

3-26-2013

# Development of Safety Performance Functions for SafetyAnalyst Applications in Florida

Jinyan Lu

*Florida International University, jlu002@fiu.edu*

**DOI:** 10.25148/etd.FI13042509

Follow this and additional works at: <https://digitalcommons.fiu.edu/etd>

 Part of the [Civil Engineering Commons](#)

---

## Recommended Citation

Lu, Jinyan, "Development of Safety Performance Functions for SafetyAnalyst Applications in Florida" (2013). *FIU Electronic Theses and Dissertations*. 880.

<https://digitalcommons.fiu.edu/etd/880>

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact [dcc@fiu.edu](mailto:dcc@fiu.edu).

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

DEVELOPMENT OF SAFETY PERFORMANCE FUNCTIONS FOR  
*SAFETYANALYST* APPLICATIONS IN FLORIDA

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CIVIL ENGINEERING

by

Jinyan Lu

2013

To: Dean Amir Mirmiran  
College of Engineering and Computing

This dissertation, written by Jinyan Lu and entitled Development of Safety Performance Functions for *SafetyAnalyst* Applications in Florida, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

---

L. David Shen

---

Mohammed Hadi

---

Zhenmin Chen

---

Albert Gan, Major Professor

Date of Defense: March 26, 2013

The dissertation of Jinyan Lu is approved.

---

Dean Amir Mirmiran  
College of Engineering and Computing

---

Dean Lakshmi N. Reddi  
University Graduate School

Florida International University, 2013

## DEDICATION

I dedicate this dissertation to my family for their support and love.

## ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor and mentor, Dr. Albert Gan. During the time of my doctoral study at Florida International University, he helped me become a professional, responsible, and confident person. Dr. Gan's influence on my life has spanned many levels, at times as a professor, at other times, like a caring father. Dr. Gan taught me invaluable wisdom and problem-solving skills, both of which have been of great benefit to me during my studies, and will certainly continue to do so as I enter the engineering profession. He even took time out of his busy schedule to show me how to write professionally, from emails to technical papers. With Dr. Gan's help, I received many university, state, and international honors and awards. He has shown me that nothing is impossible if I simply believe and try my best. I am deeply honored to have been one of Dr. Gan's students.

I am also deeply grateful for my committee members, Dr. David Shen, Dr. Mohammed Hadi, and Dr. Zhenmin Chen. I sincerely appreciate their interest in my research, and am thankful for their wealth of knowledge, comments, and suggestions. A special mention goes to Dr. David Shen, the first person I met at Florida International University. I am indebted to him for his recruitment into the Ph.D. program, and am truly thankful for his kind recommendations for my scholarship and fellowship applications. I would also like to thank Dr. Mohammed Hadi, with whom I took several major core courses. His courses, especially the "Highway Safety Analysis," helped me prepare for this research and build a solid foundation for my future career. I am also very grateful to Dr. Zhenmin Chen. As a statistics expert, Dr. Chen provided me with proficient guidance and statistical modeling in my research. A special thanks goes to

Dr. Fang Zhao, who was a member of my dissertation committee, but passed away during this dissertation research. She devoted her whole life to teaching and research, and her dedication has set a professional and respectable example that I will carry on in my life.

I would also like to thank the following colleagues at Florida International University for collaborating with me all of these years: Dr. Kaiyu Liu, Dr. Tao Wang, Ms. Meng Ma, Dr. Wanyang Wu, Dr. Yan Xiao, Mr. Li Tang, and Mr. Dibakar Saha. Special thanks go to Dr. Kirolos Haleem and Dr. Priyanka Alluri for giving valuable comments and suggestions for my dissertation research. Additionally, Ms. Laura Osorno and Ms. Haydee Cadena provided excellent departmental service, and I am grateful for their help throughout the years of my campus life.

Furthermore, I would like to acknowledge two agencies for their financial support of this research. My sincere appreciation goes to the Florida Department of Transportation Research Center, and to Florida International University's Graduate School, who provided a Dissertation Year Fellowship, enabling me to obtain the resources necessary to study and perform the research for this dissertation.

Finally, I would like to thank my parents and my husband for their constant encouragement and love. They are the greatest family one could ever wish for. Without their patience, understanding, and support, the completion of this work would not have been possible.

ABSTRACT OF THE DISSERTATION  
DEVELOPMENT OF SAFETY PERFORMANCE FUNCTIONS FOR  
*SAFETYANALYST* APPLICATIONS IN FLORIDA

by

Jinyan Lu

Florida International University, 2013

Miami, Florida

Professor Albert Gan, Major Professor

In 2010, the American Association of State Highway and Transportation Officials (AASHTO) released a safety analysis software system known as *SafetyAnalyst*. *SafetyAnalyst* implements the empirical Bayes (EB) method, which requires the use of Safety Performance Functions (SPFs). The system is equipped with a set of national default SPFs, and the software calibrates the default SPFs to represent the agency's safety performance. However, it is recommended that agencies generate agency-specific SPFs whenever possible. Many investigators support the view that the agency-specific SPFs represent the agency data better than the national default SPFs calibrated to agency data. Furthermore, it is believed that the crash trends in Florida are different from the states whose data were used to develop the national default SPFs.

In this dissertation, Florida-specific SPFs were developed using the 2008 Roadway Characteristics Inventory (RCI) data and crash and traffic data from 2007-2010 for both total and fatal and injury (FI) crashes. The data were randomly divided into two sets, one for calibration (70% of the data) and another for validation (30% of the data). The negative binomial (NB) model was used to develop the Florida-specific SPFs

for each of the subtypes of roadway segments, intersections and ramps, using the calibration data. Statistical goodness-of-fit tests were performed on the calibrated models, which were then validated using the validation data set. The results were compared in order to assess the transferability of the Florida-specific SPF models.

The default *SafetyAnalyst* SPFs were calibrated to Florida data by adjusting the national default SPFs with local calibration factors. The performance of the Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data were then compared using a number of methods, including visual plots and statistical goodness-of-fit tests. The plots of SPFs against the observed crash data were used to compare the prediction performance of the two models. Three goodness-of-fit tests, represented by the mean absolute deviance (MAD), the mean square prediction error (MSPE), and Freeman-Tukey  $R^2$  ( $R^2_{FT}$ ), were also used for comparison in order to identify the better-fitting model. The results showed that Florida-specific SPFs yielded better prediction performance than the national default SPFs calibrated to Florida data.

The performance of Florida-specific SPFs was further compared with that of the full SPFs, which include both traffic and geometric variables, in two major applications of SPFs, i.e., crash prediction and identification of high crash locations. The results showed that both SPF models yielded very similar performance in both applications. These empirical results support the use of the flow-only SPF models adopted in *SafetyAnalyst*, which require much less effort to develop compared to full SPFs.



## TABLE OF CONTENTS

CHAPTER	PAGE
1. INTRODUCTION .....	1
1.1. Background .....	1
1.2. Research Needs .....	4
1.3. Research Objective .....	5
1.4. Organization .....	7
2. LITERATURE REVIEW .....	8
2.1. Introduction .....	8
2.2. <i>SafetyAnalyst</i> .....	8
2.3. Full SPFs .....	11
2.3.1. <i>Roadway Segments</i> .....	11
2.3.2. <i>Intersections</i> .....	21
2.4. Simple SPFs .....	23
2.4.1. <i>Roadway Segments</i> .....	23
2.4.2. <i>Intersections</i> .....	26
2.4.3. <i>Ramps</i> .....	29
2.5. Segmentation .....	30
2.6. Identification of High Crash Locations .....	32
2.7. Summary .....	34
3. METHODOLOGY .....	37
3.1. Introduction .....	37
3.2. Segmentation .....	38
3.2.1. <i>Fixed Length and Variable Length</i> .....	38
3.2.2. <i>Fisher's Clustering Method</i> .....	39
3.3. Model Development .....	42
3.3.1. <i>Negative Binomial (NB) Models</i> .....	42
3.3.2. <i>SPF Functional Form for Roadway Segments and Ramps</i> .....	43
3.3.3. <i>SPF Functional Form for Intersections</i> .....	44
3.4. Model Tests .....	45
3.5. Identification of High Crash Locations .....	47
3.6. Summary .....	49
4. DATA PREPARATION .....	50
4.1. Introduction .....	50
4.2. Data Acquisition .....	50
4.3. Roadway Segments .....	51
4.3.1. <i>Interchange Influence Areas</i> .....	51
4.3.2. <i>Roadway Segmentation</i> .....	54
4.3.3. <i>Data used for Roadway Segments</i> .....	60
4.4. Intersections .....	61
4.5. Ramps .....	63

4.6. Advanced Segmentation for Identifying HCLs .....	65
4.6.1. <i>Introduction</i> .....	65
4.6.2. <i>Data and Model</i> .....	68
4.6.3. <i>High Crash Locations</i> .....	74
4.7. Summary .....	75
5. MODEL DEVELOPMENT .....	78
5.1. Introduction.....	78
5.2. Roadway Segments.....	78
5.3. Intersections .....	93
5.4. Ramps .....	97
5.5. Summary.....	99
6. MODEL COMPARISON .....	106
6.1. Introduction.....	106
6.2. Roadway Segments.....	106
6.3. Intersections .....	121
6.4. Ramp SPFs.....	129
6.5. Summary .....	136
7. MODEL APPLICATIONS .....	137
7.1. Introduction.....	137
7.2. Application Overview.....	137
7.3. Data and models.....	138
7.4. Crash Prediction.....	144
7.5. Identification of High Crash Locations.....	146
7.6. Summary .....	150
8. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS.....	152
8.1. Summary and Conclusions .....	152
8.2. Recommendations.....	155
REFERENCES .....	157
VITA.....	166

## LIST OF TABLES

TABLE	PAGE
Table 1-1 Categories of SPFs in <i>SafetyAnalyst</i> .....	6
Table 2-1 Regression Coefficient Estimates.....	23
Table 2-2 Regression Coefficient Estimates.....	25
Table 2-3 Crash Rate (Crashes/Million Entering Vehicles).....	28
Table 2-4 Regression Coefficient Estimates.....	29
Table 4-1 Attributes in RCI for Segmentation.....	56
Table 4-2 Summary Statistics of Roadway Segments .....	60
Table 4-3 Summary Statistics of Signalized Intersections .....	63
Table 4-4 Ramp Subtypes in <i>SafetyAnalyst</i> vs. in Florida’s RCI Database.....	63
Table 4-5 Summary Statistics of Ramps.....	65
Table 4-6 Summary Statistics for Three Segmentation Methods .....	69
Table 4-7 Rear-end NB Crash Frequency Models for Segmentation Methods .....	70
Table 4-8 Prediction Performance of Three NB Models for I-95.....	73
Table 4-9 Prediction Performance of Three NB Models for I-75.....	73
Table 4-10 Prediction Performance of Three NB Models for I-4.....	73
Table 4-11 Top Ten HCLs on the 17.3-mile Stretch of I-95 .....	75
Table 5-1 Florida-Specific SPFs for Freeways.....	79
Table 5-2 Goodness-of-Fit Tests for Freeways .....	87
Table 5-3 Florida-Specific SPFs for Urban Arterial Streets and Rural Roads .....	88
Table 5-4 Goodness-of-Fit Tests for Urban Arterial Streets and Rural Roads.....	93
Table 5-5 Florida-Specific SPFs for Intersections.....	94

Table 5-6 Goodness-of-Fit Tests for Intersections .....	97
Table 5-7 Florida-Specific SPFs for Ramps .....	98
Table 5-8 Goodness-of-Fit Tests for Ramps.....	104
Table 6-1 Florida-specific SPFs and <i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data for Freeways .....	108
Table 6-2 Statistical Comparison for Freeways.....	115
Table 6-3 Florida-specific SPFs and <i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data for Urban Arterial Streets and Rural Roads.....	116
Table 6-4 Statistical Comparison for Urban Arterial Streets and Rural Roads .....	121
Table 6-5 Florida-specific SPFs and <i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data for Signalized Intersections .....	122
Table 6-6 Statistical Comparison for Signalized Intersections.....	128
Table 6-7 Florida-specific SPFs and <i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data for Ramps.....	130
Table 6-8 Statistical Comparison for Ramps .....	135
Table 7-1 Descriptive Statistics of Explored Variables.....	142
Table 7-2 Full SPFs for Total and FI Crashes .....	143
Table 7-3 Simple SPFs for Total and FI Crashes .....	144
Table 7-4 Prediction of Full and Simple Models for Total and FI Crashes.....	145
Table 7-5 Ranking of Top 28 HCLs for Total Crashes .....	148
Table 7-6 Ranking of Top 28 HCLs for FI Crashes .....	149
Table 7-7 Paired t-Test Results.....	150

## LIST OF FIGURES

FIGURE	PAGE
Figure 3-1 Variable Length Segmentation Method .....	39
Figure 3-2 Sample Data Illustrating Fisher’s Clustering Method.....	42
Figure 3-3 Illustration of the EB Method and PSI.....	48
Figure 4-1 Interchange Influence Area.....	52
Figure 4-2 Freeways with Interchange Influence Area.....	53
Figure 4-3 Estimation of Milepost of the Intersecting Point .....	54
Figure 4-4 DySeg RCI User Interface .....	55
Figure 4-5 DySeg General User Interface .....	55
Figure 4-6 Data Processing Framework .....	62
Figure 4-7 VRICS User Interface .....	64
Figure 4-8 Comparison of Two Screening Methods.....	68
Figure 4-9 Observed vs. Predicted Crash Frequency from Three NB Models.....	73
Figure 4-10 HCLs on the 17.3-mile Stretch of I-95 in Miami-Dade.....	76
Figure 5-1 CURE Plots for Urban 4-Lane Freeways.....	81
Figure 5-2 CURE Plots for Urban 6-Lane Freeways.....	82
Figure 5-3 CURE Plots for Urban (8+)-Lane Freeways.....	83
Figure 5-4 CURE Plots for Rural 4-Lane Freeways.....	84
Figure 5-5 CURE Plots for Rural (6+)-Lane Freeways.....	85
Figure 5-6 CURE Plots for Urban 2-Lane Arterial Streets.....	89
Figure 5-7 CURE Plots for Urban Multilane Arterial Streets.....	90
Figure 5-8 CURE Plots for Urban One-Way Arterial Streets.....	91

Figure 5-9 CURE Plots for Rural Roads.....	92
Figure 5-10 CURE Plots for Urban Intersections.....	95
Figure 5-11 CURE Plots for Rural Intersections.....	96
Figure 5-12 CURE Plots for Urban Diamond Ramps.....	100
Figure 5-13 CURE Plots for Urban Partial Diamond Ramps.....	101
Figure 5-14 CURE Plots for Urban Partial Cloverleaf Ramps.....	102
Figure 5-15 CURE Plots for Rural Diamond Ramps.....	103
Figure 6-1 Observed Crashes and SPFs for Urban 4-Lane Freeways.....	109
Figure 6-2 Observed Crashes and SPFs for Urban 6-Lane Freeways.....	110
Figure 6-3 Observed Crashes and SPFs for Urban (8+)-Lane Freeways.....	111
Figure 6-4 Observed Crashes and SPFs for Rural 4-Lane Freeways.....	113
Figure 6-5 Observed Crashes and SPFs for Rural (6+)-Lane Freeways.....	114
Figure 6-6 Observed Crashes and SPFs for Urban 2-Lane Arterial Streets.....	117
Figure 6-7 Observed Crashes and SPFs for Urban Multilane Undivided Arterial Streets.....	118
Figure 6-8 Observed Crashes and SPFs for Urban Multilane Divided Arterial Streets.....	118
Figure 6-9 Observed Crashes and SPFs for Urban One-Way Arterial Streets.....	119
Figure 6-10 Observed Crashes and SPFs for Rural 2-Lane Roads.....	120
Figure 6-11 Observed Crashes and SPFs for Rural Multilane Divided Roads.....	120
Figure 6-12 SPFs for Urban Three-Leg Signalized Intersections.....	123
Figure 6-13 SPFs for Urban Four-Leg Signalized Intersections.....	124
Figure 6-14 SPFs for Rural Three-Leg Signalized Intersections.....	125
Figure 6-15 SPFs for Rural Four-Leg Signalized Intersections.....	126

Figure 6-16 Observed Crashes and SPFs for Urban Diamond Ramps .....	131
Figure 6-17 Observed Crashes and SPFs for Urban Partial Diamond Ramps.....	132
Figure 6-18 Observed Crashes and SPFs for Urban Partial Cloverleaf Ramps.....	133
Figure 6-19 Observed Crashes and SPFs for Rural Diamond Ramps .....	134
Figure 7-1 Observed vs. predicted crash frequency for full and simple models .....	145

## LIST OF ACRONYMS/ABBREVIATIONS

AADT	Annual Average Daily Traffic
AASHTO	American Association of State Highway and Transportation Officials
C	Calibration Factor
CAR	Crash Analysis Reporting
CART	Classification and Regression Tree
CMF	Crash Modification Factors
CURE	Cumulative Residuals
DySeg	Dynamic Segmentation
EB	Empirical Bayes
FB	Full Bayes
FDOT	Florida Department of Transportation
FHWA	Federal Highway Administration
FI	Fatal and Injury
FIU	Florida International University
GIS	Geographic Information System
GLM	Generalized Linear Model
HCL(s)	High Crash Location(s)
HSIS	Highway Safety Information System
HSM	Highway Safety Manual
HTBR	Hierarchical Tree-Based Regression
IIA	Interchange Influence Area
LCTR	Lehman Center for Transportation Research



MAD	Mean Absolute Deviance
MARS	Multivariate Adaptive Regression Splines
ML	Maximum Likelihood
MSE	Mean Square Error
MSPE	Mean Square Prediction Error
NB	Negative Binomial
NCHRP	National Cooperative Highway Research Program
ODP	Overdispersion Parameter
PDO	Property Damage Only
PSI	Potential for Safety Improvement
$R^2_{FT}$	Freeman-Tukey $R^2$ Coefficient
RCI	Roadway Characteristics Inventory
RTM	Regression-to-the-Mean
SAS	Statistical Analysis Software
SPF(s)	Safety Performance Function(s)
VRICS	Visual Roadway Inventory Collection System
ZIP	Zero-Inflated Poisson

# CHAPTER 1

## INTRODUCTION

### 1.1. Background

Traffic crashes are a major problem throughout the world, causing fatalities, physical injuries, property damage, and highway congestion. In the United States alone, around 40,000 fatalities and two million injuries are reported each year, costing the nation an estimate of over \$200 billion in damages (NHTSA, 2008). A reduction in highway traffic crashes has thus been a priority for highway agencies over the past few decades.

The roadway safety management process has played an important role in the nation's efforts to continue to improve highway traffic safety. The analysis process generally includes six major steps (Richard et al., 2010): (1) identification of high crash locations (HCLs), (2) diagnosis of safety problems at specific sites, (3) selection of countermeasures at specific sites, (4) economic appraisal for countermeasures under consideration, (5) priority rankings of improvement projects, and (6) evaluation of safety effectiveness of implemented countermeasures.

Among the six steps, identification of HCLs, also commonly referred to as network screening, is considered the most crucial step in the entire roadway safety management process. This step could be performed using either traditional methods like crash frequencies and rates, or advanced methods that use the empirical Bayes (EB) or full Bayes (FB) approach. Although traditional methods have minimal data requirements, they are fraught with several issues, limitations, and biases, including the well-recognized regression-to-the-mean (RTM) bias. This bias may cause locations with a high amount of crashes due to random fluctuations in crash occurrences to be

flagged erroneously for safety improvements; thus, reducing the cost-effectiveness of safety programs. The EB method adjusts for the RTM bias by comparing the safety performance of a specific site to several reference sites. This approach requires the use of Safety Performance Functions (SPFs).

An SPF is a mathematical model that describes the relation between crash frequency and measure of exposure. In network screening, SPFs are used with the EB method to calculate the expected crash frequency by considering both the observed and the predicted crash frequencies. Traditionally, SPFs relate crash occurrence to both roadway geometric characteristics (e.g., lane width, shoulder width, horizontal and vertical curvatures, etc.) and traffic characteristics (e.g., traffic volume and speed limit) for specific roadway types. These SPFs are referred to as full SPFs. In the full model, the predicted crash frequency is a function of both traffic and geometric characteristics, as follows:

$$N_{Predicted} = \exp (\alpha + \beta \times \ln(AADT) + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n) \quad (1-1)$$

where

$N_{Predicted}$  = predicted crash frequency per mile per year;

$AADT$  = annual average daily traffic;

$X_1, X_2, \dots, X_n$  =  $n$  roadway geometric variables; and

$\alpha, \beta, \beta_1, \beta_2, \dots, \beta_n$  = regression coefficients of the intercept,  $\ln(AADT)$  variable, and  $n$  geometric variables.

One main problem with full SPFs is that the independent variables might subject to correlation. In order to avoid the correlation problem, a new approach that links crash occurrence to traffic volume alone for specific roadway types is increasingly being

accepted. These SPFs, called simple SPFs or traffic SPFs, take the following functional form:

$$N_{Simple\ SPF} = exp(\alpha + \beta \times \ln(AADT)) \quad (1-2)$$

where

$N_{Simple\ SPF}$  = predicted crash frequency per mile per year as a function of traffic alone,

$AADT$  = annual average daily traffic, and

$\alpha, \beta$  = regression coefficients.

The Highway Safety Manual (HSM) adopts simple SPFs, which were generated using sites with “base conditions”. In other words, the sites utilized to generate base SPFs have similar “base” roadway characteristics. To account for the impacts of geometric conditions that are different from the base conditions, Crash Modification Factors (CMFs) are applied to the simple SPF in predicting the crash frequency:

$$N_{Predicted} = N_{Simple\ SPF} \times CMF_1 \times CMF_2 \times \dots \times CMF_n \quad (1-3)$$

where

$N_{Predicted}$  = predicted crash frequency per mile per year, and

$CMF_1, CMF_2, \dots, CMF_n$  = crash modification factors for  $n$  geometric conditions.

Thus, CMFs are simply multiplicative factors that are applied to adjust the predicted crash frequency obtained from the simple SPF for the effects of individual roadway geometric conditions such as lane width and shoulder width. A geometric condition that matches the base condition is assigned a CMF of 1.00. A geometric condition that results in the increase of the base crashes will have a CMF of greater than 1.00. On the other hand, a geometric condition that results in the decrease of the base crashes will have

a CMF of less than 1.00. For example, if the base lane width is 12 feet, a CMF of 1.1 for 11-foot lanes would mean that using 11-foot lanes can be expected to increase the crashes by 10% on average, as compared to those of 12-foot lanes.

Similar to the HSM, *SafetyAnalyst*, a safety analysis software system released by the American Association of State Highway and Transportation Officials (AASHTO) in 2010, also adopts simple SPFs. However, the SPFs used in *SafetyAnalyst* are generated using all sites, irrespective of base conditions. In summary, consideration of base conditions while generating SPFs is a major difference between the HSM and *SafetyAnalyst*.

## **1.2. Research Needs**

To perform the step of network screening in the roadway safety management process, *SafetyAnalyst* implements the EB method, which is data intensive, requiring the use of SPFs. *SafetyAnalyst* is equipped with a set of national default SPFs, and the software calibrates the default SPFs to represent the agency's safety performance. Agencies are recommended to generate agency-specific SPFs whenever possible. Many investigators support the view that the agency-specific SPFs represent the agency data better than the national default SPFs calibrated to agency data. Furthermore, it is believed that the crash trends in Florida are different from the states whose data were used to develop the national default SPFs.

*SafetyAnalyst* uses a very detailed classification of the road network, dividing segments, intersections, and ramps into 17, 12, and 16 subtypes, respectively. Table 1-1 lists the categories of SPFs included in *SafetyAnalyst*. Furthermore, *SafetyAnalyst* is equipped with SPFs for all of the subtypes for both total and fatal and injury (FI) crashes.

Thus, Florida-specific SPFs should be developed for both total and FI crashes for all roadway subtypes in *SafetyAnalyst* for which data are available.

### **1.3. Research Objective**

The objective of this dissertation research is to develop Florida-specific SPFs for different roadway categories, including roadway segments, intersections, and ramps, in *SafetyAnalyst* for applications in the roadway safety management process. Accordingly, the specific tasks of this dissertation are to:

1. develop Florida-specific SPFs for different types of roadways and assess the performance of the models;
2. calibrate the *SafetyAnalyst* default SPFs to Florida data;
3. compare the performance of the two SPFs (i.e., Florida-specific SPFs vs. *SafetyAnalyst* default SPFs calibrated to Florida data);
4. apply Florida-specific SPFs for crash prediction performance and identification of high crash locations; and
5. explore an alternative method that uses the clustering algorithm in segmentation for identification of high crash locations.

**Table 1-1 Categories of SPFs in *SafetyAnalyst***

<b>Roadway Segments-SPFs for Specific Site Subtypes</b>
<ul style="list-style-type: none"> <li>- Rural 2-lane roads</li> <li>- Rural multilane undivided roads</li> <li>- Rural multilane divided roads</li> <li>- Rural freeways-4 lanes</li> <li>- Rural freeways-6+ lanes</li> <li>- Rural freeways within interchange area-4 lanes</li> <li>- Rural freeways within interchange area-6+ lanes</li> <li>- Urban 2-lane arterial streets</li> <li>- Urban multilane undivided arterial streets</li> <li>- Urban multilane divided arterial streets</li> <li>- Urban one-way arterial streets</li> <li>- Urban freeways-4 lanes</li> <li>- Urban freeways-6 lanes</li> <li>- Urban freeways-8+ lanes</li> <li>- Urban freeways within interchange area-4 lanes</li> <li>- Urban freeways within interchange area-6 lanes</li> <li>- Urban freeways within interchange area-8+ lanes</li> </ul>
<b>Intersections-SPFs for Specific Site Subtypes</b>
<ul style="list-style-type: none"> <li>- Rural three-leg intersection with minor-road STOP control</li> <li>- Rural three-leg intersection with all-way STOP control</li> <li>- Rural three-leg intersection with signal control</li> <li>- Rural four-leg intersection with minor-road STOP control</li> <li>- Rural four-leg intersection with all-way STOP control</li> <li>- Rural four-leg intersection with signal control</li> <li>- Urban three-leg intersection with minor-road STOP control</li> <li>- Urban three-leg intersection with all-way STOP control</li> <li>- Urban three-leg intersection with signal control</li> <li>- Urban four-leg intersection with minor-road STOP control</li> <li>- Urban four-leg intersection with all-way STOP control</li> <li>- Urban four-leg intersection with signal control</li> </ul>
<b>Ramps-SPFs for Specific Site Subtypes</b>
<ul style="list-style-type: none"> <li>- Rural diamond off-ramp</li> <li>- Rural diamond on-ramp</li> <li>- Rural parclo loop off-ramp</li> <li>- Rural parclo loop on-ramp</li> <li>- Rural free-flow loop off-ramp</li> <li>- Rural free-flow loop on-ramp</li> <li>- Rural free-flow outer connection ramp</li> <li>- Rural direct or semi direct connection</li> <li>- Urban diamond off-ramp</li> <li>- Urban diamond on-ramp</li> <li>- Urban parclo loop off-ramp</li> <li>- Urban parclo loop on-ramp</li> <li>- Urban free-flow loop off-ramp</li> <li>- Urban free-flow loop on-ramp</li> <li>- Urban free-flow outer connection ramp</li> <li>- Urban direct or semi direct connection</li> </ul>

#### **1.4. Organization**

The rest of the dissertation is organized as follows. Chapter 2 provides a comprehensive review of the existing SPFs, the segmentation methods, and the high crash location identification process. Chapter 3 presents the methodology used in this research to achieve the stated objective. Chapter 4 describes the data acquisition and processing. Chapter 5 is devoted entirely to the development of Florida-specific SPFs for applications of *SafetyAnalyst*. Chapter 6 compares the performance of the Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data. Chapter 7 presents two applications of Florida-specific SPFs, i.e., the crash prediction and the identification of high crash locations. Finally, Chapter 8 draws conclusions, summarizes the main contribution, and provides recommendations for future research.



## CHAPTER 2

### LITERATURE REVIEW

#### 2.1. Introduction

This chapter reviews past research efforts associated with the development and application of SPFs. Section 2.2 introduces a safety analysis software system known as *SafetyAnalyst*. An extensive review of the literature on the existing SPFs, existing segmentation methods, and high crash location identification process are subsequently provided in detail. Sections 2.3 and 2.4 provide a comprehensive review of studies about SPFs, including full and simple SPFs, respectively, for various roadway types. Section 2.5 is a review of segmentation methods, an essential step for developing SPFs. Section 2.6 presents the network screening methods for identifying high crash locations, one of the main applications of SPFs. A summary is provided in the last section to highlight the necessity of conducting this research.

#### 2.2. *SafetyAnalyst*

In 2010, the AASHTO released a safety analysis software system known as *SafetyAnalyst*. *SafetyAnalyst* is a cooperative effort by the Federal Highway Administration (FHWA), 27 participating states in the U.S., and interested local agencies. *SafetyAnalyst* “provides state-of-the-art analytical tools for use in the decision-making process to identify and manage a system-wide program of site-specific improvements to enhance highway safety by cost-effective means” (FHWA, 2010). *SafetyAnalyst* uses a set of default SPFs developed using available data for four states, which include California, Minnesota, Ohio, and Washington.

*SafetyAnalyst* is a state-of-the-art analytical tool for making system-wide safety decisions. The software provides a suite of analytical tools to identify and manage system-wide safety improvements by incorporating all of the steps in the roadway safety management process. It incorporates the EB approach for network screening. It also includes many modules and can act as a complete “safety toolbox” for traffic safety agencies. The modules in *SafetyAnalyst* include (FHWA, 2010):

1. *Network Screening Module*: It identifies and ranks sites with potential for safety improvements.
2. *Diagnosis and Countermeasure Selection Module*: The diagnosis module is used to diagnose the nature of safety problems at specific sites. The countermeasure selection module assists users in selecting countermeasures that reduce crash frequency and severity at specific sites.
3. *Economic Appraisal and Priority Ranking Module*: The economic appraisal module performs an economic appraisal of a specific countermeasure or several alternative countermeasures for a specific site, while the priority ranking module provides a priority ranking of sites and proposed improvement projects based on the benefit and cost estimates determined by the economic appraisal tool.
4. *Countermeasure Evaluation Module*: The countermeasure evaluation module has the capability to conduct before/after evaluations of implemented safety improvements.

*SafetyAnalyst* includes four tool components. They are the Data Management Tool, the Analytical Tool, the Administration Tool, and the Implemented Countermeasure Tool. Together, they can perform the complete roadway safety management process. The Data Management Tool is used to import, post-process, and calibrate data. The Analytical Tool is used to perform analysis of the data. All of the modules of *SafetyAnalyst* discussed earlier can be performed with this tool. The Administration

Tool is used to perform a variety of tasks, such as adding and removing data items (with an exception of mandatory data elements). Data recording of various data elements' attributes can also be performed. This tool also provides access to the national default SPFs used within the software, which can be replaced with agency-specific SPFs, whenever available. Furthermore, diagnostic questions and countermeasures can also be edited within this tool.

In summary, *SafetyAnalyst* is a suite of software tools that implement the advanced EB method and automate all of the steps of the roadway safety management process. Even though its data requirements are rigorous, once the data are imported, the analysis procedures are relatively easy, requiring minimum statistical expertise.

The EB approach requires either site-specific SPFs or default SPFs calibrated to local data. The SPFs available in the HSM were generated using sites with a set of base conditions such as those with 12 feet lanes. As these base SPFs were generated using other states' data, they needed to be calibrated to represent local agencies' safety trends. The HSM recommends generating calibration factors for each subtype at least once every 2-3 years using data from 30-50 sites where a total of 100 crashes occur per year. On the contrary, the default SPFs used in *SafetyAnalyst* were generated using all sites, regardless of base conditions. Similar to the HSM, *SafetyAnalyst* uses a calibration factor to account for differences between the default SPFs and the agencies' safety performance. This calibration factor is calculated as the ratio of observed to predicted crashes for all sites.

In the EB approach, CMFs are used in two instances: to account for the variations in base conditions, and to select and evaluate countermeasures. Since the SPFs in the

HSM were generated using sites with base conditions, deviation of the target sites from predefined base conditions must be addressed using CMFs. Furthermore, CMFs are used to evaluate the performance of one countermeasure over the other, and therefore play a vital role in selecting and evaluating countermeasures, as well as in benefit-cost analysis. In *SafetyAnalyst*, CMFs are only used for countermeasure selection and evaluation, as SPFs to be used with *SafetyAnalyst* were generated without accounting for base conditions.

### **2.3. Full SPFs**

The HSM (2010) defines SPFs as regression models for estimating the predicted average crash frequency of individual roadway sections or intersections. There are two main types of SPFs: full SPFs and simple SPFs. A review of the full SPFs for roadway segments and intersections is presented in the following sections.

#### ***2.3.1. Roadway Segments***

##### *2.3.1.1 Arterials*

The mathematical relationships between crashes and roadway geometric design features (lane width, shoulder width, horizontal curvature, vertical grade, etc.) are discussed in many studies. As summarized in the National Cooperative Highway Research Program (NCHRP) Report 197 (1978), multiple linear regression models were employed frequently in establishing the relationship between crashes and geometric features. However, the undesirable outcome of using a multiple linear regression model was evidenced by the following studies. For example, results from multiple linear regression models used by Zegeer et al. (1990) showed that more than half of the

roadway sections had no crashes during the observed period. Jovanis and Chang (1986) discussed that the distribution of crash occurrences is positively skewed, and that the underlying normal distributional assumption for linear regression is not a good approximation for investigating this relationship.

In contrast to multiple linear regression models, the Poisson regression models have become widely used for modeling crashes and influencing factors. Joshua and Garber (1990) provided Poisson regression models to establish mathematical relationships between truck crashes on rural highways and traffic and geometric variables in the state of Virginia. The length of each roadway segment was restricted to a maximum of two miles. All of the selected sites were grouped into three environments by roadway configurations and traffic volumes: Environment I (primary highways, undivided, four-lane and two-lane, with AADT less than 15,000); Environment II (primary highways, divided, four-lane, with AADT less than or equal to 15,000); and Environment III (Interstate/ primary highways, divided, four-lane, with AADT more than 15,000). The prediction model for Environment I, II, and III are given in Equations (2-1) to (2-3) below, respectively:

$$Crashes/year = 0.015237 \times (SCR)^{0.0577} \times (AADT)^{0.5024} \times (TPERCNT)^{0.5731} \quad (2-1)$$

$$Crashes/year = 9 \times 10^{-8} \times (SCR)^{0.0471} \times (AADT)^{1.4358} \times (TPERCNT)^{1.5232} \times (SEGLN)^{0.3826} \quad (2-2)$$

$$Crashes/year = 0.001465 \times (CCR)^{0.0336} \times (AADT)^{0.7086} \times (TPERCNT)^{0.2064} \times (SEGLN)^{0.3318} \times (SPDIFSQ)^{0.0475} \quad (2-3)$$

where

*SCR* = slope change rate,

*AADT* = average annual daily traffic,

*TPERCNT* = truck percentage,

*SEGLLEN* = segment length,

*CCR* = curvature change rate, and

*SPDIFSQ* = speed difference.

These models indicated that the slope change rate, AADT, truck percentage and speed differences between trucks and non-trucks influenced the crash occurrence. However, these models did not consider any exposure (AADT or segment length) for the crash occurrence, which leads to the conclusion that the crash frequency would be zero if any of the variables' values were equal to zero (e.g., crashes would occur on a roadway section without slope changes).

Miaou et al. (1992) presented empirical relationships obtained through Poisson regression analyses, relating the truck crashes with key highway geometric design variables by using a data source administered by the FHWA, from the Highway Safety Information System (HSIS). The descriptions of the HSIS database are available in a study by Miaou et al. (1991). In the aforementioned study, four models were developed using different horizontal curvature and vertical grade measures. The following model is an example of these types of models, for rural interstate highways based on the data from 1985 to 1987, cited as such:

$$Crashes = \exp(-14.6833 + 0.044691X_1 + 0.172513X_2 + 0.162218X_3 + 0.03859X_4) \quad (2-4)$$

where

$X_1$  = AADT,

$X_2$  = horizontal curvature,

$X_3$  = vertical grade, and

$X_4$  = deviation from ideal shoulder width.

The final model suggested that AADT, horizontal curvature, and vertical grade were significantly correlated with truck crash occurrences, but that shoulder width had comparably less correlation. Due to the use of the exponential form in the model, it did not predict zero crashes when the variables' values were equal to zero. Accordingly, it was found to be more reasonable than the Joshua and Garber models cited above.

Maher and Summersgill (1996) indicated the weaknesses and technical difficulties in the application of the pure Poisson model, such as the low mean value problem, overdispersion, the disaggregation of data over time, the uncertainty of predictions, random errors in the flow estimates, and aggregation of predictions. Given these shortcomings, they emphasized that the technique of generalized linear models (GLMs) with Poisson error structure offered the most appropriate approach for data analysis.

A known limitation in applying the Poisson regression model is that the variance is restrained to be equal to the mean of the dataset (Dean, 1989). However, unlike the property of the most common count-data modeling approach, the variance of the crash counts for crash frequency exceeds the mean. Therefore, "when overdispersed data are present, the Poisson regression model will result in biased and inconsistent parameter estimates, which in turn could lead to erroneous inferences regarding the factors that

determine crash frequencies” (Lord and Mannering, 2010). In a study by Hadi et al. (1995), tests for overdispersion were conducted to decide whether Poisson or Negative binomial (NB) regression should be used for model development. Roadway segments from the state of Florida’s roadway system were used in the study. Models were developed for each of nine highway categories for total number of crashes, fatal crashes, and injury crashes, respectively. The nine highway categories are two-lane rural, four-lane rural divided, two-lane urban undivided, four-lane urban undivided, four-lane urban divided, six-lane urban divided roadway segments, four- and six-lane rural, four-lane urban, and six-lane urban freeways. One result showed that increasing lane width, median width, and/or shoulder width were effective in reducing crashes, depending on the highway type.

Miaou (1994) evaluated and compared the performance of Poisson, zero-inflated Poisson (ZIP), and NB regression models in establishing the relationship between truck crashes and the geometric design of road sections. The HSIS data were used to estimate the performance of these models, and unknown parameters were estimated by maximum likelihood (ML). It was recommended that if the overdispersion of crash frequency data was found to be moderate or high, the ZIP or NB regression models should be considered.

The ZIP regression model was considered when the data were characterized by a significant amount of zeros. Qin et al. (2004) used ZIP to estimate crash predicting models for crash types (single-vehicle, multi-vehicle same direction, multi-vehicle opposite direction, and multi-vehicle intersecting) as a function of AADT, segment length, speed limit, and roadway width for roadway segments in the state of Michigan using



crash and roadway characteristics data from HSIS. They concluded that, as opposed to the relationship between crashes and segment length for all crash types, the relationship between crashes and AADT is non-linear and varies by crash type.

The NB model is an extension of the Poisson model that was created to overcome possible overdispersion in the data. Overdispersion occurs when the crash data variance is greater than the mean. The NB distribution contains two parameters; the mean  $\mu$  and the dispersion parameter  $\alpha$  or its inverse ( $1/\alpha$ ), with the dispersion parameter used to capture the extra-variation observed in the crash data. Miaou (1996) found that the dispersion parameter  $\alpha$  can be used as a measure of goodness-of-fit. Wood (2005) also used the dispersion parameter to estimate the confidence intervals for the Poisson mean and gamma mean. Furthermore, Park and Lord (2008) used simulation to adjust the maximum likelihood estimate of the NB dispersion parameter, where simulation scenarios were used to develop a relationship between the estimated and the true dispersion parameters. Finally, Zhang et al. (2007) used the bootstrapped maximum likelihood method to estimate the dispersion parameter of the NB model for analyzing crash data.

Sawalha (2002) collected the crash data of 58 arterials from the cities of Vancouver, Canada, and Richmond, Virginia, for years 1994 through 1996, and generated a prediction model using the NB method. Crashes that occurred at signalized intersections were not included in this sample. The collected crash data consisted of crash locations, severities, and crash types, as well as the light, weather, and road conditions at time of crash. The model form is adopted from the following:

$$\lambda = a_0 \times L^{a_1} \times V^{a_2} \times e^{\sum_{j=1}^m b_j x_j} \quad (2-5)$$

where

$L$  = section length,

$V$  = section annual average daily traffic,

$x_j$  = any of  $m$  variables additional to  $L$  and  $V$ , and

$a_0, a_1, a_2, b_j$  = model parameters.

The roadway sections were divided into segments between consecutive signalized intersections. The geometric data considered were segment length, number of lanes, number of driveways, number of bus stops, number of crosswalks, median types, land use, etc. This crash prediction model was one of the strongest, as it demonstrated a robust goodness-of-fit.

Various studies have been conducted on the relationship between crash occurrences and relative variables for specific roadway facilities and characteristics using the NB regression model. Based on the NB regression, Bowman et al. (1995) generated vehicle crash models for different median types (raised median, two-way-left-turn median, and undivided cross section) in urban and suburban unlimited access arterials. The prediction ratio plots derived from the data displayed a relatively equal distribution of predicted vehicle crashes. The prediction models further showed that the raised curb and undivided cross-section models had the largest deviation, while the two-way-left-turn median had the smallest. The number of signalized intersections was not included in the models because it was not as significant as the other independent variables. Moreover, crash prediction models for roads with minor junctions at both single and

dual-carriageway roads in urban and rural areas in the United Kingdom were developed by Mountain et al. (1996). This study, which used the NB regression model based on data for 3,800 km of highway, included more than 5,000 minor junctions. Furthermore, safety analysis employing the NB regression model, in addition to the aforementioned research, can also be found in many other references (Anastasopoulos and Mannering, 2009; Abdel-Aty and Radwan, 2000; Sawalha and Sayed, 2006).

### *2.3.1.2 Freeways*

Crash prediction models for freeways have been deemed important since the 1990s, although there were few related studies beforehand. Persaud and Dzbik (1993) developed the generalized linear crash prediction models. Their studies showed that the crash patterns on freeway sections during congested periods differ from that during uncongested periods. Resende and Benekohal (1997) calibrated the multiple linear crash prediction models for rural freeways based on the ratio of volume to capacity. The result showed that the capacity was the crucial variable in the model.

Kraus et al. (1993) explored the relationship between crashes and variables, including geometric features, time of day, and traffic flow rate by developing a non-linear prediction model for urban freeway sections. Khan et al. (1999) also developed a non-linear regression model, but focused on the relationship between crash severities and traffic volume and segment length. Traffic volume, topological characteristics, and weather conditions were considered independent variables by Konduri and Sinha (2002) in a crash prediction model they developed using a non-linear modeling approach.

Garber and Ehrhart (2000) developed mathematical relationships that describe the combined influence that traffic and geometric characteristics have on crash occurrences.

The data of this study were obtained from roadways in the state of Virginia, with speed limits of 55 or 65 mph. Three types of models, i.e., linear regression, robust regression, and multivariate ratio of polynomials, were developed using the variables of traffic flow, speed, lane width, and shoulder width. The variables  $R^2$  and minimum AIC were used as standards for selecting the best-fitted models. For example, the best-fitted model for freeway with a 65-mph speed limit is:

$$\begin{aligned} \ln(CRASHRATE) = & (2629.7-0.4 \times SD^2) - (5.4E-04) \times SD^4 - 2254323.0 \times (1/FPL^2) + 4.5 \times SD^2 \times \\ & (1/FPL^2) - (5.4E+08) \times [(1/FPL^2)]^2 - 510.7 \times SQRT(MEAN) + (5.2E-02) \times \\ & SD^2 \times SQRT(MEAN) + 224565.2 \times (1/FPL^2) \times SQRT(MEAN) + 24.7 \times [S \\ & QRT(MEAN)]^2 \end{aligned} \quad (2-6)$$

where

$CRASHRATE$  = crash rate,

$SD$  = standard deviation of speed,

$FPL$  = flow per lane, and

$MEAN$  = mean speed.

Although complex, these models showed the relationship between crash rates and the independent variables of standard deviation for speed, mean speed, and flow per lane. These models also showed that the crash rate is not solely determined by only one independent variable, but by a complex interaction of multiple independent variables.

Several new ideas and techniques were used in the safety analysis of freeways. Golob et al. (2004) assessed the freeway safety in terms of crash types, locations, and severities by using a clustering method. Lord et al. (2005) studied the crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments located both downtown and outside of Montreal, in Quebec, Canada. These results showed that single- and multi-vehicle crashes should be separated in the

processing of modeling. The prediction model using vehicle density and V/C ratio as a covariate, thus offered a better characterization of each crash.

Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means results in a system that can improve its own performance. Machine learning has increasingly received more attention from transportation researchers as a promising technique in safety analysis. Employed by Chang and Chen (2005) to analyze freeway crash frequency, the classification and regression tree (CART) is one of the most widely applied data mining techniques. The results indicated that the traffic volume and precipitation variables were the key determinants for freeway crash frequencies.

Pande and Abdel-Aty (2006) used both historical crash and real-time traffic parameters obtained from loop detectors to calibrate the neural network (NN) models to predict the occurrence of lane-change-related freeway crashes. The results indicated that these models may be applied for identifying real-time traffic conditions prone to lane-change-related crashes. Relative studies of the NN model on freeways are found in several references (Cheu and Ritchie, 1995; Chang, 2005; Kononov et al., 2008). Abdel-Aty et al. (2004) also applied another machine learning method, the matched case-control logistic regression, for predicting freeway crashes based on loop detector data.

In addition to safety analysis on freeways, the machine learning technique is also widely used in other roadway facilities and fields related to traffic safety. For example, Kuhnert et al. (2000) employed logistic regression, CART, and Multivariate Adaptive

Regression Splines (MARS) to analyze motor-vehicle injury data. Karlaftis and Golias (2002) applied hierarchical tree-based regression (HTBR) to analyze the effects of road geometric and traffic characteristics on crash rates for rural two-lane and multilane roads. Haleem et al. (2010) fitted and compared the NB and MARS models by using data collected on unsignalized intersections in Florida. Their results showed that MARS is a promising technique for predicting crashes, especially for continuous response variables.

### ***2.3.2. Intersections***

Crashes commonly occur at intersections, which may be due to both design characteristics and the fact that entry by a vehicle from one roadway could be expected to result in conflicting movements with a vehicle entering from opposite roadways. Severe crashes tend to occur within the intersection buffer because left-turn and angle collisions are the most frequent crash types at intersections. Thus, it is important to identify the methods that can assess the effects that geometric, traffic flow, traffic control, environmental, and operational characteristics have on traffic crashes at intersections (Abdel-Aty and Keller, 2005).

Various models were developed to study the relationship between crashes at intersections and influencing factors. Bauer and Harwood (2000) applied multiple linear regression analysis in developing crash prediction models for at-grade intersections in California, using three years of crash data (1990 to 1992), as well as geometric design, traffic control, and traffic volume data from a database provided by the California Department of Transportation (Caltrans). Five types of intersections were modeled: rural four-leg stop-controlled intersections, rural three-leg stop-controlled intersections, urban four-leg stop-controlled intersections, urban three-leg stop-controlled intersections,

and urban four-leg signalized intersections. The multiple linear regression was used for urban four-leg stop-controlled and signalized intersections, while Poisson and NB regression were used for the remaining types.

The advantage of using Poisson and NB regression models over the multiple linear regression models has been confirmed by previous investigations. Several studies (Poch and Mannering, 1996; Sayed and Rodriguez, 1999; Harnen et al., 2003; and Salifu, 2004) presented the empirical relationships obtained through the Poisson regression analyses and/or NB techniques, relating the crashes with traffic flow, traffic control, and key highway geometric design variables. The results indicated that roadway geometric, vehicular, and operational features had an effect on crash frequency. Therefore, those factors that significantly affect crashes should be given more attention in crash analyses at intersections (Nambuusi et al., 2008).

In addition to widely used Poisson and NB regression models, the machine learning technique was also used in crash analysis at intersections. Lau et al. (1989) introduced the crash prediction models for urban intersections based on the CART technique, which is used to group crashes based on crash and intersection types by splitting the data into branches on a tree diagram. The prediction model involved three levels: Level 1, Generation of the base model; Level 2, Grouping intersections by CART; and Level 3, Adjustment by Crash History. This technique can be used to obtain the number of crashes for each injury severity, as well as the number of crashes for each intersection type.

Various techniques were used to assess the goodness-of-fit of different models, including the deviance, the Chi-square statistic, the adjusted R-square, and the pseudo

R-square (likelihood ratio index). It was difficult to compare the goodness-of-fit among the models because different measures were used, and the fitted models had different objectives (Nambuusi et al., 2008).

## 2.4. Simple SPFs

Simple SPFs, also called traffic SPFs, are mathematical models that link crash occurrence to traffic volume at specific roadway types. The simple SPFs take the form of an NB model, as adopted in the HSM. For simple SPFs, a set of AMFs is applied for prediction purposes to adjust from base conditions to prevailing conditions.

### 2.4.1. Roadway Segments

#### 2.4.1.1 Arterials

Persaud (1992) developed an SPF using the data from 1988 to 1989 for rural two-lane roads in Ontario, Canada. The functional form used is shown as follows:

$$\text{Crashes/km/year} = a \times (\text{AADT}/1000)^b \quad (2-7)$$

The regression coefficients  $a$  and  $b$  are given in Table 2-1.

**Table 2-1 Regression Coefficient Estimates**

	Total Crashes				Fatal and Injury Crashes			
Lane width	<6.1 m	<6.1 m	>6.1 m	>6.1 m	<6.1 m	<6.1 m	>6.1 m	>6.1 m
Shoulder width	<1.8 m	>1.8 m	>1.8 m	<1.8 m	<1.8 m	>1.8 m	>1.8 m	<1.8 m
$b$	0.733	0.892	0.892	0.733	0.783	0.971	0.971	0.783
$a$	0.00287	0.00096	0.00069	0.0025	0.00067	0.00018	0.00012	0.00054

Using data on two-lane rural roads in State of New York for the period of 1971 through 1987, Hauer (1994) developed the following SPF to estimate total crashes, excluding crashes at intersections:

$$\text{Crashes/km/year} = 0.00244 \times (\text{AADT})^{0.776} \quad (2-8)$$



Using the 1988 and 1989 data for urban two-lane roads in Ontario, Canada, Persaud (1992) developed the following SPF:

$$\text{Crashes/km/year} = 0.00369 \times (\text{AADT})^{0.72} \quad (2-9)$$

Persaud (1992) further developed the SPF below using the 1988 and 1989 data for multi-lane roads with no full access control in Ontario, Canada:

$$\text{Crashes/km/year} = a \times (\text{AADT})^b \quad (2-10)$$

#### 2.4.1.2 Freeways

Persaud (1991) presented a method for estimating the underlying crash potential on freeways. The method first used NB regression models to produce an initial estimate of a segment's crash potential on the basis of its traffic. This estimate was then refined by being combined with the segment's crash count, using an EB procedure. The NB prediction model is as follows:

$$\text{Crashes} = 0.6278 \times L \times (\text{AADT}/1000)^{1.024} \quad (2-11)$$

where

$L$  = segment length, and

$\text{AADT}$  = the annual average daily traffic.

The author emphasized that the geometric features were not considered in the prediction model because these variables occurred with remarkable consistency on freeways with higher design criteria. The precision of the model, therefore, would not be improved with additional variables.

Persaud (1992) developed the SPF shown below for freeways based on the 1988 and 1989 data from Ontario, Canada. The regression coefficients  $a$  and  $b$  are given in Table 2-2.

$$\text{Crashes/km/year} = a \times (\text{AADT})^b \quad (2-12)$$

**Table 2-2 Regression Coefficient Estimates**

Crash Type	Number of Lanes	a	b
Total	4	0.0000474	1.155
Total	>4	0.0000978	1.113
Fatal + Injury	4	0.0000206	1.136
Fatal + Injury	>4	0.0000122	1.212

Huang et al. (1991) developed the following two SPFs for total and fatal and injury crashes, respectively, based on data from California freeways:

$$\text{Total crashes} = 0.65 + 0.666 \times \text{Million Vehicle Miles} \quad (2-13)$$

$$\text{Fatal + Injury crashes} = 0.166 + 0.263 \times \text{Million Vehicle Miles} \quad (2-14)$$

Kiattikomol et al. (2008) generated regression models for crash prediction on interchange and non-interchange segments for urban freeways on a planning level. The impacts of interchanges on freeways were discussed, and prediction models were generated for interchange segments and non-interchange segments, respectively. For example, the models used for Tennessee freeways are as follows:

For non-interchange segments:

$$N = a(L)^{b_1} (\text{AADT})^{b_2} \quad (2-15)$$

For interchange segments with four lanes:

$$N = (L)^{b_1} (\text{AADT})^{b_2} \quad (2-16)$$

where

$N$  = expected number of crashes in a three-year period,

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

$L$  = segment length (miles), and

$a, b_1,$  and  $b_2$  = estimated parameters.

Freeway segments are characterized by the presence of interchange areas. Traditionally, there are two approaches of analyzing freeway interchange areas with respect to their mainline counterparts. The first approach is to consider both freeway mainlines and interchange areas as one entity while developing SPFs (Kraus et al., 1993; Tegge and Ouyang, 2007; Kononov and Allery, 2004; Persaud and Dzbik, 1993; and Konduri and Sinha, 2002). For example, one of the known methods is to develop SPFs using freeway segments identified from the center of one interchange to the next (Persaud and Dzbik, 1993; Konduri and Sinha, 2002).

The second approach is to consider the freeway mainlines and interchange influence areas as two separate entities in the analysis, as recommended by *SafetyAnalyst* (2010). Very few previous studies on freeways have considered interchange influence area as a specific analysis category. For example, Kiattikomol et al. (2008) developed planning level prediction models on interchange and non-interchange segments (i.e., basic freeway segments) for the states of North Carolina and Tennessee. In their study, it was concluded that the models for interchange and non-interchange freeway sections should be developed separately. Lu et al. (2012) also emphasized that interchange influence areas should be considered a separate category instead of developing freeway SPFs, regardless of the influence of interchanges, as the crash performance of freeway sections with interchange influence areas is different from that of basic freeway mainline segments.

#### **2.4.2. Intersections**

Lau et al. (1989) used the 1986 through 1988 data for signalized intersections from the state of California to develop SPFs for intersections. Separate models were

developed for fatal, injury, and PDO crashes, and three levels of estimation were used, given the availability of information for the intersections. At level 1, the following three SPFs were used if the volume of traffic entering an intersection is known:

$$\text{Fatal Crashes/year} = 0.018 \quad (2-17)$$

$$\text{Injury Crashes/year} = 0.61856 + 0.16911 \times \text{Million Entering Vehicles} \quad (2-18)$$

$$\text{PDO Crashes/year} = 4.6029 + 0.5142 \times \text{Million Entering Vehicles} \quad (2-19)$$

If further information is available about an intersection, such as design and control characteristics, proportion of cross street traffic, and environmental features, then level 2 estimates were used. At level 2, intersections were classified by group, and a “group constant” was added to the value estimated by the SPFs in Equations (2-17), (2-18), and (2-19). Groups were separated by fatal, injury, and PDO crashes. Level 3 was used when the individual crash history of an intersection was available, in addition to the information for levels 1 and 2. An important point to note is that level 3 was based on EB, and the results represent future safety estimates of existing intersections. Moreover, Lau et al. (1989) used a four-legged urban signalized intersection with AADT of 49,000 and 10,000 for major and minor streets, respectively. All approaches in this investigation were two-lane, and the signal control was a pre-timed cycle with permitted left turns. The design speed was 50-54 mph.

McDonald (1953) used data from rural unsignalized intersections located on divided highways to develop the following SPF:

$$\text{Crashes/year} = 0.000783 \times (AADT_{\text{major}})^{0.455} \times (AADT_{\text{minor}})^{0.633} \quad (2-20)$$

Using the HSIS data from 1985 to 1987 for 125 rural unsignalized intersections in the state of Minnesota, Bonneson and McCoy (1993) developed the following SPF:

$$Crashes/year = 0.000379 \times (AADT_{major})^{0.256} \times (AADT_{minor})^{0.831} \quad (2-21)$$

Webb (1955) used data from 96 signalized intersections on high-speed rural state roadways in State of California. The following SPF was developed:

$$Crashes/year = 0.00703 \times (AADT_{major})^{0.51} \times (AADT_{minor})^{0.29} \quad (2-22)$$

Using the HSIS data for rural signalized intersections, Bonneson et al. (1993) developed the following SPF:

$$Crashes/year = 0.00703 \times (AADT_{major})^{0.7213} \times (AADT_{minor})^{0.3663} \quad (2-23)$$

McGee et al. (1989) developed the following crash rates in Table 2-3 using data from urban unsignalized intersections in the cities of Seattle, Milwaukee, Rapid City, and Madison.

**Table 2-3 Crash Rate (Crashes/Million Entering Vehicles)**

Major street AADT	Minor street AADT						
	100	300	500	700	900	1250	2000
250	2.19	2.09	2.01	1.99	2.03	1.72	1.22
750	1.06	1.44	1.53	1.57	1.58	1.49	1.14
1250	0.73	1.15	1.25	1.31	1.34	1.36	1.09
1750	0.64	0.92	1.12	1.26	1.19	1.17	0.91
2500	0.53	0.73	0.90	1.02	1.04	0.99	0.88
3500	0.43	0.57	0.69	0.80	0.83	0.81	0.75

Using data on urban signalized intersections of one-way streets in Philadelphia, Persaud et al. (1995) developed the SPF, as follows:

$$Crashes/year = a \times (AADT_{major})^b \times (AADT_{minor})^c \quad (2-24)$$

the regression coefficients a, b, and c are shown in Table 2-4.

**Table 2-4 Regression Coefficient Estimates**

Parameters	Crash Types		
	Right-angle and Turn	Rear-end	Pedestrian
<i>a</i>	0.0002037	0.0002099	0.0009039
<i>b</i>	0.5941	0.6758	0.5150
<i>c</i>	0.354	0	0

**2.4.3. Ramps**

Jovanis and Chang (1986) used Poisson regression to model the relationship between crash frequency, traffic volumes, and weather conditions. A more general form of the Poisson regression and NB models was later used to explore the relationship between crash frequencies, daily traffic, and highway geometric design variables (Miaou, 1994; Le and Porter, 2012). In the NB model, the expected number of crashes of type *i* on segment *j* is expressed as:

$$\mu_{ij} = E(Y_{ij}) = \exp[(\beta x_j + \ln(L_j))] \tag{2-25}$$

where

$\mu_{ij}$  = the expected number of crashes of type *i* on segment *j*,

$\beta$  = regression coefficients estimated with maximum likelihood that quantify the relationship between  $E(Y_{ij})$  and variables in *X*,

$x_j$  = a set of traffic and geometric variables characterizing segment *j*, and

$L_j$  = length of segment *j*.

Parajuli et al. (2006) developed simple SPFs with the following form for ramps considering AADT to be the only independent variable:

$$\text{Crashes/year} = a \times (\text{AADT})^b \times e^{(\text{length})} \tag{2-26}$$

where

*AADT* = annual average daily traffic volume (vehicles per day),

$Length$  = ramp length in mile, and  
 $a, b$  = regression coefficients.

## 2.5. Segmentation

An essential step in generating an SPF is to divide the roadway into individual segments. The usual approach is to divide road sections into segments of variable length according to roadway features (Geyer et al., 2008; Vogt and Bared, 1998). This is a desirable feature, as it allows a specific level of crash statistics to be associated with a uniform set of roadway conditions. The advantage of this method is that each divided segment has a set of unique attributes. However, if a large number of attributes are included, the method could quickly result in many short segments that are more vulnerable to random crash fluctuations. On the other hand, attributes such as number of lanes and shoulder width, which often occur with high consistency along freeways and expressways, may result in very long segments. In addition, locations with high crash rates, due to changes in a geometric attribute, such as changing from three lanes to two lanes, may not be detected because the crashes are split between two segments that the variable-length method created as a result of the change in the number of lanes (Zhong et al., 2007).

Other studies have used a fixed length (Shankar et al., 1995) to divide roadway sections. Using this method, roadway sections are divided into non-overlapping segments of a constant length. For example, a 0.3-mile segment was used in a study in the state of New York by Geyer (2008). Many states (e.g., Kansas) used pre-determined segments, which are defined by jurisdiction boundaries (Geyer, 2008). Although the fixed-length method is simple, there is potential for the correlation between crash

occurrence and roadway characteristics to be ignored, as different values of the same attribute may be aggregated into one segment, or in the case of a roadway section, the section with the unchanged attribute value may be divided into two segments.

A newer method for dividing segments is through clustering. Despite its promise in reducing the heterogeneity in crash data (Wong and Chung, 2008), it has not been extensively applied for roadway segmentation. Fraley and Raftery (2002) argued that clustering is based on maximizing the similarity between in-cluster elements, and the dissimilarity between inter-cluster elements. Xu and Wunsch (2005) indicated that clustering has been applied in vast fields, ranging from engineering (machine learning, pattern recognition, mechanical engineering), computer sciences (web mining), medical sciences (biology, microbiology), social sciences (sociology, psychology), and economics (marketing).

Depaire et al. (2008) used the latent class clustering method for segmenting traffic crashes. The authors segmented the data into clusters representing specific crash types. They concluded that the clustering revealed the effect of a single independent variable, which differed for different traffic crashes. Karlaftis and Tarko (1998) argued that clustering methods are considered a useful tool to segment crash data, thereby rendering the modeled data homogeneous. The first attempt to use Fisher's clustering method (Fisher, 1958) for roadway segmentation can be found in Zhong et al. (2007). They indicated that this specific clustering concept could be used to divide the roadway sections for generating a crash prediction model. They then used the crash frequency per mile distributed along the roadway as their clustering index. Zhong et al. found that the crash frequency of each segment obtained by this method could better follow a



probability distribution, which should help improve the calibration of crash prediction models.

Other applications of Fisher's technique can be found in Xu (2005), and Wang and Tong (2009). Xu (2005) applied this technique to detect abnormal fluctuations of temperature based on grouping the boiler pipe temperature, which is a set of ordinal samples. The same technique was applied by Wang and Tong (2009) to estimate the priority of road sections for maintenance based on specific road pavement indices.

Apart from being a clustering method in general, Fisher's algorithm is designed for situations in which the data points are ordered (Cappelli et al. 2005), or in other words, in which the sample sequence will not be disrupted during the clustering process. The units of roadway section can be considered a set of ordered samples, while crash frequency or crash severity can be considered the crash indices. The units of neighboring sections with similar crash indices are integrated into the same segment, while those with an obvious difference in crash indices are assigned separate segments. As crashes directly reflect the geometric, vehicular, and operational characteristics of roadways where crashes occur, it is possible to measure the influence of these independent variables (geometric, vehicular, and operational characteristics) on the dependent variable (e.g., crash frequency) and, thus, improve the precision of the SPF calibration (Zhong et al. 2007).

## **2.6. Identification of High Crash Locations**

Various network screening methods have been used to identify high crash locations. Among these methods, crash frequency (Zegeer, 1986) is the simplest. In this method, locations are ranked by descending crash frequency. Those with more than

a predetermined number of crashes are classified as high crash locations. While this simple method has limited data requirements, it does not account for exposure (e.g., traffic volume) and threshold (HSM, 2000). This has led to the use of the crash rate method (Morin, 1967), which attempts to account for the difference in crash experience due to traffic exposure. However, this method may result in the selection of low-volume sites that have relatively few crashes. The frequency-rate method (Laughland et al., 1975) was introduced in an attempt to correct the shortcomings of both methods. This method first ranks the location by crash frequency and removes those with less than a certain number of crashes. The remaining locations are re-ranked using crash rates; those with less than a certain crash rate are then removed, leaving locations that are assured of having a minimum crash frequency and crash rate.

Examples of other methods used include the equivalent property damage only (EPDO) method (Donnell and Mason, 2004), and rate quality control method (Stokes and Mutabazi, 1996; Norden et al., 1956). The EPDO method compares the relative importance of crashes that result in property damage with that of fatal and injury crashes. This method considers crash severity, but also does not account for traffic exposure and threshold. The rate quality control method has been commonly used in practice. It compares the observed crash rate at each site with a critical crash rate for similar locations. Sites that have crash rates greater than the critical crash rate are identified as high crash locations. This method considers exposure and provides the statistical reliability level.

While these existing methods may differ in their approaches, all share the same problem of not accounting for the RTM bias. Also known as a “regression artifact” or

“selection bias,” the RTM bias is a statistical phenomenon that may lead to the identification of locations with high crashes due merely to random fluctuations in crash numbers. The EB method (Shen and Gan, 2003; Hauer, 1997) was introduced as a way to account for the RTM bias associated with the traditional network screening methods. The EB method adjusts for the RTM bias by making use of the crash information from reference sites, which is measured through an SPF. As noted previously, an SPF is most effective when calibrated to local conditions.

## **2.7. Summary**

In this chapter, a comprehensive literature review was conducted to investigate the development of SPFs and its applications. The purposes of the review are to understand the current status of the research area, identify the pending problems to be solved, determine the objectives, and form the research framework and tasks for this dissertation. The major findings of the literature review are summarized below.

With the recent release of advanced safety analysis tools, including *SafetyAnalyst*, various states are trying to assess the need for developing agency-specific SPFs. Even though *SafetyAnalyst* provides national default SPFs calibrated to local agency data, it is believed that the use of state-specific SPFs might better identify problematic sites.

The literature review extensively highlighted various studies addressing SPFs. Numerous methods and techniques were employed by both researchers and agencies to establish the relationship between traffic crashes and highway geometric variables, traffic characteristics, and environmental factors. However, the models used for calibrating SPFs were varied based on specific cases, and thus had limited applicability. The most recently released safety analysis tool of *SafetyAnalyst* remedies this limitation by

providing the SPFs for specified roadway types in a standardized form, at the national level. It is recommended to recalibrate the agency SPFs based on local data, following the criteria and techniques provided therein. In order to better represent local crash experience in this dissertation, the local SPFs, specifically reflective of Florida's conditions, were calibrated in this research.

Among all categories in *SafetyAnalyst*, freeways within interchange influence areas were in need of special discussion. Although there was evidence that the crash frequency for freeways within interchange influence areas was significantly different from basic freeway segment, the two types of segments were not analyzed separately. Very few previous studies have discussed the process of separation of interchange influence areas from basic freeway segments. Lack of accepted guidance on the definition of an interchange influence area and difficulty in obtaining roadway inventory databases with information associated with the start and end mileposts or the interchange length are two main reasons that this has not been done. Considering the fact that traffic flow characteristics for interchange influence areas and basic freeway segments are considerably different, developing separate models for these two categories is clearly needed.

The applications provided by SPFs were discussed, including developing crash prediction models and identifying high crash locations. Crash prediction models and the identification of high crash locations are closely related to the precision of SPFs. High crash locations are sites that experienced an abnormally high number of crashes when compared to similar sites. Various network screening methods have been used to identify high crash locations. While these existing methods may differ in their approaches, they

all share a common limitation, i.e., they do not account for the RTM bias. The EB method was introduced as a way to account for the RTM bias. The EB method adjusts for the RTM bias by making use of the crash information, not only from the treatment site, but also from reference sites. The crash information from the reference sites is measured via a SPF.

Roadway segmentation is an essential step in the SPF calibration. The traditional approach is to divide roadway sections into segments of variable length according to roadway features, or of a fixed length. This chapter also introduced and provided a clustering method developed by Fisher as a way to improve existing segmentation methods for both SPF calibration and screening process, for the purpose of identifying high crash locations.

## CHAPTER 3

### METHODOLOGY

#### 3.1. Introduction

This chapter describes the methodology applied in this dissertation research. The first step involves the acquisition of the crash and geometric data. The crash data are extracted from the Florida Department of Transportation's (FDOT) Crash Analysis Reporting (CAR) system, and the geometric data are extracted from the Roadway Characteristics Inventory (RCI) database. Both sets of data are then imported into a program called Dynamic Segmentation (DySeg), which was developed in-house at the Lehman Center for Transportation Research (LCTR) at Florida International University.

An essential step in calibrating an SPF is the segmentation of roadway sections. DySeg is able to perform roadway segmentation based on fixed and variable length. During the segmentation process, crashes associated with each segmented facility (segment or intersection) are summarized. The program produces an output file that contains both roadway and crash information. This output data set is then used to develop SPFs. In addition, an alternative segmentation method based on Fisher's clustering algorithm is examined in this step to improve the development of SPFs, for the purpose of identifying high crash locations. The segmentation methods are described in Section 3.2.

The main objective of this dissertation is to develop Florida-specific SPFs for different roadway types in *SafetyAnalyst*. The SPF functional forms for both roadway segments and intersections through the use of NB models are introduced in Section 3.3. Calibration factor, which is used for calibrating the national default SPFs to Florida data,

is also described in this section. Assessment of the performance of SPFs was performed by several goodness-of-fit tests, as shown in Section 3.4. Also included are the overdispersion parameter, the cumulative residuals (CURE) plot, the mean square error (MSE), the mean absolute deviance (MAD), the mean square prediction error (MSPE), and the Freeman-Tukey  $R^2$  coefficient ( $R^2_{FT}$ ).

Section 3.5 presents the EB method and the potential for safety improvements (PSI), which are applied to the process of identifying HCLs in *SafetyAnalyst*, utilizing an SPF that combines the crash history data of a treated site with the predicted crash frequency of similar reference sites.

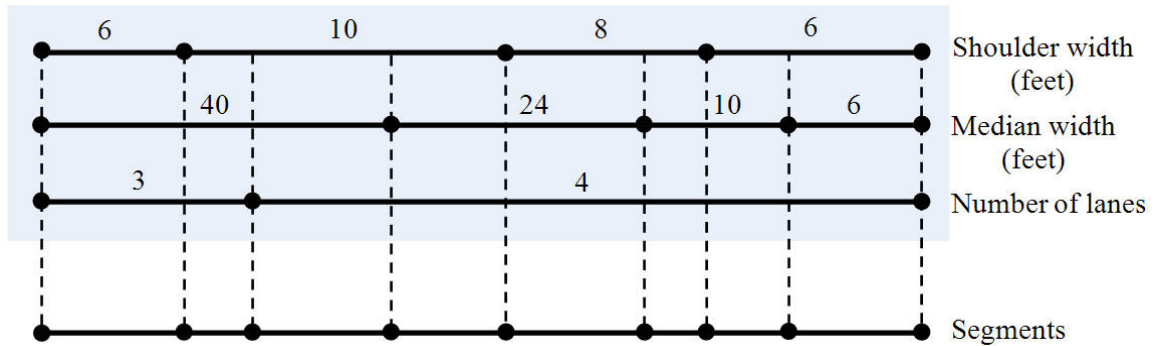
## **3.2. Segmentation**

In this step, the roadway is divided into individual sites consisting of homogenous roadway segments and intersections. Intersections are discrete entities. For intersections in this research, the prediction model estimates both the crashes that occur within the limits of an intersection and intersection-related crashes that occur on the intersection legs located 250 feet from the center. For roadway segments, there are three methods for segmentation: variable length, fixed length, and Fisher's clustering method. The limitations and strengths of the three methods were discussed in the literature review.

### ***3.2.1. Fixed Length and Variable Length***

The fixed-length method consists of roadway sections that are divided into segments of equal length. The variable length method consists of roadway sections that are divided into homogeneous segments of variable lengths, with the segment length defined by any change in the geometric or traffic characteristics. For example, if there

is a change in the shoulder width, median width, number of lanes, or other related characteristics, a new segment is defined. Figure 3-1 shows how such segments are created using this method.



**Figure 3-1 Variable Length Segmentation Method**

### ***3.2.2. Fisher's Clustering Method***

A clustering method based on a grouping strategy developed by Fisher (1958) is applied to improve the shortcomings of existing segmentation methods. Fisher's algorithm is designed for situations in which the data points are ordered. The units of roadway section can be considered as a set of ordered samples, while crash frequency or crash severity can be considered as crash indices. The units of neighboring section with similar crash indices are integrated into the same segment, while those with an obvious difference in crash indices are assigned separate segments. As crashes directly reflect the geometric, vehicular, and operational characteristics of roadways where crashes occur, it is possible to measure the influence of these independent variables (geometric, vehicular, and operational characteristics) on the dependent variable (e.g., crash frequency) and, thus, improve the precision of the SPF calibration. If a roadway attribute (e.g., shoulder width) has more than one value within the same divided segment,



the weighted average method is implemented to estimate the final value. Specifically, the final weighted average value is estimated, taking into account the attribute value and the length it occupies within the segment.

In Fisher's clustering method,  $x_1, x_2, \dots, x_n$  are assumed to be ordered samples (total crash frequency in each unit of neighboring roadway section in this research). The  $n$  samples are divided into  $k$  groups and  $\{x_i, x_{i+1}, \dots, x_j\}$  ( $1 \leq i \leq j \leq n$ ) represents one of these groups.

For Fisher's algorithm, "squared distance" is used to measure the within-group difference, as follows:

$$D(i, j) = \sum_{l=i}^j (x_l - \bar{x}_{ij})^2 \quad (3-1)$$

where  $x_l$  is the characteristic of interest used for grouping,  $\bar{x}_{ij}$  is the average of  $x_i, x_{i+1}, \dots, x_j$ , and  $D$  is the sum of squares within groups in the sense of the variance analysis (also called squared distance). The smaller the squared distance, the smaller the within-group difference is.

Assuming the function  $p(n, k): \{i_1, i_1+1, \dots, i_2-1\}, \{i_2, i_2+1, \dots, i_3-1\}, \dots, \{i_k, i_k+1, \dots, n\}$ , Fisher's clustering groups  $n$  ordered samples into  $k$  mutually exclusive subsets ( $k \leq n$ ) by minimizing the sum of squared distance  $D$  for each group, as shown in the following objective function:

$$f[p(n, k)] = \min \left\{ \sum_{j=1}^k D(i_j, i_{j+1} - 1) \right\} \quad (3-2)$$

where  $D(i_j, i_{j+1} - 1)$  is the squared distance for the  $j^{\text{th}}$  group ( $1 \leq j \leq k$ ).

Using Equation (3-2), a set of recursion formulas can be derived to obtain the optimal division. The optimal division for dividing  $n$  samples into two groups ( $k=2$ ) (i.e., the optimal 2-division) is:

$$f[p(n,2)] = \min_{2 \leq j \leq n} \{D(l, j-1) + D(j, n)\} \quad (3-3)$$

The optimal division for dividing  $n$  samples into three groups ( $k=3$ ) (i.e., the optimal 3-division) is:

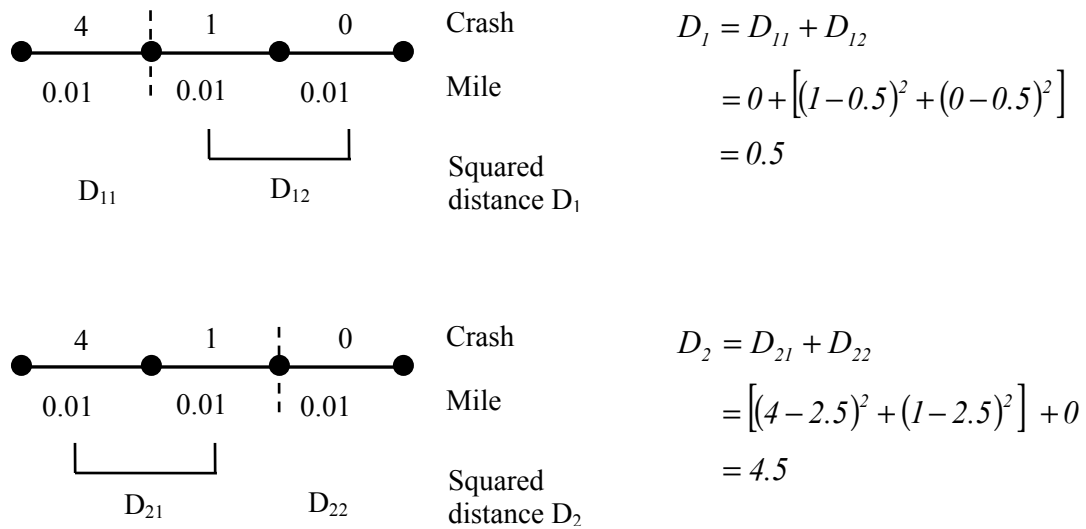
$$f[p(n,3)] = \min_{3 \leq j \leq n} \{f[p(j-1,2)] + D(j, n)\} \quad (3-4)$$

Based on Equations (3-3) and (3-4), the general formula for the optimal  $k$ -division (i.e., the optimal division for dividing  $n$  samples into  $k$  groups) is:

$$f[p(n,k)] = \min_{k \leq j \leq n} \{f[p(j-1, k-1)] + D(j, n)\} \quad (3-5)$$

Figure 3-2 shows an example on how to obtain the optimal division. In this figure, three samples ( $n=3$ ) are required to be divided into two groups ( $k=2$ ). Following the method mentioned above, the squared distances  $D$  for each potential division are calculated. The division with the smaller sum of squared distance (i.e.,  $D_j$ ) is the optimal division.

To obtain the optimal  $k$  subset, the process is repeated until the difference between the current optimal  $(k+1)$ -division, i.e.,  $f[p(n, k+1)]$ , and the previous optimal  $k$ -division, i.e.,  $f[p(n, k)]$ , is less than a small value, and the resulting  $k$  value in the final iteration is treated as the optimal number of subsets.



**Figure 3-2 Sample Data Illustrating Fisher's Clustering Method**

### 3.3. Model Development

After segmentation, SPFs for different categories can be developed. Analysis is performed on both total and FI crashes. The NB model is used to develop the SPFs.

#### 3.3.1. Negative Binomial (NB) Models

The NB, or Poisson-gamma regression model is the most widely used method in predicting crash frequency. The characteristics of crash frequency, specifically that it follows a gamma distribution and that the variance of the crash counts exceeds the mean, determine the priority of the NB regression when generating a crash prediction model. The formula for the expected crash number is given below (Miaou, 1994):

$$\lambda_i = \exp(\beta x_i + \varepsilon_i) \quad (3-6)$$

where  $\exp(\varepsilon_i)$  is a gamma distributed error term with mean 1 and variance  $\alpha$ . The addition of this term allows the variance to differ from the mean as  $Var = \lambda_i + \alpha \lambda_i^2$ , where  $\alpha$  is the overdispersion parameter. The Poisson regression model is a limiting

model of negative binomial regression models as  $\alpha$  approaches zero. The probability function of NB distribution is shown as follows:

$$P(Y = y_i) = \frac{\Gamma(y_i + (1/\alpha))}{y_i! \Gamma(1/\alpha)} \cdot \left( \frac{\alpha \lambda_i}{1 + \alpha \lambda_i} \right)^{y_i} \cdot \left( \frac{1}{1 + \alpha \lambda_i} \right)^{1/\alpha} \quad (3-7)$$

where

$\Gamma(x)$  = gamma function, and

$y_i$  = number of crashes per period for roadway segment  $i$ .

The likelihood function results from the probability function, as shown:

$$L(\lambda_i) = \prod_i \frac{\Gamma(y_i + (1/\alpha))}{y_i! \Gamma(1/\alpha)} \cdot \left( \frac{\alpha \lambda_i}{1 + \alpha \lambda_i} \right)^{y_i} \cdot \left( \frac{1}{1 + \alpha \lambda_i} \right)^{1/\alpha} \quad (3-8)$$

The advantage of this technique is that the NB model can overcome the possibility of overdispersion in crash frequency data, which generally follows a gamma probability distribution.

### 3.3.2. *SPF Functional Form for Roadway Segments and Ramps*

Roadway segments and ramps shared the same functional form. The SPF functional form for roadway segments and ramps is as follows:

$$N_{predicted} = e^a \times AADT^b \quad (3-9)$$

To generate regression models, Equation (3-9) is rewritten as:

$$N_{predicted} = \exp(a + b \times \ln(AADT)) \quad (3-10)$$

where

$N_{predicted}$  = predicted crash frequency per mile per year,

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

$a, b$  = regression coefficients.

The overdispersion parameter, which indicates the statistical reliability of the SPF, was used to account for dispersion in the data. The closer the overdispersion parameter is to zero, the more statistically reliable the SPF. The goodness-of-fit statistic, Freeman-Tukey  $R^2$  coefficient ( $R^2_{FT}$ ), was used to assess NB regression performance.

Calibration of the default SPFs was performed by multiplying the default SPFs by a “calibration factor”,  $C$ , as follows:

$$C = \frac{\sum_{All\ sites} observed\ crashes}{\sum_{All\ sites} predicted\ crashes} \quad (3-11)$$

As shown in Equation (3-11), the calibration factor is calculated as the ratio of the total observed crashes to total predicted crashes obtained from the default national SPFs. Note that “*All sites*” in the equation refers the reference sites within a specific category. Moreover, the calibration factor,  $C$ , is not needed if a local jurisdiction chooses to calibrate its own SPFs as the local safety trends are inherently addressed in the coefficients per Equation (3-10).

### 3.3.3. *SPF Functional Form for Intersections*

The model form similar to the model used to generate default SPFs (used within *SafetyAnalyst*) was used to generate Florida-specific SPFs for intersections:

$$N_{predicted} = e^a \times AADT_{major}^b \times AADT_{minor}^c \quad (3-12)$$

For regression modeling, Equation (3-12) is rewritten as:

$$N_{predicted} = \exp(a + b \times \ln(AADT_{major}) + c \times \ln(AADT_{minor})) \quad (3-13)$$

where

$N_{predicted}$  = predicted target crash frequency per intersections per year,  
 $AADT_{major}$  = average annual daily traffic volume on the major-road approaches,  
 $AADT_{minor}$  = average annual daily traffic volume on the minor-road approaches,  
 and  
 $a, b, c$  = regression coefficients that are estimated from the available data.

### 3.4. Model Tests

Several goodness-of-fit measures were used to assess the performance of SPFs, including the overdispersion parameter, CURE plot, MSE, MAD, mean MSPE, and Freeman-Tukey  $R^2$  coefficient.

CURE plot is used to assess how well predictions fit the data over the full range of an independent variable. In this method documented by Hauer and Bamfo (1997), the cumulative residuals, which are defined as the cumulative differences between the observed crashes and predicted crashes for each site, are plotted in increasing order of AADT. Well-fitted models are identified by the cumulative residuals that oscillate around 0 and fall within two standard deviations ( $+2\sigma$  and  $-2\sigma$  boundaries).

The MSE is applied to the calibration data (Young and Park, 2012). The MSE is typically a measure of model error associated with the calibration data (Washington et al., 2005). The following equation shows the formula of MSE:

$$MSE = \frac{1}{n - p} \sum (N_{observed_i} - N_{predicted_i})^2 \quad (3-14)$$

where

$n$  = number of sites in the calibration dataset,

$p$  = number of parameters in the statistical model,

$N_{observed_i}$  = observed crash frequency for site  $i$ , and

$N_{predicted_i}$  = predicted crash frequency for site  $i$ .

To assess the significant prediction performance, two evaluation criteria, the MAD and the MSPE, are applied to the validation (prediction) data. The two criteria are respectively defined in the equations below (Haleem et al., 2010; Li et al., 2008):

$$MAD = \frac{1}{n} \sum |N_{observed_i} - N_{predicted_i}| \quad (3-15)$$

$$MSPE = \frac{1}{n} \sum (N_{observed_i} - N_{predicted_i})^2 \quad (3-16)$$

where

$n$  = number of sites in the prediction dataset,

$N_{observed_i}$  = observed crash frequency for site  $i$ , and

$N_{predicted_i}$  = predicted crash frequency for site  $i$ .

MAD is used to estimate the prediction deviation, while MSPE is employed for determining the variance of the difference between predicted and observed results. The prediction performance is better if the values of MAD and MSPE are smaller. MSPE can be compared with the MSE to check for over-fitted models ( $MSPE > MSE$ ) or under-fitted models ( $MSPE < MSE$ ) (Young and Park, 2012).

The Freeman-Tukey  $R^2$  coefficient generally describes how good the model fit is, where a higher value indicates a good fit and vice-versa. Fridstrom et al. (1995) discussed this statistic in detail. The  $R^2$  for the Freeman-Tukey transformed variables is:

$$R^2_{FT} = 1 - \frac{\sum \hat{e}_i^2}{\sum (f_i - \bar{f})^2} \quad (3-17)$$

In Equation (3-17), the variance stabilizes transformation of the variable  $y_i$ :

$$f_i = \sqrt{y_i} + \sqrt{y_i + 1} \quad (3-18)$$

The statistic is approximately normally distributed with mean:

$$\phi_i = \sqrt{4\hat{y}_i + 1} \quad (3-19)$$

The deviation is estimated by the corresponding residual:

$$\hat{e}_i = f_i - \phi_i \quad (3-20)$$

where

- $y_i$  = observed crashes at site  $i$ ,
- $\hat{y}_i$  = mean of observed crashes at sites similar to site  $i$ , and
- $\bar{f}$  = average of  $f_i$  for sites considered.

### 3.5. Identification of High Crash Locations

The EB method calculates the weighted average of the crash history of a site and the predicted crash frequency of similar reference sites (calculated using an SPF). Existing studies that have used the EB method for network screening include Persaud et al. (1999), Hauer et al. (2002), and Elvik (2007). The expected crash frequency is calculated as the weighted average of predicted and observed crash frequencies:

$$N_{expected} = w \times N_{predicted} + (1-w) \times N_{observed} \quad (3-21)$$

where

- $N_{expected}$  = expected crash frequency at the treatment site,
- $N_{predicted}$  = predicted crash frequency based on an SPF,



$N_{observed}$  = observed crash frequency at the treatment site, and

$w$  = weighting factor that is estimated based on the following equation:

$$w = \frac{1}{1 + ODP \times (\sum_{\substack{\text{all} \\ \text{study} \\ \text{years}}} N_{predicted} \times L)} \quad (3-22)$$

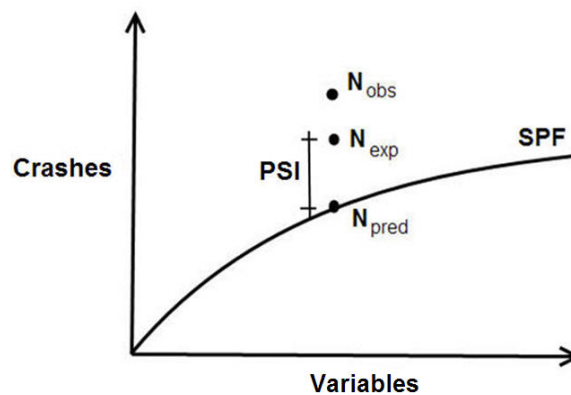
where

$ODP$  = overdispersion parameter estimated from the associated SPF, and

$L$  = segment length in miles.

The ranking of HCLs is based on the PSI (FHWA, 2010; Gan et al., 2012), which is calculated as the difference between the expected and predicted crash frequencies.

An illustration of the EB method and PSI is shown in Figure 3-3. The predicted crash frequency ( $N_{predicted}$ ) based on the fitted SPF on similar sites is plotted against the observed crash frequency ( $N_{observed}$ ) on treatment sites. Depending on a weighting factor  $w$ , the expected crash frequency,  $N_{expected}$ , can take on a number between  $N_{predicted}$  and  $N_{observed}$ . Note that  $w$  can range between 0 and 1. The higher the  $w$  value, the closer the value of  $N_{expected}$  to  $N_{predicted}$ , and vice versa. PSI is thus the difference between  $N_{expected}$  and  $N_{predicted}$ .



**Figure 3-3 Illustration of the EB Method and PSI**

### 3.6. Summary

This chapter proposed a complete methodology used in this dissertation research. The methodology is composed of four parts: segmentation, model development, model tests, and identification of high crash locations.

Segmentation is an essential step in data preparation for developing the SPFs. Three segmentation methods were described in this chapter: fixed length, variable length, and Fisher's clustering algorithm. According to *SafetyAnalyst*, SPFs are developed for homogenous segments of variable length. Thus, variable length was used as the segmentation method for developing Florida-specific SPFs in this research. In addition, an alternative segmentation method based on Fisher's clustering algorithm is examined in this step to improve the development of SPFs for the purpose of identifying high crash locations.

Florida-specific SPFs are developed using NB models. Several goodness-of-fit tests were then used to assess the performance of the SPFs, including the overdispersion parameter, CURE plot, MSE, MAD, MSPE, and  $R^2_{FT}$ .

The *SafetyAnalyst* default SPFs are calibrated to Florida data using calibration factors. The performance of the two models (i.e., Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida) is then compared by using several goodness-of-fit tests, such as MAD, MSPE, and  $R^2_{FT}$ .

The EB method, which requires the use of SPFs, is employed in identifying high crash locations, which is a main application of the development of SPFs.

## CHAPTER 4

### DATA PREPARATION

#### 4.1. Introduction

This chapter describes the data preparation of different roadway types for developing SPFs. Section 4.2 presents the data acquisition and the requirement of sample size for the calibration process. Sections 4.3, 4.4, and 4.5 present the data processing steps and provide the summary statistics for roadway segments, intersections, and ramps, respectively. Section 4.6 provides an alternative method that uses the clustering algorithm in segmentation for identifying high crash locations.

#### 4.2. Data Acquisition

Four years of crash data, from 2007 to 2010, were extracted from FDOT's CAR system. Other geometric, roadway, and traffic data were obtained from FDOT's RCI database.

Several requirements must be met during the calibration of SPFs. As recommended in *SafetyAnalyst*, crash data should be provided for a minimum of three years. For this research, although crash data from years 2000 through 2010 are available, only four of these years were considered for the purpose of minimizing the effects related to changes in external factors. The CAR database records the crash information about roadway ID, crash location, date, time, crash type, number of fatalities and/or injuries, number of vehicles involved, roadway surface conditions, weather conditions, and lighting conditions. Crash severity (fatal, injury, and property damage only (PDO)) can be derived from the number of fatalities or injuries. Since crash

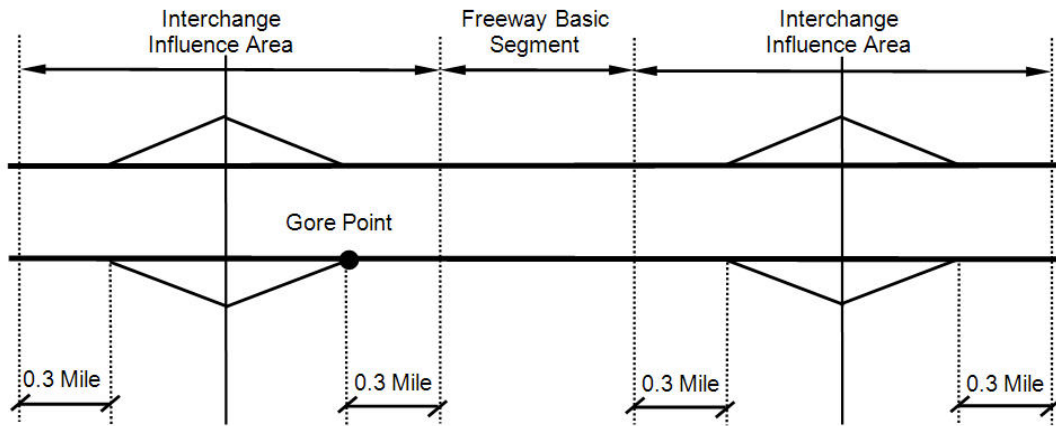
records of fatal and injury are known to be more accurate than that of PDO, the Florida-specific SPFs were developed for both total crashes and the sum of fatal and injury crashes.

As recommended by HSM (2010), for each facility type, the desirable minimum sample size for the calibration data set is 30 to 50 sites, and the entire group of calibration sites should represent a total of at least 100 crashes per year. For those categories in *SafetyAnalyst*, of which the sample sizes were insufficient in Florida to accurately estimate regression coefficients, SPFs may not be specifically developed due to lack of data. The processed data for each subtype were divided into two datasets of 70% and 30% for calibration and validation to be described in Chapter 5, respectively.

### **4.3. Roadway Segments**

#### ***4.3.1. Interchange Influence Areas***

According to the *SafetyAnalyst* user's manual (2010), interchange influence area is defined as shown in the schematic sketch of Figure 4-1. The interchange influence area of a particular interchange mainly covers the length of the freeway section extending approximately 0.3 miles upstream of the gore point of the first exit/entrance ramp, to approximately 0.3 miles downstream of the gore point of the last entrance/exit ramp on the same interchange. The area between two successive interchange influence areas is defined as basic freeway segment.



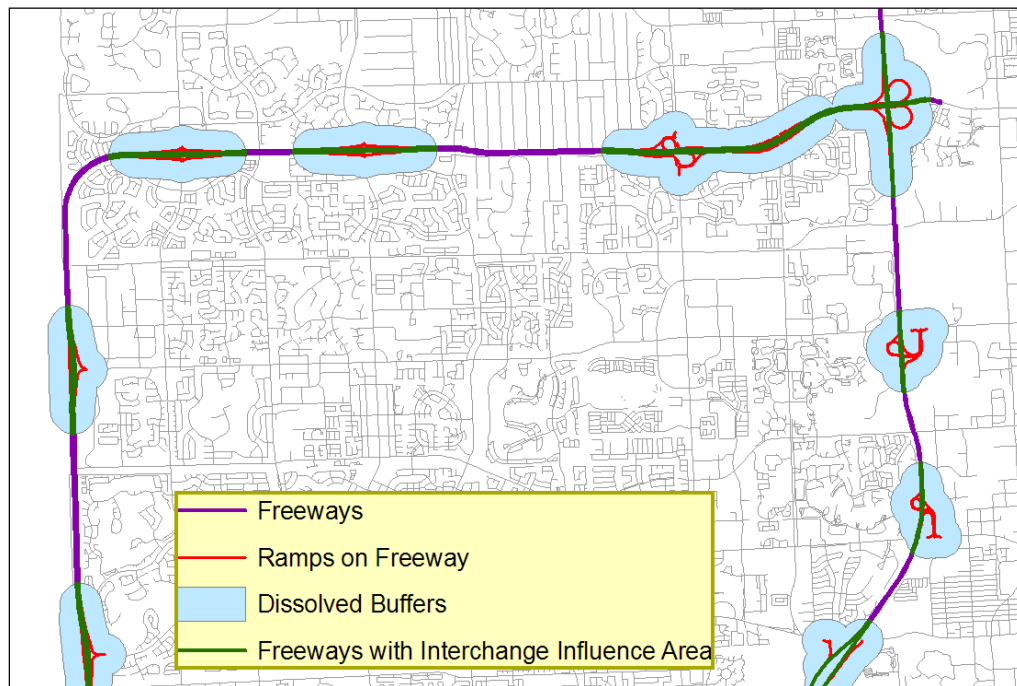
**Figure 4-1 Interchange Influence Area**

*SafetyAnalyst* categorizes freeways based on area type, number of lanes, and presence of interchange influence area (IIA). Therefore, IIA is a required data variable. However, interchange influence areas were not explicitly identified within the roadway inventory database. Therefore, the separation of interchange influence areas from the freeways segments was performed first. Identification and segregation of interchange influence areas was performed in GIS using the following steps:

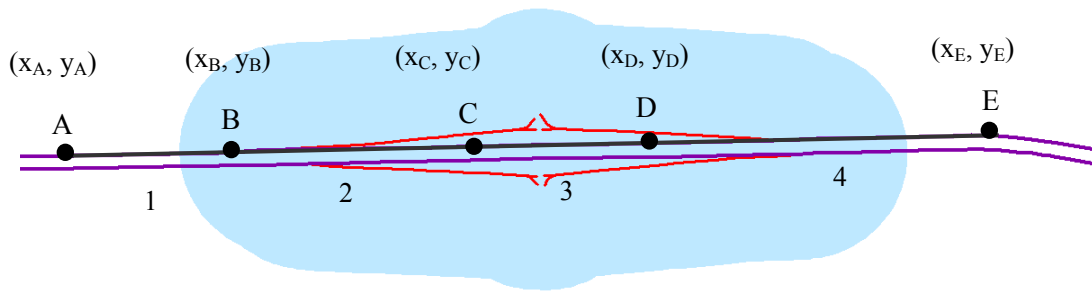
1. *Geometric Data Acquisition:* Geometric data for freeways and all access ramps were extracted from roadway shapefiles maintained by FDOT.
2. *Interchange Influence Area Identification:* A 0.3-mile buffer for each ramp of the interchange was created, and the overlapped buffers were dissolved. The dissolved buffer areas were considered interchange influence areas.
3. *Milepost Estimation:* The milepost of each crossing point between the dissolved buffer and the connected freeway basic segment was estimated. In this step, the dissolved buffer layer and the freeway layer were intersected (see Figure 4-2). The coordinates of all points within each intersected line, along with the two

newly identified endpoints of each intersected line, were recalculated. Figures 4-3(a) and (b) show original and newly identified segments, respectively.

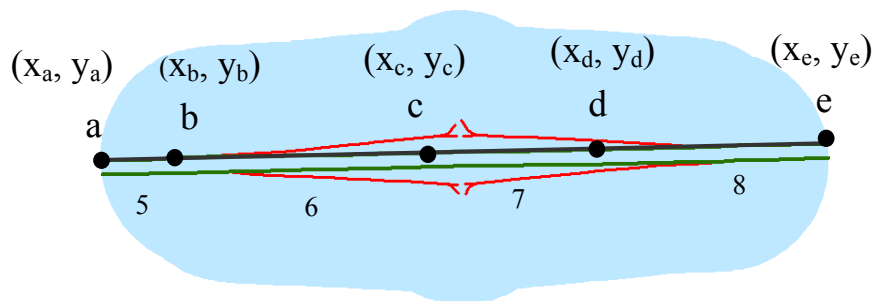
4. *Basic Freeway Segments and Freeways within Interchange Influence Area Identification:* Interchange influence areas were identified by spatially comparing the coordinates of original freeway segments and coordinates of the previously identified interchanges. Roadway characteristics databases, constituting of roadway ID and the beginning and ending mileposts of segments, were created for freeway basic segments and freeway segments within IIA.



**Figure 4-2 Freeways with Interchange Influence Area**



(a) Original Segment



(b) Intersect Area

Figure 4-3 Estimation of Milepost of the Intersecting Point

#### 4.3.2. Roadway Segmentation

The Dynamic Segmentation (DySeg) program was used to divide the road network into segments. DySeg can dynamically divide the roadway sections based on several categories, including equal length segments, uniform segments with unique attributes, or according to the specified range of segment lengths, desired roadway features and crash types, and can also compute the crash number associated with each roadway segment. Figure 4-4 shows the screen capture of the RCI user interface for specifying the input for geometric variables. Figure 4-5 shows the general user interface of DySeg for specifying the crash years, location, and severity.

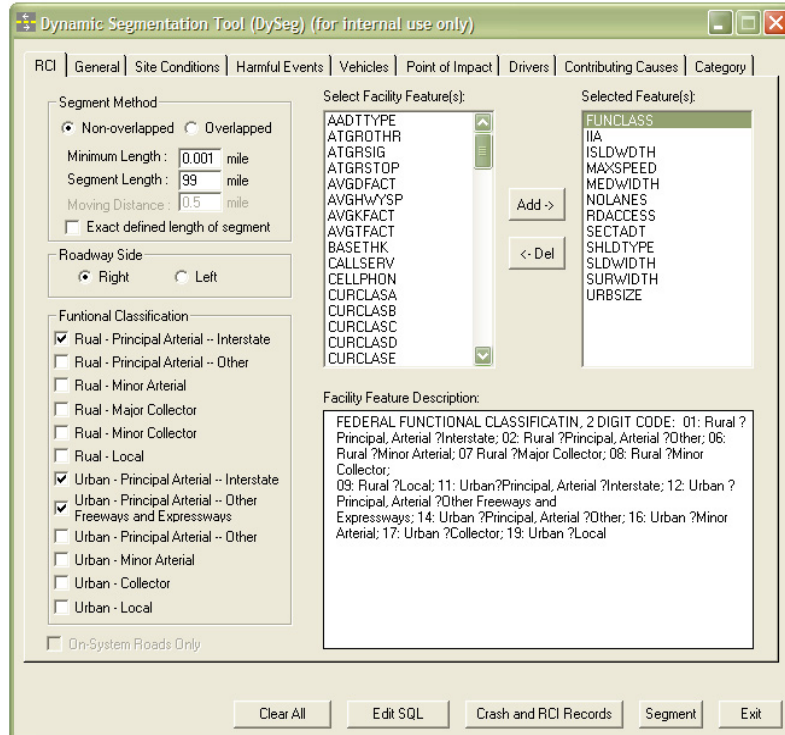


Figure 4-4 DySeg RCI User Interface

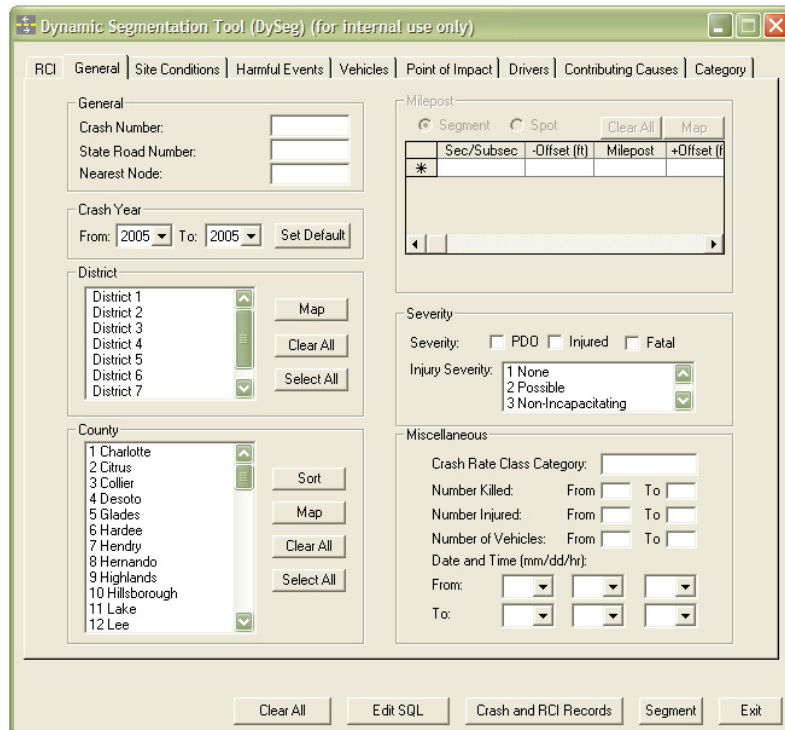




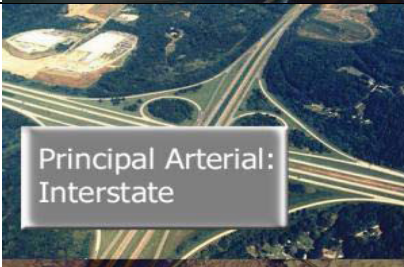


Figure 4-5 DySeg General User Interface




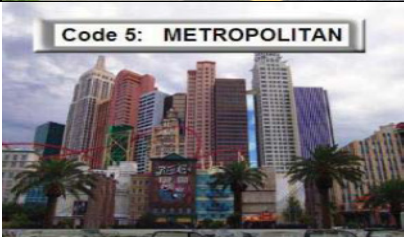



The RCI database is used to store, process, and report physical and administrative data on the Florida state highway system. Physical features, surface data features, and traffic monitoring site features are included in this database. In this research, the attributes for segmentation were selected based on their significance on crash occurrence.

Table 4-1 shows the attributes used for segmentation.

**Table 4-1 Attributes in RCI for Segmentation**

	<b>Attribute</b>	<b>Description</b>	<b>Unit</b>	<b>Illustrative Snapshot</b>
1	NOLANES	Number of lanes		
2	SURWIDTH	Surface width	ft	
3	FUNCLASS	Functional classification		
4	MAXSPEED	Maximum speed limit	mph	
5	MEDWIDTH	Median width	ft	

6	SHLDTYPE	Shoulder type		
7	SLDWIDTH	Shoulder width	ft	
8	ISLDWDTH	Inside shoulder width	ft	
9	URBSIZE	Urban size		
10	RDACCESS	Access control type		
11	SECTADT	Section ADT		

Note: For freeways and undivided arterial streets, the attribute “road side (ROADSIDE)” has to be considered because the geometric and operational features might vary on each side of the roadway.

Table 4-1 displays four categorical variables: functional classification, shoulder type, urban size, and roadway access. The associated levels for each variable are briefly discussed herein. There are 12 levels for the functional classification (FUNCLASS) variable, as documented in the 2009 RCI Field Handbook (the numbers are RCI codes):

- 01 – Rural principal arterial - interstate

- 02 – Rural principal arterial - other
- 06 – Rural minor arterial
- 07 – Rural major collector
- 08 – Rural minor collector
- 09 – Rural local
- 11 – Urban principal arterial - interstate
- 12 – Urban principal arterial - other freeways and expressways
- 14 – Urban principal arterial - other
- 16 – Urban minor arterial
- 17 – Urban collector
- 19 – Urban local

For the purpose of calibrating the freeway SPFs, the codes “01” and “11” will be used for data screening. For calibrating the expressway SPFs, the codes “02,” “12,” or “14” can be used with some caution, since these codes might also include arterials. This will be taken into consideration in the next data preparation procedure.

The shoulder type attribute (SHLDTYPE) has nine levels, as follows (the numbers are RCI codes):

- 0 – Raised curb – no shoulder exists
- 1 – Paved with or without striping – including bike slots
- 2 – Paved with warning device – raised or indented strips
- 3 – Lawn
- 4 – Gravel/Marl

- 5 – Valley gutter (not a barrier)
- 6 – Curb and gutter
- 7 – Other
- 8 – Curb with resurfaced gutter

The urban size (URBSIZE) attribute has five levels, as follows (the numbers are RCI codes):

- 1 – Rural
- 2 – Small urban – 5,000:49,999 population
- 3 – Small urbanized – 50,000:199,999 population
- 4 – Large urbanized – 200,000:499,999 population
- 5 – Metropolitan – 500,000 or more population

To differentiate between rural and urban facilities, the rural facilities will only include the code “1,” while the urban facilities will include “2,” “3,” “4,” or “5.”

The roadway access (RDACCESS) attribute has three levels, as follows (the numbers are RCI codes):

- 1 – Full control – interstate, turnpike or similar road with all access via grade-separated interchanges
- 2 – Partial control – some grade-separated interchanges and some direct access roads or driveways
- 3 – No access control – non-grade-separated interchanges

For calibrating the freeway and expressway SPFs, the roadway access should be only full control or partial control (mostly full control). For calibrating the arterials, the roadway access should be mostly non-access controlled (code “3”).

### 4.3.3 Data used for Roadway Segments

Crash and traffic data from the years 2007 to 2010 and roadway inventory data from 2008 was used to develop Florida-specific SPFs. Table 4-2 provides the summary statistics of segments used to generate SPFs based on segment subtypes identified in *SafetyAnalyst*. As mentioned previously, the data for each subtype were divided into two datasets of 70% and 30% for calibration and validation in Chapter 5, respectively.

**Table 4-2 Summary Statistics of Roadway Segments**

Category	Total Length of Segments (miles)	# of Sites	Crash Data (2007-2010)		
			Total Crashes	Fatal and Injury Crashes	
<b>Rural</b>					
2-lane Roads	3257.44	2408	15703	9610	
Multilane Undivided Roads*	9.80	37	118	63	
Multilane Divided Roads	1189.03	1048	11432	6630	
Basic Freeway Segment	4 Lanes	442.86	264	6222	3587
	6+ Lanes	181.73	101	2400	994
Segments within Interchange Influence Area	4 Lanes	49.70	156	717	423
	6 Lanes	38.38	69	502	228
<b>Urban</b>					
2-lane Arterial Streets	802.71	2038	17643	9695	
Multilane Undivided Arterial Streets	63.23	245	3562	1801	
Multilane Divided Arterial Streets	2473.98	6923	124154	63563	
One-way Arterial Streets	126.72	433	5319	1989	
Basic Freeway Segment	4 Lanes	319.26	375	8592	4223
	6 Lanes	198.118	272	10317	4694
	8+ Lanes	42.28	75	2229	1010
Segments within Interchange Influence Area	4 Lanes	280.58	620	11210	5404
	6 Lanes	263.71	558	27115	11851
	8+ Lanes	125.46	330	25748	12311

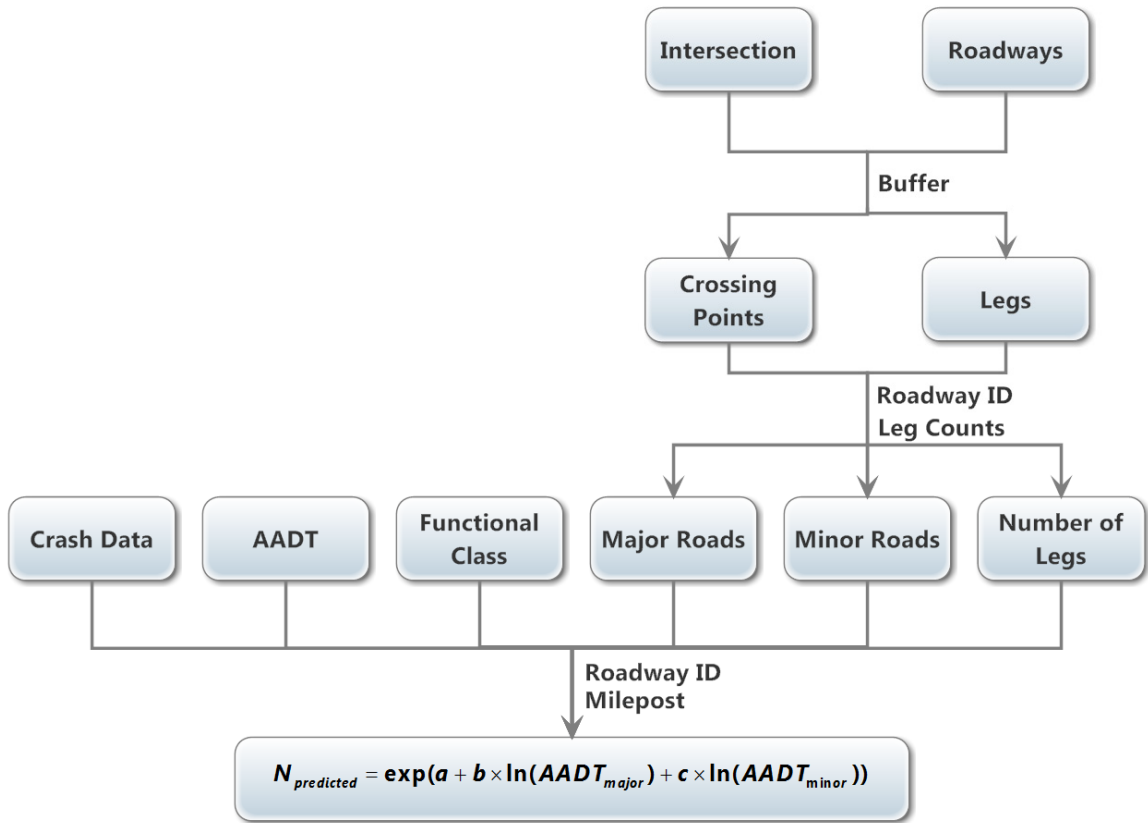
\* sample size was insufficient to accurately estimate regression coefficients

#### 4.4. Intersections

Compared to segments, the data requirements to generate intersection SPFs are more rigorous. One of the required variables that allow *SafetyAnalyst* to divide intersections into subtypes is traffic control type; however, it is not available in the details required in the RCI database. Due to data constraints, SPFs were developed for only four types of signalized intersections (rural and urban, with three-leg and four-leg each). At this point, analysis of unsignalized intersections is not possible due to the lack of detailed data on traffic control type in the RCI. The 2008 RCI data were used to identify signalized intersections. In addition to intersection-related crashes, those that occurred either at an intersection or within 250 feet from the center of an intersection were included in the analysis, as illustrated in *SafetyAnalyst*.

Figure 4-6 illustrates the steps followed to process the data. The first step involves identification of locations of signalized intersections. The location data was extracted from FDOT's Geographic Information System - Traffic Signal Location database. In this database, a signalized intersection is represented by a set of roadway IDs associated with corresponding mileposts of one of the roadways that cross the intersection. Generally, the set of roadway ID and milepost used to represent an intersection belongs to the major road, compared to minor roads. A GIS-aided process was then used to merge the location information with geographical coordinates. By using geographical coordinates, information of minor roads was linked to the corresponding intersection. During this process, all roadways crossing the intersection were summarized. This data set was then joined with AADT, leg-count, functional class, and crash data. The output data was used to generate Florida-specific SPFs for

signalized intersections.



**Figure 4-6 Data Processing Framework**

Unsignalized intersections could not be analyzed due to the following reasons:

- data limitations in the RCI database,
- fewer number of all-way stop-controlled intersections, and
- unavailability of AADT data for minor approaches, since most of these approaches are located on local roads.

Florida-specific SPFs were developed for only four subtypes: rural, urban, and three-leg and four-leg intersections. Table 4-3 shows the summary statistics of signalized intersections used to generate SPFs.

**Table 4-3 Summary Statistics of Signalized Intersections**

Area Type	Category	# of Sites	Crash Data (2007-2010)	
			Total Crashes	Fatal and Injury Crashes
Rural	Three-Leg	88	1,781	866
	Four-Leg	111	3,877	1,736
Urban	Three-Leg	314	11,471	4,783
	Four-Leg	641	39,517	16,400

#### 4.5. Ramps

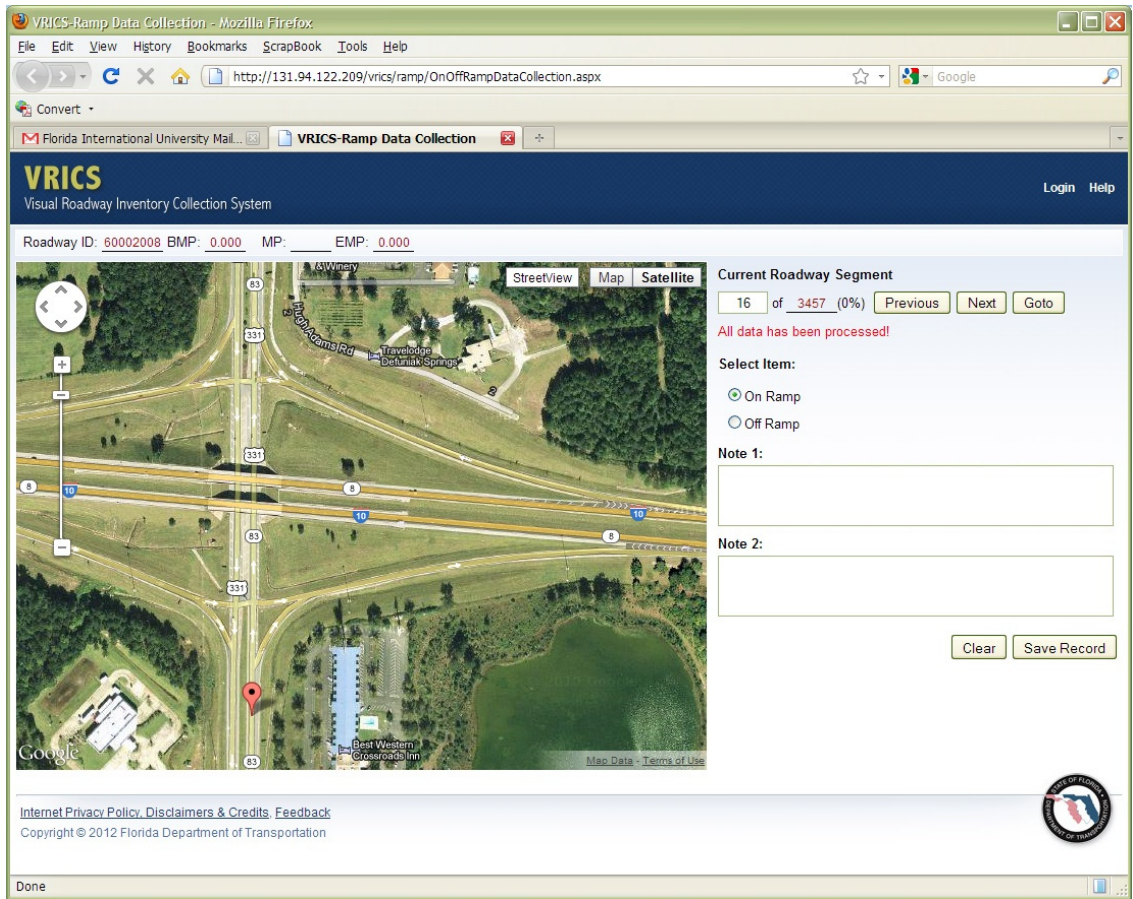
SPFs were developed for different types of ramps. *SafetyAnalyst* categorizes ramps into 16 subtypes, based on ramp type and configuration. However, it was observed that Florida’s classification of ramps is different from the *SafetyAnalyst*’s classification. Table 4-4 lists the ramp configurations used in *SafetyAnalyst*, and in Florida. To analyze the maximum number of ramps, SPFs for ramps were generated based on the classification used in Florida.

**Table 4-4 Ramp Subtypes in *SafetyAnalyst* vs. in Florida’s RCI Database**

Ramp Configuration in <i>SafetyAnalyst</i>	INTERCHG (Type of Interchange) Variable in Florida RCI Database
1 – Diamond	01 – Diamond
2 – Parclo loop	02 – Partial Diamond
3 – Free-flow loop	03 – Trumpet
4 – Free-flow outer connection	04 – Y Intersection
5 – Direct or semi-direct connection	05 – 2 Quadrant Cloverleaf or Partial Cloverleaf
6 – C-D road or other connector	06 – 4 Quadrant Cloverleaf with Collector Road
0 – Other	07 – 4 Quadrant Cloverleaf
99 – Unknown	08 – Direct Connection Design
	09 – Other

Off-ramp and on-ramp information is incomplete in the RCI database. Therefore, the first step in data processing dealt with this classification. A Visual Roadway Inventory Collection System (VRICS) program developed by the LCTR at FIU was used to collect on-ramp and off-ramp information. Figure 4-7 shows the screen capture of the VRICS user interface.





**Figure 4-7 VRICS User Interface**

A GIS-aided process was used to merge functional classification with ramp type information. The data set was then joined with AADT, crash data, and the off-ramp or on-ramp information extracted from VRICS. The output data was used to generate Florida-specific SPFs for ramps.

Crash and traffic data for the years 2008 to 2010, as well as processed ramp data for 2008, were used to develop Florida-specific SPFs for ramps. Traffic data is complete only for the year 2010. Therefore, AADT is assumed to be the same for the earlier years (i.e., 2008 and 2009). Table 4-5 provides the summary statistics of ramps used to generate SPFs.

**Table 4-5 Summary Statistics of Ramps**

Category		Total Length of Segments (miles)	# of Ramps	Crash Data (2008-2010)	
				Total Crashes	Fatal and Injury Crashes
<b>Rural</b>					
Diamond	off-ramp	27.99	74	262	113
	on-ramp	27.68	73	134	71
Partial Cloverleaf (Parclo Loop) *	off-ramp	7.03	20	101	41
	on-ramp	7.23	20	33	12
<b>Urban</b>					
Diamond	off-ramp	167.41	389	5143	2372
	on-ramp	151.35	382	2117	978
Partial Diamond	off-ramp	59.87	148	1152	516
	on-ramp	54.47	134	1122	474
Trumpet*	off-ramp	15.54	30	190	93
	on-ramp	12.28	26	108	46
Partial Cloverleaf (Parclo Loop)	off-ramp	91.35	200	1887	805
	on-ramp	92.35	204	1304	634
Direct Connection*		21.92	33	285	111

\* sample size was insufficient to accurately estimate regression coefficients

## 4.6. Advanced Segmentation for Identifying HCLs

### 4.6.1. Introduction

This section introduces the application of a clustering method, developed by Fisher (1958), to roadway segmentation, in place of the traditional fixed-length and variable-length segmentation methods, so as to improve the calibration of SPFs for identifying high crash locations. The clustering approach helps reduce crash heterogeneity for within-group elements by grouping roadway segments with similar crash distributions into homogeneous groups.

In this research, crash frequencies distributed in each unit of neighboring roadway sections (0.01 mile) are seen as ordered samples, and are expected to be grouped into  $k$  segments. The neighboring section units with similar total crash frequency are grouped

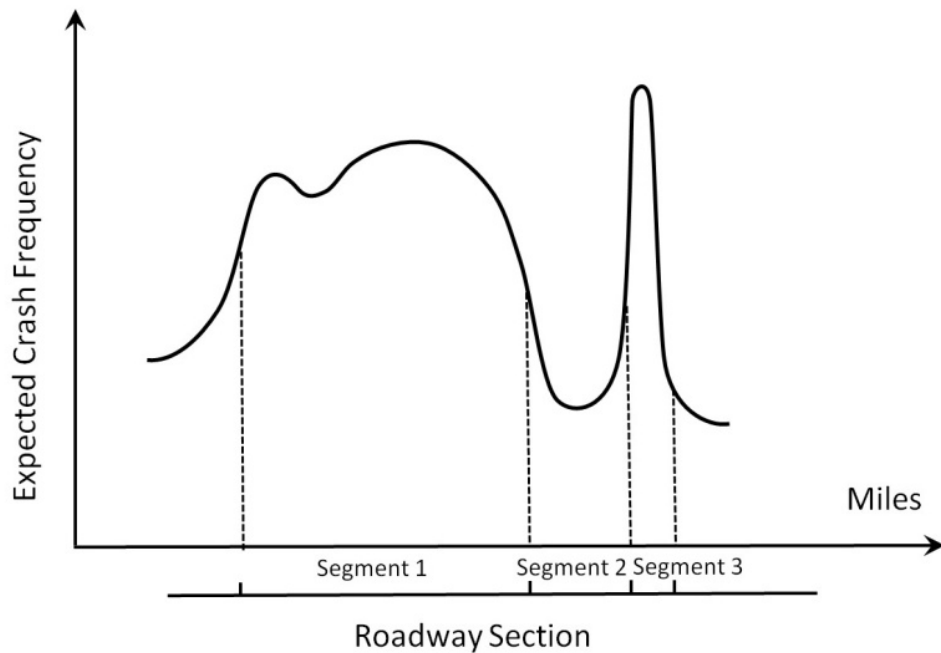
into the same segment, while those with high crash frequency differences are assigned to separate segments. As aforementioned, Fisher's clustering method avoids the shortcomings of existing techniques used for segmentation by taking into account the crash characteristics. It expands on studies using Fisher's clustering and applying this method to roadway segmentation to calibrate SPFs, which are then used together with the EB method to identify high crash locations.

Figure 4-8 shows the advantage of this approach in the screening process of identifying high crash locations relative to the commonly used sliding window method. As shown in Figure 4-8 (a), the EB network screening method is applied to each segment divided by Fisher's clustering approach, and the segments with higher EB-adjusted crash rates are classified as high crash locations.

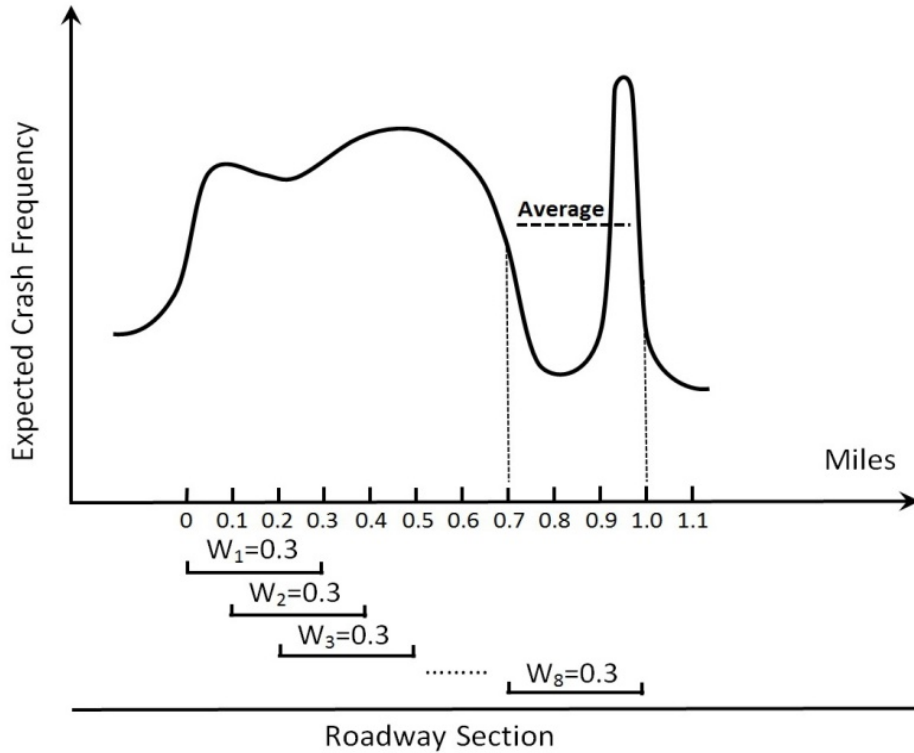
Figure 4-8 (b) shows the sliding window method. A window of a specified length moves along the road section, from beginning to end, in increments of a specified size. The selected screening technique (EB method in this research) is applied for each window, with segments ranked according to the most critical window (HSM, 2010). The sliding window is the most widely used screening method for segments, yet its weakness lies in the fact that the window size is constant; thus, high crash locations and their rankings will differ when different window sizes are selected. For example, for the eighth sliding window "W<sub>8</sub>," the use of a window size with a smaller range (e.g., 0.1-mile) can identify this site as hazardous, but a 0.3-mile window size would conceal this fact because the average crash frequency is considered in the window. Also, the sites (W<sub>1</sub>, W<sub>2</sub>, and W<sub>3</sub>) within the same roadway facility exhibit similar crash frequency. However, because of the small window size, they are assigned to different sections and ranked on

different levels. In fact, the first 0.7-mile segment could be roughly considered to be one entity if the window size is set at 0.7 miles or larger.

As aforementioned, Figure 4-8 (a) attempts to screen the roadway sections based on the segment divided by Fisher's clustering, which overcome the shortcomings of the sliding window method. In short, segmentation using Fisher's clustering method could potentially yield, not only more accurate SPFs, but also a more precise screening process for identifying high crash locations.



**(a) Site Screening by Fisher's Clustering Method**



**(b) Site Screening by Sliding Window Method**  
**Figure 4-8 Comparison of Two Screening Methods**

**4.6.2. Data and Model**

An 88.6-mile stretch of the Interstate-95 (I-95) freeway in the South Florida region was selected as the study area. This region traverses three counties: Miami-Dade, Broward, and Palm Beach, in order from south to north.

The freeway stretch was divided into segments using three methods. For the variable length method, the freeway was divided into homogeneous segments of variable lengths, with the segment length defined by any change in the geometric or traffic characteristics. For example, if there is a change in the shoulder width, median width, number of lanes, AADT, or other related characteristics, a new segment is defined. A fixed 0.5-mile segment was used for the fixed-length method. For Fisher’s clustering

method, the segmentation was evaluated based on the crash frequency distribution along the freeway, where the optimum segment length and number of segments were determined by aggregating sections with similar distributions within the same segment.

The data of the 71.3-mile section in Broward and Palm Beach counties were used for SPF calibration, while the data for the remaining 17.3-mile section in Miami-Dade were used for validation and prediction. As a test case, rear-end crashes were used in the analysis because this crash type occurred with the greatest frequency. Summary statistics for the variable-length, fixed-length, and Fisher’s clustering methods used in the datasets for both calibration (Broward and Palm Beach counties) and prediction (Miami-Dade County) are shown in Table 4-6.

**Table 4-6 Summary Statistics for Three Segmentation Methods**

<b>Calibration Data in Broward and Palm Beach (71.3 miles)</b>				
<b>Segmentation Method</b>	<b>Number of Segments</b>	<b>Total Segment Length (mile)</b>	<b>Min. Segment Length (mile)</b>	<b>Max. Segment Length (mile)</b>
Variable length	471	71.3	0.002	1.024
Fixed length	142		0.500	0.500
Fisher’s clustering	209		0.050	1.950
<b>Prediction Data in Miami-Dade (17.3 miles)</b>				
<b>Segmentation Method</b>	<b>Number of Segments</b>	<b>Total Segment Length (mile)</b>	<b>Min. Segment Length (mile)</b>	<b>Max. Segment Length (mile)</b>
Variable length	121	17.3	0.002	2.000
Fixed length	34		0.500	0.500
Fisher’s clustering	54		0.050	1.350

After using the three segmentation techniques to divide the 71.3-mile freeway stretch along Broward and Palm Beach counties, each method was examined to determine whether it could outperform the others, so as to select a better method that

would yield a better SPF. This was accomplished by fitting an NB model to rear-end crashes.

There were seven variables measured in the modeling procedure: the natural logarithm of AADT, natural logarithm of the segment length, right shoulder width, inside (left) shoulder width, median width, percentage of trucks, and speed limit. These variables are believed to have an impact on rear-end crashes, with those employed resembling other studies that analyzed rear-end crash counts (e.g., Haleem et al. 2010). There was almost no significant variation in the shoulder width variable throughout the stretch, nor was there significant variation in the speed limit (mostly 105 km/h or 65 mph). Other variables, such as lane width, could not be used since the average lane width was set at approximately 312 feet throughout the stretch. The three NB models are shown in Table 4-7.

**Table 4-7 Rear-end NB Crash Frequency Models for Segmentation Methods**

Variable Description	Variable Length		Fixed Length		Fisher's Clustering	
	Estimate (Standard Error)	P-Value	Estimate (Standard Error)	P-Value	Estimate (Standard Error)	P-Value
Intercept	-7.7310 (2.1441)	0.0003	-4.6615 (4.0681)	0.2519	-16.9474 (4.5857)	0.0002
Natural logarithm of AADT	0.4705 (0.1686)	0.0052	0.6496 (0.3302)	0.0491	1.3767 (0.3647)	0.0002
Natural logarithm of segment length	0.8737 (0.0455)	<0.0001	N/S *		0.4703 (0.0647)	<0.0001
Inside shoulder width	-0.0432 (0.0104)	<0.0001	0.0789 (0.0204)	0.0001	-0.0494 (0.0198)	0.0126
Median width	-0.0113 (0.0036)	0.0019	N/S		N/S	
Truck percentage	N/S		N/S		0.0780 (0.0329)	0.0178
Dispersion ( <i>d</i> )	0.8700 (0.0869)		0.5721 (0.0682)		1.2106 (0.1124)	
Pseudo R-square (likelihood ratio index)	0.084		0.011		0.035	

\* N/S = Not significant

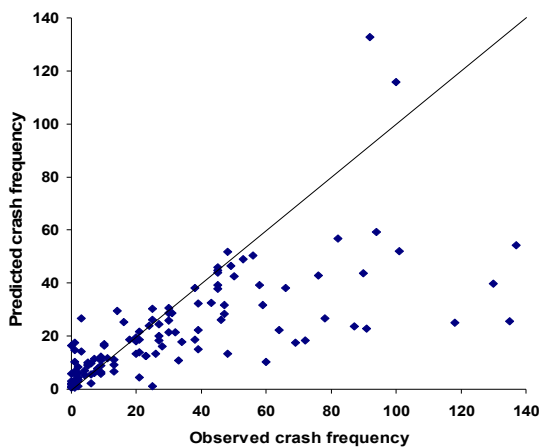
In Table 4-7, the signs of the coefficients are almost identical for the three methods (except for the inside shoulder width variable). There is a significant increase in rear-end crashes by increasing AADT, as rear-end crashes always occur at high traffic volumes, or congested situations. This is consistent with other studies (Wang and Abdel-Aty 2006; Haleem et al. 2010). There is also a significant increase in rear-end crashes with the increase of segment length. This could be due to the probability of a greater volume of vehicles being included for relatively large segments and, thereby, resulting in a higher rear-end crash risk, as is consistent with findings by Abdel-Aty et al. (2009).

An intuitive finding from the NB models for the variable length and Fisher's clustering segmentation is that rear-end crashes may be reduced by increasing the left shoulder width near the median. The relatively large shoulder width could act as a rear-end crash-prevention method for any vehicle about to hit the leading vehicle in the lane beside the median. In other words, the following driver could easily avoid rear-ending the lead vehicle by turning into wider inside-shoulder areas. An interesting finding is that an increase in the median width is associated with a reduction in rear-end crashes. This is more common with vehicles traveling in the leftmost lanes, as these vehicles could avoid rear-end crashes by diverting to the wider medians.

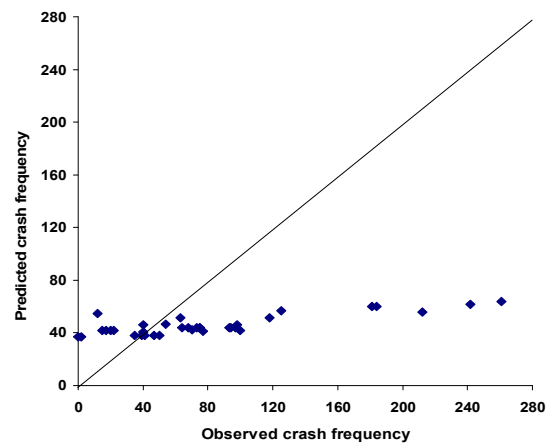
An expected finding is that rear-end crashes increase by increasing truck percentage in the fleet. The NB model shows that increasing the truck percentage by 1% increases rear-end crashes by  $e^{0.0780}$  (1.08) times. This increase could be due to the vertical and horizontal obstruction caused by trucks in the vicinity of vehicles.



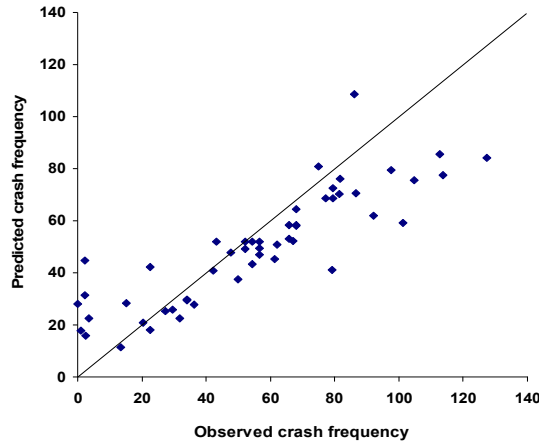
As previously discussed, the performance assessment of the fitted NB models was performed on the 17.3-mile stretch in Miami-Dade after segmenting it based on the same division methods. The plot of the observed versus predicted crashes for each of the three models is shown in Figure 4-9. From the three plots in the figure, it is obvious that the NB model with Fisher's clustering, in which the observed and predicted rear-end crashes are well-matched, performs much better than the other two models. Furthermore, to assess the prediction performance of the three models, the MAD and MSPE values for the three models are shown in Table 4-8. The table shows significantly lower MAD and MSPE values for the clustering-based model, resulting in the potential to improve the SPF calibration, compared to the variable- and fixed-length model, which came in second and third, respectively.



**(a) Variable Length Segmentation**



**(b) Fixed Length Segmentation**



**(c) Fisher’s Clustering Segmentation**

**Figure 4-9 Observed vs. Predicted Crash Frequency from Three NB Models**

**Table 4-8 Prediction Performance of Three NB Models for I-95**

<b>NB Models</b>	<b>MAD</b>	<b>MSPE</b>
Variable-length segmentation	14.17	633.27
Fixed-length segmentation	50.66	6076.23
Fisher’s clustering segmentation	6.73	87.42

For further validation, the same clustering-based technique was applied to two other freeway facilities in Florida, including an 80.5-mile stretch of I-75 in the Tampa area, and a 66.5-mile stretch of I-4 in the Orlando area in Florida. Tables 4-9 and 4-10 compare the MAD and MSPE results from the three methods for the two facilities, respectively. The results show that the clustering method produced significantly better-fitted SPFs than those produced by either the variable or the fixed-length method.

**Table 4-9 Prediction Performance of Three NB Models for I-75**

<b>NB Model</b>	<b>MAD</b>	<b>MSPE</b>
Variable-length segmentation	5.02	33.92
Fixed-length segmentation	5.93	44.22
Fisher’s clustering segmentation	2.66	10.37

**Table 4-10 Prediction Performance of Three NB Models for I-4**

<b>NB Model</b>	<b>MAD</b>	<b>MSPE</b>
Variable-length segmentation	9.06	177.13
Fixed-length segmentation	13.30	268.34
Fisher’s clustering segmentation	8.92	149.41

### ***4.6.3. High Crash Locations***

In this section, the superior clustering-based SPF is applied to identify high rear-end crash locations. The study area was a 17.3-mile stretch of I-95 in Miami-Dade County. The screening sites were the same non-overlapping segments that were divided by Fisher's clustering method.

Using the EB method, the expected rear-end crash frequency for each of the 54 screening sites was estimated by accounting for the corresponding observed and predicted frequencies. The weighting factor was estimated from the over-dispersion parameter, which was 1.2106 in this research. Table 4-11 shows the ranking and specific areas of the top ten high rear-end crash locations along the 17.3-mile stretch. The zero-mile post is the start point of I-95, and the mile post increases from south to north until the end of the 17.3 miles of Miami-Dade County. The beginning and ending mile posts shown are relative to the zero-mile post (start point of the study area). Furthermore, Figure 4-10 shows the locations with high rear-end crashes. These locations correspond to the known congested sections with stop-and-go traffic, a condition susceptible to rear-end crashes.

**Table 4-11 Top Ten HCLs on the 17.3-mile Stretch of I-95**

<b>Rank</b>	<b>Beginning Mile Post</b>	<b>Ending Mile Post</b>	<b>Segment Length (mile)</b>	<b>Expected Crash Count</b>	<b>Expected Crash Count per Mile</b>	<b>Expected Crash Count per Kilometer</b>
1	4.35	4.45	0.10	29.9	299.1	185.7
2	11.35	11.55	0.20	27.0	135.0	83.9
3	10.00	10.40	0.40	53.8	134.5	83.5
4	3.95	4.30	0.35	38.7	110.6	68.7
5	16.10	16.3	0.20	18.1	90.5	56.2
6	0.80	0.95	0.15	13.5	90.0	55.9
7	5.75	5.95	0.20	17.7	88.5	55.0
8	3.40	3.95	0.55	48.0	87.3	54.2
9	8.50	9.35	0.85	65.6	77.2	47.9
10	5.40	5.70	0.30	22.3	74.3	46.2

**4.7. Summary**

This chapter described the data preparation of all roadway types for developing SPFs. Florida-specific SPFs were developed using the 2008 RCI data and four years of crash data (from 2007 to 2010), for both total and FI crashes.



Figure 4-10 HCLs on the 17.3-mile Stretch of I-95 in Miami-Dade

As per the predefined subtypes used in *SafetyAnalyst*, roadway segments were divided into 17 site subtypes based on area type, functional classification, and number of lanes. Summary statistics were provided to develop Florida-specific SPFs.

The data requirements used to generate intersection SPFs are rigorous, compared to roadway segments. The traffic control type, an important required variable in *SafetyAnalyst* for dividing intersections into subtypes, is not available in the RCI database. Thus, SPFs for unsignalized intersections cannot be developed due to the lack of detailed data on traffic control type. Therefore, SPFs were developed for only four types of signalized intersections (rural and urban, with three-leg and four-leg each).

*SafetyAnalyst* classifies ramps into 16 subtypes. However, Florida has completely different ramp classifications from *SafetyAnalyst*. Therefore, SPFs for ramps were generated using Florida-specific subtypes.

The last section explored Fisher's clustering method as an improved method for segmentation for the calibration SPF, for use in high crash location identification. This method avoids the shortcomings of traditional division methods (fixed and variable length) by dividing segments into homogeneous clusters while accounting for specific crash characteristics. Furthermore, to expound upon the methodological research and increase the benefits derived from its application, the same method was used in the site screening process while applying the EB approach, which succeeded in identifying sites with promise for improvement. This method cannot be used for developing SPFs for the purpose of predicting crash frequency because the crash data are required for segmentation during the data processing.

## CHAPTER 5

### MODEL DEVELOPMENT

#### 5.1. Introduction

As mentioned in Chapter 1, the main objective of this research is to develop Florida-specific SPFs to be applied with *SafetyAnalyst*. In Chapter 3, the methodology of the proposed models for developing SPFs was described. This chapter describes the development of Florida-specific SPFs for every proposed roadway subtype described in Chapter 4. The Statistical Analysis Software (SAS) was used to calculate the coefficients and corresponding overdispersion parameters for each subtype, using the calibration data. The p-value, overdispersion, and CURE plot were used to test the significance, reliability, and goodness-of-fit of the models. The statistical goodness-of-fit tests provided in Chapter 3 were then performed on both the calibration and validation to assess the performance of the SPFs developed.

#### 5.2. Roadway Segments

Using calibration data, Florida-specific SPFs were developed for each of the 17 categories of segments, for both total crashes and FI crashes. Table 5-1 shows the coefficients and corresponding overdispersion parameters for each of the SPFs, specifically for freeways. In Table 5-1, parameters for all models were statistically significant at the 0.01% level, except for the intercept term ( $\alpha$ ) for the urban (8+)-lane basic freeway segment subtype, which had a p-value of about 0.2%. The overdispersion parameter was used to account for dispersion in the data. The closer the overdispersion parameter is to zero, the more statistically reliable the SPF. As shown in Table 5-1, the

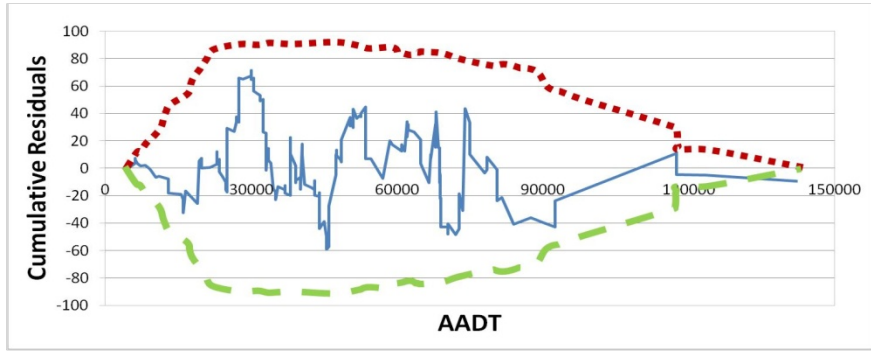
overdispersion values for all categories were less than 1, and most of the values were lower than 0.5, which means the overdispersion parameters were close to zero, and the corresponding Florida-specific SPFs were reliable. Furthermore, in order to determine which model is more reliable, the overdispersion will be compared to Florida-specific SPFs and *SafetyAnalyst* default SPFs for each category in Chapter 6. The model with a lower overdispersion is preferred to a model with higher overdispersion (Washington et al., 2005).

**Table 5-1 Florida-Specific SPFs for Freeways**

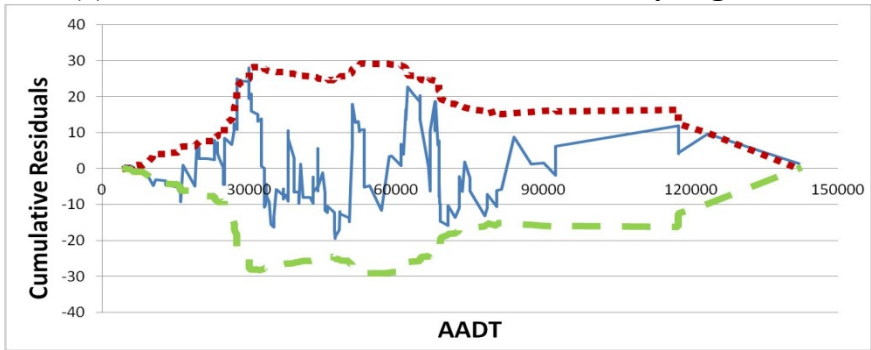
Category	Severity	Coefficient				Over-dispersion Parameter
		$\alpha$	P-Value	$\beta$	P-Value	
<b>Urban</b>						
4-Lane Basic Freeway Segments	Total	-9.372	<0.0001	1.086	<0.0001	0.633
	FI	-10.745	<0.0001	1.144	<0.0001	0.565
4-Lane Segments within Interchange Influence Area	Total	-11.656	<0.0001	1.302	<0.0001	0.355
	FI	-12.143	<0.0001	1.281	<0.0001	0.310
6-Lane Basic Freeway Segments	Total	-13.407	<0.0001	1.458	<0.0001	0.645
	FI	-14.548	<0.0001	1.487	<0.0001	0.611
6-Lane Segments within Interchange Influence Area	Total	-15.088	<0.0001	1.602	<0.0001	0.364
	FI	-15.820	<0.0001	1.595	<0.0001	0.307
(8+)-Lane Basic Freeway Segments	Total	-6.847	0.0019	0.907	0.0002	0.725
	FI	-7.239	0.0015	0.872	<0.0001	0.707
(8+)-Lane Segments within Interchange Influence Area	Total	-5.430	0.0003	0.791	<0.0001	0.520
	FI	-7.544	<0.0001	0.903	<0.0001	0.445
<b>Rural</b>						
4-Lane Basic Freeway Segments	Total	-11.412	<0.0001	1.238	<0.0001	0.233
	FI	-11.024	<0.0001	1.145	<0.0001	0.213
4-Lane Segments within Interchange Influence Area	Total	-10.572	<0.0001	1.184	<0.0001	0.312
	FI	-10.467	<0.0001	1.119	<0.0001	0.244
(6+)-Lane Basic Freeway Segments	Total	-11.522	<0.0001	1.234	<0.0001	0.231
	FI	-13.991	<0.0001	1.379	<0.0001	0.168
(6+)-Lane Segments within Interchange Influence Area	Total	-11.610	0.0003	1.273	<0.0001	0.316
	FI	-12.063	0.0002	1.244	<0.0001	0.249



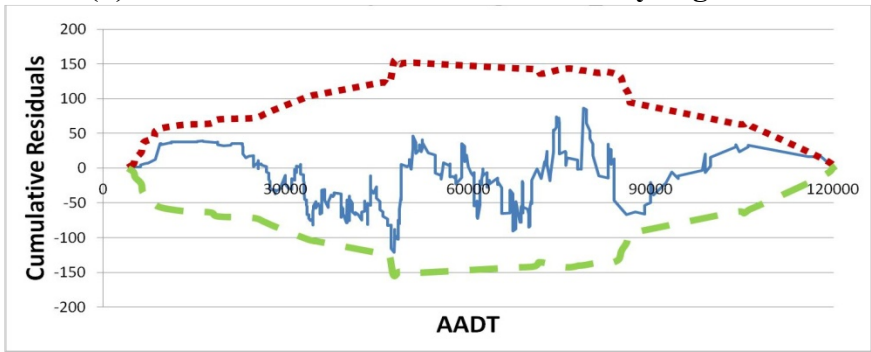
Figures 5-1 to 5-5 plot the cumulative residuals of total and FI crashes, as a function of AADT, for urban and rural basic freeway segments and interchange influence areas. As noted in these figures, the solid blue line is the cumulative residuals; the shorter red dash line and the longer green dash line represent the two standard deviations ( $+2\sigma$  and  $-2\sigma$  boundaries), respectively. Figures 5-1 and 5-2 indicate that for urban 4-lane and 6-lane freeway segments and interchange influence areas, the Florida-specific models are well-fitted as the cumulative residuals oscillate around 0 and do not stray beyond the  $\pm 2\sigma$  boundaries. The cumulative residuals in Figure 5-3, especially the cumulative residual of total crashes for urban (8+)-lane freeway segments, as shown in Figure 5-3 (a), occasionally stray beyond the  $\pm 2\sigma$  boundaries. In the CURE plots, urban (8+)-lane freeways produced the worse results when compared to urban 4-lane and 6-lane freeways, which are similar to the earlier results based on the p-value and the overdispersion values (the overdispersion values for urban (8+) -lane freeway segments are relatively higher than that for other categories). In Figures 5-4 and 5-5, some bias is apparent in the plots of several categories, specifically in Figures 5-4 (c) and (d), as well as Figures 5-5 (a) and (c). In these plots, the line of cumulative residuals consistently drifts below 0, and is indicative of a model that consistently overestimates the crash count. However, the overall fit to the data is still preferable, especially considering other statistical measures, including the p-value and the overdispersion parameter.



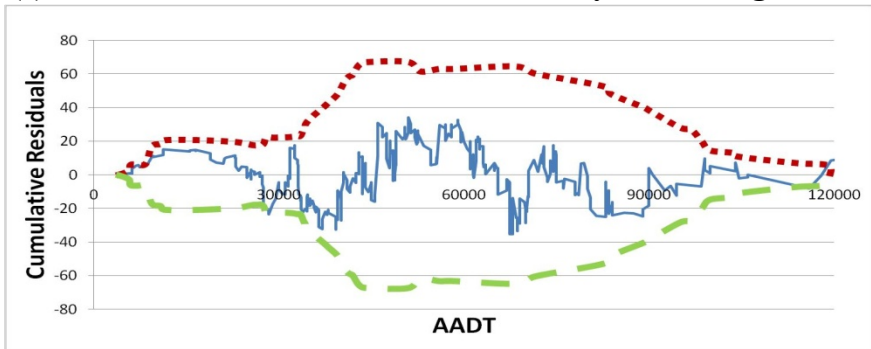
**(a) Total Crashes on Urban 4-Lane Freeway Segments**



**(b) FI Crashes on Urban 4-Lane Freeway Segments**



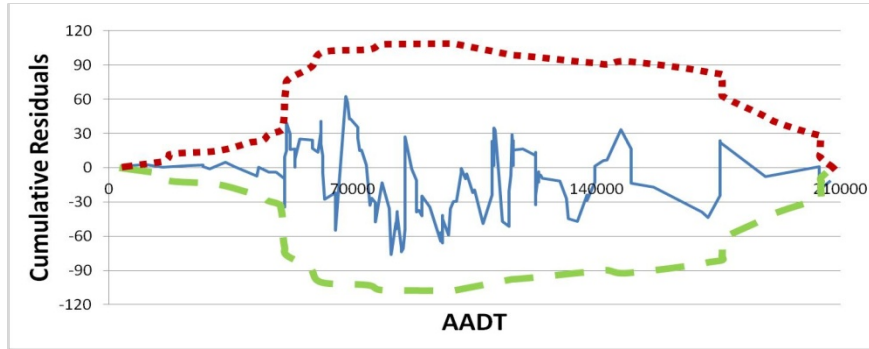
**(c) Total Crashes on Urban 4-Lane Freeway Interchange Areas**



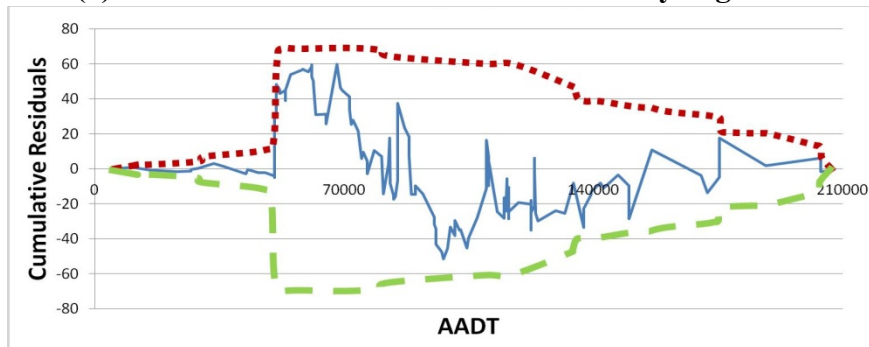
**(d) FI Crashes on Urban 4-Lane Freeway Interchange Areas**

— Cumulative Residuals      ······ + 2σ      - - - - - 2σ

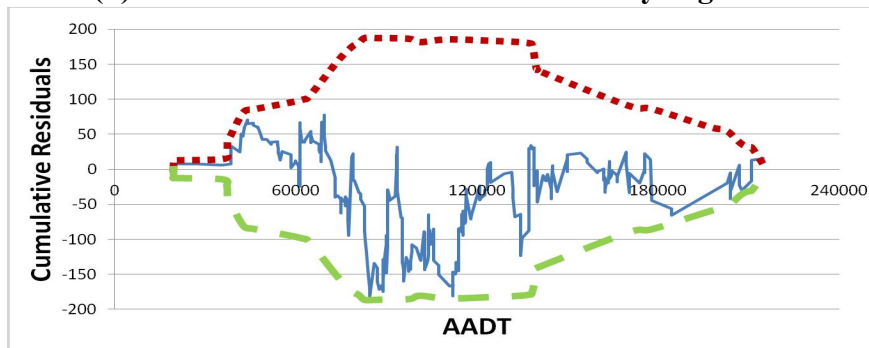
**Figure 5-1 CURE Plots for Urban 4-Lane Freeways**



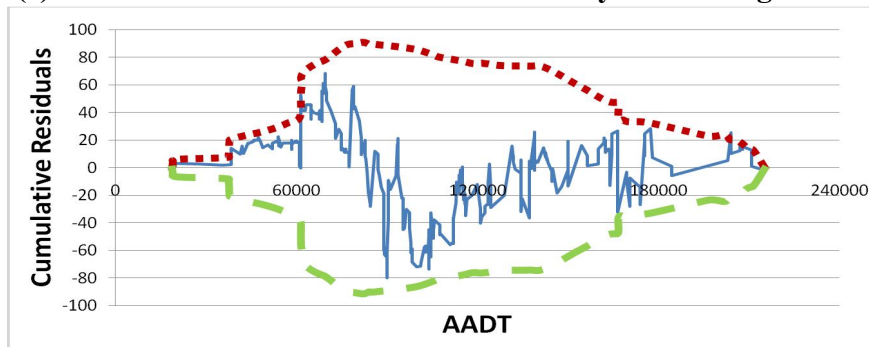
(a) Total Crashes on Urban 6-Lane Freeway Segments



(b) FI Crashes on Urban 6-Lane Freeway Segments



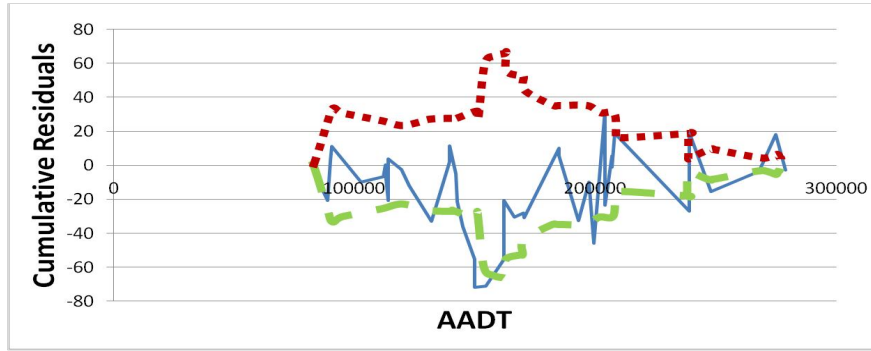
(c) Total Crashes on Urban 6-Lane Freeway Interchange Areas



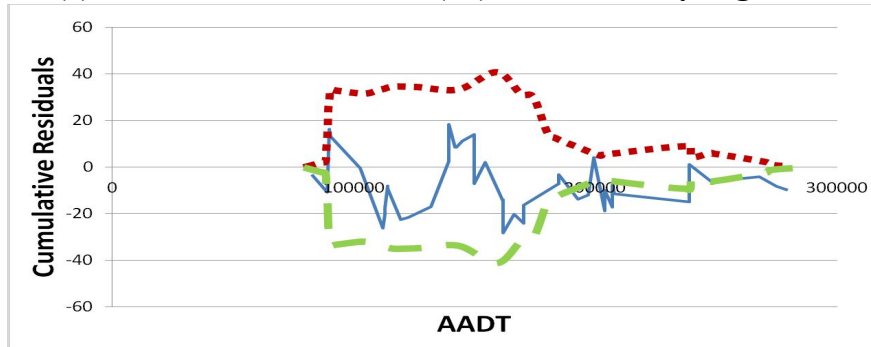
(d) FI Crashes on Urban 6-Lane Freeway Interchange Areas

— Cumulative Residuals      ····· + 2σ      - - - - - 2σ

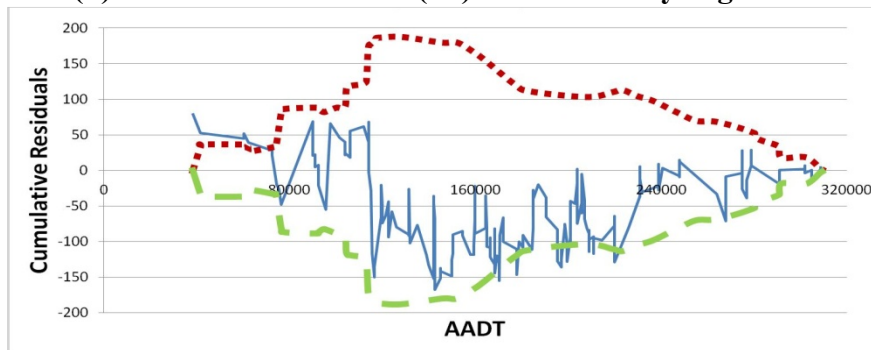
Figure 5-2 CURE Plots for Urban 6-Lane Freeways



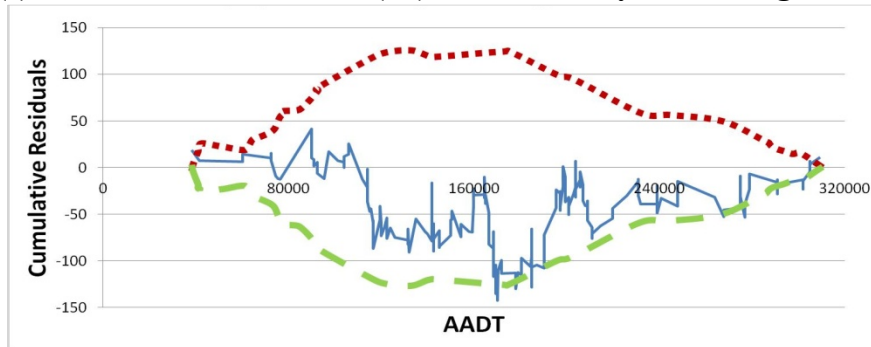
(a) Total Crashes on Urban (8+)-Lane Freeway Segments



(b) FI Crashes on Urban (8+)-Lane Freeway Segments



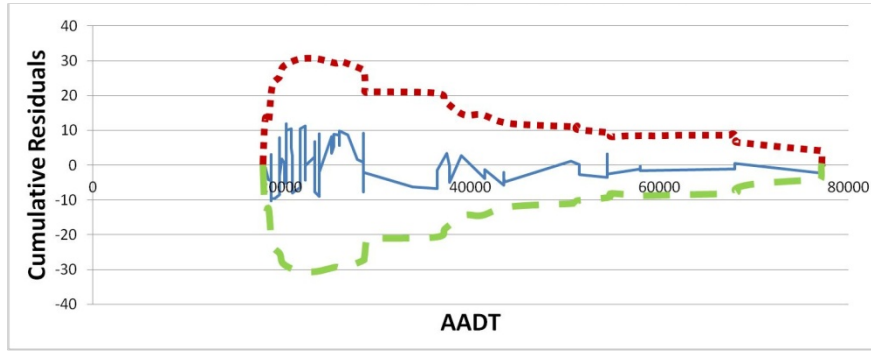
(c) Total Crashes on Urban (8+)-Lane Freeway Interchange Areas



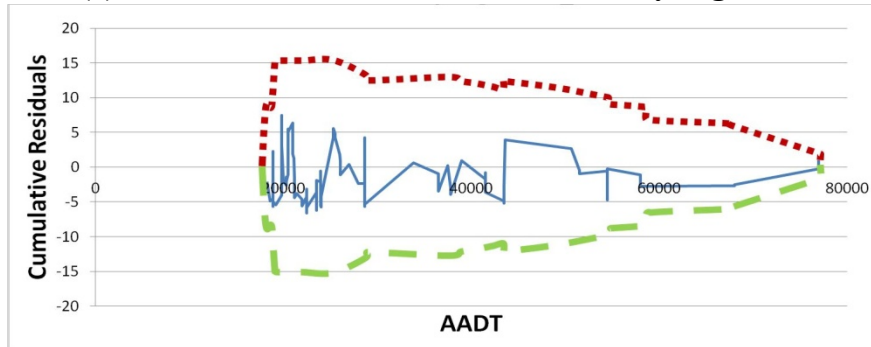
(d) FI Crashes on Urban (8+)-Lane Freeway Interchange Areas

— Cumulative Residuals      ····· + 2σ      - - - - - 2σ

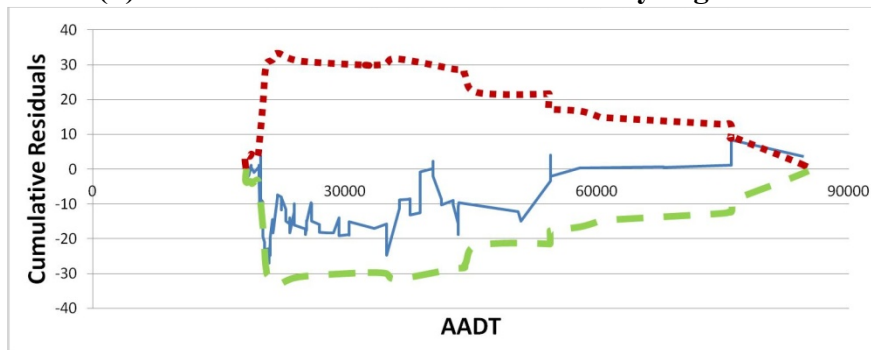
Figure 5-3 CURE Plots for Urban (8+)-Lane Freeways



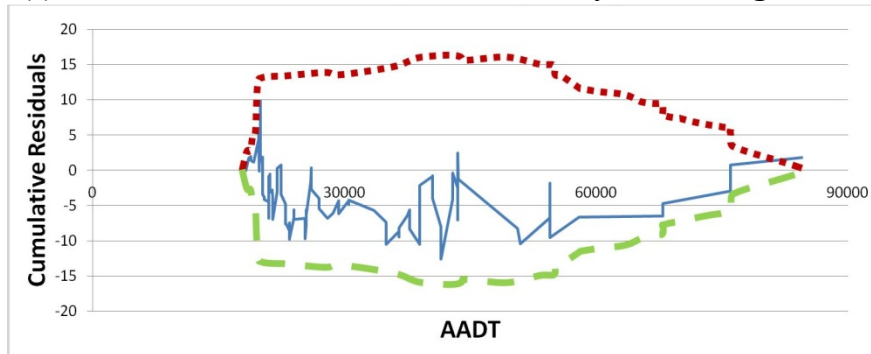
(a) Total Crashes on Rural 4-Lane Freeway Segments



(b) FI Crashes on Rural 4-Lane Freeway Segments



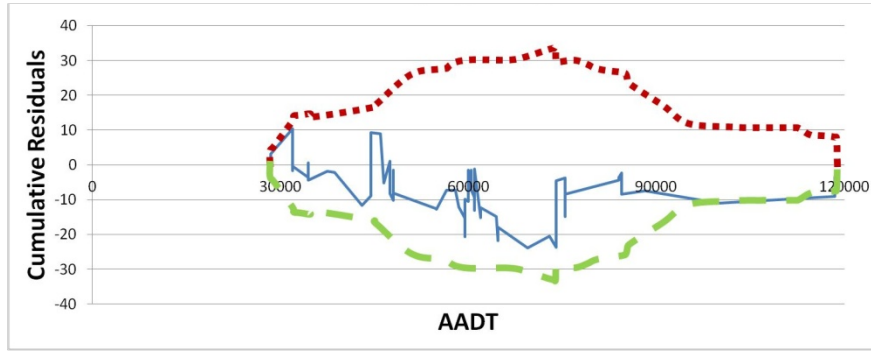
(c) Total Crashes on Rural 4-Lane Freeway Interchange Areas



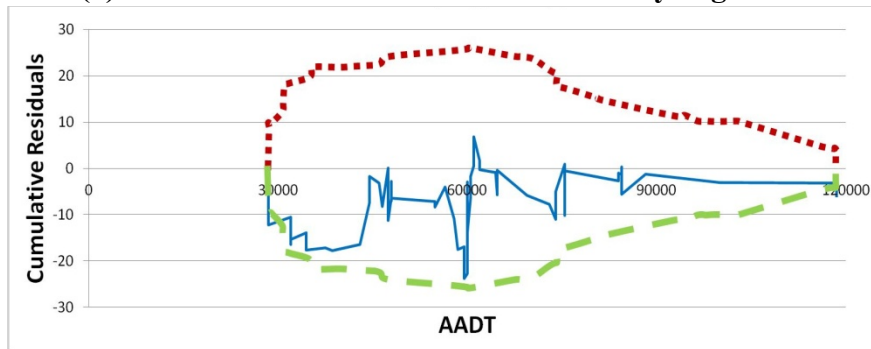
(d) FI Crashes on Rural 4-Lane Freeway Interchange Areas

— Cumulative Residuals      ······ + 2σ      - - - - - 2σ

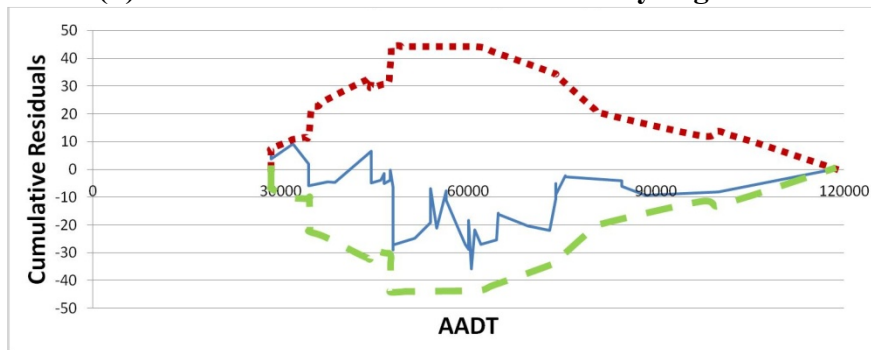
Figure 5-4 CURE Plots for Rural 4-Lane Freeways



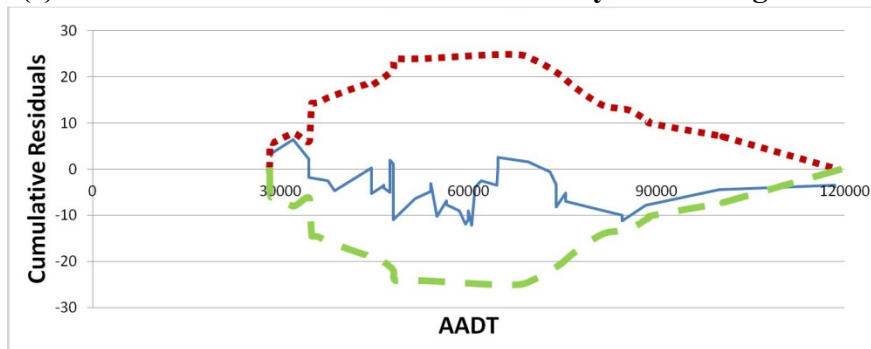
(a) Total Crashes on Rural 6-Lane Freeway Segments



(b) FI Crashes on Rural 6-Lane Freeway Segments



(c) Total Crashes on Rural 6-Lane Freeway Interchange Areas



(d) FI Crashes on Rural 6-Lane Freeway Interchange Areas

— Cumulative Residuals      ····· + 2σ      - - - - - 2σ

Figure 5-5 CURE Plots for Rural (6+)-Lane Freeways

Comparisons of goodness-of-fit tests were performed on both the calibration and validation data for freeways, as shown in Table 5-2. The purpose of this comparison is to assess the transferability of the Florida-specific SPF models, which were developed using the calibration data, to the validation data (Young and Park, 2012). The values of MSE and MSPE are similar in magnitude, particularly for models of urban 4-lane and 6-lane, and rural 4-lane freeway segments and interchange influence areas, implying that the validation data fit the model similar to the calibration data (Washington et al., 2005). In most cases, the  $R^2_{FT}$  coefficients are a slightly higher for the calibration data than that for the validation data. This could be due to the smaller sample size in the validation data than in the calibration data (Young and Park, 2012). For MSE, MSPE, and MAD, lower values are preferable to higher values; for  $R^2_{FT}$ , higher values indicate a better fit. However, it is difficult to assess if the values of MSE, MSPE, and MAD are higher or lower since there are no reference values of comparison. In Chapter 6, the values of these goodness-of-fit measures will be compared for Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data, so as to determine the best-fitted SPFs for different roadway categories in Florida.

**Table 5-2 Goodness-of-Fit Tests for Freeways**

Category	Severity	Calibration Data (70%)		Validation Data (30%)		
		MSE	R <sup>2</sup> <sub>FT</sub>	MAD	MSPE	R <sup>2</sup> <sub>FT</sub>
<b>Urban</b>						
4-Lane Basic Freeway Segments	Total	31.87	0.378	3.12	31.64	0.318
	FI	11.92	0.402	1.88	10.21	0.311
4-Lane Segments within Interchange Influence Area	Total	27.30	0.455	3.05	25.74	0.402
	FI	10.05	0.541	1.09	8.19	0.456
6-Lane Basic Freeway Segments	Total	25.88	0.433	3.06	25.90	0.339
	FI	8.77	0.447	1.05	8.12	0.324
6-Lane Segments within Interchange Influence Area	Total	51.09	0.505	3.75	49.82	0.486
	FI	13.96	0.512	2.19	13.25	0.444
(8+)-Lane Basic Freeway Segments	Total	115.08	0.060	6.35	94.91	0.052
	FI	26.64	0.087	2.88	19.90	0.114
(8+)-Lane Segments within Interchange Influence Area	Total	38.63	0.294	3.27	36.94	0.232
	FI	12.12	0.257	2.00	11.37	0.248
<b>Rural</b>						
4-Lane Basic Freeway Segments	Total	122.08	0.231	7.83	120.61	0.210
	FI	21.14	0.156	2.90	20.92	0.144
4-Lane Segments within Interchange Influence Area	Total	135.23	0.247	8.01	133.33	0.252
	FI	27.56	0.178	3.09	27.43	0.181
(6+)-Lane Basic Freeway Segments	Total	229.71	0.267	8.57	192.78	0.234
	FI	31.68	0.204	3.10	28.55	0.214
(6+)-Lane Segments within Interchange Influence Area	Total	301.66	0.167	9.33	257.48	0.162
	FI	38.72	0.185	3.18	32.85	0.203

Table 5-3 shows the coefficients and corresponding overdispersion parameters for each of the SPFs for urban arterial streets and rural roads. In Table 5-3, the coefficients for the rural multilane undivided road category were unavailable because the sample size could not produce an accurate estimation. For other categories of urban arterial streets and rural roads shown in Table 5-3, parameters for all models were statistically significant at the 0.01% level, except for the intercept term ( $\alpha$ ) in the multilane undivided arterial streets subtype, which had a slightly higher p-value. In Table 5-3, the



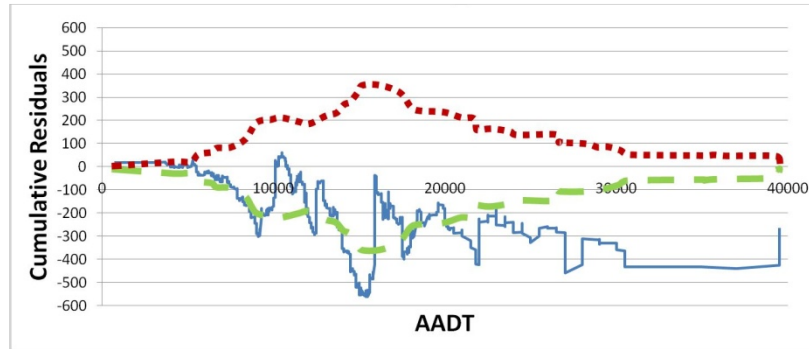
overdispersion values associated with Florida-specific SPFs for every category were low, which means the Florida-specific SPFs were reliable.

**Table 5-3 Florida-Specific SPFs for Urban Arterial Streets and Rural Roads**

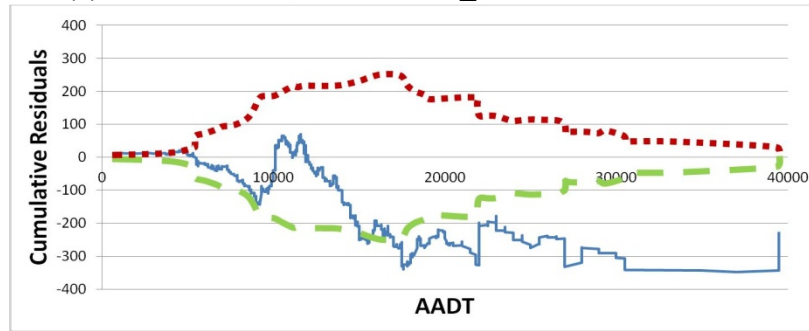
Category	Severity	Coefficient				Over-dispersion Parameter
		$\alpha$	P-Value	$\beta$	P-Value	
<b>Urban</b>						
2-lane Arterial Streets	Total	-5.480	<0.0001	0.788	<0.0001	0.757
	FI	-6.035	<0.0001	0.778	<0.0001	0.677
Multilane Undivided Arterial Streets	Total	-4.520	0.0031	0.759	<0.0001	0.736
	FI	-3.577	0.0094	0.584	<0.0001	0.568
Multilane Divided Arterial Streets	Total	-7.062	<0.0001	0.939	<0.0001	0.667
	FI	-7.783	<0.0001	0.940	<0.0001	0.561
One-way Arterial Streets	Total	-3.151	0.0003	0.612	<0.0001	0.822
	FI	-2.912	0.0007	0.467	<0.0001	0.815
<b>Rural</b>						
2-lane Roads	Total	-6.829	<0.0001	0.862	<0.0001	0.488
	FI	-7.267	<0.0001	0.848	<0.0001	0.486
Multilane Undivided Roads	Total	-	-	-	-	-
	FI	-	-	-	-	-
Multilane Divided Roads	Total	-5.213	<0.0001	0.673	<0.0001	0.461
	FI	-5.799	<0.0001	0.671	<0.0001	0.428

Figures 5-6 to 5-9 illustrate the CURE plots for urban arterial streets and rural roads. As noted in these figures, the solid blue line represents the cumulative residuals; the shorter red dash line and the longer green dash line represent the two standard deviations ( $+2\sigma$  and  $-2\sigma$  boundaries), respectively. Figures 5-7 and 5-9 show that for urban and rural multilane arterial streets, the models fluctuate around the horizontal axis and lie between the  $\pm 2\sigma$  boundaries, which are an indication of well-fitted models. In Figures 5-9 (a) and (b), the cumulative residuals consistently drift upwards in the plots of rural 2-lane roads, which indicate that the model underestimated the crash count. In Figure 5-6, for urban 2-lane arterial streets, most of the cumulative residuals of both total crashes and FI crashes exceed the  $\pm 2\sigma$  boundary. In Figure 5-8, the cumulative residual

of FI crashes for urban one-way arterial streets consistently drifts above 0, which shows that the model consistently underestimates the crash count. From the CURE plots, urban 2-lane and one-way arterial streets produced the worse results when compared with other categories, which is similar to the earlier results based on the overdispersion values (the overdispersion values for the two subtypes are relatively higher than that for other categories).



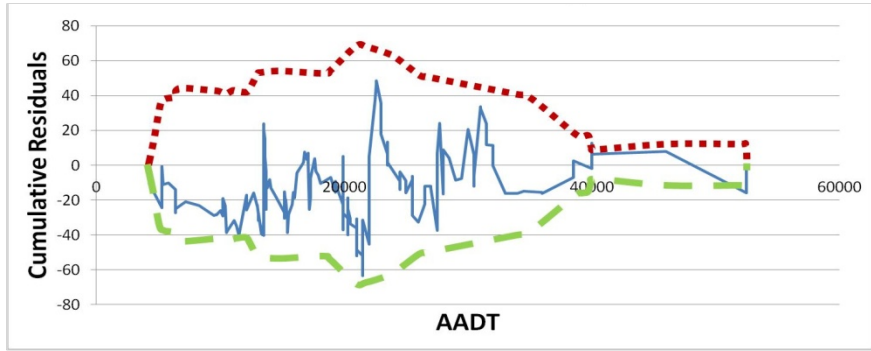
**(a) Total Crashes on Urban 2\_Lane Arterial Streets**



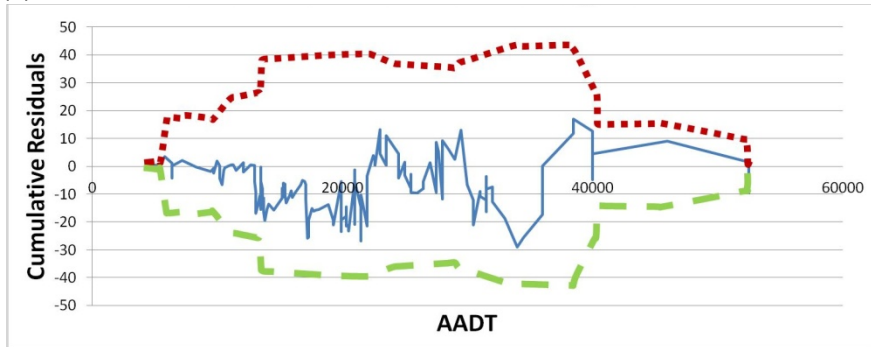
**(b) FI Crashes on Urban 2\_Lane Arterial Streets**

— Cumulative Residuals      ····· + 2σ      - - - - - 2σ

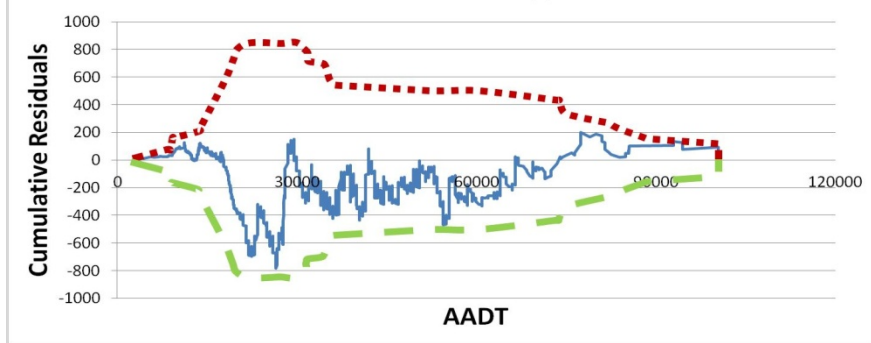
**Figure 5-6 CURE Plots for Urban 2-Lane Arterial Streets**



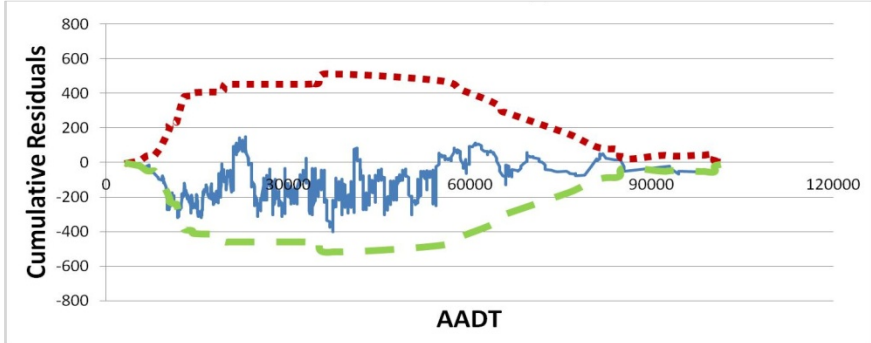
(a) Total Crashes on Urban Multilane Undivided Arterial Streets



(b) FI Crashes on Urban Multilane Undivided Arterial Streets



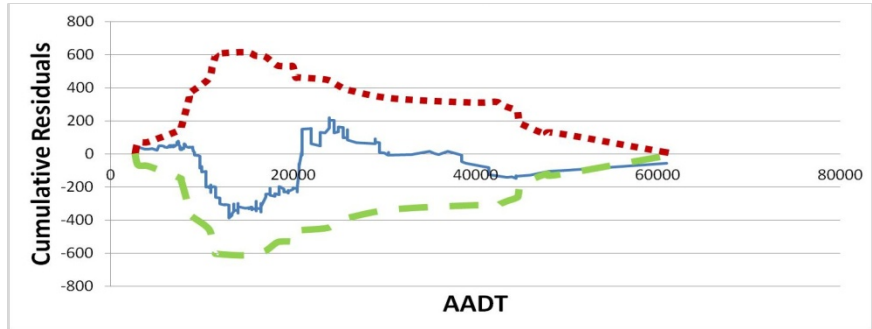
(c) Total Crashes on Urban Multilane Divided Arterial Streets



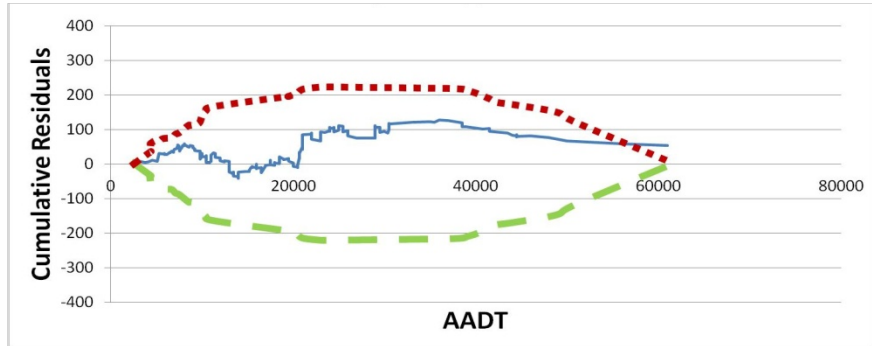
(d) FI Crashes on Urban Multilane Divided Arterial Streets

— Cumulative Residuals      ····· + 2σ      - - - - - 2σ

Figure 5-7 CURE Plots for Urban Multilane Arterial Streets



**(a) Total Crashes on Urban One-Way Arterial Streets**

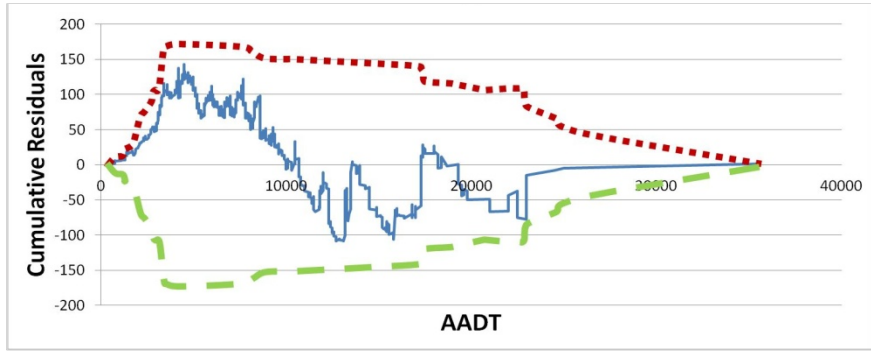


**(b) FI Crashes on Urban One-Way Arterial Streets**

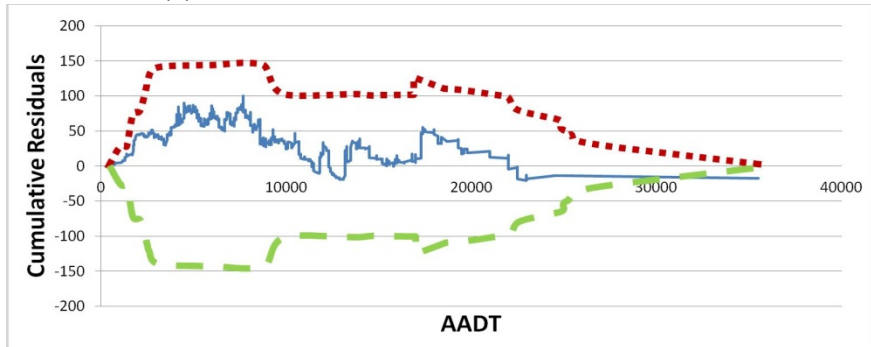
— Cumulative Residuals      ······ + 2σ      - - - - - 2σ

**Figure 5-8 CURE Plots for Urban One-Way Arterial Streets**

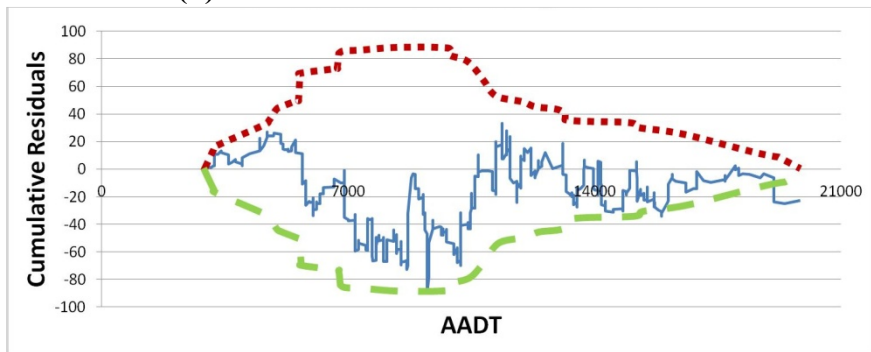
Comparisons of goodness-of-fit tests were performed on both the calibration data and validation data for urban arterial streets and rural roads. Table 5-4 shows the results of the comparisons. For the models for urban divided arterial streets, rural divided roads, and rural 2-lane roads, the values of MSE and MSPE are very similar, which represents a high level of transferability of the developed models using the calibration data to the validation data. The same goes for freeways: the  $R^2_{FT}$  coefficients are slightly higher for the calibration data than that for the validation data on both urban arterial streets and rural roads, which most of the time, may be the result of the relatively smaller sample size in the validation data than the calibration data.



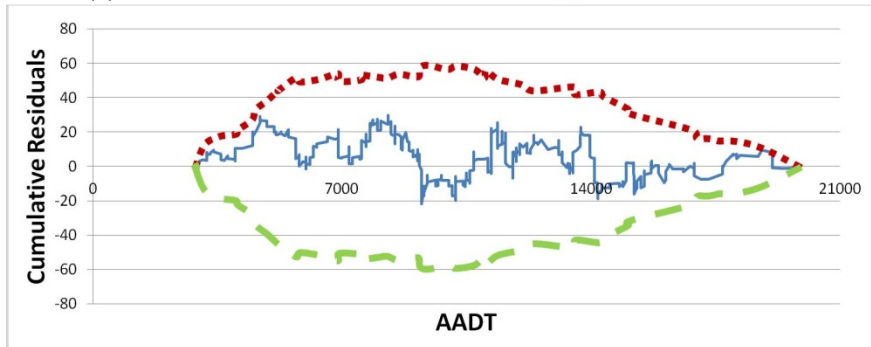
**(a) Total Crashes on Rural 2-Lane Roads**



**(b) FI Crashes on Rural 2-Lane Roads**



**(c) Total Crashes on Rural Multilane Divided Roads**



**(d) FI Crashes on Rural Multilane Divided Roads**

— Cumulative Residuals      - - - + 2σ      - - - - 2σ

**Figure 5-9 CURE Plots for Rural Roads**

**Table 5-4 Goodness-of-Fit Tests for Urban Arterial Streets and Rural Roads**

Category	Severity	Calibration Data (70%)		Validation Data (30%)		
		MSE	R <sup>2</sup> <sub>FT</sub>	MAD	MSPE	R <sup>2</sup> <sub>FT</sub>
<b>Urban</b>						
2-lane Arterial Streets	Total	288.31	0.195	10.67	275.12	0.160
	FI	91.17	0.161	6.02	91.94	0.124
Multilane Undivided Arterial Streets	Total	125.10	0.143	7.11	121.05	0.130
	FI	32.22	0.157	3.05	30.41	0.119
Multilane Divided Arterial Streets	Total	57.04	0.294	5.24	57.47	0.295
	FI	19.63	0.278	2.07	18.92	0.220
One-way Arterial Streets	Total	126.37	0.112	7.31	123.44	0.094
	FI	35.79	0.126	3.23	35.16	0.107
<b>Rural</b>						
2-lane Roads	Total	102.77	0.275	6.35	102.55	0.288
	FI	25.78	0.228	2.90	25.09	0.230
Multilane Divided Roads	Total	95.36	0.281	6.14	94.36	0.244
	FI	23.55	0.253	2.55	23.18	0.207

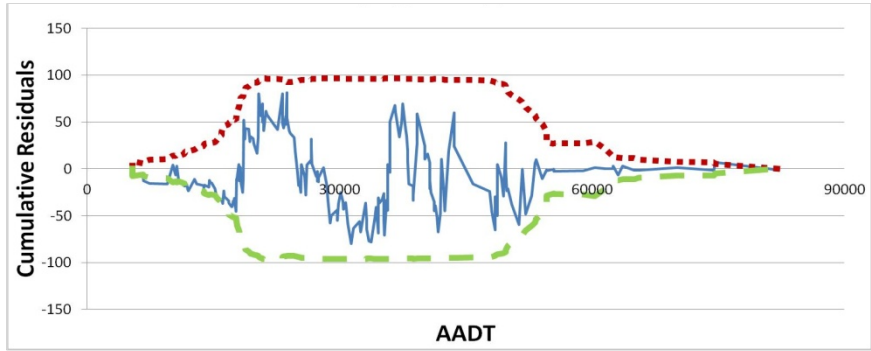
### 5.3. Intersections

Florida-specific SPFs were developed for each of the four categories of signalized intersections, for both total crashes and FI crashes. Table 5-5 shows the coefficients and corresponding overdispersion parameters for each of the SPFs for urban and rural signalized intersections. Most of the model parameters of urban signalized intersections were found to be statistically significant at the 0.01% level, while the coefficients of major AADT ( $\beta_1$ ) and minor AADT ( $\beta_2$ ) had fewer higher p-values for rural signalized intersections. This result is likely due to the smaller sample size for rural signalized intersections than urban signalized intersections. However, these parameters still meet the requirement of the significance level. In Table 5-5, the overdispersion parameters for the four categories of signalized intersections were low (most of the overdispersion parameters were less than 0.5), which means the Florida-specific SPFs were reliable.

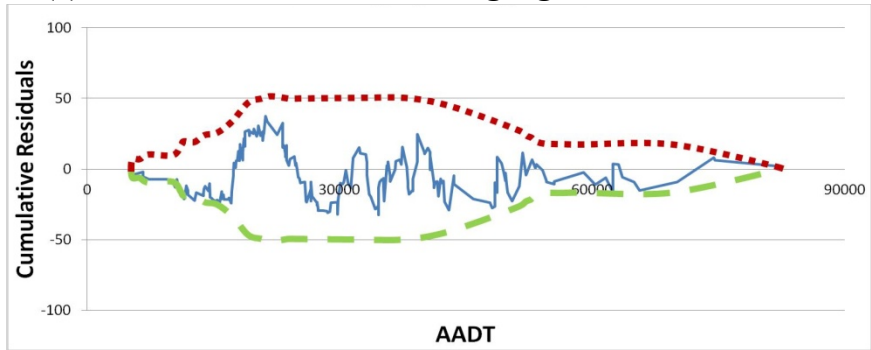
**Table 5-5 Florida-Specific SPFs for Intersections**

Category	Severity	Coefficient						Over-dispersion Parameter
		$\alpha$	P-Value	$\beta_1$	P-Value	$\beta_2$	P-Value	
<b>Urban</b>								
3-Leg	Total	-9.589	<0.0001	0.725	<0.0001	0.453	<0.0001	0.404
	FI	-8.354	<0.0001	0.605	<0.0001	0.360	<0.0001	0.310
4-Leg	Total	-8.877	<0.0001	0.740	<0.0001	0.404	<0.0001	0.457
	FI	-9.104	<0.0001	0.674	<0.0001	0.408	<0.0001	0.349
<b>Rural</b>								
3-Leg	Total	-9.376	<0.0001	0.664	0.0003	0.495	0.0012	0.515
	FI	-9.784	<0.0001	0.640	0.0007	0.488	0.0027	0.468
4-Leg	Total	-7.143	<0.0001	0.488	0.0074	0.481	0.0021	0.440
	FI	-6.770	<0.0001	0.351	0.0076	0.505	0.0028	0.436

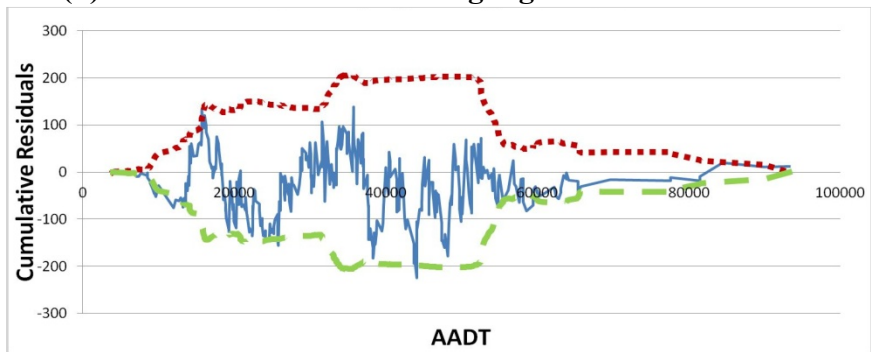
Figures 5-10 and 5-11 show CURE plots for urban 3-leg signalized, urban 4-leg signalized, rural 3-leg signalized, and rural 4-leg signalized intersections, based on both total and FI crashes. As noted in the figures, the solid blue line is the cumulative residuals, and the shorter red dash line and longer green dash line represent the two standard deviations, respectively. The figures show that for all four categories, the cumulative residuals of Florida-specific SPFs are confined within the two standard deviations along the entire range of major AADT. Among the four categories, for urban signalized intersections, the cumulative residuals oscillate above and below 0, which indicates a well-fitted model. For rural signalized intersections, some bias is apparent in the plots, which is indicative of a model over- or under-estimating the predicted number of crashes for a consistent period. This may be due to the smaller sample size for rural intersections; however, the overall fit to the data is still considered good.



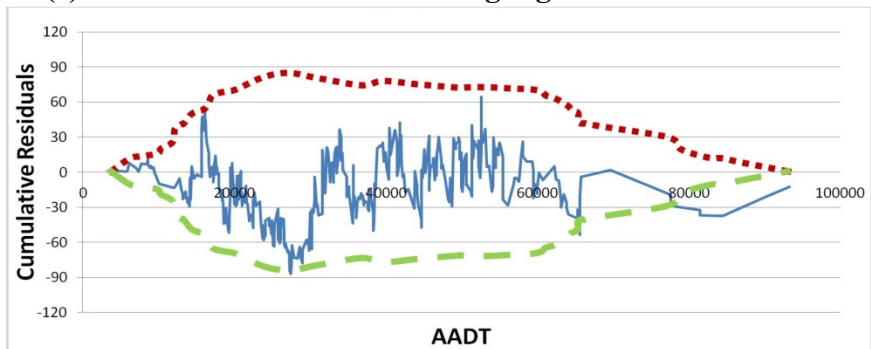
**(a) Total Crashes on Urban 3-Leg Signalized Intersections**



**(b) FI Crashes on Urban 3-Leg Signalized Intersections**



**(c) Total Crashes on Urban 4-Leg Signalized Intersections**

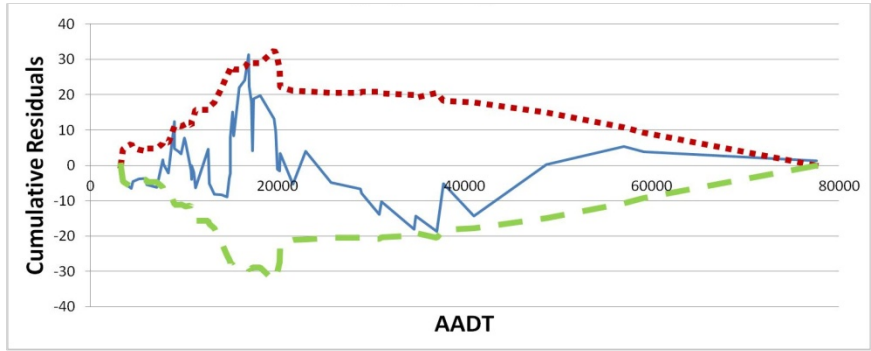


**(d) FI Crashes on Urban 4-Leg Signalized Intersections**

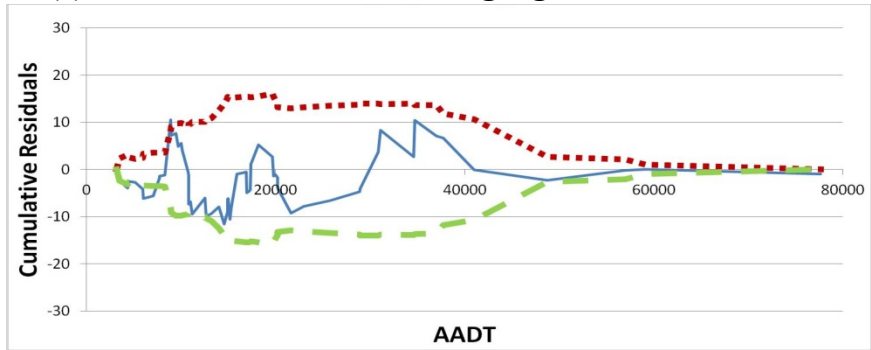
— Cumulative Residuals      ····· + 2σ      - - - - - 2σ

**Figure 5-10 CURE Plots for Urban Intersections**

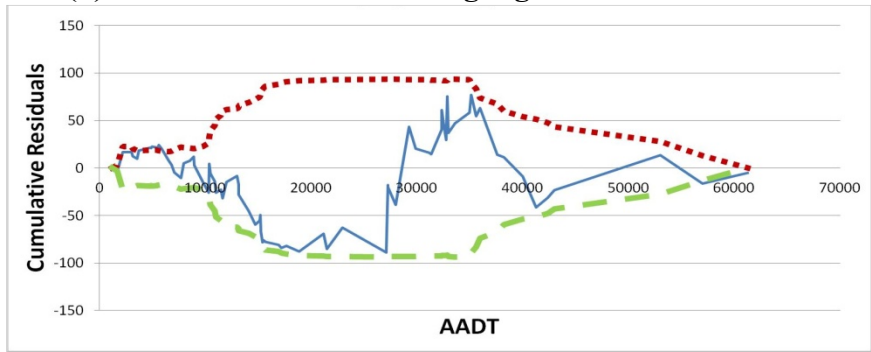




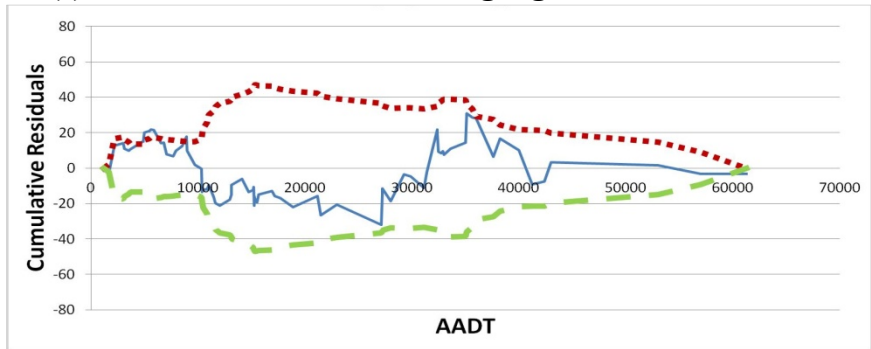
**(a) Total Crashes on Rural 3-Leg Signalized Intersections**



**(b) FI Crashes on Rural 3-Leg Signalized Intersections**



**(c) Total Crashes on Rural 4-Leg Signalized Intersections**



**(d) FI Crashes on Rural 4-Leg Signalized Intersections**

— Cumulative Residuals     
 - - -  $+ 2\sigma$      
 - - -  $- 2\sigma$

**Figure 5-11 CURE Plots for Rural Intersections**

Comparisons of goodness-of-fit tests were also performed between the calibration data and the validation data for signalized intersections. Table 5-6 shows that the values of MSE and MSPE calculated from calibration and validation data are similar in magnitude. The Freeman-Tukey  $R^2$  coefficients are slightly higher for the calibration data than that for the validation data in most cases. These results for signalized intersections are the same as that for roadway segments mentioned in the previous section.

**Table 5-6 Goodness-of-Fit Tests for Intersections**

Category	Severity	Calibration Data (70%)		Validation Data (30%)		
		MSE	$R^2_{FT}$	MAD	MSPE	$R^2_{FT}$
<b>Urban</b>						
3-Leg	Total	30.68	0.431	3.13	30.04	0.414
	FI	9.50	0.387	1.06	8.73	0.362
4-Leg	Total	38.18	0.485	3.25	37.20	0.450
	FI	10.22	0.469	1.17	9.95	0.427
<b>Rural</b>						
3-Leg	Total	192.47	0.347	8.21	187.14	0.287
	FI	26.77	0.384	3.09	26.26	0.356
4-Leg	Total	57.36	0.633	5.61	54.25	0.643
	FI	14.20	0.602	2.18	13.69	0.583

#### 5.4. Ramps

Florida-specific SPFs were developed for each category of ramps, for both total crashes and FI crashes. As mentioned earlier, the ramp classification in Florida is different from the default classification used in *SafetyAnalyst*. Therefore, the Florida-specific SPFs of only the following subtypes were developed: diamond ramps for both rural and urban areas, urban partial diamond ramps, and urban partial cloverleaf ramps. The sample size of urban trumpet ramps, urban direct connections, and rural partial cloverleaf ramps was inadequate; the sample size could not accurately estimate regression coefficients. Table 5-7 shows the coefficients and corresponding

overdispersion parameters for each of the SPFs for ramps. In Table 5-7, parameters for all models were found to be statistically significant. The overdispersion parameter was used to account for dispersion. A model with a lower value of the overdispersion parameter is more reliable than that with a higher overdispersion parameter. In Table 5-7, the overdispersion parameters associated with Florida-specific SPFs for rural ramps were lower than that for urban ramps; however, the overall magnitude for urban ramps is still preferable.

**Table 5-7 Florida-Specific SPFs for Ramps**

Category	Severity	Coefficient				Over-dispersion Parameter
		$\alpha$	P-Value	$\beta$	P-Value	
<b>Urban Diamond</b>						
Off-ramp	Total	-2.700	0.0004	0.574	<0.0001	0.815
	FI	-3.286	<0.0001	0.554	<0.0001	0.749
On-ramp	Total	-3.419	0.0002	0.561	<0.0001	0.971
	FI	-4.402	<0.0001	0.582	<0.0001	1.063
<b>Urban Partial Diamond</b>						
Off-ramp	Total	-3.768	0.0006	0.633	0.0011	1.002
	FI	-4.105	0.0009	0.594	0.0015	0.987
On-ramp	Total	-7.398	<0.0001	1.044	<0.0001	1.776
	FI	-7.304	<0.0001	0.941	<0.0001	1.749
<b>Urban Partial Cloverleaf (Parclo Loop)</b>						
Off-ramp	Total	-3.059	0.0007	0.581	<0.0001	0.919
	FI	-4.416	<0.0001	0.633	<0.0001	0.808
On-ramp	Total	-5.223	<0.0001	0.774	<0.0001	0.718
	FI	-6.453	<0.0001	0.832	<0.0001	0.770
<b>Rural Diamond</b>						
Off-ramp	Total	-5.280	<0.0001	0.823	<0.0001	0.634
	FI	-5.299	0.0005	0.727	0.0002	0.605
On-ramp	Total	-8.287	<0.0001	1.103	<0.0001	0.419
	FI	-10.069	<0.0001	1.247	<0.0001	0.362

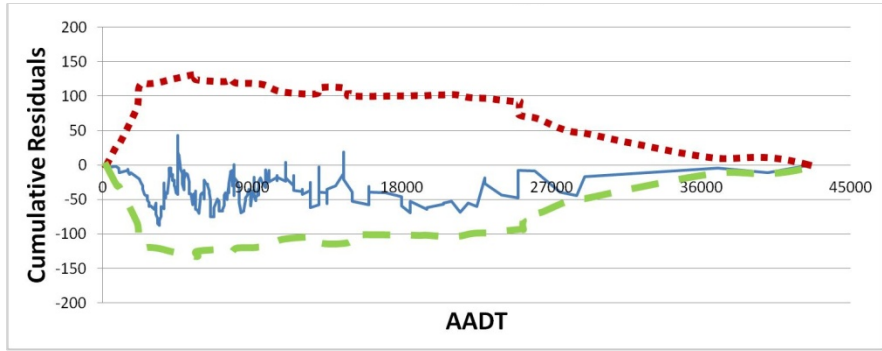
CURE plots (Figures 5-12 to 5-15) were also performed for ramps. As noted in the figures, the solid blue line represents the cumulative residuals, and the shorter red dash line and longer green dash line represent the two standard deviations, respectively. Figures 5-12 to 5-15 show that for most of the subtypes, the Florida-specific models are

good-fitted since the cumulative residuals fluctuate around 0 and fall within the  $\pm 2\sigma$  boundaries for the entire range of AADT. The cumulative residuals in Figures 5-12 (a) and (d), 5-13 (c), 5-14 (a), as well as 5-15 (a) and (c), show that some bias is apparent in the plots. This is indication of a model that consistently underestimates crash counts, but the overall fit to the data is still acceptable, especially considering other statistical measures like the overdispersion parameter.

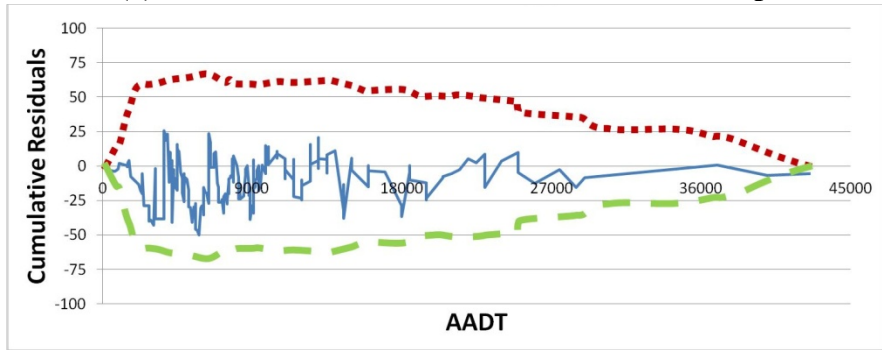
Comparisons of goodness-of-fit tests were also performed on both the calibration data and the validation data for ramps. Table 5-8 displays the results of these comparisons. The values of MSE and MSPE are similar in magnitude, particularly for models of urban partial diamond and urban partial cloverleaf ramps, implying that the validation data fit the model similar to the calibration data. The  $R^2_{FT}$  coefficients are slightly higher for the calibration data than that for the validation data in most cases, which could be due to the smaller sample size in the validation data than in the calibration data. These results for ramps are the same as that for roadway segments and intersections.

## **5.5. Summary**

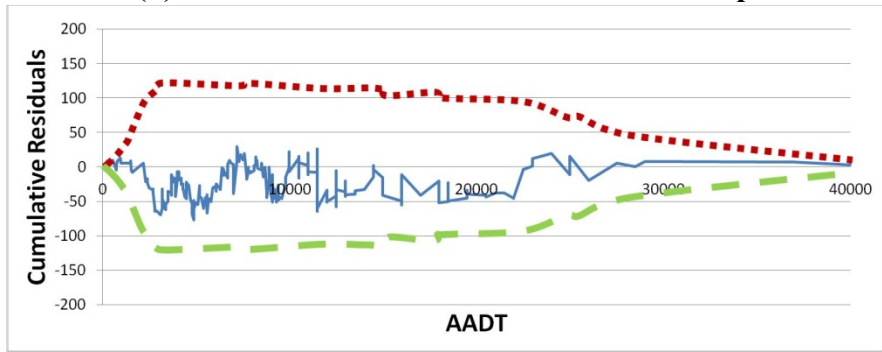
In this chapter, Florida-specific SPFs for each of the subtypes of roadway segments, intersections and ramps, for both total crashes and FI crashes, were developed using calibration data. The coefficients and corresponding overdispersion parameters of developed SPFs for each subtype were provided. Parameters for all models were found to be statistically significant, and overdispersion parameters for all subtypes were low, which indicates that the Florida-specific SPFs were reliable.



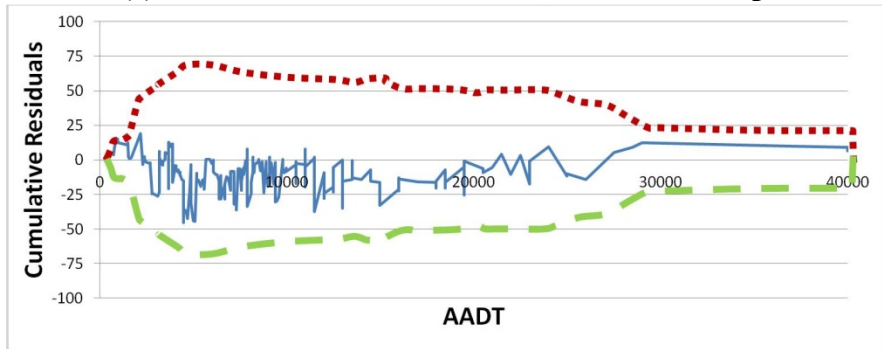
**(a) Total Crashes on Urban Diamond Off-Ramp**



**(b) FI Crashes on Urban Diamond Off-Ramp**



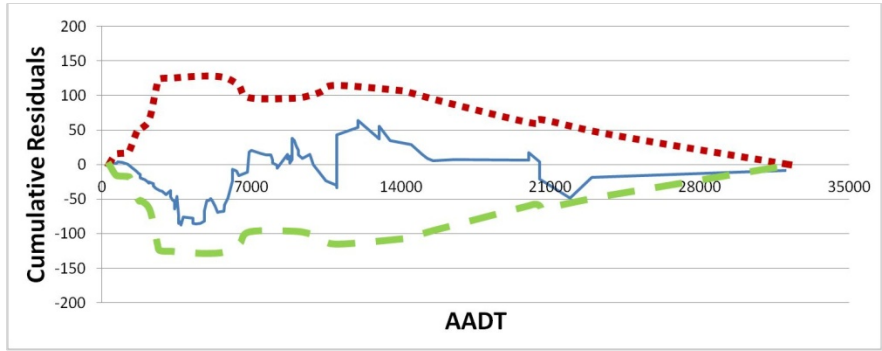
**(c) Total Crashes on Urban Diamond On-Ramp**



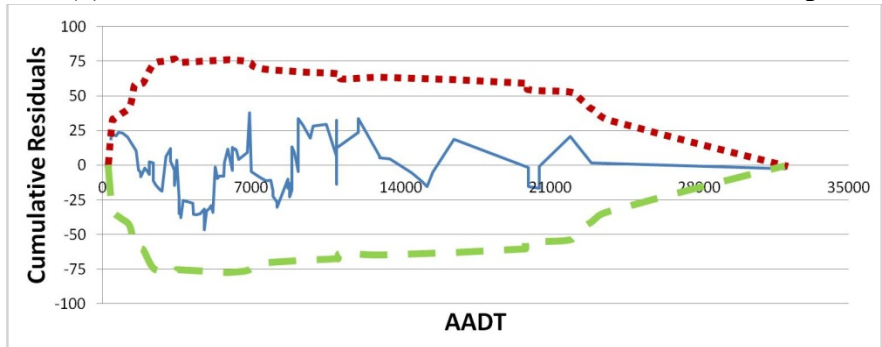
**(d) FI Crashes on Urban Diamond On-Ramp**

— Cumulative Residuals      ····· + 2σ      - - - - - 2σ

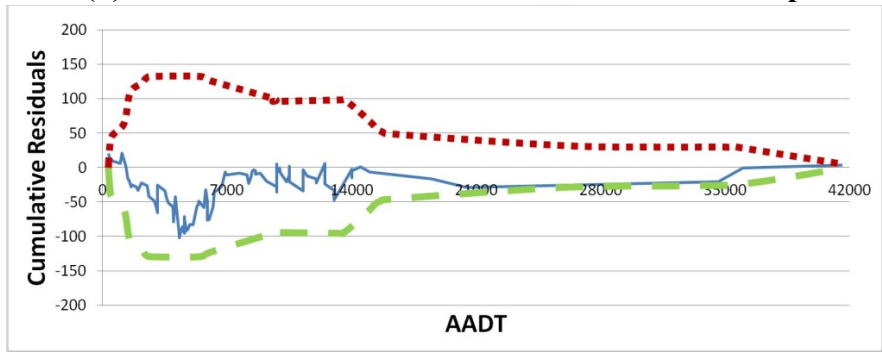
**Figure 5-12 CURE Plots for Urban Diamond Ramps**



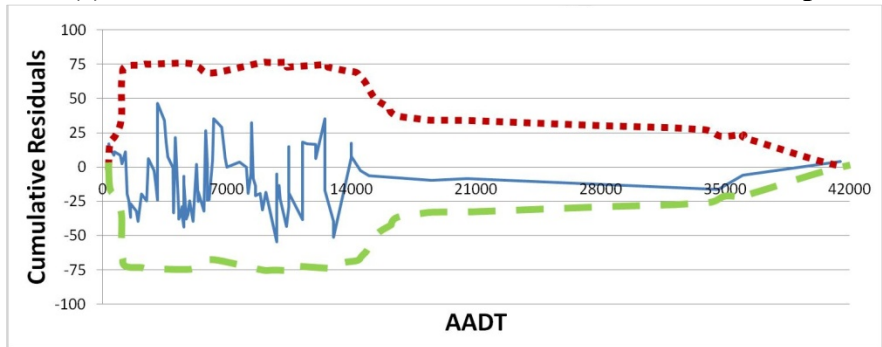
**(a) Total Crashes on Urban Partial Diamond Off-Ramp**



**(b) FI Crashes on Urban Partial Diamond Off-Ramp**



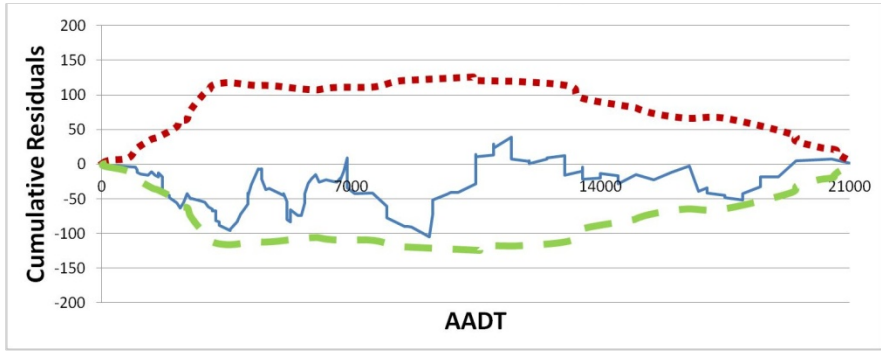
**(c) Total Crashes on Urban Partial Diamond On-Ramp**



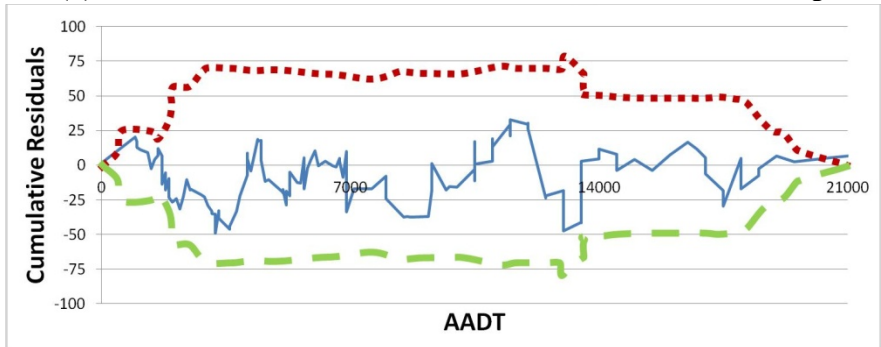
**(d) FI Crashes on Urban Partial Diamond On-Ramp**

— Cumulative Residuals      ····· + 2σ      - - - - - 2σ

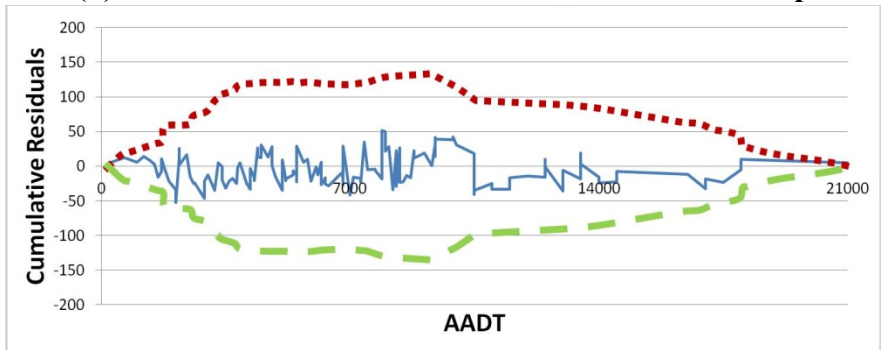
**Figure 5-13 CURE Plots for Urban Partial Diamond Ramps**



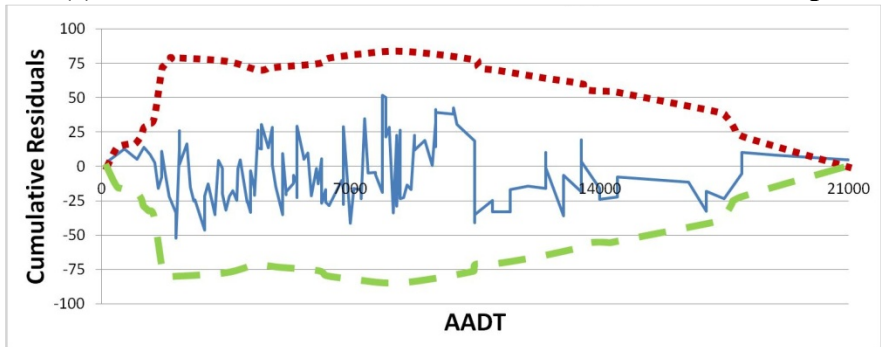
**(a) Total Crashes on Urban Partial Cloverleaf Off-Ramp**



**(b) FI Crashes on Urban Partial Cloverleaf Off-Ramp**



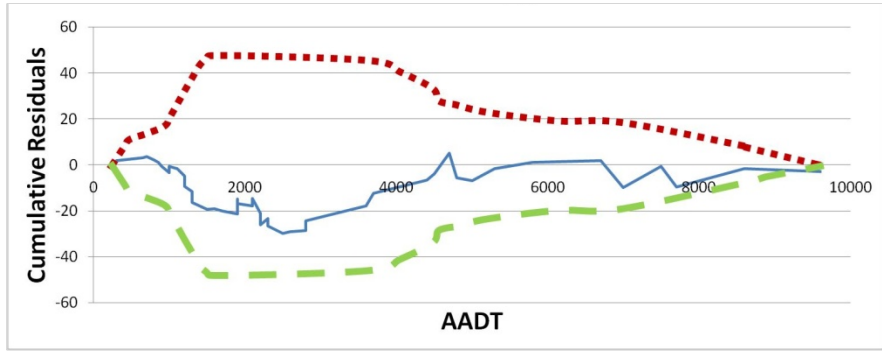
**(c) Total Crashes on Urban Partial Cloverleaf On-Ramp**



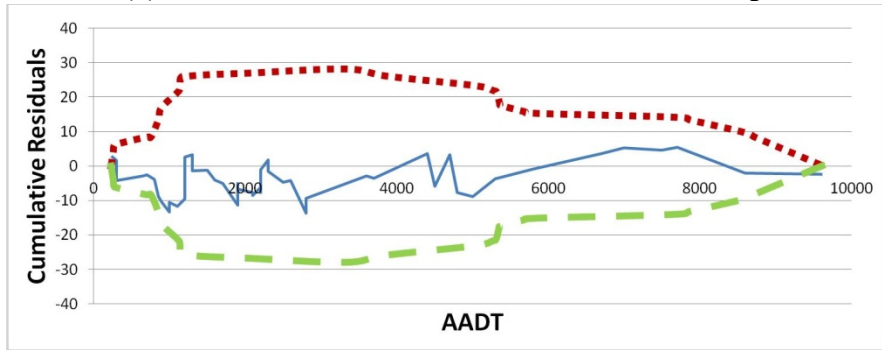
**(d) FI Crashes on Urban Partial Cloverleaf On-Ramp**

— Cumulative Residuals      ····· + 2σ      - - - -2σ

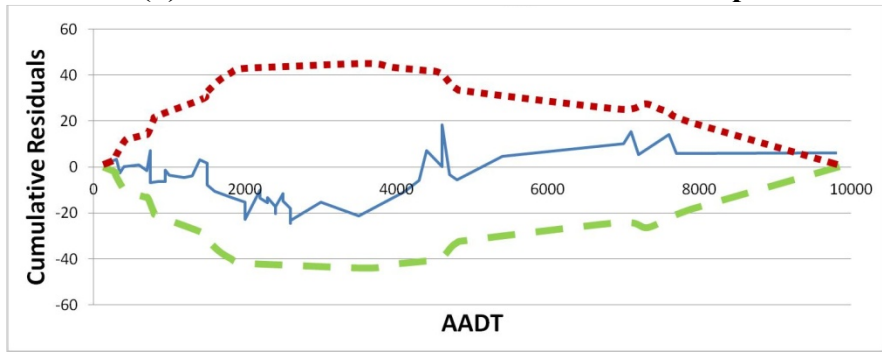
**Figure 5-14 CURE Plots for Urban Partial Cloverleaf Ramps**



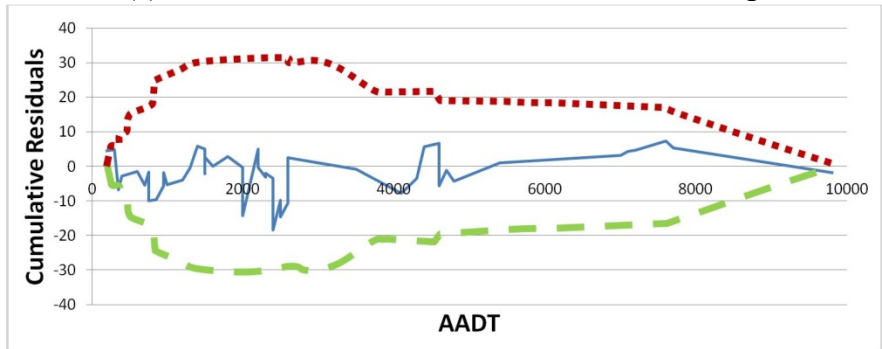
**(a) Total Crashes on Rural Diamond Off-Ramp**



**(b) FI Crashes on Rural Diamond Off-Ramp**



**(c) Total Crashes on Rural Diamond On-Ramp**



**(d) FI Crashes on Rural Diamond On-Ramp**

— Cumulative Residuals      ····· + 2σ      - - - - - 2σ

**Figure 5-15 CURE Plots for Rural Diamond Ramps**



**Table 5-8 Goodness-of-Fit Tests for Ramps**

Category	Severity	Calibration Data (70%)		Validation Data (30%)		
		MSE	R <sup>2</sup> <sub>FT</sub>	MAD	MSPE	R <sup>2</sup> <sub>FT</sub>
<b>Urban Diamond</b>						
Off-ramp	Total	163.25	0.142	8.05	150.20	0.073
	FI	12.38	0.176	2.26	10.85	0.140
On-ramp	Total	182.23	0.191	8.91	178.54	0.134
	FI	25.06	0.159	3.33	24.17	0.166
<b>Urban Partial Diamond</b>						
Off-ramp	Total	207.63	0.267	9.07	207.27	0.199
	FI	31.88	0.224	4.15	30.22	0.201
On-ramp	Total	153.35	0.261	8.23	158.07	0.234
	FI	12.49	0.230	2.78	11.33	0.218
<b>Urban Partial Cloverleaf (Parclo Loop)</b>						
Off-ramp	Total	47.32	0.275	5.83	44.92	0.249
	FI	9.15	0.263	1.37	8.53	0.233
On-ramp	Total	33.05	0.337	4.45	35.84	0.305
	FI	6.00	0.278	1.06	5.87	0.292
<b>Rural Diamond</b>						
Off-ramp	Total	140.18	0.381	7.62	131.01	0.281
	FI	10.73	0.290	1.85	9.02	0.236
On-ramp	Total	164.75	0.465	8.24	158.41	0.374
	FI	13.07	0.444	2.92	11.84	0.400

In addition, for most of the subtypes, the Florida-specific models are good-fitted because the cumulative residuals oscillate around 0 and stay inside the  $\pm 2\sigma$  boundaries. The cumulative residuals for some subtypes sometimes consistently over- or under-estimate the crash count or stray beyond the  $\pm 2\sigma$  boundaries. However, the overall fit to the data is still preferable, especially considering other statistical measures, such as the overdispersion parameter.

Comparisons of goodness-of-fit tests were performed on both the calibration data and the validation data for all subtypes. For most of the subtypes, the models present a high level of transferability, as the MSE and MSPE values are similar in magnitude. For most of the subtypes, the MSPE values are lower than the MSE values, signifying that the crash counts were not over-fitted by the regression models. Generally, when

sample sizes are smaller in the validation data, the  $R^2_{FT}$  coefficients are somewhat elevated in the calibration data. For MSE, MSPE, and MAD, lower values are preferable to higher values since  $R^2_{FT}$ , higher values indicate a better fit. In order to determine the best-fitted SPFs for different categories in Florida the values of these goodness-of-fit measures will be compared for Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data in Chapter 6.

## CHAPTER 6

### MODEL COMPARISON

#### 6.1. Introduction

The objective of this chapter is to compare the performance of the Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data. First, calibration factors were calculated as the ratio of the total observed crashes to total predicted crashes obtained from the default national SPFs for all roadway subtypes. The Florida-specific SPFs were then compared to *SafetyAnalyst* default SPFs (both calibrated to Florida data and original default model) using a number of methods, including visual plots and statistical goodness-of-fit tests. The plots of SPFs against the observed crash data were used to compare the prediction performance of the two models. Three goodness-of-fit tests represented by MAD, MSPE, and Freeman-Tukey  $R^2$  were also used for comparison in order to identify the better-fitting model.

#### 6.2. Roadway Segments

The model of *SafetyAnalyst* default SPFs calibrated to Florida data was estimated by adjusting the national default SPFs with calibration factors. Calibration factors were calculated as the ratio of the total observed crashes to predicted crashes obtained from the national default SPFs. Table 6-1 shows the two models (Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data) for each category of freeways, for both total crashes and FI crashes, using the calibration data (70% of the data). The “ODP” represents the overdispersion parameter; and “C” represents the calibration factor. As Table 6-1 shows, calibration factors, which are greater than 1, indicate that the

*SafetyAnalyst* default SPFs underestimate the crash frequency. The calibration factors, which are less than 1, indicate that the *SafetyAnalyst* default SPFs overestimate the crash frequency.

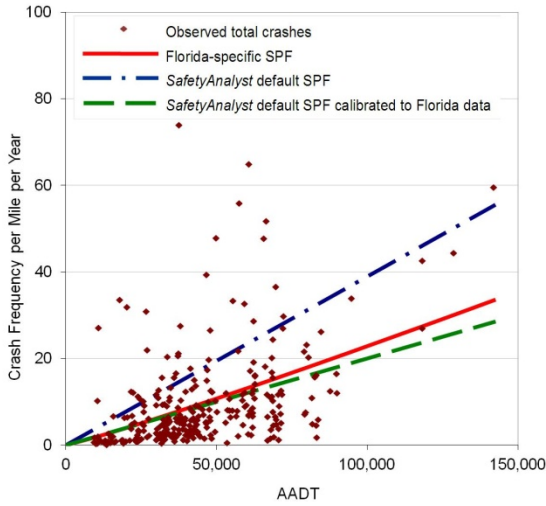
The overdispersion parameter, which indicates the statistical reliability of the SPF, was used to account for dispersion in the data. The closer the overdispersion parameter is to zero, the more statistically reliable the SPF. For urban 4-lane freeways (both segments and interchange influence areas) and urban 6-lane freeway interchange influence areas, the overdispersion values associated with Florida-specific SPFs are much lower than the corresponding default models, demonstrating that the Florida-specific SPFs are statistically more reliable. The remaining categories have moderately higher overdispersion values.

In addition, three SPFs (Florida-specific SPF, *SafetyAnalyst* default SPF, and *SafetyAnalyst* default SPF calibrated to Florida data) were plotted against the observed crash data for rural and urban freeways, for both total and FI crashes. Figures 6-1 to 6-5 show the plots of the predicted annual crash frequency per mile against AADT and observed crash frequency for all categories of freeways. As noted in the figures, the plotted data points are the observed crash frequency based on Florida data. The solid red line represents the Florida-specific SPF, the blue dash dot line represents the national default SPF used in *SafetyAnalyst*, and the green dash line represents the *SafetyAnalyst* default SPF calibrated to Florida data using a calibration factor.

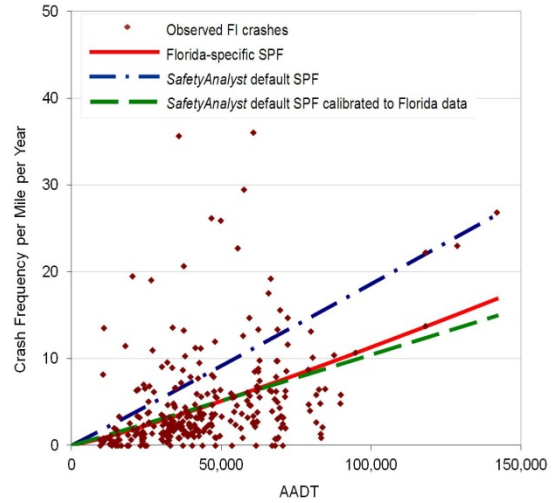
**Table 6-1 Florida-specific SPFs and *SafetyAnalyst* Default SPFs Calibrated to Florida Data for Freeways**

		Florida-Specific SPFs			<i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data			
Category	Severity	Coefficient		ODP	Coefficient		ODP	C
		$\alpha$	$\beta$		$\alpha$	$\beta$		
<b>Urban</b>								
4-Lane Basic Freeway Segments	Total	-9.372	1.086	0.633	-7.850	1.000	0.990	0.514
	FI	-10.745	1.144	0.565	-8.820	1.020	1.150	0.561
4-Lane Segments within Interchange Influence Area	Total	-11.656	1.302	0.355	-11.230	1.300	0.810	0.651
	FI	-12.143	1.281	0.310	-12.890	1.380	0.790	0.793
6-Lane Basic Freeway Segments	Total	-13.407	1.458	0.645	-5.960	0.780	0.480	1.372
	FI	-14.548	1.487	0.611	-7.600	0.850	0.540	1.504
6-Lane Segments within Interchange Influence Area	Total	-15.088	1.602	0.364	-11.250	1.280	0.600	0.926
	FI	-15.820	1.595	0.307	-13.620	1.420	0.550	0.844
(8+)-Lane Basic Freeway Segments	Total	-6.847	0.907	0.725	-16.240	1.670	0.450	0.939
	FI	-7.239	0.872	0.707	-19.160	1.850	0.520	0.918
(8+)-Lane Segments within Interchange Influence Area	Total	-5.430	0.791	0.520	-26.760	2.580	0.520	0.667
	FI	-7.544	0.903	0.445	-25.630	2.420	0.530	0.625
<b>Rural</b>								
4-Lane Basic Freeway Segments	Total	-11.412	1.238	0.233	-6.820	0.810	0.170	1.027
	FI	-11.024	1.145	0.213	-8.820	0.890	0.160	1.856
4-Lane Segments within Interchange Influence Area	Total	-10.572	1.184	0.312	-7.760	0.970	0.150	0.638
	FI	-10.467	1.119	0.244	-8.860	0.960	0.240	1.212
(6+)-Lane Basic Freeway Segments	Total	-11.522	1.234	0.231	-8.280	0.940	0.090	1.087
	FI	-13.991	1.379	0.168	-10.250	1.030	0.090	1.054
(6+)-Lane Segments within Interchange Influence Area	Total	-11.610	1.273	0.316	-9.630	1.060	0.210	1.741
	FI	-12.063	1.244	0.249	-10.480	1.040	0.200	1.934

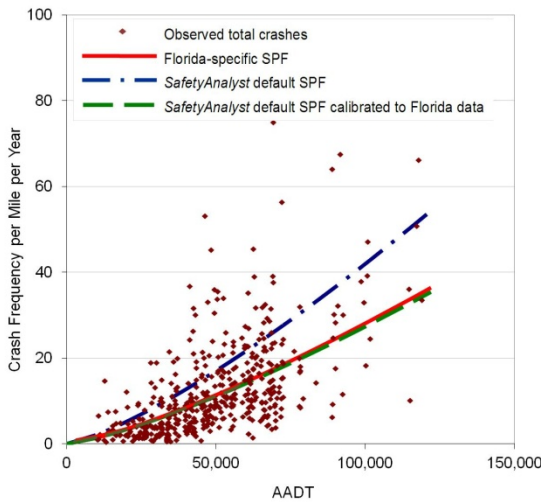
Figure 6-1 plots the SPFs for urban 4-lane freeways. The observation made from the plots is that the predicted crash frequency is overestimated by *SafetyAnalyst* default models. However, the plots of the default models calibrated to Florida data and Florida-specific SPFs share similarities for both basic segments and interchange areas. Therefore, either of the two models is recommended for urban 4-lane freeways.



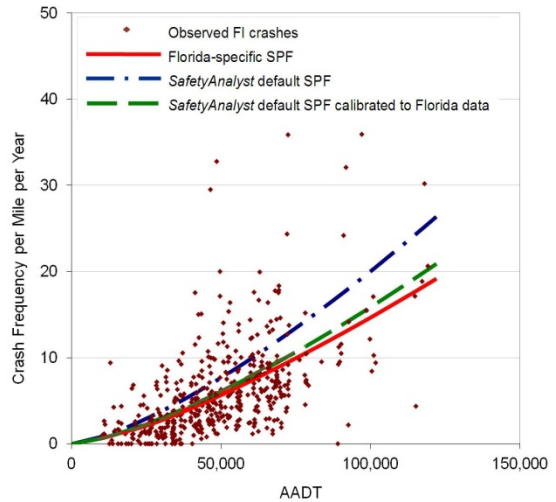
**(a) Total Crashes for Basic Segments**



**(b) FI Crashes for Basic Segments**



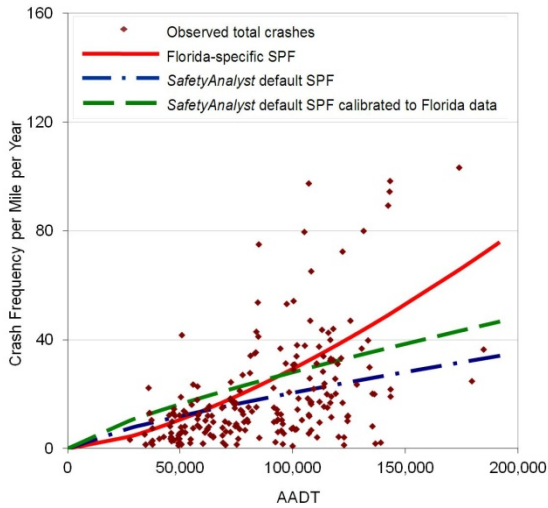
**(c) Total Crashes for Interchange Areas**



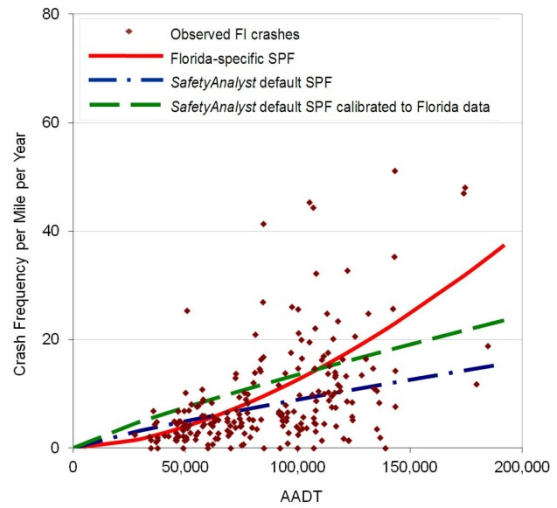
**(d) FI Crashes for Interchange Areas**

**Figure 6-1 Observed Crashes and SPFs for Urban 4-Lane Freeways**

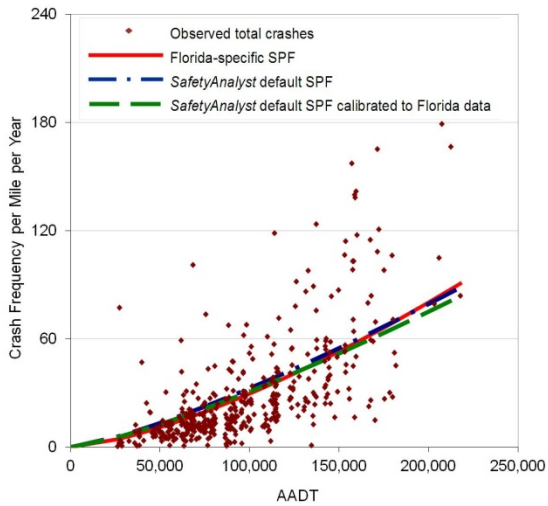
Figure 6-2 displays the SPFs for urban 6-lane freeways. The plots show that Florida-specific SPFs better represent the observed crash data for basic freeway segments, while the default models calibrated to Florida data and Florida-specific SPFs show the similarities for freeway segments within interchange areas.



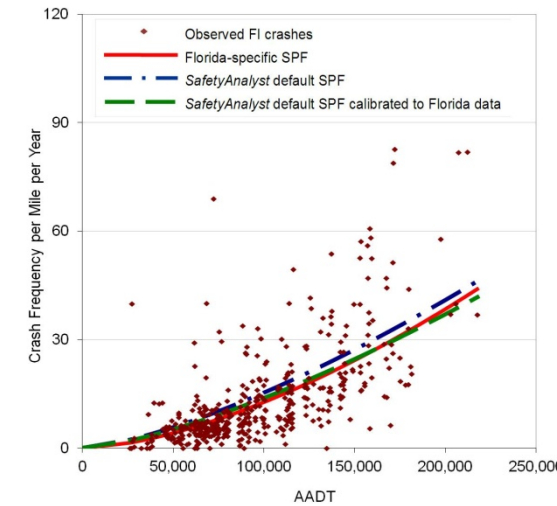
**(a) Total Crashes for Basic Segments**



**(b) FI Crashes for Basic Segments**



**(c) Total Crashes for Interchange Areas**

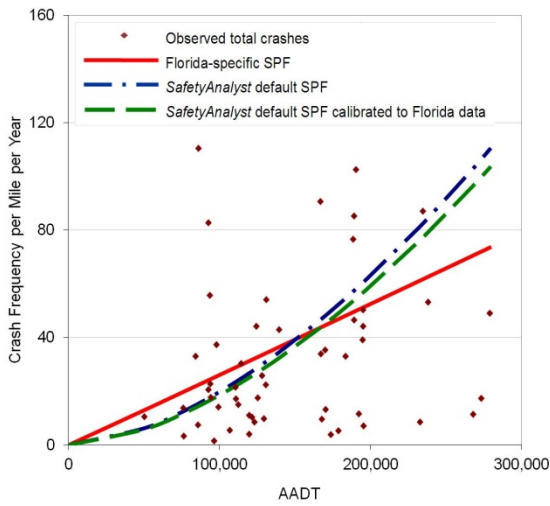


**(d) FI Crashes for Interchange Areas**

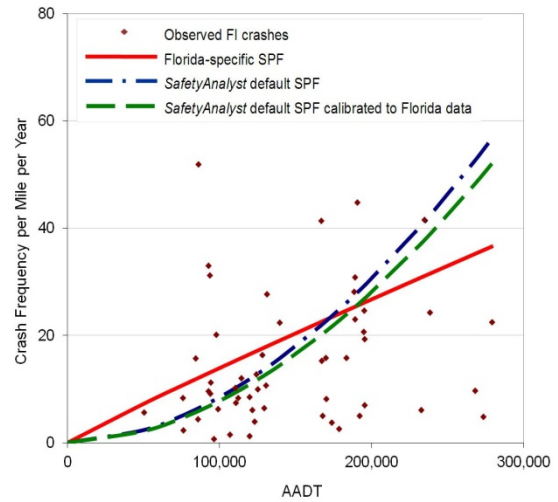
**Figure 6-2 Observed Crashes and SPFs for Urban 6-Lane Freeways**

Unlike the scenario for 4-lane and 6-lane freeways, the default models calibrated to Florida data and Florida-specific SPFs are different for (8+)-lane freeways in urban areas (both basic segments and interchange areas), as shown in Figure 6-3. The differences in total and FI crash trends for both Florida-specific SPFs and calibrated default SPFs might be due to the complexity of traffic characteristics at these facilities.

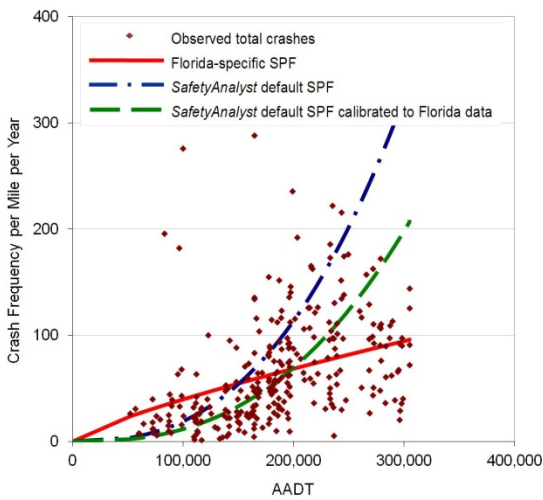
In urban areas, 8+ lane freeways (both basic segments and interchange areas), Florida-specific SPF with very low overdispersion parameters accurately represent observed crash data for both total and FI crashes. Therefore, Florida-specific SPFs is recommended for urban 8-lane freeways, as opposed to using default SPFs calibrated to Florida data.



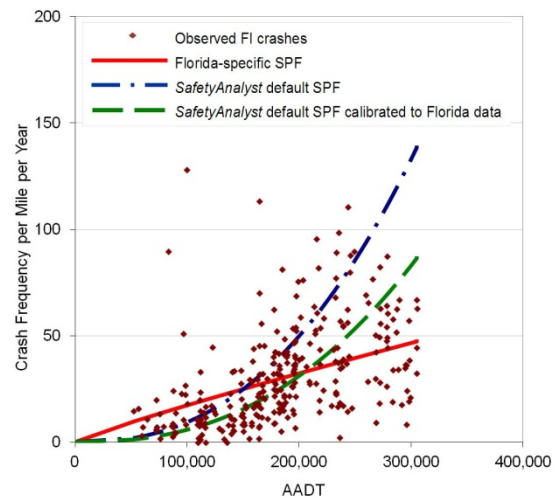
**(a) Total Crashes for Basic Segments**



**(b) FI Crashes for Basic Segments**



**(c) Total Crashes for Interchange Areas**



**(d) FI Crashes for Interchange Areas**

**Figure 6-3 Observed Crashes and SPFs for Urban (8+)-Lane Freeways**

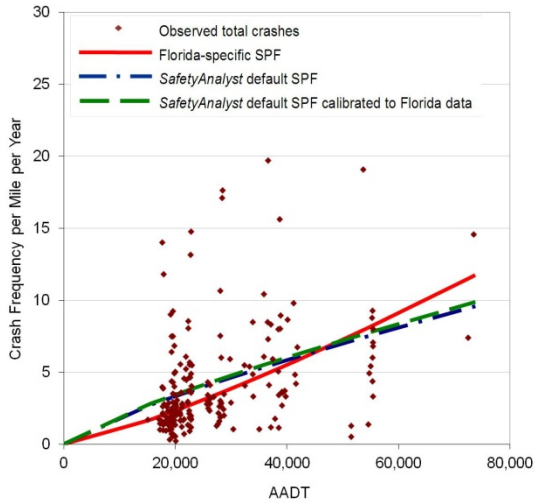


Furthermore, Figures 6-1, 6-2, and 6-3, displayed at the same level of AADT, show the predicted crash frequency for segments within interchange influence areas as higher than that for basic freeway segments. This is likely as a result of multiple conflict points due to the high tendency of weaving (merging and diverging) maneuvers within these interchange influence areas. These results point to the importance of considering the interchange influence area as a separate category instead of developing freeway SPFs, regardless of the influence of interchanges.

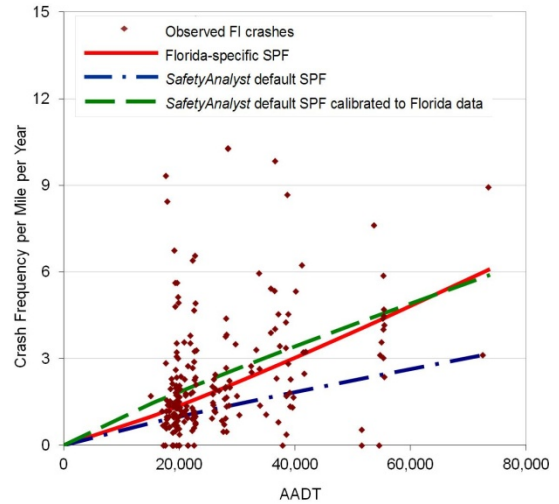
Figures 6-4 and 6-5 display the SPFs for rural freeways with 4 lanes and 6+ lanes, respectively. The curves of default models calibrated to Florida data and Florida-specific SPFs are well matched, especially for interchange areas. It can be concluded that, for these roadway types, there is no significant benefit in developing Florida-specific SPFs.

In order to identify the better-fitted model, statistical goodness-of-fit tests were performed on the validation data for both Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data, for comparison. Three goodness-of-fit tests represented by MAD, MSPE, and  $R^2_{FT}$  were used. The results of the comparison are given in Table 6-2. The highlighted boxes identify the models that performed more effectively for each test. It can be concluded from Table 6-2 that Florida-specific SPFs presented lower values for both MAD (other than urban (8+)-lane segments total crashes, rural (6+)-lane segments FI crashes, and rural (6+)-lane interchanges total crashes), and MSPE (other than urban 4-lane segments FI crashes, urban (8+)-lane segments FI crashes and rural 4-lane segments total crashes). The Florida-specific SPFs also presented higher  $R^2_{FT}$  values (other than urban 6-lane interchanges FI crashes, rural 4-lane segments FI crashes

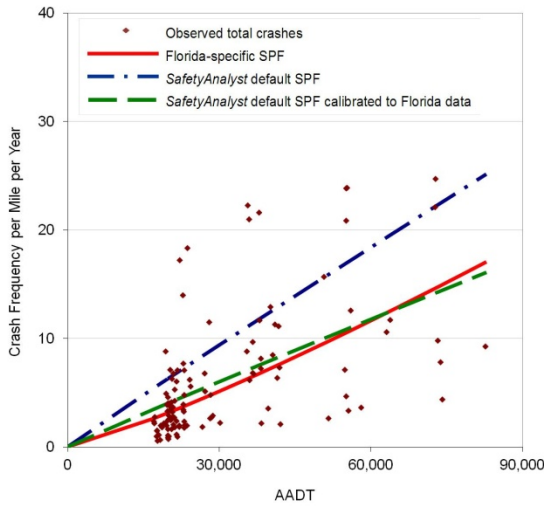
and rural (6+)-lane segments FI crashes) than the *SafetyAnalyst* default SPFs calibrated to Florida data. Based on the results, for freeways, Florida-specific SPFs yielded better prediction performance than the *SafetyAnalyst* default SPFs calibrated to Florida data.



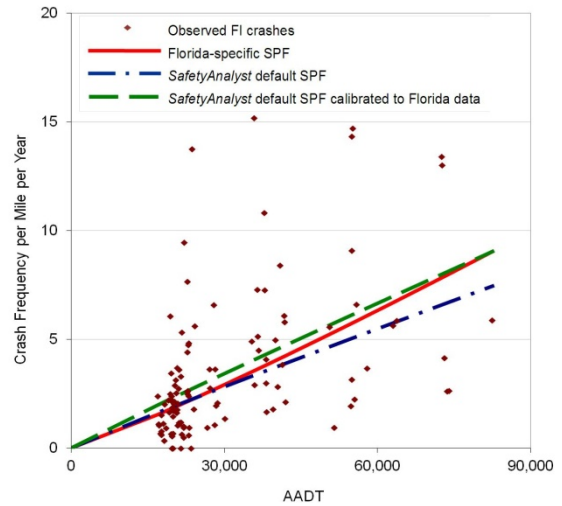
**(a) Total Crashes for Basic Segments**



**(b) FI Crashes for Basic Segments**

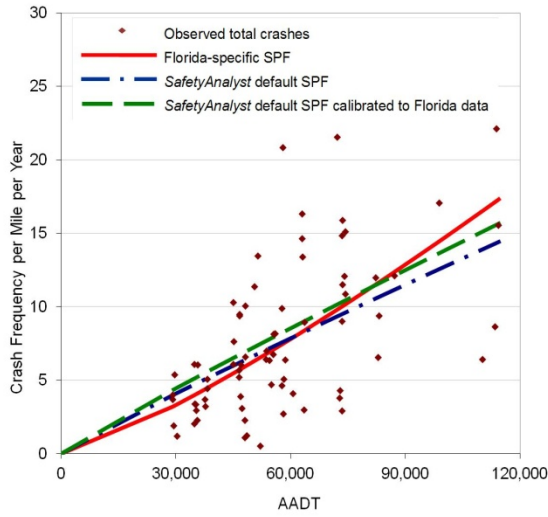


**(c) Total Crashes for Interchange Areas**

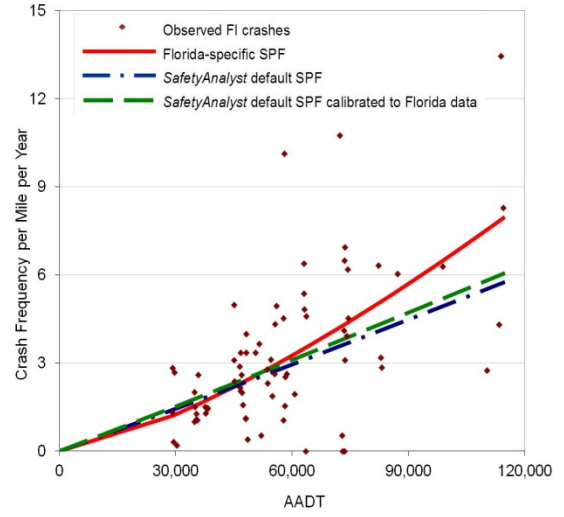


**(d) FI Crashes for Interchange Areas**

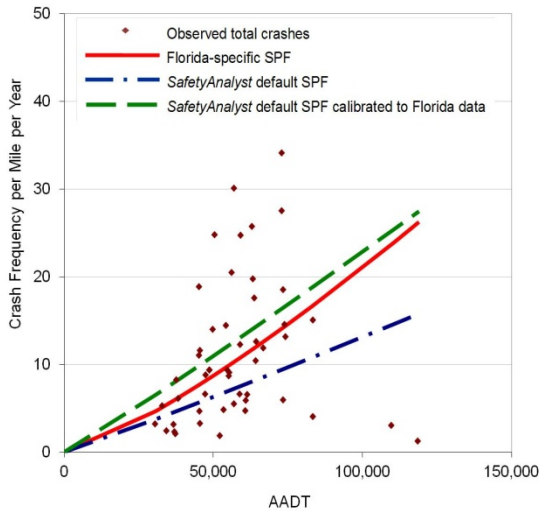
**Figure 6-4 Observed Crashes and SPFs for Rural 4-Lane Freeways**



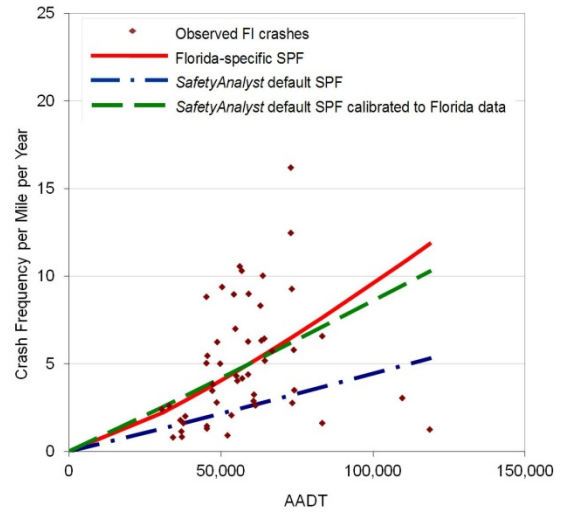
**(a) Total Crashes for Basic Segments**



**(b) FI Crashes for Basic Segments**



**(c) Total Crashes for Interchange Areas**



**(d) FI Crashes for Interchange Areas**

**Figure 6-5 Observed Crashes and SPFs for Rural (6+)-Lane Freeways**

**Table 6-2 Statistical Comparison for Freeways**

Category	Severity	Florida-Specific SPFs			<i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data		
		MAD	MSPE	R <sup>2</sup> <sub>FT</sub>	MAD	MSPE	R <sup>2</sup> <sub>FT</sub>
<b>Urban</b>							
4-Lane Basic Freeway Segments	Total	3.12	31.64	0.318	3.17	31.86	0.305
	FI	1.88	10.21	0.311	1.92	10.17	0.284
4-Lane Segments within Interchange Influence Area	Total	3.05	25.74	0.402	3.06	26.04	0.400
	FI	1.09	8.19	0.456	1.14	8.55	0.454
6-Lane Basic Freeway Segments	Total	3.06	25.90	0.339	3.10	26.32	0.304
	FI	1.05	8.12	0.324	1.09	8.35	0.235
6-Lane Segments within Interchange Influence Area	Total	3.75	49.82	0.486	3.87	49.88	0.445
	FI	2.19	13.25	0.444	2.23	13.67	0.452
(8+)-Lane Basic Freeway Segments	Total	6.35	94.91	0.052	6.30	101.42	0.039
	FI	2.88	19.90	0.114	2.94	19.83	0.092
(8+)-Lane Segments within Interchange Influence Area	Total	3.27	36.94	0.232	3.35	37.09	0.123
	FI	2.00	11.37	0.248	2.21	11.93	0.188
<b>Rural</b>							
4-Lane Basic Freeway Segments	Total	7.83	120.61	0.210	7.89	118.95	0.204
	FI	2.90	20.92	0.144	3.15	29.33	0.151
4-Lane Segments within Interchange Influence Area	Total	8.01	133.33	0.252	8.22	133.60	0.247
	FI	3.09	27.43	0.181	3.67	29.64	1.177
(6+)-Lane Basic Freeway Segments	Total	8.57	192.78	0.234	9.04	196.31	0.233
	FI	3.10	28.55	0.214	3.02	29.08	0.226
(6+)-Lane Segments within Interchange Influence Area	Total	9.33	257.48	0.162	9.27	257.58	0.134
	FI	3.18	32.85	0.203	4.05	33.71	0.203

Table 6-3 shows the two models (Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data) for each subtype of urban arterial streets and rural roads, for both total crashes and FI crashes, using the calibration data. As mentioned previously, the “ODP” represents the overdispersion parameter, and “C” represents the calibration factor.

The overdispersion parameter, which indicates the statistical reliability of the SPF, was used to account for dispersion in the data. The closer the overdispersion parameter is to zero, the more statistically reliable the SPF. For urban roadway segments, Florida-specific SPFs are statistically more reliable when the overdispersion values associated with Florida-specific SPFs are much lower than the corresponding default models. On the other hand, overdispersion values for some categories in rural roadway segments are slightly higher.

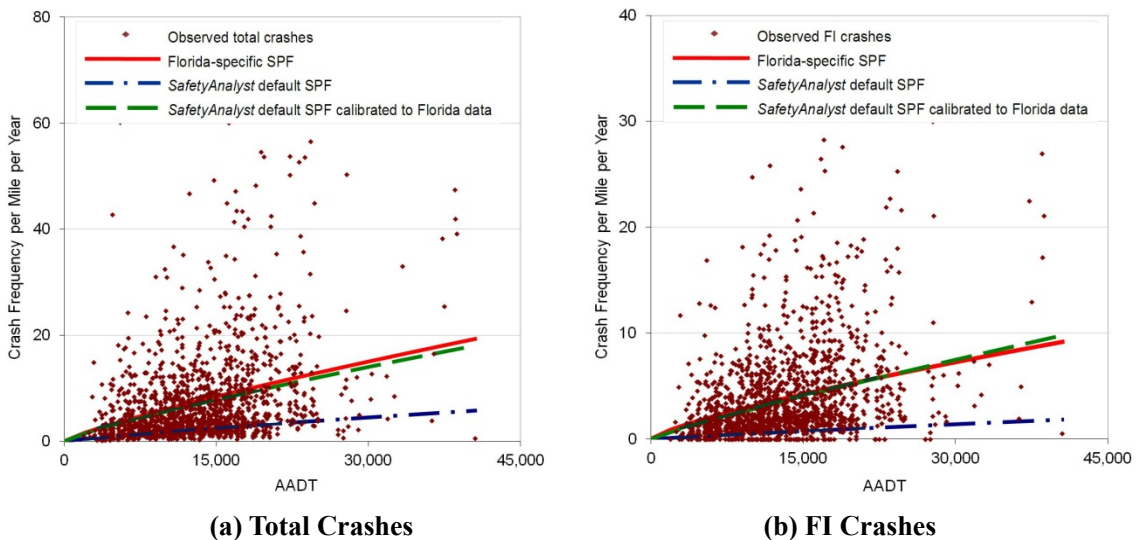
**Table 6-3 Florida-specific SPFs and *SafetyAnalyst* Default SPFs Calibrated to Florida Data for Urban Arterial Streets and Rural Roads**

		Florida-Specific SPFs			<i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data			
Category	Severity	Coefficient		ODP	Coefficient		ODP	CF
		$\alpha$	$\beta$		$\alpha$	$\beta$		
<b>Urban</b>								
2-lane Arterial Streets	Total	-5.480	0.788	0.757	-7.160	0.840	4.400	3.125
	FI	-6.035	0.778	0.677	-8.840	0.890	4.540	5.353
Multilane Undivided Arterial Streets	Total	-4.520	0.759	0.736	-10.240	1.290	0.850	1.477
	FI	-3.577	0.584	0.568	-12.070	1.390	0.810	1.495
Multilane Divided Arterial Streets	Total	-7.062	0.939	0.667	-11.850	1.340	5.910	1.786
	FI	-7.783	0.940	0.561	-14.870	1.520	5.810	2.462
One-way Arterial Streets	Total	-3.151	0.612	0.822	-3.530	0.600	1.380	1.347
	FI	-2.912	0.467	0.815	-5.150	0.650	1.450	1.718
<b>Rural</b>								
2-lane Roads	Total	-6.829	0.862	0.488	-3.630	0.530	0.500	0.926
	FI	-7.267	0.848	0.486	-4.860	0.530	0.670	1.830
Multilane Undivided Roads	Total	-	-	-	-3.170	0.490	0.530	-
	FI	-	-	-	-4.200	0.500	0.530	-
Multilane Divided Roads	Total	-5.213	0.673	0.461	-5.050	0.660	0.320	0.967
	FI	-5.799	0.671	0.428	-7.460	0.720	0.090	3.111

In addition, three SPFs (Florida-specific SPF, *SafetyAnalyst* default SPF, and *SafetyAnalyst* default SPF calibrated to Florida data) were plotted against the observed crash data for rural and urban roadway segments for both total and FI crashes. Figures 6-6 to 6-11 show the plots of the predicted annual crash frequency per mile against AADT and observed crash frequency for each subtype of urban arterial streets and rural

roads. As noted previously, the plotted data points are the observed crash frequency based on Florida data: the solid red line represents the Florida-specific SPF, the blue dash dot line represents the national default SPF used in *SafetyAnalyst*, and the green dash line represents the *SafetyAnalyst* default SPF calibrated to Florida data using a calibration factor.

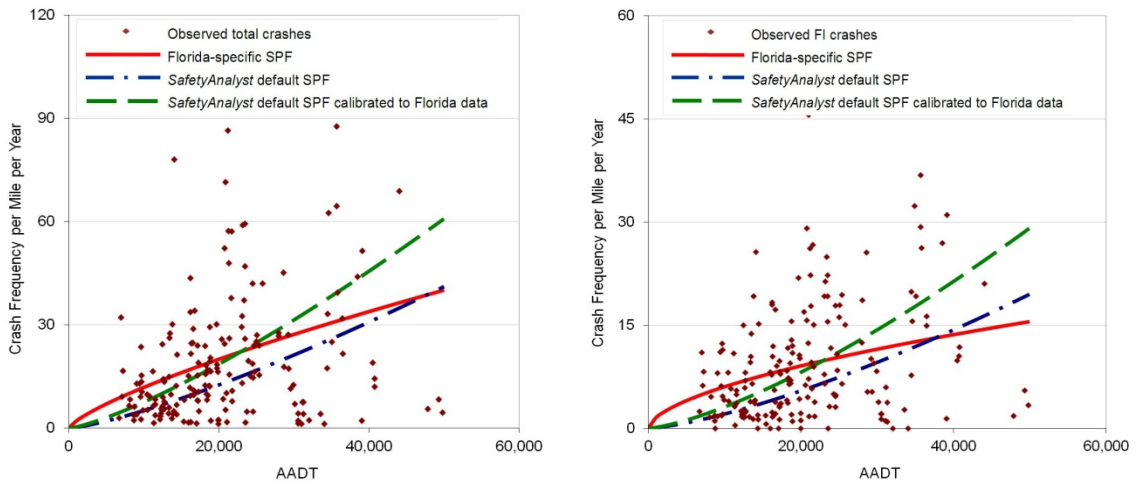
Figure 6-6 plots the SPFs for urban 2-lane arterial streets for both total crashes and FI crashes. The shapes of the default models calibrated to Florida data and Florida-specific SPFs are quite similar. Both models adequately represent observed crash data for both total and FI crashes. Predictions are underestimated by *SafetyAnalyst* default models.



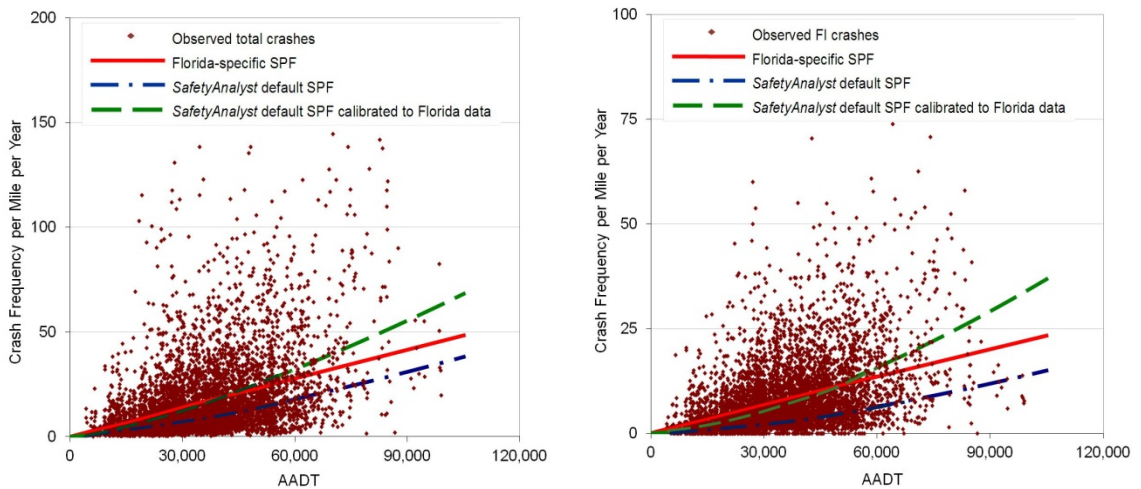
**Figure 6-6 Observed Crashes and SPFs for Urban 2-Lane Arterial Streets**

Figures 6-7 and 6-8 display the SPFs for urban multilane undivided and divided arterial streets, respectively. The observed crash data are well-represented by Florida-specific SPFs for both total and FI crashes. On the other hand, *SafetyAnalyst* default models underestimate the predicted crash frequency. *SafetyAnalyst* default

models calibrated to Florida data are closely matched with the Florida-specific SPFs; however, slight discrepancies remain. Therefore, the development of Florida-specific SPFs is recommended, as opposed to using default SPFs calibrated to Florida data, for both urban multilane undivided and divided arterial streets.



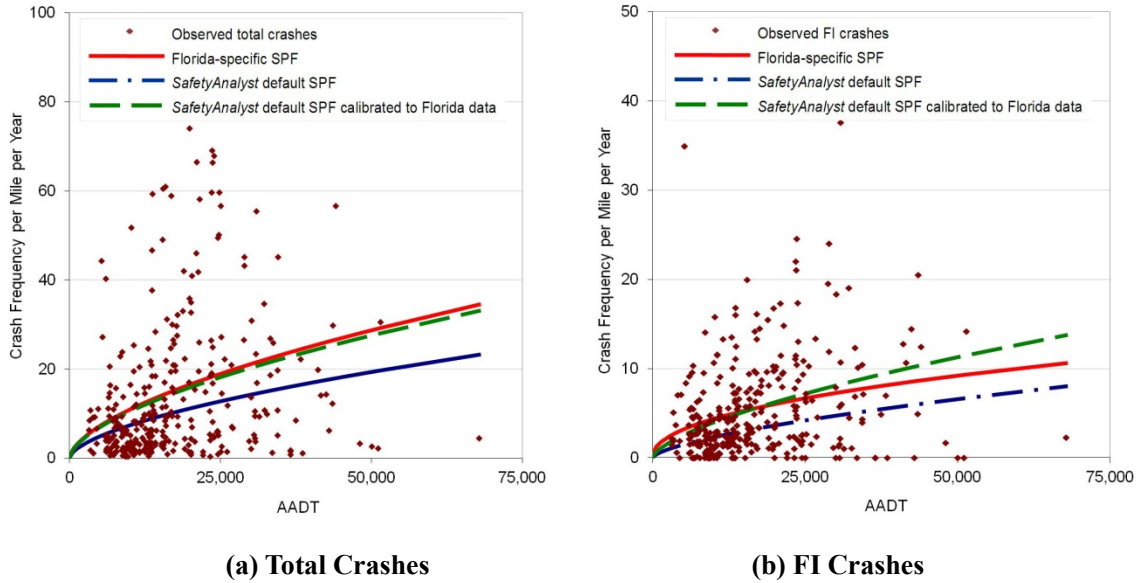
(a) Total Crashes (b) FI Crashes  
**Figure 6-7 Observed Crashes and SPFs for Urban Multilane Undivided Arterial Streets**



(a) Total Crashes (b) FI Crashes  
**Figure 6-8 Observed Crashes and SPFs for Urban Multilane Divided Arterial Streets**

Figure 6-9 shows that Florida-specific SPFs accurately represent urban one-way arterial streets, as well as observed crash data for both total and FI crashes. Similarities

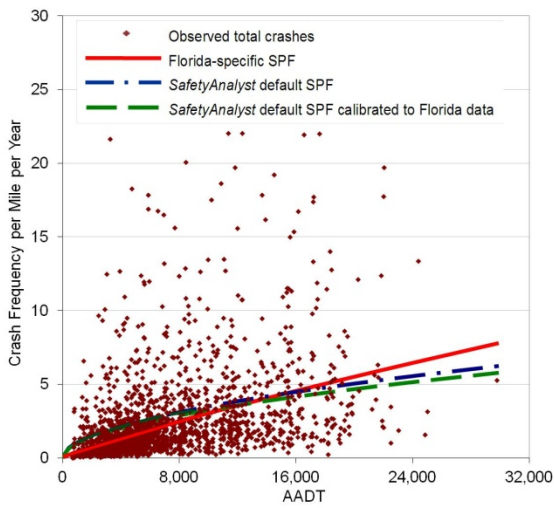
are found in the curves of default models calibrated to Florida data and calibrated Florida SPFs, primarily for total crashes. Results show that predictions are underestimated by *SafetyAnalyst* default models.



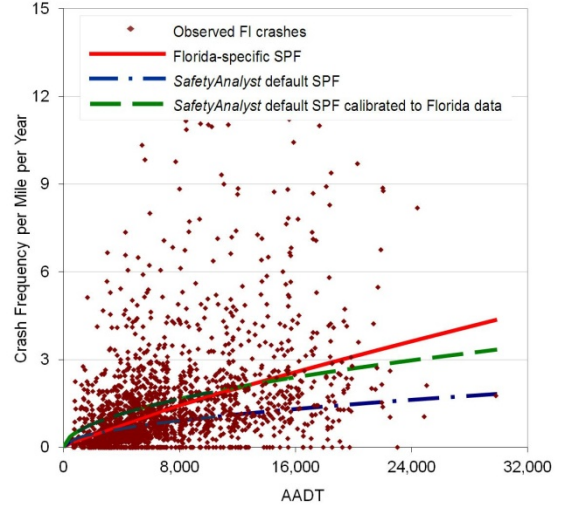
**Figure 6-9 Observed Crashes and SPFs for Urban One-Way Arterial Streets**

Figure 6-10 displays the SPFs for rural 2-lane roads for both total and FI crashes. As seen from the plots, it is difficult to say which model better represents the observed crash data. However, it is obvious that *SafetyAnalyst* default models underestimate the crash frequency of fatal and injury crashes. Figure 6-11 shows the SPFs for rural multilane divided roads. Figure 6-11 (a) shows that the observed crash data are better represented by all three models for total crashes. Similar to the SPFs for rural 2-lane roads, fatal and injury crashes are underestimated by the *SafetyAnalyst* default model.



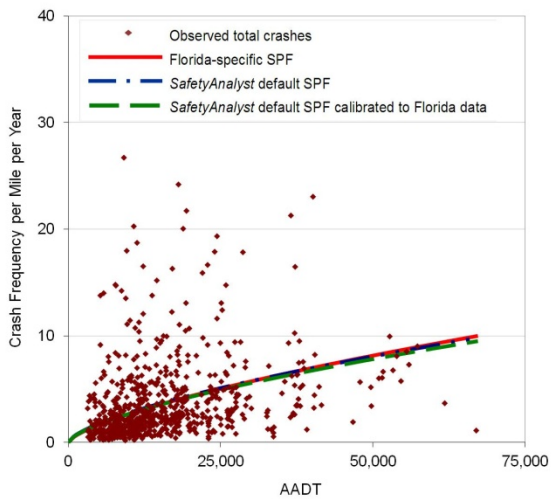


(a) Total Crashes

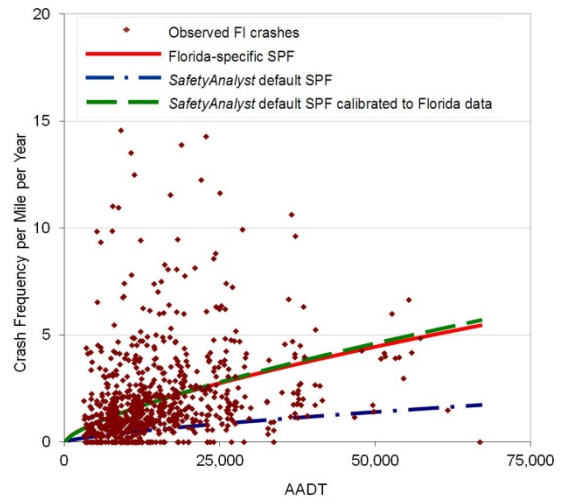


(b) FI Crashes

**Figure 6-10 Observed Crashes and SPFs for Rural 2-Lane Roads**



(a) Total Crashes



(b) FI Crashes

**Figure 6-11 Observed Crashes and SPFs for Rural Multilane Divided Roads**

The process used to identify the better-fitted model involved performing statistical goodness-of-fit tests on the validation data for both Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data, for comparison. Three goodness-of-fit tests were used, which were represented by MAD, MSPE, and  $R^2_{FT}$ . The results of the comparison are given in Table 6-4, of which the highlighted boxes identify

the models that performed more effectively for each test. In Table 6-4, for most of the subtypes, Florida-specific SPFs presented lower values for both MAD (other than urban one-way arterial streets total crashes) and MSPE (other than urban multilane undivided arterial streets total crashes and rural 2-lane roads FI crashes); while they presented higher value for  $R^2_{FT}$  (other than urban one-way arterial streets total and FI crashes, and rural 2-lane roads total crashes). Based on the results shown in Table 6-4, Florida-specific SPFs yielded better prediction performance than the national default SPFs calibrated to Florida data.

**Table 6-4 Statistical Comparison for Urban Arterial Streets and Rural Roads**

Category	Severity	Florida-Specific SPFs			<i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data		
		MAD	MSPE	$R^2_{FT}$	MAD	MSPE	$R^2_{FT}$
<b>Urban</b>							
2-lane Arterial Streets	Total	10.67	275.12	0.160	11.25	329.64	0.145
	FI	6.02	91.94	0.124	6.38	118.25	0.108
Multilane Undivided Arterial Streets	Total	7.11	121.05	0.130	7.89	109.67	0.082
	FI	3.05	30.41	0.119	4.02	31.30	0.053
Multilane Divided Arterial Streets	Total	5.24	57.47	0.295	5.33	57.56	0.279
	FI	2.07	18.92	0.220	2.67	19.07	0.173
One-way Arterial Streets	Total	7.31	123.44	0.094	7.25	131.12	0.102
	FI	3.23	35.16	0.107	3.55	37.28	0.110
<b>Rural</b>							
2-lane Roads	Total	6.35	102.55	0.288	6.80	104.49	0.293
	FI	2.90	25.09	0.230	2.93	23.75	0.218
Multilane Divided Roads	Total	6.14	94.36	0.244	6.51	95.33	0.237
	FI	2.55	23.18	0.207	3.07	25.59	0.188

### 6.3. Intersections

Florida-specific SPFs were developed for each of the four categories of signalized intersections by using calibration for both total and FI crashes. Table 6-5 shows the coefficients and corresponding overdispersion parameters for each subtype of the two models (Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data),

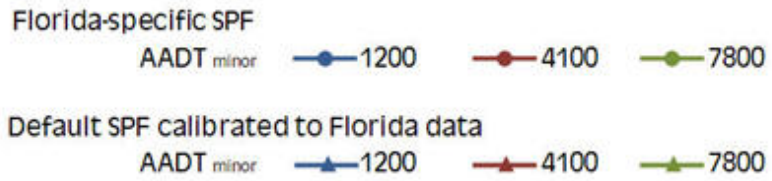
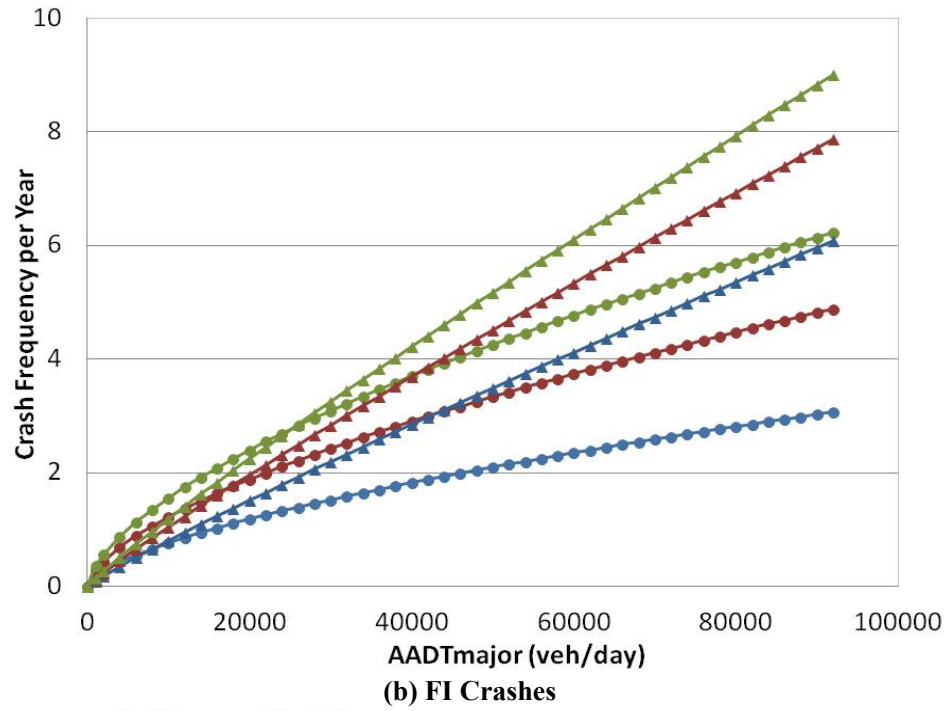
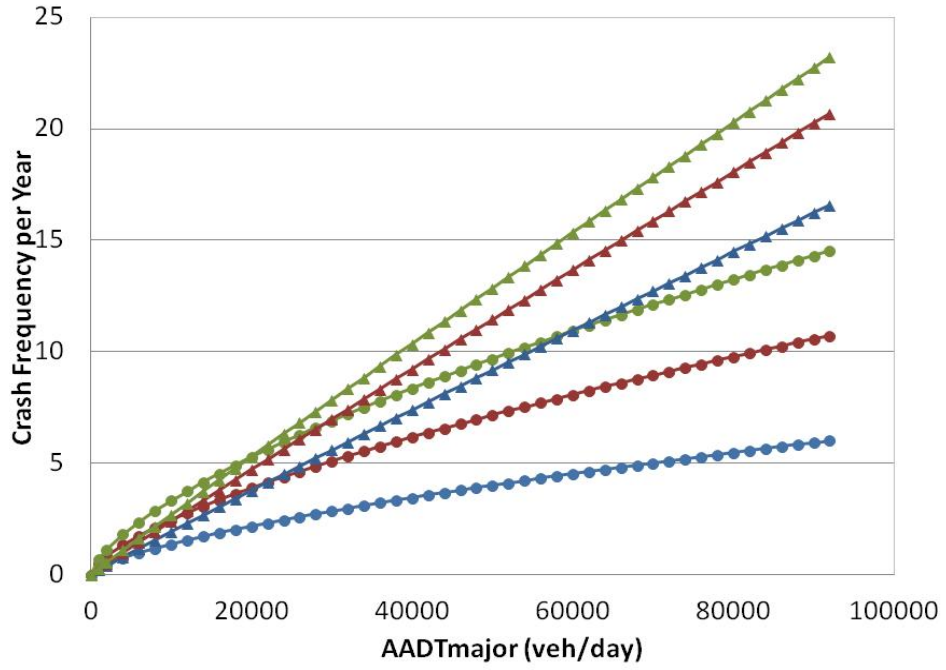
as well as the calibration factors of the default models. For signalized intersections, the overdispersion parameters of both default SPFs and the corresponding Florida-specific models are low and very similar.

**Table 6-5 Florida-specific SPFs and *SafetyAnalyst* Default SPFs Calibrated to Florida Data for Signalized Intersections**

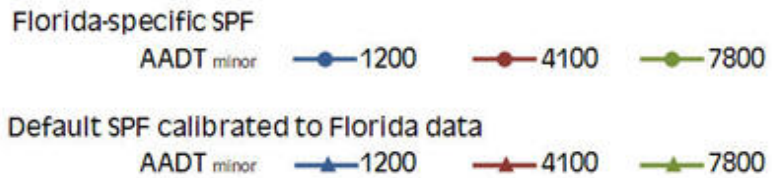
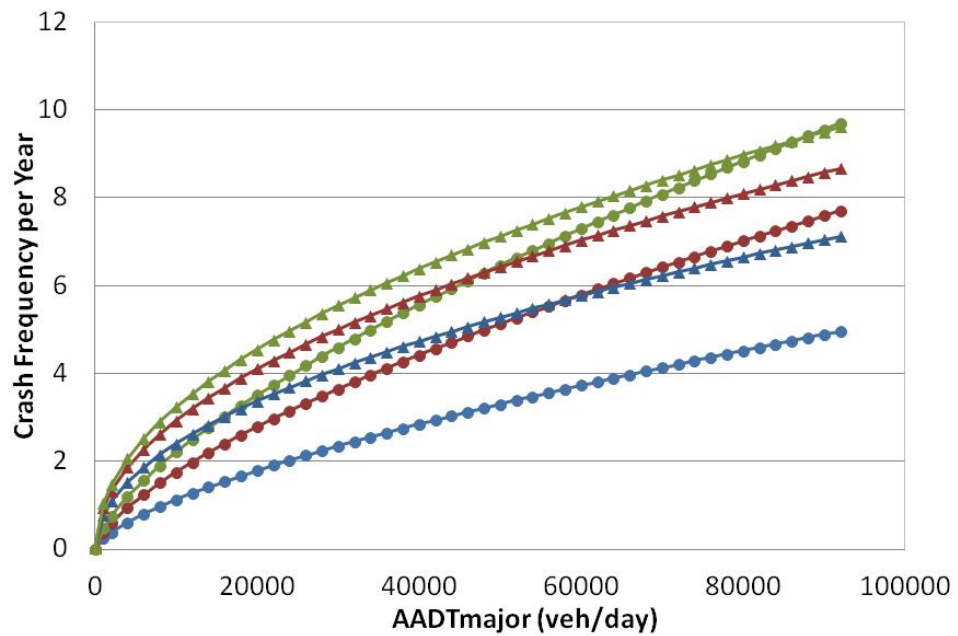
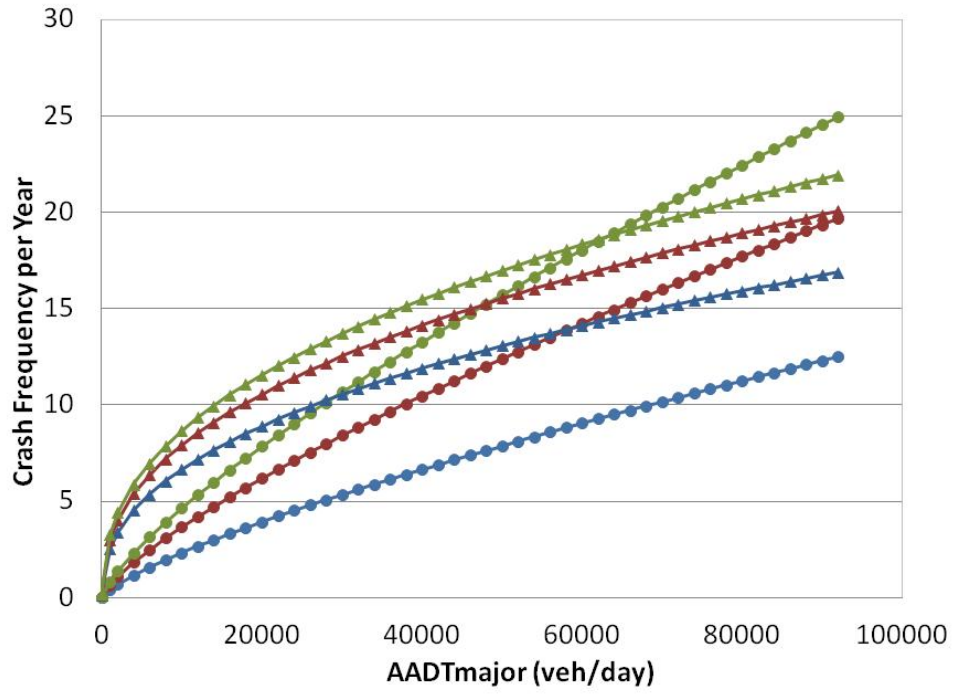
		Florida-Specific SPFs				<i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data				
Category	Severity	Coefficient			ODP	Coefficient			ODP	C
		$\alpha$	$\beta_1$	$\beta_2$		$\alpha$	$\beta_1$	$\beta_2$		
<b>Urban</b>										
3-Leg	Total	-9.589	0.725	0.453	0.404	-9.850	0.970	0.180	0.230	1.322
	FI	-8.354	0.605	0.360	0.310	-10.220	0.910	0.210	0.270	1.156
4-Leg	Total	-8.877	0.740	0.404	0.457	-3.470	0.420	0.140	0.320	1.628
	FI	-9.104	0.674	0.408	0.349	-5.110	0.490	0.160	0.300	1.404
<b>Rural</b>										
3-Leg	Total	-9.376	0.664	0.495	0.515	-6.570	0.660	0.200	0.330	0.941
	FI	-9.784	0.640	0.488	0.468	-7.830	0.750	0.140	0.500	1.197
4-Leg	Total	-7.143	0.488	0.481	0.440	-6.570	0.660	0.200	0.330	1.484
	FI	-6.770	0.351	0.505	0.436	-7.830	0.750	0.140	0.500	1.713

For each type of signalized intersection category, Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data were plotted for both total and FI crashes. Figures 6-12 to 6-15 plot the predicted annual crash frequency against AADT for major-road approaches.

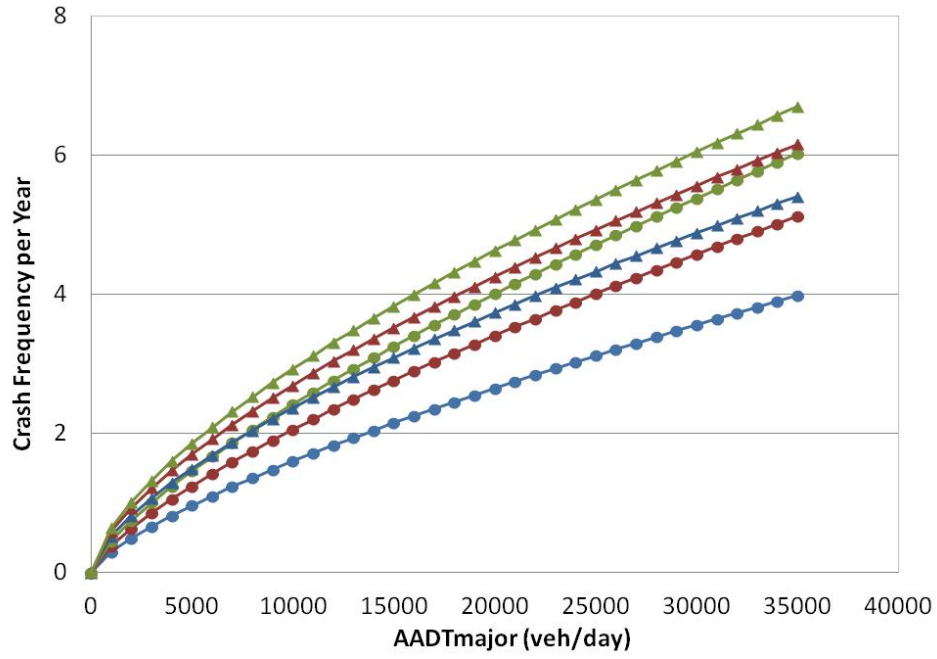
For signalized intersections, as indicated in the legend, lines with circles represent Florida-specific SPFs, and the lines with triangles represent default national SPFs calibrated to Florida data. The blue line represents SPFs assuming AADT for minor-road approaches to be 1200 veh per day. For rural intersections, the red and green lines represent SPFs assuming AADT for minor-road approaches to be 2300 and 3500 veh per day, respectively. For urban intersections, the red and green lines represent SPFs assuming AADT for minor-road approaches to be 4100 and 7800 veh per day, respectively.



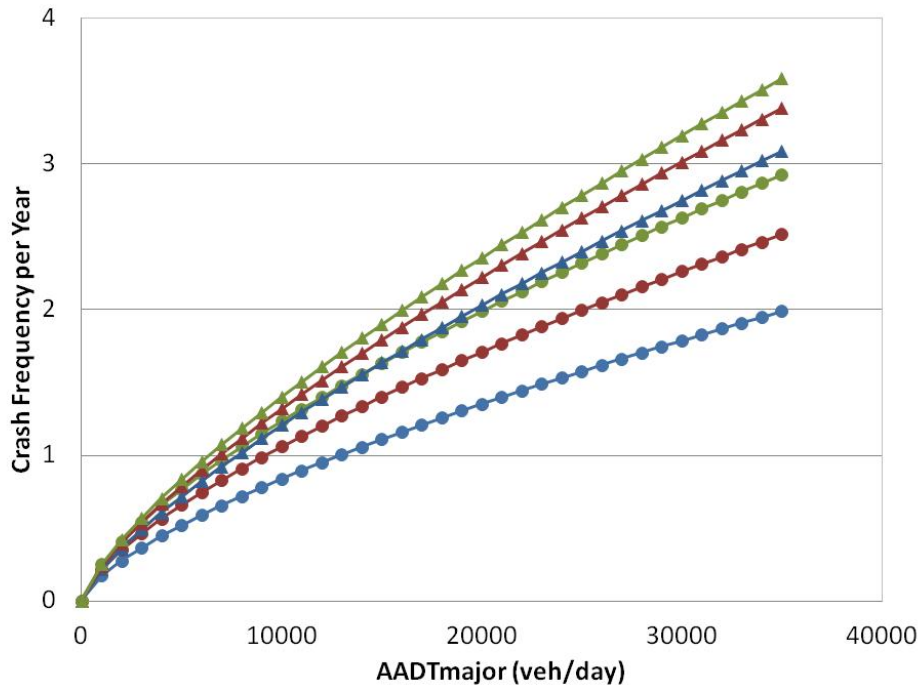
**Figure 6-12 SPFs for Urban Three-Leg Signalized Intersections**



**Figure 6-13 SPFs for Urban Four-Leg Signalized Intersections**



(a) Total Crashes



(b) FI Crashes

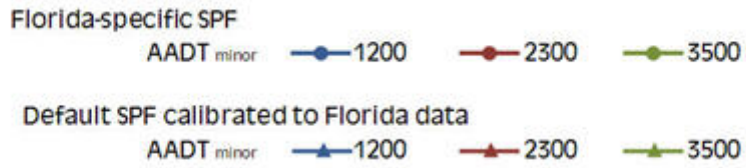
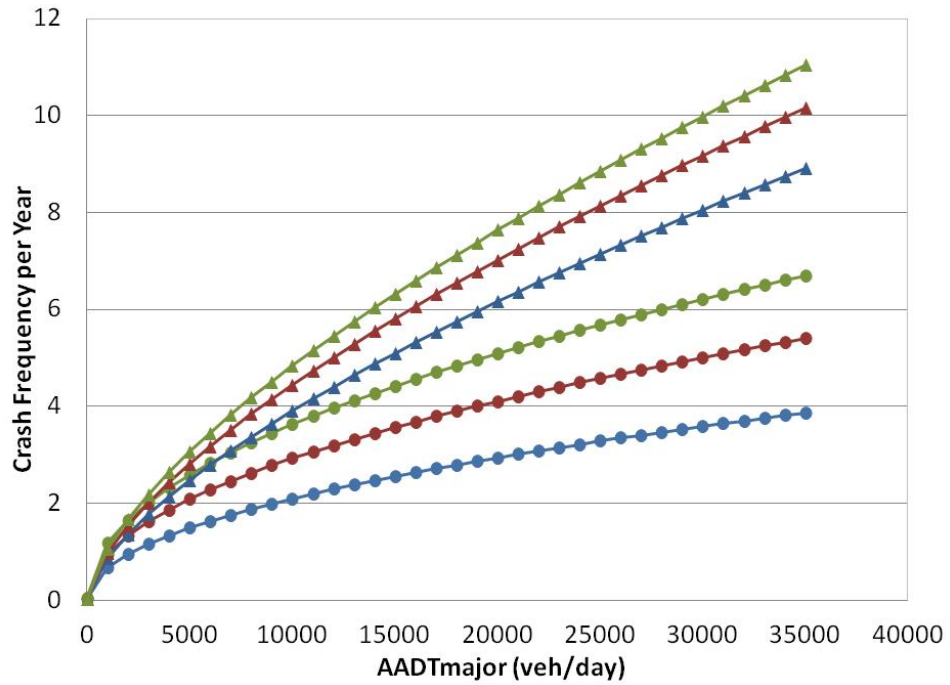
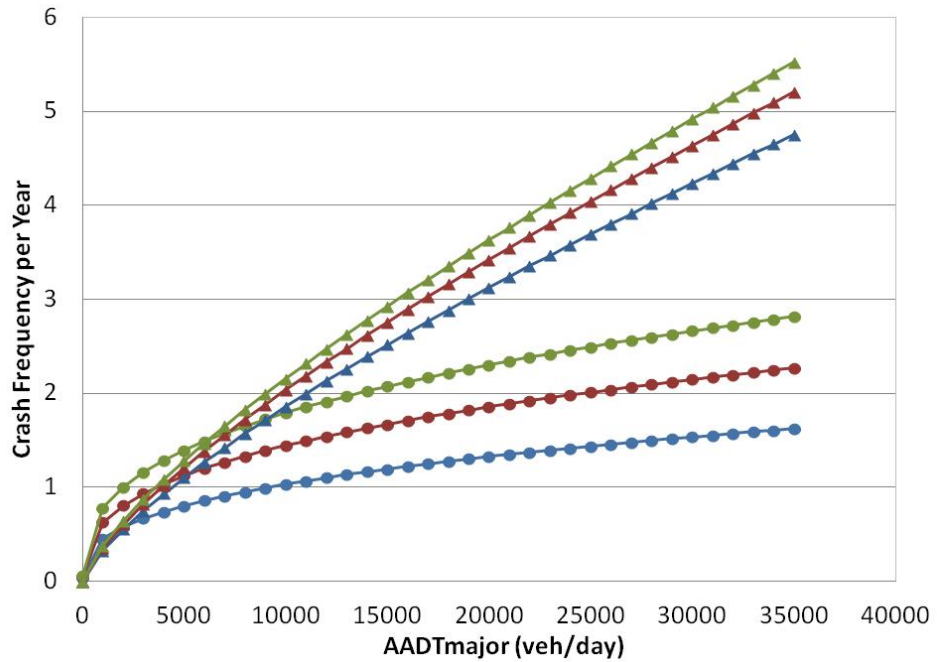


Figure 6-14 SPFs for Rural Three-Leg Signalized Intersections



(a) Total Crashes



(b) FI Crashes

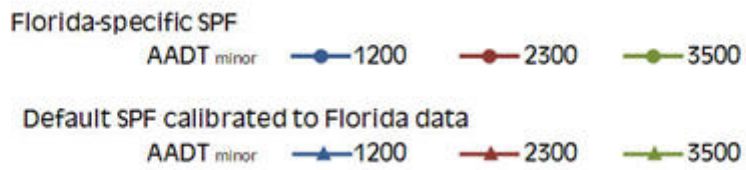


Figure 6-15 SPF for Rural Four-Leg Signalized Intersections

Figure 6-12 illustrates the Florida-specific SPFs and the *SafetyAnalyst* default SPF calibrated to Florida data and developed for total and FI crash frequency for urban three-leg signalized intersections. For both total crashes and FI crashes, the Florida-specific SPFs underestimate the predicted crash frequency when major AADTs are higher, while the predicted crash frequency is overestimated when major AADTs are lower.

Figure 6-13 presents the Florida-specific SPFs and the *SafetyAnalyst* default SPF calibrated to Florida data and developed for total and FI crash frequency for urban four-leg signalized intersections. For both total crashes and FI crashes, the Florida-specific SPFs overestimate the predicted crash frequency when both major and minor AADTs are higher, while the predicted crash frequency is underestimated when major AADTs are lower.

Figure 6-14 displays the Florida-specific SPFs and the *SafetyAnalyst* default SPF calibrated to Florida data and developed for total and FI crash frequency for rural three-leg signalized intersections. For both total crashes and FI crashes, the Florida-specific SPFs and the *SafetyAnalyst* default SPF calibrated to Florida data are closer and follow a very similar trend. The Florida-specific SPFs underestimate predicted crash frequency, compared to the default SPFs calibrated to Florida data.

Figure 6-15 shows the Florida-specific SPFs and the *SafetyAnalyst* default SPF calibrated to Florida data and developed for total and FI crash frequency for rural four-leg signalized intersections. The Florida-specific SPFs underestimate the crash frequency for total crashes, as well as for FI crashes when major AADTs are higher, while the crash frequency is overestimated for FI crashes with lower AADTs on major streets.



The better-fitted model was found by performing statistical goodness-of-fit tests on the validation data for both Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data, for comparison. Three goodness-of-fit tests represented by MAD, MSPE, and  $R^2_{FT}$  were used. The results of the comparison are given in Table 6-6, and the highlighted boxes identify the models that performed more effectively for each test. Florida-specific SPFs presented the lower values for MAD (other than rural 4-leg signalized intersections total crashes) and MSPE (other than rural 3-leg signalized intersections total and FI crashes). The Florida-specific SPFs also presented higher  $R^2_{FT}$  (rural 3-leg signalized intersections total crashes) than the *SafetyAnalyst* default SPFs calibrated to Florida data.

**Table 6-6 Statistical Comparison for Signalized Intersections**

Category	Severity	Florida-Specific SPFs			<i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data		
		MAD	MSPE	$R^2_{FT}$	MAD	MSPE	$R^2_{FT}$
<b>Urban</b>							
3-Leg	Total	3.13	30.04	0.414	3.26	30.23	0.342
	FI	1.06	8.73	0.362	1.30	8.75	0.315
4-Leg	Total	3.25	37.20	0.450	3.55	38.05	0.309
	FI	1.17	9.95	0.427	1.45	9.96	0.363
<b>Rural</b>							
3-Leg	Total	8.21	187.14	0.287	10.03	182.36	0.328
	FI	3.09	26.26	0.356	3.37	24.08	0.345
4-Leg	Total	5.61	54.25	0.643	5.60	55.72	0.592
	FI	2.18	13.69	0.583	2.39	13.84	0.540

It can be concluded that Florida-specific SPFs yielded better prediction performance than national default SPFs calibrated to Florida data for rural and urban signalized intersections, except for total crashes of rural 3-leg signalized intersections. This could be because of the fewer number of rural 3-leg sites used to develop Florida-specific SPFs. Thus, when there is a large sample size, Florida-specific SPFs are more accurately representative of Florida data than the national default SPFs

calibrated to Florida data. Therefore, developing Florida-specific SPFs, as opposed to using default SPFs calibrated to Florida data, is recommended for urban signalized intersections and rural signalized intersections with a larger sample size.

#### **6.4. Ramp SPFs**

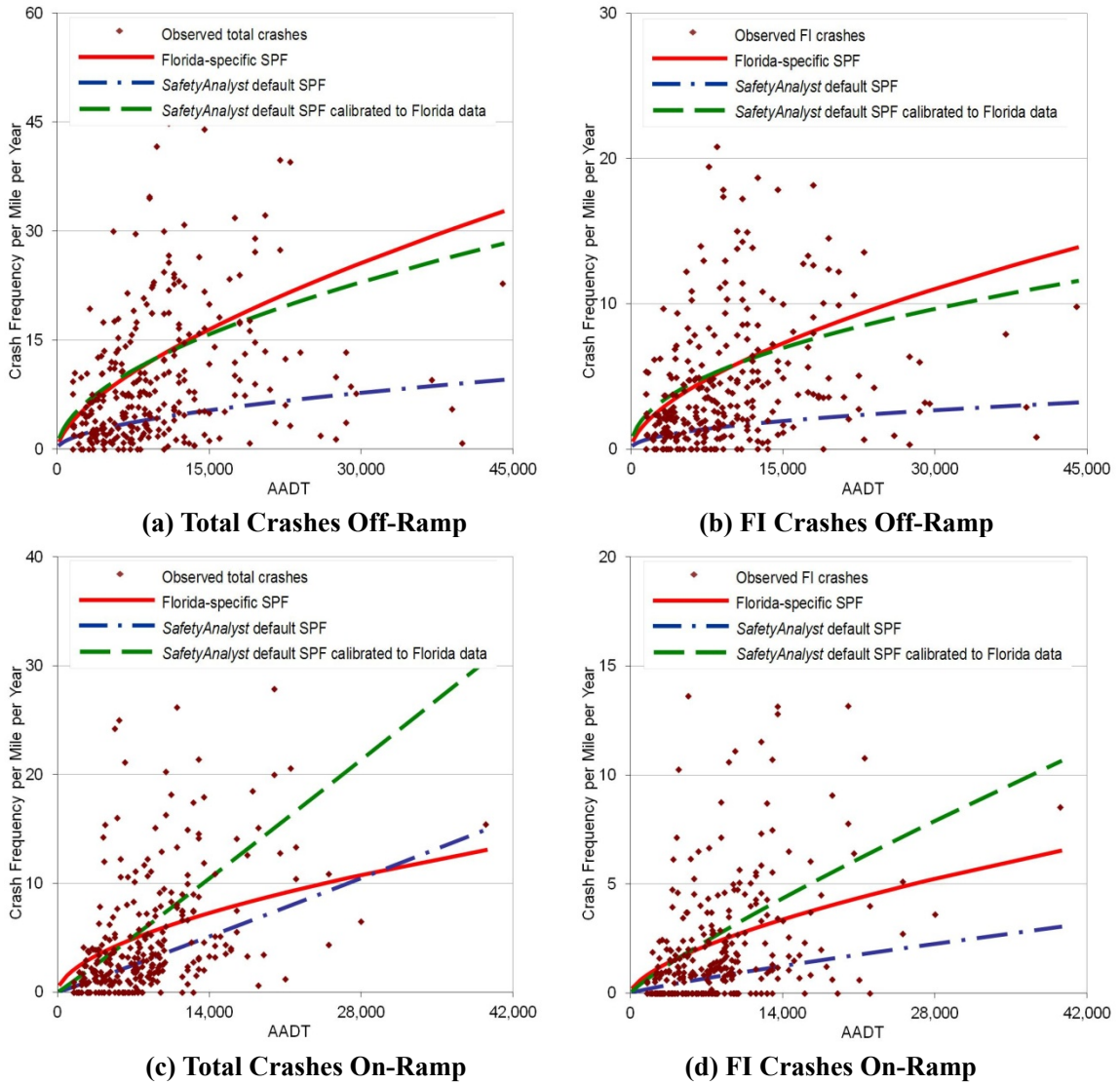
Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data were developed respectively for each of the four categories of ramps, for both total crashes and FI crashes, using the calibration data. The models are shown in Table 6-7. As mentioned earlier, the ramp classification in Florida is different from the default classification used in *SafetyAnalyst*. Therefore, the SPFs of only the following subtypes were compared: diamond ramps for both rural and urban areas, and urban partial cloverleaf ramps. The regression coefficients for urban direct connections are insignificant, while the sample size for urban trumpet ramps and rural partial cloverleaf ramps was insufficient; it did not accurately estimate regression coefficients.

In addition, three SPFs (Florida-specific SPF, *SafetyAnalyst* default SPF, and *SafetyAnalyst* default SPF calibrated to Florida data) were plotted against the observed crash data for rural and urban ramps, for both total and FI crashes. Figures 6-16 to 6-19 display the plots of the predicted annual crash frequency per mile against AADT and observed crash frequency for all the available ramp categories. As shown in these figures, the data points plotted are the observed annual crash frequency per mile, the solid red line symbolizes the Florida-specific SPF, the blue dash dot line plots the default SPF from *SafetyAnalyst* without applying the calibration factor, and the green dash line symbolizes the *SafetyAnalyst* default SPF calibrated to Florida data.

**Table 6-7 Florida-specific SPF and *SafetyAnalyst* Default SPFs Calibrated to Florida Data for Ramps**

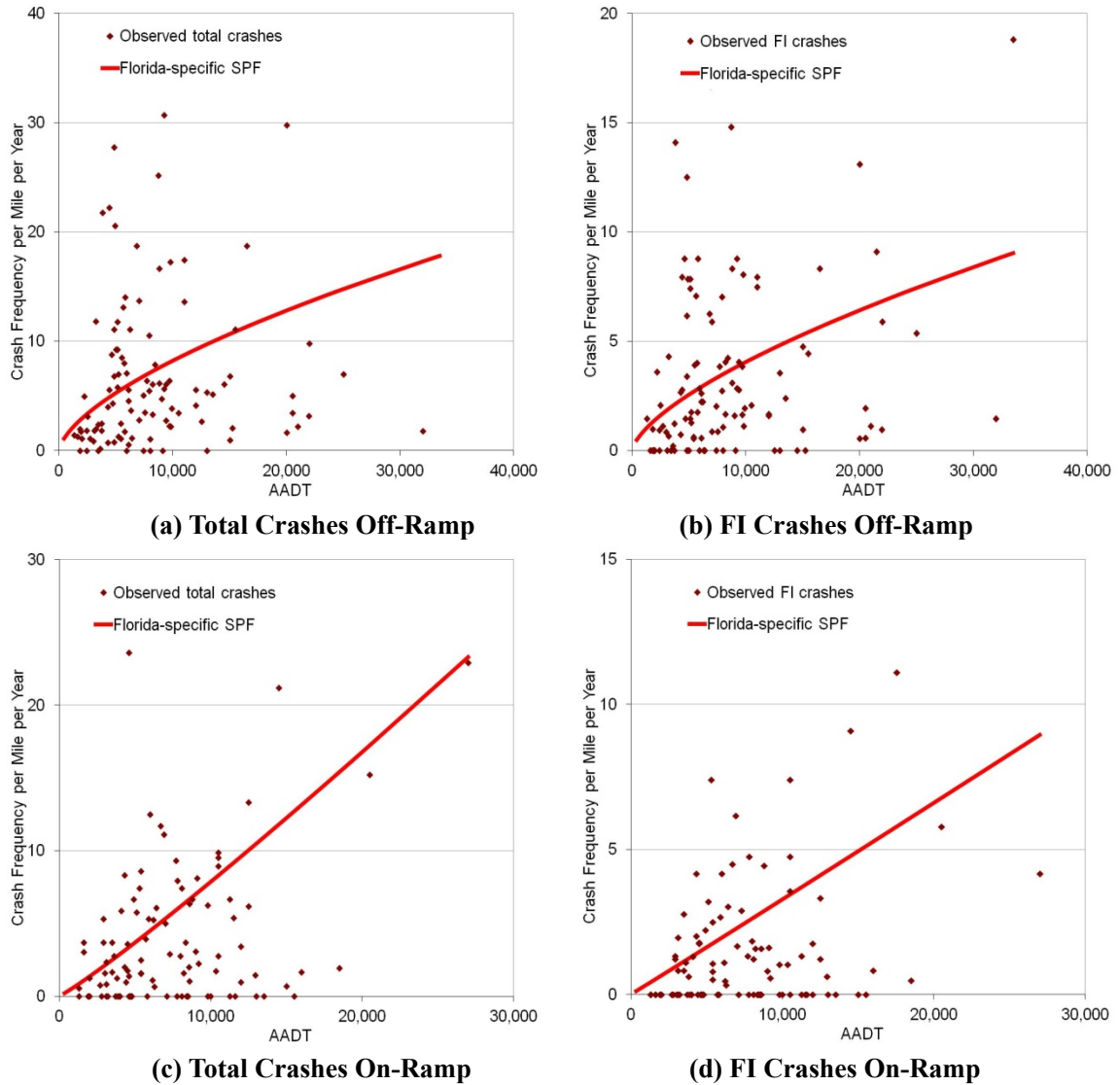
		Florida-Specific SPFs			<i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data			
Category	Severity	Coefficient		ODP	Coefficient		ODP	C
		$\alpha$	$\beta$		$\alpha$	$\beta$		
<b>Urban Diamond</b>								
Off-ramp	Total	-2.700	0.574	0.815	-3.52	0.54	1.15	2.895
	FI	-3.286	0.554	0.749	-3.86	0.47	1.94	3.461
On-ramp	Total	-3.419	0.561	0.971	-8.20	1.03	1.21	2.104
	FI	-4.402	0.582	1.063	-7.99	0.86	0.69	3.313
<b>Urban Partial Diamond</b>								
Off-ramp	Total	-3.768	0.633	1.002	---			
	FI	-4.105	0.594	0.987				
On-ramp	Total	-7.398	1.044	1.776				
	FI	-7.304	0.941	1.749				
<b>Urban Partial Cloverleaf (Parclo Loop)</b>								
Off-ramp	Total	-3.059	0.581	0.919	-1.15	0.26	0.12	2.516
	FI	-4.416	0.633	0.808	-3.68	0.53	0.67	1.197
On-ramp	Total	-5.223	0.774	0.718	-5.59	0.82	0.97	0.925
	FI	-6.453	0.832	0.770	-1.34	0.24	1.20	1.182
<b>Rural Diamond</b>								
Off-ramp	Total	-5.280	0.823	0.634	-3.07	0.46	1.34	2.088
	FI	-5.299	0.727	0.605	-4.54	0.47	2.66	3.526
On-ramp	Total	-8.287	1.103	0.419	-2.16	0.19	1.86	3.045
	FI	-10.069	1.247	0.362	-8.12	0.86	0.98	3.137

Figure 6-16 displays the SPFs for urban diamond ramps (both on-ramp and off-ramp) for both total and FI crashes. Crash data are well-presented by Florida-specific SPFs for urban diamond ramps when AADTs are lower. On the other hand, *SafetyAnalyst* default models underestimate the crash frequency. By adjusting *SafetyAnalyst* default models using the calibration factor, the curves become much closer to that of Florida-specific SPFs; however, slight discrepancies can still be found. Therefore, the development of Florida-specific SPFs, as opposed to using default SPFs calibrated to Florida data, is recommended for urban diamond ramps.



**Figure 6-16 Observed Crashes and SPFs for Urban Diamond Ramps**

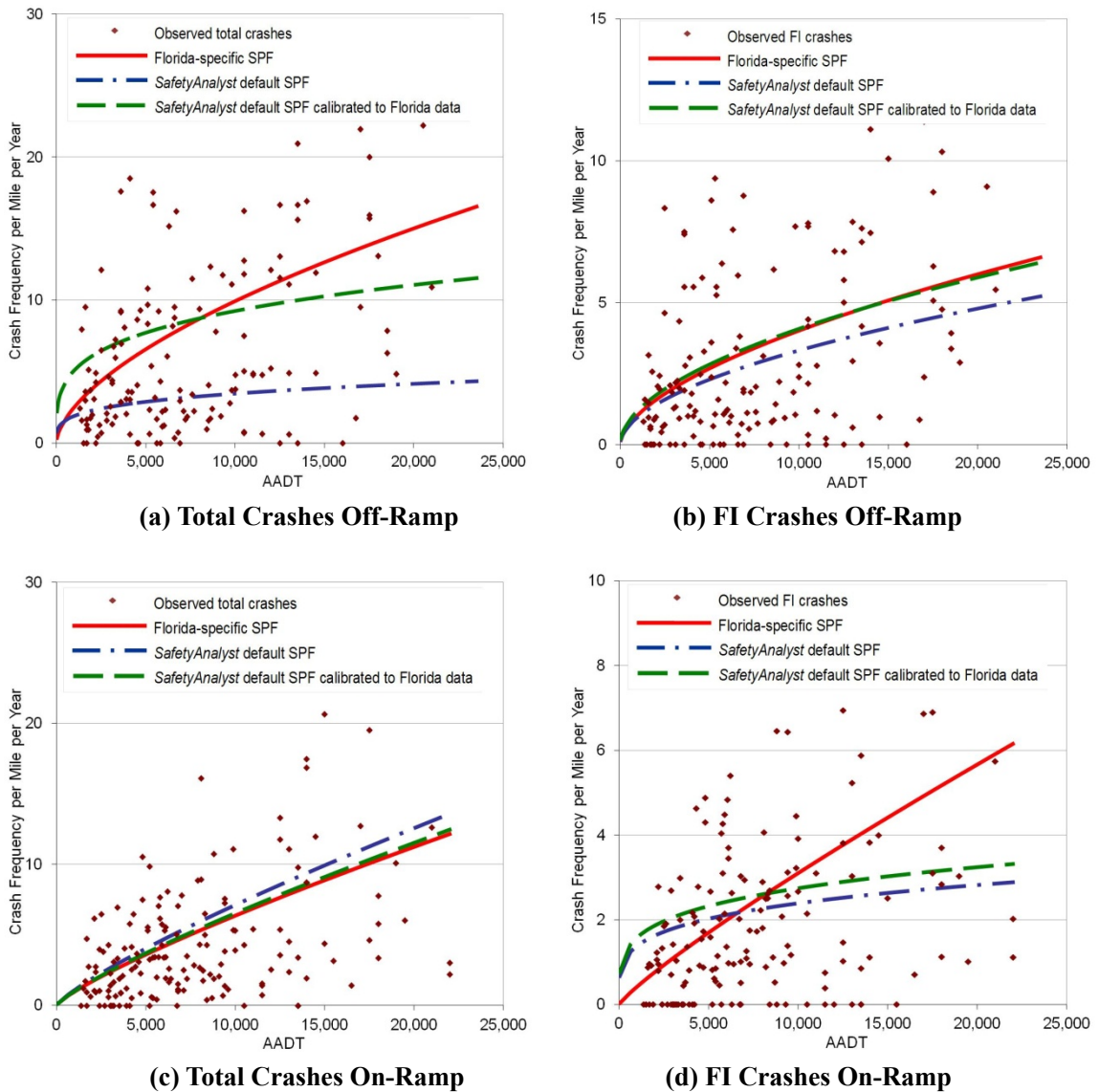
Figures 6-17 displays SPFs of urban partial diamond ramps. These categories are not found among the 16 subtypes used by *SafetyAnalyst*. Therefore, there are no *SafetyAnalyst* default SPFs to compare. Florida-specific SPFs can be used to analyze this category of ramps.



**Figure 6-17 Observed Crashes and SPFs for Urban Partial Diamond Ramps**

Figure 6-18 displays the SPFs for urban partial cloverleaf ramps (both on-ramp and off-ramp) for both total and FI crashes. The plots show that the total crashes for off-ramp are underestimated by the *SafetyAnalyst* default model. As for FI crashes on off-ramps and total crashes on on-ramps, the observed crash data are equally well-represented by both the Florida-specific SPFs and the adjusted default models. Florida-specific SPFs are recommended for FI crashes on on-ramps, due to the variation

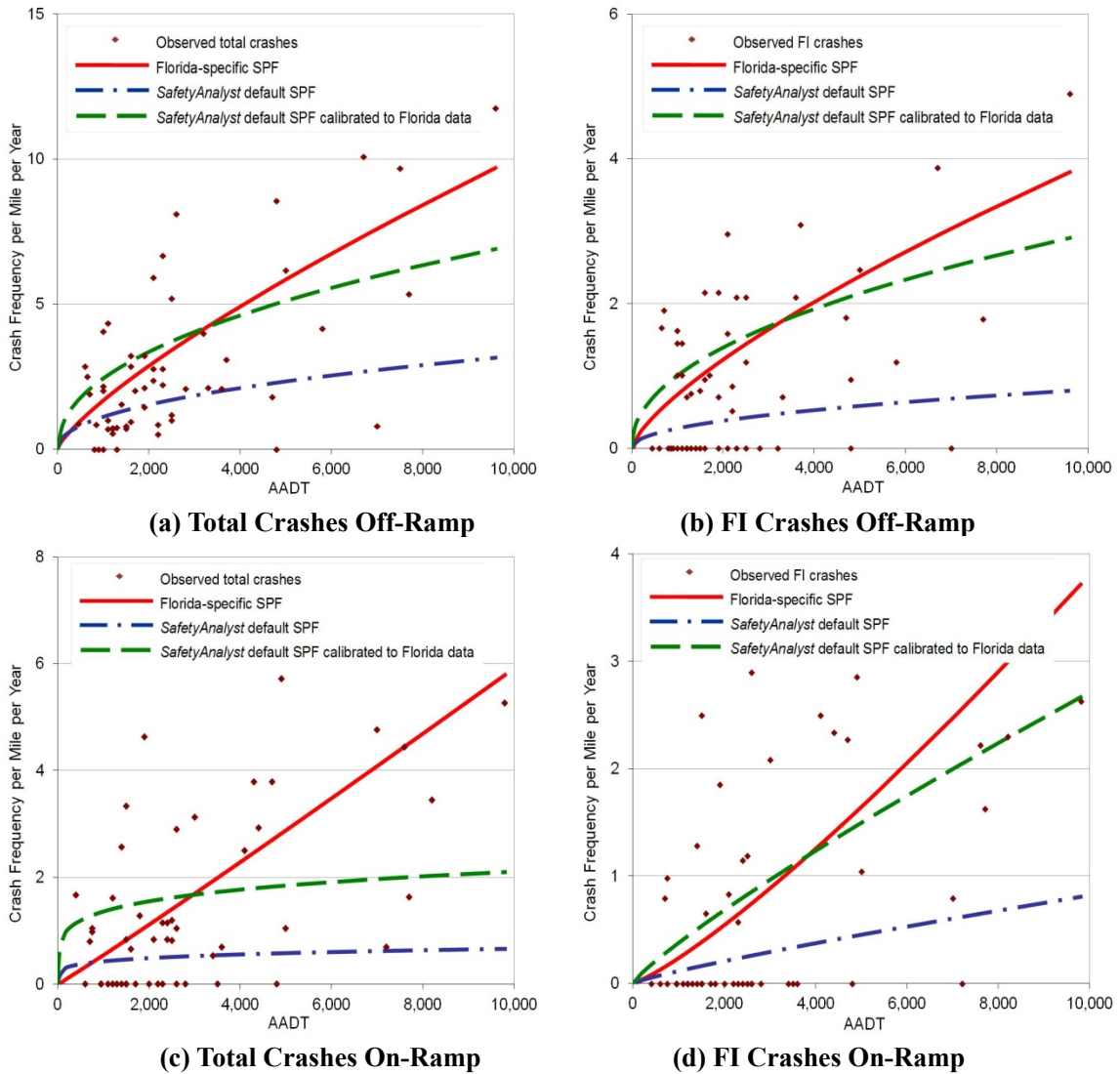
in the base shape of Florida-specific SPFs and the national default SPFs.



**Figure 6-18 Observed Crashes and SPFs for Urban Partial Cloverleaf Ramps**

Figure 6-19 displays the SPFs for rural diamond ramps (both on-ramp and off-ramp) for both total and FI crashes. For off-ramp (both total and FI crashes), the default models calibrated to Florida data and Florida-specific SPFs are better matched than that for on-ramp crashes. The on-ramp SPFs of default models calibrated to Florida data and Florida-specific SPFs are better matched for FI crashes than total crashes.

In addition, the *SafetyAnalyst* default models underestimate the crash frequency on rural diamond ramps.



**Figure 6-19 Observed Crashes and SPFs for Rural Diamond Ramps**

In order to identify the better-fitted model, statistical goodness-of-fit tests were performed on the validation data for both Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data for comparison. As mentioned earlier, the ramp classification in Florida is different from the default classification used in *SafetyAnalyst*. Therefore, the SPFs of only the following subtypes were compared: diamond ramps for

both rural and urban areas, and urban partial cloverleaf ramps. Three goodness-of-fit tests represented by MAD, MSPE, and  $R^2_{FT}$  were used. The results of the comparison are given in Table 6-8, of which the highlighted boxes identify the models that performed more effectively for each test. In Table 6-8, Florida-specific SPFs presented lower values for both MAD (other than urban and rural diamond on-ramp total crashes) and MSPE (other than urban diamond on-ramp total and FI crashes, urban partial cloverleaf off-ramp total crashes, and rural diamond on-ramp total crashes). The Florida-specific SPFs also presented higher  $R^2_{FT}$  values (other than urban diamond off-ramp total crashes). It can be concluded that Florida-specific SPFs yielded better prediction performance than the national default SPFs calibrated to Florida data for both rural and urban ramps. Even though the  $R^2_{FT}$  values of both models are low, the Florida-specific SPFs are better-fitted.

**Table 6-8 Statistical Comparison for Ramps**

Category	Severity	Florida-Specific SPFs			<i>SafetyAnalyst</i> Default SPFs Calibrated to Florida Data		
		MAD	MSPE	$R^2_{FT}$	MAD	MSPE	$R^2_{FT}$
<b>Urban Diamond</b>							
Off-ramp	Total	8.05	150.20	0.073	8.12	162.50	0.078
	FI	2.26	10.85	0.140	2.37	12.84	0.138
On-ramp	Total	8.91	178.54	0.134	8.88	176.76	0.105
	FI	3.33	24.17	0.166	3.76	22.52	0.084
<b>Urban Partial Cloverleaf (Parclo Loop)</b>							
Off-ramp	Total	5.83	44.92	0.249	6.22	44.84	0.164
	FI	1.37	8.53	0.233	1.58	9.02	0.217
On-ramp	Total	4.45	35.84	0.305	5.11	37.19	0.299
	FI	1.06	5.87	0.292	1.67	5.94	0.169
<b>Rural Diamond</b>							
Off-ramp	Total	7.62	131.01	0.281	8.34	135.27	0.210
	FI	1.85	9.02	0.236	1.87	11.78	0.185
On-ramp	Total	8.24	158.41	0.374	7.79	155.37	0.178
	FI	2.92	11.84	0.400	3.36	11.95	0.361



## 6.5. Summary

In this chapter, the model of *SafetyAnalyst* default SPFs calibrated to Florida data was first estimated by adjusting the national default SPFs via a set of calibration factors. The performance of the Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data were then compared.

The two models developed by using the calibration data were compared by overdispersion parameters and plots of SPFs against the observed crash data. Statistical goodness-of-fit tests were performed on the validation data for both models for comparison, specifically in order to identify the better-fitted model. Three goodness-of-fit tests represented by MAD, MSPE, and  $R^2_{FT}$  were used. In general, the values of MAD and MSPE for Florida-specific SPFs were lower than *SafetyAnalyst* default SPFs calibrated to Florida data, and the  $R^2_{FT}$  results for Florida-specific SPFs were higher than *SafetyAnalyst* default SPFs calibrated to Florida data. Overall, based on the comparison, Florida-specific SPFs more closely represent the Florida data than the national default SPFs calibrated to Florida data.

## CHAPTER 7

### MODEL APPLICATIONS

#### 7.1. Introduction

The roadway safety management process consists of two main applications of SPFs: crash prediction performance and identification of high crash locations. In this chapter, Florida-specific SPFs for all roadway subtypes developed in Chapter 5 were applied in the crash prediction performance and identification of high crash locations. However, Florida-specific SPFs were developed using the simple SPF models in *SafetyAnalyst*. The use of this traffic-only model in *SafetyAnalyst* may be questioned over the application of the traditional full SPFs for predicting crashes, as traffic might not be the sole predictor of crashes. Hence, this chapter compared the simple SPFs provided in *SafetyAnalyst* with full SPFs in both the crash prediction and high crash location identification applications to determine if the two models yielded similar performance of crash prediction and network screening.

#### 7.2. Application Overview

Florida-specific SPFs were developed using the simple SPF models in *SafetyAnalyst*. The models were developed, and traffic was used as the sole predictor of crashes without accounting for variations in geometric conditions. The SPF was a flow-only model calibrated by data from all sites irrespective of base conditions. The use of the simple SPFs in *SafetyAnalyst* may be questioned over the application of the traditional, full SPFs for predicting crashes, as traffic might not be the sole predictor of crashes. Hence, there is a need to examine whether the only-predictor, AADT, could

yield high prediction accuracy, when compared to the full SPFs, which consider multiple geometric and traffic predictors.

The comparison focuses on two applications: crash prediction performance and identification of high crash locations with the greatest potential for safety improvement. In order to compare the prediction performance, full and simple SPFs were estimated using data collected in Florida. Models are estimated for both total and FI crashes. The MAD and MSPE are used to assess and compare the prediction performance of the two models, and the variations in the rankings of HCLs using each model were examined.

### **7.3. Data and models**

The urban 4-lane freeway interchange influence areas in Florida were used as examples to describe the comparison of the performance of the full and simple SPFs. The freeway sections were divided into homogeneous segments of variable lengths, with the segment length defined by any change in the geometric or traffic characteristics. A total of 261 miles of freeway interchange areas were divided into 560 homogeneous segments. During the four-year study period (from 2007-2010), the 560 segments experienced 11,792 total crashes, of which 5,714 were FI crashes.

Four-fold cross-validation was chosen as the sampling method to divide the dataset into two groups: calibration and prediction. This sampling approach was previously used in safety studies (e.g., Haleem and Abdel-Aty, 2012). A simple explanation of the methodological approach of  $k$ -fold cross-validation is that the original dataset is randomly and equally divided into  $k$  subsets. Of the  $k$  subsets,  $(k-1)$  subsets are used as training data, and the remaining subset is used as validation data for testing the model. This process is then repeated  $k$  times, with each of the  $(k-1)$  subsets used

once as training data, and each of the  $k$  subsets used once as validation data. The  $k$  sets of results are then averaged to produce an entire estimate. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. In general,  $k$  is an unfixed parameter, and could be any number (Geisser, 1993).

In this research, 560 roadway segment records were randomly and equally divided into four subsets of 140 records each; as a result, four groups were created. In each group, three out of four subsets were combined as one dataset for calibration, and the remaining fourth subset was used for prediction. Thus, in each of the four different groups, 420 and 140 records were available for calibration and prediction, respectively. This process of calibration and prediction was then repeated four times. The average MAD and MSPE of all groups were then used to represent the mean prediction performance of the four groups.

After the segmentation and the sample recombination using the four-fold cross-validation approach, full and simple SPFs were developed. Analysis was performed on both total and FI crashes.

The following nine variables are believed to have an impact on the safety performance of full SPFs, and are measured in the modeling procedure:

- the natural logarithm of AADT
- lane width
- median width
- outside (right) shoulder width
- inside (left) shoulder width

- outside (right) shoulder type
- inside (left) shoulder type
- speed limit
- percentage of trucks

Table 7-1 shows the descriptive statistics of the abovementioned identified variables. Lane width was consistent across all segments, and hence was not included in the model development process. Shoulders were only paved on 4-lane freeways (either with or without warning devices). The warning device shoulder type is one that includes safety features, such as rumble strips to warn drivers encroaching onto the shoulders. Therefore, only two levels of paved shoulders were included while developing the full SPFs, where paved shoulders without warning devices were considered the baseline. Table 7-2 shows the full NB models for total crashes and FI crashes for each sample group. For example, the full SPF for total crashes in group 1 is:

$$N_{predicted} \text{ full SPF} = \exp(-10.401 + 1.188 \times \ln(AADT) + 0.025 \times \text{inside shoulder width} - 0.316 \times \text{outside shoulder type})$$

From Table 7-2, it is observed that there was no significant influence of median width, percentage of trucks, outside (right) shoulder width, and inside (left) shoulder type on crash frequency; hence, these variables were removed from the models. In every group, there is a significant increase in both total and FI crashes, with an increase in AADT, as more crashes always occur at high traffic volumes, as in congested situations. This is consistent with other studies (e.g., Pilko et al., 2007; Chen, et al., 2009). There is also an increase in crashes with the increase of the inside (left) shoulder width. The reason could be that wider inside shoulders allow drivers to park or stop on the inside

shoulder next to the fast lane (along a two-lane per direction freeway), which could negatively impact safety.

An interesting finding is that roadway segments with warning devices on the outside (right) shoulders are associated with less total and FI crashes, when compared to those without the warning device. As expected, an increased number of total and FI crashes occurred along the segments with higher speed limits. This result is supported by other studies (Vernon and Cook, 1999; Elvik, et al., 2004). It is noted that the speed limit variable is significant in only one out of four groups, for both total and FI crashes.

Simple SPFs were generated for urban 4-lane freeway interchanges. Table 7-3 shows the SPF models, along with their  $R^2_{FT}$  and overdispersion values for each group, for both total and FI crashes. For example, the simple SPF for FI crashes in group 4 is:

$$N_{predicted\ simple\ SPF} = exp(-11.886 + 1.253 \times \ln(AADT))$$

In Table 7-3, the overdispersion values are low for both total and FI crashes, indicating that the simple SPFs are reliable. Furthermore, the  $R^2_{FT}$  of the full SPFs shown in Table 7-2 are relatively similar to those of the simple SPFs shown in Table 7-3 for both total and FI crashes. In short, the two models have comparable  $R^2_{FT}$  values.

**Table 7-1 Descriptive Statistics of Explored Variables**

Variable	Descriptive Statistics	4-Fold Group			
		1	2	3	4
Number of total crashes		7630	9606	8894	9246
Number of fatal and injury crashes		3801	4713	4276	4433
Natural logarithm of AADT	Minimum value	9.148	8.476	8.476	8.476
	Maximum value	11.691	11.724	11.724	11.724
	Average value	10.752	10.805	10.774	10.761
	Standard deviation	0.448	0.480	0.496	0.457
Outside (Right) shoulder width (ft)	Minimum value	3	3	3	4
	Maximum value	20	25	25	25
	Average value	9.514	9.781	9.755	9.950
	Standard deviation	1.910	2.176	2.193	2.077
Inside (left) shoulder width (ft)	Minimum value	2	2	2	2
	Maximum value	20	17	20	20
	Average value	5.690	5.598	5.588	5.581
	Standard deviation	3.062	2.836	2.949	3.013
Lane width (ft)	Minimum value	11.5	11.5	11.5	12
	Maximum value	12.5	12.5	12.5	12.5
	Average value	12.005	12.001	12.004	12.005
	Standard deviation	0.069	0.055	0.065	0.049
Median width (ft)	Minimum value	6	6	6	8
	Maximum value	245	250	250	250
	Average value	56.064	60.202	61.310	61.731
	Standard deviation	28.153	33.639	34.207	31.946
Percentage of trucks	Minimum value	1.64	1.64	1.64	1.90
	Maximum value	34.86	34.44	34.86	34.86
	Average value	10.838	11.188	11.688	11.889
	Standard deviation	6.167	5.106	6.170	6.036
Speed limit	% of speed limit $\geq$ 65 mph	29.0	35.0	38.6	42.4
	% of speed limit $<$ 65 mph	71.0	65.0	61.4	57.6
Outside (Right) shoulder type	% of paved shoulder with warning device	71.7	74.8	74.8	79.5
	% of paved shoulder without warning device	28.3	25.2	25.2	20.5
Inside (left) shoulder type	% of paved shoulder with warning device	71.0	74.3	75.5	78.6
	% of paved shoulder without warning device	29.0	25.7	24.5	21.4

**Table 7-2 Full SPFs for Total and FI Crashes**

Variable	4-Fold Group							
	1		2		3		4	
	Estimate (Standard Error)	P-Value	Estimate (Standard Error)	P-Value	Estimate (Standard Error)	P-Value	Estimate (Standard Error)	P-Value
<b>Total Crashes</b>								
Intercept	-10.401 (0.892)	<0.0001	-10.390 (0.760)	<0.0001	-11.863 (0.756)	<0.0001	-11.278 (0.785)	<0.0001
Ln AADT	1.188 (0.082)	<0.0001	1.173 (0.069)	<0.0001	1.314 (0.070)	<0.0001	1.279 (0.072)	<0.0001
Inside (left) shoulder width	0.025 (0.012)	0.0299	0.044 (0.011)	<0.0001	0.032 (0.011)	0.0028	N/S	
Speed limit	N/S		0.181 (0.066)	0.0063	N/S		N/S	
			Baseline					
Outside (Right) shoulder type	-0.316 (0.077)	<0.0001	-0.261 (0.072)	0.0003	-0.261 (0.071)	0.0002	-0.210 (0.077)	0.0060
	Baseline		Baseline		Baseline		Baseline	
Dispersion ( <i>d</i> )	0.382 (0.031)		0.309 (0.026)		0.310 (0.026)		0.322 (0.027)	
R <sup>2</sup> <sub>FT</sub>	0.533		0.214		0.374		0.405	
<b>Fatal and Injury Crashes</b>								
Intercept	-11.824 (0.978)	<0.0001	-11.039 (0.862)	<0.0001	-12.669 (0.824)	<0.0001	-11.630 (0.846)	<0.0001
Ln AADT	1.264 (0.089)	<0.0001	1.174 (0.078)	<0.0001	1.322 (0.076)	<0.0001	1.239 (0.077)	<0.0001
Inside (left) shoulder width	N/S		0.030 (0.012)	0.0119	0.022 (0.011)	0.0373	N/S	
Speed limit	N/S		0.192 (0.069)	0.0053	N/S		N/S	
			Baseline					
Outside (Right) shoulder type	-0.246 (0.080)	0.0020	-0.244 (0.076)	0.0013	-0.187 (0.071)	0.0085	-0.135 (0.076)	0.0770
	Baseline		Baseline		Baseline		Baseline	
Dispersion ( <i>d</i> )	0.343 (0.035)		0.286 (0.029)		0.248 (0.027)		0.255 (0.028)	
R <sup>2</sup> <sub>FT</sub>	0.503		0.267		0.316		0.397	

\* N/S = Not significant

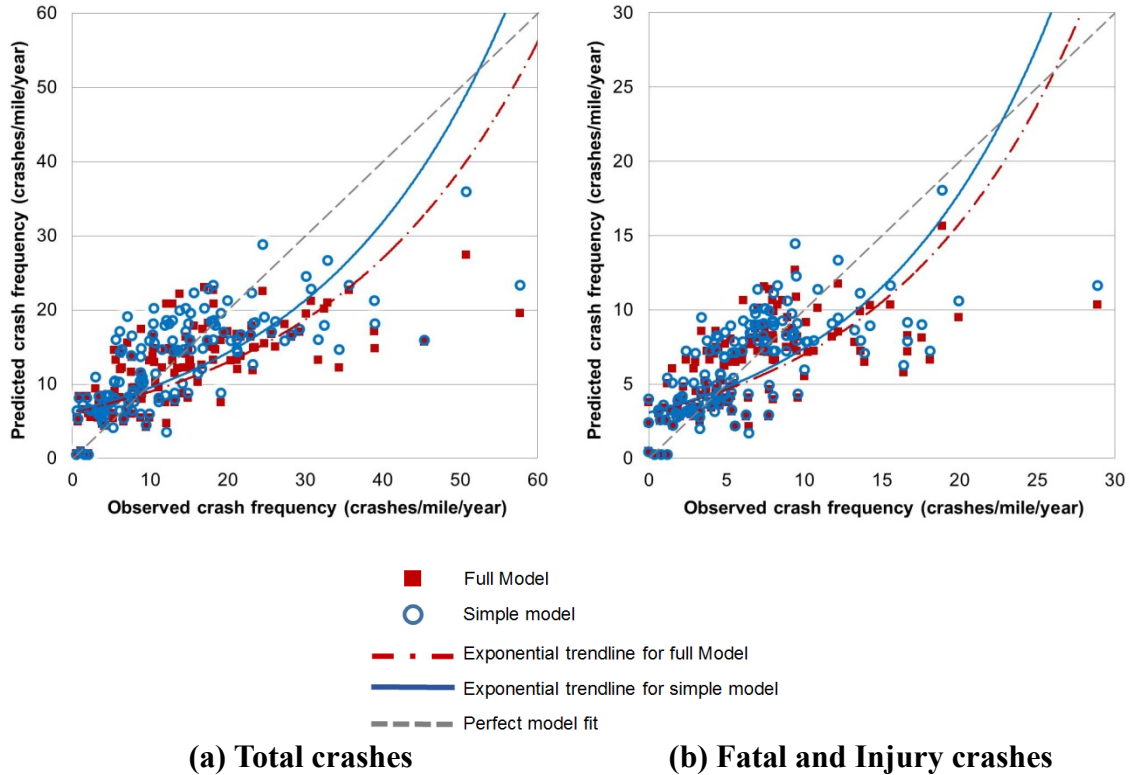


**Table 7-3 Simple SPFs for Total and FI Crashes**

4-Fold Group	Coefficient				ODP	R <sup>2</sup> <sub>FT</sub>
	$\alpha$		$\beta$			
	Estimate	P-Value	Estimate	P-Value		
<b>Total Crashes</b>						
1	-11.906	<0.0001	1.321	<0.0001	0.406	0.574
2	-11.123	<0.0001	1.253	<0.0001	0.340	0.307
3	-12.597	<0.0001	1.382	<0.0001	0.332	0.347
4	-11.621	<0.0001	1.296	<0.0001	0.328	0.381
<b>Fatal and Injury Crashes</b>						
1	-12.864	<0.0001	1.344	<0.0001	0.353	0.522
2	-11.678	<0.0001	1.238	<0.0001	0.309	0.364
3	-13.281	<0.0001	1.378	<0.0001	0.259	0.303
4	-11.886	<0.0001	1.253	<0.0001	0.257	0.382

#### 7.4. Crash Prediction

In this section, the prediction performance of the two models is compared. The plot of the observed versus predicted crashes for each model is generated. There are four sets of plots representing the different groups. Figure 7-1 shows one of the four sets (groups) as an example. As noted in the figure, the red squares represent crash frequency predicted from full models, while the blue circles represent crash frequency predicted from simple models. The red and blue lines in each plot are the exponential trend lines of the predicted crash frequency for full and simple models, respectively. The figures show that the trends and the predictions of the two models are similar.



**Figure 7-1 Observed vs. predicted crash frequency for full and simple models**

As previously discussed, four sets of MAD and MSPE were estimated from the four groups. Subsequently, the average MAD and MSPE of all groups were then used to represent the prediction performance. The MAD and MSPE values for the two models for both total and FI crashes are shown in Table 7-4. The table shows that the MAD and MSPE values for the two models for both total and FI crashes are quite similar.

**Table 7-4 Prediction of Full and Simple Models for Total and FI Crashes**

Prediction Performance Statistic	Total crashes		Fatal and Injury Crashes	
	Full Model	Simple Model	Full Model	Simple Model
<b>MAD</b>	6.35	6.27	3.15	3.13
<b>MSPE</b>	115.08	114.42	26.64	26.76

## 7.5. Identification of High Crash Locations

In this section, the EB method is applied separately using the two models to identify high crash locations for both total and FI crashes. The 560 segments, with 11,792 total crashes and 5,714 FI crashes, were analyzed. Using the EB method, the expected total and FI crash frequency for each of the 560 screening sites was estimated. The top 5% locations (28 sites) with the greatest PSI for both total and FI crashes were identified. Interpretations of different methods to identify HCLs were also included in the study by Qin et al. (2010).

Tables 7-5 and 7-6 show the ranking and HCLs identified using full and simple SPFs (with the EB approach) for total and FI crashes, respectively. In the two tables, the predicted crashes for the full models (column 6 in Tables 7-5 and 7-6) and for the simple models (column 10) were estimated using Equations (1-1) and (1-2), respectively. Furthermore, the expected crashes (columns 7 and 11 in Tables 7-5 and 7-6) were estimated using the EB approach. The PSIs in the tables were calculated as the difference between the expected and predicted crashes.

It can be seen from Tables 7-5 and 7-6 that among the top ten HCLs categorized by the full SPFs, all sites are almost similarly identified by the simple SPFs. The HCLs identified from both models are quite similar. However, a mere comparison of the top 10 or top 5% of HCLs identified using the two SPFs could be biased. Extremely unsafe sites (i.e., sites with very high observed crashes and high PSI), regardless of the model used, will top the list. In other words, some sites could be so unsafe that whichever model is chosen, they will be identified. This issue may weaken the comparison of the

two models. A statistical test was therefore applied to determine if there was a significant difference between the PSIs calculated using the full and the simple SPFs.

The differences between the two sets of PSIs for all of the 560 sites estimated from the two models were tested and found to nearly follow a normal distribution. The 198 sites with non-negative PSIs were compared. A paired t-Test with the following hypothesis was then performed on the resulting PSIs from the two models, as follows:

$H_0$ : there is no difference in means between the two sets of PSIs ( $\mu = \mu_0$ ),

$H_1$ : there are differences in means between the two sets of PSIs ( $\mu \neq \mu_0$ ).

Table 7-7 shows the results from the paired t-Test. In this table,  $H_0$  is accepted since  $t$  of 0.367 is less than the two-tail critical  $t$  of 1.972, and the p-value of 0.714 is more than the alpha of 0.05. Thus, there is sufficient evidence to support the conclusion that at a 5% significance level, there is no significant difference in the means between the two sets of PSIs. In other words, the two models produce the same performance in the network screening process.

**Table 7-5 Ranking of Top 28 HCLs for Total Crashes**

Roadway ID	Begin Mile Post	End Mile Post	Segment Length (miles)	Observed Crashes (crashes/mile/year)	Full Model				Simple Model			
					Predicted Crashes (crashes/mile/year)	Expected Crashes (crashes/mile/year)	PSI	Rank	Predicted Crashes (crashes/mile/year)	Expected Crashes (crashes/mile/year)	PSI	Rank
87470000	0.34	0.58	0.24	126.02	20.85	112.10	91.2	1	17.90	109.96	92.1	1
87005000	3.83	4.34	0.51	95.24	39.44	93.14	53.7	2	34.15	92.65	58.5	2
86471000	6.71	7.01	0.30	70.83	12.18	60.23	48.0	3	15.13	62.90	47.8	3
72040000	12.79	12.93	0.14	75.00	19.72	62.47	42.8	4	18.64	62.61	44.0	4
72270000	15.62	16.15	0.53	62.15	15.69	58.04	42.3	5	20.23	59.46	39.2	8
87005000	6.39	6.54	0.15	70.75	22.78	62.02	39.2	6	16.96	59.01	42.1	5
87075000	0.13	0.26	0.13	79.27	33.02	71.85	38.8	7	28.52	70.21	41.7	6
75008000	4.96	5.12	0.16	80.30	36.98	75.25	38.3	8	34.56	74.97	40.4	7
87471000	39.19	39.36	0.17	57.80	19.53	51.59	32.1	9	23.30	53.23	29.9	10
87260000	1.87	2.01	0.14	67.57	31.01	62.28	31.3	10	26.76	60.98	34.2	9
72090000	4.42	4.55	0.13	53.15	12.89	39.86	27.0	11	9.97	37.05	27.1	12
87005000	2.03	2.21	0.18	57.18	27.45	53.29	25.8	12	23.65	52.27	28.6	11
87021000	0.46	0.68	0.22	45.25	13.27	38.56	25.3	13	11.33	37.33	26.0	13
93470000	32.81	32.97	0.16	45.45	15.68	39.44	23.8	14	15.89	39.83	23.9	14
10075000	30.21	31.95	1.74	50.86	27.42	50.55	23.1	15	35.92	50.72	14.8	25
14140000	4.54	5.12	0.58	38.89	16.67	37.19	20.5	16	19.67	37.71	18.0	17
70220000	20.15	20.70	0.55	39.02	17.10	37.60	20.5	17	21.25	38.13	16.9	21
72040000	12.20	12.58	0.38	35.52	12.16	31.99	19.8	18	11.12	31.85	20.7	15
72040000	11.45	12.20	0.75	34.05	13.03	32.44	19.4	19	11.98	32.37	20.4	16
55320000	8.88	9.12	0.24	35.27	12.23	30.32	18.1	20	14.20	31.47	17.3	19
72001000	22.17	22.29	0.12	39.06	14.82	32.84	18.0	21	18.12	34.67	16.5	22
75471000	0.86	0.98	0.12	41.67	17.59	34.95	17.4	22	18.64	35.59	16.9	20
72040000	12.58	12.79	0.21	33.72	12.16	28.62	16.5	23	11.12	28.41	17.3	18
75470000	23.34	23.74	0.4	31.73	13.23	29.66	16.4	24	15.98	30.32	14.3	26
72001000	34.86	35.00	0.14	34.42	12.21	28.20	16.0	25	14.61	29.79	15.2	24
10470000	3.55	4.10	0.55	31.88	15.52	30.51	15.0	26	16.41	30.67	14.3	27
10002000	4.41	5.00	0.59	21.54	5.99	18.65	12.7	27	6.70	19.17	12.5	31
72270000	11.47	12.06	0.59	26.41	13.35	25.18	11.8	28	14.04	25.39	11.4	34

**Table 7-6 Ranking of Top 28 HCLs for FI Crashes**

Roadway ID	Begin Mile Post	End Mile Post	Segment Length (miles)	Observed Crashes (crashes/mile/year)	Full Model				Simple Model			
					Predicted Crashes (crashes/mile/year)	Expected Crashes (crashes/mile/year)	PSI	Rank	Predicted Crashes (crashes/mile/year)	Expected Crashes (crashes/mile/year)	PSI	Rank
75008000	4.96	5.12	0.16	59.09	17.42	48.27	30.8	1	16.64	48.05	31.4	1
87005000	3.83	4.34	0.51	43.65	17.38	41.01	23.6	2	15.85	40.63	24.8	2
87075000	0.13	0.26	0.13	48.78	14.63	36.74	22.1	3	13.31	35.56	22.2	3
87470000	0.34	0.58	0.24	35.86	9.37	27.91	18.5	4	8.49	27.11	18.6	4
87021000	0.46	0.68	0.22	32.81	6.05	21.49	15.4	5	5.45	20.59	15.1	5
87005000	2.03	2.21	0.18	33.24	12.23	26.96	14.7	6	11.11	26.21	15.1	6
86471000	6.71	7.01	0.3	26.67	6.27	19.56	13.3	7	7.30	20.74	13.4	7
87471000	39.19	39.36	0.17	28.90	10.33	23.52	13.2	8	11.62	24.40	12.8	9
87260000	1.87	2.01	0.14	32.09	13.76	26.14	12.4	9	12.52	25.35	12.8	8
87005000	6.39	6.54	0.15	29.87	10.06	22.22	12.2	10	8.18	20.64	12.5	10
14140000	4.54	5.12	0.58	21.37	8.43	19.17	10.7	11	9.49	19.61	10.1	11
72090000	4.42	4.55	0.13	29.53	5.79	15.81	10.0	12	4.82	14.41	9.6	12
70220000	20.15	20.70	0.55	19.96	9.46	18.67	9.2	13	10.58	18.95	8.4	14
72270000	15.62	16.15	0.53	18.83	7.98	16.98	9.0	14	9.81	17.62	7.8	17
79002000	23.30	23.88	0.58	16.41	5.75	14.49	8.7	15	6.23	14.74	8.5	13
75471000	0.86	0.98	0.12	24.12	8.51	16.28	7.8	16	8.82	16.60	7.8	19
72040000	11.45	12.20	0.75	15.35	6.42	13.98	7.6	17	5.84	13.87	8.0	16
72292000	0.96	1.11	0.15	24.19	12.99	20.53	7.5	18	11.81	19.90	8.1	15
70220000	40.98	41.50	0.52	16.67	8.25	15.45	7.2	19	9.15	15.70	6.6	22
10470000	3.55	4.10	0.55	16.39	7.54	14.70	7.2	20	7.80	14.80	7.0	20
55320000	8.88	9.12	0.24	17.63	6.17	13.00	6.8	21	6.86	13.66	6.8	21
75008160	1.11	1.94	0.83	14.65	7.07	13.57	6.5	22	7.32	13.65	6.3	24
87005000	6.54	7.03	0.49	17.86	10.06	16.53	6.5	23	8.18	15.98	7.8	18
72001000	34.86	35.00	0.14	18.12	6.61	13.00	6.4	24	7.23	13.59	6.4	23
75470000	10.41	10.60	0.19	17.76	8.44	14.39	6.0	25	8.50	14.59	6.1	27
93470000	32.81	32.97	0.16	16.67	7.16	13.04	5.9	26	7.87	13.56	5.7	30
72040000	12.20	12.58	0.38	14.08	5.99	11.80	5.8	27	5.43	11.61	6.2	26
72292000	1.11	1.28	0.17	20.06	11.23	16.97	5.7	28	10.19	16.40	6.2	25

**Table 7-7 Paired t-Test Results**

<b>Statistic</b>	<b>Full Model</b>	<b>Simple Model</b>
Mean	7.711675521	7.662571638
Variance	121.4353987	126.3419956
Observations	198	198
Pearson Correlation	0.985904346	
Hypothesized Mean Difference	0	
df	197	
t Stat	0.367212125	
P(T<=t) one-tail	0.35692748	
t Critical one-tail	1.65262522	
P(T<=t) two-tail	0.71385496	
t Critical two-tail	1.972078988	

## 7.6. Summary

The EB approach identifies high crash locations based on their PSI, calculated as the difference between expected and predicted crashes at the location. The expected crash frequency is calculated as the weighted average of the observed and predicted crash frequencies. The predicted crash frequency is calculated using an SPF, and depends on roadway geometric characteristics (such as lane width, shoulder width, horizontal and vertical curvatures, etc.) and traffic characteristics (such as traffic volume and speed limit).

There are two main types of SPFs: simple and full. The simple SPFs are developed using data from all sites regardless of base conditions, and consider traffic to be the sole predictor of crashes. On the other hand, the full SPFs account for both traffic and geometric conditions. The approach of using just the simple SPFs to predict crashes without accounting for changes in geometric conditions is questionable. This chapter, therefore, focused on the comparison of crash prediction performance and the ranking of HCLs generated using the full and the simple SPFs.

Compared to freeways, simple SPFs are more widely recognized and used for arterial roads and intersections than full SPFs for the use of predicting crash performance. The HSM (2010) provides simple SPF functional forms for both arterial roads and intersections based on many previous researches. However, the SPF functional form for freeways is not recommended in the first edition of the HSM. Both full SPFs and simple SPFs were found to be widely used in previous studies. Thus, freeways were used as examples to present the comparison of the performance of the full and simple SPFs in this research. A total of 261 miles of urban 4-lane freeway sections within interchange influence areas in Florida were identified. Four-year crash data from 2007-2010 were used in the analysis. Four-fold cross-validation was chosen as the sampling method to divide the dataset into calibration and prediction groups. Full and simple SPFs were developed using the NB regression method for both total and FI crashes. The MAD and the MSPE were used to compare the prediction performance of the two SPFs, as well as the variations in the rankings of the HCLs identified (using the two SPFs).

The results showed that the prediction performance of the simple SPF resembled that of the full SPF for both total and FI crashes on urban 4-lane freeway interchanges. Moreover, the lists of top HCLs generated using the two models were almost identical, and there is no significant difference between the PSIs in the network screening process when using the two models. Besides urban 4-lane freeway interchanges, analysis on other subtypes of freeways in Florida came up with same results. Thus, the results from this research show that simple models produce similar results, in spite of using AADT as the only predictor. Both the simple and the full SPFs yield similar prediction estimations on freeways in Florida.



## CHAPTER 8

### SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

#### 8.1. Summary and Conclusions

A comprehensive literature review was performed to investigate and assess the development of SPFs, as well as its applications. *SafetyAnalyst*, a safety analysis software system, was introduced and discussed. *SafetyAnalyst* uses a set of national default SPFs. Although these national default SPFs can be calibrated to reflect a state's safety data, in order to better represent local safety performance, local agencies were encouraged to develop their own SPFs. Two applications provided by SPFs, crash prediction and identification of high crash locations, were also discussed, and were found to be affiliated with the precision of SPFs. As such, the main objective of this research was to develop Florida-specific SPFs to be used with *SafetyAnalyst*.

Florida-specific SPFs were developed using RCI data from 2008, as well as crash and traffic data from 2007-2010, for both total and FI crashes. The data were divided into two sets, one for calibration and another for validation. In accordance with the structure of predefined subtypes used in *SafetyAnalyst*, roadway segments were divided into 17 site subtypes, based on area type, functional classification, and number of lanes.

Compared to roadway segments, the data requirements to generate intersection SPFs are rigorous. One of the required variables for *SafetyAnalyst* to divide intersections into subtypes is traffic control type, yet it is not available in the detail required in the RCI database. Therefore, SPFs were developed for only four types of signalized intersections: rural and urban, with three-leg and four-leg each. At present,

due to the lack of detailed traffic control type data, analysis of unsignalized intersections is not possible.

*SafetyAnalyst* classifies ramps into 16 subtypes based on ramp configuration, ramp type, and area type. This classification could not be used to generate Florida-specific SPFs, as Florida has completely different ramp classifications. Instead, Florida-specific subtypes were used to generate SPFs for ramps.

Segmentation is an essential step in data preparation for developing the SPFs. Three segmentation methods were described in this chapter, including fixed-length, variable-length, and Fisher's clustering algorithm. According to *SafetyAnalyst*, SPFs are developed for homogenous segments of variable length. Thus, variable length was used as the segmentation method for developing Florida-specific SPFs in this research. Moreover, in order to improve the development of SPFs that identify high crash locations, an alternative segmentation method was examined in this step that was based on Fisher's clustering algorithm.

The NB model was used to develop the Florida-specific SPFs for each of the subtypes of roadway segments, intersections and ramps, for both total crashes and FI crashes, using the calibration data. Most of the model parameters were statistically significant, while some coefficients had fewer higher p-values, but were still found to be acceptable. The overdispersion parameter is important in that it indicates the statistical reliability of the SPF; this parameter was used to account for dispersion in the data. Florida-specific SPFs were deemed reliable after observation of the overdispersion parameters for all subtypes were found to have low levels.

In addition, CURE plots for each subtype that showed total and FI crashes as a function of AADT were used to test the fit of the models. For most of the subtypes, the cumulative residuals for the Florida-specific models fall within the  $\pm 2\sigma$  boundaries for the entire range of AADT. The models fluctuate around zero on the horizontal axis, which is indicative of a well-fitted model. The cumulative residuals for several subtypes sometimes will consistently over- or underestimate the crash count or stray beyond falling within the  $\pm 2\sigma$  boundaries. However, the overall fit to the data is still preferable. Statistical goodness-of-fit tests were performed on both the calibration data and the validation data for all subtypes. The results were compared in order to assess the transferability of the Florida-specific SPF models, which were developed using the calibration data, to the validation data.

After the development of Florida-specific SPFs, the default *SafetyAnalyst* SPFs were calibrated to Florida data by adjusting the national default SPFs with local calibration factors. The performance of the Florida-specific SPFs and *SafetyAnalyst* default SPFs calibrated to Florida data were then compared using a number of methods, such as visual plots and statistical goodness-of-fit tests. In order to identify the better-fitted model, three goodness-of-fit tests represented by MAD, MSPE, and  $R^2_{FT}$  were performed on the validation data for both models for comparison. In most instances, Florida-specific SPFs presented lower values for both MAD and MSPE than *SafetyAnalyst* default SPFs calibrated to Florida data. The Florida-specific SPFs also presented higher  $R^2_{FT}$  results than *SafetyAnalyst* default SPFs calibrated to Florida data. Based on an overall comparison, Florida-specific SPFs yielded better prediction performance than the national default SPFs calibrated to Florida data. In general,

observed data were found to be better represented by Florida-specific SPFs than the default calibrated SPFs. However, this conclusion is dependent on the sample size. When fewer samples are available for a specific subtype, it is recommended to use the default calibrated SPFs.

In the roadway safety management process, there are two main applications of SPFs: crash prediction performance and identification of high crash locations. In this research, Florida-specific SPFs for all roadway subtypes developed were applied in crash prediction performance and identification of high crash locations. In this case, however, the simple SPF models in *SafetyAnalyst* were used to develop Florida-specific SPFs. The use of this traffic-only model in *SafetyAnalyst* may be questioned over the application of the traditional full SPFs for predicting crashes, as traffic might not be the sole predictor of crashes. Hence, this research compared the simple SPFs provided in *SafetyAnalyst* with full SPFs in both of these applications to determine if the two models yielded any similarities in the performance of crash prediction and network screening. The results showed that the two models yielded very similar performance of crash prediction and network screening. This empirical result supports the use of the flow-only SPF model adopted in *SafetyAnalyst*, which requires much less effort to develop, compared to full SPFs.

## **8.2. Recommendations**

Even though this dissertation has achieved the proposed research objective, the following areas require further research:

- For applications of Florida-specific SPFs, future research could extend the comparison of the full models and the simple models to other facilities, such as

urban multilane arterials, which have more varying geometric conditions and may potentially produce more dissimilar outcomes.

- Although the work conducted in this research has shown that Fisher's clustering method is more appropriate than the traditional segmentation methods, it is worth noting that this method was solely limited to freeway segments and should not be generalized to other roadway segments, such as arterial streets. Further research can extend the method to other locations, such as toll plazas and arterial corridors. Although the variables in this research have shown favorable model prediction performance, using more variables during the calibration procedure (e.g., interactions of two or more variables, as well as other relevant variables after segmentation) may further improve the results.

## REFERENCES

- Abdel-Aty, M. and J. Keller, "Exploring the Overall and Specific Crash Severity Levels at Signalized Intersections," *Accident Analysis & Prevention*, 37, pp. 417-425, 2005.
- Abdel-Aty, M. and E. Radwan, "Modeling Traffic Accident Occurrence and Involvement," *Accident Analysis & Prevention*, 32, pp. 633-642, 2000.
- Abdel-Aty, M., N. Uddin, A. Pande, F. Abdalla, and L. Hsia, "Predicting Freeway Crashes Based on Loop Detector Data Using Matched Case-Control Logistic Regression," *Transportation Research Record*, 1897, pp. 88-95, 2004.
- American Association of State Highway and Transportation Officials (AASHTO). *Highway Safety Manual*, 1st ed., Washington, D.C., 2009.
- American Association of State Highway and Transportation Officials (AASHTO). *SafetyAnalyst*, <http://www.safetyanalyst.org/>, Accessed June 14, 2010.
- Anastasopoulos, P. and F. Mannering, "A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models," *Accident Analysis & Prevention*, 41, pp. 153-159, 2009.
- Bauer, K. M. and D.W. Harwood, "Statistical Models of at-grade Intersections Accidents-addendum," Publication No.: FHWA-RD-99-094, FHWA, Washington D.C., 2000.
- Bonneson, J. and P. McCoy, "Estimation of Safety at Two-Way Stop-Controlled Intersection on Rural Highways," *Transportation Research Record*, 1401, pp. 83-89, 1993.
- Bowman, B. L., R. L. Vecellio, and J. Miao, "Vehicle and Pedestrian Accident Models for Median Locations," *Journal of Transportation Engineering*, 121 (6), pp. 531-537, 1995.
- Cappelli, C., R. Penny, and M. Reale, "Detecting Multiple Mean Breaks at Unknown Points with Atheoretical Regression Trees," MODSIM International Congress on Modeling and Simulation, Melbourne, Australia, pp. 974-978, 2005.
- Chang, L. Y., "Analysis of Freeway Accident Frequencies: Negative Binomial Regression versus Artificial Neural Network," *Safety Science*, 43(8), pp. 541-557, 2005.
- Chang, L.Y. and W.C. Chen, "Data Mining of Tree-Based Models to Analyze Freeway Accident Frequency," *Journal of Safety Research* 36(4), pp. 365-375, 2005.
- Chen, H., P. Liu, J. Lu, and B. Behzadi, "Evaluating the Safety Impacts of the Number and Arrangement of Lanes on Freeway Exit Ramps," *Accident Analysis & Prevention*, 41, pp. 543-551, 2009.

- Cheu, R. and S. Ritchie, "Automated Detection of Lane-blocking Freeway Incident using Artificial Neural Networks," *Transportation Research Part C: Emerging Technologies*, 3(6), pp. 371-388, 1995.
- Dean, C. and F. Lawless, "Tests for Detecting Overdispersion in Poisson Regression Models," *Journal of the American Statistical Association*, 84, pp. 467-472, 1989.
- Depaire, B., G. Wets, and K. Vanhoof, "Traffic Accident Segmentation by Means of Latent Class Clustering," *Accident Analysis & Prevention*, 40, pp. 1257-1266, 2008.
- Elvik, R., "State-of-the-Art Approaches to Road Accident Black Spot Management and Safety Analysis of Road Networks," Institute of Transport Economics, Report 883, Oslo, 2007.
- Elvik, R., P. Christensen, and A. Amundsen, "Speed and Road Accidents: An Evaluation of the Power Model," Institute of Transport Economics Report, 2004.
- Fisher, W. "On Grouping for Maximum Homogeneity," *American Statistical Association Journal*, 53, pp. 789-798, 1958.
- Fraley, C. and A. Raftery, "Model-based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97 (458), pp. 611-631, 2002.
- Fridstrom, L., J. Ifver, S. Ingebrigtsen, R. Kulmala, and L. K. Thomsen, "Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to the Variation in Road Accident Counts," *Accident Analysis & Prevention*, 27, pp. 1-20, 1995.
- Gan, A., K. Haleem, P. Alluri, J. Lu, T. Wang, M. Ma, and C. Diaz, "Preparing Florida for Deployment of SafetyAnalyst for All Road," Final Report Submitted to FDOT Research Center, Lehman Center for Transportation Research, Florida International University, 2012.
- Gan, A., J. Shen, and A. Rodriguez., "Update of Florida Crash Reduction Factors and Countermeasures to Improve the Development of District Safety Improvement Projects," Final Report Submitted to FDOT Research Center, Lehman Center for Transportation Research, Florida International University, 2005.
- Garber, N. J. and A. A. Ehrhart., "Final Report: The effect of speed, flow, and geometric characteristics on crash rates for different types of Virginia highways," Virginia Transportation Research Council, Charlottesville, Virginia, 2000.
- Geisser, S, Predictive Inference: An Introduction. New York, NY: Chapman and Hall. ISBN 0-412-03471-9, 1993.

- Geyer, J., E. Lankina, C. Chan, D. Ragland, T. Pham, and A. Sharafsaleh, "Methods for Identifying High Collision Concentration Locations for Potential Safety Improvements," California PATH Research Report, UCB-ITS-PRR-2008-35, 2008.
- Golob, T. F., W. W. Recker, and V. M. Alvarez, "Freeway safety as a function of traffic flow," *Accident Analysis and Prevention*, 36, pp. 933-946, 2004.
- Hadi, M., J. Aruldas, L.F. Chow, and J.A. Wattleworth, "Estimating the Safety Effects of Cross-Section Design for Various Highway Types Using Negative Binomial Regression," *Transportation Research Record*, 1500, pp. 169-177, 1995.
- Haleem, K., M. Abdel-Aty, and J. Santos, "Multiple Applications of the Multivariate Adaptive Regression Splines Technique in Predicting Rear-End Crashes at Unsignalized Intersections," *Transportation Research Record*, 2165, pp. 33-41, 2010.
- Harnen, S., U. R. S. Radin, S. V. Wong, and W. W. I. Hashim, "Motorcycle Crash Prediction Model for Non-Signalized Intersections," *IATSS Research*. 27(2), pp. 58-65, 2003.
- Hauer, E, "The Effect of Resurfacing on the Safety of Two-lane Rural Roads in New York State," Transportation Research Board 73<sup>rd</sup> Annual Meeting, Washington D.C., 1994.
- Hauer, E. Observational, "Before-After Studies in Road Safety," Pergamon Publications, London, 1997.
- Hauer, E., D. Harwood, F. Council, and M. Griffith, "Estimating Safety by the Empirical Bayes Method: a Tutorial," *Transportation Research Record*, 1784, pp. 126-131, 2002.
- Huang, Y., R. Cayford, and A. May, "Accident Prediction Models for Freeway Segments and Feasibility Study for Improving the Utilization of TASAS: Vol. 2 – Accident Prediction Models for Freeway Segments," Final Report, Report No. UCB-ITS-RR-91-9," University of California at Berkeley, 1991.
- Joshua, S. C. and N. J. Garber, "Estimating Truck Accident Rate and Involvements Using Linear and Poisson Regression Models," *Transportation Planning and Technology*, 15, pp. 41-58, 1990.
- Jovanis, P.P. and H. L. Chang, "Modeling the Relationship of Accident to Miles Traveled," *Transportation Research Record*, 1068, pp. 42-51, 1986.
- Karlaftis, M. G. and I. Golias, "Effects of Road Geometry and Traffic Volumes on Rural Roadway Accident Rates," *Accident Analysis & Prevention*, 34, pp. 357-365, 2002.
- Karlaftis, M. and A. Tarko, "Heterogeneity Considerations in Accident Modeling," *Accident Analysis & Prevention*, 30, pp. 425-433, 1998.



- Khan, S., R. Shanmugam, and B. Hoeschen, "Injury, Fatal, and Property Damage Accident Models for Highway Corridors," *Transportation Research Record*, 1665, 84-92, 1999.
- Kiattikomol, V., A. Chatterjee, J. Hummer, and M. Younger, "Planning Level Regression Models for Prediction of Crashes on Interchange and Noninterchange Segments of Urban Freeways," *Journal of Transportation Engineering*, 134, pp. 111-117, 2008.
- Konduri, S. and K. C. Sinha, "Statistical Models for Prediction of Freeway Incidents," the Seventh International Conference on Applications of Advances Technology in Transportation Engineering, Cambridge, MA, pp. 167-174, 2002.
- Kononov, J. and B. Allery, "Explicit Consideration of Safety in Transportation Planning and Project Scoping," Transportation Research Board 83rd Annual Meeting, Washington D.C., 2004.
- Kononov, J., B. Bailey, and B. Allery, "Exploratory Examination of the Functional Form of Safety Performance Functions of Urban Freeways," Transportation Research Board 87th Annual Meeting, Washington, D.C., 2008.
- Kraus, J. F., C. L. Anderson, S. Arzemanian, M. Salatka, P. Hemyari, and G. Sun, "Epidemiological Aspects of Fatal and Severe Injury Urban Freeway Crashes," *Accident Analysis and Prevention*, 25, pp. 229-239, 1993.
- Kuhnert, P. M., K. A. Do, and R. McClure, "Combining nonparametric models with logistic regression: An application to motor vehicle injury data," *Computational Statistics and Data Analysis*, 34(3), pp. 371-386, 2000.
- Lau, M., A. May, and R. Smith, "Application of Accident Prediction Models," *Transportation Research Record*, 1238, pp. 20-30, 1989.
- Laughland, J., L. Haefner, J. Hall, and D. Clough, "NCHRP Report 162: Methods for Evaluating Highway Safety Improvements," *Transportation Research Board*, Washington, D.C., 1975.
- Le, T. Q. and R.J. Porter. "Safety Evaluation of Geometric Design Criteria for Entrance-Exit Ramp Spacing and Auxiliary lane Use," Transportation Research Board 91st Annual Meeting, Washington, D.C., 2012.
- Li, X., D. Lord, Y. Zhang, and Y. Xie, "Predicting Motor Vehicle Crashes Using Support Vector Machine Models," *Accident Analysis & Prevention*, 40(4), pp. 1611-1618, 2008.
- Lord, D., A. Manar, and A. Vizioli. "Modeling Crash-Flow-Density and Crash-Flow-V/C Ratio Relationships for Rural and Urban Freeway Segments," *Accident Analysis and Prevention*, 37, pp. 185-199, 2005.

Lord, D. and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transport Research Part A*, 44, pp. 291-305, 2010.

Lu, J., A. Gan, K. Haleem, P. Alluri, and K. Liu, "Comparing Locally-Calibrated and SafetyAnalyst-Default Safety Performance Functions for Florida's Urban Freeways," Transportation Research Board 91st Annual Meeting, Washington D.C., 2012.

Lu, L., J. Lu, P. Lin, Z. Wang, and H. Chen. "Development of an Interface between FDOT's Crash Analysis Reporting System and the *SafetyAnalyst*," Final Report Submitted to FDOT Research Center, University of South Florida, July 2009.

Maher, M. and I. Summersgill, "A Comprehensive Methodology for the Fitting of Predictive Accident Models," *Accident Analysis & Prevention* Vol 28, pp. 281-296, 1996.

McDonald, J., "Relation between Number of Accidents and Traffic Volume at divided Highway Intersections," Highway Research Board, 1953.

McGee, H. and M. Blanskeship, "Guidelines for Converting Stop to Yield Control at Intersections," NCHRP Report 320, Transportation Research Board, Washington, D.C., 1989.

Miaou, S., "Measuring the Goodness-of-Fit of Accident Prediction Models. FHWA-RD-96-040," Federal Highway Administration, Washington, D.C., 1996.

Miaou, S., P. S. Hu, T. Wright, A. K. Rathi, and S. C. Davis, "Development of Relationships between Truck Accidents and Highway Geometric Design: Phase I," Technical Memorandum prepared by the Oak Ridge National Laboratory for the Federal Highway Administration, 1991.

Miaou, S., P. S. Hu, T. Wright, A. K. Rathi, and S. C. Davis, "Relationship between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach," *Transportation Research Record*, 1376, pp. 10-18, 1992.

Miaou, S., "The Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson versus Negative Binomial Regressions," *Accident Analysis & Prevention*, 26, pp. 471-482, 1994.

Morin, D., "Application of Statistical Concepts to Accident Data," *Highway Research Record*, 188, pp. 72-79, 1967.

Mountain, L., B. Fawaz, and D. Jarret., "Accident Prediction Models for Roads with Minor Junctions," *Accident Analysis & Prevention*, 28, pp. 695-707, 1996.

Nambuusi, B. B., T. Brijs, and E. Hermans, "A review of accident prediction models for road intersections," Steunpunt Mobiliteit & Openbare Werken – Spoor Verkeersveiligheid RA-MOW-2008-004, 2008.

*NCHRP Report 197*, "Cost and Safety Effectiveness of Highway Design Elements," Transportation Research Board, National Research Council, Washington, D.C., 1978.

NHTSA, "Technology Applications for Traffic Safety Programs: A Primer," National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, D.C., 2008.

Norden, M., J. Orlansky, and H. Jacobs, "Application of Statistical Quality-Control Techniques to Analysis of Highway-Accident Data," *Highway Research Record*, pp. 17-31, 1956.

Pande, A. and M. Abedl-Aty, "Assessment of Freeway Traffic Parameters Leading to Lane-Change Related Collisions," *Accident Analysis & Prevention*, 38, pp. 936-948, 2006.

Parajuli, B., B. Persaud, C. Lyon, and J. Munro, "Safety Performance Assessment of Freeway Interchanges, Ramps, and Ramp Terminals," Road Safety Engineering Management, Session of Annual Conference of the Transportation Association of Canada, Charlottetown, Prince Edward Island, 2006.

Park, B. and D. Lord, "Adjustment for Maximum Likelihood Estimate of Negative Binomial Dispersion Parameter," Transportation Research Board 87th Annual Meeting, Washington, D.C., 2008.

Persaud, B., "Estimating Accident Potential of Ontario Road Sections," *Transportation Research Record*, 1327, pp. 47-53, 1991.

Persaud, B., "Road Safety – a Review of the Ontario Experience and Relevant Work Elsewhere," Ontario Ministry of Transportation, 1992.

Persaud, B. and L. Dzbik, "Accident Prediction Models for Freeways," *Transportation Research Record*, 1041, pp. 55-60, 1993.

Persaud, B., E. Hauer, R. Vallurupalli and K. Mucsi, "Study of the Safety Effect of Removing Traffic Signals in Philadelphia," Draft Final Report, 1995.

Pilko, P., J. Bared, P. Edara, and T. Kim, "Safety Assessment of Interchange Spacing on Urban Freeways," Publication No. FHWA-HRT-07-031, 2007.

Poch, M. and F. Mannering, "Negative Binomial Analysis of Intersection Accident Frequencies," *Journal of Transportation Engineering*, 122(2), pp. 105-113, 1996.

- Qin, X., J. N. Ivan, and N. Ravishanker, "Selecting Exposure Measures in Crash Rate Prediction for Two-lane Highway Segments," *Accident Analysis and Prevention*, 36, pp. 183-191, 2004.
- Qin, X., M. Ng, and P. E. Reyes, "Identifying Crash-Prone Locations with Quantile Regression," *Accident Analysis & Prevention*, 42, pp. 1531-1537, 2010.
- Resende, P. T. and R. F. Benekohal, "Development of Volume-to-Capacity Based Accident Prediction Models." *Traffic Congestion and Traffic Safety in the 21st Century*, Chicago, IL, pp. 215-221, 1997.
- Salifu, M., "Accident prediction models for unsignalized urban junctions in Ghana," *IATSS Research*, 28(1), pp. 68-81, 2004.
- Sawalha, Z., "Traffic Accident Modeling: Statistical Issues and Safety Applications," The University of British Columbia, April, 2002.
- Sawalha, Z. and T. Sayed, "Transferability of Accident Prediction Models," *Journal of Safety Science*, 44(3), pp. 209-219, 2006.
- Sayed, T. and F. Rodriguez, "Accident Prediction Models for Urban Unsignalized Intersections in British Columbia," *Transportation Research Record*, 1665, pp. 93-99, 1999.
- Shankar, V., F. Mannering, and W. Barfield, "Effect of Roadway Geometrics and Environment Factors on Rural Freeway Accident Frequencies," *Accident Analysis & Prevention*, 27, pp. 371-389, 1995.
- Shen, J. and A. Gan, "Development of Crash Reduction Factors Methods, Problems, and Research Needs," *Transportation Research Record*, 1840, pp. 50-56, 2003.
- Stokes, R. and M. Mutabazi, "Rate-Quality Control Method of Identifying Hazardous Road Locations," *Transportation Research Record*, 1542, pp. 44-48, 1996.
- Tegge, R. and Y. Ouyang, "Developing Safety Performance Functions for Illinois," ICT Project R27-20 Submitted to Illinois DOT, University of Illinois at Urbana-Champaign, 2007.
- U.S. Department of Transportation and Federal Highway Administration (FHWA). *SafetyAnalyst: Software Tools for Safety Management of Specific Highway Sites*, July 2010.
- Vernon, D. and L. Cook, "Analysis of the Impact of Increased Speed Limits in Utah," U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, D.C., 1999.

- Vogt, A. and G. Bared., "Accident Models for Two-Lane Roads: Segments and Intersections," Federal Highway Administration, Report No. FHWA-RD-98-133, 1998.
- Wang, J. and T. Tong, "Research on Ordinal Cluster Division for Maintenance Road Section.," Second International Conference on Intelligent Computation Technology and Automation, pp. 707-710, 2009.
- Wang, X. and M. Abdel-Aty, "Temporal and Spatial Analyses of Rear-End Crashes at Signalized Intersections," *Accident Analysis & Prevention*, 38, pp. 1137-1150, 2006.
- Washington, S., B. Persaud, C. Lyon, and J. Oh, "Validation of Accident Models for Intersections," FHWA Report, Publication No. FHWA-RD-03-037, Federal Highway Administration, Washington, D.C., 2005.
- Webb, G., "The Relation between Accidents and Traffic Volumes at Signalized Intersections," *ITE Proceedings*, pp. 149-167, 1955.
- Wong, J. and Y. Chung, "Analyzing Heterogeneous Accident Data from the Perspective of Accident Occurrence," *Accident Analysis & Prevention*, 40, pp. 357-367, 2008.
- Wood, G., "Confidence and prediction intervals for generalized linear accident models," *Accident Analysis & Prevention*, 37, pp. 267-273, 2005.
- Xu, R. and D. Wunsch, "Survey of Clustering Algorithms," *IEEE Trans. Neural Networks*, 16 (3), pp. 645-678, 2005.
- Xu, X., "Fisher Ordered Gathering and Its Application on Periodical Pick-up of Boiler Pipe Temperature," *Journal of Guangxi Electric Power*, 5, pp. 12-14, 2005.
- Young, J. and P. Park, "Comparing the Highway Safety Manual's Safety Performance Functions with Jurisdiction-Specific Functions for Intersections in Regina," Annual Conference of the Transportation Association of Canada Fredericton, New Brunswick, 2012.
- Zegeer, C., "Methods for Identifying Hazardous Highway Elements," Transportation Research Board, Washington, D.C., 1986.
- Zegeer, C., R. Stewart, D. Reinfurt, F. M. Council, T. Neuman, E. Hamilton, T. Miller, and W. Hunter, "Cost-effective Geometric Improvements for Safety Upgrading of Horizontal Curves," Federal Highway Administration, Washington, D.C.1990.
- Zhang, Y., Z. Ye, and D. Lord, "Estimating Dispersion Parameter of Negative Binomial Distribution for Analysis of Crash Data: Bootstrapped Maximum Likelihood Method," *Transportation Research Record*, 2019, pp. 15-21, 2007.

Zhong, L., Y. Chen, X. Sun, X. Liu, and Y. He., "Research on Section Division of Freeway with Ordinal Clustering Method," International Conference on Transportation Engineering, pp. 4171-4177, 2007.

## VITA

JINYAN LU

### EDUCATION

- 1999 – 2003 B.S., Civil Engineering  
Beijing University of Technology  
Beijing, China
- 2004 – 2006 M.S., Transportation Engineering  
Beijing University of Technology  
Beijing, China
- 2007 – 2013 Doctoral Candidate in Civil Engineering  
Florida International University  
Miami, Florida

### EMPLOYMENT

- 2007– 2012 Graduate Teaching/Research Assistant  
Lehman Center for Transportation Research  
Department of Civil and Environmental Engineering  
Florida International University, Miami, Florida

### PUBLICATIONS

- **Lu, Jinyan**, Kirolos Haleem, Priyanka Alluri, and Albert Gan, “Full versus Simple Safety Performance Functions: A Comparison Based on Urban Four-Lane Freeway Interchanges in Florida”, *Journal of Transportation Research Board: Transportation Research Record (in press)*.
- **Lu, Jinyan**, Albert Gan, Kirolos Haleem, and Wanyang Wu, “Clustering-Based Roadway Segment Division for the Identification of High Crash Locations”, *Journal of Transportation Safety and Security (in press)*.
- Haleem, Kirolos, Albert Gan, and **Jinyan Lu**, “Using Multivariate Adaptive Regression Splines (MARS) to Develop Crash Modification Factors for Urban Freeway Interchange Influence Areas”, *Accident Analysis & Prevention*, Volume 55, pp. 12-21, June 2013.
- Gan, Albert, Kirolos Haleem, Priyanka Alluri, **Jinyan Lu**, Tao Wang, Meng Ma, and Claudia Diaz, “Preparing Florida for Deployment of *SafetyAnalyst* for All Roads”, Final Report, Prepared for Florida DOT, May 2012.
- **Lu, Jinyan**, “Comparison of Local and Default Safety Performance Functions for *SafetyAnalyst* Applications”, *Best Student Paper*, District 10 ITE, May 2012.
- **Lu, Jinyan**, Albert Gan, Kirolos Haleem, Priyanka Alluri, and Kaiyu Liu, “Comparing Locally-Calibrated and *SafetyAnalyst*-Default Safety Performance

- Functions for Florida's Urban Freeways", *Proceedings of the 91st Annual Meeting of the Transportation Research Board*, January 2012.
- Haleem, Kirolos, Albert Gan, and **Jinyan Lu**, "Developing Crash Modification Factors for Interchange Influence Areas on Urban Four-Lane Freeways Using Multivariate Adaptive Regression Splines", *Proceedings of the 91st Annual Meeting of the Transportation Research Board*, January 2012.
  - Sun, Lishan, Liya Yao, Jian Rong, **Jinyan Lu**, Bohua Liu, and Shuwei Wang, "Simulation Analysis on Driving Behavior during Traffic Sign Recognition", *International Journal of Computational Intelligence Systems*, Volume 4, Issue 3, pp. 353-360, 2011.
  - **Lu, Jinyan**, Kirolos Haleem, Albert Gan, and Wanyang Wu, "Application of Fisher's Clustering Method to Improve Empirical Bayes Estimation for the Identification of High Crash Locations", *Proceedings of the 90th Annual Meeting of the Transportation Research Board*, January 2011.
  - **Lu, Jinyan** and Jiancheng Weng, "Improvement of Rate Quality Control Method in Identifying Hazardous Locations", *Proceedings of the 10th International Conference of Chinese Transportation Professionals*, ASCE, Volume 382, pp. 591-602, China, 2010.
  - Weng, Jian-Cheng, Xiao-Guang Zhao, **Jinyan Lu**, and Meng-Jia Wang, "Influence Analysis of Traffic Demand Management Policy in Beijing based on FCD", *Proceedings of the 10th International Conference of Chinese Transportation Professionals*, ASCE, Volume 382, pp. 2350-2361, China, 2010.
  - **Lu, Jinyan**, Xiang Zhou, and Xiaoming Liu, "The Measurement of Critical Gap of Return Headway on Chinese Two-lane Highway Using Passing Experiment", *Proceedings of the International Conference on Transportation Engineering*, ASCE, pp. 3476-3481, China, 2007.
  - **Lu, Jinyan**, Pingnan Ruan, and Xiaoming Liu, "Study on the Way to Achieve the Maximum Social Welfare of Road Resource", *Journal of Beijing University of Technology*, No. 2, pp. 189-192, 2007.
  - Zhou, Jian, Qingxia Liu, **Jinyan Lu**, and Jiang Liu, "Design of Simulation Models for Two-lane Highway on Mountainous Areas", *China Journal of Traffic Engineering*, No. 1, pp. 16-20, 2007.

## HONORS AND AWARDS

- Bill McGrath Transportation Studies Scholarship, Florida Section of ITE, 2012
- Best Student Paper Award, District 10 ITE, 2012
- Dissertation Year Fellowship, University Graduate School, FIU, since May 2012
- Student Poster Competition (second place), Florida Section of ITE & ITS Florida, 2012
- Helene M. Overly Memorial Scholarship, WTS Advancing Women in Transportation, 2011
- Book Scholarship Essay Competition (third place), ITE Gold Coast Chapter, 2010
- AASHTO Francis B. Francois Scholarship Award, 2009



