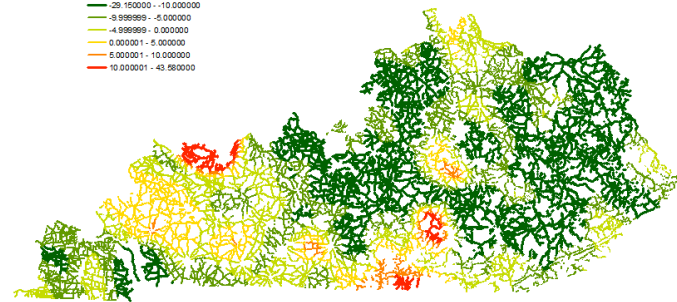
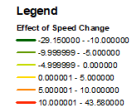
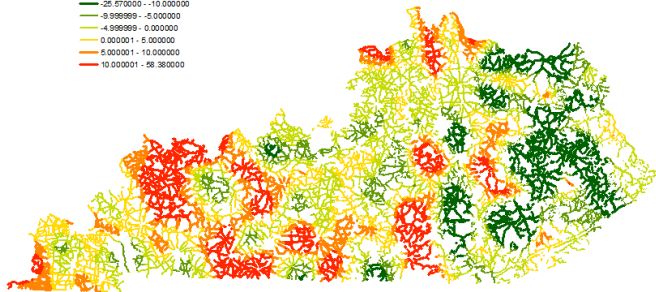
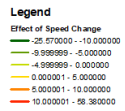
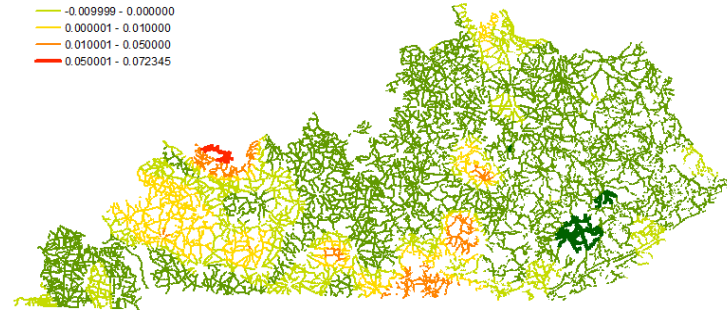
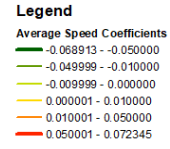
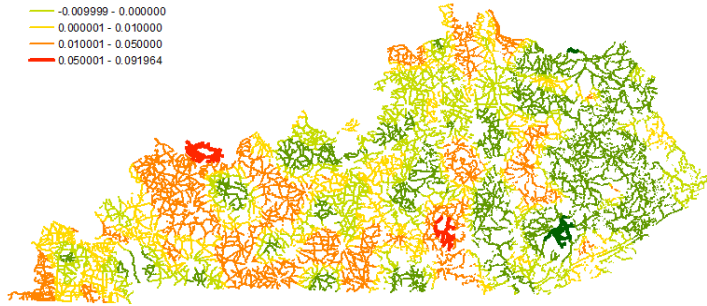
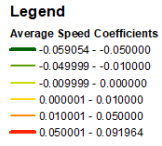
In Western Kentucky, a one-degree change in curvature has a higher effect on the KABC crashes compared to the PDO crashes. This area is mostly flat terrain with straight sections (Figure 21(b) and Figure 19(d)). Drivers do not expect to see sharp curves in this area. An increase in the curvature may make the crash more severe. To minimize the severity level in this area, curvature should be taken into consideration while applying safety measures.

For most of the Northern and Southern parts, the results show a similar case as Western Kentucky in terms of severity levels. However, for each severity level, the effect is lower in these areas than in Western Kentucky.

GWP Model

KABC

PDO



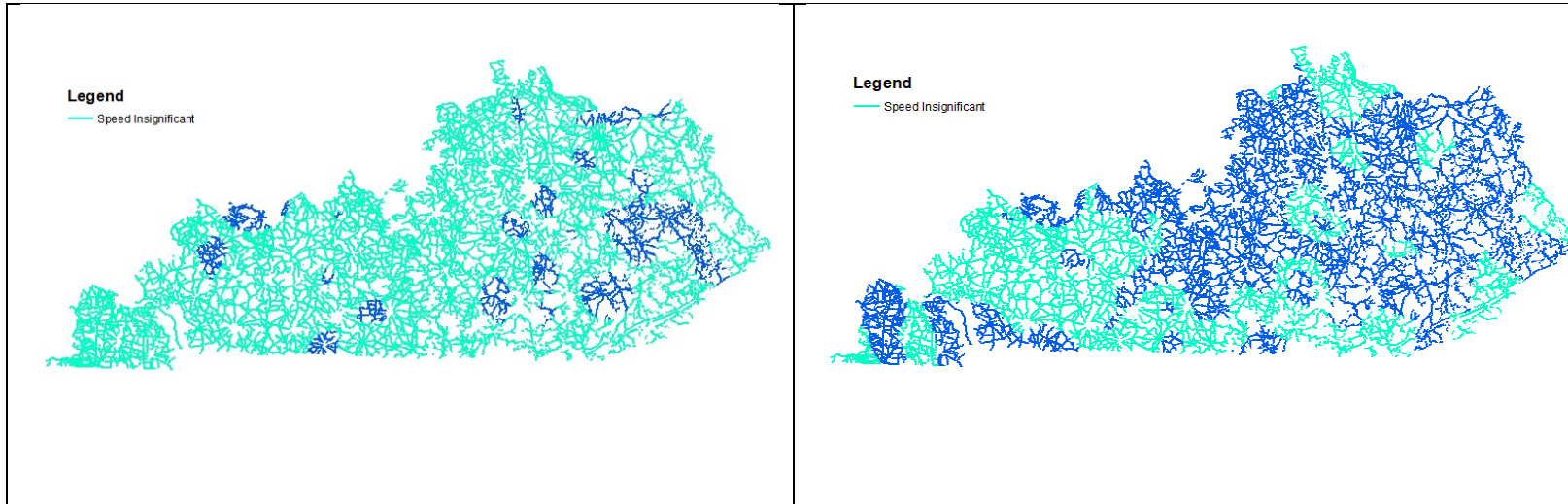
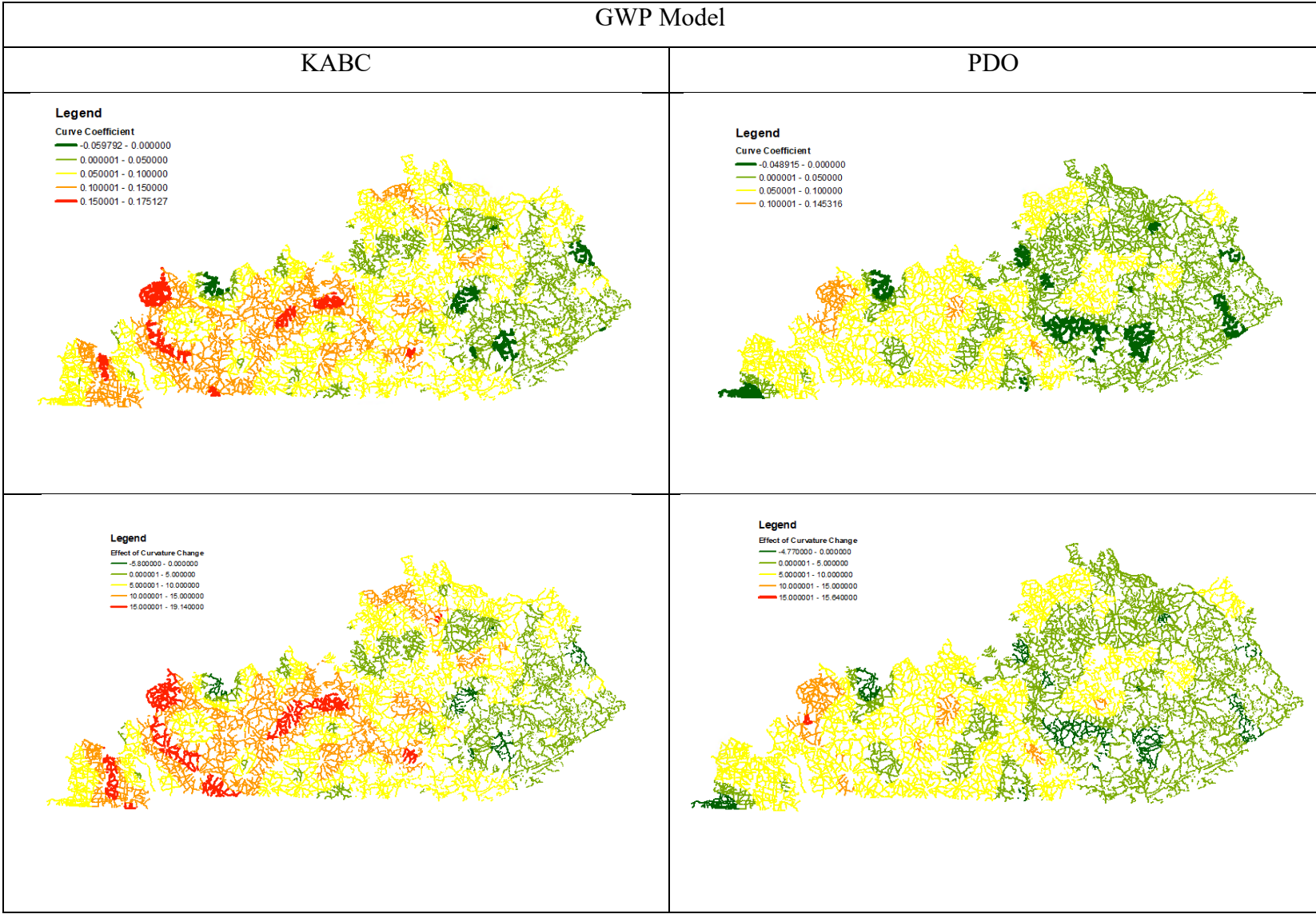


Figure 37 Spatial Distribution of the Coefficients for Average Speed and Effect of Speed Changes



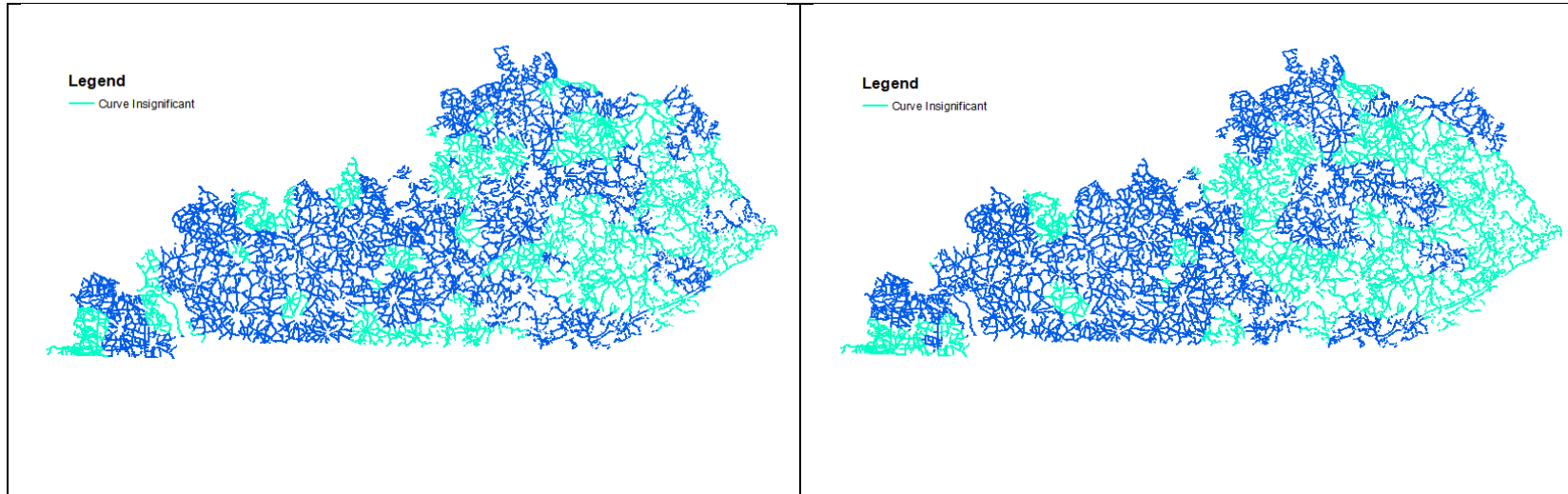


Figure 38 Spatial Distribution of the Coefficients for Degree of Curvature and Effect of Curvature Changes

The above spatial analysis reveals the spatially varying effects of AADT, Average Speed, and Degree of Curvature at different severity levels. Practitioners and local agencies can utilize such results from the spatial analysis to identify effective improvement measures and reduce the severity of crashes on a segment based on the geological location.

## 7.5 Major Findings and Significance of the Analysis

This chapter investigated the effect of speed in addition to geometric (i.e., L, Shoulder Width, Lane Width, and Degree of Curvature) and traffic volume on the KABC and PDO crashes for rural two-lane segments. The analyses were separated into traditional count models and spatial models.

Initially, separate ZIP-based models were developed to predict the number of KABC and PDO crashes by utilizing all the rural two-lane segments. The explanatory variables included AADT, L, Shoulder Width, Lane Width, Degree of Curvature, and speed measures. As the speed measure, Average Speed was identified as significant for the KABC and PDO crashes of rural two-lane highways. Here are major observations from the analysis:

- Average Speed was positively related to the KABC crashes, which is consistent with Wang et. al. (71). For PDO crashes, the association was negative consistent with Dutta and Fontaine (27). This difference in the effect of speed at different severity levels was not captured by previously developed total crash prediction models.
- Further investigation on the speed categorizer-based separate models revealed that the influence of speed can be different at different speed ranges for each severity level.
  - For low-speed roads, speed showed a positive association with KABC crashes. However, it was insignificant for the PDO crashes. These roads had poor geometric conditions (narrow shoulder or lane and presence of sharp curves). It is understandable that a crash can be severe when speed goes up under such restrictive geometric conditions of the roads.

- For medium-speed roads, speed showed a positive association for KABC crashes, whereas, a negative association for PDO crashes. The findings are consistent with the single model, i.e., Average Speed Model.
- For high-speed roads, speed showed a negative relation to both KABC and PDO crashes. These roads have better geometric standards (wider shoulder, straight sections, average lane width higher than 10 ft). The number of severe crashes tends to be low under standard geometric conditions.
- The combined performance of low, medium, and high-speed models outperformed the single model as well as the fixed proportion-based HSM model. For KABC crashes, the improvement was a maximum of 47% compared to the HSM model and 19% compared to the single model, and, for PDO crashes, there was a maximum of 22% improvement with respect to HSM and a maximum of 6.5% improvement compared to the single model.

Overall, the varying effect of speed on KABC and PDO was captured after separating the models based on speed ranges. In addition, it showed improvement over the single model as well as fixed proportion-based HSM models. This suggests developing separate models for KABC and PDO instead of applying the fixed proportions of different severity to the total number of crashes predicted by the HSM model. In addition, HSM and policymakers can adopt speed as a categorizer variable while developing models for each severity level to achieve further improvement and better assess the safety of the rural two-lane highways. Additionally, speed can be used as a surrogate for the geometric conditions of low-speed roads to take safety measures since geometric attributes may not be always available.

Later, spatial modeling approaches (GWP and GWZIP) were adopted to investigate the spatially varying effects of the explanatory variables at different levels of severity. GWP model outperformed the GWZIP model. Further analysis based on GWP model revealed some interesting localized effects of AADT, Average Speed, and Degree of Curvature on KABC and PDO crashes. These are:



- From Figure 36, Eastern Kentucky has the highest impact of AADT on PDO crashes and less impact on KABC crashes in this region. The roads in this area are mostly below standard, and the average speed is low to medium (Figure 19(c)). An increase in AADT has a high effect on the number of crashes, especially for the PDO. Western Kentucky showed a similar picture for KABC and PDO crashes in Figure 36. Even though these are high-speed roads, they seem to have less severe crashes and more PDO crashes with increasing AADT. The possible reason can be the better geometric conditions with flat terrain, wider lanes, and shoulders on these roads (Figure 21(b) and Figure 19(e)-(f)).
- For Eastern Kentucky, the Average Speed is mainly negatively associated with both KABC and PDO crashes (Figure 37). The roads in this area have poor geometric conditions. For example, the shoulders are mostly narrow (0-2 ft), and the presence of sharp curves in this area. The improvement measures for these areas should consider these geometric conditions to minimize both KABC and PDO crashes. In other regions of Kentucky, Average Speed seems to be positively associated with KABC, but negatively with PDO crashes. However, majority of the positive effects on KABC crashes are insignificant, as shown in Figure 37.
- As shown in Figure 38, the effect of curvature is increasing from East to West for both KABC and PDO crashes. In Eastern Kentucky, Degree of Curvature has the lowest effect and is mostly insignificant for both severity levels. In Western Kentucky, a one-degree change in curvature has a higher effect on the KABC crashes compared to the PDO crashes. This area is mostly flat terrain with straight sections. Drivers do not expect to see sharp curves in this area. An increase in the curvature may make the crash more severe. To minimize the severity level in this area, curvature should be taken into consideration while applying safety measures in this region. For most of the Northern and Southern parts, the results show a similar case as Western Kentucky in terms of severity levels. However, for each severity level, the effect is lower in these regions compared to Western Kentucky.

In summary, this chapter provides an understanding of the factors at different severity levels. Results from both the traditional count model and spatial model results

can help the practitioners adopt strategies for minimizing crashes, especially severe ones. Agencies can use this to evaluate alternative road designs and ensure better safety. Especially by utilizing the spatial models, they can provide localized treatment to address the severity of a crash.

## CHAPTER 8. MACHINE LEARNING MODEL-BASED ANALYSIS

Previously, this study explored traditional count models and spatial models to investigate the effect of speed on crashes. The modeling steps required different trials to come up with the final model with the significant variables. In addition, the models are susceptible to multicollinearity issues, and a presumption on the model form was required. To address such issues, this chapter explores an RF-based machine learning model for predicting crashes by incorporating speed as one of the factors.

### 8.1 Objectives

This analysis adopts RF based modeling technique to develop a crash prediction model for rural two-lane highways by incorporating speed along with traffic and geometric attributes. Below are the objectives:

- Investigate the effect and importance of speed measures in crash prediction of rural two-lane highways based on RF model.
- Compare the performance of the RF model with the previously experimented ZINB model for total number of crashes.

While little has been done to investigate the effect of speed on the crashes of rural two-lane highways based on ML models, this analysis attempts to fill that gap (97).

### 8.2 Dataset and Variables

The dataset (Section 6.2) used for the analysis in Chapter 6 was utilized for this analysis. The dataset contains 53,208 segments with a total of 65,091 crashes aggregated from both directions of the road. Table 36 presents the statistics of geometric, traffic, and speed attributes on these segments. L, Degree of Curvature, Lane Width, and Shoulder Width represent the geometrics of the roads, whereas, Average Speed, Speed Limit, the 85<sup>th</sup> Percentile speed, and Std of speed represent the speed attributes on these roads. In addition, the table shows the summary statistics for the total number of crashes in 5 years.

Table 36 Summary Statistics of the Road Attributes

Variables	Unit	Statistics			
		Min.	Max.	Mean	Standard Deviation
AADT	vehicle	2	19619	1355	1772
Segment Length (L)	mile	0.10	2.97	0.26	0.21
Degree of Curvature ( $Cu$ )	degrees	0	63.81	2.42	3.80
Lane Width (LW)	ft	6	18	9.42	1.14
Shoulder Width (SW)	ft	0	14	3.51	2.06
Average Speed ( $V_a$ )	mph	5.37	61.76	39.92	9.87
Speed Limit ( $V_{sp}$ )	mph	15	55	53.89	4.22
Standard Deviance (Std) of Speed	mph	6.20	36.74	16.42	4.30
The 85 <sup>th</sup> Percentile Speed ( $V_{85}$ )	mph	12.88	67.33	48.84	8.02
Number of Crashes in 5 years		0	161	1.22	2.85

### 8.3 Analysis and Results

The “RandomForestRegressor” package in Python was used to develop RF regression model. The geometric, traffic, and speed variables listed in Table 36 were used as the input variables and total number of crashes in 5 years as the output. Note that, in the previous analysis based on count models, this study could not include all the speed measures (the 85<sup>th</sup> Percentile Speed, Average Speed, etc.) in the same model due to high multicollinearity among the speed measures although the measures were found to be significant in separate models. Since the RF model can handle the multicollinearity among the explanatory variables, this analysis included all the speed variables to

investigate the effect of all the speed measures in addition to other factors in the same model.

For the model calibration, 70% of the dataset was used as the training set and the rest as the testing dataset. Following the calibration process described in Section 4.4.3.1, the best combination of hyperparameters shown in Table 37 was estimated. The RF model was built using these hyperparameters.

Table 37 Best Combination of Hyperparameters

<b>Hyperparameters</b>	<b>Optimum Value</b>
n_estimators	10,000
max_features	$p$
max_depth	10
min_samples_leaf	4
min_sample_split	2

Based on the RF model, the VI of geometric, traffic, and speed variables were determined. Among these variables, Speed Limit has the lowest contribution (0.31% as VI) to the model outcome. Since it does not contribute much to model prediction, it was excluded from the final model. Table 38 presents the final set of variables with their rankings. It turns out that AADT and L are the top two variables in the list, which is not surprising since these are the exposure variables. The third variable is Shoulder Width. Speed measures such as the 85<sup>th</sup> Percentile Speed and Average Speed were found as the fourth and fifth variables, respectively. Their total contribution is 11.5% in the model. The rest of the variables (Degree of Curvature, Std of speed, Lane Width) seem to have low importance in the model. It appears that speed measures especially, the 85<sup>th</sup> Percentile Speed and Average Speed are more important than some of these geometric features.

Table 38 Ranking of the Variables

Variables	VI (%)	Rank
AADT	44.1	1
Length	26.6	2
Shoulder Width	11.9	3
The 85 <sup>th</sup> Percentile Speed	8.9	4
Average Speed	2.6	5
Degree of Curvature	2.1	6
Std of Speed	2.0	7
Lane Width	1.7	8

To further assess the influence of each variable on the number of crashes based on RF model, this study looked at the partial dependence plots (PDPs) of the variables introduced by Friedman (143). These plots help to reveal the functional relationship between the explanatory and response variables and show the marginal effect of individual variables on the response. The interpretation of these plots is similar to the coefficients provided by the traditional statistical models. They can be utilized to find out whether the relationship between an explanatory variable and a response is linear or non-linear.

PDPs provide the causal effect of individual variables assuming that each variable is independent. PDPs calculate the average marginal effect corresponding to the given values of a target variable while keeping the actual values for other variables. The mathematical function for estimating partial dependence for a target variable using the training dataset is as follows (143):

$$f_t(X_t) = \frac{1}{N} \sum_{i=1}^N f(X_t, X_o^{(i)}) \quad (56)$$

Where,

$f_t$  = partial dependence function for a target variable

$X_t$  = the target variable for which the PDF is plotted

$N$  = number of observations in the training dataset

$f$  = RF model

$X_o^{(i)}$  = actual values of the other variables

Figure 39 displays the relationship between the number of crashes and the explanatory variables using PDPs. It looks like most of the factors have a non-linear effect on the number of crashes except for Lane Width, Shoulder Width, and Std of speed. For AADT, the predicted number of crashes increases with an increase in AADT at an exponential rate. However, the influence of AADT is lower after an AADT of around 6000. This kind of fluctuation in the effect of AADT on the number of crashes is also observed in an existing study by Saha et al. (101). The effect of Length seems to be almost linear based on the PDP which is expected. For the 85<sup>th</sup> Percentile Speed and Average Speed, the trend is downward with nearly a non-linear relationship. This negative relationship is generally consistent with existing studies (10; 27; 28; 30; 48). Based on the study data, the rural two-lane highways with higher speeds tend to be the corridors with better geometric conditions. For the Degree of Curvature, the association is positive from the PDP and it seems to show a slight jump after 7.5 degrees. This implies that number of crashers is more influenced by the Degree of Curvature if it is between Class D and Class F. For Shoulder Width, the trend seems to be flat indicating no significant effect on the number of crashes. This somehow contradicts the ranking of this variable as shown in Table 38. This difference can be due to assuming the effect of the variable as independent and ignoring interactions with other features while calculating the average predictions for PDP (144). If the variable is not correlated with other explanatory variables, PDP provides a better interpretation of the effect of the variables. However, in this study, Shoulder Width tends to be correlated with the AADT and speed measures. For Lane Width and Std of speed, the effect is almost flat indicating no substantial effect on the number of crashes. These are consistent with their rankings from Table 38.

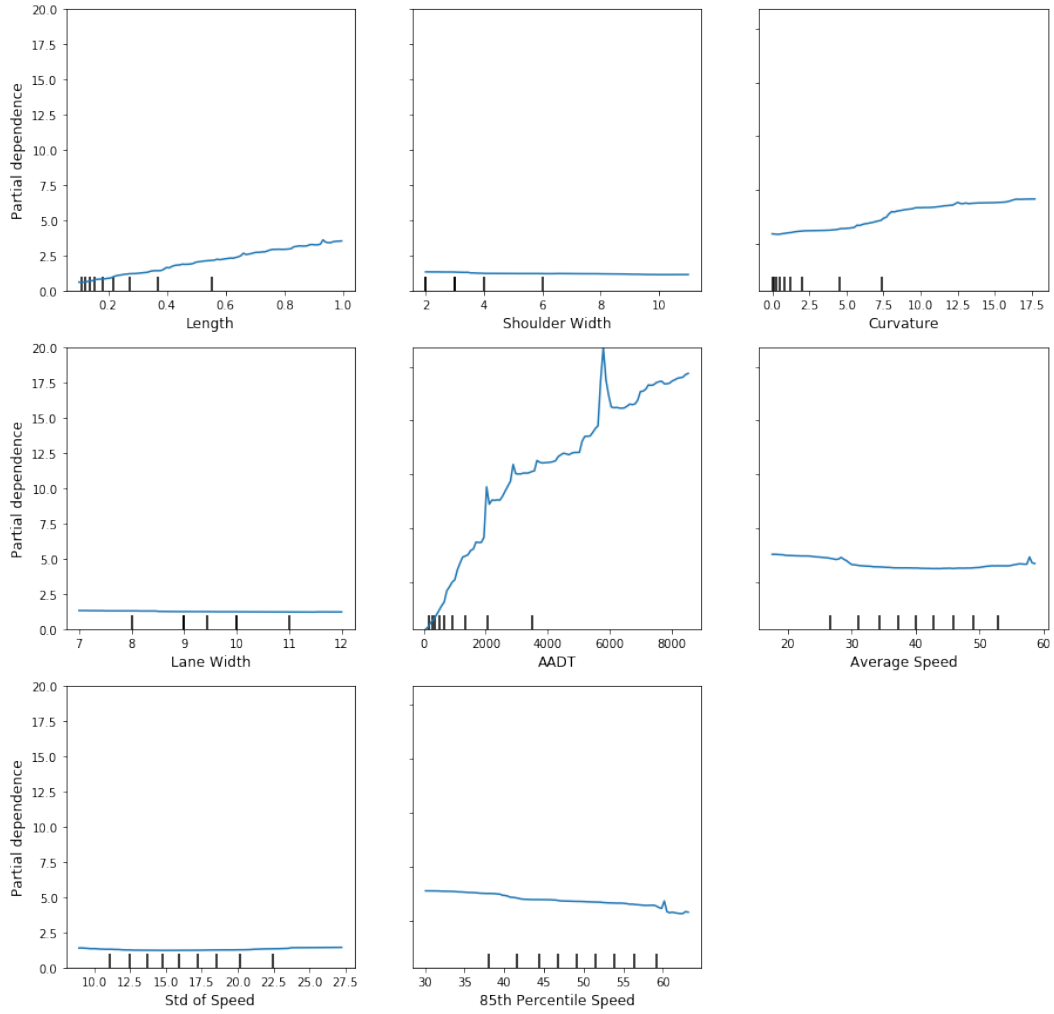


Figure 39 Partial Dependence Plots

Further, this analysis evaluated the performance of the RF model in predicting total number of crashes. The performance was also compared with the ZINB model. As the performance measures,  $R^2$ , MAPE, RMSE, and MAD were utilized. Table 39 presents the performance measures for the RF model developed with the eight variables listed in Table 38. The performances are close between the training and testing data implying no significant overfitting or underfitting by the trained model. Results from the RF model and ZINB model comparison are discussed in the sub-section below.



Table 39 Performance of the RF Model

<b>Measures</b>	<b>Training Set</b>	<b>Testing Set</b>
$R^2$	0.57	0.40
<i>MAPE</i>	54.95%	61.85%
<i>RMSE</i>	1.89	2.01
<i>MAD</i>	0.88	0.96

### 8.3.1 Model Comparison

To test how the RF-based crash prediction model performs compared to the traditional model (i.e., ZINB), this analysis developed an RF model with AADT, L, Average Speed, and Degree of Curvature. Since ZINB model could only include these variables after accounting for multicollinearity, the RF model considered the same set of variables to directly compare with the ZINB. Overall, the RF outperformed the ZINB model, as shown in Table 40. Especially in case of testing data, the maximum improvement is around 13%. In addition, Figure 40 shows a comparison between the predicted and observed number of crashes for each model. Compared to ZINB, more data are in the diagonal line for RF model clearly indicating better prediction performance by RF.

Table 40 Comparison of Model Performance

<b>Measures</b>	<b>RF Model</b>		<b>ZINB Model</b>	
	<b>Training Set</b>	<b>Testing Set</b>	<b>Training Set</b>	<b>Testing Set</b>
$R^2$	0.54	0.36	0.27	0.32
<i>MAPE</i>	56.01%	62.53%	63.38	63.88
<i>RMSE</i>	1.97	2.07	2.47	2.13
<i>MAD</i>	0.90	0.98	1.04	1.02

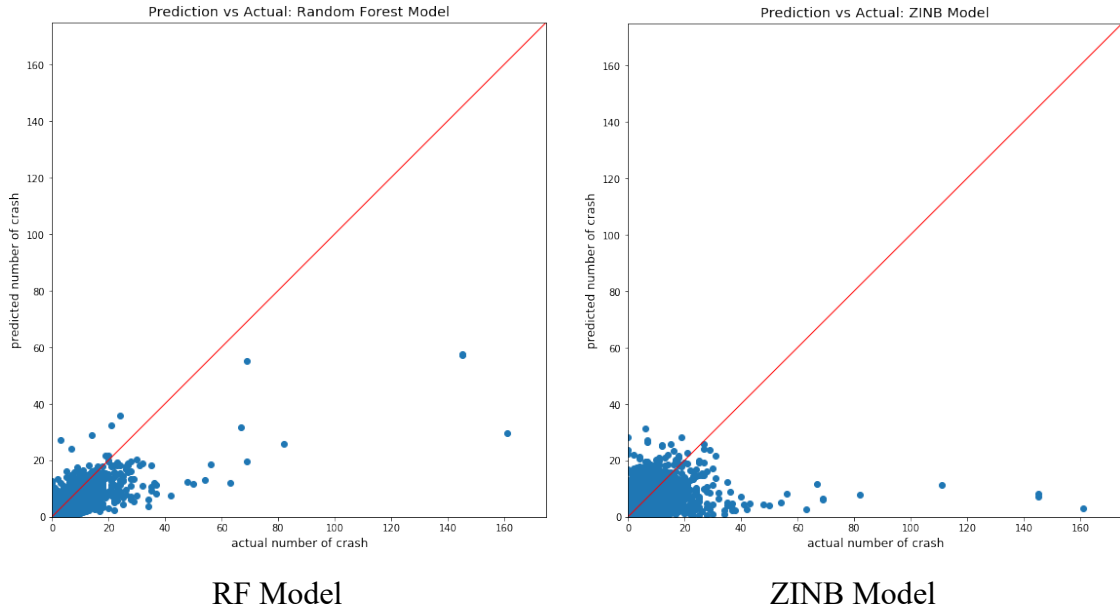
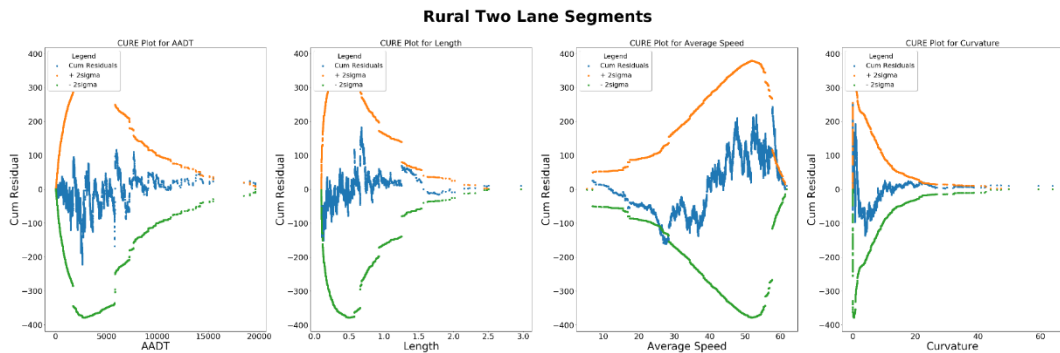


Figure 40 Comparison between predicted and actual number of crashes

To further compare the model fits between RF and ZINB models, this analysis also developed CURE plots, as presented in Figure 41. Clearly, RF model fits the data significantly better than the ZINB model considering each of the four variables. While ZINB requires separate models based on speed and AADT ranges, RF model seems to perform well without stratifying the data.



(a) RF Model

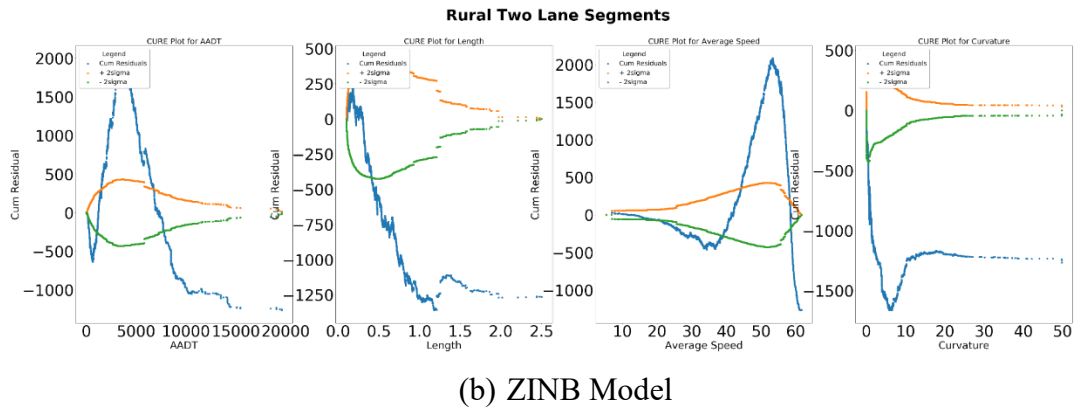


Figure 41 Comparison of CURE Plots

In addition, this analysis compared the ranking of Average Speed between each model. Table 41 displays the variable ranking based on VI for each model. In RF model, Average Speed is the third variable, whereas, it is the least important variable in ZINB model. As a non-parametric model, RF method can capture the variability in the data better than the ZINB model, therefore, the importance of speed is better assessed through RF model.

Table 41 Variable Importance from RF and ZINB model

Variables	RF Model		ZINB Model	
	VI (%)	Rank	VI (%)	Rank
AADT	52.2	1	50.5%	1
Length	31.6	2	26.1%	2
Average Speed	12.5	3	2.1%	4
Degree of Curvature	3.7	4	21.2%	5

#### 8.4 Findings and Significance of the Analysis

This chapter experimented with the RF model as one of the data mining techniques since it can deal with multicollinearity between explanatory variables, therefore, allowing for additional variables such as Shoulder Width, Lane Width,

different speed measures along with Average Speed. The model revealed the importance of each variable in the crash prediction of rural two-lane highways. In case of speed measures, the 85<sup>th</sup> percentile speed and average speed turned out to be the fourth and fifth most important variables. From the ranking of these speed measures, speed seemed to be more important than some of the geometric factors such as Lane Width and Degree of Curvature for the rural two-lane highways used in this study. Further comparison with the ZINB model, it was found that RF model significantly performs better than the ZINB model and does not require further splitting of the dataset based on speed or AADT (Table 40 and Figure 41).

While traditional statistical models have advantages like ease of better transformability and applicability, machine learning models like RF model can provide better accuracy in crash prediction. Not only improved predictions but also the influence of the variables can be directly assessed without any assumptions on the functional form. However, the application of this model can be limited by computational cost in terms of hardware requirements or computational time. In that case, practitioners can adopt this method to identify the most important variables and use those variables in developing statistical models for crash prediction of rural two-lane highways.

## CHAPTER 9. CONCLUSION

The purpose of this study was to evaluate the effect of speed on the crashes of rural two-lane highways by utilizing the measured speed dataset. To incorporate the speed in the crash prediction model, different modeling techniques such as traditional count model, spatial model, and machine learning model were adopted. Analyzing the results from the models identified a varying effect of speed at different speed categories of the roads and different locations of the state. Capturing such varying effects of speed provided improvement in the model performance. The results based on the models from this study can be utilized while adopting safety countermeasures and improvement strategies for rural two-lane highways.

This chapter summarizes the research with major findings and provides recommendations for future work.

### 9.1 Summary

Due to lack of measured data, past research was limited for exploring the effect of speed on the crashes of rural two-lane highways. With the advancements in GPS technologies, speed data availability has become better than before on these roads. This study utilized such dataset to estimate different speed measures and incorporated them into the crash prediction model for rural two-lane highways. The primary goal is to investigate the role of speed on the crashes of these roads utilizing measured data.

#### Investigating Significance of Speed:

At first, this study explored the effect of speed from an operational perspective. The ZINB-based model was adopted for this analysis since the zero inflation models can address the overdispersion due to the presence of excess zero crashes. This study incorporated different speed measures (Average Speed, the 85<sup>th</sup> Percentile Speed, etc) along with AADT and length in the model. In each model speed measures were found significant. Average Speed was chosen for additional analysis since it better represents the operating condition of these roads. A varying effect of speed was observed from low-speed to high-speed roads when separate models were developed for each speed range.

This implies that speed has a subgroup effect on the crashes of rural two-lane highways. To capture that, developing separate SPFs based on the speed of these roads can be considered. Even though including the speed variable in the model may not always add a significant improvement in the prediction performance, considering the speed during splitting the data for developing separate models can improve the overall performance by 11.3%. For the safety assessment rural two-lane roads, DOTs and agencies can adopt such an approach of separating the model for different speed ranges.

This study further incorporated speed differential between consecutive segments in predicting crashes of rural two-lane highways. The analysis showed that more crashes tend to occur when the 85<sup>th</sup> percentile speed differential between consecutive segments increases. However, the application of speed differential-based model to identify hot spots revealed that the higher crash location in this study may not be always involved with speed inconsistency. Therefore, speed differential may not be a suitable factor of predicting crashes in this study, rather it can be useful to take measures for further design improvement of the roads.

#### Spatial Varying Effect of the Factors on Crashes:

Traditional count models (such as ZINB) assume a stationary pattern of the variables over the spatial domain. Such models estimate single coefficient values as the average effect of the variables on crashes. However, the effect may show spatial heterogeneity considering the spatial dependency of crashes and the road attributes. This study incorporated such spatial dependency utilizing spatial models like GWP and GWZIP and investigated the spatially varying effect of speed in addition to other factors for the rural two-lane segments. Both GWZIP and GWP models outperformed global models (i.e., Poisson and ZIP) by a maximum of 35.9% and 32% improvement, respectively. For further analysis, GWZIP was selected as it showed slightly better performance. The results from this model helped to diagnose the localized influence of the predictor variables. Some of the interesting findings from the analysis can be listed below:

- The analysis showed a spatial pattern of the significance of speed. Its significance varied at different locations, which was not observed in the global model.
- Speed was found significant for most of Eastern and Northern Kentucky. These are the roads with poor geometric conditions. To further enhance safety in those areas, measures can be taken to improve road geometrics.
- Some areas (Figure 24(a)) in Western Kentucky showed that speed affects the crashes positively and speed was the top-ranked variable. Considering the standard geometric conditions and low traffic on those roads, speed is clearly the main factor if there are any crashes.

While traditional models identify the same factors as significant over the state, the spatial models can be adopted to diagnose local factors that can be significant for one region but may not be in another region in the same jurisdiction. Such results can be used to prioritize the important local factors of crashes for a road in a certain area. The most important variable in that area can be utilized to plan an efficient improvement strategy. In addition, this analysis provided local models for each road. Practitioners can utilize the model of a certain road for analyzing the safety performance of a new road within its close proximity.

#### Effect of Speed at Different Levels of Crash Severity

Further, this study investigated the effect of speed in addition to geometric and traffic factors on the KABC and PDO crashes for rural two-lane segments. The analyses were separated into traditional count and spatial modeling. Similar to the models for total number of crashes, this analysis also revealed the subgroup effect of speed in developing models for KABC and PDO crashes. Based on the speed ranges, a varying effect of speed was found for the KABC and PDO crashes. Therefore, models were separated for low, medium, and high-speed roads. The key findings can be listed below.

- For low-speed roads, crashes can be severe when speed goes up under poor geometric conditions. Speed seemed to be a better surrogate of the geometrics of these roads compared to the medium speed road.

- The high-speed roads have better geometric standards (wider shoulder, straight sections, average lane width higher than 10 ft). The number of severe crashes tends to be low under standard geometric conditions.

Overall, the varying effects of speed on KABC and PDO were captured after separating the models based on speed ranges. In addition, it showed improvement over the single model as well as fixed proportion-based HSM models. For KABC crashes, the improvement was a maximum of 47% compared to the HSM model and 19% compared to the single model, and, for PDO crashes, there was a maximum of 22% improvement with respect to HSM and a maximum of 6.5% improvement compared to the single model. These suggest developing separate models for KABC and PDO instead of applying the fixed proportions of different severity to the total number of crashes predicted by HSM model. In addition, HSM and policymakers can adopt speed as a categorizer variable while developing models for each severity level to achieve further improvement and better assess the safety of the rural two-lane highways. Moreover, speed can be used as a surrogate for the geometric conditions of low-speed roads to take safety measures since geometric attributes may not be always available.

Later, spatial modeling approaches (GWP and GWZIP) were adopted to investigate the spatially varying effects of the explanatory variables at different levels of severity. The analysis based on the GWP model revealed some interesting localized effects of the factors on KABC and PDO crashes. These are:

- AADT had a higher impact on PDO crashes than KABC mostly in Eastern and Western Kentucky regions.
- Both KABC and PDO crashes seemed to be influenced by the low speed of the roads in a region. This mainly draws attention to the geometric condition in that area. The improvement measures for such areas should consider geometric conditions.
- Degree of curvature had a higher effect in areas with flat terrain with straight sections. Drivers do not expect to see sharp curves in this area. An increase in the curvature may make the crash more severe. To minimize the severity level in this



area, curvature should be taken into consideration while applying safety measures in this region.

The analysis of different severity levels provides an understanding of the factors at different severity levels. Both the traditional count and spatial modeling results can help practitioners adopt strategies for minimizing crashes, especially severe ones. Agencies can use this to evaluate alternative road designs and ensure better safety. Especially by utilizing the spatial models, they can provide localized treatment to address the severity of a crash.

#### Machine Learning Model-based Analysis

This study also experimented with the RF model as one of the data mining techniques since it can deal with multicollinearity between explanatory variables and requires no presumption on the functional form. The model revealed the importance of each variable in the crash prediction of rural two-lane highways. In case of speed measures, the 85<sup>th</sup> percentile speed and average speed turned out to be the fourth and fifth most important variables. From the ranking of these speed measures, speed seemed to be more important than some of the geometric factors such as Lane Width and Degree of Curvature for the rural two-lane highways used in this study. Further comparison with the ZINB model, it was found that RF model significantly performs better than the ZINB model and does not require further splitting of the dataset based on speed or AADT.

While traditional statistical models have advantages like ease of better transformability and applicability, machine learning models like RF model can provide better accuracy in crash prediction. Not only improved predictions but also the influence of the variables can be directly assessed without any assumptions on the functional form. However, the application of this model can be limited by computational cost in terms of hardware requirements or computational time. In that case, practitioners can adopt this method to identify the most important variables and use those variables in developing statistical models for crash prediction on rural two-lane highways.

## 9.2 Study Limitations and Future Work

This study was limited to the crash data aggregated for 5 years and speed data aggregated for 3 years. As future work, the study can collect more data disaggregated by year and with better coverage. The data can be used for highway project improvement. For example, the data can help to identify certain improvements for a road, and then the safety benefit can be quantified with those improvements. Furthermore, future work can utilize machine learning-based techniques to develop models for KABC and PDO crashes on these roads. Additional variables like functional class, median type, access control type, and surface condition can be collected and the influence of these variables can be assessed in addition to the speed variables for rural two-lane highways.

The different analyses conducted in this study had some limitations, especially considering the data issues. Future work can also look into the data issue and revisit the models. For example:

- Analysis in Section 5.1 is limited due to the HIS database and speed dataset. Even though the 85<sup>th</sup> percentile speed-based model in Table 7 was the best model considering the predictive performance, this analysis did not select it as calculating the 85<sup>th</sup> percentile speed requires a large amount of dataset. This study will further look at the 85<sup>th</sup> percentile speed model when more speed data become available in the future. Moreover, the dataset contained some low functional class roads with lower average speeds although the speed limits from HIS database were 55 mph. It requires further verification of the HIS database and revisiting the models. In addition, some of the average speeds of the roads seemed to be affected by conflation issue of the speed network. In future, this type of issue will be further investigated to see how it affects the accuracy of the crash prediction models.
- The analysis in Section 5.2 has multiple limitations in terms of datasets. It could not explore the effect of Speed Differential for each direction of the road since the crash dataset came into an aggregated format regardless of the directions. If directional crash data can be collected, the analysis can be revised further. Moreover, 92% of the data was from curve Class A and majority of the segments

had a good design. This requires further looking into the analysis if more data for other curve classes are available.

- The aggregation process of the segments based on the same curvature class may affect the analysis in Chapter 6 to Chapter 8, especially while looking into the effect of the curvature on crashes. For future analysis, this study will include a more precise measurement of the curvature before developing models.

## APPENDIX

### Curve Class (Source: HPMS Field Manual)

Curve Class	Degree of Curvature Range
A	<3.5
B	3.5 – 5.4
C	5.4 – 8.4
D	8.5 – 13.9
E	14 – 27.9
F	$\geq 28$

## REFERENCES

- [1] Amaliana, L., and A. Fernandes. Comparison of Two Weighting Functions in Geographically Weighted Zero-Inflated Poisson Regression on Filariasis Data. In *Journal of Physics: Conference Series*, No. 1097, IOP Publishing, 2018. p. 012070.
- [2] 2010 Census urban and rural classification and urban area criteria. <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html>.2022.
- [3] *Assessing and managing the ecological impacts of paved roads, Ch. 2 History and status of the US road system*. National Research Council, 2005.
- [4] Agent, K. R., J. G. Pigman, and N. Stamatiadis. Countermeasures for fatal crashes on two-lane rural roads. In, Kentucky Transportation Cabinet, 2001. p. 55.
- [5] Glennon, J. C., T. R. Neuman, and J. E. Leisch. *Safety and operational considerations for design of rural highway curves*. Leisch (Jack E.) and Associates, 1985.
- [6] Solomon, D. Accidents on main rural highways related to speed, driver, and vehicle. In, U.S. Department of Commerce, 1964. p. 50.
- [7] *Highway Safety Manual Part C*. AASHTO, 2010.
- [8] Cooper, P. J. The relationship between speeding behaviour (as measured by violation convictions) and crash involvement. *Journal of Safety Research*, Vol. 28, No. 2, 1997, pp. 83-95.
- [9] Fildes, B., G. Rumbold, and A. Leening. Speed Behavior and Drivers' Attitude to Speeding (Report 16). *Monash University Accident Research Center, Monash, Victoria, Australia, June, 1991*.
- [10] Garber, N. J., and R. Gadiraju. Factors affecting speed variance and its influence on accidents. *Transportation Research Record*, Vol. 1213, 1989, pp. 64-71.
- [11] Kloeden, C., A. McLean, V. Moore, and G. Ponte. Traveling Speed and the Risk of Crash Involvement. In, No. 1, NHMRC Road Accident Research Unit. The University of Adelaide, 1997. p. 69.
- [12] Kockelman, K. M., and J. Ma. Freeway speeds and speed variations preceding crashes, within and across lanes. In *Journal of the Transportation Research Forum*, No. 46, 2007. p. 43.
- [13] Hossain, F., and J. Medina. Effects of Operating Speed and Traffic Flow on Severe and Fatal Crashes using usRAP. In *the 99th Annual Meeting of the Transportation Research Board*, Washington D.C., 2020. p. 18.
- [14] Hutton, J., D. Cook, J. Grotheer, and M. Conn. Evaluation of the Relationship Between Driving Speed and Crashes on Urban and Suburban Arterials In *99th Annual Meeting of the Transportation Research Board*, Washington D.C., 2020.
- [15] Kloeden, C. N., J. McLean, and G. F. V. Glonek. Reanalysis of travelling speed and the risk of crash involvement in Adelaide South Australia. In, Australian Transport Safety Bureau, 2002. p. 40.
- [16] Kononov, J., C. Lyon, and B. K. Allery. Relation of flow, speed, and density of urban freeways to functional form of a safety performance function. *Transportation Research Record*, Vol. 2236, No. 1, 2011, pp. 11-19.
- [17] Kweon, Y. J., and K. M. Kockelman. Safety effects of speed limit changes: Use of panel models, including speed, use, and design variables. *Transportation Research Record*, Vol. 1908, No. 1, 2005, pp. 148-158.

- [18] Taylor, M. C., D. Lynam, and A. Baruya. The effects of drivers' speed on the frequency of road accidents. In, Transport Research Laboratory Crowthorne, 2000. p. 50.
- [19] Banihashemi, M., M. Dimaiuta, A. Zineddin, B. D. Spear, O. Smadi, and Z. Hans. Using linked SHRP2 RID and NPMRDS data to study speed- safety relationships on urban interstates and major arterials. In *98th Annual Meeting of the Transportation Research Board*, Washington DC, USA, 2019. p. 21.
- [20] Wang, X., Q. Zhou, M. Quddus, and T. Fan. Speed, speed variation and crash relationships for urban arterials. *Accident Analysis & Prevention*, Vol. 113, 2018, pp. 236-243.
- [21] Anderson, I. B., and R. A. Krammes. Speed reduction as a surrogate for accident experience at horizontal curves on rural two-lane highways. *Transportation Research Record*, Vol. 1701, No. 1, 2000, pp. 86-94.
- [22] Cafiso, S., A. Di Graziano, G. Di Silvestro, G. La Cava, and B. Persaud. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *Accident Analysis & Prevention*, Vol. 42, No. 4, 2010, pp. 1072-1079.
- [23] De Oña, J., and L. Garach. Accidents prediction model based on speed reduction on Spanish two-lane rural highways. *Procedia-Social and Behavioral Sciences*, Vol. 53, 2012, pp. 1010-1018.
- [24] Llopis-Castelló, D., F. Bella, F. J. Camacho-Torregrosa, and A. García. New consistency model based on inertial operating speed profiles for road safety evaluation. *Journal of Transportation Engineering, Part A: Systems*, Vol. 144, No. 4, 2018, p. 10.
- [25] Ng, J. C. W. Quantifying the relationship between geometric design consistency and road safety. In *Department of Civil Engineering*, University of British Columbia, 2002. p. 103.
- [26] Boonsiripant, S. Speed profile variation as a surrogate measure of road safety based on GPS-equipped vehicle data. In *Civil Engineering*, Georgia Institute of Technology, 2009. p. 297.
- [27] Dutta, N., and M. D. Fontaine. Improving freeway segment crash prediction models by including disaggregate speed data from different sources. *Accident Analysis & Prevention*, Vol. 132, 2019, p. 16.
- [28] Pei, X., S. Wong, and N.-N. Sze. The roles of exposure and speed in road safety analysis. *Accident Analysis & Prevention*, Vol. 48, 2012, pp. 464-471.
- [29] Ederer, D., M. Rodgers, M. Hunter, and K. Watkins. Probe-speed based safety performance metrics in Georgia: A case study. In *the 99th Annual Meeting of the Transportation Research Board*, Washington D.C., 2020. p. 16.
- [30] Stipancic, J., E. B. Racine, A. Labbe, N. Saunier, and L. Miranda-Moreno. Relating traffic flow to crashes using Massive GPS Data: Smartphones and usage-based insurance data agree In *the 99th Annual Meeting of the Transportation Research Board*, Washington D.C., 2020.
- [31] Das, S., S. Geedipally, R. Avelar, L. Wu, K. Fitzpatrick, M. Banihashemi, and D. Lord. Rural speed safety project for USDOT safety data initiative. In, Texas A & M Transportation Institute 2020. p. 122.
- [32] Kockelman, K. M., and Y.-J. Kweon. Driver injury severity: an application of ordered probit models. *Accident Analysis Prevention*, Vol. 34, No. 3, 2002, pp. 313-321.

- [33] Quddus, M. Exploring the relationship between average speed, speed variation, and accident rates using spatial statistical models and GIS. *Journal of Transportation Safety & Security*, Vol. 5, No. 1, 2013, pp. 27-45.
- [34] Wang, C., M. A. Quddus, and S. G. Ison. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis & Prevention*, Vol. 43, No. 6, 2011, pp. 1979-1990.
- [35] Kweon, Y.-J., and K. M. Kockelman. Safety effects of speed limit changes: Use of panel models, including speed, use, and design variables. *Transportation Research Record*, Vol. 1908, No. 1, 2005, pp. 148-158.
- [36] Lave, C. A. Speeding, coordination, and the 55 mph limit. *The American Economic Review*, Vol. 75, No. 5, 1985, pp. 1159-1164.
- [37] Wang, K., J. N. Ivan, N. Ravishanker, and E. Jackson. Multivariate poisson lognormal modeling of crashes by type and severity on rural two lane highways. *Accident Analysis Prevention*, Vol. 99, 2017, pp. 6-19.
- [38] Ma, J., K. M. Kockelman, and P. Damien. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis Prevention*, Vol. 40, No. 3, 2008, pp. 964-975.
- [39] Stapleton, S. Y., A. J. Ingle, M. Chakraborty, T. J. Gates, and P. Savolainen. Safety performance functions for rural two-lane county road segments. *Transportation Research Record*, Vol. 2672, No. 52, 2018, pp. 226-237.
- [40] West, R., D. French, R. Kemp, and J. Elander. Direct observation of driving, self reports of driver behaviour, and accident involvement. *Ergonomics*, Vol. 36, No. 5, 1993, pp. 557-567.
- [41] Maycock, G., P. Brocklebank, and R. Hall. Road layout design standards and driver behaviour. In, 1998. p. 49.
- [42] Richards, D., R. Cuerden, and G. Britain. *The relationship between speed and car driver injury severity*. Department for Transport London, 2009.
- [43] Anderson, I., K. Bauer, D. Harwood, and K. Fitzpatrick. Relationship to safety of geometric design consistency measures for rural two-lane highways. *Transportation Research Record*, No. 1658, 1999, pp. 43-51.
- [44] Wu, K.-F., E. T. Donnell, S. C. Himes, and L. Sasidharan. Exploring the association between traffic safety and geometric design consistency based on vehicle speed metrics. *Journal of Transportation Engineering, Part A: Systems*, Vol. 139, No. 7, 2013, pp. 738-748.
- [45] Nancy Dutta, and M. D. Fontaine. DEVELOPING RURAL FOUR LANE FREEWAY CRASH PREDICTION MODELS USING HOURLY FLOW PARAMETERS. *TRB 2019 Annual Meeting*, 2019.
- [46] Stout, T. B. Speed metrics and crash risks: statistical assessment and implications for highway safety policy. In *Civil, Construction, and Environmental Engineering*, No. *Doctor of Philosophy*, Iowa State University, 2005. p. 171.
- [47] Finch, D., P. Kompfner, C. Lockwood, and G. Maycock. Speed, speed limits and crashes. In, No. 58, 1994.
- [48] Baruya, A. Speed-accident relationships on European roads. In *9th International Conference on Road Safety in Europe*, 1998. pp. 1-19.

- [49] Wang, X., T. Fan, M. Chen, B. Deng, B. Wu, and P. Tremont. Safety modeling of urban arterials in Shanghai, China. *Accident Analysis & Prevention*, Vol. 83, 2015, pp. 57-66.
- [50] Najjar, Y. M., and S. Mandavilli. Data mining the Kansas traffic-crash database: summary. In, Kansas. Dept. of Transportation. Bureau of Materials & Research, 2009.
- [51] Elvik, R. *The Power Model of the relationship between speed and road safety: update and new analyses*. 2009.
- [52] Stipanovic, J., E. Racine, A. Labbe, N. Saunier, and L. Miranda-Moreno. Relating Traffic Flow to Crashes Using Massive GPS Data: Smartphones and Usage-Based Insurance Data Agree. In *The 99th Annual Meeting of Transportation Research Board* Washington D. C., 2020.
- [53] Llopis-Castelló, D., D. J. Findley, and A. Garcia. Comparison of the highway safety manual predictive method with safety performance functions based on geometric design consistency. *Journal of Transportation Safety Security*, Vol. 13, No. 12, 2021, pp. 1365-1386.
- [54] Ottesen, J. L., and R. A. Krammes. Speed-profile model for a design-consistency evaluation procedure in the United States. *Transportation Research Record*, Vol. 1701, No. 1, 2000, pp. 76-85.
- [55] Polus, A., K. Fitzpatrick, and D. B. Fambro. Predicting operating speeds on tangent sections of two-lane rural highways. *Transportation Research Record*, Vol. 1737, No. 1, 2000, pp. 50-57.
- [56] Abdelwahab, W., M. Aboul-Ela, and J. J. R. Morrall. Geometric design consistency based on speed change on horizontal curves. *Road Transport Research*, Vol. 7, No. 1, 1998.
- [57] Igene, M., and O. Ogirigbo. Evaluating the geometric design consistency and road safety on two-lane single carriageways using operating speed criteria. *J. Sci. Technol. Res*, Vol. 3, No. 1, 2021, pp. 90-98.
- [58] Gemechu, S. M., and G. S. Tulu. Safety effects of geometric design consistency on two-lane rural highways: the case of Ethiopia. *American journal of traffic transportation engineering*, Vol. 6, No. 4, 2021, pp. 107-115.
- [59] Garber, N. J., and A. A. Ehrhart. Effect of speed, flow, and geometric characteristics on crash frequency for two-lane highways. *Transportation Research Record*, Vol. 1717, No. 1, 2000, pp. 76-83.
- [60] Taylor, M. C., A. Baruya, and J. V. Kennedy. *The relationship between speed and accidents on rural single-carriageway roads*. TRL, 2002.
- [61] Council, F. M., M. Reurings, R. Srinivasan, S. Masten, and D. Carter. Development of a Speeding-Related Crash Typology:[Summary Report]. In, Turner-Fairbank Highway Research Center, 2010.
- [62] Bornheimer, C. Finding a new safety performance function for two-way, two-lane highways in rural areas. In *Civil Engineering*, University of Kansas, 2011. p. 130.
- [63] Dell'Acqua, G., and F. Russo. Safety performance functions for low-volume roads. *The Baltic Journal of Road and Bridge Engineering*, Vol. 6, No. 4, 2011, pp. 225-225.
- [64] Fitzpatrick, K., L. Elefteriadou, D. W. Harwood, J. M. Collins, J. McFadden, I. B. Anderson, R. A. Krammes, N. Irizarry, K. D. Parma, and K. M. Bauer. Speed prediction for two-lane rural highways. In, United States. Federal Highway Administration, 2000. p. 213.



- [65] Montella, A., and L. L. Imbriani. Safety performance functions incorporating design consistency variables. *Accident Analysis Prevention*, Vol. 74, 2015, pp. 133-144.
- [66] McFadden, J., and L. Elefteriadou. Evaluating horizontal alignment design consistency of two-lane rural highways: Development of new procedure. *Transportation Research Record*, Vol. 1737, No. 1, 2000, pp. 9-17.
- [67] Dhahir, B., and Y. Hassan. Using horizontal curve speed reduction extracted from the naturalistic driving study to predict curve collision frequency. *Accident Analysis & Prevention*, Vol. 123, 2019, pp. 190-199.
- [68] Hauer, E. Statistical road safety modeling. *Transportation Research Record*, Vol. 1897, No. 1, 2004, pp. 81-87.
- [69] Huang, H., and H. C. Chin. Modeling road traffic crashes with zero-inflation and site-specific random effects. *Statistical Methods & Applications*, Vol. 19, No. 3, 2010, pp. 445-462.
- [70] Sharma, A., and V. Landge. Zero inflated negative binomial for modeling heavy vehicle crash rate on Indian rural highway. Presented at International Journal of Advances in Engineering & Technology, 2013.
- [71] Wang, C., M. Quddus, and S. Ison. The effects of area-wide road speed and curvature on traffic casualties in England. *Journal of transport geography*, Vol. 17, No. 5, 2009, pp. 385-395.
- [72] Xu, J., K. M. Kockelman, and Y. Wang. Modeling crash and fatality counts along mainlanes and frontage roads across texas: the roles of design, the built environment, and weather In *93rd Annual Meeting of the Transportation Research*, 2014. p. 24.
- [73] Yan, X., B. Wang, M. An, and C. Zhang. Distinguishing between rural and urban road segment traffic safety based on zero-inflated negative binomial regression models. *Discrete Dynamics in Nature Society*, Vol. 2012, 2012, p. 11.
- [74] Lave, C. A. J. T. A. E. R. Speeding, coordination, and the 55 mph limit. Vol. 75, No. 5, 1985, pp. 1159-1164.
- [75] Ma, J., K. M. Kockelman, and P. Damien. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention*, Vol. 40, No. 3, 2008, pp. 964-975.
- [76] Huang, H., M. A. Abdel-Aty, and A. L. Darwiche. County-level crash risk analysis in Florida: Bayesian spatial modeling. *Transportation Research Record*, Vol. 2148, No. 1, 2010, pp. 27-37.
- [77] Liu, J., and A. J. Khattak. Gate-violation behavior at highway-rail grade crossings and the consequences: using geo-spatial modeling integrated with path analysis. *Accident Analysis Prevention*, Vol. 109, 2017, pp. 99-112.
- [78] Xu, P., and H. Huang. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis Prevention*, Vol. 75, 2015, pp. 16-25.
- [79] Quddus, M. A. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. *Accident Analysis Prevention*, Vol. 40, No. 4, 2008, pp. 1486-1497.
- [80] Wang, X., J. Liu, A. J. Khattak, and D. Clarke. Non-crossing rail-trespassing crashes in the past decade: A spatial approach to analyzing injury severity. *Safety science*, Vol. 82, 2016, pp. 44-55.
- [81] Fotheringham, A. S., C. Brunson, and M. Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

- [82] Gomes, M. J. T. L., F. Cunto, and A. R. da Silva. Geographically weighted negative binomial regression applied to zonal level safety performance models. *Accident Analysis Prevention*, Vol. 106, 2017, pp. 254-261.
- [83] Hadayeghi, A., A. S. Shalaby, and B. N. Persaud. Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. *Accident Analysis Prevention*, Vol. 42, No. 2, 2010, pp. 676-688.
- [84] Hezaveh, A. M., R. Arvin, and C. R. Cherry. A geographically weighted regression to estimate the comprehensive cost of traffic crashes at a zonal level. *Accident Analysis Prevention*, Vol. 131, 2019, pp. 15-24.
- [85] Iyanda, A. E., and T. Osayomi. Is there a relationship between economic indicators and road fatalities in Texas? A multiscale geographically weighted regression analysis. *GeoJournal*, Vol. 86, No. 6, 2021, pp. 2787-2807.
- [86] Ji, S., Y. Wang, and Y. Wang. Geographically weighted poisson regression under linear model of coregionalization assistance: Application to a bicycle crash study. *Accident Analysis Prevention*, Vol. 159, 2021, p. 106230.
- [87] Li, Z., W. Wang, P. Liu, J. M. Bigham, and D. R. Ragland. Using geographically weighted Poisson regression for county-level crash modeling in California. *Safety science*, Vol. 58, 2013, pp. 89-97.
- [88] Liu, J., A. Hainen, X. Li, Q. Nie, and S. Nambisan. Pedestrian injury severity in motor vehicle crashes: an integrated spatio-temporal modeling approach. *Accident Analysis Prevention*, Vol. 132, 2019, p. 105272.
- [89] Liu, J., A. J. Khattak, and B. Wali. Do safety performance functions used for predicting crash frequency vary across space? Applying geographically weighted regressions to account for spatial heterogeneity. *Accident Analysis Prevention*, Vol. 109, 2017, pp. 132-142.
- [90] Mathew, S., S. S. Pulugurtha, and S. Duvvuri. Exploring the effect of road network, demographic, and land use characteristics on teen crash frequency using geographically weighted negative binomial regression. *Accident Analysis Prevention*, Vol. 168, 2022, p. 106615.
- [91] Mohammadnazar, A., I. Mahdinia, N. Ahmad, A. J. Khattak, and J. Liu. Understanding how relationships between crash frequency and correlates vary for multilane rural highways: Estimating geographically and temporally weighted regression models. *Accident Analysis Prevention*, Vol. 157, 2021, p. 14.
- [92] Wachnicka, J., and K. Palikowska. APPLICATION OF THE GWR MODEL FOR PREDICTING THE ROAD FATALITIES RATE ON THE ROAD NETWORK IN THE NUTS 3 REGIONS IN EUROPE ON THE EXAMPLE OF KUYAVIAN--POMERANIAN VOIVODESHIP. *Journal of KONBiN*, Vol. 50, 2020, pp. 215-236.
- [93] Wang, C., S. Li, and J. Shan. Non-Stationary Modeling of Microlevel Road-Curve Crash Frequency with Geographically Weighted Regression. *ISPRS International Journal of Geo-Information*, Vol. 10, No. 5, 2021, p. 286.
- [94] Cardozo, O. D., J. C. García-Palomares, and J. Gutiérrez. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, Vol. 34, 2012, pp. 548-558.
- [95] Das, S., X. Sun, and M. Sun. Rule-based safety prediction models for rural two-lane run-off-road crashes. *International journal of transportation science technology*, Vol. 10, No. 3, 2021, pp. 235-244.

- [96] Iranitalab, A., and A. Khattak. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, Vol. 108, 2017, pp. 27-36.
- [97] Wei, Z., S. Das, and Y. Zhang. Short Duration Crash Prediction for Rural Two-Lane Roadways: Applying Explainable Artificial Intelligence. *Transportation Research Record*, 2022, p. 03611981221096113.
- [98] Wen, X., Y. Xie, L. Jiang, Y. Li, and T. Ge. On the interpretability of machine learning methods in crash frequency modeling and crash modification factor development. *Accident Analysis Prevention*, Vol. 168, 2022, p. 106617.
- [99] Zhang, X., S. T. Waller, and P. Jiang. An ensemble machine learning-based modeling framework for analysis of traffic crash frequency. *Computer-Aided Civil Infrastructure engineering*, Vol. 35, No. 3, 2020, pp. 258-276.
- [100] Saha, D., P. Alluri, and A. Gan. A random forests approach to prioritize Highway Safety Manual (HSM) variables for data collection. *Journal of Advanced Transportation*, Vol. 50, No. 4, 2016, pp. 522-540.
- [101] ---. Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees. *Accident Analysis Prevention*, Vol. 79, 2015, pp. 133-144.
- [102] Basso, F., L. J. Basso, F. Bravo, and R. Pezoa. Real-time crash prediction in an urban expressway using disaggregated data. *Transportation research part C: emerging Technologies*, Vol. 86, 2018, pp. 202-219.
- [103] Yu, R., and M. Abdel-Aty. Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis Prevention*, Vol. 51, 2013, pp. 252-259.
- [104] Ijaz, M., M. Zahid, and A. Jamal. A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. *Accident Analysis Prevention*, Vol. 154, 2021, p. 106094.
- [105] Lee, C., and X. Li. Predicting driver injury severity in single-vehicle and two-vehicle crashes with boosted regression trees. *Transportation Research Record*, Vol. 2514, No. 1, 2015, pp. 138-148.
- [106] Qu, X., W. Wang, W. Wang, and P. Liu. Real-time freeway sideswipe crash prediction by support vector machine. *IET Intelligent Transport Systems*, Vol. 7, No. 4, 2013, pp. 445-453.
- [107] Ossiander, E. M., and P. Cummings. Freeway speed limits and traffic fatalities in Washington State. *Accident Analysis Prevention*, Vol. 34, No. 1, 2002, pp. 13-18.
- [108] Aguiro-Valverde, J., and P. P. Jovanis. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. *Transportation Research Record*, Vol. 2136, No. 1, 2009, pp. 82-91.
- [109] Anarkooli, A. J., B. Persaud, M. Hosseinpour, and T. Saleem. Comparison of univariate and two-stage approaches for estimating crash frequency by severity—Case study for horizontal curves on two-lane rural roads. *Accident Analysis Prevention*, Vol. 129, 2019, pp. 382-389.
- [110] Ahmadi, A., A. Jahangiri, V. Berardi, and S. G. Machiani. Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. *Journal of Transportation Safety*, Vol. 12, No. 4, 2020, pp. 522-546.
- [111] Shankar, V., and F. Mannering. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. *Journal of Safety Research*, Vol. 27, No. 3, 1996, pp. 183-194.

- [112] O'donnell, C., and D. H. Connor. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis Prevention*, Vol. 28, No. 6, 1996, pp. 739-753.
- [113] Shankar, V., F. Mannering, and W. Barfield. Statistical analysis of accident severity on rural freeways. *Accident Analysis Prevention*, Vol. 28, No. 3, 1996, pp. 391-401.
- [114] *Fatality Facts 2018 Urban/rural comparison*. <https://www.iihs.org/topics/fatality-statistics/detail/urban-rural-comparison>.
- [115] Chen, M., X. Zhang, F. Rahman, J. Brashear, and R. R. Souleyrette. Measuring Congestion for Strategic Highway Investment for Tomorrow (SHIFT) Implementation (PL-32).In, Kentucky Transportation Center Research Report, 2019. p. 42.
- [116] Hauer, E., and J. Bamfo. Two tools for finding what function links the dependent variable to the explanatory variables.In *Proceedings of the ICTCT 1997 Conference, Lund, Sweden*, 1997. p. 18.
- [117] Srinivasan, R., and K. M. Bauer. Safety performance function development guide: Developing jurisdiction-specific SPFs.In, United States. Federal Highway Administration. Office of Safety, 2013. p. 47.
- [118] Lamm, R., B. Psarianos, and T. Mailaender. *Highway design and traffic safety engineering handbook*. 1999.
- [119] Fitzpatrick, K. Evaluation of design consistency methods for two-lane rural highways: executive summary.In, United States. Federal Highway Administration, 2000.
- [120] Hauke, J., and T. Kossowski. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, Vol. 30, No. 2, 2011, pp. 87-93.
- [121] Cliff, A., and J. Ord. Spatial processes: models and applications. London: Pion. Cressie, N.(1991). Statistics for spatial data.In, Wiley, New York, 1981.
- [122] AASHTO, M. 320. *Standard Specification for Performance-Graded Asphalt Binder*, American Association of State Highway and Transportation Officials, 2010.
- [123] Kononov, J., and B. Allery. Level of service of safety: conceptual blueprint and analytical framework. *Transportation Research Record*, Vol. 1840, No. 1, 2003, pp. 57-66.
- [124] Greene, W. H. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. 1994.
- [125] Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, Vol. 34, No. 1, 1992, pp. 1-14.
- [126] Yusuf, O., and T. Bello. Zero inflated poisson and zero inflated negative binomial models with application to number of falls in the elderly. *Biostatistics Biometrics Open Access Journal*, Vol. 1, No. 4, 2017, pp. 69-75.
- [127] Ridout, M., J. Hinde, and C. G. Demétrio. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, Vol. 57, No. 1, 2001, pp. 219-223.
- [128] Breiman, L. Random forests. *Machine learning*, Vol. 45, No. 1, 2001, pp. 5-32.
- [129] Svetnik, V., A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information computer sciences*, Vol. 43, No. 6, 2003, pp. 1947-1958.

- [130] Han, S., B. D. Williamson, and Y. Fong. Improving random forest predictions in small datasets from two-phase sampling designs. *BMC medical informatics decision making*, Vol. 21, No. 1, 2021, pp. 1-9.
- [131] Parmar, H., S. Bhandari, and G. Shah. Sentiment mining of movie reviews using Random Forest with Tuned Hyperparameters. In *International Conference on Information Science*, Kerala, 2014. pp. 1-6.
- [132] Probst, P., M. N. Wright, and A. L. Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining knowledge discovery*, Vol. 9, No. 3, 2019, p. e1301.
- [133] Genuer, R., J. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31, 2225 10.1016. *J. PATREC*, Vol. 14, 2010.
- [134] Han, H., X. Guo, and H. Yu. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, IEEE, 2016. pp. 219-224.
- [135] Sharma, A., and V. Landge. Pedestrian accident prediction model for rural road. *International Journal of Science and Advanced Technology*, Vol. 2, No. 8, 2012, pp. 66-72.
- [136] Lewis, F., A. Butler, and L. Gilbert. A unified approach to model selection using the likelihood ratio test. *Methods in Ecology Evolution*, Vol. 2, No. 2, 2011, pp. 155-162.
- [137] Lamm, R., E. M. Choueiri, J. C. Hayward, and A. Paluri. Possible design procedure to promote design consistency in highway geometric design on two-lane rural roads. *Transportation Research Record*, Vol. 1195, 1988, p. 111.
- [138] Pirdavani, A., T. Bellemans, T. Brijs, and G. Wets. Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. *Journal of Transportation Engineering*, Vol. 140, No. 8, 2014, p. 04014032.
- [139] *Strategic Highway Investment Formula for Tomorrow (SHIFT)*, Kentucky Transportation Cabinet. <https://transportation.ky.gov/SHIFT/Pages/default.aspx2021>.
- [140] Kentucky Terrain In, Kentucky Geological Survey (KGS) [https://kgs.uky.edu/kgsweb/download/mc187\\_12.pdf](https://kgs.uky.edu/kgsweb/download/mc187_12.pdf).
- [141] El-Basyouny, K., and T. Seyed. Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis Prevention*, Vol. 41, No. 4, 2009, pp. 820-828.
- [142] Geedipally, S. R., J. A. Bonneson, M. P. Pratt, and D. Lord. Severity distribution functions for freeway segments. *Transportation Research Record*, Vol. 2398, No. 1, 2013, pp. 19-27.
- [143] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 2001, pp. 1189-1232.
- [144] Molnar, C. *Interpretable machine learning*. Lulu. com, 2020.

## VITA

### EDUCATION

---

**Ph.D. in Civil Engineering**, GPA 4.00/4.00 Expected August  
2022

**Concentration:** Transportation  
Engineering

University of Kentucky, Lexington, Kentucky

**M.Sc. in Civil Engineering**, GPA 4.00/4.00 May 2019

University of Kentucky, Lexington, Kentucky

**B.Sc. in Civil Engineering**, GPA 3.68/4.00 March, 2016

Bangladesh University of Engineering and Technology, Dhaka,  
Bangladesh

### PROFESSIONAL EXPERIENCES

---

**University of Kentucky, Lexington, KY** [August 2017 –  
Graduate Research Assistant Present]

**University of Kentucky, Lexington, KY** [January 2019 –  
Graduate Teaching Assistant May 2022]

**Bangladesh University of Engineering and Technology, Dhaka,  
Bangladesh** [April 2016 – July  
Undergraduate Research Assistant 2017]

### LICENCE & CERTIFICATION

---

Fundamentals of Engineering (FE) Civil - Passed December 2020

### LEADERSHIP EXPERIENCE & EXTRA-CURRICULAR ACTIVITIES

---

- ❖ Vice President -Student Chapter Institute of Transportation Engineers (ITE) at UKY, 2020-present
- ❖ Treasurer -Student Chapter Institute of Transportation Engineers (ITE) at UKY, 2019-2020
- ❖ Rotaract Club of Banani Model Town, Dhaka, Bangladesh—Joint Secretary, 2015-2017

## **AWARD & SCHOLARSHIP**

---

- ❖ Kentucky Section Institute of Transportation Engineers William (Bill) Seymour Scholarship, 2020
- ❖ People's Choice Award—3-Minute Thesis Competition, UK GradResearch Live!, 2020
- ❖ University of Kentucky Graduate Student Block Funding Awards, 2020
- ❖ Runners-up at Institute of Transportation Engineers (ITE) International Collegiate Traffic Bowl, Austin, Texas, 2019
- ❖ Runners-up at Southern District Institute of Transportation Engineers (SDITE) Traffic Bowl, Mobile, Alabama, 2018
- ❖ Thomas J. and Viva B. Timmons Graduate Fellowship, 2017-2021

## **PUBLICATIONS**

---

### **Journals Published:**

- ❖ Haque, N., Hadiuzzaman, M., Rahman, F. and Siam, M. R. K (2019). Real-time motion trajectory based head-on crash probability estimation on two-lane undivided highway. *Journal of Transportation Safety & Security*.
- ❖ Hadiuzzaman, M., Siam, M. R. K, Haque, N., Shimu, T. H. and Rahman, F. (2018). Adaptive Neuro-Fuzzy Approach for Modeling Equilibrium Speed-Density Relationship. *Transportmetrica A: Transport Science*, ISSN: 2324–9935.
- ❖ Khan, M. M. I., Hadiuzzaman, M., Das, T., Rahman, F., Shimu, T. H. (2018). A structural equation approach in modeling perceived service quality of passenger ferry. *Management Research and Practice*, Issue 1.
- ❖ Haque, N., Rahman, F., Hadiuzzaman, M., Hossain, S., Siam, M. R. K. and Qiu, T. Z. (2017). Pixel Based Heterogeneous Traffic Measurement considering Shadow and Illumination Variation. *Signal, Image and Video Processing*. ISSN:1863–1711, pp 1-8.
- ❖ Muniruzzaman, S. M., Haque, N., Rahman, F., Siam, M. R. K., Musabbir, R., Hadiuzzaman, M. and Hossain, S. (2016). Deterministic algorithm for traffic detection in free-flow and congestion using video sensor. *Journal of Built Environment, Technology and Engineering*, ISSN: 0128–1003, Vol.1, Issue 1, pp 111-130, no 14.
- ❖ Muniruzzaman, S. M., Hadiuzzaman, M., Rahman, F., Hasan, T. (2016). Calibration and Validation of Microscopic Simulation Model for Non-Lane Based Heterogeneous Traffic Stream of Developing Country. *Journal of Built Environment, Technology and Engineering*, ISSN: 0128–1003, Vol.1, Issue 1, pp 244-251, no 29.