CRASH PREDICTION MODELS ON TRUCK-RELATED CRASHES ON TWO-LANE RURAL HIGHWAYS WITH VERTICAL CURVES

A Thesis
Presented to
The Graduate Faculty of The University of Akron

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

Srutha Vavilikolanu
December, 2008
CRASH PREDICTION MODELS ON TRUCK-RELATED CRASHES ON TWO-LANE RURAL HIGHWAYS WITH VERTICAL CURVES

Srutha Vavilikolanu

Thesis

Approved: 

Advisor
Dr. William H. Schneider IV

Accepted: 

Department Chair
Dr. Wieslaw K. Binienda

Committee Member
Dr. Ping Yi

Dean of the College
Dr. George K. Haritos

Committee Member
Dr. Richard Steiner

Dean of the Graduate School
Dr. George R. Newkome

Date
ABSTRACT

According to Federal Motor Carrier Safety Administration (FMCSA), truck involvement in fatal crashes is more on rural areas than on urban areas. The Fatality Analysis Reporting System (FARS) encyclopedia also indicates that truck involvement in fatal crashes are approximately 12% of the total fatal crashes in the nation and 14 % in The State of Ohio. One area for potential concern is the role of vertical curves on truck crashes. In the design of vertical curves stopping distance, grade and length of the curve are important factors taken into consideration. Vehicle operations on vertical curves are influenced by the grade of the curve, stopping sight distance and vehicle speed. These factors may create operational issues for vehicles traveling on vertical curves and in turn increase the likelihood for crashes. Truck specific studies in the past have focused on geometric roadway factors associated with crashes on vertical curves. Most of the research studies are focused on crest curve truck crashes, and little research has been done on crashes on vertical sag curves.

The main research goal of the study is to develop prediction models to evaluate the impact of geometry, traffic volumes and speed on truck-related crashes on two-lane rural vertical curves. The accomplishment of the research goal is achieved by setting five objectives. The first objective is to develop three crash prediction models using negative binomial regression model. These models are 1. Full model – for all vertical curves 2.
Reduced model I - for crest curves only and 3. Reduced model II - for sag curves only.
The dataset includes 1,935 vertical curve segments with 205 truck crashes from 2002-2006. In second and third objective, Full Bayes approach is used to enhance the results obtained in the Reduced Models I and II. These results are then compared to the initial models. The fourth objective is evaluating the vertical curve variables which are statistically significant with truck-related crashes. These results show that higher grade change for the length of the vertical curve, total width in the range of 24 to 26ft, more number of passenger cars and trucks, increases the truck-related crashes on both crest and sag curves. Low speed limit on crest curves and high speed limit on sag curves increases truck-related crashes which may seem counter intuitive. The fifth objective is to provide suggestions on effective methods to reduce truck related crashes and improve safety. Some potential areas for design improvement may include flattening of steep vertical curves, advisory speed signs and increasing the roadway width on rural vertical curves in Ohio.
ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my advisor, Professor Dr. William H. Schneider IV, for his valuable guidance, continuous support and encouragement throughout this study.

The useful data information from Ohio Department of Transportation and Department of Public Safety is really appreciated. Acknowledgements are also extended to my committee members, Dr. Ping Yi, Dr. Richard Steiner, for reviewing my work and helpful recommendations.

Special thanks are given to my fellow graduate students, John Tsapakis and Darren N. Moore for their useful discussions related to the topic of data validation, and discrete choice models. The sincere friendship and support from them always gave me energy and impetus to finish this dissertation.

My deepest gratitude goes to my family and friends who provide love and support more than I could ever expect.
TABLE OF CONTENTS

LIST OF TABLES .............................................................................................................. x
LIST OF FIGURES .......................................................................................................... xii
LIST OF EQUATIONS ................................................................................................... xiv
LIST OF ACRONYMS .................................................................................................. xvii

CHAPTER

I. INTRODUCTION ........................................................................................................... 1
  1.1 Objectives ............................................................................................................. 3
  1.2 Overview of Thesis ............................................................................................. 3
    1.2.1 Chapter II ................................................................................................. 4
    1.2.2 Chapter III ................................................................................................ 4
    1.2.3 Chapter IV ................................................................................................ 4
    1.2.4 Chapter V ................................................................................................. 4
    1.2.5 Chapter VI ............................................................................................... 5

II. LITERATURE REVIEW ............................................................................................. 6
  2.1 Introduction .......................................................................................................... 6
    2.1.1 The National and State Trends ................................................................. 7
  2.2 State of Practice on the Analysis of Crashes ....................................................... 9
    2.2.1 Discrete Choice Models ......................................................................... 11
4.1.3 Overdispersion ................................................................. 45

4.2 Negative Binomial Regression Model ........................................ 46

4.2.1 Probability Density Function of Negative Binomial Regression
Model .......................................................................................46

4.2.2 Negative Binomial Regression Model Form ........................... 49

4.3 Maximum Likelihood Estimation For Negative Binomial Dispersion
Parameter ................................................................. 51

4.4 Negative Binomial Based on Gibbs Sampling .......................... 53

4.5 Selection of Variable Criteria .................................................. 56

4.6 Testing for Zero-Inflated Models ............................................... 56

4.7 General Assessment of the Model .............................................. 57

4.7.1 Goodness-of-fit ................................................................. 57

4.7.2 Overdispersion Criteria ...................................................... 58

V. MODEL DEVELOPMENT AND RESULTS ........................................ 60

5.1 Model Assessment and Model Development .............................. 60

5.2 Model Assessment ................................................................. 61

5.2.1 General Assessment of the Model ....................................... 62

5.2.2 Variable Coefficient Interpretation ....................................... 63

5.3 Model Development ............................................................... 63

5.3.1 Full Model Using Negative Binomial Regression Model ........ 64

5.3.2 Reduced Model I Using Negative Binomial Regression Model ... 65

5.3.3 Reduced Model II Using Negative Binomial Regression Model .... 66

5.3.4 Reduced Models I and II Using Full Bayes ............................ 67
5.4 Results and Inference........................................................................................................ 68
  5.4.1 Full Model................................................................................................................. 68
  5.4.2 Reduced Model I ..................................................................................................... 71
  5.4.3 Improvements in the Reduced Model I for the Crest Curves .............. 75
  5.4.4 Reduced Model II ................................................................................................. 79
  5.4.5 Improvements in the Reduced Model II for Sag Curves .................... 82

5.5 Summary of Results ................................................................................................. 86
  5.5.1 Geometric Characteristics ................................................................. 86
  5.5.2 Traffic Volumes ...................................................................................... 87
  5.5.3 Speed Variable ......................................................................................... 88

VI. CONCLUSION AND RECOMMENDATIONS ....................................................... 89
  6.1 Model Development ......................................................................................... 90
  6.2 Improvement Using Bayesian Method ................................................................. 90
  6.3 Summary of Findings .................................................................................... 90
    6.3.1 Geometric Characteristics ..................................................................... 90
    6.3.2 Traffic Volumes ..................................................................................... 91
    6.3.3 Speed Variable ......................................................................................... 91
  6.4 Improving Safety .............................................................................................. 92

BIBLIOGRAPHY ............................................................................................................. 94

APPENDIX ..................................................................................................................... 102


<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1. Total Fatal Crashes in United States and Ohio from 2002 to 2005</td>
<td>7</td>
</tr>
<tr>
<td>2.2. Large Trucks Involved in Fatal Crashes in United States and Ohio from 2002 to 2005</td>
<td>8</td>
</tr>
<tr>
<td>2.3. Fatal Crashes by Urban vs. Rural Location in United States and Ohio from 2002 to 2005</td>
<td>8</td>
</tr>
<tr>
<td>2.4. Large Truck Involved in Fatal Crashes in Percentage by Urban vs. Rural Location in US and Ohio from 2003 to 2006</td>
<td>8</td>
</tr>
<tr>
<td>2.5. Total Crashes Modeled Using the Negative Binomial Regression Model Developed for Specific Geometric Segment</td>
<td>17</td>
</tr>
<tr>
<td>2.6. Truck-Related Crashes Modeled Using the Negative Binomial Regression Model</td>
<td>19</td>
</tr>
<tr>
<td>2.7. Sample Sizes for Modeling Crashes Using the Negative Binomial Regression Model</td>
<td>21</td>
</tr>
<tr>
<td>3.1. Modified data variables</td>
<td>33</td>
</tr>
<tr>
<td>3.2. Descriptive Statistics for Entire Modified Dataset</td>
<td>41</td>
</tr>
<tr>
<td>3.3. Descriptive Statistics for the Crest Vertical Curve Data</td>
<td>41</td>
</tr>
<tr>
<td>3.4. Descriptive Statistics for the Sag Vertical Curve Data</td>
<td>42</td>
</tr>
<tr>
<td>5.1 Statistical Summary of Full Model Using Negative Binomial Regression Model</td>
<td>70</td>
</tr>
<tr>
<td>5.2. Statistical Summary of Reduced Model I Using Negative Binomial Regression Model</td>
<td>73</td>
</tr>
<tr>
<td>5.3 Statistical Summary of Reduced Model I using Full Bayes Technique</td>
<td>74</td>
</tr>
</tbody>
</table>
5.4 Statistical Summary of Reduced Model II Using Negative Binomial Regression Model ................................................................. 80

5.5 Statistical Summary of Reduced Model II Using Full Bayes Technique. ..................... 81

5.6 Statistical Summary of Reduced Models I and II. ................................................................. 86
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1. Types of Crash Analysis</td>
<td>10</td>
</tr>
<tr>
<td>2.2. Illustration of the Multinomial Logit Framework</td>
<td>12</td>
</tr>
<tr>
<td>3.1. Structure of the Given Dataset</td>
<td>25</td>
</tr>
<tr>
<td>3.2. State route 2 in The State of Ohio obtained from ARC Map- GIS</td>
<td>28</td>
</tr>
<tr>
<td>3.3. Horizontal Curve Check for Route 2 in Defiance County</td>
<td>29</td>
</tr>
<tr>
<td>3.4. US Route 23 in The State of Ohio Obtained from ARC Map- GIS</td>
<td>30</td>
</tr>
<tr>
<td>3.5. Interstate Route 71 in The State of Ohio Obtained from ARC Map- GIS</td>
<td>30</td>
</tr>
<tr>
<td>3.6. Route 81 in the City of Ada in The State of Ohio Quad Map</td>
<td>32</td>
</tr>
<tr>
<td>3.7. Criteria for Forming a Vertical Curve</td>
<td>34</td>
</tr>
<tr>
<td>3.8. Types of Crest Curves</td>
<td>35</td>
</tr>
<tr>
<td>3.9. Types of Sag Curves</td>
<td>36</td>
</tr>
<tr>
<td>3.10. Structure of Study Area Concern</td>
<td>36</td>
</tr>
<tr>
<td>5.1. Sensitivity Analysis at Each Chain for Reduced Model I</td>
<td>75</td>
</tr>
<tr>
<td>5.2. Gelman Rubin Plots for the Variable Estimates for Reduced Model I</td>
<td>77</td>
</tr>
<tr>
<td>5.3. Kernel Density Plots for Reduced Model I</td>
<td>78</td>
</tr>
<tr>
<td>5.4. Sensitivity Analysis at Each Chain for Reduced Model II</td>
<td>82</td>
</tr>
<tr>
<td>5.5. Gelman Rubin Plots for the variable estimates for Reduced Model II</td>
<td>84</td>
</tr>
<tr>
<td>5.6. Kernel Density Plots for Reduced model II</td>
<td>85</td>
</tr>
</tbody>
</table>
6.1 Mild and Steep Vertical Crest Curves. ................................................................. 92
**LIST OF EQUATIONS**

<table>
<thead>
<tr>
<th>Equation</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Crash Frequency</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Crash Density</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Crash Rate</td>
<td>10</td>
</tr>
<tr>
<td>2.4 The Poisson Equation</td>
<td>14</td>
</tr>
<tr>
<td>2.5 The General Equation for the Negative Binomial Regression Model</td>
<td>14</td>
</tr>
<tr>
<td>3.1 Degree of Curvature</td>
<td>27</td>
</tr>
<tr>
<td>3.2 Algebraic Difference Between the Gradients</td>
<td>37</td>
</tr>
<tr>
<td>3.3 Rate of Vertical Curvature</td>
<td>37</td>
</tr>
<tr>
<td>3.4 Grade Change for the Length of the Curve</td>
<td>38</td>
</tr>
<tr>
<td>4.1 Probability Density Function of Poisson Distribution</td>
<td>44</td>
</tr>
<tr>
<td>4.2 Poisson Distribution with Explanatory Variables</td>
<td>44</td>
</tr>
<tr>
<td>4.3 Mean and Variance Relationship in Poisson Distribution</td>
<td>45</td>
</tr>
<tr>
<td>4.4 General Form of the Poisson Regression Model</td>
<td>45</td>
</tr>
<tr>
<td>4.5 The Probability Distribution of the Negative Binomial Regression Model</td>
<td>46</td>
</tr>
<tr>
<td>4.6 The Unconditional Probability Distribution of the Negative Binomial Regression Model</td>
<td>47</td>
</tr>
<tr>
<td>4.7 The Probability Distribution in terms of Inverse Dispersion Parameter</td>
<td>47</td>
</tr>
<tr>
<td>4.8 Normal Form of Density Function of the Gamma Distribution</td>
<td>48</td>
</tr>
</tbody>
</table>
5.4 Parameter Estimates of Full Model................................................................. 68
5.5 Parameter Estimates of Reduced Model I..................................................... 71
5.6 Parameter Estimates of Reduced Model II ............................................... 79
LIST OF ACRONYMS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT</td>
<td>Annual Average Daily traffic</td>
</tr>
<tr>
<td>AASHTO</td>
<td>American Association of State Highway and Transportation Officials</td>
</tr>
<tr>
<td>ADT</td>
<td>Average Daily Traffic</td>
</tr>
<tr>
<td>ATSSA</td>
<td>American Traffic Safety Services Association</td>
</tr>
<tr>
<td>BMV</td>
<td>Bicycle Motor Vehicle</td>
</tr>
<tr>
<td>CUY</td>
<td>Cuyahoga County</td>
</tr>
<tr>
<td>DEF</td>
<td>Defiance County</td>
</tr>
<tr>
<td>ERI</td>
<td>Erie County</td>
</tr>
<tr>
<td>FUL</td>
<td>Fulton County</td>
</tr>
<tr>
<td>FARS</td>
<td>Fatality Analysis Reporting System</td>
</tr>
<tr>
<td>FMCSA</td>
<td>Federal Motor Carrier Safety Administration</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>GLMs</td>
<td>Generalized Linear Models</td>
</tr>
<tr>
<td>GVWR</td>
<td>Gross Vehicle Weight Rating</td>
</tr>
<tr>
<td>LAK</td>
<td>Lake County</td>
</tr>
<tr>
<td>LUC</td>
<td>Lucas County</td>
</tr>
<tr>
<td>NHTSA</td>
<td>National Highway Traffic Safety Administration</td>
</tr>
<tr>
<td>NLF</td>
<td>Network Linear Feature</td>
</tr>
<tr>
<td>ODOT</td>
<td>Ohio Department of Transportation</td>
</tr>
<tr>
<td>ODPS</td>
<td>Ohio Department of Public Safety</td>
</tr>
<tr>
<td>OTT</td>
<td>Ottawa County</td>
</tr>
<tr>
<td>PDO</td>
<td>Property Damage Only</td>
</tr>
<tr>
<td>QCQA</td>
<td>Quality Control Quality Assurance</td>
</tr>
<tr>
<td>RIC</td>
<td>Richland County</td>
</tr>
<tr>
<td>SUV</td>
<td>Sport Utility Vehicle</td>
</tr>
<tr>
<td>TSP</td>
<td>Transportation Safety Planning</td>
</tr>
<tr>
<td>US</td>
<td>United States of America</td>
</tr>
<tr>
<td>USGS</td>
<td>United States Geological Survey</td>
</tr>
<tr>
<td>WIL</td>
<td>Williams County</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Trucks carry more than 11 billion tons of goods per year and account for less than 5% of all highway crashes [1]. However, the injury-severity results from these crashes include approximately 13% of all highway fatalities [1]. Zaloshnja estimated that, in 2004, the average single commercial vehicle crashes cost more than $59,000, and multi-trailer crashes cost closer to $88,000 [2].

In recent years, there has been much research on improving commercial fleet safety. Some of this research includes the development of probabilistic crash injury-severity models which identify key parameters such as high speed limits and type of crash [3], the impact of urban and rural settings [4], environmental conditions including wind speed [5] and more traditional predictive models [6,7]. Other research has tried to identify causality including driver fatigue [8], the driver’s medical condition [9], and regulatory compliance [10]. These studies and others provide valuable insight into the safety of heavy-duty vehicles.

One area for potential concern is the role of vertical curves on crashes. Vertical curves consist of crest and sag curves. Hassan 2003, stated that in the design of a vertical curve, the effect of grade and the minimum length of the curve are essential components effecting stopping distance [11].
Labi 2006, concluded that on crest curves, the grade of the curve and sight distance are main concerns in the design of vertical curves. In the case of sag curves, sight distance as a result of driver headlights is the main concern at night. Steep grades cause insufficient sight distance on crest curves and inadequate passing opportunities on sag curves [12]. Jessen, 2001 investigated the influence of roadway, traffic, and speed characteristic on vehicle speeds traveling on vertical crest curves. The result from this study show that posted speeds on rural two-lane crest vertical curves was found to have most influence on the operating speed [13]. The impact of grade, stopping sight distance, and vehicle speed create operational issues for vehicles traveling on vertical curves and in turn may increase the likelihood for crashes. Truck specific studies in the past have focused on geometric roadway factors associated with crashes on vertical curves. Miaou, 1994 found that vertical grade’s greater than 2 % have increasing effect on truck crash involvements [14]. Daniel, 2004 found that truck-related crashes on vertical curves with a 0.005 mean rate of vertical grade per 100 ft is a critical factor (at $\alpha = 0.1$) for improving safety [15], while Zhang, 2005 focused more specific on head-on-crashes on two-lane rural vertical curves. Zhang’s findings suggested that the effective way of reducing head-on-crashes on vertical crest curves is to soften the vertical alignment rather than widening the road [16]. Most of the research studies on vertical curves have focused on crest curve truck crashes, and there is little research on what factors are associated with truck crashes on vertical sag curves.
1.1 Objectives

There are five objectives in this study and there objectives are stated as follows:

Objective One: The first objective is to develop full model, reduced model I and reduced model II using negative binomial regression model for evaluating the impact of vertical curve elements, roadway and traffic features on rural two-lane vertical curves in Ohio.

Objective Two: The second objective is to develop Full Bayesian methodology on the developed reduced models I and II in objective one.

Objective Three: The third objective is to evaluate the effectiveness of the Full Bayesian statistical techniques on the model improvement over the initial models

Objective Four: The fourth objective is to evaluate the vertical curve elements including geometric roadway features and traffic characteristics, which are statistically significant.

Objective Five: The fifth objective is to provide suggestions on the most effective way to improve the safety of these truck-related crashes, which occur on rural two-lane vertical curves within Ohio.

1.2 Overview of Thesis

A brief outline of the chapters within this thesis are described in sections 1.2.1 through 1.2.5. Detailed discussion of these chapters II through IV are provided later in the thesis.
1.2.1 Chapter II

The literature review in Chapter II describes the crash trends nationwide with wide Ohio crashes. The material within this chapter includes the type of crashes, crash trends in Nation and in Ohio, basic practices used in the past to present to model crashes, the factors associated with modeling crashes in particular, model techniques, size of data, and empirical on Full Bayesian statistical methodology.

1.2.2 Chapter III

Chapter III describes the Empirical data that are used in this study. Additional information in this chapter includes the data cleaning and validation steps.

1.2.3 Chapter IV

Chapter IV describes the statistical methodology used for modeling truck crashes. This statistical methodology includes Poisson regression models, negative binomial regression models, and Full Bayes models along with the general assessment for each of the models.

1.2.4 Chapter V

The chapter V is the Model Development and Results chapter. In this chapter, the dataset described in the Chapter III are used to develop three negative binomial regression models. These models include full model, reduced model I for crest curves and reduced model II for sag curves; as well as two Full Bayes methods are developed for reduced models I and II. The final results also include the impact of variable specific parameters on truck crash prediction.
1.2.5 Chapter VI

Chapter VI is the conclusion and recommendation chapter. In this chapter, summary of the model results for the improved crest and sag models, conclusions and recommendations based on these summary results.
CHAPTER II
LITERATURE REVIEW

2.1 Introduction

Crashes in general are occurrences between two or more vehicles, vehicles with fixed objects, and vehicles with pedestrians or animals. There are many ways to describe crashes some examples include collision type, severity and vehicle type. Collision types include head-on, rear-end, sideswipe (same and opposite direction), angle, backing, rollover, fixed object, animal, pedestrian, and motorcycle crashes [17]. The second common way to describe crashes is crash severity. The level of crash severity is commonly described by the health condition of the driver or the worst-case condition of a passenger within the vehicle. Crash injury severity descriptions are Property Damage Only (PDO), possible injuries described as non-visible injuries, non-incapacitating injuries refer to minor visible injuries, incapacitating injuries refer to serious visible injuries, and fatal injuries [18]. In the decision-making process for budget allocations or design changes it may be more effective to focus on areas with higher injury severities than areas with higher frequencies of crashes but relatively low impact on human welfare. In addition to these losses, including both personal injury and property damage, mobility is reduced causing increased travel times and congestion along the roadway or corridor.
Vehicle classification is the third traditional way of describing crashes, because of the underlying trends that may be associated with a particular crash type or severity. The severity of crash may vary with respect to the size of the vehicle and type of collision. As a result of which, many research studies develop models exclusively for one type of vehicle classification. For example truck-related crashes in many cases produce mild injuries in a single vehicle crash. This phenomenon may be explained by the amount of mass protecting the driver. In a two vehicle crash however, this added mass of the semi-truck will have a negative impact on the second vehicle.

2.1.1 The National and State Trends

According to The National Highway Traffic Safety Administration (NHTSA), the total fatal crashes and large trucks involved in fatal crashes in The United States and in Ohio from 2002 to 2005 are summarized in Tables 2.1 and 2.2. Truck involvement crashes refer to involvement of trucks in a multivehicle crashes. Fatal crashes and large truck involved in fatal crashes by percentage for urban vs. rural in The United States and in The State of Ohio for 2003 to 2006 are given in Tables 2.3 and 2.4.

Table 2.1. Total Fatal Crashes in United States and Ohio from 2002 to 2005.

<table>
<thead>
<tr>
<th>Year</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>39,491</td>
<td>38,477</td>
<td>38,444</td>
<td>39,252</td>
</tr>
<tr>
<td>Ohio</td>
<td>1,285</td>
<td>1,165</td>
<td>1,163</td>
<td>1,223</td>
</tr>
</tbody>
</table>

NOTE: These values are taken from Fatality Analysis Reporting System (FARS) encyclopedia and Federal Motor Carrier Safety Administration (FMCSA) [19].
Table 2.2. Large Trucks Involved in Fatal Crashes in United States and Ohio from 2002 to 2005.

<table>
<thead>
<tr>
<th>Year</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>4,587</td>
<td>4,721</td>
<td>4,902</td>
<td>4,932</td>
</tr>
<tr>
<td>Ohio</td>
<td>189</td>
<td>147</td>
<td>179</td>
<td>174</td>
</tr>
</tbody>
</table>

NOTE: These values are taken from large truck crash facts, 2005 by FMCSA. According to FARS a large truck is defined as a truck with a gross vehicle weight rating (GVWR) of more than 10,000 pounds [19,20].

According to NHTSA and as seen in Table 2.4, there is a higher percentage of truck involvement crashes on rural roadways than on urban roadways both nationally and in Ohio from 2003 to 2006 [21]. This may be the result of higher speeds and different design standards for vehicles traveling in rural areas.

Table 2.3. Fatal Crashes by Urban vs. Rural Location in United States and Ohio from 2002 to 2005.

<table>
<thead>
<tr>
<th>Year</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>Rural</td>
<td>22,758</td>
<td>21,995</td>
<td>22,144</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>15,652</td>
<td>16,352</td>
<td>16,234</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>81</td>
<td>130</td>
<td>66</td>
</tr>
<tr>
<td>Ohio</td>
<td>Rural</td>
<td>837</td>
<td>606</td>
<td>709</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>448</td>
<td>559</td>
<td>454</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

NOTE: Source from FMCSA and represents fatal crashes [22].

Table 2.4. Large Truck Involved in Fatal Crashes in Percentage by Urban vs. Rural Location in US and Ohio from 2003 to 2006.

<table>
<thead>
<tr>
<th>Year</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>Rural</td>
<td>66.6</td>
<td>66.6</td>
<td>62.5</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>33</td>
<td>33.3</td>
<td>36.8</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>0.4</td>
<td>0.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Ohio</td>
<td>Rural</td>
<td>52.2</td>
<td>73.2</td>
<td>69.5</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>40.8</td>
<td>26.8</td>
<td>30.5</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

NOTE: The percentages of large trucks involved in fatal crashes are presented in this table. Source of these values are from crash reports by FMCSA [19,20].
2.2 State of Practice on the Analysis of Crashes

The development of crash analysis is comprised of crash records provided by the Ohio Department of Public Safety (ODPS) and roadway geometry from Ohio Department of Transportation (ODOT). Perhaps one of the biggest impacts of crash analysis is the improvement in technology. The impact of technology, through advanced computers, has created the capability to store and process huge data sources, enable the development of Geographic Information Systems (GIS) which may track the exact crash location, and produce new statistical software which may enhance the process of data mining analysis of crash data.

Crashes may be analyzed based on crash frequency, crash density, and crash rates [21]. Crash frequency in Equation (2.1) refers to the number of crashes at a location, and is given as:

\[
Crash\ Frequency = \frac{N}{I} \tag{2.1}
\]

where \( N \) = number of crashes, and

\( I \) = vertical curve segment.

Crash density in Equation (2.2), is the number of crashes per mile along a section of roadway and is given as:

\[
Crash\ Density = \frac{N}{L \times Y} \tag{2.2}
\]

where \( Y \) = number of years,

\( N \) = number of crashes, and

\( L \) = segment length in miles.
Crash rate in Equation (2.3), is the number of crashes occurring per million vehicle miles traveled for a location and is given as:

\[
\text{Crash Rate} = \frac{10^6 \times N}{L \times Y \times \text{ADT} \times 365 \text{ days/year}}
\]  

(2.3)

where

- ADT= Average Daily Traffic along the segment,
- \( Y = \) number of years,
- \( N = \) number of crashes, and
- \( L = \) segment length in miles.

There are two common statistical techniques, discrete choice models and prediction models, traditionally used for analyzing crashes as shown in Figure 2.1. The third technique includes other less commonly used models.

![Figure 2.1. Types of Crash Analysis.](image)
As seen in Figure 2.1, the discrete choice models are developed based on the assumption that the crash has occurred, while prediction crash models are developed to predict the future crashes based on current statistically significant variables. In the following sections, discrete choice models and prediction models are discussed in more detail.

2.2.1. Discrete Choice Models

Crash counts are discrete values and are often divided into crash severity categories. Discrete-choice models may be both ordered and unordered probability models. The ordered probit model (ordered probability) and the multinomial logit model (unordered probability) are two of the most prominent discrete outcome modeling frameworks utilized in injury severity analysis.

Ordered probability models have frequently been utilized by researchers for injury severity analysis. There are two potential drawbacks for utilizing ordered probability models [23,24,25,26,27]. First, ordered probability models constrain variable influence to be consistently positive (or negative) for increasing levels of injury severity [26]. Second, crashes of a lower injury severity level are more likely to be underreported in crash data [28]. Ordered probability models can lead to biased or inconsistent model coefficient estimates because the true rate of underreporting is unknown. This drawback does not occur in unordered probability models [29].

Multinomial logit models are obtained by assuming error terms in ordered probability model to be a generalized extreme value [29]. Its probability is estimated by standard maximum likelihood techniques. Since the estimated models are conditioned on crashes occurring, there is no interpretation of crash risk in terms of exposure. The models
identify significant factors linked with increasing the probability of a particular crash severity category, given that a crash occurred. In order to arrive at an estimate of total risk, further information is needed about the probability of a driver being involved in a crash. Figure 2.3, provides a sample illustration of the process of developing the three layers used within multinomial logit models.

Figure 2.2: Illustration of the Multinomial Logit Framework.

Vehicle description provides detail about the type of vehicle and the safety attributes of the vehicle, including airbag deployment and seat belt usage, as well as the impact area and the overall condition of the vehicle after the crash. The second category, driver attributes, includes the driver’s age, gender, and personal crash severity level. Other variables included are alcohol or drug involvement, driver inattention or fatigue, the procurement of insurance, and the driver’s vehicle speed. The environmental category includes the weather conditions, time of day, day of week, road surface conditions (dry or wet pavement), and street lighting conditions. Other information provided includes the object stuck by the vehicle, such as a tree, sign, ditch or animal.
2.2.2 Prediction Models

The second common statistical techniques used for analyzing crashes are crash prediction models. With this technique, crash prediction models are developed to estimate the safety performance of certain geometric design features of an existing or planned roadway. These models include conventional linear regression models and Generalized Linear Models (GLMs), including Poisson regression models and negative binomial regression models. In these models the response variable is discrete and at times categorical and the traffic volumes in the form of ADTs are required to model crashes. The following section provides additional discussion on the various prediction models.

Conventional linear regression model

Conventional linear regression models, which include linear regression models and multiple linear regression models, allow for data transformations such as square or square root, inverse, and natural logarithm of variables [30,48]. There are three limitations of these models. First, the response variable assumed is normally distributed which is not always statistically valid for crash count data [32,33]. Therefore, the assumption of normally distributed errors and homoscedacity are not necessarily appropriate for crash count data. Second, this model uses the least square method to estimate variable coefficients, which is not appropriate for this type of discrete data [34]. Third, the potential multicollinearity is shown among independent variables [30].

Generalized Linear Models

Generalized linear modeling techniques, which include the Poisson model for count data, are an extension of linear model methods and multivariate responses. These models
encompass non-normal response distributions [32,33]. The Poisson model uses a Poisson distribution for the response variable and assumes the mean and the variance of the response variable to be equal [32].

The Poisson equation is given as follows:

\[
\ln \mu = \beta_0 + \beta_1 x
\]  

(2.4)

where \( \mu \) = the mean of the response variable,

\( \beta_0, \beta_1 \) = estimated parameters in vector form, and

\( x \) = explanatory or independent variable [35,36,37].

The major limitation of this model is its assumption of equality between expected mean and variance of the response variable [36,38,39]. In the case of crash data quite often the variance is not equal to the mean, which results in overdispersion [33]. The overdispersion of the model violates one of the assumptions of the model and therefore the Poisson model would not be considered an appropriate model form.

**Negative Binomial Regression Model**

The overdispersion associated with count data is common and research has suggested an alternative model form called the negative binomial regression model. This model is a conjugate mixture distribution of count data and is an extension of the Poisson regression model [36,40]. The general equation for the negative binomial regression model is given as:

\[
\ln \lambda = \beta_0 + \beta_1 x + \varepsilon
\]  

(2.5)

where \( \lambda \) = expected mean value of the response variable,
\[ \beta_0, \beta_1 = \text{estimated parameters in vector form,} \]

\[ x = \text{explanatory variable or independent variable, and} \]

\[ \varepsilon_i = \text{error term, where } \exp(\varepsilon_i) \text{ has a standard gamma distribution with} \]

\[ \text{mean of one and variance } \sigma^2 [35, 36, 40]. \]

In this model, the response variable, also called as the Poisson gamma distribution, follows negative binomial distribution. Many researchers have used this model to analyze crashes [41, 39, 42, 43]. One of the limitations of this model is that it does not incorporate the possibility of the occurrence of zero inflated variables. The inflation of zeros is due to the probability of under-recording crashes or the probability of the location being safe.

One additional form of the negative binomial regression model that does incorporate the possibility of zeros is called the zero inflated model [44, 45, 44, 44].

2.3 Negative Binomial Regression Model in the Present Study

Some of the research studies from the past to the present have suggested that the negative binomial regression model is the most efficient for predicting crashes. Miaou, 1994 established relationship between truck accidents and geometric design of road sections using Poisson, Zero-Inflated and negative binomial regression models. His main focus was on roadway sections and considered geometric variables such as horizontal curvature and vertical grade as categorical variables. Abdel-Aty, 2000, modeled crash frequencies, and crash involvement of roadway characteristics mainly including horizontal alignment and traffic flow on rural/urban arterial roads [35]. Berhanu, 2004 modeled the effect of crashes with the roadway geometry and traffic flows on arterial roadway segments [48]. Zhang, 2005 developed models to estimate the geometric roadway factors mainly include
horizontal and vertical curves and traffic characteristics influencing head-on- crashes on two-lane rural roadways [16]. Oh, 2004 developed prediction models on rural highways for three types of intersections influencing roadway features which include horizontal and vertical curves and also detailed geometric and traffic characteristics [37]. Daniel, 2004 developed crash prediction models on urban arterials with heavy truck volumes on signalized intersections and roadway segments mainly on horizontal and vertical curves. In all these research studies, the negative binomial regression model was preferred over the Poisson regression model and concentrated on a particular segment of roadway or intersection.

2.4 Geometric Influence on Crash Prediction

The roadway geometry may also influence the potential for crashes. Some of these geometric features include horizontal and vertical curvature, sight distance, superelevation, lane width, shoulder width, median width, and number of lanes [17,37]. The combinations of these variables provide visual queues for the driver of the vehicle. The more challenging or difficult queues to process create opportunities for driver error, and loss of vehicle control which often result in crashes. The roadway and traffic variables used in the previous studies to model total crashes and truck related crashes using negative binomial regression model along with the suggested results are discussed in Table 2.4 and 2.5.
Table 2.5. Total Crashes Modeled Using the Negative Binomial Regression Model Developed for Specific Geometric Segment.

<table>
<thead>
<tr>
<th>Objective of Study</th>
<th>Variables</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>To develop macrolevel crash prediction models between crash occurrence and traffic and geometric features on rural signalized intersections.</td>
<td>53 explanatory variables were considered which includes different traffic volumes, geometric elements and driver characteristics. Geometric elements include vertical and horizontal curves, sight distances, lane and shoulder widths, driveway intensity, and signal control.</td>
<td>It was concluded that traffic flow variables significantly affected the traffic safety for all the intersection types. The effect of geometric variables on crashes varied across intersection types [37].</td>
</tr>
<tr>
<td>To predict models relating total crashes with road environment and traffic flows on arterial roads for undivided and divided roads.</td>
<td>The variables include road segment, traffic crashes, roadway geometry and traffic data.</td>
<td>Results proved that wider roadway width, wider and paved sidewalks and raised curbs reduce the traffic crashes. Increase in ADT and pedestrian traffic tends to increase crash frequencies [48].</td>
</tr>
<tr>
<td>To develop crash prediction models on rural two lane segments and intersections.</td>
<td>Variables include traffic, horizontal and vertical alignments, lane and shoulder widths, roadside hazard rating, channelization and number of driveways.</td>
<td>Results proved that segment crashes depended significantly on roadway variables and intersection crashes primarily on traffic. Adjustment factors for different regions were developed [49].</td>
</tr>
<tr>
<td>To model crash frequencies, crash involvement of roadway and traffic flow on rural/urban arterial roads.</td>
<td>Variables include log of segment length, log of AADT per lane, degree of horizontal curve, shoulder, median and lane width.</td>
<td>Heavy traffic volume, speeding, narrow lane width, more number of lanes, narrow shoulder and median widths on urban section tends to increase crash involvement [35].</td>
</tr>
<tr>
<td>To develop crash prediction models to model the effect of geometric features on head-on-crashes on two lane roads.</td>
<td>Variables that influence the head-on-crashes are speed limit, sum of absolute change rate of horizontal curvature, degree of curvature (max) and sum of absolute change rate of vertical curvature.</td>
<td>The results suggest that reduction in number and degree of horizontal and vertical curves reduce the head-on-crashes rather than widening of pavement [16].</td>
</tr>
<tr>
<td>To develop crash prediction models for multilane roads for total and severe crashes.</td>
<td>Variables include traffic flow, road geometry, sight distance, pavement surface friction and rain precipitation.</td>
<td>Results show that for horizontal curves, length, curvature and AADT were significant [33].</td>
</tr>
</tbody>
</table>
Table 2.5. Total Crashes Modeled Using the Negative Binomial Regression Model Developed for Specific Geometric Segment. (Cont.)

| To model the effect of road network infrastructure and demographic changes with total crashes and fatalities. | Variables include log of mean of number of lanes, lane width, outer shoulder width, and horizontal deflection angle, number of horizontal and vertical curves per mile. | Increase in lane widths and decrease in outer shoulder widths increase the crashes. The median width, shoulder width, vertical, and horizontal curvature were not significant [42]. |
| To develop crash prediction models to estimate the nature of overdispersion in motor vehicle crashes on roadway geometric variables. | Variables include AADT, median, and shoulder width, speed, absolute change in grade and other categorical variables. | Results show that in case of overdispersion, the model shows the ideal way of crash prediction [40]. |

Table 2.5 and 2.6 show that there are few truck related crash analysis based on the roadway features. As discussed earlier on crash trends, the traffic flow on two lanes, which is characteristic of most rural roads are an area of interest in the present study [19]. In this case roadways with vertical curves tend to experience more crashes near locations with high grade values and small horizontal radii. In many of the research papers, these vertical alignment variables are not statistically significant as in case of Daniel, 2002 [50]. In case of research finding of Daniel, 2004, the truck crashes are concentrated on crest curves at intersections, and on sag curves are not studied [15]. The challenging part is the sample size required to model truck crashes with roadway geometric and traffic data using negative binomial regression models. Section 2.5 discusses the sample sizes considered for modeling similar kind of crash prediction models.
Table 2.6. Truck-Related Crashes Modeled Using the Negative Binomial Regression Model.

<table>
<thead>
<tr>
<th>Objective of Study</th>
<th>Variables</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>To develop crash prediction models to identify the factors that impact the truck</td>
<td>Variables include segment length, AADT, degree of horizontal curvature, length of curve, crest curve grade rate, length of vertical curve, posted speed, pavement width, number of lanes, interchanges within segment and signalized intersection within the segment.</td>
<td>Results show that segment length, number of lanes, pavement width and posted speed were significant for the full model. For the reduced model number of lanes, pavement width and vertical curve rate were significant [15].</td>
</tr>
<tr>
<td>crash occurrence on urban arterials with signalized intersections and roadways</td>
<td>The variables considered in the model include AADT per lane, horizontal curvature (= 0 if curve&lt;=1), vertical grade in the form of categorical variable (=0 if grade &lt;=2), deviation from ideal shoulder with of 12 ft per direction.</td>
<td>It was concluded that NB model performed the best in the estimating the truck crash frequency of road sections with zero truck crash involvement [14].</td>
</tr>
<tr>
<td>To establish the relationship between truck accidents and geometric design of</td>
<td>Variables include segment length, AADT, degree of horizontal curve, length of horizontal curve, crest curve grade rate, length of vertical curve, posted speed, number of lanes, number of signals, number of interchanges within the segment and pavement width.</td>
<td>The results proved that segment length, AADT, length of horizontal curve, crest curve grade rate, length of vertical curve, number of lanes, number of signals within the segment and pavement width are significant using Poisson regression model. Using negative binomial regression model on segment length was significant [50].</td>
</tr>
<tr>
<td>road sections using Poisson, Zero-Inflated and negative binomial regression models.</td>
<td>Variables mainly include percentage of trucks, truck lane restriction, AADT per lane, and free flow speed.</td>
<td>Truck lane restriction variable in the model was negative but was insignificant [51].</td>
</tr>
</tbody>
</table>
2.5 Sample Sizes

For developing any kind of model, data are required and heavily influenced by the sample size. The question of how much data are sufficient for the development of negative binomial regression model are based on guidance from previous studies. The results to this question are shown in Table 2.7. In Table 2.7, the model sample size, length of study period, and research objectives with concluded results are given and typically three to five years of data are used by majority of the researchers. Other trends show common sample size to be greater than 1,000 crash records. Many research studies have developed models with less than 100 observations to greater than 10,000 observations.

Apart from these prediction models, there are few techniques which enhance the predictive capabilities in the model. These techniques include Empirical Bayes and Full Bayes statistical methodologies and are described in the following section.
Table 2.7. Sample Sizes for Modeling Crashes Using the Negative Binomial Regression Model.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Study period</th>
<th>Research Objectives</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>8,263 homogeneous road segments</td>
<td>5 years</td>
<td>To establish the relationship between truck accidents and geometric design of road sections using Poisson, Zero-Inflated and negative binomial regression models.</td>
<td>It was concluded that negative binomial model performed the best in the estimating the truck crash frequency of road sections with zero truck crash involvement [14].</td>
</tr>
<tr>
<td>54 study sites, 3,364 traffic crashes</td>
<td>2 years</td>
<td>To predict models relating total crashes with road environment and traffic flows on arterial roads for undivided and divided roads.</td>
<td>Results proved that wider roadway width, wider and paved sidewalks and raised curbs reduce the traffic crashes. Increase in ADT and pedestrian traffic tends to increase crash frequencies [48].</td>
</tr>
<tr>
<td>136 three-legged intersections, 124 four-legged intersections on multilane and 100 four-legged intersections on rural areas</td>
<td>6 years</td>
<td>To develop macrolevel crash prediction models between crash occurrence and traffic and geometric features on rural signalized intersections.</td>
<td>It was concluded that traffic flow variables significantly affected the traffic safety for all the intersection types. The effect of geometric variables on crashes varied across intersection types [37].</td>
</tr>
<tr>
<td>1,331 segments</td>
<td>5 and 3 years</td>
<td>To develop crash prediction models on rural two lane segments and intersections.</td>
<td>Results proved that segment crashes depended significantly on roadway variables and intersection crashes primarily on traffic. Adjustment factors for different regions were developed [49].</td>
</tr>
<tr>
<td>655 highway segments</td>
<td>6 years</td>
<td>To develop crash prediction models to model the effect of geometric features on head-on-crashes on two lane roads.</td>
<td>The results suggest that reduction in number and degree of horizontal and vertical curves reduce the head-on-crashes rather than widening of pavement [16].</td>
</tr>
<tr>
<td>102 counties</td>
<td>4 years</td>
<td>To model the effect of road network infrastructure and demographic changes with total crashes and fatalities.</td>
<td>Increase in lane widths and decrease in outer shoulder widths increase the crashes. The median width, shoulder width, vertical, and horizontal curvature were not significant [42].</td>
</tr>
</tbody>
</table>
Table 2.7. Sample Sizes for Modeling Crashes Using the Negative Binomial Regression Model. (Cont.)

<table>
<thead>
<tr>
<th>38 counties.</th>
<th>2 years</th>
<th>To develop crash prediction models to estimate the nature of overdispersion in motor vehicle crashes on roadway geometric variables.</th>
<th>Results show that in case of overdispersion, the model shows the ideal way of crash prediction [40].</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,916 crashes, 265 road segments.</td>
<td>5 years</td>
<td>To develop crash prediction models for multilane roads for total and severe crashes.</td>
<td>Results show that for horizontal curves, length, curvature and AADT were significant [33].</td>
</tr>
<tr>
<td>117, 258, 1101, 1359 observations for unified, signal, roadway and combined segment models respectively.</td>
<td>3 years</td>
<td>To develop crash prediction models to identify the factors that impact the truck crash occurrence on urban arterials with signalized intersections and roadways segments.</td>
<td>Results show that segment length, number of lanes, pavement width and posted speed were significant for the full model. For the reduced model number of lanes, pavement width and vertical curve rate were significant [15].</td>
</tr>
<tr>
<td>39 observations.</td>
<td>2 years</td>
<td>To identify factors that contributes to truck crashes on roadways influenced by signalized intersections.</td>
<td>The results proved that segment length, AADT, length of horizontal curve, crest curve grade rate, length of vertical curve, number of lanes, number of signals within the segment and pavement width are significant using Poisson regression model. Using negative binomial regression model on segment length was significant [50].</td>
</tr>
</tbody>
</table>

Note – Study period refers to number of years for which data were collected and used for each respective study.
Empirical Bayes and Full Bayes Techniques

Empirical Bayes and Full Bayes techniques are used to enhance the predictive power of the model. In the Empirical Bayes setting, ordinary Bayesian analysis techniques are developed for parameter estimates of the prior distribution and weighted factors are used to obtain posterior estimates [32,52]. This technique combines the information of the prior groups of similar characteristics to interpret the information of the specific study under consideration [52]. The Empirical Bayes paradigm has long been used for such applications as estimating a vector of means or binomial proportions, modeling traffic crash-flow relationships for intersections [53], ranking sites for safety investigation at signalized intersections and highways [54], effective incident management [54], and ranking high traffic crash risks sites [56]. The major limitation of this method is that it does not take into account the source of variability due to the substitution of estimates for prior parameters.

Many researchers have found added benefits of incorporating the Full Bayes statistical approach in lieu of Empirical Bayes [56,57,58]. Full Bayesian framework uses the Markov Chain Monte Carlo (MCMC) algorithm and full posterior inference for parameters in complex model settings [57]. MCMC is used to sample posterior probability distribution, an approach that is better than Empirical Bayes technique which underestimates uncertainties in the model [57,57]. Bayesian estimates are well suited for crash rate prediction when large number of crashes are within a region; otherwise, few crashes in a region may cause inflated variance estimates [57]. Hierarchical modeling is done in three stages. The first stage is to setup the model’s joint probability distribution for the estimated and predicted quantities. The second stage is the calculation and
interpretation of posterior distribution from the model in the previous stage. The third stage is the evaluation of the fit of model for the resulting posterior distribution [59,57].

2.6 Summary

In the present study, the negative binomial regression models using Gibbs sampling and Full Bayes are used to model truck crashes on two-lane rural highways in Ohio. The previous research studies on truck crashes develop negative binomial regression models or zero inflated models to develop prediction models [15]. Most of the truck-related crash studies are concentrated on urban roadways and mostly at the intersections [33,50,15]. The vertical crest curves are considered to be an influential factor in truck-related crashes but are not statistically significant in most of the cases and there are no findings on truck-related crashes on vertical sag curves. For this reason, the present study is concerned with truck-related crashes on geometric and traffic characteristics on two-lane rural vertical curves including both crest and sag curves. The next Chapter III discusses the validation of the dataset and finally presents the set of data used in the process of model development and the methodology used is illustrated in Chapter IV.
CHAPTER III

STUDY DATA

3.1 Given Data

Crash data for The State of Ohio are provided by the ODPS from 2002 to 2006 and roadway geometric data are provided by the ODOT. The information from these two sources are grouped by NLF id and are given in the access format. NLF id is route identification field which is a text string describing jurisdiction, county, type of route, route number and other route related details. For example SRICUS00042**C where S indicates state jurisdiction, RIC indicates Richland county, US00042 indicates the US route 42, * indicates regular route and C indicates cardinal direction of travel. The structure of the given dataset is described in Figure 3.1.

Figure 3.1. Structure of the Given Dataset.
3.2 Data Validation

In order to validate the given dataset, log points or station miles which describe the geometric roadway location are used in concert with Geographic Information System (GIS) maps. The data provided in the GIS maps are posted on the ODOT website in the form of shape files depicting counties, State Routes, US routes, Interstates, townships and cities in Ohio. These files are joined by ARC Map software to form a single map of Ohio. Using this map, the location of the particular route is identified and the spatial information including the log point and roadway geometry are provided in the map. The spatial information in the map are matched with the given dataset. Routes such as state routes, interstate routes and US routes are also chosen along with road contours to facilitate data filtration. Three Quality Control Quality Assurance (QCQA) checks are performed using data tables provided in these maps.

- First stage checks are performed on all horizontal curves.
- Second stage checks are conducted on all angle points given in the data.
- Third stage the checks are conducted on vertical curves.

The following sections from 3.2.1 through 3.2.3 provide more detailed information on the three QCQA checks.

3.2.1 Validation on Horizontal Curves

Validation on horizontal curves is performed by comparing the log points of GIS maps with the log point numbering for each county and tracking the path of each route. The first check is conducted in two steps, step one uses the radius and degree of curvature and
step two validates the physical location of horizontal curves. First step is performed by using the Equation (3.1) as:

\[
D_c = \frac{5,729.58}{R}
\]  

(3.1)

where \( R \) = radius in feet, and

\( D_c \) = degree of curvature [60].

The second step is performed by using ArcMap GIS software. A particular route is selected and the log points are matched with the GIS map. The matched log points are tracked for horizontal curvature. This check validates the existence of horizontal curves per location. The geometric details of the validated dataset are then matched. This is illustrated in the example considered below with the Figure 3.3. In this example State route 2 which passes through Defiance (DEF), Williams (WIL), Fulton (FUL), Lucas (LUC), Ottawa (OTT), Erie (ERI), Cuyahoga (CUY), and Lake (LAK) counties as seen in Figure 3.2.
Figure 3.2. State route 2 in The State of Ohio obtained from ARC Map-GIS.

From the given dataset, route 2 passing through Defiance County with a beginning log point of 5.12 and an ending log point of 5.16 is considered and the horizontal curve of radius associated with crash is 818.5 ft. These data are compared and matched with Arc Map for the same route and county with beginning log point at 5.09 and ending log point at 5.31 as shown in Figure 3.3.
Figure 3.3. Horizontal curve check for route 2 in Defiance County.

The segment of horizontal curve circled in Figure 3.3 is obtained by interpolation of distances on map scale. The details such as speed limit, lanes, lane widths, shoulder widths, median widths, and ADT are matched. The same procedure is followed for state route 2 passing through other counties. Similar checks are conducted on US route 23 and Interstate route 71. These routes are chosen because these routes pass through several counties as shown in Figure 3.4 and 3.5 respectively. The other routes are randomly chosen and spot checked.
Figure 3.4. US Route 23 in The State of Ohio Obtained from ARC Map- GIS.

Figure 3.5. Interstate Route 71 in The State of Ohio Obtained from ARC Map- GIS.
3.2.2 Validation on Angles Points

The angle on the road is the point of intersection of two road segments of different orientation. These are similar to horizontal curves with sharp degree of curvature. In the given dataset, the degree of angle and the direction of the angle which includes left and right are provided for each segment. For these angle points, validation is performed at every single angle point with the directions matched using the GIS maps and are accomplished by selecting routes and then spot checking the locations similar to horizontal curve validation.

3.2.3 Validation on Vertical Curves

The major components of vertical curves are grade, length of curve, and elevation of the curve. The grade or the gradient is designated as plus ($+g_1$) for upgrade and minus ($-g_2$) for downgrade. The vertical curves depending on the gradient are divided into crest and sag curves.

The check on the vertical curves is conducted by using the topography maps along with U.S. Geological Survey (USGS) quadrangle maps provided on the ODOT website [61]. These quadrangle maps which contain the contour elevation are used for the validation of vertical curves. This is accomplished by selecting routes and matching the contour elevation in the quadrangle maps with that of the given dataset. In this study, state routes 8, 53, 81, and 235, US routes 68 and 224 and interstate routes 70 and 77 are validated with the contours provided by ODOT maps and the other routes are spot checked. One example of this validation technique is shown by state route 81, which passes through the cities of Willshire, Elgin, Lima, Lafayette, Ada, Dunkirk, and Patterson. By using the
quadrangle maps as shown in Figure 3.6, and tracking contours along route 81 for The City of Ada, the vertical curve elevations are matched with the given dataset. The log points and elevation based on the contour calculated to the scale are checked to match the given dataset.

![Figure 3.6. Route 81 in the City of Ada in The State of Ohio Quad Map [61].](image)

### 3.2.4 Validation Issues

All the GIS maps used for validation of horizontal curves and angle points are matched correctly with the given data but in case of validation of vertical curves, some of the topography maps did not match with the given data because the data in these maps are not updated. For instance, route 8, 235, 68, 53, 81 and many more are not updated. Validations on these routes are not done accurately due to the limitation of the reference data. A superficial check is done by using latitude and longitude of the GIS map. This
information is applied to vertical curve data and location is tracked using Google earth and accordingly location points are consistent with the dataset.

After the validation, the inappropriate data were discarded. Missing information related to functional classification of roadways are added to the dataset in order to differentiate between rural and urban roadways. Section 3.3 and 3.4 discusses the two parts of the modified data.

3.3 Modified Data – Part I

The modified dataset along with the addition of new variables are shown in Table 3.1.

Table 3.1. Modified data variables.

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable (unit)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key variables</td>
<td>County name</td>
<td>88 counties</td>
</tr>
<tr>
<td></td>
<td>Calendar year</td>
<td>2002 – 2006</td>
</tr>
<tr>
<td>Truck Crashes</td>
<td>Number of truck crashes</td>
<td>Total count of truck crashes on each segment</td>
</tr>
<tr>
<td>Roadway functional classification</td>
<td>2 = rural principal arterial- (other), 6 = rural minor arterial, 7= rural major collector, 8= rural minor collector, 9= rural local, 11= urban principal arterial (other freeways and expressways), 12 = urban principal arterial- (other), 16 = urban minor arterial, 17= urban collector, 19= urban local</td>
<td></td>
</tr>
<tr>
<td>Roadway features</td>
<td>Segment length (mile)</td>
<td>Remained same as the previously provided data</td>
</tr>
<tr>
<td></td>
<td>Lanes</td>
<td>2 to 8 lanes directional</td>
</tr>
<tr>
<td></td>
<td>Median width (ft)</td>
<td>2 ft to &gt;1,000 ft</td>
</tr>
<tr>
<td></td>
<td>Lane width (ft)</td>
<td>Total lane width inclusive of median and surface width without the median</td>
</tr>
<tr>
<td></td>
<td>Shoulder width (ft)</td>
<td>Inner and outer shoulder with for both left and right lanes</td>
</tr>
<tr>
<td></td>
<td>Horizontal width (ft)</td>
<td>Degree of curvature</td>
</tr>
<tr>
<td></td>
<td>Vertical curvature</td>
<td>Degree of slope</td>
</tr>
<tr>
<td></td>
<td>Speed limit</td>
<td>20 mph to 65 mph</td>
</tr>
<tr>
<td>Average Daily Traffic</td>
<td>ADT</td>
<td>Total ADT/year/segment</td>
</tr>
<tr>
<td></td>
<td>Passenger cars ADT/year/segment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Truck ADT/year/segment</td>
<td></td>
</tr>
</tbody>
</table>

Note: All the variables of modified data are not included. For detail set of modified data variables see the Appendix.
The focus of the present study is on two lane rural highways and on vertical curves. Vertical curve data are separated from the entire dataset based on the degree of slope. These data points contain 2 to 8 lanes as mentioned in Table 3.1 and two lane highways are separated from this dataset. These data points which include vertical curve data on two lane highways are considered as a new dataset. In the new dataset, functional classes of 2, 6, 7 and 8 are selected to obtain the rural roadways. These rural roadways are further divided into the crest and sag curves are divided discussed in section 3.3.2.

3.3.1 Criteria for Separation of Vertical Curves

The new dataset with vertical curve data includes upgrade and downgrade segments but not as a curve. In order to form a curve, a new variable is created to determine the difference between two successive segments as shown in Figure 3.7 with yellow colored blocks forming a curve. If the length is less than or equal to 0.04 miles then it is said to form a vertical curve.

![Figure 3.7. Criteria for Forming a Vertical Curve.](image)
3.3.2 Crest and Sag curve Separation

The crest curves are divided into three types as shown in Figure 3.8, where (i) represents crest curve with upgrade segment (+g₁) followed by a downgrade segment (-g₂), (ii) represents crest curve with upgrade segment (+g₁) followed by another upgrade segment (+g₂) with |g₁|>|g₂| and (iii) represents crest curve with downgrade segment (-g₁) followed by another downgrade segment (-g₂) with |g₂|>|g₁| [63].

![Diagram of Crest Curves](image)

NOTE: These are referred from AASHTO [63].

Figure 3.8. Types of Crest Curves.

Similarly the sag curves are also divided into three types as shown in Figure 3.9, where (i) represents sag curve with downgrade segment (-g₂) followed by an upgrade segment (+g₁), (ii) represents sag curve with downgrade segment (-g₁) followed by another downgrade segment (-g₂) with |g₁|>|g₂|, and (iii) represents sag curve with upgrade segment (+g₁) followed by another upgrade segment (+g₂) with |g₂|>|g₁| [63].
Figure 3.9. Types of Sag Curves.

Based on these distinctions, the new dataset is divided into crest and sag curves. The structure of the study data concerned is given in Figure 3.10.

Figure 3.10. Structure of Study Area Concern.
In addition to the variables mentioned earlier, new variables are calculated which describe the rate of vertical curvature, grade rate change only in case of crest and sag curves.

### 3.3.3 Calculations Related to Vertical Curves

**Rate of Vertical Curvature**

The algebraic difference between the gradients is calculated by using the Equation (3.2).

\[ A = |g_2 - g_1| \]  

(3.2)

where \( A \) = algebraic difference between gradients in percent,

\( g_2 \) = gradient of second segment in curve in percent, and

\( g_1 \) = gradient of first segment in curve in percent.

The rate of vertical curvature \( (K) \) is measured by dividing length of the curve in feet \( (L) \) by the algebraic difference \( (A) \) between the gradients given in Equation (3.3). This represents the horizontal distance required to effect a 1-percent change in the gradient along the curve and is constant along the length of curve [63].

\[ K = \frac{L}{A} \]  

(3.3)

where \( K \) = rate of vertical curvature,

\( L \) = total length of vertical curve in feet, and

\( A \) = algebraic difference in gradients in percent [62,63].
**Grade Change for the length of the curve**

Gradient of a curve is measured in percent, that is, the number of feet of rise or fall in a 100-foot horizontal road. The rise or fall for the entire length of the curve also termed as grade change for the length of the curve is given by the algebraic difference between the gradients multiplied by the total length of the curve given in Equation (3.4).

\[ G_C = A \times L \]  

(3.4)

where \( G_C \) = Grade change for the length of the curve,

\( A \) = algebraic difference in gradients in percent, and

\( L \) = total length of vertical curve in feet [63]

A curve is said to be homogenous when the traffic volume is same throughout the curve. Dummy variables are created to test the homogeneity of the curves and non-homogenous curves are discarded. Truck crashes for the entire curve are calculated by the summation of truck crashes on each side of the curve. The new dataset which is focused on two lane rural highways with only the vertical curve data is modified in section 3.4 according to American Association of State Highway and Transportation Officials (AASHTO) standards.

3.4 Modified Data – Part II

The second part of the section describes modified data of all the geometric variables included in the dataset according to AASHTO standards and relevant modifications. The roadway geometric variables include posted speed, vertical alignment including maximum grade and rate of vertical curvature, surface width, shoulder width and total
width are discussed based on AASHTO standards for rural two lane vertical curves. Section 3.5 discusses the descriptive statistics of final dataset along with crest and sag data which are used to develop the negative binomial regression model.

3.4.1 Roadway Related Variables

The roadway geometric variables included in the dataset are posted speed limit, vertical alignment, surface width, shoulder width, and total width for two lane rural roadways with vertical curves. These rural roadways are discussed based on the each functional class of rural roadway mentioned in section 2.9.

Posted Speed limit

The posted speed limit for rural roadways depends on the design traffic volume and the roadway terrain. For different roadway classification, the minimum speed limit of 20 mph on rural local roads to a maximum of 65 mph on rural arterials and freeways. In the given dataset, the speed ranges from 25 mph to 55 mph on rural roads for both the crest and sag vertical curves [63].

Vertical Alignment

Maximum grade and the rate of vertical curvature define the vertical alignment of roadway. The maximum grade values according to AASHTO for local and collector roads range from 5 % to 12 %, and for rural arterials and freeways, it ranges from 3 % to 6 % for the speed range of 25 to 65 mph [63]. In the given dataset, the grade values are discarded if they exceed the maximum grade AASHTO standards.

The rate of vertical curvature for the speed range of 25 to 65 mph range from 289 ft per % to 1865 ft per % for the crest curves and for the sag curves, it ranges from 10 ft per %
to 181 ft per % [63]. In the modified dataset, the calculated K values are discarded if they exceed 2,000 ft per % for crest vertical curves and 200 ft per % for sag vertical curves.

**Surface Width, Shoulder Width and Total Width**

The surface width of roadway is the traveled way width, while the total width of roadway is the surface width plus the shoulder width on both sides. According to AASHTO on rural local and collector roads, the minimum surface and total width are 20 ft and 24 ft respectively, with a minimum shoulder width of 2 ft on each side. On rural arterials, the minimum surface width is 22 ft with minimum shoulder width of 4 ft on each side and on rural freeways for design speeds greater 50 mph, the usable shoulder widths on each side should be at least 10 ft [63]. In the modified dataset, the surface widths less than 20 ft are discarded.

3.5 Final Dataset

The final dataset with all the modification completed is presented in Tables 3.2 through 3.4. Table 3.2 presents the descriptive statistics of the entire modified data, and Tables 3.3 and 3.4 presents the descriptive statistics of crest and sag curve data respectively. Using this modified dataset and the methodology used in Chapter IV, negative binomial regression models are developed in Chapter V.
Table 3.2. Descriptive Statistics for Entire Modified Dataset

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate of Vertical curvature (k in ft per %)</td>
<td>147.16</td>
<td>229.78</td>
<td>4.22</td>
<td>1,927.2</td>
<td></td>
</tr>
<tr>
<td>Grade Change for the Curve length (ft-%)</td>
<td>3.04</td>
<td>2.58</td>
<td>0.07</td>
<td>24.64</td>
<td></td>
</tr>
<tr>
<td>Speed Limit (mph)</td>
<td>52.26</td>
<td>6.07</td>
<td>25</td>
<td>55</td>
<td>25 – 45 mph – 20 %</td>
</tr>
<tr>
<td>Total Width (ft)</td>
<td>26.47</td>
<td>4.49</td>
<td>20</td>
<td>54</td>
<td>&lt; 24 ft – 21 %</td>
</tr>
<tr>
<td>Surface Width (ft)</td>
<td>22.34</td>
<td>2.57</td>
<td>20</td>
<td>45</td>
<td>20 – 22 ft – 62 %</td>
</tr>
<tr>
<td>Average Shoulder Width (ft)</td>
<td>2.07</td>
<td>2.36</td>
<td>0</td>
<td>11</td>
<td>≤ 4 ft – 87 %</td>
</tr>
<tr>
<td>Truck Crashes</td>
<td>0.11</td>
<td>0.38</td>
<td>0</td>
<td>4</td>
<td>&gt; 4 ft – 13 %</td>
</tr>
<tr>
<td>ADT of Trucks</td>
<td>138.96</td>
<td>194.62</td>
<td>10</td>
<td>1,910</td>
<td></td>
</tr>
<tr>
<td>ADT of Passenger Cars</td>
<td>1,842.70</td>
<td>1,794.02</td>
<td>60</td>
<td>16,500</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3. Descriptive Statistics for the Crest Vertical Curve Data

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate of Vertical curvature (k in ft per %)</td>
<td>137.76</td>
<td>213.83</td>
<td>4.22</td>
<td>1,927.2</td>
<td></td>
</tr>
<tr>
<td>Grade Change for the Curve length (ft-%)</td>
<td>3.23</td>
<td>2.80</td>
<td>0.10</td>
<td>24.64</td>
<td></td>
</tr>
<tr>
<td>Speed Limit (mph)</td>
<td>52.02</td>
<td>6.27</td>
<td>25</td>
<td>55</td>
<td>25 – 45 mph – 20 %</td>
</tr>
<tr>
<td>Total Width (ft)</td>
<td>26.11</td>
<td>4.37</td>
<td>20</td>
<td>54</td>
<td>&lt; 24 ft – 25 %</td>
</tr>
<tr>
<td>Surface Width (ft)</td>
<td>22.40</td>
<td>2.51</td>
<td>20</td>
<td>40</td>
<td>20 – 22 ft – 61 %</td>
</tr>
<tr>
<td>Average Shoulder Width (ft)</td>
<td>1.85</td>
<td>2.28</td>
<td>0</td>
<td>11</td>
<td>≤ 4 ft – 89 %</td>
</tr>
<tr>
<td>Truck Crashes</td>
<td>0.10</td>
<td>0.36</td>
<td>0</td>
<td>3</td>
<td>&gt; 4 ft – 11 %</td>
</tr>
<tr>
<td>ADT of Trucks</td>
<td>128.50</td>
<td>188.02</td>
<td>10</td>
<td>1,910</td>
<td></td>
</tr>
<tr>
<td>ADT of Passenger Cars</td>
<td>1,658.91</td>
<td>1,638.37</td>
<td>90</td>
<td>12,400</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.4. Descriptive Statistics for the Sag Vertical Curve Data

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate of Vertical curvature (k in ft per %)</td>
<td>154.68</td>
<td>249.56</td>
<td>13.89</td>
<td>1,900.8</td>
<td></td>
</tr>
<tr>
<td>Grade Change for the Curve length (ft-%)</td>
<td>2.84</td>
<td>2.27</td>
<td>0.07</td>
<td>21.58</td>
<td></td>
</tr>
<tr>
<td>Speed Limit (mph)</td>
<td>52.72</td>
<td>5.56</td>
<td>25</td>
<td>55</td>
<td>25 – 45 mph – 16 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>45 – 55 mph – 84 %</td>
</tr>
<tr>
<td>Total Width (ft)</td>
<td>27.04</td>
<td>4.39</td>
<td>22</td>
<td>44</td>
<td>&lt; 24 ft – 37 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24 – 26 ft – 22 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&gt; 26 ft – 41 %</td>
</tr>
<tr>
<td>Surface Width (ft)</td>
<td>22.33</td>
<td>2.56</td>
<td>20</td>
<td>36</td>
<td>20 – 22 ft – 61 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&gt; 22 ft – 39 %</td>
</tr>
<tr>
<td>Average Shoulder Width (ft)</td>
<td>2.36</td>
<td>2.41</td>
<td>0</td>
<td>10</td>
<td>≤ 4 ft – 86 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&gt; 4 ft – 14 %</td>
</tr>
<tr>
<td>Truck Crashes</td>
<td>0.10</td>
<td>0.35</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>ADT of Trucks</td>
<td>148.70</td>
<td>192.92</td>
<td>10</td>
<td>1,820</td>
<td></td>
</tr>
<tr>
<td>ADT of Passenger Cars</td>
<td>1,993.00</td>
<td>1,914.07</td>
<td>60</td>
<td>16,500</td>
<td></td>
</tr>
</tbody>
</table>

Finally after the three sets of validation performed on the given dataset, the vertical curve dataset contains 1,935 vertical curve segments with 205 truck crashes. The vertical curve dataset includes 1,065 crest curve segments with 114 truck crashes and 870 sag curve segments with 91 truck crashes. These data along with the methodology in Chapter IV, are used to develop negative binomial regression models and Bayesian models for two lane rural highways in Chapter V.
CHAPTER IV
METHODOLOGY

Crash counts, which are discrete values, are used in Poisson and negative binomial regression crash prediction models. The methodology of these models is discussed in sections 4.1 and 4.2. The estimation of the dispersion parameter is discussed in section 4.3 along with derivation proved by several researchers. The methodology of negative binomial model using Gibbs sampling is discussed in section 4.4. Sections 4.5 through 4.7 outline the methodology used for testing the models. Sections 4.5 through 4.7 discuss the criteria for selection of variables and for testing the models.

4.1 Poisson Regression Model

The Poisson distribution is typically used to model count data including crashes. Count data cannot be modeled as continuous data by applying standard least-square regression because regression models predict values that are non-integers or negative values, both of which are inconsistent with count data. These limitations make standard regression analysis inappropriate for modeling count data without modifying dependent variables [32,33,26].
4.1.1 Probability Density Function of Poisson Regression Model

In the Poisson regression model the response variable follows a Poisson distribution, with the following probability density function shown in Equation (4.1) as:

\[
P(y_i \mid \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}
\]  (4.1)

where \( y_i \) = independent Poisson outcomes denoting the number of truck-related crashes observed during a period of time,

\( \mu_i \) = mean of the number of truck-related crashes, and

\( i = \) ranges from 1 to N observations [35,36,39,44,64,16].

In this study \( y_i \) denotes the number of truck-related crashes considered by the function of regressor or explanatory variables \( X_i \) for \( i \) observations.

Thus, the Poisson distribution accordingly is given in Equation (4.2) as:

\[
P(Y = y_i) = \frac{e^{-\mu_i(X_i, \beta)} [\mu_i(X_i, \beta)]^{y_i}}{y_i!}
\]  (4.2)

where \( y_i \) = independent Poisson outcomes denoting the number of truck-related crashes observed during a period of time,

\( \mu_i \) = mean of the number of truck-related crashes,

\( X_i \) = explanatory variables responsible for truck-related crashes,

\( \beta \) = parameter explaining the explanatory variable, and

\( i = \) range from 1 to N observations [35,48].
In the Poisson distribution, the important characteristic is that the variance of response variable \( y_i \) is equal to its mean \( \mu_i \) \( [33,39,44,65,40] \); that is,

\[
\text{Var}(y_i) = E(y_i) = \mu_i
\]

(4.3)

where \( \mu_i \) = mean of the response variable \( y_i \),

\[
\text{Var}(y_i) = \text{variance of the response variable } y_i, \text{ and }
\]

\[
E(y_i) = \lambda_i = \text{expected value of the response variable } y_i [33,57].
\]

### 4.1.2 Model Form

The general form of the Poisson regression model in exponential form is given by

\[
\lambda_i = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}
\]

(4.4)

where \( \lambda_i \) = has a Poisson distribution with expected mean number of truck-related crashes on vertical curves,

\( \beta_0 \) = constant term,

\( \beta_1,\ldots,\beta_n \) = estimated parameters in vector form, and

\( x_1,\ldots,x_n \) = explanatory variables on rural two lane vertical curve \( i [35,36,37] \).

### 4.1.3 Overdispersion

When the conditional variance of the response variable \( y_i \) is greater than the conditional mean of the response variable \( y_i \), then the model is said to have overdispersion. When the conditional variance of the response variable \( y_i \) is less than the conditional mean of the response variable \( y_i \), then it is said to have underdispersion \([36,66,39]\). The
conditional variance of $y_i$ is denoted by $\text{Var}(y_i | x)$ or $\sigma_i^2$ and the conditional mean of $y_i$ is denoted by $E(y_i | x)$ or $\mu_i$. For most of the crash data, overdispersion is present and the mean is not equal to the variance of the outcome. The overdispersion in crash data may be as a result of missing relevant variables, as well as uncertainty in regressor variables with a larger variance in the response variable [67,40]. When the Poisson regression model is used in the presence of overdispersion, the maximum likelihood parameter estimates are consistent, but the estimation of the variance of these parameters is inconsistent [67]. To rectify this overdispersion, the Poisson regression model is replaced by the negative binomial regression model [35,36,33,40,43].

4.2. Negative Binomial Regression Model

In the negative binomial regression model, the mean and the variance are not the same and vehicle crashes are better represented by a negative binomial distribution [35,36,33,40,43].

4.2.1 Probability Density Function of Negative Binomial Regression Model

The probability distribution of the negative binomial regression model is given by Equation (4.5) as:

$$P(Y = y_i | \epsilon_i) = \frac{\exp[-\lambda_i \exp(\epsilon_i)]\lambda_i^{y_i}}{y_i!}$$

(4.5)

where $y_i = \text{number of truck-related crashes on vertical curve } i \text{ over a time period}$, $\epsilon_i = \text{error term}$, and
\( \lambda_i \) = expected value of the mean number of truck-related crashes on vertical curve \( i \)[35,36,40].

To obtain the unconditional distribution of \( y_i \), the integral of Equation (4.5) is taken, resulting in the following distribution in Equation (4.6),

\[
P(Y = y_i \mid x_i) = \frac{\Gamma(y_i + \theta)}{y_i! \Gamma(\theta)} \left[ \frac{\theta}{\theta + \lambda_i} \right]^\theta \left[ \frac{\lambda_i}{\theta + \lambda_i} \right]^{-\theta} \]

For \( \theta > 0, \lambda_i > 0 \) (4.6)

where \( y_i \) = number of truck-related crashes on vertical curve \( i \) over a time period,

\( \lambda_i \) = expected value of the mean number of truck-related crashes on vertical curve \( i \),

\( \Gamma() \) = gamma distribution function, and

\( \theta \) = dispersion parameter [35,48,33,66,69,44,52,42,37,64,16].

When dispersion parameter \( \theta \) is equal to zero, this results in Poisson distribution [35,70,71,72,73]. In Equation (4.6) by replacing \( \theta \) with \( \alpha^{-1} \), the probability in terms of inverse dispersion parameter \( \alpha \) is written in Equation (4.7) as:

\[
P(Y = y_i \mid x_i) = \frac{\Gamma(y_i + \alpha^{-1})}{y_i! \Gamma(\alpha^{-1})} \left[ 1 + \alpha \lambda_i \right]^{-1/\alpha} \left[ \frac{\alpha \lambda_i}{1 + \alpha \lambda_i} \right]^{y_i} \]

(4.7)

where \( y_i \) = number of truck-related crashes on vertical curve \( i \) over a time period,

\( \lambda_i \) = expected mean number of truck-related crashes on vertical curve \( i \),

\( \Gamma() \) = gamma distribution function, and

\( \alpha \) = inverse dispersion parameter [66,68,44].
The gamma distribution function has two parameters, the shape parameter $a$ and the scale parameter $b$, represented by $\text{Gamma}(a,b)$. The density function of the gamma distribution with parameters $a$ and $b$ is given in two forms, the normal form given in Equation (4.8) and the inverted form given in Equations (4.9):

$$f(x) = \frac{1}{\Gamma(b)a^b}x^{b-1}e^{-x/a}, x < 0, a, b > 0$$ (4.8)

$$[x | a, b] = \text{Gamma}(x | a, b) = \frac{a^b}{\Gamma(b)}x^{b-1}e^{-ax}$$ (4.9)

where $a =$ scale parameter of the gamma distribution function, $b =$ shape parameter of the gamma distribution function, and $\Gamma () =$ gamma function [74].

The negative binomial distribution is also represented as given in Equation (4.10) as:

$$y_i \sim \text{NB}(\lambda_i, \theta)$$ (4.10)

where $y_i =$ number of truck-related crashes on vertical curve $i$, $\lambda_i =$ expected mean of the number of truck-related crashes, and $\theta =$ dispersion parameter [33,40].

Equation (4.10) indicates that the unconditional distribution of the response variable $y_i$ follows negative binomial distribution with expected mean number of truck-related crashes $\lambda_i$ on vertical curve $i$ and dispersion parameter $\theta$. When the dispersion parameter $\theta$ tends to zero, then the distribution represents Poisson distribution [35,33,40].
4.2.2 Negative Binomial Regression Model Form

The negative binomial regression model is given in Equation (4.11) as

\[ y_i = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n + \varepsilon_i) \] (4.11)

where \( y_i \) = number of truck-related crashes on vertical curve \( i \), with negative binomial distribution

conditional on \( \varepsilon_i \),

\( \beta_0 \) = constant term,

\( \beta_1, \ldots, \beta_n \) = the estimated parameters in vector form,

\( x_1, \ldots, x_n \) = the explanatory variables influencing truck crashes on vertical curve \( i \),

\( \varepsilon_i \) = the error term associated with the variance of \( y_i \), and

\( \exp(\varepsilon_i) \) = has a standard gamma distribution with mean 1 and variance

\( \sigma^2 \) [35,36,39,44,40,42,37,64].

The relationship between the mean and the variance of response variable in the negative binomial regression model is given in Equations (4.12) as:

\[ \text{Var}(y_i) = \lambda_i + \theta \lambda_i^2 \] (4.12)

where \( \theta \) = dispersion parameter,

\( \lambda_i \) = expected mean number of truck-related crashes on vertical curve \( i \), and

\( \text{Var}(y_i) \) = variance of number of truck-related crashes \( y_i \)

[35,36,48,33,68,69,39,44,37,64,16].

From Equation (4.12), variance over the mean is called the overdispersion rate and is given in Equation (4.13) as:
\[
\frac{Var(y_i)}{\lambda_i} = 1 + \theta \lambda_i
\]  
(4.13)

where \( \theta \) = dispersion parameter,

\( \lambda_i \) = expected mean number of truck-related crashes on vertical curve \( i \), and

\( Var(y_i) \) = variance of number of truck-related crashes \( y_i \)

[35,36,48,33,68,69,39,44,37,64,16].

If dispersion parameter \( \theta \) is zero, then the variance of number of truck-related crashes would be equal to expected mean number of truck-related crashes; that is, the negative binomial regression model changes to the Poisson regression model. The negative binomial regression model works efficiently only if \( \theta \) is significantly different from zero [35,36,48,33].

The parameters of the negative binomial regression model can be effectively estimated by the maximum likelihood parameter with corresponding likelihood function shown in Equation (4.14) as:

\[
L(\lambda_i) = \frac{N}{\Pi_{i=1}^{N}} \frac{\Gamma(y_i + \theta)}{y_i \Gamma(\theta)} \left[ \frac{\theta}{\theta + \lambda_i} \right]^\theta \left[ \frac{\lambda_i}{\theta + \lambda_i} \right]^{y_i}
\]  
(4.14)

where \( \lambda_i \) = expected value of the mean number of truck-related crashes on vertical curve \( i \),

\( y_i \) = number of truck-related crashes on vertical curve \( i \) over a time period \( t \),

\( N \) = total number of vertical curves,

\( \Gamma() \) = gamma function,

\( \theta \) = dispersion parameter [35,48,33,66,68,42].
The likelihood function is maximized to obtain the estimated coefficients. The derivation of the maximum likelihood function for the dispersion parameter as derived previously by researchers is shown in section 4.3.

4.3 Maximum Likelihood Estimation For negative binomial Dispersion Parameters

Earlier researchers [74,52,74] have applied the derivation of the maximum likelihood estimation to estimate the negative binomial dispersion parameter. The log-likelihood function of negative binomial regression model, which is used to estimate the dispersion parameter, is given by the Equation (4.15) as:

$$l(\alpha, \lambda_i) = \frac{1}{N} \left[ \sum_{i=1}^{N} \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})} + \bar{y} \log \{\lambda_i\} - (\bar{y} + \alpha^{-1}) \log \{1 + \alpha \lambda_i\} \right]$$  \hspace{1cm} (4.15)$$

where $N =$ total number of vertical curves,
$\alpha =$ inverse dispersion parameter,
$\lambda_i =$ expected mean number of truck-related crashes with sample size $n$,
$\Gamma(\cdot) =$ gamma function,
$y_i =$ number of truck-related crashes on vertical curve $i$ over a time period, and
$\bar{y} =$ observed mean number of truck-related crashes [66,74,52,37,74,64].

Equation (4.16), which is the simplified log-likelihood function without the gamma function, is obtained from Equation (4.15) and is given as:

$$l(\alpha, \lambda_i) = \frac{1}{N} \left[ \sum_{i=1}^{N} \sum_{\kappa=0}^{y_i-1} \log \{1 + \alpha \kappa\} + \bar{y} \log \{\lambda_i\} - (\bar{y} + \alpha^{-1}) \log \{1 + \alpha \lambda_i\} \right]$$  \hspace{1cm} (4.16)$$

where $\alpha =$ inverse dispersion parameter,
$\lambda_i =$ expected mean number of truck-related crashes,
\( N \) = total number of vertical curves,
\( k \) = a substitute for \((y_i - 1)\),
\( y_i \) = number of truck-related crashes on vertical curve \( i \) over a time period, and
\( \bar{y} \) = observed mean number of truck-related crashes \([66,74,52,74]\).

When \( y_i - 1 \) is less than, the summation of \( k \) from 0 to \( y_i - 1 \) is equal to zero. After applying the first and second derivatives of the log-likelihood function \( l(\alpha, \lambda_i) \), gradient elements were obtained as given in Equations (4.17) and (4.18) as:

\[
\nabla_{\lambda_i} l = \frac{\bar{y}}{\lambda_i} \left( \frac{1 + \alpha \bar{y}}{1 + \alpha \lambda_i} \right) \quad (4.17)
\]

\[
\nabla_{\alpha} l = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{y_i - 1}{\kappa = 0} \left[ \sum_{\kappa = 0}^{\kappa} \frac{\kappa}{1 + \alpha \kappa} \right] + \alpha^{-2} \log \{1 + \alpha \lambda_i\} - \frac{\lambda_i (\bar{y} + \alpha^{-1})}{1 + \alpha \lambda_i} \right) \quad (4.18)
\]

where \( \bar{y} \) = observed mean number of truck-related crashes,
\( \nabla_{\lambda_i} l \) = gradient element of the mean number of truck-related crashes,
\( \nabla_{\alpha} l \) = gradient element of the dispersion parameter,
\( \lambda_i \) = expected mean number of truck-related crashes,
\( \hat{\lambda}_i \) = estimated mean number of truck-related crashes,
\( \alpha \) = inverse dispersion parameter,
\( N \) = total number of vertical curves,
\( k \) = a substitute for \((y_i - 1)\), and
\( y_i \) = number of truck-related crashes on vertical curve \( i \) over a time period \([66,74,52,74]\).
When the gradient element of the mean number of truck-related crashes is equal to zero, the expected mean number of truck-related crashes on vertical curve \( i \) will be equal to that of the observed mean. When the gradient element of the dispersion parameter is equal to zero at expected mean number of truck-related crashes equal to estimated number of truck-related crashes, gives the maximum likelihood estimate of dispersion parameter \( \alpha \) [66,74,52,74].

4.4 Negative Binomial Based on Gibbs Sampling

A Bayesian methodology with Gibbs sampling technique is adopted after developing the basic negative binomial regression model in order to obtain the predictive posterior simulation of truck-related crashes. To obtain these posterior distributions, Markov Chain Monte Carlo (MCMC) and Gibbs sampling techniques are used. The Monte Carlo simulation and Markov chains, posterior distribution is sampled from a distribution that depends on the previous sampled value [31,59,75,77]. The Gibbs Sampler works by repeatedly sampling each of these conditional distributions. The negative binomial model may be parameterized as a gamma mixture of Poisson distribution [40,45,57,57,77]. The parameterization refers to the ratio of mean of the response variable to the dispersion index and is used for the interpretation of parameters [31,59,74,77]. For the parameterization of the Poisson-gamma distribution, the mean should be equal to \( a*b \) and variance equal to \( a*b^2 \) where \( a \) and \( b \) are shape and scale parameters of the gamma distribution [31,59,77]. The Bayesian model framework to obtain negative binomial model for truck crashes are shown in Equations (4.19) and (4.20).
\[ y_i \sim Poisson(\hat{\lambda}_i) \]  \hspace{1cm} (4.19)

where \( y_i \) = number of truck-related crashes on vertical curve \( i \) over a time period, and
\[ \hat{\lambda}_i = \text{estimated mean number of truck crashes on vertical curve } i \ [73,77]. \]

Equation (4.19) states that the number of truck crashes has a Poisson distribution with parameter \( \hat{\lambda}_i \), which is the estimated mean number of truck-related crashes given in Equation (4.20) as:
\[ \hat{\lambda}_i = \exp(\beta_i X_i) \times \exp(\epsilon_i) \]  \hspace{1cm} (4.20)

where \( y_i \) = number of truck-related crashes on vertical curve \( i \) over a time period, \[ \hat{\lambda}_i = \text{estimated mean number of truck-related crashes on vertical curve } i, \]
\[ \beta_i = \text{estimated parameters in vector form associated with } X_i, \]
\[ X_i = \text{explanatory variables influencing truck crashes}, \]
\[ \exp(\epsilon_i) = \text{gamma distributed with mean 1 and variance } \sigma^2, \text{ and } \]
\[ \theta = \text{dispersion parameter } [40,57,57]. \]

Dispersion parameter \( \theta \) is taken into account to measure the variance and is taken as variance over mean [40]. The inverse of variance is termed as precision. In fully Bayesian framework for modeling and inference of estimated parameters, the Bayesian model specification requires a likelihood function and prior distribution to obtain the posterior density of estimated parameters from the given data. The posterior distribution, according to Monte Carlo methods, is proportional to the product of the prior distribution and the likelihood function, which is given in Equation (4.21) as:
\[ p(y \mid x) \propto p(y) \times p(x \mid y) \]  

where \( p(y \mid x) \) = posterior probability density function for parameter \( y \) given \( x \),

\( p(y) \) = prior probability density function for parameter \( y \) and

\( p(x \mid y) \) = likelihood function of parameter \( y \) [31,59].

Prior distribution uses the probability as a means to quantify uncertainty in the parameter \( y \) before taking the data, likelihood function relates the data to the unknown parameters by Full bayes technique and posterior distribution expresses the uncertainty in the parameter \( y \) after taking the data [31,59].

Predictive distributions for the negative binomial model for truck crashes are calculated using a Monte Carlo simulation. In the Monte Carlo simulation, non-informative priors are assigned to the parameter coefficients to obtain the marginal posterior simulations. Using Monte Carlo simulation and three Markov chains, predictive posterior distribution are sampled from a distribution that depends on the previous sampled value. Replicates are created for the truck crashes and the output are obtained in three successive chains. The initial chain uses the non-informative priors to obtain the posteriors and saves them as priors for the second chain. These newly created priors are then used in the second chain to obtain the second set of posterior values, which will then be used as priors for the third chain. The posteriors obtained in the third chain are used to obtain the predictive posterior simulations of truck crashes [31,59,77]. Once enough samples are taken to summarize the posterior distribution, the model is considered converged. Convergence of the model is assessed using Gelman-Rubin statistics, Kernel density, autocorrelation,
trace plots and times series discussed in Chapter V of Model Development [59, 74, 79, 73, 77].

4.5 Selection of Variable Criteria

The initial step in the model development is the selection of variables. The preliminary selection of variables in a model is to avoid redundancy in the model by variables that are functions of one another. The second criteria for section of variables in the model is that the variable should be significant at 95 % confidence interval, that is, at $\alpha = 0.05$ value. Including many variables in the model may cause a problem of overshadowing the effect of other variables. Thus, relevant variables are only chosen and tested for the above mentioned criteria of variable selection [33, 50, 15, 51].

4.6 Testing for Zero-Inflated Models

Number of truck crashes in the dataset includes many vertical curves with zero crashes. The Vuong’s statistics is used to test whether zero inflated Poisson, zero inflated negative binomial or negative binomial regression model are appropriate. Vuong’s statistics is stated in Equations (4.22) and (4.23) as:

$$
\nu = \frac{\sqrt{n} \left[ \frac{1}{n} \sum_{i=1}^{n} m_i \right]}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (m_i - \bar{m})^2}} \quad (4.22)
$$

$$
m_i = \log(f_i(y_i \mid x_i)/f_z(y_i \mid x_i)) \quad (4.23)
$$

where $f_j(y_i \mid x_i) =$ probability of predicting that the truck-related crashes is equal to
assumed zero inflated model (equal to 1), and negative binomial model (equals to 2) [15,59,80]

4.7 General Assessment of the Model

The model developed using the relevant statistically significant variables are further tested for the goodness-of-fit which includes Pearson statistics, deviance statistics along with Chi-squared and G-squared test discussed in section 4.7.1. Chi-squared and G-squared test also test the overdispersion criteria discussed in section 4.7.2.

4.7.1 Goodness-Of-Fit

The goodness of fit for negative binomial regression model cannot be obtained from $R^2$ value as in least square regression model, instead the Pearson $R_p^2$ statistics and Deviance $R_d^2$ statistics are used to assess the improvement in the model fit. Pearson statistics are calculated as given in Equation (4.24):

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n} \left( \frac{y_i - E(y_i)}{\sqrt{E(y_i)}} \right)^2}{\sum_{i=1}^{n} \left[ 1 - \frac{\mu_i}{\sqrt{\mu_i}} \right]^2}$$  (4.24)

where $y_i =$ observed number of truck-related crashes,

$\mu_i =$ mean number of truck-related crashes,

$E(y_i) =$ expected number of truck-related crashes, and

$n =$ number of observations [59,74,79,73,77].
Deviance statistics are calculated as given in Equation (4.25):

\[
R_d^2 = 1 - \sum_{i=1}^{n} \left[ y_i \log \left( \frac{y_i}{E(y_i)} \right) - \left( y_i - E(y_i) \right) \right] - \frac{\sum_{i=1}^{n} y_i \log \left( \frac{y_i}{E(y_i)} \right)}{n}
\]  

(4.25)

where \( y_i \) = observed number of truck-related crashes,

\( E(y_i) \) = expected number of truck-related crashes, and

\( n \) = number of observations \([59,74,79,73,77]\).

The chi-squared and G-squared values are used to assess the goodness-of-fit of the model and overdispersion. These are discussed in the section 4.6.2.

4.7.2 Overdispersion Criteria

Pearson Chi-Squared

Pearson chi-squared statistics are used to assess the presence of overdispersion in the model and is given in Equation (4.26) as:

\[
\chi^2 = \sum_{i=1}^{n} \left[ \frac{y_i - \lambda_i}{\lambda_i} \right]^2
\]  

(4.26)

where \( y_i \) = observed number of truck crashes,

\( \lambda_i \) = expected number of truck-related crashes, and

\( n \) = number of observations \([48,67,16]\).
**G-Squared Statistics**

G-squared statistics are used to assess the fit of the model and overdispersion. The G-squared value is sum of deviance and is defined as twice the difference between the maximum log likelihood with no parameter restriction (that is, with all parameters included) and maximum log likelihood with parameters restricted (that is, only with the constant term). The G-squared statistics is given by the Equation (4.27) as:

\[
G^2 = \sum_{i=1}^{n} d_i = 2 \sum_{i=1}^{n} y_i \ln(y_i / E(y_i))
\]  

(4.27)

where \( d_i \) = sum of deviance,

\( y_i \) = observed number of truck crashes,

\( E(y_i) \) = expected number of truck-related crashes, and

\( n \) = number of observations [59,74,79,73,77].

In summary, for the truck crashes which are count data Poisson regression model is applied and tested for overdispersion. The overdispersion is tested using Pearson chi-squared and deviance G-squared statistics. The presence of overdispersion may be overcome by using negative binomial regression model. The model fit is tested by Pearson \( R_p^2 \) and Deviance \( R_d^2 \) statistics. The truck crashes often contain zero crashes and are accordingly tested for zero inflated model by using Vuong’s statistics. The model is then improved using Full Bayes technique.
CHAPTER V
MODEL DEVELOPMENT AND RESULTS

5.1 Model Assessment and Model Development

The objective of this study is to develop truck-related crash prediction models for vertical curves on two-lane rural highways in Ohio using negative binomial regression model. This objective is achieved by developing a full model including both the crest and sag curves with all relevant variables that influence truck-related crashes and using the data as mentioned in Chapter III. There are two more models developed in addition to the full model. First model is the reduced model I, which refers to model developed for the number of truck crashes and relevant variables that influence truck-related crashes on crest curves. Second model is the reduced model II developed for the number of truck crashes and relevant variables that influence truck-related crashes on sag curves. The variables chosen for each of the models are based on the selection of variables criteria explained in section 4.5 and section 5.2. The criteria for the consideration of zero-inflated models performed using the Vuong’s Statistics already described in section 4.6 is discussed in section 5.2. The criterion for the general assessments is discussed in section 5.2.1. The interpretation of the variable coefficient is mentioned in section 5.2.2. The model developed using negative binomial regression model is followed by reduced models I and II using Full Bayes technique.
In section 5.3 of the model development, sections 5.3.1 though 5.3.3 concentrate on the model development and descriptive statistics of variables considered for full model, reduced models I and II using negative binomial regression. Section 5.3.4 concentrates on reduced models I and II using Full Bayes models along with a general discussion on assessment of convergence. Section 5.5 discusses the results obtained from negative binomial regression models for the full model and the reduced models I and II with assessments for each model. This is followed by discussion of the results obtained from Full Bayes models for the reduced models I and II along with the assessments. Section 5.6 compares the negative binomial models with the Full Bayes models for reduced models I and II. The improvements in the models are assessed based on the general assessment criteria for each of the two reduced models by using various statistical tests discussed in section 5.3. Lastly section 5.7 discusses the summary of results for reduced models I and II using Gibbs sampling.

5.2 Model Assessment

Given the dataset of truck crash counts and section details, the first step is to test the presence of overdispersion in order to decide between Poisson and negative binomial regression model. This is done by testing the basic assumption of the Poisson model for overdispersion as discussed earlier in section 4.1.3. In all the three models which include the full model, reduced model I and reduced model II, the variance of truck crashes is greater than that of its mean, thus indicating presence of overdispersion and the use of negative binomial regression model [33,50,15,51]. The Vuong’s statistics $|v|$ discussed in section
4.6 should be greater than +1.96 to favor a zero inflated Poisson model [15,59]. As already mentioned in section 4.5, the variables with \( p \)-value less than 0.05 significance level are considered to be significant and selected in the model. The variables with \( p \)-values greater than 0.05 significance level are removed from model [33,50,15,51].

5.2.1 General Assessment of the Model

Goodness-of-fit

The goodness of fit for negative binomial regression model as discussed in section 4.7.1 and 4.7.2 using Pearson \( R^2_p \) statistics and Deviance \( R^2_d \) statistics closer to one indicates a better model fit. If the \( G \)-squared value is equal to zero, the model is considered to be a perfect fit model. Thus the lowest \( G \)-squared value is considered to have a better fit. Likelihood ratio for goodness-of-fit is given by 2 times the difference between log likelihood function and restricted log likelihood function or the model with only the constant. Its \( p \)-values should be closer to 0.0 [59,74,79,73,77].

Overdispersion Criteria

Pearson chi-squared statistics discussed in section 4.7.2 are used to assess if the model is overdispersed at the 0.05 significance level. If the ratio of chi-squared over \((n-p)\) which is the degree of freedom, is greater than 1, then the model is overdispersed [51,59,74,79,73,77]. Where \( n \) is the number of observations of truck-related crashes and \( p \) is the number of parameters estimated in full model, reduced model I and reduced model II respectively.

\( G \)-squared statistics discussed in section 4.7.2 is used to test overdispersion. The mean deviance, that is, \( G \)-squared value over degree of freedom which is \((n-p)\), is greater than
one indicates the data to be overdispersed [51,74,79,73,77]. Where \( n \) is the number of observations of truck-related crashes and \( p \) is the number of parameters estimated in full model, reduced model I and reduced model II respectively.

Likelihood ratio for overdispersion is given similar to that likelihood ratio for goodness-of-fit. If its \( p \)-value is less than 0.0 then it rejects the null hypotheses of dispersion parameter \( \theta = 0 \) and signifies the evidence of overdispersion with the preference of negative binomial regression model over Poisson model [51,59,74,79,73,77].

5.2.2 Variable Coefficient Interpretation

The coefficient of each variable influencing the truck crashes in each of the three models gives the size of the exponential effect of a particular variable on the number of truck-related crashes. The coefficients bearing a sign of plus (+) or minus (-) indicates the direction of the effect, that is, positive indicating increase in truck-related crashes and negative indicating decrease with increase in the variable coefficient. However, the interpretation is done by effect of a unit change in the variable would affect the truck-related crashes by exponential power of that variable coefficient, and considering other variables as constants [51,59,74,79,73,77].

5.3 Model Development

Three negative binomial regression models are developed, the full model, reduced model I for crest curve data and reduced model II for the sag curve data. The variables included in each of the models with model equations given in Equations (5.1) through (5.3) are mentioned in sections 5.4.1 through 5.4.3.
5.3.1 Full Model Using Negative Binomial Regression Model

In this negative binomial regression model is fitted to the entire vertical curve data which includes both crest and sag curve data. The variables selected are according to the criteria of variable selection as discussed in section 5.2. At significance level, the variables that are significant include grade change for the length of the vertical curve \((G_{C})\), total width \((TW)\), ADT of trucks \((ADT_{TK})\) and ADT of passenger cars \((ADT_{P})\). The equation of the full model using negative binomial regression model is given in Equation (5.1) as:

\[
TK_{ACC} = \exp(\beta_0 + \beta_1 * G_{C} + \beta_2 * TW + \beta_3 * ADT_{TK} + \beta_4 * ADT_{P})
\]  

(5.1)

where \(\beta_0\) = constant term,

\(\beta_1, ..., \beta_4\) = coefficients of variable parameters, and

\(TK_{ACC}\) = number of truck-related crashes,

\(G_{C}\) = grade change for the length of vertical curve,

\(TW\) = total width,

\(ADT_{TK}\) = ADT of trucks, and

\(ADT_{P}\) = ADT of passenger cars.

In this model, total width of the roadway is taken as categorical variables in the range of 24 ft to 26 ft as mentioned in Table 3.2. The parameter estimates for this model along with the inference of results are discussed in section 5.5.1. The variables signifying the type of vertical curve are not significant at 0.05 significance level. Due to this drawback, separate negative binomial regression models are developed for crest and sag curves in the form of reduced models I and II respectively.
5.3.2 Reduced Model I Using Negative Binomial Regression Model

The negative binomial regression model developed separately for crest curves uses the data and variables already mentioned in Table 3.3 to obtain the geometric influence on truck-related crashes. All the variables mentioned in Table 3.3 are used to develop the model. According to the variable selection criteria mentioned in section 5.2 only few variables are found significant. The significant variables include grade change for the length of the vertical curve ($G_C$), posted speed (SPD), total width (TW), ADT of trucks (ADT_TK) and ADT of passenger cars (ADT_P). The equation of the reduced model I for the crest curves using negative binomial regression is given in Equation (5.2) as:

$$TK_{ACC} = \exp(\beta_0 + \beta_1 * G_C + \beta_2 * SPD + \beta_3 * TW + \beta_4 * ADT_{TK} + \beta_5 * ADT_P)$$  \hspace{1cm} (5.2)

where $\beta_0$ = constant term,

$\beta_1,...,\beta_5$ = coefficients of variable parameters, and

TK_ACC = number of truck-related crashes,

$G_C$ = grade change for the length of vertical curve,

SPD = posted speed,

TW = total width,

ADT_TK = ADT of trucks, and

ADT_P = ADT of passenger cars.

The posted speed limit and total width are taken as categorical variables in the range of 25 to 45 mph and 24 to 26 ft respectively, as mentioned in Table 3.3. The parameter estimates along with the inference for the reduced model I are given later in section 5.4.2.
5.3.3 Reduced Model II Using Negative Binomial Regression Model

Reduced model II for sag curves is developed similar to that of reduced model I for the crest curves by using variables mentioned in Table 3.4. By using the variable selection criteria at 0.05 significance level, the significant variables include grade change for the length of the vertical curve ($G_C$), posted speed (SPD), total width – 22 to <24 ft (TW1), total width – 24 to 26 ft (TW2), ADT of passenger cars (ADT_P) and offset of ADT of trucks (LN(ADT_TK)). The equation of the reduced model II for the sag curves using negative binomial regression is given in Equation (5.3) as:

$$TK_{ACC} = \exp(\beta_0 + \beta_1 \cdot G_C + \beta_2 \cdot SPD + \beta_3 \cdot TW1 + \beta_4 \cdot TW2 + \beta_5 \cdot ADT_P + LN(\beta_6 \cdot ADT_{TK}))$$  \hspace{1cm} (5.3)

where $\beta_0 = \text{constant term}$,

$\beta_1, ..., \beta_6 = \text{coefficients of variable parameters}$, and

$TK_{ACC} = \text{number of truck-related crashes}$,

$G_C = \text{grade change for the length of vertical curve}$,

$SPD = \text{posted speed}$,

$TW1 = \text{total width in the range of 22 to 24 ft}$,

$TW2 = \text{total width in the range of 24 to 26 ft}$,

$ADT_{TK} = \text{ADT of trucks}$, and

$ADT_P = \text{ADT of passenger cars}$.

The categorical variable in the model are the posted speed limit in the range of 45 to 55 mph and total width of 22 to <24 ft and 24 to 26 ft as mentioned in Table 3.4. The ADT of trucks are taken as offset values, which gives a linear effect on the truck-related
crashes. The parameter estimates along with the inference for the reduced model II for
the sag curves are later discussed in section 5.4.4.

5.3.4 Reduced Models I and II Using Full Bayes

As discussed in section 4.4 MCMC technique and Gibbs sampling are used to obtain
predictive distributions for truck-related crashes. The variable parameters considered for
developing Full Bayes reduced models I and II are same as that of reduced models I and
II using negative binomial regression models. In the Monte Carlo simulation, the
parameter priors for both reduced models I and II are taken as non-informative priors
with a normal distribution of mean 0.0, precision ($\tau$) as 1.0E-6, and a gamma distributed
dispersion parameter with shape and scale parameter as 0.1 and 0.001 respectively. Three
Markov chains are used to estimate the truck-related crashes in order to assure
convergence and to obtain predictive posterior distribution in the form of replicates
created for truck-related crashes. For both the models, the initial ‘burn-ins’ of 5000 are
discarded. Finally, 30,000 iterations are performed for each chain. After obtaining the
posterior distribution for reduced models I and II, the models are tested for convergence.
Convergence of the model is performed by visual examination of MCMC, Gelman-Rubin
statistics, Kernel density, autocorrelation, trace plots and times series [74,79,77]. The
results of these Full Bayes models along with inference of results are discussed in
sections 5.4.2 and 5.4.4.

Assessing Convergence

The increase in the number of Markov chains increases convergence of the model in each
chain. The mean of the coefficient values of the parameters and 2.5 % to 97.50 %
quantiles of each of the coefficient, should be close to median quantile to signify the convergence of the model. The Gelman-Rubin Statistics with the ratio of the normalized width of the central 80% interval of the pooled runs over the normalized average width of the 80% intervals within the individual runs is approximately equal to 1 and signifies the convergence. This may be seen by a red colored line in the Gelman-Rubin statistical plot. Kernel density plots show tendency of convergence by bell shaped plots. Autocorrelation plots indicate that the model parameters are independent variables with less correlation among themselves [74,79,77].

5.4 Results and Inference

The results for each of the three models using negative binomial regression model and two models developed using Full Bayes models are discussed along with inference of results in this section. The improvements in the model by using the negative binomial based on Gibbs sampling are also discussed in sections 5.4.3 and 5.4.5 respectively.

5.4.1 Full Model

*Using Negative Binomial Regression Model*

According to the parameter estimates obtained in this model, the model form can be written as in Equation (5.4):

\[
    \text{TK}_\text{ACC} = \exp(-3.84 + 0.18 \times G_c - 0.32 \times \text{TW} + 0.002 \times \text{ADT}_\text{TK}
    + 0.0002 \times \text{ADT}_\text{P})
\]  

(5.4)

where \(G_c\) = grade change for the length of vertical curve,

\(\text{TK}_\text{ACC}\) = number of truck-related crashes,

\(\text{TW}\) = total width,
ADT_TK = ADT of trucks, and
ADT_P = ADT of passenger cars.

The coefficients of full model show that total width has negative sign, signifying that for a unit increase in total width the truck-related crashes decreases by $e^{0.32}$ by considering the effect from other variables constant. The other variables which include grade change for the length of the curve, ADT of trucks and passenger cars are consistent with the findings of Daniel, 2002 and Daniel, 2004 [50,15]. As the grade change increases, which is the change in grade or steepness of the curve for the total length of the vertical curve, the model signifies that the truck-related crashes increases.
### Table 5.1 Statistical Summary of Full Model Using Negative Binomial Regression Model

| Variable   | Coefficient | SE  | b/SE  | P[|Z|>|z|] | Mean  |
|------------|-------------|-----|-------|-----------|-------|
| Constant   | -3.84       | 0.18| -21.16| 0         |       |
| \(G_{C}\) | 0.18        | 0.02| 9.10  | 0         | 3.04  |
| TW         | -0.32       | 0.16| -2.02 | 0.04      | 0.44  |
| ADT_TK     | 0.002       | 0.0003| 6.30  | 0         | 138.96|
| ADT_P      | 0.0002      | .35D-04| 4.66  | 0         | 1842.70|
| Dispersion Parameter | 0.55     | 0.29  | 1.90  | 0.06      |       |

#### Poisson Regression
- Iterations: 9
- Number of Observations: 1935
- Chi squared: 224.6974
- Prob[ChiSqd > value] = 0.0000000
- Log likelihood function: -578.9657
- Restricted log likelihood: -691.3144
- Chi-squared = 2431.7533
- G-squared = 2180.9201
- RsqP= 0.2805
- RsqD= 0.2189
- Overdispersion tests: g=mu(i) : 1.591
- Overdispersion tests: g=mu(i)^2: 1.674

#### Negative Binomial Regression
- Iterations: 8
- Prob[ChiSqd > value] = 0.2261390E-01
- Log likelihood function: -576.3668
- Restricted log likelihood: -578.9657
- Chi squared: 5.197923

**NOTE:**
- \(G_{C}\) represents grade of curve for the length of the curve, TW is the total width of two lane rural roadway, ADT_TK represents ADT of trucks and ADT_P represents ADT of passenger cars.

Table 5.1 shows the results from Poisson regression model and also from negative binomial regression model. The Pearson Chi-square over \((n - p)\) is 1.26 and \(G\)-squared statistics over \((n - p)\) is 1.13 which indicates the overdispersion and the preference of negative binomial regression model over Poisson model. The likelihood ratio for overdispersion mentioned in section 5.2.1 and in Table 5.1 with likelihood ratio as 224.69 and \(p < 0.0000\) statistically signifies the evidence of overdispersion. The Pearson \(R_p^2\) and
\( R^2 \) for this model are 0.28 and 0.21 respectively indicating a poor model fit when Poisson regression model is used. The likelihood ratio for goodness-of-fit by using negative binomial model is given as 5.198 with \( p \) value 0.022 which is closer to one and signifies a better fit than that by Poisson regression model. The Vuong’s statistics \(|\nu|\) is -1.64 which is less than +1.96 and does not favor the zero inflated negative binomial model.

In this full model, vertical curve indicator variable which indicates the crest and sag curves is found to be insignificant and is not included in the model. Due to this, the influence of crest and sag curves on truck crashes cannot be obtained from this model. For this reason separate models for crest and sag curves are developed. The results of these models are discussed in sections 5.4.2 and 5.4.4.

5.4.2 Reduced Model I

Using Negative Binomial Regression Model

In this model, the parameter estimates obtained may be written as in Equation (5.5):

\[
\text{TK\_ACC} = \exp(-4.18 + 0.17G_c + 0.54\text{SPD} - 0.59\text{TW} \\
+ 0.002\text{ADT\_TK} + 0.0002\text{ADT\_P})
\]

(5.5)

where \( G_c \) = grade change for the length of vertical curve,

TK\_ACC = number of truck-related crashes,

SPD = posted speed,

TW = total width,

ADT\_TK = ADT of trucks, and

ADT\_P = ADT of passenger cars.
The coefficients of reduced model I for crest curves show that total width ranging from 24 to 26 ft and the constant term have negative sign. Total width signifies that for every unit increase in total width, the truck crashes on crest curves decreases by $e^{0.59}$ by considering the effect from other variables constant. The other variable which include grade change for the length of the curve, ADT of trucks and passenger cars are consistent with the Daniel’s findings [50,15]. Truck-related crashes increases with increase in these variables.

Table 5.2 is similar to Table 5.1 and shows the results from Poisson regression model as well as the negative binomial regression model. The Pearson Chi-square over $(n-p)$ is 1.84 and $G$-squared statistics over $(n-p)$ is 1.34 which are significantly greater than one indicating overdispersion. Also, the likelihood ratio for overdispersion as seen in Table 5.2 has a value 107.69 with $p < 0.0000$ statistically signifying the evidence of overdispersion and preference of negative binomial regression over Poisson regression.

The Pearson $R^2_p$ and $R^2_d$ for this model are 0.34 and 0.31 respectively indicating better fit when compared to the full model. The likelihood ratio for goodness-of-fit by using negative binomial regression model is given as 6.465 with $p$ value 0.01 which is closer to one and signifies a better fit than that by Poisson regression model and also better fit when compared to the full model. The Vuong’s statistics $|\nu|$ for the reduced model I is -1.03 which is less than $+1.96$ and does not favor the zero inflated negative binomial model.
Table 5.2. Statistical Summary of Reduced Model I Using Negative Binomial Regression Model

| Variable          | Coefficient | SE   | b/SE  | P[|Z|>z] | Mean   |
|-------------------|-------------|------|-------|---------|--------|
| Constant          | -4.18       | 0.29 | -14.16| 0       |        |
| GC                | 0.17        | 0.03 | 6.24  | 0       | 3.23   |
| SPD – 25 to 45 mph| 0.54        | 0.25 | 2.11  | 0.03    | 0.20   |
| TW – 24 to 26 ft  | 0.59        | 0.23 | 2.53  | 0.01    | 0.45   |
| ADT_TK            | 0.002       | 0.0004 | 4.36  | 0       | 128.50 |
| ADT_P             | 0.0002      | .58D-04 | 3.03  | 0.002   | 1658.91|
| Dispersion Parameter | 0.89   | 0.47 | 1.9   | 0.05    |        |

**Poisson Regression**

- Iterations : 9
- Chi squared : 107.6981
- Log likelihood function : -301.2409
- Chi-squared = 1953.90764
- RsqP= 0.3393
- Overdispersion tests: g=mu(i) : 2.006
- Overdispersion tests: g=mu(i)^2: 1.652

**Negative Binomial Regression**

- Iterations : 14
- Chi squared : 6.464799
- Log likelihood function : -298.0085
- Restricted log likelihood : -301.2409

**NOTE:**

GC represents grade of curve for the length of the curve, SPD is posted speed, TW is the total width of two lane rural roadway, ADT_TK represents ADT of trucks and ADT_P represents ADT of passenger cars.
Using Full Bayes

Table 5.3 Statistical Summary of Reduced Model I using Full Bayes Technique.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>MC error</th>
<th>2.50 %</th>
<th>Median</th>
<th>97.50 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.10</td>
<td>0.30</td>
<td>0.01</td>
<td>-4.71</td>
<td>-4.08</td>
<td>-3.45</td>
</tr>
<tr>
<td>GC</td>
<td>0.17</td>
<td>0.03</td>
<td>9.22E-04</td>
<td>0.11</td>
<td>0.17</td>
<td>0.23</td>
</tr>
<tr>
<td>SPD – 25 to 45 mph</td>
<td>0.51</td>
<td>0.25</td>
<td>7.86E-03</td>
<td>0.02</td>
<td>0.51</td>
<td>0.98</td>
</tr>
<tr>
<td>TW – 24 to 26 ft</td>
<td>0.54</td>
<td>0.21</td>
<td>7.60E-03</td>
<td>0.16</td>
<td>0.54</td>
<td>0.99</td>
</tr>
<tr>
<td>ADT_TK</td>
<td>0.0019</td>
<td>4.40E-04</td>
<td>1.43E-05</td>
<td>1.09E-03</td>
<td>1.83E-03</td>
<td>2.87E-03</td>
</tr>
<tr>
<td>ADT_P</td>
<td>0.0002</td>
<td>6.03E-05</td>
<td>2.19E-06</td>
<td>3.71E-05</td>
<td>1.66E-04</td>
<td>2.84E-04</td>
</tr>
<tr>
<td>α</td>
<td>4.24</td>
<td>6.85</td>
<td>0.28</td>
<td>0.54</td>
<td>1.8</td>
<td>25.41</td>
</tr>
</tbody>
</table>

NOTE:
K refers to inverse dispersion parameter which is equal to 0.24, GC represents grade of curve for the length of the curve, SPD is posted speed, TW is the total width of two lane rural roadway, ADT_TK represents ADT of trucks and ADT_P represents ADT of passenger cars. α is the inverse dispersion parameter, which is 1/θ.

In this reduced model I using Full Bayes technique, the variable coefficients are improved by using MCMC method and three chains as mentioned earlier in section 5.3.4.

From the three chains the predictive posterior simulation is obtained by following the same procedure discussed in section 5.3.4. For each chain the estimated values of the variable coefficients are saved and are plotted in Figure 5.1. The values obtained in the third chain are used as proper prior and predictive posterior distributions as shown in Table 5.3 and are then tested for convergence.
Figure 5.1. Sensitivity Analysis at Each Chain for Reduced Model I.

5.4.3 Improvements in Reduced Model I for the Crest Curves

Using the assessment of convergence criteria as discussed in section 5.3.4, the estimated variable coefficients in Table 5.3 are tested for convergence. Table 5.3 shows the ADT of trucks and passenger attained convergence as the values at 2.5 % and 97.5 % quantile are very close to the median and mean value. The variables which include grade change for the length of crest curve, total width of roadway, and posted speed also attained convergence but the inverse dispersion parameter value as seen in the Table 5.3 is tending to converge after 20,000 iterations in each chain. Convergence may also be seen
in the Gelman Rubin statistic plots shown in Figure 5.2, where the ratio represented by
the red color line is closer to one, for all the parameters except for the inverse dispersion
parameter which is tending to get closer to one. The kernel density plots as shown in
Figure 5.3 also signify the same conclusion. In Table 5.3 the inverse dispersion parameter
is 4.24 which signifies that the dispersion parameter as one over 4.24 equal to 0.23. This
value is compared to the dispersion parameter of 0.89 obtained using negative binomial
regression model in Table 5.2. This signifies that reduced model I using Full Bayes has
lesser dispersion than that using negative binomial regression model. Thus these results
obtained in the Table 5.3 which have converged for all the roadway and traffic
parameters for the reduced model I using Full Bayes are considered as the improved
model over the reduced model I using negative binomial regression model.
NOTE: Beta1 to Beta 5 correspond to the variable coefficients in the reduced model I mentioned in Table 5.3. and K refers to inverse dispersion parameter.

Figure 5.2. Gelman Rubin Plots for the Variable Estimates for Reduced Model I.
NOTE: Beta1 to Beta 5 correspond to the variable coefficients in the reduced model I mentioned in Table 5.3. and K refers to inverse dispersion parameter. Except for K all the other variables estimates have converged completely.

Figure 5.3. Kernel Density Plots for Reduced Model I.
5.4.4 Reduced Model II

Using Negative Binomial Regression Model

According to the parameter estimates the model may be written as in Equation (5.6):

\[
\begin{align*}
\text{TK}_{\text{ACC}} &= \exp(-7.44 + 0.19G_c + 1.26\text{SPD} - 0.88\text{TW1} - 0.51\text{TW2} \\
& \quad + 0.0002\text{ADT}_P + \ln(0.49\text{ADT}_\text{TK})
\end{align*}
\] (5.6)

where \(G_c\) = grade change for the length of vertical curve,

\(\text{TK}_{\text{ACC}}\) = number of truck-related crashes,

\(\text{SPD}\) = posted speed,

\(\text{TW1}\) = total width in the range of 22 to 24 ft,

\(\text{TW2}\) = total width in the range of 24 to 26 ft,

\(\text{ADT}_\text{TK}\) = ADT of trucks, and

\(\text{ADT}_P\) = ADT of passenger cars.

The coefficients of reduced model II for sag curves show similar signs of the variable coefficients used as in the reduced model I for crest curves. In this model, total width ranging from 22 to less than 24 ft, 24 to 26 ft, and the constant term have negative sign bearing the coefficient and signifying the similar impact on truck-related crashes as in reduced model I. These results on total width are consistent with Daniel 2002 and Daniel 2004 [50,15]. The other variables behave similar to that of the reduced model I except for the offset. The offset which is the logarithmic of ADT of trucks influence the truck-related crashes linearly rather than exponentially and with a positive sign.
Table 5.4 Statistical Summary of Reduced Model II Using Negative Binomial Regression Model

| Variable          | Coefficient | SE     | b/SE | P[|Z|>z] | Mean  |
|-------------------|-------------|--------|------|---------|-------|
| Constant          | -7.44       | 1.01   | -7.38| 0       |       |
| $G_C$             | 0.19        | 0.03   | 6.81 | 0       | 2.84  |
| SPD – 45 to 55 mph| 1.26        | 0.59   | 2.13 | 0.03    | 0.93  |
| TW1 – 22 to 24 ft | 0.88        | 0.42   | 2.10 | 0.04    | 0.14  |
| TW2 – 24 to 26 ft | 0.51        | 0.26   | 1.98 | 0.05    | 0.45  |
| ADT_P             | 0.0002      | 0.0005 | 3.91 | 0.0001  | 1993.00|
| LN(ADT_TK)        | 0.49        | 0.17   | 2.93 | 0.003   | 4.43  |
| Dispersion Parameter | 0.69   | 0.92   | 0.75 | 0.05    |       |

**Poisson Regression**
- Iterations: 9
- Number of Observations: 870
- Chi squared: 96.4811
- Prob[ChiSqd > value] = 0.0000000
- Log likelihood function: -231.1993
- Restricted log likelihood: -279.4399
- Chi- squared = 1746.353
- RsqP= 0.258
- RsqD= 0.2357
- Overdispersion tests: g=mu(i) : -3.068
- Overdispersion tests: g=mu(i)^2: -2.645

**Negative Binomial Regression**
- Iterations: 10
- Log likelihood function: -231.51

**NOTE:**
- $G_C$ represents grade of curve for the length of the curve, SPD is posted speed, TW1 and TW2 represent two different ranges of total width of two lane rural roadway, LN(ADT_TK) represents the offset that is log of ADT of trucks and ADT_P represents ADT of passenger cars.

Table 5.4 similar to Table 5.2 shows the results from Poisson regression model and also from negative binomial regression model with Pearson Chi-square over $(n-\rho)$ as 2.01 and
deviance $G$-squared statistics over $(n-p)$ as 1.53. These values signify overdispersion and are relative higher than those obtained from reduced model I. The likelihood ratio for overdispersion as seen in Table 5.4 has a value 96.48 with $p <0.0000$, also signifying the evidence of overdispersion and the preference of negative binomial regression over Poisson regression. The Pearson $R_p^2$ and $R_d^2$ for this model are 0.26 and 0.24 respectively. This model is also tested for zero inflated models using Vuong’s statistics with $|\nu|$ of -1.32 which does not favor the zero inflated negative binomial models.

Using Full Bayes Model

Table 5.5 Statistical Summary of Reduced Model II Using Full Bayes Technique.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>MC error</th>
<th>2.50 %</th>
<th>Median</th>
<th>97.50 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-7.55</td>
<td>0.75</td>
<td>3.0E-02</td>
<td>-9.01</td>
<td>-7.61</td>
<td>-6.07</td>
</tr>
<tr>
<td>$G_C$</td>
<td>0.19</td>
<td>0.03</td>
<td>1.0E-03</td>
<td>0.13</td>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>SPD – 45 to 55 mph</td>
<td>1.30</td>
<td>0.40</td>
<td>1.6E-02</td>
<td>0.58</td>
<td>1.25</td>
<td>2.16</td>
</tr>
<tr>
<td>TW1 – 22 to 24 ft</td>
<td>0.81</td>
<td>0.41</td>
<td>1.5E-02</td>
<td>-0.04</td>
<td>0.83</td>
<td>1.59</td>
</tr>
<tr>
<td>TW2 – 24 to 26 ft</td>
<td>0.52</td>
<td>0.24</td>
<td>8.7E-03</td>
<td>8.7E-03</td>
<td>0.52</td>
<td>0.96</td>
</tr>
<tr>
<td>ADT_P</td>
<td>0.0002</td>
<td>4.38E-05</td>
<td>1.5E-06</td>
<td>1.1E-04</td>
<td>2.0E-04</td>
<td>2.9E-04</td>
</tr>
<tr>
<td>LN(ADT_TK)</td>
<td>0.50</td>
<td>0.14</td>
<td>5.3E-03</td>
<td>0.26</td>
<td>0.51</td>
<td>0.73</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>8.87</td>
<td>7.48</td>
<td>0.25</td>
<td>1.33</td>
<td>6.62</td>
<td>28.54</td>
</tr>
</tbody>
</table>

NOTE:
K refers to the inverse dispersion parameter, that is dispersion parameter is equal to 0.11, $G_C$ represents grade of curve for the length of the curve, SPD is posted speed, TW1 and TW2 represents two different ranges of total width of two lane rural roadway, LN(ADT_TK) represents the offset that is log of ADT of trucks and ADT_P represents ADT of passenger cars. $\alpha$ is the inverse dispersion parameter, which is $1/\theta$. 

81
In this reduced model II for the sag curves using Full Bayes technique, the variable coefficients are estimated using MCMC methods and three Markov chains. Figure 5.4 shows the plot of estimated variable parameters which are saved for each chain of the three chains and are used to obtain the predictive posterior distributions shown in Table 5.5 which are then tested for convergence.

![Figure 5.4. Sensitivity Analysis at Each Chain for Reduced Model II.](image)

### 5.4.5 Improvements in Reduced Model II for the Sag Curves

The variable coefficients as seen in Table 5.5 for 2.5 % to 97.5 % quantile, the ADT of passenger cars, grade change for the length of sag curve, offset in the form of log of ADT of trucks, followed by total width and speed have converged completely as the values are closer to median quantiles and mean values, except for inverse dispersion parameter.
which is tending to converge. Convergence may also be seen in the Gelman Rubin statistic plots shown in Figure 5.5, where the ratio represented by the red color line is closer to one, for all the parameters. The kernel density plots as shown in Figure 5.6 signify that coefficients of grade change for the length of the sag curve, total width of roadway and ADT of passenger cars have bell shaped curves confirming convergence. The rest of the variable coefficients in Figure 5.6 show the tendency to converge. In Table 5.5 the inverse dispersion parameter is 8.89 which signifies the dispersion parameter as one over 8.89 equals 0.11. This value when compared to the dispersion parameter of 0.69 obtained using negative binomial regression model in Table 5.4 concludes that reduced model II using Full Bayes has lesser dispersion than that using negative binomial regression model. Thus, these results obtained in the Table 5.5 which have roadway and traffic parameters converged and tending to converge for the reduced model II using Full Bayes are considered as an improved model over the reduced model II using negative binomial regression model.
NOTE: Beta1 to Beta 6 correspond to the variable coefficients in the reduced model II mentioned in Table 5.5. and K refers to inverse dispersion parameter.

Figure 5.5. Gelman Rubin Plots for the variable estimates for Reduced Model II.
NOTE: Beta1 to Beta 6 correspond to the variable coefficients in the reduced model I mentioned in Table 5.5, and K refers to inverse dispersion parameter. Except for Beta1, Beta3, Beta4 and then followed by Beta 5 and K, all the other variables estimates are tending to converge.

Figure 5.6. Kernel Density Plots for Reduced model II.
5.5 Summary of Results

Table 5.6 Statistical Summary of Reduced Models I and II.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Crest Curves</th>
<th></th>
<th>Sag Curves</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial model</td>
<td>Improved model</td>
<td>Initial model</td>
<td>Improved model</td>
</tr>
<tr>
<td>Geometry Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-4.18</td>
<td>-4.10</td>
<td>-7.44</td>
<td>-7.55</td>
</tr>
<tr>
<td>GC</td>
<td>0.17</td>
<td>0.17</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>TW1 – 22 to 24 ft</td>
<td>0.59</td>
<td>0.54</td>
<td>0.51</td>
<td>0.52</td>
</tr>
<tr>
<td>TW – 24 to 26 ft</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic Volumes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADT_TK</td>
<td>0.002</td>
<td>0.0019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LN(ADT_TK)</td>
<td></td>
<td></td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>ADT_P</td>
<td>0.0002</td>
<td>0.0002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speed Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPD – 25 to 45 mph</td>
<td>0.54</td>
<td>0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPD – 45 to 55 mph</td>
<td></td>
<td></td>
<td>1.26</td>
<td>1.30</td>
</tr>
<tr>
<td>Dispersion Parameter</td>
<td>0.89</td>
<td>0.23</td>
<td>0.69</td>
<td>0.11</td>
</tr>
</tbody>
</table>

NOTE: GC represents grade of curve for the length of the curve, SPD is posted speed, TW1 and TW2 represents two different ranges of total width of two lane rural roadway, LN (ADT_TK) represents the offset that is log of ADT of trucks and ADT_P represents ADT of passenger cars. When possible categorical variables are to be the same.

5.5.1 Geometric Characteristics

Grade Change for the Length of the Curve

As the grade changes for the length of the curve, which is the measure of the steepness of the curve increases there are more number of truck-related crashes. This is demonstrated by decrease in stopping distance and vehicle differential speed. Of the two models results shown in Table 5.6, truck-related crashes on the sag curves are slightly higher than on crest curves. Miaou, 1994 defines steepness of the curve as categorical variable and
showed that truck involvement crashes increases with steepness greater than 2% and Daniel, 2004 defines vertical curve rate as continuous variable but not statistically significant. These findings are consistent with the present findings.

Total Width
The results of the crest curve show that the middle range total width appears to have higher influence on truck-related crashes than the other two sections as shown in Table 5.6. No additional information can be gathered for higher and lower ranges. For the sag curves, the results of the three categorical range show that as total width decreases the number of truck-related crashes increases. These results suggest that total width decreases the margin for human errors as well as reduces the ability of the driver to maneuver. These results are consistent with Miaou’s and Daniel’s findings on total width considered as continuous, which show that narrowed total width of roadway tends to increase truck crashes [15].

5.5.2 Traffic Volumes
For the both the models, the results show that truck-related crashes increases with increase in truck ADT and passenger car ADT. One interesting finding is the interaction between trucks and passenger cars. According to highway capacity manual, the capacity of road is lowered with increase in truck volumes. This effect creates a higher volume to capacity ratio and is associated with more interaction between truck and passenger cars, resulting in increase in likelihood of crashes. These results support the findings of Lord, 2005 that increase in vehicle density and volume to capacity ratio increases the number of single- and multi-vehicle crashes and crash risk [81].
5.5.3 Speed Variable

The results obtained from the reduced model I using Full Bayes approach shown in Table 5.3 signify that posted speeds greater than 45 mph increases the truck-related crashes on vertical crest curves. The results obtained from the reduced model II using Full Bayes approach shown in Table 5.5 signify that posted speeds greater than 55 mph tends to increase the truck-related crashes on vertical sag curves.

The results for the crest curve may seem counter intuitive. This may be due to an underlying trend such as, lower design standards with higher operational vehicle speeds. Fitzpatrick’s findings suggest that operational speed of the vehicle and posted speed are not same on crest curves. Sag curves on the other hand, show an increase in truck-related crashes as posted speed limits increase. This may be explained as the speed of the vehicle increases, sight distance increases and the vehicle travels further in the direction. The impact of speed on crashes has been shown in Jessen’s, Fitzpatrick’s, and Collin’s findings [13,83].
CHAPTER VI

CONCLUSIONS AND RECOMMENDATIONS

Roadway crashes cause substantial economic and life loss each year. Truck crashes account for less than 5% of all highway crashes but account for 13% of all highway fatalities. Zaloshnja estimated that, in 2005, the average truck involvement crashes cost closer to $91,000 [2]. These research studies provide valuable insight into the safety of trucks. Other area of potential concern is the role of vertical curves on crashes. Previous research studies have stated in the design of vertical curves, vertical grade, and length of the curve are essential components effecting stopping distance. Jessen, Hassan, and Labi have stated that the impact of grade, stopping sight distance, and vehicle speed create operational issues for vehicles travelling on vertical curves and in turn may increase the likelihood for crashes. In many of these research findings, the geometric and traffic variables are not statistically significant and are not used in their final crash prediction model. The present study successfully predicts the impact of truck-related crashes, based upon geometric, traffic, and speed variables on rural two-lane vertical curves in Ohio. The dataset is cleaned and validated and includes, 1,935 vertical curve segments with 205 truck crashes from 2002 to 2006. The vertical curve dataset includes 1,065 crest curve segments with 114 truck crashes and 870 sag curve segments with 91 truck crashes.
6.1 Model Development

Three prediction models are developed using negative binomial regression. These models are 1. Full model – for all vertical curves 2. Reduced model I - for crest curves only and 3. Reduced model II - for sag curves only. These models are used to evaluate geometric, traffic and speed factors on truck-related crashes on vertical curves. All three models indicate the presence of overdispersion. These models when tested using Vuong’s statistics did not favor zero-inflated negative binomial model. The vertical curve indicator variable is found to be insignificant.

6.2 Improvement Using Bayesian Method

The reduced models I and II are improved using Bayesian method. The assessment of convergence criteria is used to obtain the improvement over the initial models. This method gives improved estimate parameters. These models when compared to initial models, improved by 20 %.

6.3 Summary of Findings

Based on the results obtained in Table 5.6, the summary of geometric, traffic and speed characteristics are discussed in sections 6.3.1 through 6.3.3.

6.3.1 Geometric Characteristics

Grade Change for the Length of the Curve

As seen in Table 5.6, as the grade change for the length of the curve increases, there is an increase in number of truck-related crashes. This may be the result of a decrease in
stopping sight distance and vehicle differential speed. Of the two models results shown in Table 5.6, truck-related crashes on the sag curves are slightly higher than the crest curves.

Total Width

The results of the crest curve show that the middle range total width appears to have statistically significant influence on truck-related crashes. For the sag curves, the results of the three categorical range show that as total width decreases, the number of truck-related crashes increases. These results suggest that total width decreases the margin for human error as well as reduces the ability of the driver to maneuver and avoid potential conflicts.

6.3.2 Traffic Volumes

For both models, the results show that truck-related crashes increase with the increase in truck ADT as well as passenger car ADT. In the case of sag curves, truck ADT is developed as an offset. One interesting finding is the negative interaction between trucks and passenger cars. According to highway capacity manual, the capacity of road is lowered with increase in truck volumes which increases volume to capacity ratio. The volume to capacity ratio is associated with more interactions between truck and passenger cars, resulting in the increase in likelihood of crashes.

6.3.3 Speed Variable

The speed results for the crest curve may seem counter intuitive. This may be due to an underlying trend such as, lower design standards with higher operational vehicle speeds. Sag curves on the other hand, show an increase in truck-related crashes as posted speed
limits increase. This may be explained as the speed of the vehicle increases, sight distance increases and the vehicle travels further in the direction.

6.4 Improving Safety

Truck-related crashes on vertical curves are influenced by grade change, total width of the roadway, traffic volumes of trucks and passenger cars and posted speed. The most statistically significant variables on crest and sag curves are the steepness of the curve, narrow roadway width and speed of the vehicle. Three effective ways are suggested to improve safety on these vertical curves.

1. Curve Flattening

![Figure 6.1. Mild and Steep Vertical Crest Curves.](image)

As seen in Figure 6.1, (i) shows a steep crest curve, whereas (ii) shows a mild crest curve. Steep curve in (i) has shorter sight distance when compared to the mild curve in (ii). This may increase the likelihood of crashes on steep curves than on mild curves. These suggest that by flattening steep curves, the safety on vertical curves may increase.
2. Increasing the Width of the Roadway

By increasing the width of the roadway on vertical curves the safety is improved as it increases the margin for human error as well as the ability of the driver to maneuver and in turn reducing potential conflicts.

3. Placing Advisory Speed Signs.

On the vertical crest curves, lower design speeds with higher operational vehicle speed result in truck-related crashes. Higher operational speed of the vehicle is also a major issue on sag curves. Advisory speed signs may be placed on the vertical crest and sag curves to reduce the operational speed of the vehicles and to improve the safety.
BIBLIOGRAPHY


17. Crash Tractape Layout. *Ohio Department of Public Safety*.


60. Roadway Standards, horizontal and vertical curve design. Technical report, Ohio Department of Transportation.
61. Quad maps. GIS / Mapping Section Downloads. Ohio Department of Transportation. Available as of 06/01/08:
http://www.dot.state.oh.us/aerial/QuadList.asp?D=100dpi&F=GIF

http://www.sddot.com/pe/roaddesign/docs/rdmanual/rdmch06.pdf


## APPENDIX

### LIST OF VARIABLES INCLUDED IN THE DATASET

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>County</td>
<td>This field contains abbreviations of counties. It is three character county code.</td>
</tr>
<tr>
<td>District</td>
<td>This is district where the segment of road exits. This is an integer data type. 1 to 12, each district has many counties.</td>
</tr>
</tbody>
</table>
| ML_CL         | Mile Class - Indicates if the road is inside or outside,  
  1. Rural  
  2. Municipal (incorporated)  
  3.  
  4. Rural and municipal (Split) |
| Route         | This is route number of the roadway. All US, State and Interstate routes in Ohio. This is integer data type. |
| Route_X       | Route suffix designator as follows:  
  A- Alternatives  
  B- Bypass  
  C- Spur or Connector  
  D- Directional Alternate (1st within county)  
  E- East Bound  
  F- Directional Alternate (2nd within county)  
  G- Directional Alternate (3rd within county)  
  J- Future Route (New alignment or Journalized alignment)  
  K- Turnpike  
  N- North Bound  
  P- Proposed (not built)  
  R- Regular  
  S- South Bound  
  T- Temporary Route  
  W- West Bound  
  X- Regular route due for abandonment  
  Y- DA route due for abandonment  
  A, B, C, D, F, G, J, R, T are the ones in the data set. |
| Logpt         | Station or mile point from county line or other beginning point. The values range from single digit to four digits. |
Mod_Logpt

Log pt divided by 100. This field data type is a decimal field with two decimal places. This filed data type was changed from “nvarchar” to decimal field with three decimal places.

Begin_Point

This field contains the exactly same information as the Mod_Logpt.

End_Point

This field is the ending log point. This field data type is a decimal field with two decimal places.

Mod_Length

This field is the difference between the begin_point and the end_point. This field data type is a decimal field with two decimal places. Numeric format: 99v99 (nearest one-hundredth of a mile).

Length

This field is Mod_length field multiplied by 100.

DESC

This field is the general description of the location of vertical and horizontal curves.

SYS_CL

This field is the system classification of roadways as I (Interstate), M (Major thorough fare), A (Auxiliary), L (Local/State) and Blank.

BRK

This field contains NULL and #. # is PAS/NHS intersection marker.

SEQ

This field contains sequential number for identical record types at log point.

TCODE

This field indicates the record type as A-Bridges.

Vert_Hor

This field indicates if it is a Grade, Curve or Angle.

Slope

This field indicates Crest and Sag curves with Plus for Crest and Minus for Sag.

Degree_Curve

This field gives the values of the degree of curvature for Horizontal curves in percentages.

Radius

This field is obtained from Degree of Curve using the formula, $DC=R/5280 \Rightarrow R=DC* 5280 \text{ (ft)}$.

Degree_Slope

This field gives the values of the degree of slope for Vertical curves in percentages.

Speed_post

This field gives the speed postings for each section of the roadway.

Lanes

Two digit number of lanes field. Numeric 1-19, and Blank.

Lt_Out_Shldr_width

This field will be coded for all records (divided and undivided roadways). Only paved shoulders will be included. Units in feet.

Lt_Lane_width

This field indicates one width (left lane width) of the
<table>
<thead>
<tr>
<th>Field Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lt_In_Shldr_width</td>
<td>This field will be coded for divided highways only. Only paved shoulders will be included. Units in feet.</td>
</tr>
<tr>
<td>Rt_In_Shldr_width</td>
<td>This field will be coded for divided highways only. Only paved shoulders will be included. Units in feet.</td>
</tr>
<tr>
<td>Rt_Lane_width</td>
<td>This field indicates one width (right lane width) of the combination base or divided highway. Units in feet.</td>
</tr>
<tr>
<td>Rt_Out_Shldr_width</td>
<td>This field will be coded for all records (divided and undivided roadways). Only paved shoulders will be included. Units in feet.</td>
</tr>
<tr>
<td>Median</td>
<td>This field is coded for divided highways and blank for undivided highways. Numeric value is greater than 00. Units in feet.</td>
</tr>
<tr>
<td>Median_width</td>
<td>This field indicates the distance between the hard inside edges of a through highway and it includes the width of the inside shoulders. Units in feet. 000 will be coded for undivided or non-applicable highways. 999 will be coded for median width of 1000 ft or more.</td>
</tr>
</tbody>
</table>
| Median_type                | 0- Not applicable  
1- Curbed  
2- Positive Barrier  
3- Unprotected  
4- None |
| Old_width                  | This field indicates Highway surface width plus shoulders (in feet). Median width is not indicated. It contains numeric or blank.          |
| Old_s_width                | This field indicates surface width or traveled way. Units are in feet.                                                                        |
| Population                 | This field contains population figures in hundreds. This is only derived for Municipal sections i.e., ML_CL = 2                             |
| ADT_Total                  | This field indicates the Average daily traffic (Vehicles per day). Weighted average- traffic Sections.                                           |
| Urb_Loc                    | 0- Not applicable  
1- Central Business District (CBD)  
2- High density Business / commercial center (Excluding the CBD)  
3- Low density commercial  
4- High density residential  
5- Low density Residential  
6- Other  |
| NLF_ID                     | NLF- Network Linear identification. A combination of the county and route field combined into a single text string which is used by the GIS software to identify individual GIS strings. |
Example: SBROUS00062**C

<table>
<thead>
<tr>
<th>FUNC_CLASS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Principal Arterial- Interstate (Rural)</td>
</tr>
<tr>
<td>2</td>
<td>Principal Arterial- Other (Rural)</td>
</tr>
<tr>
<td>6</td>
<td>Minor Arterial (Rural)</td>
</tr>
<tr>
<td>7</td>
<td>Major Collector (Rural)</td>
</tr>
<tr>
<td>8</td>
<td>Minor Collector (Rural)</td>
</tr>
<tr>
<td>9</td>
<td>Local (Rural)</td>
</tr>
<tr>
<td>11</td>
<td>Principal Arterial- Interstate (Urban)</td>
</tr>
<tr>
<td>12</td>
<td>Principal Arterial- Other Freeways and Expressway (Urban)</td>
</tr>
<tr>
<td>14</td>
<td>Principal Arterial- Other (Urban)</td>
</tr>
<tr>
<td>16</td>
<td>Minor Arterial (Urban)</td>
</tr>
<tr>
<td>17</td>
<td>Collector (Urban)</td>
</tr>
<tr>
<td>19</td>
<td>Local (Urban)</td>
</tr>
</tbody>
</table>

Num_Truck Crashes  This field indicates the total number of truck-related crashes for that particular segment of length.