

To the Graduate Council:

I am submitting herewith a dissertation written by Vasin Kiattikomol entitled "Freeway Crash Prediction Models for Long-Range Urban Transportation Planning." I have examined the final electronic copy of this dissertation for form and content and recommended that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Civil Engineering.

Arun Chatterjee
Major Professor

We have read this dissertation
and recommend its acceptance:

Thomas Urbanik II

Lee D. Han

Mary Sue Younger

Accepted for the Council:

Anne Mayhew
Vice Chancellor and Dean of
Graduate Studies

(Original signatures are on file with official student records)

**FREEWAY CRASH PREDICTION MODELS FOR LONG-RANGE
URBAN TRANSPORTATION PLANNING**

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Vasin Kiattikomol
August 2005

DEDICATION

This dissertation is dedicated to my parents:

Kraiwood and Siritas Kiattikomol,

my sister:

Thidarat Kiattikomol,

and my grandmothers.

My destiny would not go as far as this without them.

ACKNOWLEDGMENT

The author would like to express his sincere gratitude and appreciation to all persons who contributed to the completion of this dissertation. The author would like to thank his major professor, Dr. Arun Chatterjee, for his patience and kindness in providing guidance and great support in the writing of this dissertation. The author is also greatly indebted to the other members of his committee: Dr. Thomas Urbanik II, Dr. Lee D. Han, and Dr. Mary Sue Younger for their time and guidance.

The author is grateful to Dr. Joseph E. Hummer, Professor at the North Carolina State University, for his support and advice for this research. The author also thanks Mr. Tony W. Ku and Mr. Kevin J. Lacy of North Carolina Department of Transportation for providing the data for North Carolina. The author also likes to thank Tennessee Department of Transportation for providing the data for Tennessee. The author is indebted to Ms. Vickie Schultze of the University of Tennessee for her help with processing of data for North Carolina and Tennessee.

The author is grateful to the research sponsor Southeastern Transportation Center and its Director, Ms. DeAnna Flinchum, for the financial support for this research.

Finally, the author wishes to express his appreciation to his family in Thailand for their love and endless encouragement for making this work possible. The author also likes to express his gratitude to Pornpat Wiwattarakul for her endless support and encouragement.

ABSTRACT

The need for safety assessment tools for long-range transportation planning was not seriously recognized until the Transportation Equity Act for Twenty First Century (TEA-21) established a requirement related to safety considerations in the planning process of Metropolitan Planning Organizations (MPO). Currently, most MPOs do not assess the safety consequences of alternative transportation systems. The goal of this dissertation is to develop practical tools for assessing safety consequences of freeways in the context of long-range urban transportation plans. Data for freeway segments and crashes were obtained from North Carolina Department of Transportation and Tennessee Department of Transportation. Three different modeling approaches were utilized for analyzing crash and freeway data. These approaches are the analysis of variance, regression analysis, and classification tree analysis. Separate models were developed for each crash type by severity for non-interchange segments and interchange segments. Appropriate independent variables and model forms were selected based on the availability of future data for the variables and statistical measures. The prediction performance of the different developed models was assessed and compared. The research reveals characteristics and patterns of crashes on urban freeways and provides crash prediction models that can be utilized in the long-range transportation planning process.

TABLE OF CONTENTS

1. INTRODUCTION AND MOTIVATION OF THE STUDY	1
OBJECTIVE OF THIS RESEARCH	2
ASSUMPTIONS AND SCOPE OF THIS RESEARCH.....	2
BENEFITS OF THIS RESEARCH.....	3
2. LITERATURE REVIEW	5
CRASH PREDICTION MODELS DURING 1980'S AND 1990'S.....	6
CRASH PREDICTION MODELS DURING 2000'S	8
MODEL STRUCTURE FOR PLANNING LEVEL APPLICATIONS	10
FINDINGS FOR DISTRIBUTION OF CRASHES	11
DEVELOPMENT OF CRASH PREDICTION MODELS	13
SUGGESTED MODEL VARIABLES AND MODEL FORMS	14
CRASH PREDICTION MODELS DEVELOPED FOR FREEWAYS	17
SUMMARY OF FINDINGS FROM LITERATURES	21
3. STUDY METHODOLOGY.....	24
DATA PREPARATION AND PRELIMINARY ANALYSIS	24
MODEL DEVELOPMENT.....	25
Analysis of Variance Approach	27
Regression Analysis Approach.....	28
Classification Tree Analysis Approach	32
MODEL ASSESSMENT AND COMPARISON	35
CONCLUSIONS.....	35
4. PRELIMINARY ANALYSIS OF DATA	37
CRASH AND FREEWAY DATA FOR NORTH CAROLINA	37
CRASH AND FREEWAY DATA FOR TENNESSEE	39
COMPARISON OF CRASH RATES FOR FREEWAY SEGMENTS OF NORTH CAROLINA AND TENNESSEE.....	41

CONCLUSIONS.....	43
5. CRASH PREDICTION MODELS BASED ON ANALYSIS OF VARIANCE	
APPROACH.....	45
ANOVA MODELS FOR PREDICTING CRASH RATES FOR NORTH CAROLINA	
FREEWAYS.....	45
ANOVA MODELS FOR PREDICTING CRASH RATES FOR TENNESSEE FREEWAYS...	50
CONCLUSIONS.....	55
6. CRASH PREDICTION MODELS BASED ON REGRESSION ANALYSIS	
APPROACH.....	56
CRASH PREDICTION MODELS BASED ON REGRESSION ANALYSIS FOR NORTH	
CAROLINA FREEWAYS	56
Regression Models for Non-Interchange Segments of North Carolina	57
Regression Models for Interchange Segments of North Carolina	59
Pattern of Variation in Crashes with Traffic Volume based on North Carolina	
Models.....	61
CRASH PREDICTION MODELS BASED ON REGRESSION ANALYSIS FOR TENNESSEE	
FREEWAYS	64
Regression Models for Non-Interchange Segments of Tennessee	64
Regression Models for Interchange Segments of Tennessee.....	66
Pattern of Variation in Crashes with Traffic Volume based on Tennessee	
Models.....	68
SUMMARY FOR THE FORMS OF REGRESSION MODELS.....	70
CONCLUSIONS.....	72
7. CRASH PREDICTION MODELS BASED ON CLASSIFICATION TREE	
ANALYSIS APPROACH.....	73
CLASSIFICATION TREE MODELS FOR NORTH CAROLINA FREEWAYS	73
Classification Tree Model for Non-Interchange Segments of North Carolina	74
Classification Tree Model for Interchange Segments of North Carolina	79
CLASSIFICATION TREE MODELS FOR TENNESSEE FREEWAYS	86

Classification Tree Model for Non-Interchange Segments of Tennessee	86
Classification Tree Model for Interchange Segments of Tennessee.....	92
CONCLUSIONS.....	98
8. ASSESSMENT AND COMPARISON OF CRASH PREDICTION MODELS FOR FREEWAYS OF NORTH CAROLINA AND TENNESSEE	100
ASSESSMENT AND COMPARISON OF CRASH PREDICTION MODELS DEVELOPED FOR NORTH CAROLINA FREEWAYS	101
Assessment and Comparison of Models for Non-Interchange Segments with Four Lanes	102
Assessment and Comparison of Models for Non-Interchange Segments with More than Four Lanes	103
Assessment and Comparison of Models for Interchange Segments with Four Lanes	104
Assessment and Comparison of Models for Interchange Segments with More than Four Lanes.....	105
ASSESSMENT AND COMPARISON OF CRASH PREDICTION MODELS DEVELOPED FOR TENNESSEE FREEWAYS	106
Assessment and Comparison of Models for Non-Interchange Segment with Four Lanes	107
Assessment and Comparison of Models for Non-Interchange Segments with More than Four Lanes	108
Assessment and Comparison of Models for Interchange Segments with Four Lanes	109
Assessment and Comparison of Models for Interchange Segments with More than Four Lanes.....	109
CONCLUSIONS.....	111
9. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	112
GENERAL CONCLUSIONS.....	112
CONTRIBUTIONS OF THIS RESEARCH	116

LIMITATIONS AND DIFFICULTIES OF THIS RESEARCH	117
FURTHER EXTENSIONS OF THE RESEARCH	118
LIST OF REFERENCES	119
VITA.....	127

LIST OF TABLES

Table 3-1 Categories of Crash Prediction Models for Urban Freeways	26
Table 3-2 Regression Model Forms for Non-Interchange Segments and Interchange Segments	29
Table 4-1 Summary of Freeway Segments for Six Counties in North Carolina	38
Table 4-2 Freeway Crashes by Severity Type for Six Counties in North Carolina (2000- 2002)	39
Table 4-3 Summary of Freeway Segments for Four Counties in Tennessee.....	40
Table 4-4 Freeway Crashes by Severity Type for Four Counties in Tennessee (2000- 2002)	40
Table 4-5 Crash Rates for Non-Interchange Segments in North Carolina and Tennessee	42
Table 4-6 Crash Rates for Interchange Segments in North Carolina and Tennessee	42
Table 5-1 Injury Crash Rates for North Carolina	47
Table 5-2 PDO Crash Rates for North Carolina	48
Table 5-3 Fatal and Injury Crash Rates for North Carolina.....	49
Table 5-4 Injury Crash Rates for Tennessee.....	51
Table 5-5 PDO Crash Rates for Tennessee.....	53
Table 5-6 Fatal and Injury Crash Rates for Tennessee	54
Table 6-1 Injury Crash Models for Non-Interchange Segments of North Carolina	57
Table 6-2 PDO Crash Models for Non-Interchange Segments of North Carolina	58
Table 6-3 Fatal and Injury Crash Models for Non-Interchange Segments of North Carolina.....	58
Table 6-4 Injury Crash Models for Interchange Segments of North Carolina	59
Table 6-5 PDO Crash Models for Interchange Segments of North Carolina	60
Table 6-6 Fatal and Injury Crash Models for Interchange Segments of North Carolina..	60
Table 6-7 Injury Crash Models for Non-Interchange Segments of Tennessee.....	64
Table 6-8 PDO Crash Models for Non-Interchange Segments of Tennessee	65
Table 6-9 Fatal and Injury Crash Models for Non-Interchange Segments of Tennessee.	65

Table 6-10 Injury Crash Models for Interchange Segments for Tennessee.....	66
Table 6-11 PDO Crash Models for Interchange Segments for Tennessee	67
Table 6-12 Fatal and Injury Crash Models for Interchange Segments for Tennessee.....	67
Table 6-13 Summary of Selected Model Forms for Non-Interchange Segments.....	71
Table 6-14 Summary of Selected Model Forms for Interchange Segments.....	71
Table 7-1 Crash Rates for the Classification Tree Model for Non-Interchange Segments of North Carolina	78
Table 7-2 Crash Rates for the Classification Tree Model for Interchange Segments of North Carolina	85
Table 7-3 Crash Rates for the Classification Tree Model for Non-Interchange Segments of Tennessee.....	93
Table 7-4 Crash Rates for the Classification Tree Model for Interchange Segments of Tennessee	97
Table 8-1 Number of Actual Crashes by Severity and Number of Segments for North Carolina Freeways	102
Table 8-2 Number of Actual Crashes by Severity and Number of Segments for Tennessee Freeways	107

LIST OF FIGURES

Figure 2-1 Relationship between Crashes and Vehicle Miles Traveled (VMT)	11
Figure 2-2 Freeway Segmentation Schemes Found in Previous Studies.....	23
Figure 3-1 Coverage Areas for Non-Interchange Segments and Interchange Segments..	27
Figure 3-2 An Example of 1-Level CART Tree.....	34
Figure 3-3 Diagram for Crash Prediction Modeling Process for Non-Interchange Segments	36
Figure 3-4 Diagram for Crash Prediction Modeling Process for Interchange Segments .	36
Figure 6-1 Predicted Pattern of Annual 'Fatal and Injury' Crashes based on North Carolina Models.....	62
Figure 6-2 Predicted Pattern of Annual 'PDO' Crashes based on North Carolina Models	62
Figure 6-3 Predicted Pattern of Annual 'Total' Crashes based on North Carolina Models	63
Figure 6-4 Predicted Pattern of Annual 'Fatal and Injury' Crashes based on Tennessee Models.....	69
Figure 6-5 Predicted Pattern of Annual 'PDO' Crashes based on Tennessee Models	69
Figure 6-6 Predicted Pattern of Annual 'Total' Crashes based on Tennessee Models.....	70
Figure 7-1 Classification Tree Model for Non-Interchange Segments of North Carolina	75
Figure 7-2 Enlarged Area of the Classification Tree Model for Non-Interchange Segments with Traffic Volume Less than or Equal to 55,000 Vehicles per Day	76
Figure 7-3 Enlarged Area of the Classification Tree Model for Non-Interchange Segments with Traffic Volume More than 55,000 Vehicles per Day.....	77
Figure 7-4 Classification Tree Model for Interchange Segments of North Carolina	80
Figure 7-5 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume Less than or Equal to 69,834 Vehicles per Day.....	81
Figure 7-6 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume from 69,835 to 88,575 Vehicles per Day.....	82

Figure 7-7 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume from 88,576 to 104,375 Vehicles per Day	83
Figure 7-8 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume More than 104,375 Vehicles per Day	84
Figure 7-9 Classification Tree Model for Non-Interchange Segments of Tennessee.....	87
Figure 7-10 Enlarged Area of the Classification Tree Model for Non-Interchange Segments with Traffic Volume Less than or Equal to 54,490 Vehicles per Day	88
Figure 7-11 Enlarged Area of the Classification Tree Model for Eight-Lane Non- Interchange Segments with Traffic Volume More than 54,490 Vehicles per Day	89
Figure 7-12 Enlarged Area of the Classification Tree Model for Six-Lane and Ten-Lane Non-Interchange Segments with Traffic Volume More than 54,490 Vehicles per Day	90
Figure 7-13 Enlarged Area of the Classification Tree Model for Four-Lane Non- Interchange Segments with Traffic Volume More than 54,490 Vehicles per Day	91
Figure 7-14 Classification Tree Model for Interchange Segments of Tennessee	94
Figure 7-15 Enlarged Area of the Classification Tree Model for Six-Lane, Eight-Lane, and Ten-Lane Interchange Segments with Traffic Volume Less than or Equal to 146,021 Vehicles per Day	95
Figure 7-16 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume More than 146,021 Vehicles per Day	95
Figure 7-17 Enlarged Area of the Classification Tree Model for Four-Lane Interchange Segments with Traffic Volume Less than or Equal to 146,021 Vehicles per Day	96
Figure 8-1 Comparison of the Models for Non-Interchange Segments with Four Lanes of North Carolina	103

Figure 8-2 Comparison of the Models for Non-Interchange Segments with More than Four Lanes of North Carolina	104
Figure 8-3 Comparison of the Models for Interchange Segments with Four Lanes of North Carolina	105
Figure 8-4 Comparison of the Models for Interchange Segments with More than Four Lanes of North Carolina.....	106
Figure 8-5 Comparison of the Models for Non-Interchange Segments with Four Lanes of Tennessee	108
Figure 8-6 Comparison of the Models for Non-Interchange Segments with More than Four Lanes of Tennessee	109
Figure 8-7 Comparison of the Models for Interchange Segments with Four Lanes of Tennessee	110
Figure 8-8 Comparison of the Models for Interchange Segments with More Than Four Lanes of Tennessee	110

LIST OF ABBREVIATIONS

ADT	Average Daily Traffic
AADT	Annual Average Daily Traffic
ANOVA	Analysis of Variance
MPO	Metropolitan Planning Organization
MVMT	Million Vehicle Miles Traveled
NC	North Carolina
NCDOT	North Carolina Department of Transportation
OLS	Ordinary Least Squares
PDO	Property Damage Only
RMSE	Root Mean Squared Error
TDOT	Tennessee Department of Transportation
TEA-21	Transportation Equity Act for the Twenty First Century
TN	Tennessee
VMT	Vehicle Miles Traveled

1. INTRODUCTION AND MOTIVATION OF THE STUDY

Predictive models for assessing safety impacts of different types of highway facilities have been of interest to transportation and traffic engineers for a long time. The survey by Satterthwaite (1981) reveals over 80 previous papers on the investigation of the relationship between crashes and traffic volume and other variables. Recent studies have established even more detailed relationships among vehicle crashes, geometric features of roadways, and a variety of other factors, which make the recent developed models more accurate than ones in the past in predicting the number of crashes for designated roadway facilities. However, these models with more predictor variables related to roadway characteristics, drivers, weather conditions, and environmental factors are found to create difficulties with regard to forecasting all those required input data for long-range transportation planning. Chatterjee, et al. (2003) pointed out that predictor variables commonly needed for these models are related to site-specific detailed information, which is not generally available for future transportation system plans that are examined in the long-range planning process.

While there are several research studies on developing crash prediction models for roadway facilities, most of them are associated with non-freeway facilities. Fazio, Holden, and Roupail (1993) gave an explanation why the number of studies on freeway crashes is low. They pointed out that higher average crash rates are normally observed on non-freeway roadways and especially intersections. This seems to encourage researchers and traffic safety experts to dedicate more attention on developing predictive models for non-freeway facilities and intersections.

Among the limited number of existing crash prediction models for freeways, the majority appear to have been developed for the purpose of either identifying highly hazardous locations, traditionally called black-spots, or evaluating the safety effectiveness of safety treatments, i.e., countermeasures. These models are suitable for corridor and short range

planning involving specific road segments. The need for safety assessment tools at the long-range planning level was not seriously recognized until Transportation Equity Act for the Twenty First Century (TEA-21) established a requirement related to safety to be considered in long-range transportation planning in the form of one of the seven planning factors. TEA-21 has called for Metropolitan Planning Organizations (MPOs) in urbanized areas to take into account the safety related aspects of alternative urban transportation systems in the metropolitan planning process. However, having no reliable analytical tools, as Khan, Shanmugam, and Hoeschen (1999) observed, MPOs currently evaluate safety of alternative plans based on “rules of thumb” and expert judgment. There is a need to develop tools, standards, or criteria to assess safety impacts and their economical consequences in the long-range transportation planning process. To provide urban transportation planners with simple and practical forms of crash models for suitable for long-range transportation planning applications is the main motivation of this research study.

Objective of this Research

The objective of this dissertation is to develop simple forms of predictive models for identifying safety impacts for urban freeways. These crash prediction models are intended to be used to assess the consequences related to traffic safety associated with various alternative types of highway networks for a long-range transportation plan. The independent variables used in the models are not too difficult to forecast for future scenarios. These models will be useful when used in conjunction with other predictive models for non-freeway arterial highways.

Assumptions and Scope of this Research

The research plan is formulated based on the hypothesis that the risk of crashes is different between freeway segments near interchanges and freeway segments outside interchange influence areas, respectively. This is because opportunities for conflicts are

normally higher in areas near interchanges than those in areas farther away. Therefore, developing crash prediction models separately for non-interchange segments and interchange segments is expected to improve the predictive capability of the models representing the different traffic patterns in these two regimes. Another assumption made is that separate models should be developed for crashes of different severity of injury. This actually is recognized in other crash related researches.

Unlike the common approach for crash prediction modeling, this study does not try to identify and select all significant predictor variables that are available in database files. This study is interested in predicting crashes using terms of different levels of traffic volume and some other predictor variables for which long-range forecasts are usually available. In addition, this study focuses on developing simple and practical forms of crash prediction models that can be used in urban transportation planning.

The crash prediction models for long-range planning applications are not intended to replace the detailed models previously developed by other researchers since the application of these two categories of models are different. The crash prediction models of this study are expected to remove the lack of tools for evaluating alternative long-range transportation plans. Many facilities included in long-range plans do not exist at the present time and their future geometric and traffic characteristics are not known at a detailed level. The existing prediction models in most cases are intended to evaluate safety impacts of projects at the design stage or for corridor planning when detailed design features of alternative projects are known.

Benefits of this Research

The prediction models to be developed are expected to strengthen the understanding of transportation planners and their ability to respond with regard to several issues and questions:

- Better understanding of the nature of crashes on urban freeways and the effect of interchanges on crash occurrences on freeways.
- The ability to predict crashes by severity types on new and existing urban freeways for estimated traffic volumes.
- Better understanding of the safety consequences of adding or eliminating an interchange on a proposed freeway segment. The developed crash prediction models are expected to allow planners to predict crashes separately for non-interchange segments and interchange segments.
- The opportunity to use different forms of crash prediction techniques for estimating freeway crashes in long-range urban transportation system alternatives.

However, the primary benefit of this study on crash prediction models is that it provides transportation planners tools for assessing the safety impact of alternative freeway networks in urban areas and also that these models can be applied utilizing the outputs of traditional transportation demand forecasting models.

2. LITERATURE REVIEW

The main sources used for the literature search for this study include three engineering databases and one electronic journal subscribed by the University of Tennessee. The three databases found to be useful are Compendex, Web of Science, and Transportation Research Information Services (TRIS). The subscribed electronic journal's title is Accident Analysis and Prevention. The searches returned nearly 40 articles representing different forms of crash prediction models, different types of applications for crash prediction models, and different classes of roadway facilities for which models were developed.

Chatterjee, et al. (2003) classified crash prediction model forms into either simple or complex types. The simple model form was described as a cross-classification table where the cell values are mean crash rates and the classification is based on different predictor variables such as traffic volumes, roadway characteristics, and some other crash related factors. The complex model form referred to mathematical formulations that correlated crash frequencies with various predictor variables using advance statistical techniques. Four common statistical approaches for developing complex model formulations were found from previous studies and these are simple linear regression, non-linear regression, neural network, and nested logic.

Khan, Shanmugam, and Hoeschen (1999) pointed out that existing crash prediction models have been used for three main common applications, which include planning, preliminary design, and before-and-after evaluation. In addition, Lau, May, and Smith (1989) determined more detailed types of applications of crash prediction models including large-scale regional transportation studies, estimation of safety performance of new facilities, estimation of safety performance of redesigned facilities, network simulation and optimization studies, crash surveillance, estimation of safety of existing facilities with crash history, and before-and-after studies.

Most of previous researches were found to focus on developing crash prediction models for specific types of roadway facilities such as two-lane rural roads, freeway sections, non-signalized intersections, and signalized intersections. More aggregate types of crash prediction models for analysis zones or larger areas were also found in the studies by Frantzeskakis, Assimakopoulos, and Kindinis (1994) and Guevara, Washington, and Oh (2004).

This literature review focuses on the methodologies and applications presented in various articles as well as the problems encountered in these studies with regard to developing predictive models for estimating safety impacts. The review first presents an overview of the previous work and then presents an in depth analysis of the work involving crash prediction modeling related to freeways.

Crash Prediction Models During 1980's and 1990's

In the 1980s, some of the research studies dealt with developing crash prediction models in connection with before-and-after evaluation applications (Maycock and Hall 1984, Jovanis and Chang 1986, Hauer, Ng, and Lovell 1988, and Lau, May, and Smith 1989). There were investigations on safety at intersections also, and these were more interesting (Maycock and Hall 1984, Hauer, Ng, and Lovell 1988, and Lau, May, and Smith 1989). There was one study by Jovanis and Chang (1986) that investigated the safety on general highway sections. The search could not identify any previous work related to prediction of safety on freeway facilities during this time period. Lau, May, and Smith (1989) presented an idea on developing flexible models, a three-level prediction procedure, for a wide range of applications on safety of intersections. Their study was the only one found during this time period that touched on safety prediction at the planning level. Most of the prediction models were developed for assessing total crash frequencies regardless of crash types or crash severity while Lau, May, and Smith (1989) took the severity of crashes into account in their prediction models.

During the 1980's, Maycock and Hall (1984), Jovanis and Chang (1986), and Hauer, Ng, and Lovell (1988) examined the relationship between crashes and traffic flows and found out that they were not linearly related. Jovanis and Chang (1986) clearly illustrated the phenomenon that led to the use of Poisson regression approach instead of linear regression approach. Maycock and Hall (1984) demonstrated how the Negative Binomial modeling could be used as an extension to Poisson modeling. The argument against using linear regression modeling was strongly made during this time period. A generalized linear modeling approach was utilized by Hauer, Ng, and Lovell (1988), which incorporated a non-linear function. Lau, May, and Smith (1989) also utilized a univariate analysis approach for planning level models.

In 1990s, the non-linear relationship between the crash occurrence and a traffic volume was widely accepted and the attention of researchers was drawn on developing appropriate mathematical equations that expressed crash frequencies as a function of the traffic volume and some other significant roadway characteristics for several types of roadway facilities. Crash prediction models were developed for different types of facilities and situations as follows:

- For general roadway sections by Persaud (1991), Abdel-Aty, and Radwan (1999), and Khan, Shanmugam, and Hoeschen (1999)
- For rural sections by Persaud (1994)
- For freeway sections by Fazio, Holden, and Rouphail (1993), Kraus, et al. (1993), Persaud and Dzbik (1993), Shankar, Mannering, and Barfield (1996), and Resende and Benekohal (1997)
- For general intersections by Maher and Summersgill (1996)
- For urban non-signalized intersections by Sayed and Rodriguez (1999)
- For rural non-signalized intersections by Bonneson and McCoy (1993)
- For minor intersections by Mountain, Fawaz, and Jarrett (1996)
- For road networks in states or provinces by Frantzeskakis, Assimakopoulos, and Kindinis (1994)

Crash prediction models found during 1990's were developed mainly for the purpose of assessing future safety of existing facilities given historical crash data, or for before-and-after evaluation application. The investigations during this time period also detected non-linear relationships between the crash and the traffic volume and agreed with the forerunners on the idea of using the non-linear regression modeling approach. The most common approach used by previous work is the generalized linear modeling with either Poisson distribution assumption (Bonneson and McCoy 1993, Maher and Summersgill 1996, and Sayed and Rodriguez 1999) or Negative Binomial distribution assumption (Persaud 1991 and 1994, Persaud and Dzbik 1993, Mountain, Fawaz, and Jarrett 1996, and Sayed and Rodriguez 1999). However, Frantzeskakis, Assimakopoulos, and Kindinis (1994), Resende and Benekahal (1997), and Abdel-Aty and Radwan (1999) were found to have used the traditional multiple linear regression approach in modeling their crash prediction models. Among the studies found during this time, only the one developed by Frantzeskakis, Assimakopoulos, and Kindinis (1994) was found to be appropriate for application at the planning level. Their models did not use the traffic volume as an independent variable and, instead, the number of crashes by crash types was correlated with a linear function of variables such as tourist activities, population, and the length of primary and secondary roadway networks. However, it should be noted that urban transportation planning models normally forecast the traffic volume on each link of road networks and that the predictor variables used in the models developed by Frantzeskakis, Assimakopoulos, and Kindinis (1994) are not dealt with by travel demand forecasting models.

It is found that during the 1980's and 1990's there were hardly any studies that focused on developing crash prediction models for long-range planning applications.

Crash Prediction Models During 2000's

During the past three to four years, the crash prediction modeling efforts have included different facility types as follow:

- General roadway sections (Saccomanno and Fu 2003)
- Urban roadway sections (Miller, et al. 2001, Greibe 2003, Chatterjee, et al. 2003, Hauer, Council, and Mohammedshah 2004, and Lord and Persaud 2004)
- Rural roadway sections (Harwood, et al. 2000, Ivan, Wang, and Bernardo 2000, Ivan 2004, and Maze, et al. 2005)
- Freeways (Konduri and Sinha 2002, Golob, Recker, and Alvarez 2004, and Kononov and Allery 2004)
- Urban signalized intersections (Lord and Persaud 2000 and 2004, Abdelwahab and Abdel-Aty 2001, Chatterjee, et al. 2003, and Greibe 2003)
- Urban non-signalized intersections (Chatterjee, et al. 2003 and Greibe 2003)
- Rural signalized intersections (Harwood, et al. 2000, Lord and Persaud 2000 and 2004, Abdelwahab and Abdel-Aty 2001, and Chatterjee, et al. 2003)
- Rural non-signalized intersections (Harwood, et al. 2000, and Chatterjee, et al. 2003)

Most of the researchers during this time period used the non-linear relationship among the crash and the traffic volume as well as some related factors. A few researchers used fairly advanced approaches such as neural network (Abdelwahab and Abdel-Aty 2001) and nonparametric technique (Golob, Recker, and Alvarez 2004). It was found that more researchers were of the need for crash prediction models specifically for planning applications after the passage of TEA-21 (Chatterjee, et al. 2003, Guevara, Washington and Oh 2004, Kononov and Allery 2004, and Schwetz, Reiff, and Chatterjee 2004). The search for significant variables to be included in crash prediction models was typically motivated by the need to improve the accuracy of models. However, the disadvantages of including many variables in the models for predicting crashes for long-range planning, as pointed out by Chatterjee, et al. (2003), are that the explanatory (or independent) variables used by these models are not always easy to forecast and thus the application of those models is difficult.

With regard to the use of simple crash rates Chatterjee, et al. (2003) pointed out the advantages of the cross-classification table is that it avoids the formulation of mathematical equations and the requirement of several independent variables. On the other hand, they were also aware of the downside of the cross-classification table since the independent variables need to be grouped into categories using specific range of values and the strength of correlation between each independent variables and crash rates, the dependent variables, is not usually determined unless separate statistical techniques are used.

Model Structure for Planning Level Applications

Past attempts at developing crash prediction models for long-range planning applications were found to have used two different approaches. The common approach is to estimate crashes on highway facilities based on measures of a traffic exposure, which represents the traditional procedure for highway safety analysis (Lau, May, and Smith 1989, Chatterjee, et al. 2003, Kononov and Allery 2004, and Schwetz, Reiff, and Chatterjee 2004). The other approach is to correlate crashes with demographic and socioeconomic characteristics of each zone in a study area (Frantzeskakis, Assimakopoulos, and Kindinis 1994; and Guevara, Washington, and Oh 2004). The common link among these two approaches is that the demographic and socioeconomic variables used for zonal level models are related to a travel demand or level of the traffic exposure used for models at the highway facility level. Both modeling approaches are considered to be useful for planning applications. However, the scope of their applications is different as they help assess different types of alternatives. Highway planners commonly deal with alternatives at the facility level and therefore, the first approach seems to be more appropriate for highway planning. The second approach at the zonal level would be useful for assessing land use alternatives.

Findings for Distribution of Crashes

Most of previous researchers found that crashes were non-linearly related to the traffic volume, and thus non-linear modeling approaches with non-normal error assumption were more appropriate in developing crash prediction models. Jovanis and Chang (1986) gave a clear illustration presented in Figure 2-1 that not only a crash frequency increased with increasing vehicle miles traveled (VMT), but the variance of the crash frequency was also found to increase with increasing VMT. Jovanis and Chang (1986) determined that the relationship between the number of crashes and VMT violated the homoscedasticity assumption, for which error terms are assumed to have equal variances for the entire range of the predictor variable, of a linear regression. The linear regression approach normally requires random variables to satisfy a normal distribution pattern known as Gaussian pattern, for which median, mean and mode are likely to be very close (Khan, Shanmugam, and Hoeschen 1999).

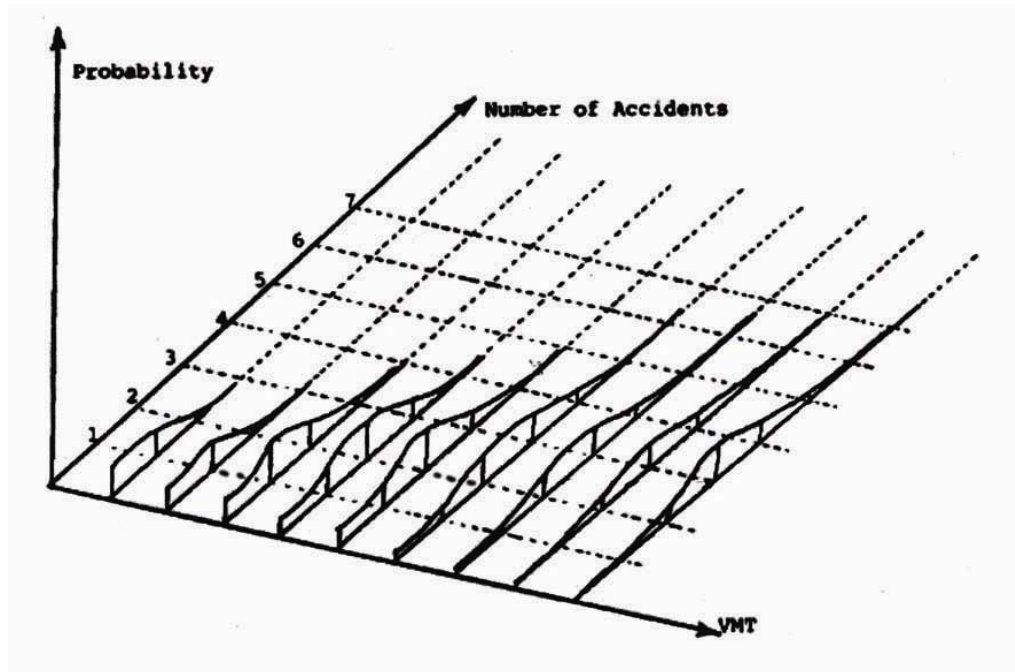


Figure 2-1 Relationship between Crashes and Vehicle Miles Traveled (VMT)

In addition, the equal variance (of crashes) assumption in the traditional linear regression approach may result in estimating negative number of crashes at a low level of VMT, which makes the models generally fail at low traffic volumes. However, they suggested that the violation of the assumption of equal variances of error terms did not affect the parameter estimation, but it altered the correct confidence intervals of the estimators and invalidated any hypothesis tests concerning the significance of estimated parameters.

Lord and Persaud (2000) recommended that the coefficients of the mathematical form of crash prediction models should not be estimated by traditional ordinary least squares estimation method since its assumption was clearly violated by the discrete, non-negative nature of the crash occurrence, non-normally distributed error structure, and non-constant variance in the number of crashes. The use of Poisson regression approach and the maximum likelihood estimation method were introduced to overcome these limitations of traditional linear regression approach and least squares estimation method (Jovanis and Chang 1986, Bonneson and McCoy 1993, Kraus, et al. 1993, Maher and Summersgill 1996, Abdel-Aty and Radwan 1999, Sayed and Rodriguez 1999, Ivan, Wang, and Bernardo 2000, Greibe 2003, and Saccomanno and Fu 2003). It also has been observed in recent researches that the crash occurrence should be considered as a negative binomial random variable (Maycock and Hall 1984, Hauer, Ng, and Lovell 1988, Maher and Summersgill 1996, and Sayed and Rodriguez 1999, Ivan, Wang, and Bernardo 2000, and Maze, et al. 2005). The negative binomial error specification follows on assumptions that crash counts are gamma distributed and follow the Poisson probability. The negative binomial distribution is considered to be more appropriate for the crash occurrence because it is found that the variance of the observed number of crashes is generally larger than the expected value, and this known as an over-dispersion phenomenon. This indicates that the negative binomial distribution assumption is more realistic than pure Poisson distribution assumption. However, Greibe (2003) suggested that the over-dispersion incident does not influence the estimation of coefficients of parameters, but it causes the standard errors to be underestimated.

Development of Crash Prediction Models

Crash prediction model formulations that have been commonly used by most of the researchers are derived based on basic concepts of crash potential and risk. Ivan (2004) gave a clear explanation of the crash risk (R) which is defined as the number of expected crashes per unit of a traffic exposure in a given time period, or it can be equivalently defined as the number of events (N) divided by the number of trials (T) where N represents the number of crashes and T represents an exposure measure. In some cases, T may be an aggregated traffic volume (V) over the analysis period, and in some cases the traffic volumes are multiplied by corresponding segment lengths to determine vehicle miles traveled (VMT). The classic representation of crash risk/crash rate is shown in Equation 2-1.

$$R = \frac{N}{T} \quad (2-1)$$

where R is commonly called “crash rate” or “crash risk”

For crash prediction models, the dependent variable can be either crash rate (R) or number of crashes (N). When the crash rate is used as the dependent variable, the exposure measure (VMT or number of vehicles), which is used to develop the rate, is incorporated in the independent variable. However, if the number of crashes is used as the dependent variable, the crash rate equation rearranges to the terms as shown in Equation 2-2.

$$N = RT \quad (2-2)$$

This new model form deals with a shortcoming that it has to assume the crash rate (R) to be constant and imply a linear relationship between the number of crashes and the exposure measures. This hypothesis is questionable. Maycock and Hall (1984), Jovanis

and Chang (1986), Bonneson and McCoy (1993), and Ivan (2004) suggested that a relationship between the number of crashes and the traffic volume is non-linear.

Majority of researchers used the number of crashes N as the dependent variable and modified the linear prediction model form to one permitting non-linear relationship between the number of crashes and the exposure measure as presented in Equation 2-3.

$$N = T^{\beta} e^{x'\beta} \quad (2-3)$$

where $e^{x'\beta} \cong R$ and $T =$ either V or VMT

In Equation 2-3, the crash rate or crash risk (R) is represented by an exponential function of road characteristics x' multiplied by a coefficient of parameter (β). Common road characteristics that have been commonly considered include lane width, median width, shoulder width, type of pavement, degree of curvature, and type of terrain. The exposure measure (T) is considered in the form of either traffic volume (V) or vehicle miles traveled (VMT) depending on whether the model is for intersection crashes or segment crashes. Equation 2-3 represents a non-linear relationship between the number of crashes and the traffic volume. This model form is found in recent crash prediction modeling (Hauer, Ng, and Lovell 1988, Lau, May, and Smith 1989, Persaud 1991 and 1994, Bonneson and McCoy 1993, Persaud and Dzbik 1993, Maher and Summersgill 1996, Mountain, Fawaz, and Jarrett 1996, Ivan, Khan, Shanmugam, and Hoeschen 1999, Sayed and Rodriguez 1999, Lord and Persaud 2000 and 2004, Ivan, Wang and Bernardo 2000, Konduri and Sinha 2002, Chatterjee, et al. 2003, Greibe 2003, Hauer, Saccomanno, Fu, and Roy 2003, Hauer, Council and Mohammedshah 2004, Ivan 2004, Kononov and Allery 2004, and Maze, et al. 2005).

Suggested Model Variables and Model Forms

When developing crash prediction models for crashes by types of severity, Chatterjee, et al. (2003) experienced difficulty with developing a model for fatal crashes and suggested

that combining fatal and injury crashes for a single model should be appropriate for planning level applications. Khan, Shanmugam, and Hoeschen (1999) found that Poisson regression with logarithmic transformation of predictor variables performed well for all levels of severity except for fatal crashes. They also found that an exposure measuring predictors, i.e., vehicle miles traveled, was significant in the fatal crash prediction model whereas segment lengths and traffic volumes were found to be significant in crash prediction models for injury crashes and property-damage-only crashes.

The review of various models showed that the traffic volume was a significant variable in every type of crash prediction models. The accuracy of these crash prediction models was improved by adding other variables such as roadway characteristics, weather conditions, and other related factors into the models. Ivan (2004) further investigated and found that the number of crashes varied with traffic volumes whether they were represented as exposure counts (representing number of trials), flow rates (representing traffic intensity), or volume-to-capacity ratios (representing level of services).

The classic crash prediction model form permits non-linear relationship between the number of crashes and the traffic volume and the model presented by Ivan (2004) expresses this form. This model is shown in Equation 2-4.

$$N = V^{\beta} e^{x'\beta} \quad (2-4)$$

N = number of crashes

V = traffic volume

β = estimated exponent of traffic volume

$e^{x'\beta}$ = crash risk or exponential function of road characteristics x' and vector of parameters β

x' = road characteristics such as lane width, median width, shoulder width, degree of curvature, and grade

Although this simplified model form has been modified and commonly used in crash prediction modeling, this model form may not very applicable for long-range planning since it still requires a consideration on some detailed road characteristics for the future roadways, which are normally not specified in a long-range plan. For example, Harwood, et al. (2000) developed models for predicting the number of non-intersection crashes on two-lane rural highway segments using independent variables such as average daily traffic, land width, shoulder width, driveway density, roadside hazard rating, horizontal curve lengths and radii, vertical curve lengths and grade. Another set of models developed by Harwood, et al. (2000) for predicting intersection crashes on rural highways included average daily traffic, roadside hazard rating, exclusive right turn lane, intersection angle, driveways on approaches, protected phases of traffic signals, left turn percentage, grades, and percent trucks. More applicable forms for crash prediction models for long-range planning are models for predicting crashes on segments developed by Persaud (1991 and 1994), Persaud and Dzbik (1993), Lord and Persaud (2004), and Maze, et al. (2005). Another applicable model form for predicting crashes at intersections was developed by Hauer, Ng, and Lovell (1988), Bonneson and McCoy (1993), Sayed and Rodriguez (1999), and Lord and Persaud (2000 and 2004). The independent variables involved in their models are average daily traffic (ADT) and a length of sections. These model forms are presented in Equation 2-5 through Equation 2-7.

$$E(m) = a(L)(ADT)^b \quad (2-5)$$

$$E(m) = a(L)^{b_1} (ADT)^{b_2} \quad (2-6)$$

$$E(m) = a(F_1)^{b_1} (F_2)^{b_2} \quad (2-7)$$

$E(m)$	= expected number of crashes
L	= section length
ADT	= traffic volume in both directions
F_1 and F_2	= entering traffic volume of major and minor roads in ADT
a , b , b_1 , and b_2	= estimated parameters depending on road classes

Although there is the lack of theory to suggest the form for crash prediction models for long-range planning applications, Equation 2-5 and Equation 2-6 are common for models for estimating crashes for segments while Equation 2-7 is common for models for estimating crashes for intersections. The independent variables required by these models are known from travel demand forecasting models used in the traditional planning process, and the classes of future roadways are also known.

Crash Prediction Models Developed for Freeways

Since this study's primary interest is limited to crash prediction for freeways, research studies involving freeway crash prediction models performed by other researchers was examined carefully. This section presents a review of these studies.

Fazio, Holden, Roupail (1993) developed a tool for assessing crash rates in weaving sections of freeways. They proposed an idea on using conflict rates in weaving sections as an indicator of crash rates. By relating crash rates, which were calculated using traditional crash rate analysis procedures, to conflict rates, which were estimated by using simulation program, they found that the conflict rate had a significant correlation with the crash rate and used the conflict rate as a measure of level of service as well as level of safety of freeway weaving sections. The simulation approach as used in their study seems to be promising for estimating the crash rate for future facilities. However, details of future freeway networks, which are required in the process, are not available in long-range plan alternatives.

Kraus, et al. (1993) examined urban freeway crash rates by collision type, e.g., rear-end and run-off-the-road, and their relation with features of urban freeways. Their modeling approach included both bivariate and multivariate analyses with independent variables such as physical characteristics of freeways, time of day, and traffic flow. The multivariate analyses were based on nonlinear modeling and assumed Poisson distribution for crashes. It is obvious that these researchers were trying to identify the

related factors that might be altered to improve a freeway safety. Their crash prediction models were developed for urban freeway sections regardless of their location in relation to interchanges.

Persaud and Dzbik (1993) developed crash prediction models for freeway sections at both macro level, crashes per unit length per year, and micro level, crashes per unit length per hour using a generalized linear modeling approach with a negative binomial error structure for the dependent variable. They noticed that existing models were likely to be macroscopic since the average daily traffic rather than specific flow rate at the time of crashes were commonly included in the models. They pointed out the weakness of macroscopic models that the models could not clearly capture the difference in crash potential between freeways with the intense traffic flow during peak hours and freeways with the same average daily traffic, but with the flow evenly spread out during the day. Their studies confirmed that the crash pattern on freeway sections during congested periods and that during uncongested periods were different.

Shankar, Mannering, and Barfield (1996) developed models for crashes by severity using a nested logit approach, which included both single and interaction terms of predictor variables. The variables included in their model were geometric features, weather conditions, pavement surface conditions, vehicle conditions, and driver related factors. Their model was developed based on evenly distributed freeway segments of equal distance and did not distinguish between non-interchange segments and interchange segments. These models are applicable in situations given that crashes have occurred. The limitation of the model applications and the use of some independent variables requiring detailed data made this modeling approach difficult to be used in long-range transportation planning.

Resende and Benekohal (1997) developed a crash prediction model for rural freeways based on a volume-to-capacity ratio. Their modeling technique is the multiple linear

regression analysis with the number of crashes per lane per mile as the dependent variable and the daily volume-to-capacity ratio as well as a few other variables such as median width and surface rating, as independent variables. They found that crash rates could be correlated with either traffic volumes or daily volume-to-capacity ratios. Their results suggested that crash prediction models should include the capacity as an independent variable. In addition, they found that a percentage of heavy vehicles did not vary proportionally with annual average daily traffic volume, and so the use of an adjustment factor for heavy vehicles were suggested to be essential. The capacity of freeways was considered to be a crucial variable in their model. However, it should be pointed out that the capacity may be difficult to estimate for future freeways since their geometric features such lane width, shoulder lateral clearance and grade are not usually selected in a long-range transportation planning study although the number of lanes are known. Actually, the number of lanes of a freeway reflects its capacity in a crude way.

Khan, Shanmugam, and Hoeschen (1999) explored relationship of crash frequencies stratified by types of severity with traffic volume, segment length, and vehicle miles traveled using several methods of non-linear regressions. Their findings suggested that Poisson regression with logarithmic transformation of predictor variables performed better for all types of severity except for fatal crashes. The segment length and the traffic volume were found to be significant in the models predicting injury crashes and property-damage-only crashes while the exposure-based predictor (vehicle miles traveled) was found to be significant in models predicting fatal crashes. The freeway segments of their study included segments from both inside and outside interchange influence areas, their work did not separately investigate the effects of interchanges on crash frequencies.

Konduri and Sinha (2002) developed crash prediction models for freeway segments between interchanges and these segments covered the lengths from the center of one interchange to that of an adjacent interchange. The predictor variables included in their models were traffic volume, geometric characteristics, and weather conditions. Crash

rates and the number of crashes were predicted by time of the day using the non-linear modeling approach. They found that both the methods of least sum of squares and maximum likelihood estimation provided nearly the same coefficient values for the models. They interestingly found that the traffic volume and the section length could not be combined as a single variable (vehicle miles traveled) in crash prediction modeling. One limitation of this work is that the freeway segments of their study included some effects of interchanges on both ends and that there was no separate treatment of the effects of interchanges on crashes.

Chatterjee, et al. (2003) developed crash prediction models specifically for applications in long-range transportation planning. Their modeling methods include cross-classification tables and linear regression analysis. The cross-classification table was considered to have advantage over the regression approach in that it was easier to be interpreted and it avoided mathematical formulation of equations. Their regression models used the logarithmic transformation of variables to express a non-linear relation in a linear form. The crash prediction models for freeways were developed for segments that stretched from the center of one interchange to that of another. Their model did not treat interchange related crashes and segment related crashes separately.

Golob, Recker, and Alvarez (2004) developed a tool for real time assessment of the level of safety on freeways in terms of crash types, crash locations, and severity of crashes. The clustering technique was used to categorize real time traffic data, collected from inductance loop detectors, into homogeneous groups of traffic flow conditions, which were statistically analyzed to describe their relationships with crash typology. Their models were purposely developed for crash surveillance on freeways and required real time traffic flows from detectors as an input, and this approach is not meant for long-range planning.

Kononov and Allery (2004) developed safety-based performance standards for urban freeway segments based on crash prediction models using generalized linear modeling approach. They developed crash prediction models using a non-linear function between a crash rate as the dependent variable and an annual average daily traffic as one of independent variables. The scope of their study is similar to this research. However, their models were developed based on data for urban six-lane freeways only. Each segment for their model included one interchange and the distance up to the mid-points of the sections between adjacent interchanges. This is illustrated in a conclusion of this literature review.

None of the studies involving crash prediction models for urban freeways, which this literature search identified, treated interchange related crashes separately from those that occur away from interchanges. Intuitively, this appears to be a weakness of the past approach and this study would attempt to fill this gap in the state of the practice.

Summary of Findings from Literatures

The attempts to develop crash prediction models for certain types of highway facilities were made since a long time. Several of them involved freeways while the majority dealt with non-freeways. Initially, the modeling efforts used the traditional linear regression approach. However, it was found later that the non-linear modeling approach is more appropriate. Developing accurate prediction models for crashes was difficult due to several reasons – a discrete nature of crashes, a non-negative nature of crash occurrences, a non-normal distribution of errors, and a non-constancy of variances. These conditions were not compatible with the assumptions required for ordinary least squares regression, which requires an error term to be normally distributed and a variance to be constant. The use of Poisson regression approach was introduced to overcome these limitations. However, crash prediction modeling based on Poisson distribution still suffered from violating the requirement that the count data should have its mean equal to its variance. It was found that crash occurrences generally have variances larger than expected means. This condition was known as an over-dispersion incidence, which would influence the

estimation of standard errors, but not the estimation of model coefficients. Due to this reason, the negative binomial distribution was considered to be more appropriate and used in many of the previous studies (Hauer, Ng, and Lovell 1988, Persaud 1991 and 1994, Persaud and Dzbik 1993, Mountain, Fawaz, and Jarrett 1996, Sayed and Rodriguez 1999, Lord and Persaud 2000 and 2004, Guevara, Washington, and Oh 2004, Hauer, Council, and Mohammedshah 2004, and Maze, et al. 2005). The negative binomial distribution assumption was commonly substituted to Poisson distribution because the negative binomial allows variance to differ from mean, which relaxes the limitation required by Poisson distribution.

For previous studies dealing with crash prediction models for freeways, not only statistical approaches were found to be different, but the model organization and the freeway segmentation were also found to be different. Some researchers developed models for freeway segments which were defined as segments between two successive interchanges and included one-half of the interchanges at the two ends on the segment (Persaud and Dzbik 1993, Konduri and Sinha 2002, and Chatterjee, et al. 2003). Some developed models for freeway segments which included interchanges in full (Kraus, et al. 1993 and Kononov and Allery 2004). Some developed models for crash rates in weaving areas (Fazio, Holden, and Roupail 1993). Some developed models for assessing crashes at specific locations on freeway segments (Golob, Recker, and Alvarez 2004). And some developed models for freeway segments without considering the influence of interchanges (Shankar, Mannering, and Barfield 1996). Figure 2-2 illustrates the schemes of segmentation found in previous crash prediction models for freeways.

It should be pointed out that none of the previous studies developed separate crash prediction models for two different areas on freeways – non-interchange segments and interchange segments. It is learned from literatures that the idea of developing different models for non-interchange segments and interchange segments should be pursued.

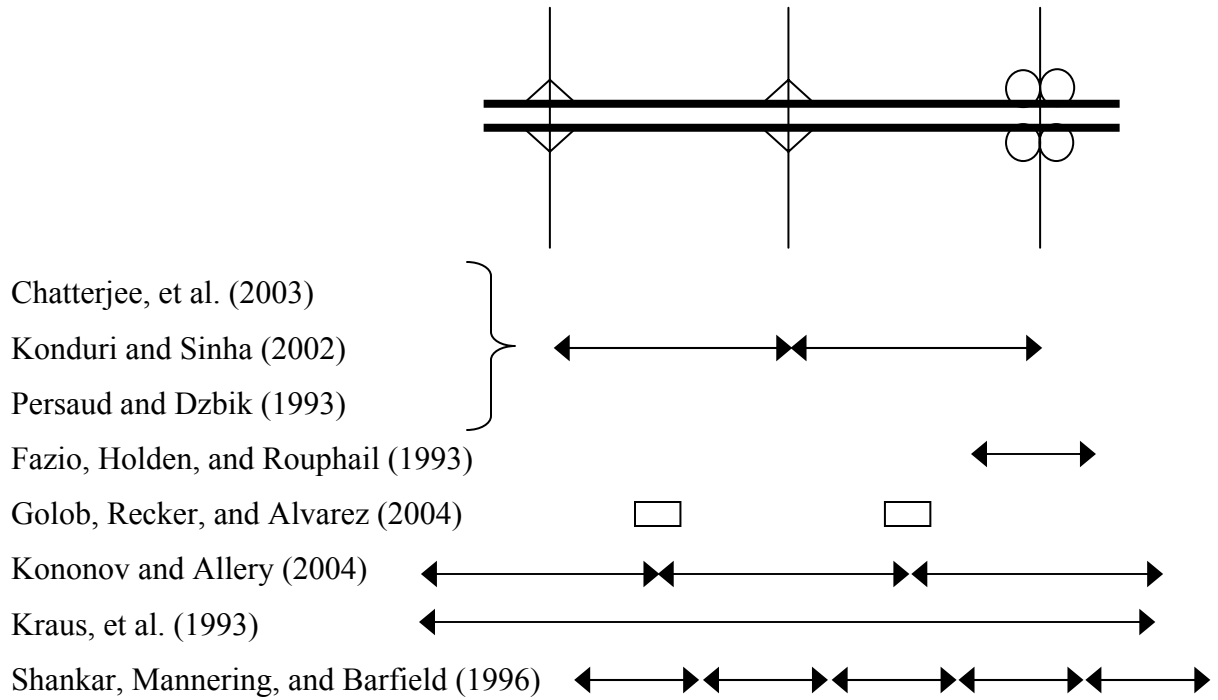


Figure 2-2 Freeway Segmentation Schemes Found in Previous Studies

3. STUDY METHODOLOGY

This chapter presents the conceptual details of various tasks performed in this study. Three main tasks are: data preparation, model development, and model assessment and comparison. The actual analyses performed and the results are presented in subsequent chapters.

Data Preparation and Preliminary Analysis

For the purpose of statistical analysis for model development, data from freeway inventory files and freeway crash records for selected urban areas of North Carolina and Tennessee were processed by matching the milepost and the segment identification number included in freeway inventory data tables and crash data tables respectively. Two types of tables – 1) segment based tables and 2) individual crash based tables – are needed for modeling. Segment based tables include crash counts for each freeway segment and for each interchange area along with their physical characteristics and traffic volume counts. Segment based tables are further grouped into non-interchange segments and interchange segments. These separate data sets are needed for modeling freeway segments inside and outside interchange influence areas respectively. Crashes occurring on freeway segments within approximately 1,500 feet from the middle of interchanges are considered as interchange freeway crashes. Crashes occurring on freeway segments beyond 1,500 feet from the middle of interchanges are considered as non-interchange freeway crashes.

Data included in segment based tables are to be used for regression and analysis of variance models. The data tables that include individual crash records are needed for the classification tree models.

Model Development

Since the goal of this research is to develop application oriented models, the model formulation process was designed to recognize issues involving the availability of data for the independent variables, difficulties of making forecasts for the independent variables, and the use of the model in the context of urban transportation planning. Safety consequences usually are incorporated in the alternative evaluation process as costs expressed in monetary terms. In transportation planning, each crash is assigned a cost value in terms of cost per incident and/or cost per person based on the severity level of crash as described in Manual of Transportation Engineering Studies (Robertson, Hummer, and Nelson, 2000), and for this purpose, three categories of severity of crashes are commonly used:

- Fatal (F)
- Injury (I)
- Property Damage Only (PDO)

Crash prediction models are expected to be developed separately for North Carolina urban freeways and Tennessee urban freeways respectively. Two main categories of crash prediction models are to be developed – one for non-interchange segments and the other one for interchange segments. Each of them is further grouped based on the number of freeway lanes – four-lane and more-than-four-lane. Further, the model development was attempted for each crash type by severity. However, it is learned from previous studies that developing a prediction model for ‘fatal’ crashes is difficult. Chatterjee, et al. (2003) suggested that developing a model for ‘fatal and injury’ crashes combined can overcome the difficulty. Therefore, developing prediction models was mainly attempted for three crash types – ‘injury’, ‘PDO’, and ‘fatal and injury’ combined. Table 3-1 summarizes the model categories that were considered for developing crash prediction models for North Carolina and Tennessee urban freeways.

Table 3-1 Categories of Crash Prediction Models for Urban Freeways
(N = North Carolina and T = Tennessee)

Model Categories	Freeway Lane Categories	Models for Injury	Models for PDO	Models for Fatal & Injury
Non-Interchange Freeways	Four Lanes	N&T	N&T	N&T
	More than Four Lanes	N&T	N&T	N&T
Interchange Freeways	Four Lanes	N&T	N&T	N&T
	More Than Four Lanes	N&T	N&T	N&T

Figure 3-1 illustrates the coverage areas used for crash prediction models for non-interchange segments and interchange segments. Non-interchange segments are considered as segments between two successive interchanges with the back-off distance of approximately 1,500 feet from the middle of interchanges on both ends of segments. It is assumed that crashes occurring within non-interchange segments are associated with a traffic pattern along the segments with full access control. Interchange-freeway-segment crashes are crashes occurring on freeway lanes within 1,500 feet on either side from middle of an interchange.

Three different approaches including the analysis of variance model, the regression model, and the classification tree model were utilized in developing crash prediction models for urban freeways for each state using data from corresponding areas. Each approach is to be examined for appropriate crash prediction models for non-interchange segments, and interchange segments. The details of these approaches are discussed in the following sections.

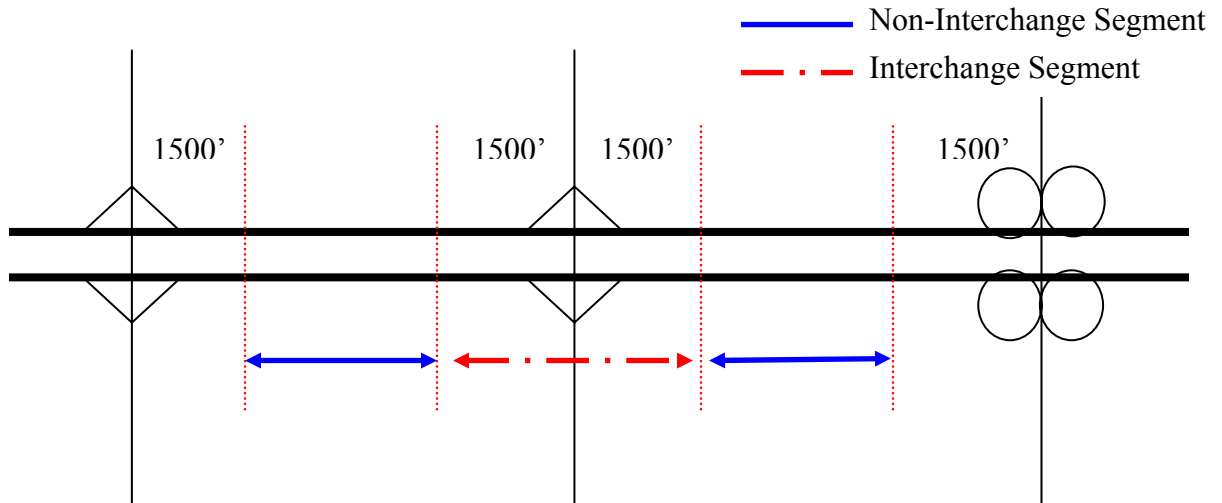


Figure 3-1 Coverage Areas for Non-Interchange Segments and Interchange Segments

Analysis of Variance Approach

Analysis of variance (ANOVA) technique was applied to develop prediction models, which are used for making a prediction in terms of crash rate. The crash rate is expressed as the number of crashes per million vehicle miles traveled for freeway segments.

For developing ANOVA models for both non-interchange segments and interchange segments, the independent variables are freeway traffic volume, type of freeway segments, and number of freeway lanes. The freeway traffic volume is classified into five or six ranges increasing at 25,000 vehicles per day. The type of segments is classified into non-interchange segments and interchange segments. The number of lanes is classified into four-lane and more-than-four-lane. Separate ANOVA models were developed for each crash type by severity – ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes.

Regression Analysis Approach

Because the number of crashes is a count variable, assumptions for linear regression based on the ordinary least squares (OLS) method are found to be invalid. Jovanis and Chang (1986) pointed out the following characteristics of crashes that violate the assumptions of OLS regression:

- Non-constant variances of crashes with increasing traffic volumes (heteroscedasticity)
- Non-negativity of crash occurrences
- Non-normally distributed error terms

A non-linear regression approach or generalized linear regression approach is common for developing crash prediction models. Investigations by Maycock and Hall (1984), Jovanis and Chang (1986), and Hauer, Ng, and Lovell (1988) suggested that crashes are non-linearly related with predictor variables such as a traffic volume, which is the major predictor variable in most of the previous studies. A multiplicative form of model was considered in order to yield valid parameter estimates. Persaud (1991 and 1994) and Lord and Persaud (2004) suggested two model forms for predicting crashes on segments or links as shown in Equation 2-5 and Equation 2-6.

Several variations of the above suggested model forms are created and analyzed for identifying the most appropriate set of models. Table 3-2 presents the model forms of crash prediction models for non-interchange segments and interchange segments that are used for developing regression models. Only a multiplicative form using traffic volumes and other independent variables and/or exponential terms of independent variables is applicable. The multiplicative model form is to be used to predict no crashes when no traffic exists on freeway segments. The maximum likelihood estimation method was used to estimate all parameters in the regression models. It is found from the literature review that the maximum likelihood estimation is common when the probability distribution of the error term is specified to be either Poisson or negative binomial distribution.

Table 3-2 Regression Model Forms for Non-Interchange Segments and Interchange Segments

Regression Model Forms for Non-Interchange and Interchange Segments
$\mu = (\text{AADT})^b$
$\mu = a (\text{AADT})^b$
$\mu = (\text{Segment Length}) (\text{AADT})^b$
$\mu = a (\text{Segment Length}) (\text{AADT})^b$
$\mu = (\text{Segment Length})^{b1} (\text{AADT})^{b2}$
$\mu = a (\text{Segment Length})^{b1} (\text{AADT})^{b2}$

The assumption of negative binomial distribution is used because it allows greater variance in the crash data and can overcome the over-dispersion phenomenon, which is likely to arise with crash data.

Because of the availability of good statistical software, negative binomial regression is of interest and performed in the regression analysis. The statistical software package, SAS, was used for estimating parameter coefficients of each model in Table 3-2. The GENMOD procedure with specified option DIST = NB in the MODEL statement produces maximum likelihood estimates of parameters coefficients using negative binomial distribution. The negative binomial probability distribution of the response ‘Y’, where ‘y’ represents an observed value of the response ‘Y’, ‘μ’ represents a mean of ‘Y’, and ‘k’ represents a dispersion parameter, is available in the GENMOD procedure as presented in Equation 3-1 (SAS Institute Inc., 1999).

$$f(y) = \frac{\Gamma\left(y + \frac{1}{k}\right)}{\Gamma(y+1)\Gamma\left(\frac{1}{k}\right)} \frac{(k\mu)^k}{(1+k\mu)^{y+\frac{1}{k}}}, \text{ for } y = 0, 1, 2, \dots \quad (3-1)$$

μ = Mean of Y

k = Dispersion

Variance function of Y or $\text{Var}(Y) = \mu + k\mu^2$

It is noted from the variance function of 'Y', $\text{Var}(Y)$, that the negative binomial distribution is reduced to Poisson distribution when a dispersion parameter 'k' approaches to zero. A zero value of a dispersion parameter 'k' will indicate an equality condition between the variance and the mean of response 'Y'. The specified log-likelihood function in SAS for negative binomial distribution is presented in Equation 3-2.

$$l_i = y\log(k\mu) - \left(y + \frac{1}{k}\right)\log(1 + k\mu) + \log\left(\frac{\Gamma\left(y + \frac{1}{k}\right)}{\Gamma(y+1)\Gamma\left(\frac{1}{k}\right)}\right) \quad (3-2)$$

The Γ function represents a gamma function and can be defined that $\Gamma(y+1) = y!$. The parameter 'k' is the negative binomial dispersion parameter that is estimated during the process. By default, the GENMOD procedure estimates the parameter 'k' using maximum likelihood technique (SAS Institute Inc., 1999). Researchers at the University of Kentucky (Social Sciences Teaching and Research Statistics) suggested the procedure to fit negative binomial regression in SAS using a log link function to create a log of the mean (μ), which is assumed to be a linear function of the independent variables. This log link function is presented in Equation 3-3.

$$\log(\mu) = \text{intercept} + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_nX_n \quad (3-3)$$

The log link function in Equation 3-3 was modified for each of model forms presented in Table 3-2. For example, the link function for the model with the variable ‘AADT’ and the parameters ‘a’ and ‘b’ is shown in Equation 3-4.

$$\log(\mu) = \log(a) + b \log(\text{AADT}) \quad (3-4)$$

The goodness-of-fit of a negative binomial model, which is an indicator of how well the models can perform on predicting, is reported in the form of deviance statistics. McCullagh and Nelder (1989) pointed out that a deviance function behaves in the same way as the residual sum of squares in ordinary linear models and it can be used for indicating the adequacy of the fitted model. The deviance for the negative binomial distribution is calculated during the estimation process in GENMOD procedure by the Equation 3-5 (SAS Institute Inc., 1999).

$$\text{Deviance} = 2 \sum_i \left[y \log\left(\frac{y}{\mu}\right) - \left(y + \frac{1}{k}\right) \log\left(\frac{y + \frac{1}{k}}{\mu + \frac{1}{k}}\right) \right] \quad (3-5)$$

The rule of thumb indicates that if the values of Deviance divided by the degrees of freedom are close to one, a model can be accepted as fitting the data well (SAS Institute Inc., 1999).

Another approach for measuring the goodness-of-fit for a negative binomial model was developed by Miaou (1996). This technique is based on a dispersion parameter ‘k’ and can be used equivalently to the estimation of a coefficient of determination (R^2). The indicator based on the parameter ‘k’ is calculated using the Equation 3-6.

$$R_k^2 = 1 - \frac{k}{k_{\max}} \quad (3-6)$$

The term 'k' is the dispersion parameter estimated in the negative binomial model and another term ' $k_{\max}$ ' is the dispersion parameter estimated in the same negative binomial model with only an intercept term and a dispersion parameter. This measure technique is helpful in estimating a percentage of variation explained by each regression model.

Classification Tree Analysis Approach

Classification tree modeling was performed to examine an alternative form of a crash prediction model. The tree model is interesting because it can be interpreted easily and it functions similarly to a cross-classification table, which is used commonly in planning applications. In addition, one appealing feature of the classification tree approach is that there are no distribution assumptions. Unlike the cross-classification table in which the categories are specified, the classification tree model will automatically determine the most appropriate splits or threshold values of each variable at each level to stratify the data into different categories. As explained by Neville (1999), the tree model divides a set of data into groups. Starting from a root, the tree splits the entire data into a number of branches. Each branch or 'parent node' may be either left alone as a final group or further split into 'child nodes'. Tree leaves are defined as those 'parent nodes' without 'child nodes'.

The typical algorithm of tree models used in the statistical community is known as Classification and Regression Trees (CART). The CART algorithm is a binary decision tree algorithm introduced by Breiman et al. (1984). The CART tree is created by splitting data, starting from the entire data set, using all desired predictor variables to repeatedly generate two 'child nodes' under each of 'parent nodes'. The goal is to produce child nodes which are as homogeneous as possible with respect to the target variable. The target or dependent variable used for this study is number of crashes designated by severity types. Each split is determined by maximizing the decrease in impurity of crashes by severity. All the predictor variables are evaluated and compared, and the one

with the best improvement is selected for each split. The measure of impurity is represented as a Gini index, which can be calculated using Equation 3-7.

$$g(t) = 1 - \sum_j p_j^2(t) \quad (3-7)$$

The term $p_j(t)$ in Equation 3-7 is shown in Equation 3-8 as the fitted probability of each crash severity type in node 't' or the ratio between number of crashes for each severity type 'j' (denoted as $n_j(t)$) and the total number of crashes in each node 't' (denoted as $n(t)$). The range of Gini index ($g(t)$) varies from maximum value of $1-1/k$, where 'k' is the number of severity types, to minimum value of zero, when there is a single severity type.

$$p_j(t) = \frac{n_j(t)}{n(t)} \quad (3-8)$$

The improvement in the Gini index for a split 's' at node 't' is calculated using Equation 3-9.

$$l(s, t) = g(t) - p_L g(t_L) - p_R g(t_R) \quad (3-9)$$

The subscripts 'L' and 'R' indicate the left node and the right node respectively. The split 's' that obtains the maximum improvement is selected.

Figure 3-2 illustrates an example of 1-level CART tree algorithm with three crash severity types. The first best split in the Figure 3-2 is based on the predictor variable traffic volume (AADT). The tree separates crashes into a group of AADT less than or equal to 50,000 vehicles per day and a group of AADT greater than 50,000 vehicles per day. The Gini indexes of node 0, 1, and 2 in the Figure 3-2 are calculated using Equation 3-7 and Equation 3-8 and their results are presented as follows:

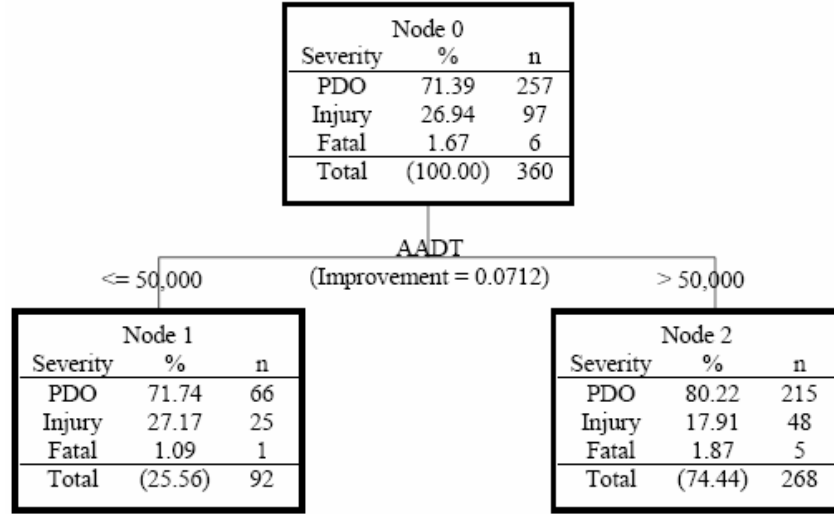


Figure 3-2 An Example of 1-Level CART Tree

$$g(\text{node } 0) = 1 - \left[\left(\frac{257}{360} \right)^2 + \left(\frac{97}{360} \right)^2 + \left(\frac{6}{360} \right)^2 \right] = 1 - 0.5825 = 0.4175$$

$$g(\text{node } 1) = 1 - \left[\left(\frac{66}{92} \right)^2 + \left(\frac{25}{92} \right)^2 + \left(\frac{1}{92} \right)^2 \right] = 1 - 0.5886 = 0.4114$$

$$g(\text{node } 2) = 1 - \left[\left(\frac{215}{268} \right)^2 + \left(\frac{48}{268} \right)^2 + \left(\frac{5}{268} \right)^2 \right] = 1 - 0.6760 = 0.3240$$

The improvement in the Gini index of the first split at node 0 is calculated using Equation 3-9 and the result is as follows:

$$l(s, 0) = g(0) - p_1 g(1) - p_2 g(2) = 0.4175 - \left(\frac{92}{360} \right) (0.4114) - \left(\frac{268}{360} \right) (0.3240) = 0.0712$$

Detailed information about decision tree modeling is provided by Giudici (2003), Neville (1999), Wilkinson (1992), and Breiman, et al. (1984).

Classification tree models are developed for non-interchange segments and interchange segments separately for North Carolina and Tennessee. The CART algorithm in the statistical software package, SPSS AnswerTree, was used to generate all tree models.

Model Assessment and Comparison

The prediction models were evaluated to examine their prediction performances. The comparisons were made using a percent error criterion, which is a percent of a ratio between a root mean squared error (RMSE) and number of crashes for an average segment. The smaller the percent error indicates the better the performance of the model. RMSE is a square root of the mean squared difference between predicted value and actual value of different crash types for each segment.

Conclusions

Developing crash prediction models for long-range transportation planning is difficult due to some considerations and constraints. The modeling processes were organized and simplified into three main tasks - data preparation, model development, and model assessment and comparison.

Figure 3-3 and Figure 3-4 represents detailed processes for developing crash prediction models for non-interchange segments and interchange segments, respectively. Data of interest are the individual crash table and the segment table, which includes information such as traffic volumes and lane numbers. Information was linked using the segment identification number and the milepost information. Three modeling approaches – ANOVA models, regression models, and classification tree models – were performed. The model performance was assessed and compared in terms of the percent error.

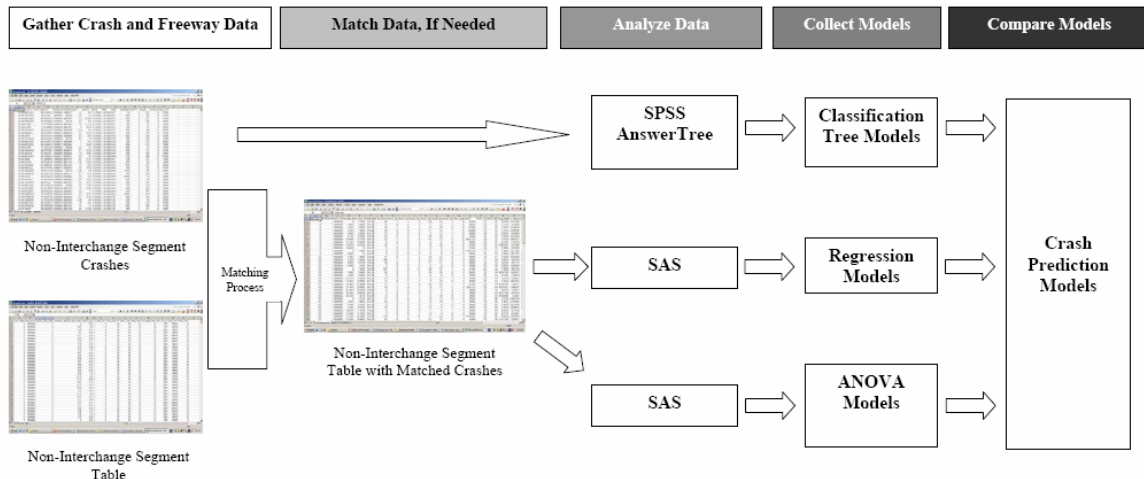


Figure 3-3 Diagram for Crash Prediction Modeling Process for Non-Interchange Segments

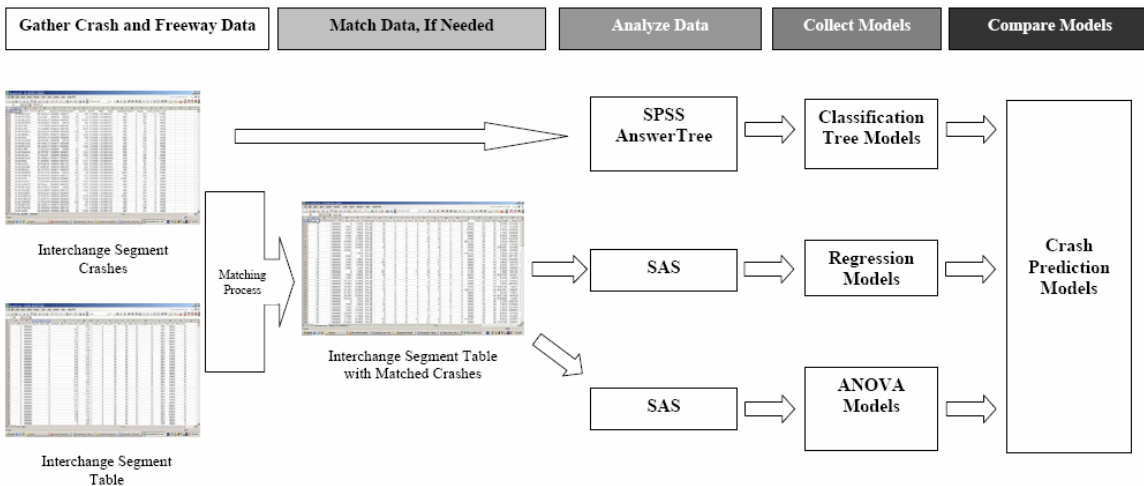


Figure 3-4 Diagram for Crash Prediction Modeling Process for Interchange Segments

4. PRELIMINARY ANALYSIS OF DATA

Crash prediction models were developed based on crash and freeway data which were obtained from North Carolina Department of Transportation (NCDOT) and Tennessee Department of Transportation (TDOT). All crashes used in the analysis are for three years. Details of crash and freeway data for both states are presented in this chapter.

Crash and Freeway Data for North Carolina

The database from NCDOT includes freeway inventory data and crash records for the years 2000, 2001, and 2002. The NCDOT database contains separate tables for different types of information. The tables for freeway segment related information includes traffic volumes, length of each freeway segment, and route identification number. The table for crash related information includes crash records for freeway segments. Different data tables were linked to each other based on a county code, a segment or interchange identification number, a milepost, and a route number.

Of all 99 counties in North Carolina, ten include large to medium size urban areas. Of these ten counties, six were selected for this study because of the richness of complete data for these counties. These six counties are Alamance, Buncombe, Cumberland, Durham, Gaston, and Wake. The summary of freeway segments in six urban counties of North Carolina is shown in Table 4-1. All freeway segments selected for this study represent 2-digit freeways. Examples of 2-digit interstate highways include I-40 and I-85. Examples of 3-digit highways, which are not included, are I-440 and I-240. For the analysis, the freeway data were classified into two types – non-interchange segments and interchange segments – based on the criteria described in chapter 3. Each type is further classified into two groups based on the number of lanes – four-lane and more-than-four-lane.

Table 4-1 Summary of Freeway Segments for Six Counties in North Carolina

Counties	Type of Freeway Segments	Number of Lanes	Number of Segments	Total Length (miles)
Alamance Buncombe Cumberland	Non-Interchange	Four	36	77.34
		More Than Four	37	59.76
		Total	73	137.10
Durham Gaston Wake	Interchange	Four	33	18.54
		More Than Four	32	18.00
		Total	65	36.54

Another set of North Carolina freeway data is related to individual crash records for crashes occurring on freeway segments. A summary of this data is presented in Table 4-2. It should be pointed out that crash severity types were coded in the original data as fatal, A-injury, B-injury, C-injury, property-damage-only, and unknown type. The crash severity types A-injury, B-injury, and C-injury were combined into a single type of injury crashes to make it similar to the classification used for Tennessee data and also to avoid excessive stratification. Furthermore, the crashes with unknown severity type were excluded from the modeling process since it provides no information for estimating safety related costs in the planning process. It is noted that crash data shown in Table 4-2 represents approximately 80% of total crashes in the NCDOT's crash database. Crash rates were calculated using the traditional approach for crash analysis for roadway segments, which is presented in Equation 4-1. The rate is in terms of crashes per million vehicle miles traveled (MVMT).

$$\text{Crash Rate} = \frac{\text{Crashes (in 3 years)} \times 10^6}{\text{Segment Length} \times \text{AADT} \times 365 \times 3 \text{ years}} \quad (4-1)$$

**Table 4-2 Freeway Crashes by Severity Type for Six Counties in North Carolina
(2000-2002)**

Crash Severity Type	Non-Interchange Segments			Interchange Segments		
	Crashes	%	Rates (Crashes /MVMT)	Crashes	%	Rates (Crashes /MVMT)
PDO	2,787	66.80	0.2805	2,989	70.25	0.9892
Injury	1,355	32.48	0.1364	1,246	29.28	0.4124
Fatal	30	0.72	0.0030	20	0.47	0.0066
Total	4,172	100	0.4198	4,255	100	1.4081

It can be seen in Table 4-2 that the proportions of PDO crashes, injury crashes, and fatal crashes for freeway segments are quite similar. PDO crashes are found to be about 70% while injury crashes are about 30% and fatal crashes are less than 1% of the total. Furthermore, PDO crash rates and injury crash rates for interchange segments are about three times as high as those for non-interchange segments. Fatal crash rates for interchange segments are about two times as high as those for non-interchange segments.

Crash and Freeway Data for Tennessee

The freeway inventory and crash data for Tennessee cover four urban counties of Knox, Davidson, Shelby, and Hamilton. Similar to the case of North Carolina, two types of data tables – the table for non-interchange/interchange segments and the table for individual crashes – are needed for developing data for modeling. The summary of Tennessee freeway segments that are used in the analysis is shown in Table 4-3.

Table 4-4 presents summaries of crash rates by crash severity types on urban freeway segments of four counties in Tennessee. The injury crash type for Tennessee data has only one combined level without separating injuries into types A, B, and C like the crash

Table 4-3 Summary of Freeway Segments for Four Counties in Tennessee

County	Type of Freeway Segments	Number of Lanes	Number of Segments	Total Length (miles)
Davidson Hamilton Knox Shelby	Non-Interchange	Four	36	45.65
		More Than Four	106	98.86
		Total	142	144.51
	Interchange	Four	15	15.05
		More Than Four	50	44.12
		Total	65	59.17

Table 4-4 Freeway Crashes by Severity Type for Four Counties in Tennessee (2000-2002)

Crash Severity Type	Non-Interchange Segments			Interchange Segments		
	Crashes	%	Rates (Crashes /MVMT)	Crashes	%	Rates (Crashes /MVMT)
PDO	7,372	71.19	0.5820	8,501	71.81	1.4462
Injury	2,949	28.48	0.2328	3,306	27.92	0.5624
Fatal	34	0.33	0.0027	32	0.27	0.0054
Total	10,355	100	0.8175	11,839	100	2.0141

data for North Carolina. All crash rates for freeway segments shown in Table 4-4 are in the unit of crashes per million vehicle miles traveled (MVMT). Similar to the findings for North Carolina, it is found for Tennessee crash data that the proportions of crash severities are similar for non-interchange segments and interchange segments. PDO crashes have the highest proportion of more than 70%. Injury crashes are nearly 30% and fatal crashes are less than 1% of the total. It is noticed that crash rates for interchange segments are higher than those for non-interchange segments.

It can be seen in Table 4-4 that PDO crash rates for interchange segments are almost three times of those for non-interchange segments. In addition, injury crash rates and fatal crash rates for interchange segments are approximately twice of those for non-interchange segments.

Comparison of Crash Rates for Freeway Segments of North Carolina and Tennessee

This section presents the comparison between crash rates for North Carolina freeways and those for Tennessee freeways. For each group of freeway segments, crashes by crash severity type were converted into crash rates using Equation 4-1 to make them comparable. The test of the hypothesis regarding the difference between mean crash rates for North Carolina freeways and those for Tennessee freeways was performed using the t-test. The null hypothesis (H_0) and the alternative hypothesis (H_a) for comparing crash rates are as follows:

H_0 : mean crash rates for North Carolina freeways and those for Tennessee are equal

H_a : mean crash rates for North Carolina freeways are not equal to those for Tennessee

The test was performed for each crash severity type and for each type of segments – non-interchange segments and interchange segments. The comparison results are presented in Table 4-5 for non-interchange segments and in Table 4-6 for interchange segments.

Table 4-5 Crash Rates for Non-Interchange Segments in North Carolina and Tennessee

Type of Crashes	Numbers of NC Non-Interchange Segments	Number of TN Non-Interchange Segments	Means of NC Non-Interchange Segments Crash Rates	Means of TN Non-Interchange Segments Crash Rates	P-Value for the Equality of Means	Equality of Means
Fatal	73	142	0.0026121	0.0034519	0.6612	Yes
Injury	73	142	0.1230622	0.2630360	<0.0001	No
PDO	73	142	0.2581848	0.6705869	<0.0001	No
Total	73	142	0.3838591	0.9370748	<0.0001	No

Table 4-6 Crash Rates for Interchange Segments in North Carolina and Tennessee

Type of Crashes	Numbers of NC Interchange Segments	Number of TN Interchange Segments	Means of NC Interchange Segments Crash Rates	Means of TN Interchange Segments Crash Rates	P-Value for the Equality of Means	Equality of Means
Fatal	65	65	0.0064741	0.0076702	0.0468	No
Injury	65	65	0.4358883	0.5257547	0.0849	Yes
PDO	65	65	0.9485421	1.3597077	0.0065	No
Total	65	65	1.3909045	1.8931325	0.0147	No

It is found in Table 4-5 that the mean rates for all crash types except ‘fatal’ crashes for non-interchange segments of North Carolina and those non-interchange segments of Tennessee are significantly different. Further, crash rates for non-interchange segments of Tennessee are found to be at least two times of those for non-interchange segments of North Carolina. Table 4-6 indicates that the rates for ‘fatal’ crashes, ‘PDO’ crashes, and ‘total’ crashes for interchange segments of Tennessee are significantly higher than those for interchange segments of North Carolina. However, these differences are found to be smaller than those for non-interchange segments.

Conclusions

Data for crashes and freeways used in this study were obtained from NCDOT and TDOT. The crash data are in three-year basis for the years 2000, 2001, and 2002. Individual crash records and freeway inventory data were available in two separated tables for each state. Crash and freeway data were linked based on information for county codes, identification numbers, mileposts, and route numbers. Data for as many as possible for freeway segments from selected urban areas of North Carolina and Tennessee were included in the analysis to overcome uncertainties of high variability of crash data and to strengthen the crash prediction models.

The investigation on crash data for North Carolina and Tennessee revealed that the proportions of crashes by severity are nearly similar for non-interchange and interchange segments for both states of North Carolina and Tennessee. Typically, PDO crashes were found to account for approximately 70% while injury crashes account for about 30% and fatal crashes were found to be less than 1% of the total. These proportions of crash severity were found consistent with those reported by NCDOT’s Traffic Safety Systems Management.

Further, it was found that the crash rates for Tennessee freeways and those for North Carolina freeways are significantly different. This finding indicated that the crash rates

are unique for each state and, therefore, crash prediction models should be developed separately for North Carolina freeways and Tennessee freeways.

5. CRASH PREDICTION MODELS BASED ON ANALYSIS OF VARIANCE APPROACH

Developing crash prediction models based on the analysis of variance (ANOVA) technique is the first of three proposed modeling approaches used in this study. ANOVA model represents a simple model form, which is easy to be interpreted and applied in the long-range transportation planning process. ANOVA models are utilized to provide values of crash rates by severity type of crashes for different freeway categories and different levels of traffic volumes. For each crash severity type, a crash rate for each non-interchange or interchange segment was calculated using Equation 4-1.

Factors or independent variables of interest are the freeway traffic volume, the type of segments and the number of lanes. Freeway traffic volumes were classified into groups increasing at an increment of 25,000 vehicles per day. Types of segments were classified into non-interchange segments and interchange segments. Numbers of lanes were classified into four-lane and more-than-four-lane.

Three-way ANOVA was utilized using the three factors – traffic volume, type of segments, and number of lanes. The statistical F-test is used as an indicator for developing conclusions on the significance of the models, effect of each factor, and interaction effect between factors used in the models.

ANOVA Models for Predicting Crash Rates for North Carolina Freeways

This section presents the results of the analysis of variance performed with North Carolina data. It is noticed that ‘fatal’ crash data are not sufficient to create a reliable model for predicting ‘fatal’ crash type alone. However, it was suggested by Chatterjee, et al. (2003) that combining ‘fatal’ crashes with ‘injury’ crashes can help overcome this problem. Accordingly, ANOVA models were developed for ‘injury’, ‘PDO’, and ‘fatal and injury’ crash types. Further, it is found that the actual traffic volumes for North

Carolina freeways vary between 25,000 and 149,999 vehicles per day. Five levels of traffic volumes were defined at increments of 25,000 and used for the analysis.

Table 5-1 shows the model for estimating ‘injury’ crash rates. Sufficient data were not available for cell representing interchange segments with more than four lanes and with AADT between 50,000 and 74,999 vehicles per day. The effect of the type of segments is found to be significant, which indicates that the rates for ‘injury’ crashes for non-interchange segments are different from those for interchange segments. The highest injury crash rate for non-interchange segments and that for interchange segments occur on four-lane segments at volumes in the range of 50,000 to 74,999 vehicles per day. However, no underlying trend is found for ‘injury’ crash rates as traffic volume increases.

Table 5-2 shows the ANOVA model for ‘PDO’ crashes for North Carolina freeways. The effect of the type of segments is found to be significant, which indicates that the rates for ‘PDO’ crashes for non-interchange segments are different from those for interchange segments. However, without sufficient evidence, the effect of traffic volume and that of three-way interaction among three factors may be found to be significant by chance. In this research, a chance of making a Type I error was set to be 5% ($\alpha = 0.05$). As it can be seen in Table 5-2, the highest ‘PDO’ crash rate for non-interchange segments occurs on more-than-four-lane segments in the volume ranges of 50,000 to 74,999 vehicles per day while the highest rate for interchange segments occurs on four-lane segments in the volume ranges of 75,000 to 99,999 vehicles per day. However, no apparent underlying trend for ‘PDO’ crash rates is found in this model.

The ANOVA model for predicting the rates for ‘fatal and injury’ crashes combined is shown in Table 5-3. Combining ‘fatal’ crashes with ‘injury’ crashes can be a good practical strategy for modeling to overcome the difficulty for developing a model for ‘fatal’ crashes alone.

Table 5-1 Injury Crash Rates for North Carolina

Freeway Volumes (Vehicles/Day)	Injury Crash Rates (Crashes/MVMT)			
	Non-Interchange Segments		Interchange Segments	
	Four Lanes	More Than Four Lanes	Four Lanes	More Than Four Lanes
25,000-49,999	0.1136	0.0383	0.3284	0.3285
50,000-74,999	0.1704	0.2206	0.6103	*
75,000-99,999	0.0758	0.0980	0.6094	0.3939
100,000-124,999	*	0.1798	*	0.4582
125,000-149,999	*	0.1108	*	0.3439
*No freeway segment for this cell				P-Value
Model				<0.0001
Type of Segments				<0.0001
Number of Lanes				0.4346
Traffic Volume				0.1469
Interaction				
Type of Segments * Number of Lanes				0.4938
Type of Segments * Traffic Volume				0.4109
Number of Lanes * Traffic Volume				0.7369
Type of Segments * Number of Lanes * Traffic Volume				0.1879

Table 5-2 PDO Crash Rates for North Carolina

Freeway Volumes (Vehicles/Day)	PDO Crash Rates (Crashes/MVMT)			
	Non-Interchange Segments		Interchange Segments	
	Four Lanes	More Than Four Lanes	Four Lanes	More Than Four Lanes
25,000-49,999	0.2195	0.0957	0.5154	0.7048
50,000-74,999	0.3331	0.5820	1.1923	*
75,000-99,999	0.2223	0.1863	1.6182	0.9032
100,000-124,999	*	0.3990	*	1.0951
125,000-149,999	*	0.2157	*	1.0707
*No freeway segment for this cell				P-Value
Model				<0.0001
Type of Segments				<0.0001
Number of Lanes				0.4831
Traffic Volume				0.0246
Interaction				
Type of Segments * Number of Lanes				0.4416
Type of Segments * Traffic Volume				0.0807
Number of Lanes * Traffic Volume				0.0880
Type of Segments * Number of Lanes * Traffic Volume				0.0383

Table 5-3 Fatal and Injury Crash Rates for North Carolina

Freeway Volumes (Vehicles/Day)	Fatal and Injury Crash Rates (Crashes/MVMT)			
	Non-Interchange Segments		Interchange Segments	
	Four Lanes	More Than Four Lanes	Four Lanes	More Than Four Lanes
25,000-49,999	0.1162	0.0383	0.3384	0.3285
50,000-74,999	0.1711	0.2217	0.6193	*
75,000-99,999	0.0763	0.1006	0.6149	0.3989
100,000-124,999	*	0.1871	*	0.4609
125,000-149,999	*	0.1125	*	0.3480
*No freeway segment for this cell				P-Value
Model				<0.0001
Type of Segments				<0.0001
Number of Lanes				0.4136
Traffic Volume				0.1477
Interaction				
Type of Segments * Number of Lanes				0.4708
Type of Segments * Traffic Volume				0.4062
Number of Lanes * Traffic Volume				0.7643
Type of Segments * Number of Lanes * Traffic Volume				0.1979

The effect of the type of segments is found to be significant, which indicates that ‘fatal and injury’ crash rates for non-interchange segments and those for interchange segments are different. These rates are similar to those for ‘injury’ crashes only. The highest ‘fatal and injury’ crash rate for non-interchange segments occurs on more-than-four-lane segments in the volume ranges of 50,000 to 74,999 vehicles per day while the highest rate for interchange segments occurs on four-lane segments in the volume ranges of 50,000 to 74,999 vehicles per day.

ANOVA Models for Predicting Crash Rates for Tennessee Freeways

Similar to the case of North Carolina, it was difficult to develop ANOVA model for ‘fatal’ crashes for Tennessee freeways. Thus, only models for three types of crash severity – ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes were developed. The variables traffic volume, type of segments, and number of lanes were classified into groups similar to those for the models for North Carolina. However, the traffic volumes for Tennessee freeways were classified into the total of six groups – five groups increasing at an increment of 25,000 vehicles per day and another group for traffic volume equal to or higher than 150,000 vehicles per day.

The ANOVA model for ‘injury’ crashes for Tennessee freeways are presented in Table 5-4. The effect of the type of segments is found to be significant, which indicates that the rates for ‘injury’ crashes for non-interchange segments and those for interchange segments are significantly different. The injury crash rates for non-interchange segments with four lanes increase as freeway traffic volume increases to 99,999 vehicles per day and those for non-interchange segments with more than four lanes increase as freeway traffic volume increases up to 124,999 vehicles per day, after which the rate decreases. However, no trend of ‘injury’ crash rates is seen for interchange segments for both groups of number of lanes.

Table 5-4 Injury Crash Rates for Tennessee

Freeway Volumes (Vehicles/Day)	Injury Crash Rates (Crashes/MVMT)			
	Non-Interchange Segments		Interchange Segments	
	Four Lanes	More Than Four Lanes	Four Lanes	More Than Four Lanes
25,000-49,999	0.1111	0.1659	0.3510	0.4486
50,000-74,999	0.2460	0.2274	0.6205	0.5806
75,000-99,999	0.4892	0.2574	0.4658	0.4384
100,000-124,999	0.4569	0.3210	0.2306	0.6333
125,000-149,999	*	0.1862	*	0.7283
>=150,000	*	0.2503	*	0.5298
*No freeway segment for this cell				P-Value
Model				<0.0001
Type of Segments				0.0013
Number of Lanes				0.8374
Traffic Volume				0.6461
Interaction				
Type of Segments * Number of Lanes				0.1238
Type of Segments * Traffic Volume				0.2559
Number of Lanes * Traffic Volume				0.4308
Type of Segments * Number of Lanes * Traffic Volume				0.3757

The highest ‘injury’ crash rate for non-interchange segments occurs on four-lane segments at volumes in the range of 75,000 to 99,999 vehicles per day. The highest rate for interchange segments occurs on more-than-four-lane segments at volumes in the range of 125,000 to 149,999 vehicles per day.

The ANOVA model for ‘PDO’ crashes is presented in Table 5-5. The significant effect of the type of segments indicates that ‘PDO’ crash rates for non-interchange segments and those for interchange segments are different. The table values illustrate that non-interchange segments with four lanes and those with more than four lanes have higher ‘PDO’ crash rates as traffic volume increases to 124,999 vehicles per day. An increasing trend of ‘PDO’ crash rates is found for the group of interchange segments with more than four lanes for two separate ranges of freeway traffic volumes – between 25,000 and 74,999 vehicles per day and the other one for volumes higher than 75,000 vehicles per day. However, the increasing trend is not clearly evident for the group of interchange segments with four lanes. The highest ‘PDO’ crash rate for non-interchange segments occurs on four-lane segments at volumes in the range of 100,000 to 124,999 vehicles per day. The highest ‘PDO’ crash rate for interchange segments occurs on four-lane segments in the volume range of 50,000 to 74,999 vehicles per day.

The ANOVA model for ‘fatal and injury’ crashes for Tennessee freeways is presented in Table 5-6. The effect of the type of segments is found to be significant for this model, which indicates that the rates for ‘fatal and injury’ crashes for non-interchange segments and those for interchange segments are significantly different. The ‘fatal and injury’ crash rates for non-interchange segments with four lanes increases as freeway traffic volume increases to 99,999 vehicles per day and those for non-interchange segments with more than four lanes increase as freeway traffic volume increases up to 124,999 vehicles per day, after which the rate decreases. However, no trend is found for the rates for interchange segments as traffic volume increased.

Table 5-5 PDO Crash Rates for Tennessee

Freeway Volumes (Vehicles/Day)	PDO Crash Rates (Crashes/MVMT)			
	Non-Interchange Segments		Interchange Segments	
	Four Lanes	More Than Four Lanes	Four Lanes	More Than Four Lanes
25,000-49,999	0.3070	0.3375	1.0331	0.7719
50,000-74,999	0.7265	0.5244	1.9561	1.4741
75,000-99,999	0.7763	0.6671	0.9685	1.1313
100,000-124,999	1.0264	0.7671	0.3980	1.5306
125,000-149,999	*	0.5980	*	1.6337
>=150,000	*	0.9707	*	1.8769
*No freeway segment for this cell				P-Value
Model				0.0006
Type of Segments				0.0026
Number of Lanes				0.9931
Traffic Volume				0.1750
Interaction				
Type of Segments * Number of Lanes				0.4454
Type of Segments * Traffic Volume				0.2833
Number of Lanes * Traffic Volume				0.4276
Type of Segments * Number of Lanes * Traffic Volume				0.3211

Table 5-6 Fatal and Injury Crash Rates for Tennessee

Freeway Volumes (Vehicles/Day)	Fatal and Injury Crash Rates (Crashes/MVMT)			
	Non-Interchange Segments		Interchange Segments	
	Four Lanes	More Than Four Lanes	Four Lanes	More Than Four Lanes
25,000-49,999	0.1140	0.1794	0.3510	0.4486
50,000-74,999	0.2526	0.2310	0.6272	0.5965
75,000-99,999	0.4916	0.2599	0.4658	0.4473
100,000-124,999	0.4599	0.3233	0.2306	0.6426
125,000-149,999	*	0.1870	*	0.7381
>=150,000	*	0.2513	*	0.5320
*No freeway segment for this cell				P-Value
Model				<0.0001
Type of Segments				0.0012
Number of Lanes				0.7839
Traffic Volume				0.6325
Interaction				
Type of Segments * Number of Lanes				0.1147
Type of Segments * Traffic Volume				0.2485
Number of Lanes * Traffic Volume				0.4308
Type of Segments * Number of Lanes * Traffic Volume				0.3716

The highest ‘fatal and injury’ crash rate for non-interchange segments occurs on four-lane segments in the volume range of 75,000 to 99,999 vehicles per day. The highest rate for interchange segments occurs on more-than-four-lane segments in the volume range of 125,000 to 149,999 vehicles per day.

Conclusions

ANOVA models represent one type of prediction models for estimating crash rates (and/or crashes) for freeways. It can be used as a look-up table, and it has a simple form, which is easy to understand. Thus an ANOVA model can be attractive to transportation planners and other users. However, the selection of factors and the ranges for these factors can be difficult and will need a good understanding of the causes of crashes.

It was difficult to develop ANOVA models for ‘fatal’ crashes for both North Carolina and Tennessee. Therefore, it is recommended that ‘fatal’ crashes and ‘injury’ crashes should be combined for developing the models.

Although no underlying trend was found for crash rates for North Carolina freeways, the rates for Tennessee freeways slightly increased with increasing traffic volumes in many cases. However, the increasing of these rates was not significant. Further, it was found that the rates for ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes for non-interchange segments are significantly different (lower) from those for interchange segments. This indicated that non-interchange segments and interchange segments are associated with different levels of crash risk, and, therefore, developing separate crash prediction models for non-interchange segments and interchange segments is appropriate and should be taken seriously.

6. CRASH PREDICTION MODELS BASED ON REGRESSION ANALYSIS APPROACH

Crash prediction models for North Carolina and Tennessee urban freeways developed by the regression analysis approach are presented in this chapter along with the results of analysis. All regression models presented were developed using the GENMOD procedure in statistical software SAS. Similar to the case of ANOVA models, it was difficult to develop regression models for estimating ‘fatal’ crashes, and so a prediction model for estimating ‘fatal’ crashes and ‘injury’ crashes combined was developed. Separate regression models were developed for estimating three crash types – ‘injury’, ‘PDO’, and ‘fatal and injury’ crashes – for non-interchange segments and interchange segments. Each type of segments was further divided into sub-groups of four-lane and more-than-four-lane.

All the regression model forms presented in Table 3-2 were attempted. For each sub-group of four-lane freeways and more-than-four-lane freeways, only the models which have the ratio between Deviance and the degrees of freedom close to one and have a relatively high estimated R_k^2 , are presented in this chapter. The technique for estimating R_k^2 was introduced by Miaou (1996) and this indicator is used to measure the level of explanatory ability of each model. The R_k^2 was calculated using Equation 3-6. It should be pointed out that the expected number of crashes predicted by the regression models represents the total for three years.

Crash Prediction Models based on Regression Analysis for North Carolina Freeways

This section represents the regression models for urban freeways in North Carolina. The results of analysis reveal similar patterns of the prediction model forms. It should be noted that the variable ‘Segment Length’ used in all models is in miles. Details of the

regression models for non-interchange segments and interchange segments are presented in the following sections.

Regression Models for Non-Interchange Segments of North Carolina

Table 6-1 through Table 6-3 present the regression models for ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes, respectively. Two model forms are found to be acceptable for non-interchange segments of North Carolina. The first form is for four-lane segments while the second model form is for more-than-four-lane segments.

It can be seen in Table 6-1 through Table 6-3 that all the coefficients and the exponents for ‘Segment Length’ and ‘AADT’ are estimated with a positive sign, which indicate that longer segments and/or higher freeway traffic volumes, which increase exposure, will result in more ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes on both four-lane and more-than-four-lane non-interchange segments. In addition, it is found that the R_k^2 is higher for the models for four-lane segments.

Table 6-1 Injury Crash Models for Non-Interchange Segments of North Carolina

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k^2
Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			36.5477	1.1075	0.1729	0.90
a	1.9×10^{-7}	<0.0001				
b ₁	1.2675	<0.0001				
b ₂	1.6010	<0.0001				
More Than Four Lanes						
$\mu = (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			42.4394	1.2126	1.7306	0.41
b ₁	1.0788	<0.0001				
b ₂	0.2283	<0.0001				

Table 6-2 PDO Crash Models for Non-Interchange Segments of North Carolina

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k²
Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			46.5763	1.4114	0.1282	0.92
a	4.9×10^{-8}	<0.0001				
b ₁	1.1043	<0.0001				
b ₂	1.8007	<0.0001				
More Than Four Lanes						
$\mu = (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			43.1936	1.2341	1.8649	0.41
b ₁	1.2127	<0.0001				
b ₂	0.2935	<0.0001				

Table 6-3 Fatal and Injury Crash Models for Non-Interchange Segments of North Carolina

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k²
Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			36.4848	1.1056	0.1755	0.90
a	2.3×10^{-7}	<0.0001				
b ₁	1.2811	<0.0001				
b ₂	1.5838	<0.0001				
More Than Four Lanes						
$\mu = (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			42.4883	1.2140	1.7530	0.40
b ₁	1.0626	<0.0001				
b ₂	0.2305	<0.0001				

Regression Models for Interchange Segments of North Carolina

Only one form of models was selected for interchange segments of North Carolina. The variable ‘AADT’ is found significant for these models while the variable ‘Segment Length’ is omitted because all interchange segments were specified to have the same length of 3,000 feet.

In addition, it should be pointed out that the variable ‘AADT’ used in the models for interchange segments represents the average of traffic volumes for freeway lanes near interchange areas. The coefficients and the exponents for crash prediction models for interchange segments of North Carolina are presented in Table 6-4 through Table 6-6.

As seen in Table 6-4 through Table 6-6, the magnitudes and the signs of all the coefficients and the exponents for ‘AADT’ indicate that higher freeway traffic volumes will result in more ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes on interchange segments. In addition, the models for four-lane segments were found to have higher R_k^2 than the models for more-than-four-lane segments.

Table 6-4 Injury Crash Models for Interchange Segments of North Carolina

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k^2
Four Lanes						
$\mu = a (\text{AADT})^b$			34.7071	1.1196	0.2811	0.49
a	5.5×10^{-7}	<0.0001				
b	1.5756	<0.0001				
More Than Four Lanes						
$\mu = a (\text{AADT})^b$			33.4307	1.1144	0.2164	0.27
a	4.0×10^{-5}	0.0120				
b	1.1589	0.0010				

Table 6-5 PDO Crash Models for Interchange Segments of North Carolina

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k²
Four Lanes						
$\mu = a(AADT)^b$			33.3565	1.0760	0.1789	0.76
a	2.1×10^{-10}	<0.0001				
b	2.3557	<0.0001				
More Than Four Lanes						
$\mu = a(AADT)^b$			34.0315	1.1344	0.2957	0.27
a	1.6×10^{-5}	0.0065				
b	1.3167	0.0002				

Table 6-6 Fatal and Injury Crash Models for Interchange Segments of North Carolina

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k²
Four Lanes						
$\mu = a(AADT)^b$			34.7258	1.1202	0.2704	0.50
a	6.8×10^{-7}	<0.0001				
b	1.5578	<0.0001				
More Than Four Lanes						
$\mu = a(AADT)^b$			33.4711	1.1157	0.2135	0.28
a	3.8×10^{-5}	0.0113				
b	1.1629	0.0009				

Pattern of Variation in Crashes with Traffic Volume based on North Carolina Models

A question that is of interest to transportation engineers and planners is: how does the number of crashes on a freeway segment increase with the increase of traffic volume – at a constant rate, increasing rate, or decreasing rate? Another related question is whether the severity types of crashes vary in their proportions to total crashes as traffic volume increases; for example, does the proportion of fatal crashes increase or decrease with the increase of traffic volume?

In order to provide some information/answers to these questions, the regression models for each state were used to estimate the number of crashes for a unit segment length for traffic volumes increasing at an increment of 10,000 vehicles per day. These values were plotted in graphical forms and the graphs for ‘fatal and injury’ crashes, ‘PDO’ crashes, and ‘total’ crashes are presented in Figure 6-1 through Figure 6-3, respectively. The maximum AADT used for four-lane segments was 100,000 vehicles per day, and that for more-than-four-lane segments was 150,000 vehicles per day. It should be pointed out that the curves of ‘total’ crashes were developed from combining the curves of ‘fatal and injury’ crashes and those of ‘PDO’ crashes.

The shape of the curves of ‘fatal and injury’ and ‘PDO’ crashes is found to be similar to that of ‘total’ crashes. For all graphs, the curve for interchange segment with four lanes stands out distinctly from others showing increasing slope as AADT increases, which indicates an increasing crash rate. For range of AADT between 50,000 and 100,000 vehicles per day, the number of crashes on these segments also is greater than that on others. The implication of this finding is that interchange segments with four lanes are most likely to have higher crashes than others as AADT exceeds 50,000 vehicles per day, and the likelihood increases as AADT increases.

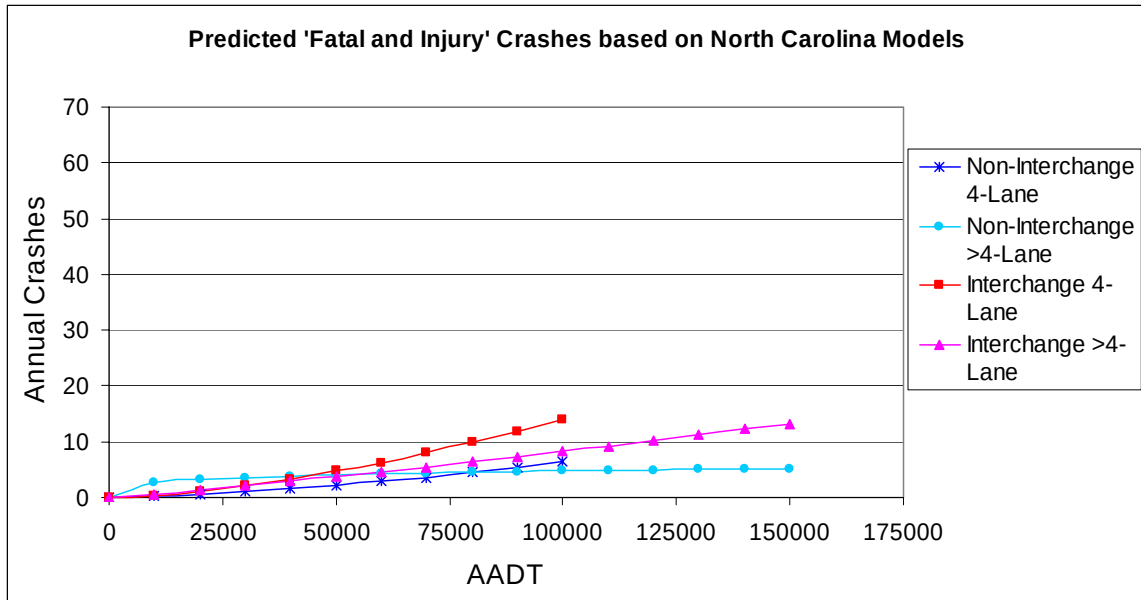


Figure 6-1 Predicted Pattern of Annual 'Fatal and Injury' Crashes based on North Carolina Models

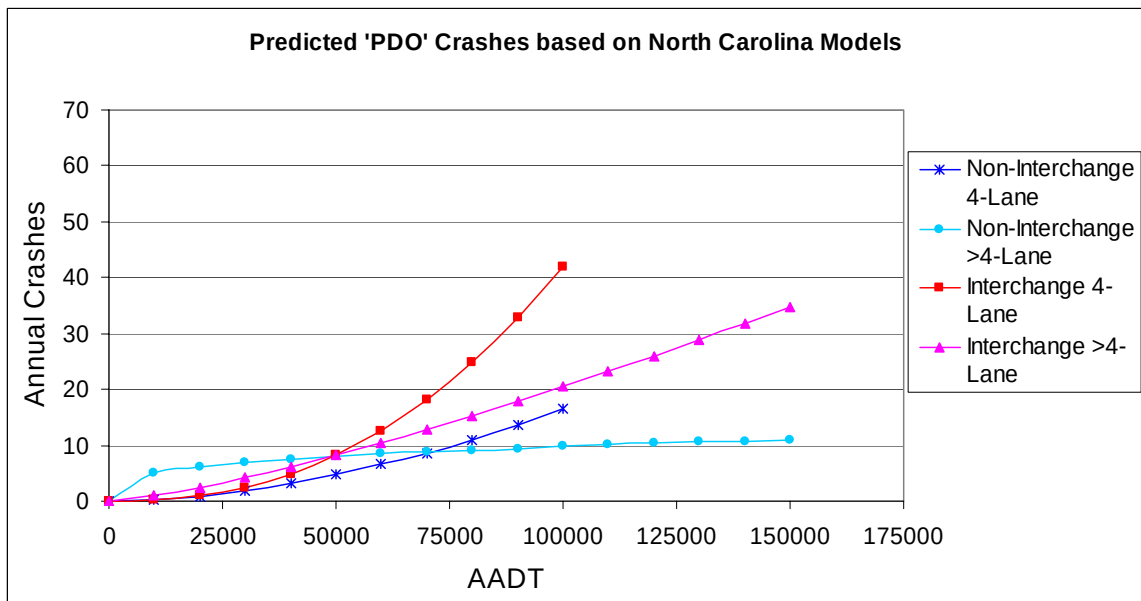


Figure 6-2 Predicted Pattern of Annual 'PDO' Crashes based on North Carolina Models

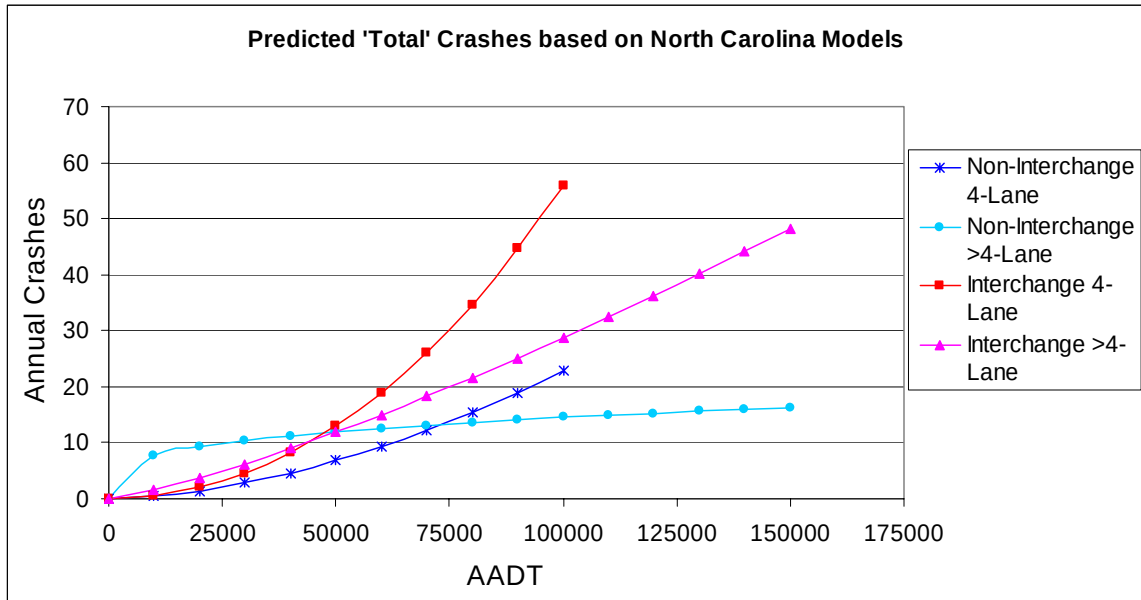


Figure 6-3 Predicted Pattern of Annual 'Total' Crashes based on North Carolina Models

Crash Prediction Models based on Regression Analysis for Tennessee Freeways

The modeling approach used for Tennessee freeways is similar to that used for North Carolina freeways. However, the models for Tennessee freeways are found to have slightly different forms from those for North Carolina freeways.

Regression Models for Non-Interchange Segments of Tennessee

Tables 6-7 through 6-9 present crash prediction models for ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes combined, respectively for non-interchange segments of Tennessee. Among all the model forms presented in Table 3-2, there is one form acceptable for regression models for estimating all three crash types. The variables ‘Segment Length’, ‘AADT’, coefficient ‘a’, and exponents ‘b₁’ and ‘b₂’ are included in the models for non-interchange segments of Tennessee. The magnitude and the sign of the exponents for ‘Segment Length’ and ‘AADT’ indicate that longer freeway segments and/or higher freeway traffic volume, which increase ‘exposure’, will result in more crashes for all three crash types on non-interchange segments with four lanes and also on those with more than four lanes. In addition, R_k^2 values are fairly close for the models for four-lane segments and those for more-than-four-lane segments.

Table 6-7 Injury Crash Models for Non-Interchange Segments of Tennessee

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k^2
Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			39.8414	1.2073	0.3540	0.65
a	2.2×10^{-10}	<0.0001				
b ₁	0.7856	<0.0001				
b ₂	2.2600	<0.0001				
More Than Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			109.041	1.0586	0.4562	0.62
a	5.0×10^{-5}	0.0001				
b ₁	0.9649	<0.0001				
b ₂	1.1480	<0.0001				

Table 6-8 PDO Crash Models for Non-Interchange Segments of Tennessee

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k²
Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			40.2107	1.2185	0.3448	0.67
a	1.7×10^{-10}	<0.0001				
b ₁	0.8078	<0.0001				
b ₂	2.3669	<0.0001				
More Than Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			116.546	1.1315	0.6248	0.53
a	1.2×10^{-5}	<0.0001				
b ₁	0.8996	<0.0001				
b ₂	1.3525	<0.0001				

Table 6-9 Fatal and Injury Crash Models for Non-Interchange Segments of Tennessee

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k²
Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			40.0045	1.2123	0.3517	0.65
a	2.7×10^{-10}	<0.0001				
b ₁	0.7906	<0.0001				
b ₂	2.2430	<0.0001				
More Than Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			108.828	1.0566	0.4389	0.64
a	6.8×10^{-5}	0.0001				
b ₁	0.9726	<0.0001				
b ₂	1.1233	<0.0001				

Regression Models for Interchange Segments of Tennessee

Tables 6-10 through 6-12 present the regression models for ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes combined, respectively for interchange segments. It should be pointed out that the lengths of interchange segments were not equal in all cases. Thus, the variable ‘Segment Length’ was used in models for interchange segments of Tennessee.

Two model forms are found acceptable for interchange segments. The first form includes the variables ‘Segment Length’, ‘AADT’, and exponents ‘ b_1 ’ and ‘ b_2 ’. This model form is for interchange segments with four lanes. As shown in Table 6-10 through Table 6-12, the magnitude and the positive sign of the exponents of ‘Segment Length’ and ‘AADT’ indicate that for all three crash types for interchange segments with four lanes, the number of crashes are higher as segment length and/or freeway traffic volume increase, which also increase ‘exposure’.

Table 6-10 Injury Crash Models for Interchange Segments for Tennessee

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k^2
Four Lanes						
$\mu = (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			15.7556	1.2120	0.1779	0.53
b_1	0.4803	0.0007				
b_2	0.2977	<0.0001				
More Than Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			51.8295	1.1028	0.2777	0.40
a	9.0×10^{-5}	0.0034				
b_1	0.8098	<0.0001				
b_2	1.1607	<0.0001				

Table 6-11 PDO Crash Models for Interchange Segments for Tennessee

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k²
Four Lanes						
$\mu = (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			15.8369	1.2182	0.3246	0.33
b ₁	0.4239	0.0089				
b ₂	0.3853	<0.0001				
More Than Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			52.1781	1.1102	0.2852	0.40
a	2.5x10 ⁻⁵	0.0006				
b ₁	0.7059	0.0006				
b ₂	1.3502	<0.0001				

Table 6-12 Fatal and Injury Crash Models for Interchange Segments for Tennessee

Regression Models	Estimate	P-value	Deviance	Deviance/DF	Dispersion (k)	R_k²
Four Lanes						
$\mu = (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			15.7645	1.2127	0.1817	0.54
b ₁	0.4919	0.0005				
b ₂	0.2989	<0.0001				
More Than Four Lanes						
$\mu = a (\text{Segment Length})^{b_1} (\text{AADT})^{b_2}$			51.7710	1.1015	0.2697	0.40
a	0.0001	0.0034				
b ₁	0.8117	<0.0001				
b ₂	1.1528	<0.0001				

The second model form, which is similar to the first model form and includes the coefficient 'a', is acceptable for interchange segments with more than four lanes. These regression models are presented in Table 6-10 through Table 6-12. The positive sign of the exponents for 'Segment Length' and 'AADT' indicates that longer segment and/or higher freeway traffic volume will result in more 'injury' crashes, 'PDO' crashes, and 'fatal and injury' crashes. In addition, the R_k^2 for the model for four-lane segments is higher than that for the model for more-than-four-lane segments for the cases of 'injury' crashes and 'fatal and injury' crashes. The R_k^2 for the model for more-than-four-lane segments is higher than that for the model for four-lane segments for the case of 'PDO' crashes.

Pattern of Variation in Crashes with Traffic Volume based on Tennessee Models

To provide similar information as those for the cases of North Carolina on how the number of crashes increases with the increase of traffic volume on Tennessee freeways, the graphs for 'fatal and injury' crashes, 'PDO' crashes, and 'total' crashes predicted for a unit segment length of Tennessee freeways were developed and they are presented in Figure 6-4 through Figure 6-6, respectively.

The pattern of increase of crashes with the increase of traffic volume as predicted by the models for Tennessee is found to be different from that for North Carolina. The graphs for Tennessee are different from those for North Carolina in several respects. As it can be seen in Figure 6-4 through Figure 6-6, the curve for interchange segments with four lanes does not stand out as prominently above the others as it does in the case of North Carolina. The slope of this curve is almost linear. The curve for interchange segments with more than four lanes stands out above others when AADT exceeds 60,000 vehicles per day. In addition, the curve for non-interchange segments with four lanes has increasing slope indicating increasing crash rate beyond an AADT of 40,000 vehicles per day.

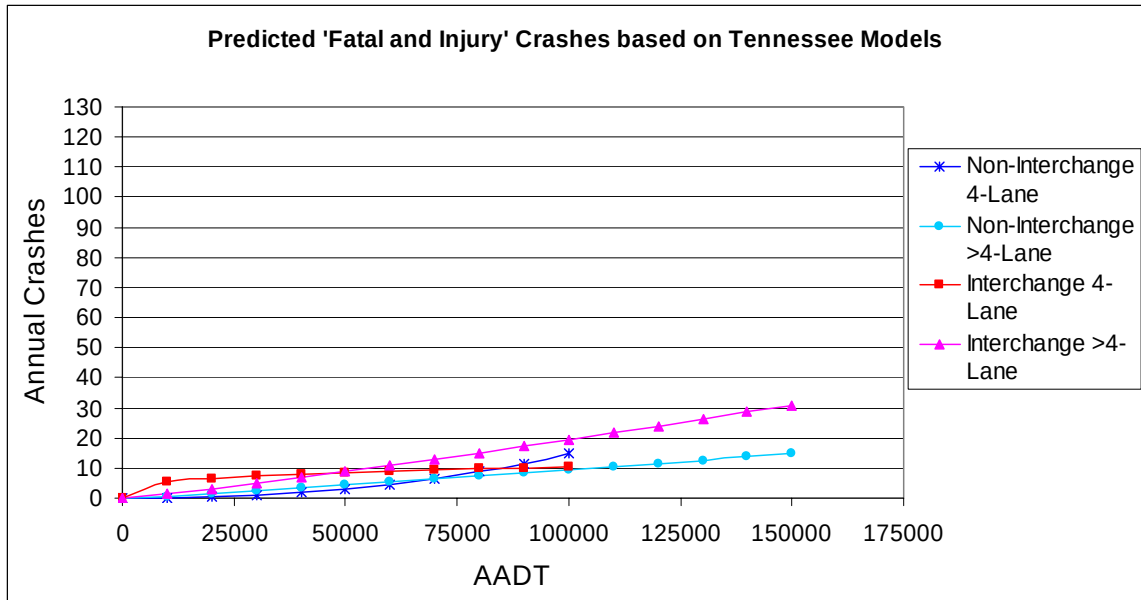


Figure 6-4 Predicted Pattern of Annual 'Fatal and Injury' Crashes based on Tennessee Models

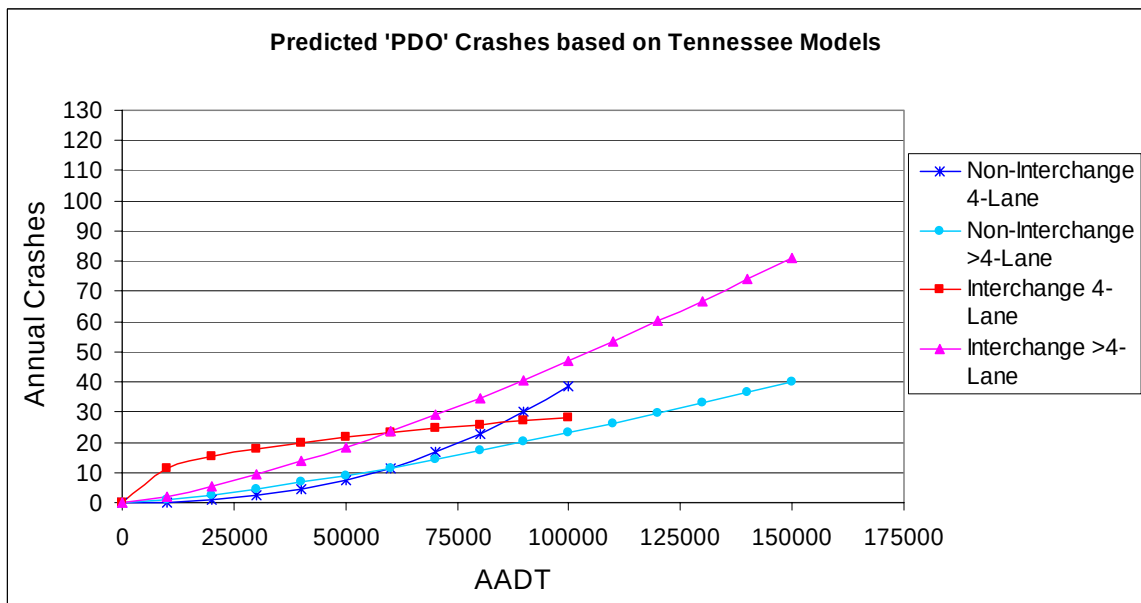


Figure 6-5 Predicted Pattern of Annual 'PDO' Crashes based on Tennessee Models

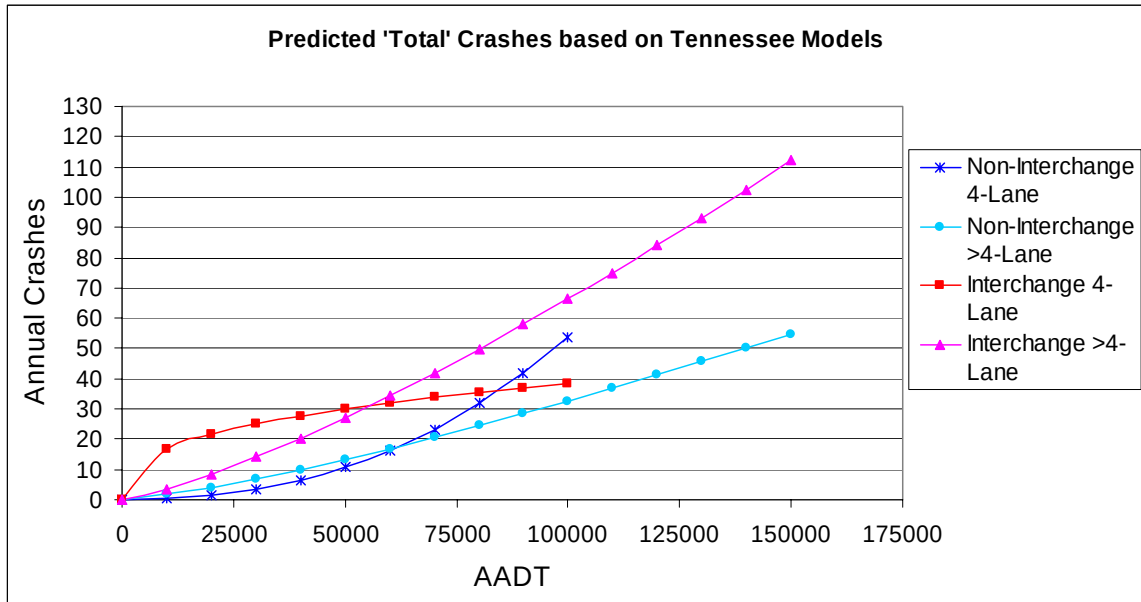


Figure 6-6 Predicted Pattern of Annual 'Total' Crashes based on Tennessee Models

Summary for the Forms of Regression Models

The regression models for non-interchange segments are presented in Table 6-13. There is only one common model form for non-interchange segments with four lanes for both states of North Carolina and Tennessee. However, model forms for all three crash types for non-interchange segments with more than four lanes of North Carolina are slightly different from those of Tennessee.

For interchange segments, there are three different forms of regression models as presented in Table 6-14. The regression models for North Carolina were developed with only the variable 'AADT' since the segment lengths were equal in all cases. However, similar model form was not appropriate for interchange segments of Tennessee since the segments did not have equal lengths in all cases. The models for Tennessee were developed using the variables 'Segment Length' and 'AADT'.

Table 6-13 Summary of Selected Model Forms for Non-Interchange Segments

Forms of Regression Models	Conditions	
	Type of Crashes	Number of Freeway Lanes (State)
1. $\mu = (\text{Segment Length})^{b1} (\text{AADT})^{b2}$	Injury PDO Fatal and Injury	More than four lanes (NC)
2. $\mu = a (\text{Segment Length})^{b1} (\text{AADT})^{b2}$	Injury PDO Fatal and Injury	Four lanes (NC & TN) More than four lanes (TN)

Table 6-14 Summary of Selected Model Forms for Interchange Segments

Forms of Regression Models	Conditions	
	Type of Crashes	Number of Freeway Lanes (State)
1. $\mu = a (\text{AADT})^b$	Injury PDO Fatal and Injury	Four lanes (NC); More than four lanes (NC)
2. $\mu = (\text{Segment Length})^{b1} (\text{AADT})^{b2}$	Injury PDO Fatal and Injury	Four lanes (TN)
3. $\mu = a (\text{Segment Length})^{b1} (\text{AADT})^{b2}$	Injury PDO Fatal and Injury	More than four lanes (TN)

Conclusions

Non-linear regression models based on the assumption that the probability distribution of crashes follows negative binomial distribution were found appropriate for developing crash prediction models. Several model forms were found to be appropriate for different crash severity types for non-interchange segments and interchange segments. However, it was found that developing models for estimating ‘fatal’ crashes was difficult. Therefore, ‘fatal’ crashes and ‘injury’ crashes were combined for developing regression models. It should be pointed out that all the regression models were developed using crash data for three years. The estimated number of crashes should be divided by the factor of three to yield annual number of crashes.

The results of regression analysis indicated that longer segment length and/or higher traffic volume will result in more crashes for non-interchange segments and interchange segments for both states of North Carolina and Tennessee. In addition, it was found in most of the cases that the models for four-lane segments performed as well as or better than the models for more-than-four-lane segments. The pattern of increase of crashes with the increase of traffic volume as predicted by the regression models was found to be different between both states. For North Carolina, it was found that crash rates for interchange segments with four lanes would increase with increasing traffic volume, and these segments may be most crash-prone when AADT exceeds 50,000 vehicles per day. For Tennessee, the models indicated that the interchange segments with more than four lanes have the highest crash rates when AADT exceeds 60,000 vehicles per day. An assessment of all the regression models as well as a comparison of the prediction performance of these models and other models – ANOVA models and classification tree models – will be presented in a latter chapter.

7. CRASH PREDICTION MODELS BASED ON CLASSIFICATION TREE ANALYSIS APPROACH

This chapter presents the results of classification tree models developed for analyzing freeway segment crashes. The classification tree approach requires a different format of data tables and it presents its results in a form different from those for ANOVA and regression models. Results of analysis for a classification tree are represented in terms of the number of crashes by severity type for each group of segments. However, to make all classification trees clearly interpretable and comparable to other models, an additional table was created for each classification tree to convert the number of crashes represented in each tree node into crash rates for each crash severity type. Only crash rates for the nodes classified as ‘leaf’ of a classification tree are presented in a crash rate table. The conversion to crash rate was performed using Equation 4-1.

Another difference between classification tree models and the other two types of prediction models is that a single classification tree model can be useful for estimating crashes for all sub-categories within the group of the same segment type. For example, one tree can be used for all sub-categories of non-interchange segments and another tree can be used for all sub-categories of interchange segments.

Classification Tree Models for North Carolina Freeways

Classification tree models were developed for estimating crashes for different locations of freeways, one tree model for non-interchange segments and another for interchange segments. Two variables – ‘AADT’ and ‘number of freeway lanes’ – were used for developing classification tree models for both types of segments.

The maximum depth of the tree level was specified to avoid complex models, which cannot be easily used in the transportation planning process. All the tree models were limited to the depth of five levels below the root node.

Classification Tree Model for Non-Interchange Segments of North Carolina

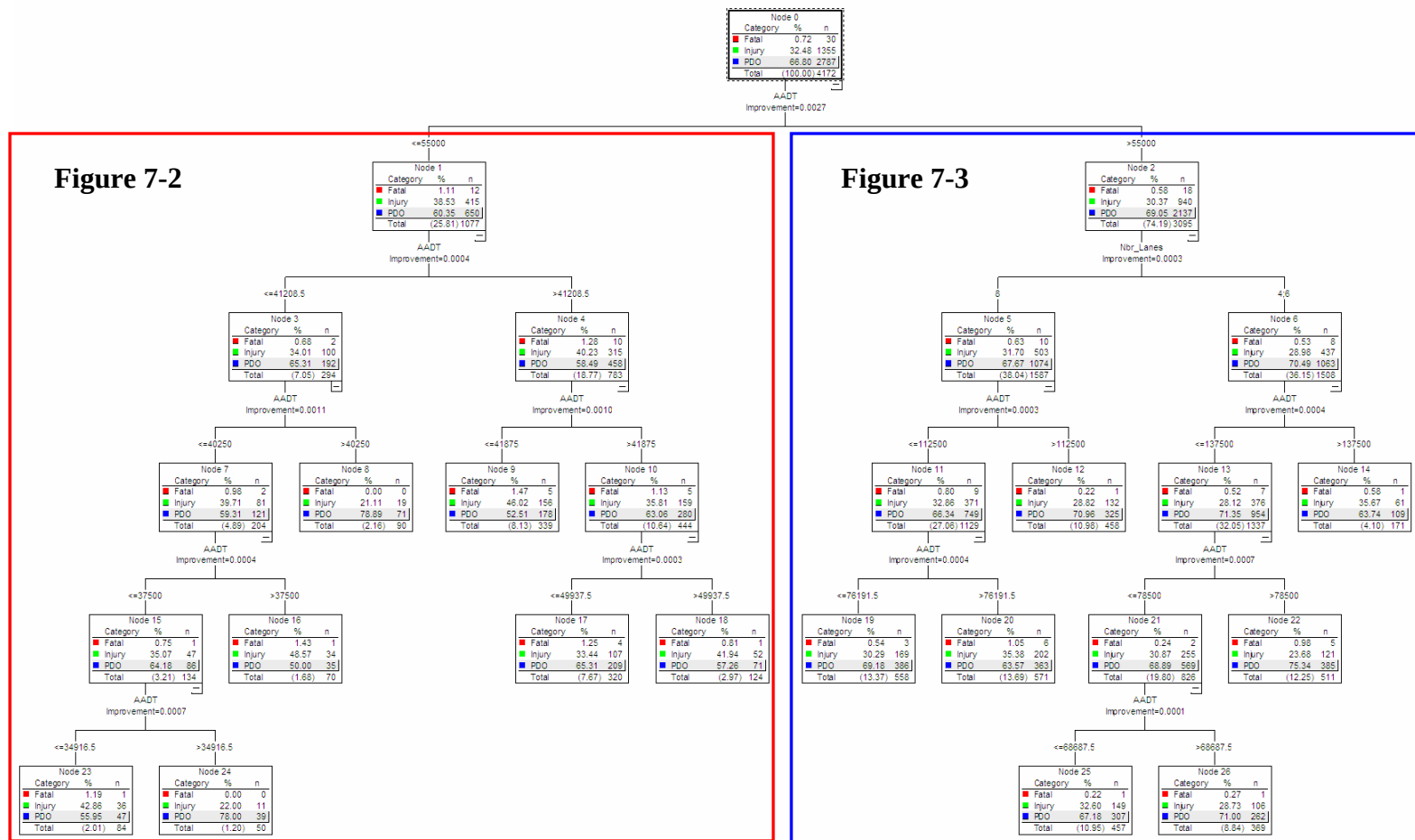
The classification tree for estimating crashes for non-interchange segments is shown in Figure 7-1 and two enlarged areas of the tree are presented in Figure 7-2 and Figure 7-3. All non-interchange segments were stratified into 14 groups with various traffic volumes and numbers of freeway lanes. It can be inferred that all the segments within each group (node) are associated with the same level of crash risk and are likely to have similar crash rates.

It is found that the variable ‘AADT’ was used in the first split, which classified the tree into two groups – one with traffic volumes less than or equal to 55,000 vehicles per day and another one with traffic volumes more than 55,000 vehicles per day. Figure 7-1 shows that the variable ‘AADT’ is also found in 12 out of 13 splits while the number of freeway lanes is found in one split.

Table 7-1 was developed for the tree model presented in Figure 7-1 to assess crash rates associated to each node. To calculate the rates for each crash type for segments in each node, the information on ‘VMT’ for a group of segments in each node must be obtained. For each segment, the ‘VMT’ can be calculated using Equation 7-1.

$$\text{VMT} = \text{Segment Length} \times \text{AADT} \times 365 \times \text{Years} \quad (7-1)$$

For example, the summation of ‘VMT’ for all segments in node 23, which have traffic volumes less than or equal to 34,917 vehicles per day, is 223.50 millions for three years as shown in Table 7-1. The rate for ‘injury’ crash type for segments in node 23 is equal to 0.1611 per million VMT, which is calculated from dividing the number of ‘injury’ crashes in node 23 (presented in Figure 7-1) by the ‘VMT’ for node 23 (presented in Table 7-1).



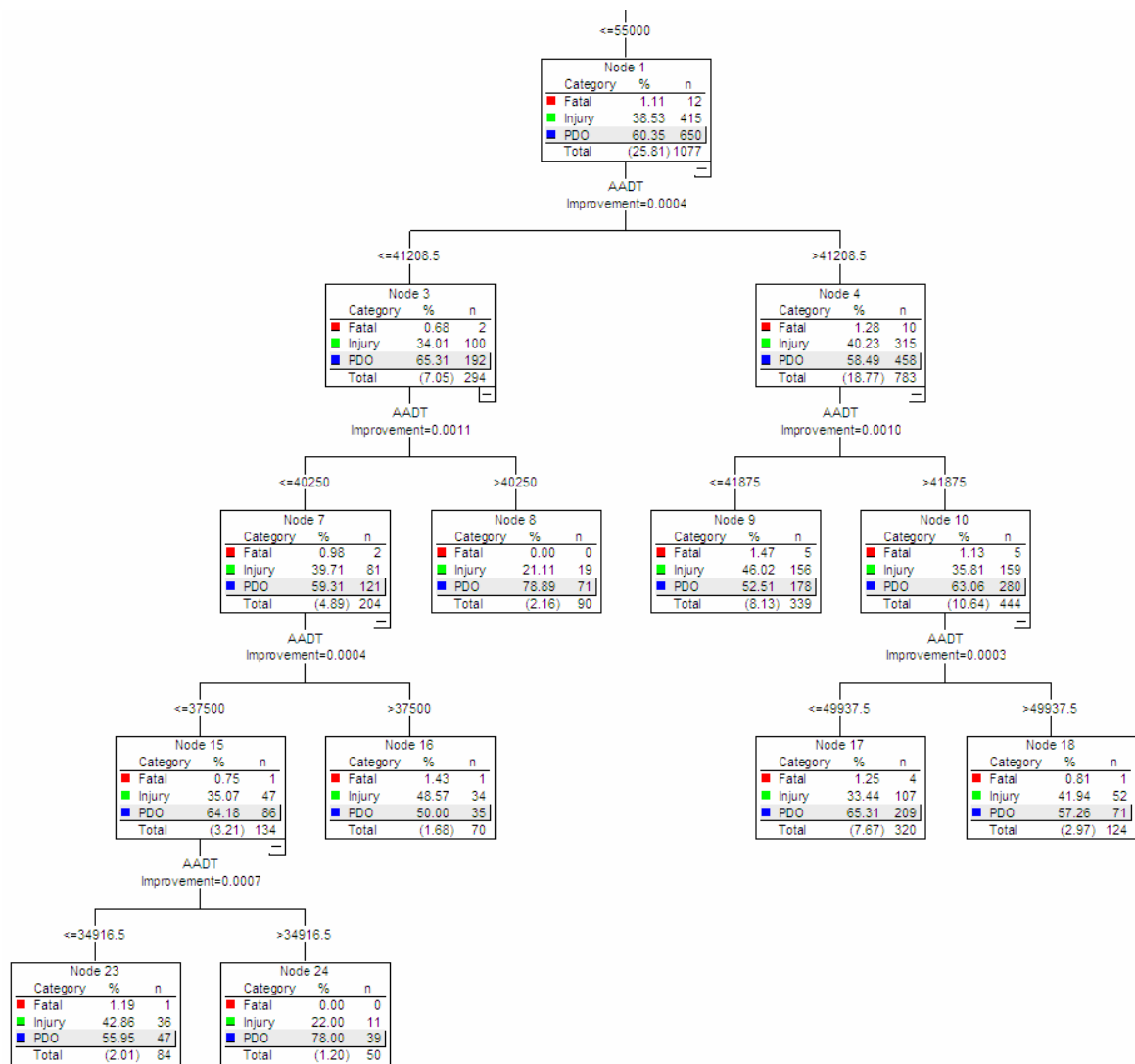


Figure 7-2 Enlarged Area of the Classification Tree Model for Non-Interchange Segments with Traffic Volume Less than or Equal to 55,000 Vehicles per Day

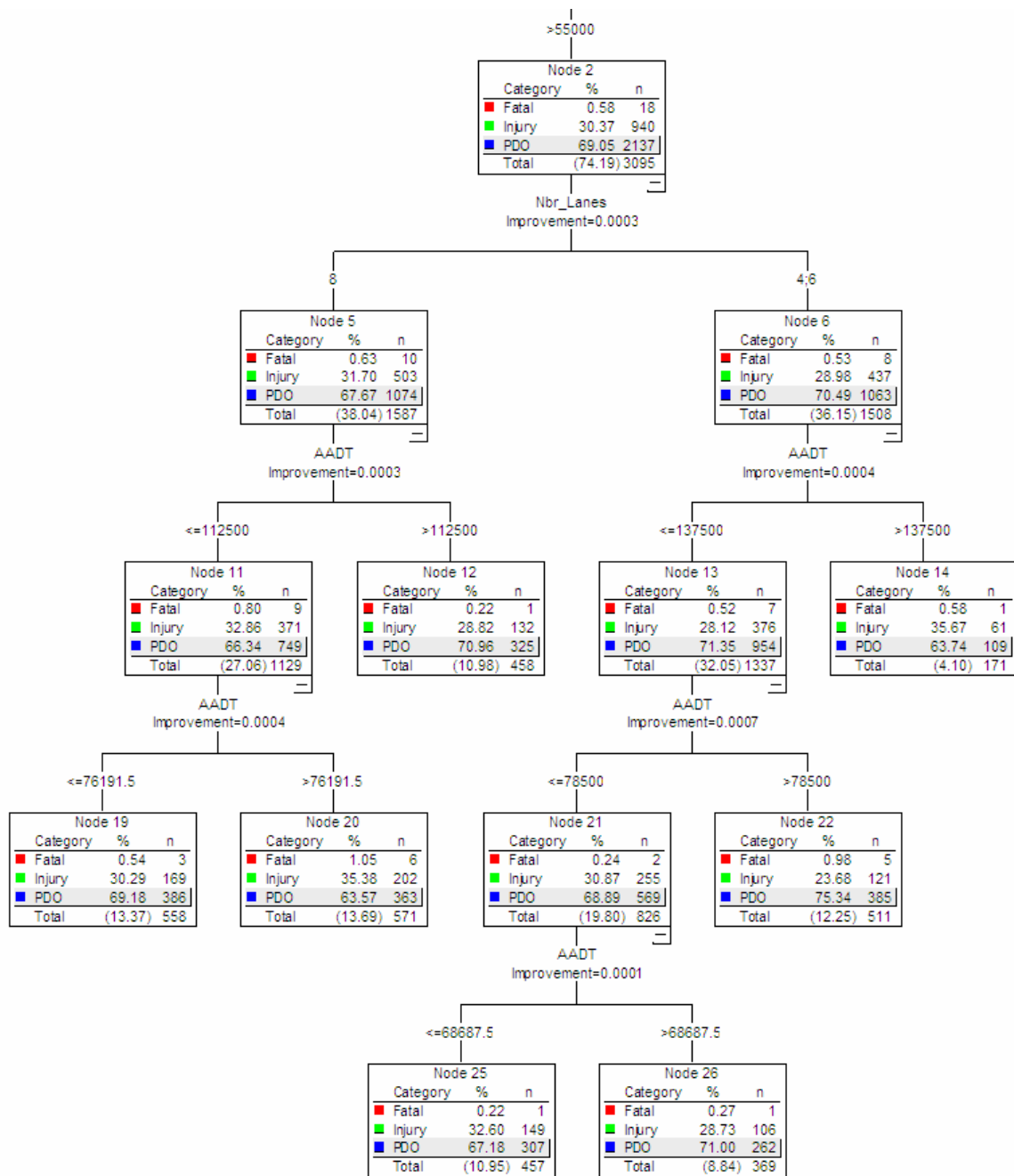


Figure 7-3 Enlarged Area of the Classification Tree Model for Non-Interchange Segments with Traffic Volume More than 55,000 Vehicles per Day

**Table 7-1 Crash Rates for the Classification Tree Model for Non-Interchange
Segments of North Carolina**

No. of Lanes	Traffic Volume (Vehicles per day)	Tree Node	VMT (Millions for 3 Years)	Crash Rates (Crashes/MVMT)			
				Fatal	Injury	PDO	Fatal and Injury
All	<=34,917	23	223.50	0.0045	0.1611	0.2103	0.1655
All	34,918-37,500	24	173.58	0.0000	0.0634	0.2247	0.0634
All	37,501-40,250	16	207.98	0.0048	0.1635	0.1683	0.1683
All	40,251-41,209	8	338.13	0.0000	0.0562	0.2100	0.0562
All	41,210-41,875	9	744.06	0.0067	0.2097	0.2392	0.2164
All	41,876-49,938	17	690.53	0.0058	0.1550	0.3027	0.1607
All	49,939-55,000	18	364.96	0.0027	0.1425	0.1945	0.1452
4;6	55,001-68,688	25	758.03	0.0013	0.1966	0.4050	0.1979
4;6	68,689-78,500	26	696.55	0.0014	0.1522	0.3761	0.1536
4;6	78,501-137,500	22	1607.68	0.0031	0.0753	0.2395	0.0784
4;6	>137,500	14	195.65	0.0051	0.3118	0.5571	0.3169
8	55,001-78,192	19	860.58	0.0035	0.1964	0.4485	0.1999
8	78,193-112,500	20	2705.17	0.0022	0.0747	0.1342	0.0769
8	>112,500	12	370.60	0.0027	0.3562	0.8770	0.3589

For non-interchange segments of North Carolina, the ‘VMT’ was calculated for each group of segments and presented in Table 7-1. The same procedure was utilized for developing all crash-rate tables presented in this chapter.

It may be noted in Figure 7-1 that all nodes have similar proportion of ‘fatal’ crashes, ‘injury’ crashes, and ‘PDO’ crashes. ‘Fatal’ crashes are approximately 1%, while ‘injury’ crashes vary between 20% and 40% and ‘PDO’ crashes vary between 60% and 80% of the total for each node. In addition, Table 7-1 shows that non-interchange segments in node 9 have the highest fatal crash rate. These are segments with traffic volumes between 41,210 and 41,875 vehicles per day. Table 7-1 also shows that the highest injury crash rate and the highest PDO crash rate are found for a group of segments in node 12, which represents eight-lane segments with traffic volumes more than 112,500 vehicles per day.

In addition, the highest total crash rate, which includes the rates for ‘fatal’, ‘injury’, and ‘PDO’ crashes, is also found for a group of segments in node 12.

Classification Tree Model for Interchange Segments of North Carolina

The classification tree model for interchange segments is presented in Figure 7-4 and four enlarged areas of the tree are presented in Figure 7-5 through Figure 7-8. The tree structure for interchange segments is different from that for non-interchange segments. As shown in Figure 7-4, the classification tree stopped growing early on the left side and held node 3 and node 4 as a ‘leaf’ of the tree. This indicates that no further significant improvement is found for these nodes, which represent a group of segments with traffic volumes less than or equal to 65,500 vehicles per day and another group with traffic volumes between 65,501 and 69,834 vehicles per day respectively. However, the stratification for those segments having traffic volumes higher than 69,834 vehicles per day was fully developed and stopped at the specified depth level of five.

The classification tree classified interchange segments into 12 groups with similar characteristics. Out of 11 splits, the variable ‘AADT’ is found in 10 splits. Only one split is associated with the variable ‘number of freeway lanes’. It may be concluded that different levels of freeway traffic volumes have more influence on interchange segments crashes than the number of freeway lanes. In addition, all tree nodes are found to have nearly similar proportions of ‘fatal’ crashes, ‘injury’ crashes, and ‘PDO’ crashes. ‘Fatal’ crashes are approximately 1% while ‘injury’ crashes are about 30% and ‘PDO’ crashes are about 70% of total crashes.

Table 7-2 presents crash rates developed for the classification tree for interchange segments of North Carolina. A group of interchange segments with traffic volumes between 65,501 and 69,834 vehicles per day, which is represented in node 4 of the classification tree, is found to have the highest fatal crash rate and the highest injury crash rate.

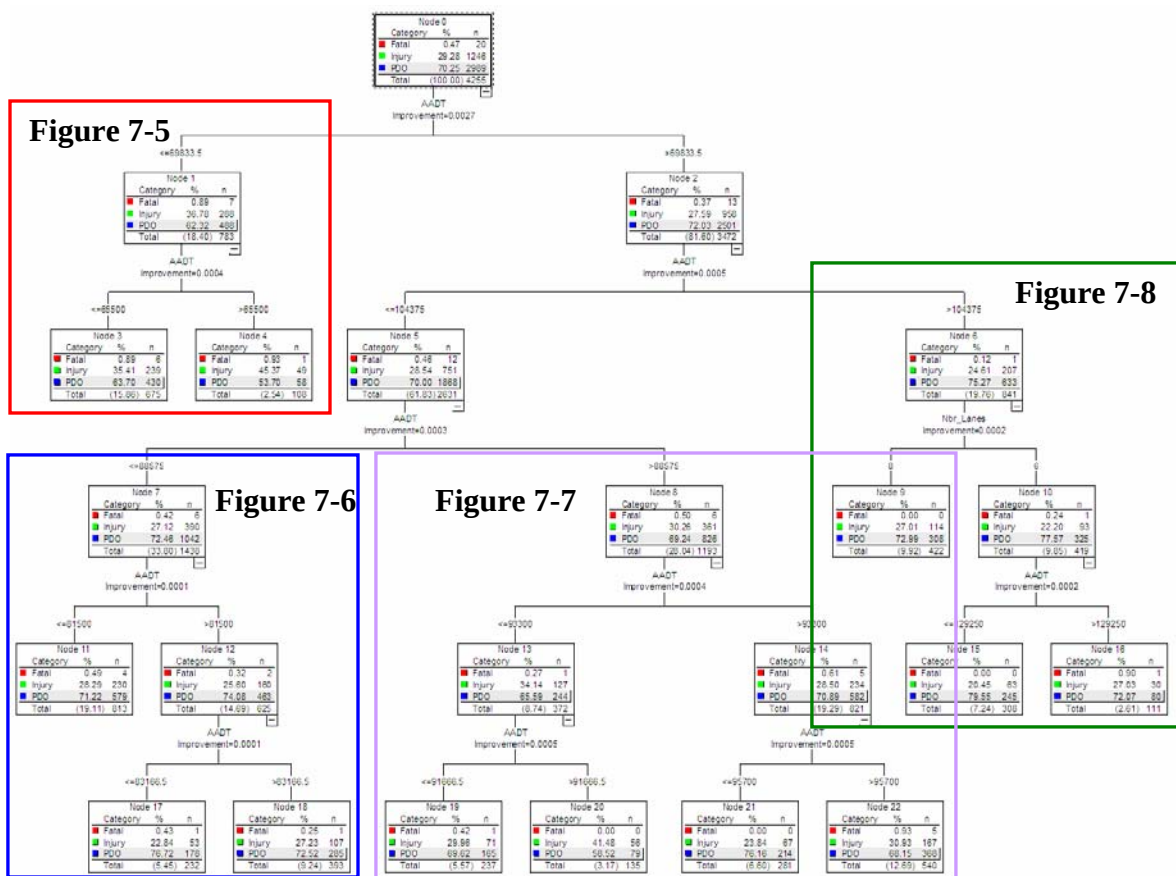


Figure 7-4 Classification Tree Model for Interchange Segments of North Carolina

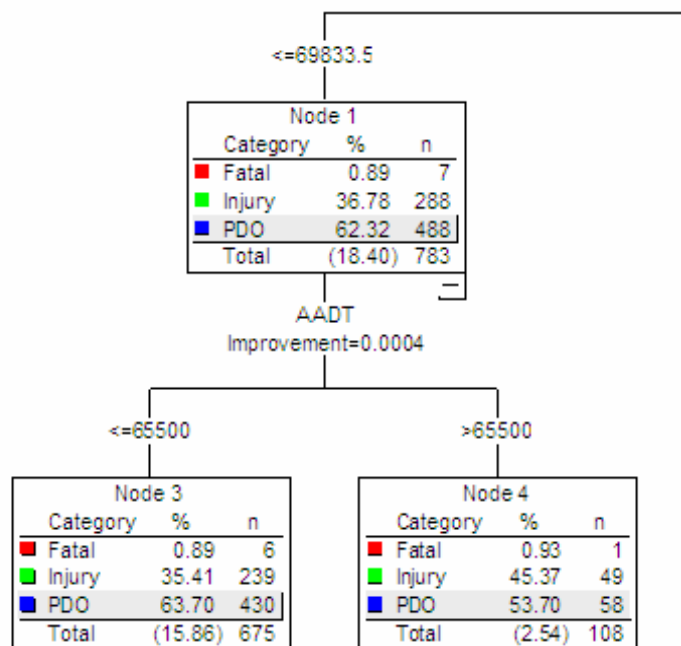


Figure 7-5 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume Less than or Equal to 69,834 Vehicles per Day

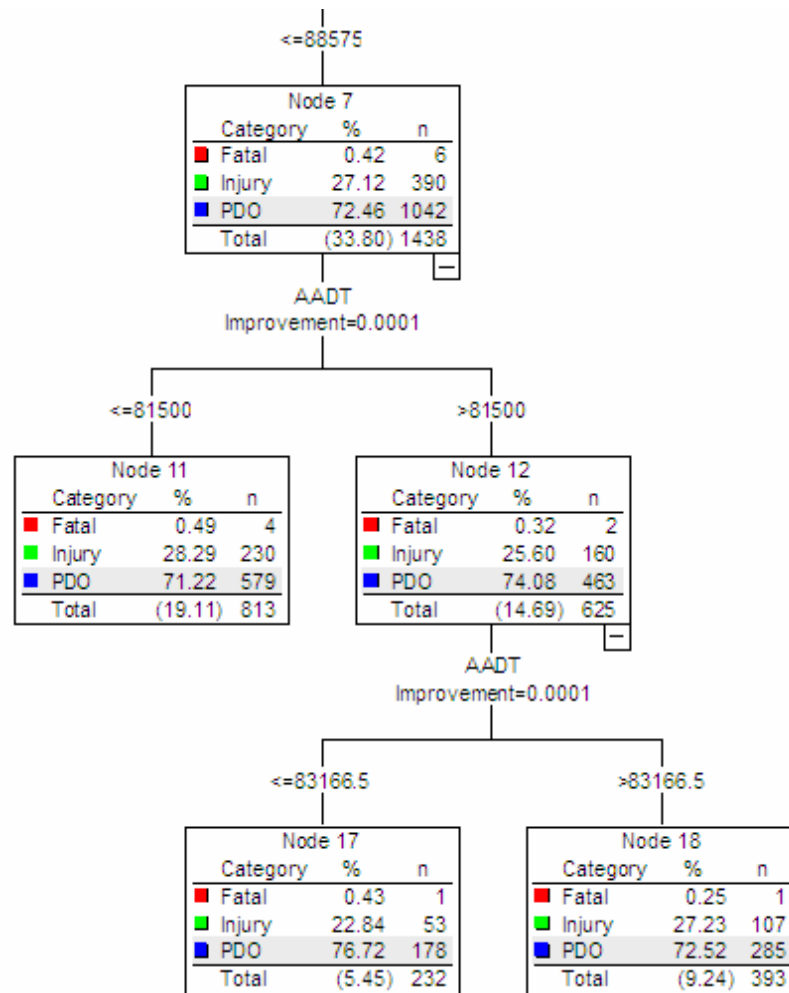


Figure 7-6 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume from 69,835 to 88,575 Vehicles per Day

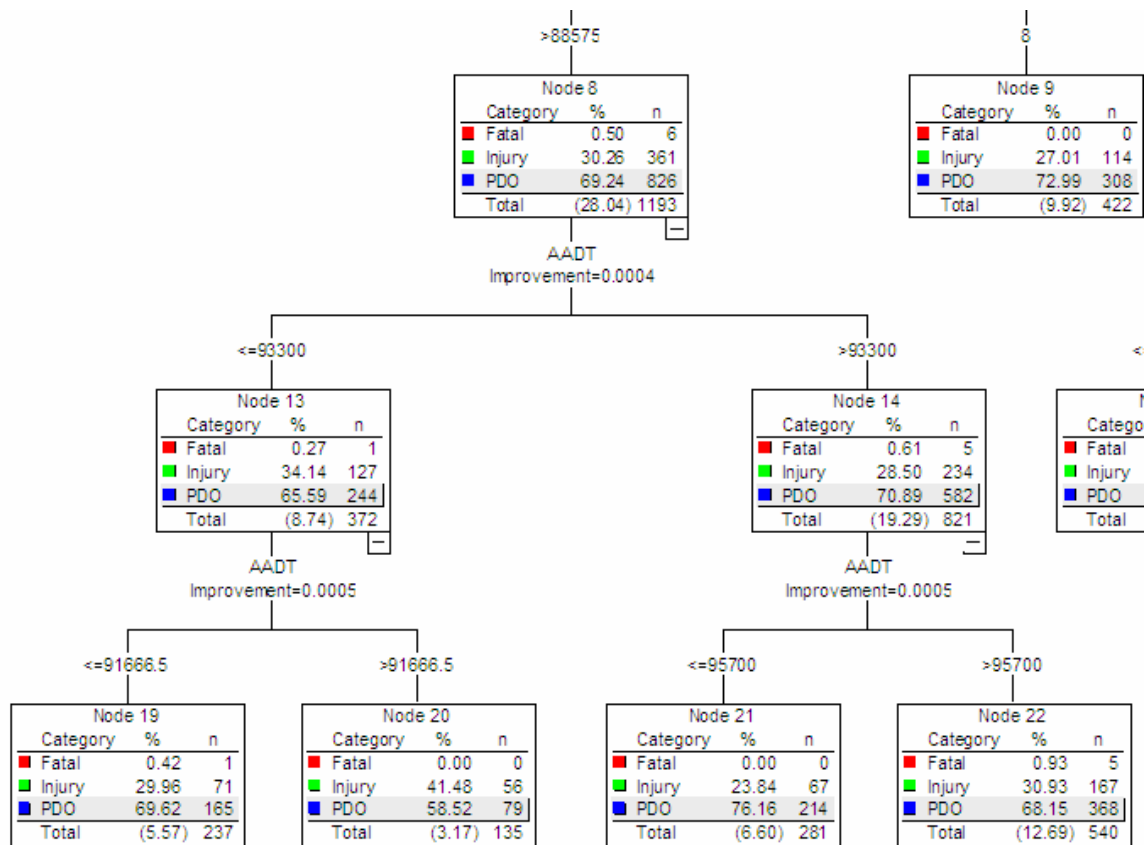


Figure 7-7 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume from 88,576 to 104,375 Vehicles per Day

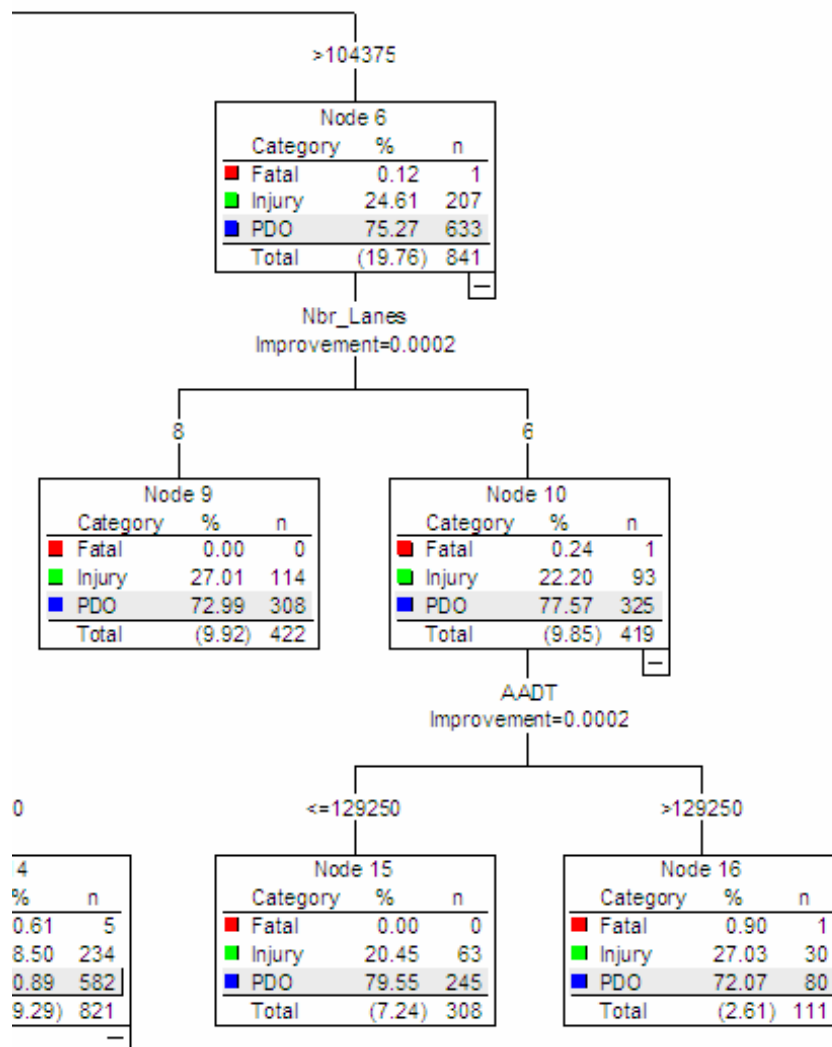


Figure 7-8 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume More than 104,375 Vehicles per Day

**Table 7-2 Crash Rates for the Classification Tree Model for Interchange Segments
of North Carolina**

No. of Lanes	Traffic Volume (Vehicles per day)	Tree Node	VMT (Millions for 3 Years)	Crash Rates (Crashes/MVMT)			
				Fatal	Injury	PDO	Fatal and Injury
All	<=65,500	3	616.34	0.0097	0.3878	0.6977	0.3975
All	65,501-69,834	4	71.86	0.0139	0.6819	0.8071	0.6958
All	69,835-81,500	11	377.09	0.0106	0.6099	1.5354	0.6205
All	81,501-83,167	17	153.68	0.0065	0.3449	1.1583	0.3514
All	83,168-88,575	18	265.44	0.0038	0.4031	1.0737	0.4069
All	88,576-91,667	19	168.70	0.0059	0.4209	0.9780	0.4268
All	91,668-93,300	20	114.85	0.0000	0.4876	0.6878	0.4876
All	93,301-95,700	21	176.39	0.0000	0.3798	1.2132	0.3798
All	95,701-104,375	22	547.87	0.0091	0.3048	0.6717	0.3139
6	104,376-129,250	15	156.48	0.0000	0.4026	1.5657	0.4026
6	>129,250	16	81.51	0.0123	0.3681	0.9815	0.3803
8	>104,375	9	291.49	0.0000	0.3911	1.0566	0.3911

Meanwhile, the highest PDO crash rate is found in node 15, which is a group of interchange segments with six lanes in the volume range of 104,376 to 129,250 vehicles per day. In addition, the highest total crash rate occurs in node 11, which represents group of interchange segments in the volume range of 69,835 to 81,500 vehicles per day.

Classification Tree Models for Tennessee Freeways

Classification trees were developed for analyzing crashes for non-interchange and interchange segments of Tennessee freeways. The tree model developed for non-interchange segments includes two variables including the traffic volume (or ‘AADT’) and the number of freeway lanes. Similar to the classification trees for North Carolina, all classification trees for Tennessee were limited to the maximum depth of five levels to minimize the complexity.

Classification Tree Model for Non-Interchange Segments of Tennessee

Figure 7-9 presents a classification tree for non-interchange segments of Tennessee and four enlarged areas of the tree are presented in Figure 7-10 through Figure 7-13. The segments were classified into 12 groups with different ranges of traffic volumes and numbers of freeway lanes. ‘AADT’ is found in 9 splits out of 11 splits while the number of lanes is found in 2 splits.

The classification tree stopped growing early on the left side and held node 3 and node 4 as a ‘leaf’ of the tree. This indicates that no further significant improvement is found by further classifying a group of non-interchange segments with traffic volumes less than or equal to 53,665 vehicles per day and another group with volumes between 53,666 and 54,490 vehicles per day. However, the stratification for other segments having volumes higher than 54,490 vehicles per day was fully developed. There are 12 nodes classified as ‘leaves’ of the classification tree in Figure 7-9.

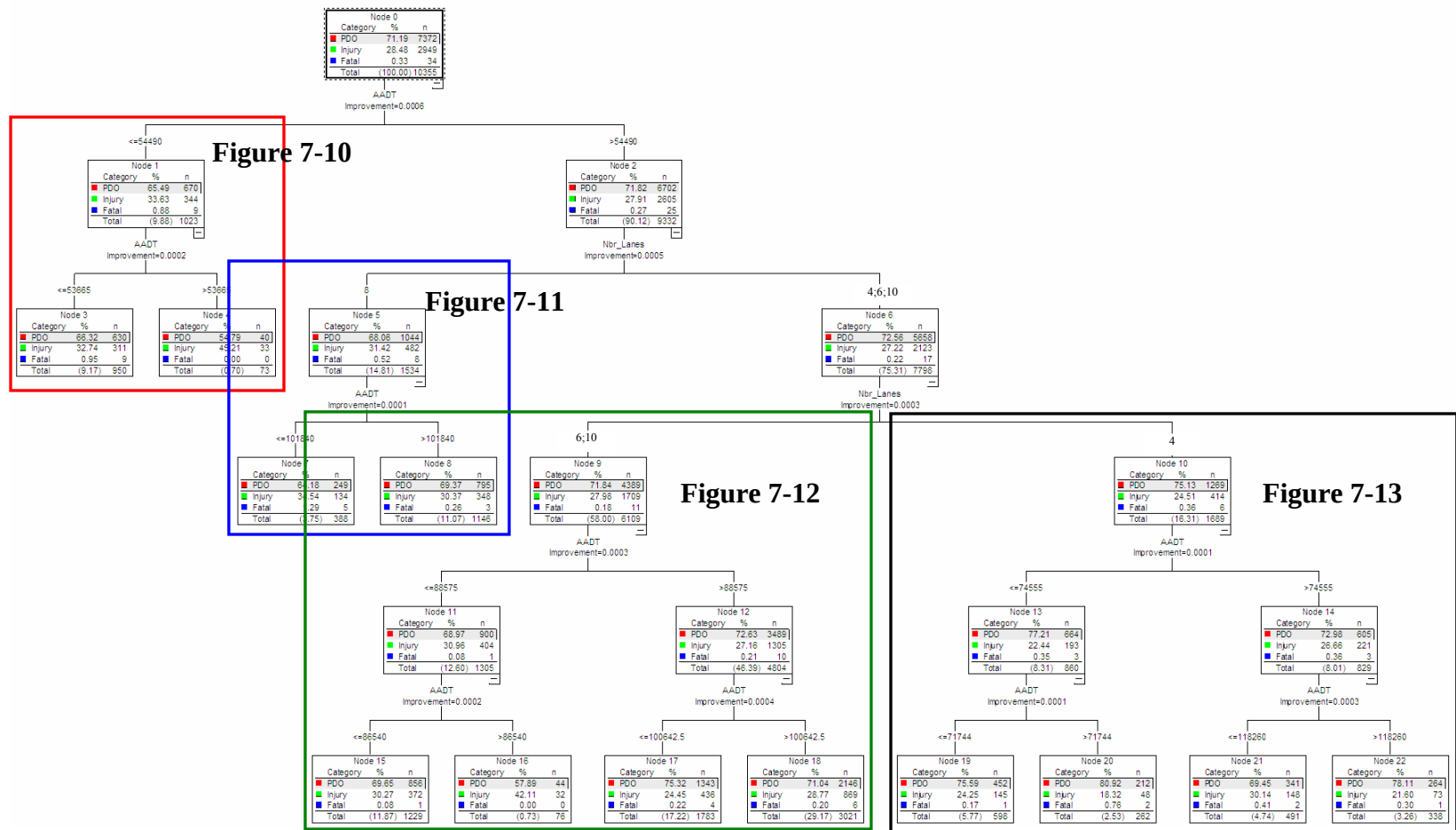


Figure 7-9 Classification Tree Model for Non-Interchange Segments of Tennessee

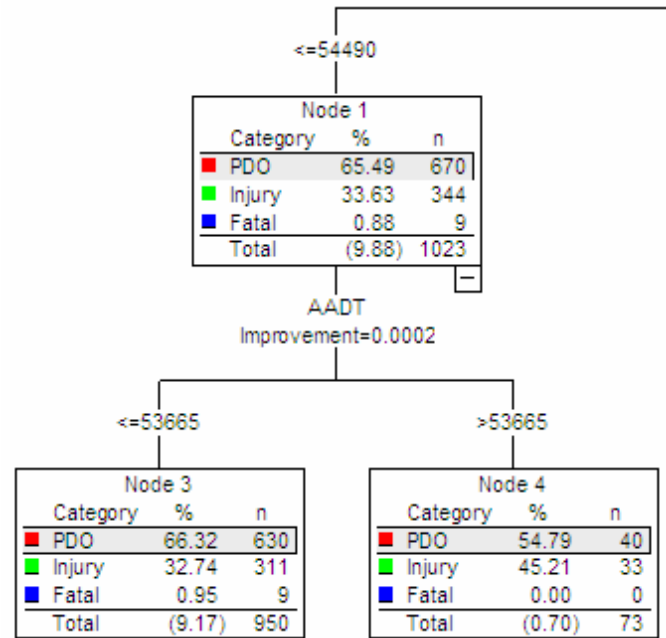


Figure 7-10 Enlarged Area of the Classification Tree Model for Non-Interchange Segments with Traffic Volume Less than or Equal to 54,490 Vehicles per Day

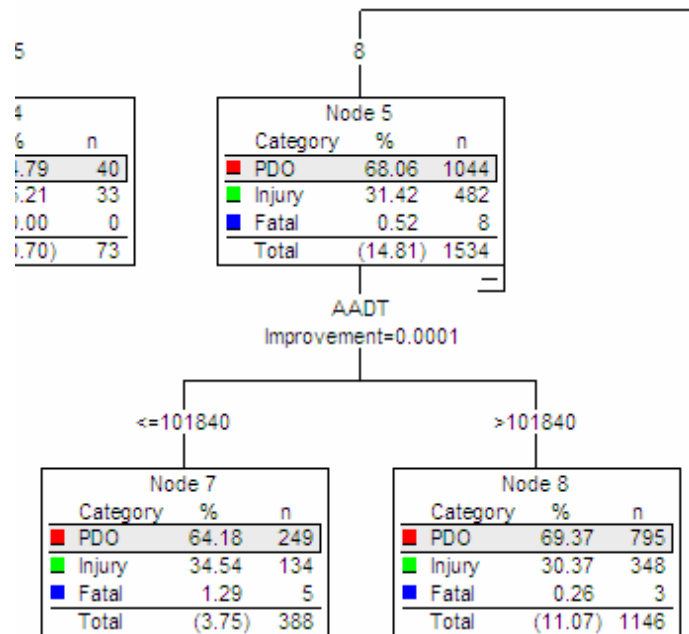


Figure 7-11 Enlarged Area of the Classification Tree Model for Eight-Lane Non-Interchange Segments with Traffic Volume More than 54,490 Vehicles per Day

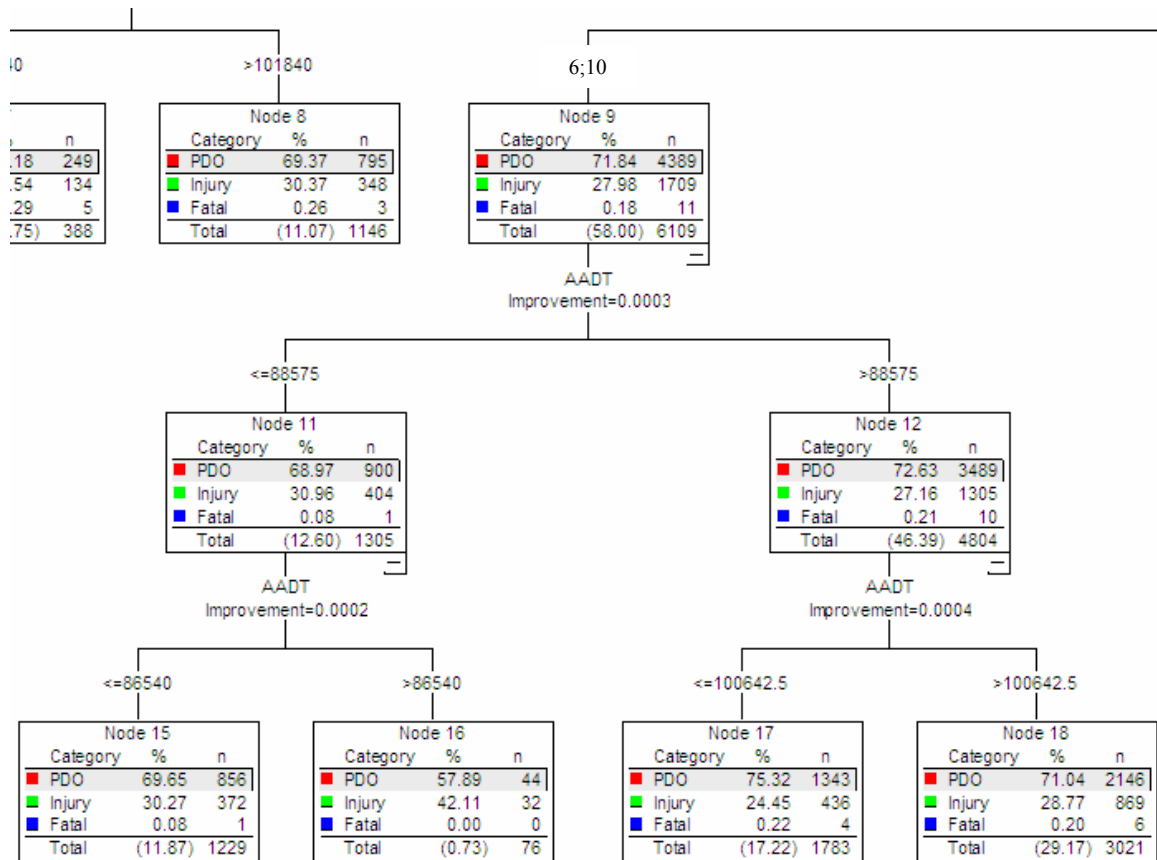


Figure 7-12 Enlarged Area of the Classification Tree Model for Six-Lane and Ten-Lane Non-Interchange Segments with Traffic Volume More than 54,490 Vehicles per Day

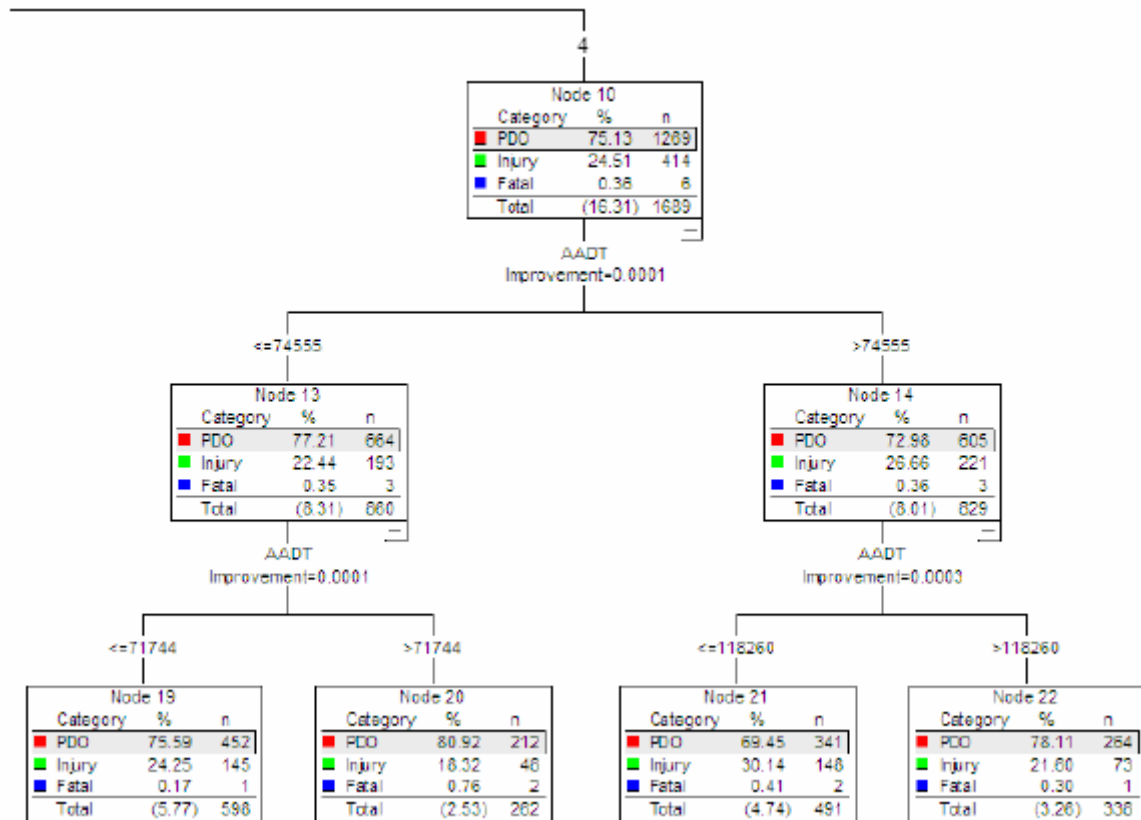


Figure 7-13 Enlarged Area of the Classification Tree Model for Four-Lane Non-Interchange Segments with Traffic Volume More than 54,490 Vehicles per Day

Similar to the trees for North Carolina freeways, the proportions of ‘PDO’ crashes, ‘injury’ crashes, and ‘fatal’ crashes for each group of non-interchange segments are similar to the overall proportion. As seen for in Figure 7-9 through Figure 7-13, ‘PDO’ crashes are found to be about 70% while ‘injury’ crashes are about 30% and ‘fatal’ crashes are less than 1% of the total for each node.

Table 7-3 presents the rates for ‘fatal’ crashes, ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes calculated for the classification tree for non-interchange segments of Tennessee. The highest fatal crash rate is found in node 20, which represents the group of segments with four freeway lanes and having traffic volumes between 71,745 and 74,555 vehicles per day. The injury crash rate and the PDO crash rate are highest in node 22, which represents the group of four-lane segments with traffic volume more than 118,260 vehicles per day. In addition, it is found that the segments in node 22 also have the highest total crash rate, which includes the rates for ‘fatal’, ‘injury’, and ‘PDO’ crashes combined.

Classification Tree Model for Interchange Segments of Tennessee

The classification tree model for interchange segments is shown in Figure 7-14. Three enlarged areas of the tree are presented in Figure 7-15 through Figure 7-17. The classification tree model classified interchange segments into 9 groups. The number of freeway lanes is found in one split while ‘AADT’ is found in all other splits of the tree. The traffic volume of 146,021 vehicles per day was specified by the tree as the cut-off volume for the first split. The classification tree classified a group of interchange segments with freeway traffic volumes more than 146,021 vehicles per day as a single group as shown in node 2 of the tree. This indicates that all interchange segments within node 2 are associated with the same level of crash risk.

**Table 7-3 Crash Rates for the Classification Tree Model for Non-Interchange
Segments of Tennessee**

No. of Lanes	Traffic Volume (Vehicles per day)	Tree Node	VMT (Millions for 3 Years)	Crash Rates (Crashes/MVMT)			
				Fatal	Injury	PDO	Fatal and Injury
All	<=53,665	3	1740.83	0.0052	0.1787	0.3619	0.1838
All	53,666-54,490	4	122.59	0.0000	0.2692	0.3263	0.2692
4	54,491-71,744	19	1045.26	0.0010	0.1387	0.4324	0.1397
4	71,745-74,555	20	273.09	0.0073	0.1758	0.7763	0.1831
4	74,556-118,260	21	608.77	0.0033	0.2431	0.5601	0.2464
4	>118,260	22	156.97	0.0064	0.4651	1.6818	0.4714
6;10	54,491-86,540	15	1963.85	0.0005	0.1894	0.4359	0.1899
6;10	86,541-88,575	16	242.76	0.0000	0.1318	0.1812	0.1318
6;10	88,576-100,643	17	2016.26	0.0020	0.2162	0.6661	0.2182
6;10	>100,643	18	4496.00	0.0013	0.1933	0.4773	0.1946
8	54,491-101,840	7	811.82	0.0062	0.1651	0.3067	0.1712
8	>101,840	8	1094.77	0.0027	0.3179	0.7262	0.3206

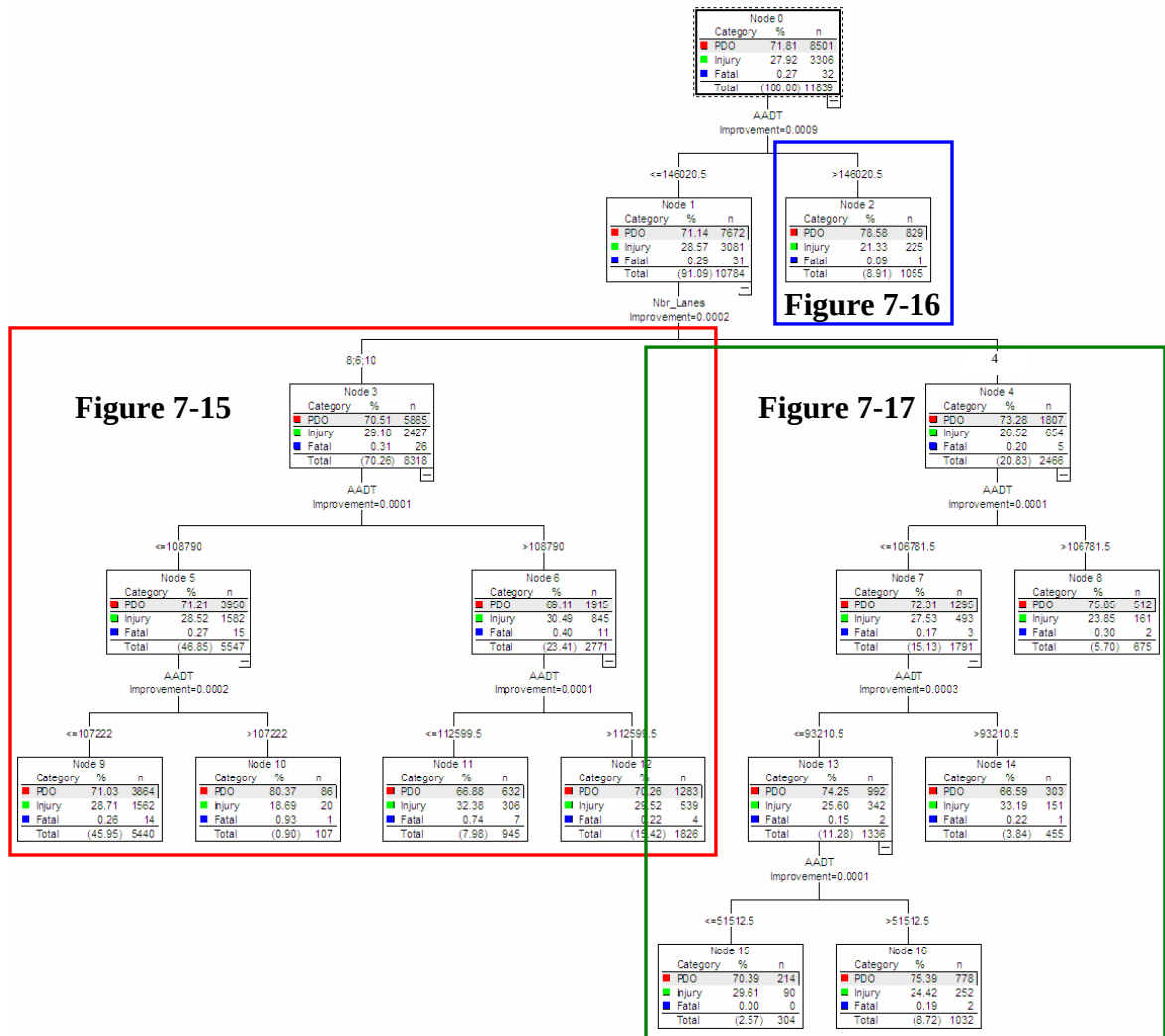


Figure 7-14 Classification Tree Model for Interchange Segments of Tennessee

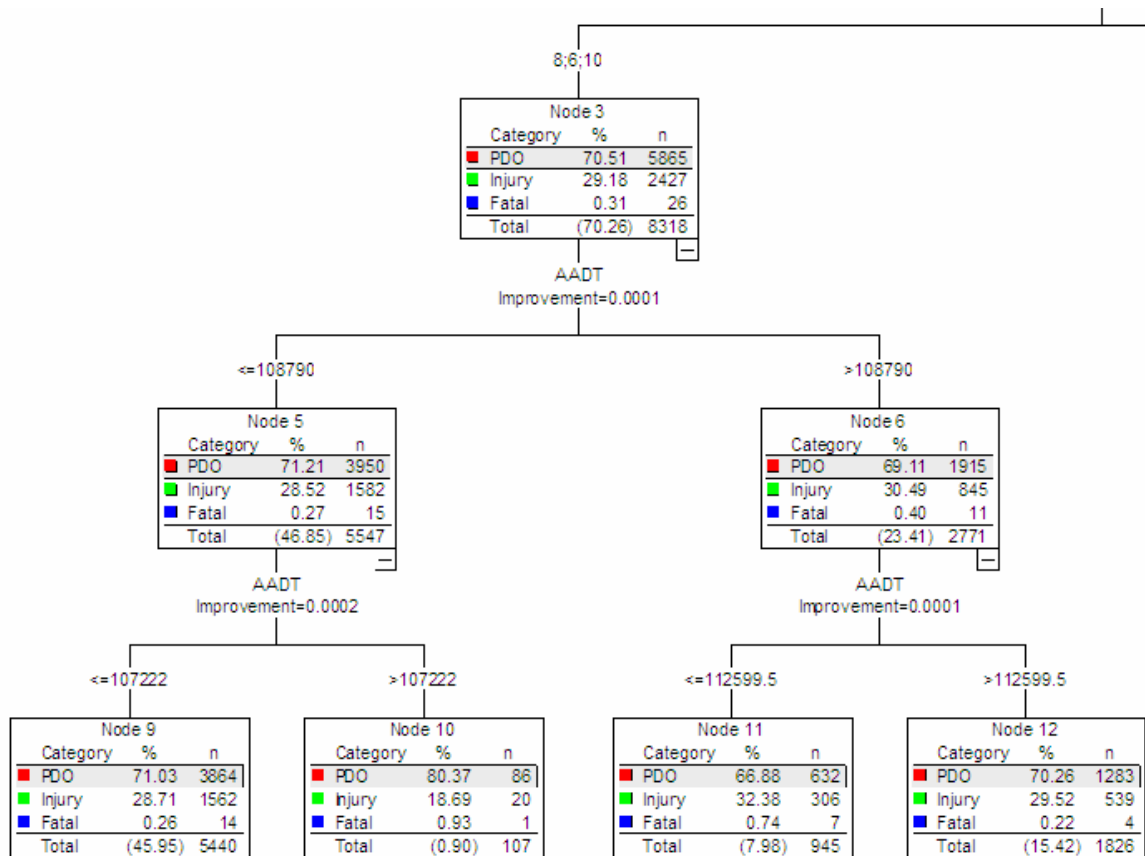


Figure 7-15 Enlarged Area of the Classification Tree Model for Six-Lane, Eight-Lane, and Ten-Lane Interchange Segments with Traffic Volume Less than or Equal to 146,021 Vehicles per Day

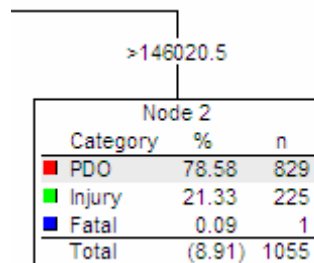


Figure 7-16 Enlarged Area of the Classification Tree Model for Interchange Segments with Traffic Volume More than 146,021 Vehicles per Day

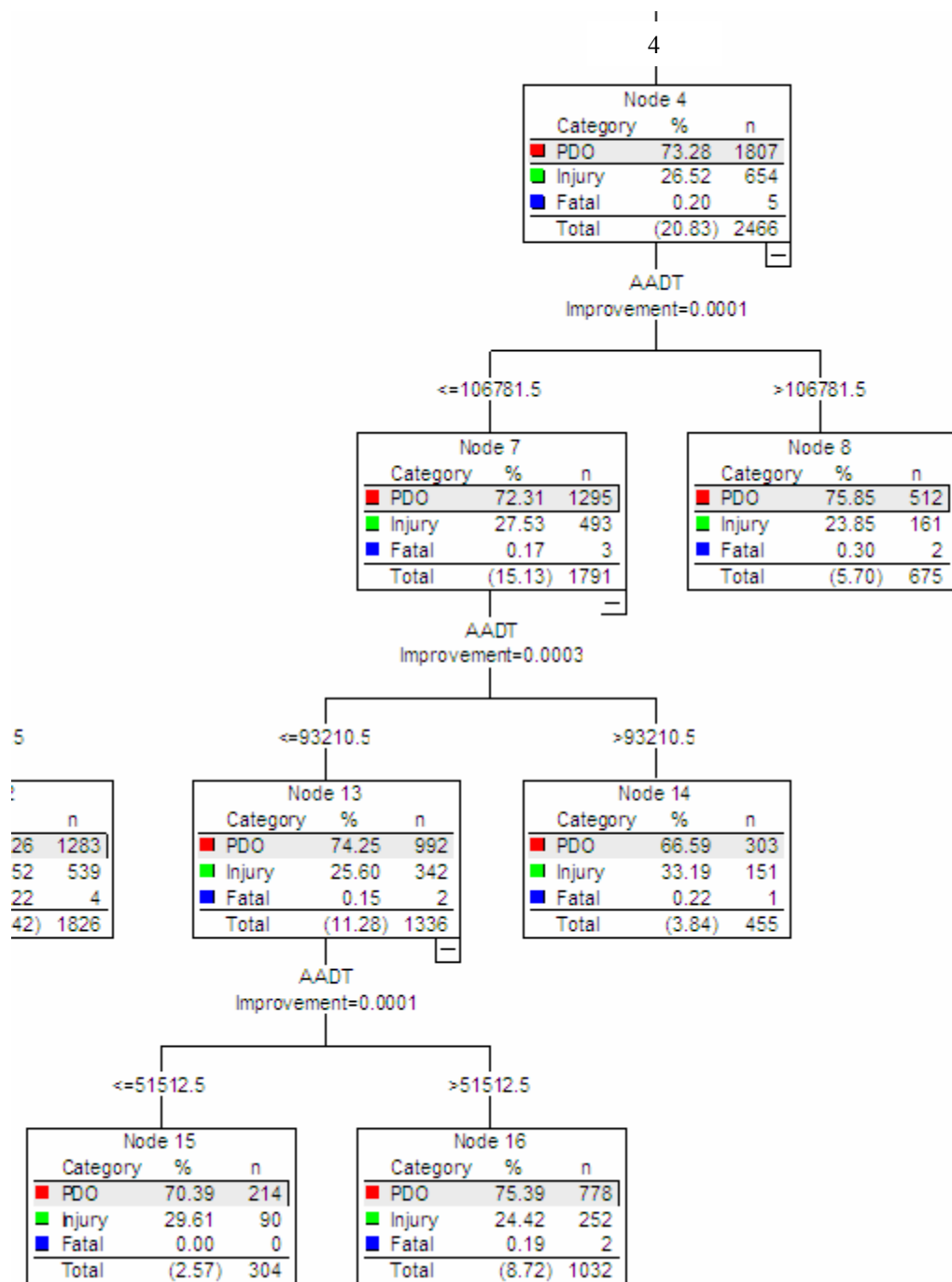


Figure 7-17 Enlarged Area of the Classification Tree Model for Four-Lane Interchange Segments with Traffic Volume Less than or Equal to 146,021 Vehicles per Day

It may be noted in Figure 7-14 that proportions of ‘PDO’ crashes, ‘injury’ crashes, and ‘fatal and injury’ crashes for all nodes are similar to the overall proportion for interchange segments of Tennessee. ‘PDO’ crashes vary between 70% and 80%, while ‘injury’ crashes vary between 20% and 30% and ‘fatal’ crashes are less than 1% of the total for each node.

Crash rates by severity were calculated and they are presented in Table 7-4. The highest rates for ‘fatal’ crashes and ‘injury’ crashes are found in node 11, which is presented in Figure 7-14 as the group of interchange segments with traffic volume in the range between 108,791 and 112,600 vehicles per day and with six, eight, or ten freeway lanes. The highest rate for ‘PDO’ crashes is found in the group of interchange segments within node 2, which represents the group of more-than-four-lane interchange segments with traffic volumes more than 146,021 vehicles per day. The group of segments in node 2 also has the highest total crash rate.

Table 7-4 Crash Rates for the Classification Tree Model for Interchange Segments of Tennessee

No. of Lanes	Traffic Volume (Vehicles per day)	Tree Node	VMT (Millions for 3 Years)	Crash Rates (Crashes/MVMT)			
				Fatal	Injury	PDO	Fatal and Injury
4	<=51,513	15	179.45	0.0000	0.5015	1.1926	0.5015
4	51,514-93,211	16	753.63	0.0027	0.3344	1.0323	0.3370
4	93,212-106,782	14	285.14	0.0035	0.5296	1.0626	0.5331
4	>106,782	8	282.86	0.0071	0.5692	1.8101	0.5763
>4	<=107,222	9	2677.81	0.0052	0.5833	1.4430	0.5885
>4	107,223-108,790	10	104.08	0.0096	0.1922	0.8263	0.2018
>4	108,791-112,600	11	384.91	0.0182	0.7950	1.6419	0.8132
>4	112,601-146,021	12	824.83	0.0048	0.6535	1.5555	0.6583
>4	>146,021	2	385.27	0.0026	0.5840	2.1517	0.5866

Conclusions

It was found that the classification tree model, which is a common data mining technique, can be used to provide useful model forms for crash prediction purposes. Crash rate tables developed using classification tree models can be applied in a manner similar to the tables for ANOVA models; both of these can be used as look-up tables.

Unlike ANOVA and regression models, classification tree models graphically represent all classified groups of freeway segments and numbers of crashes by severity type – ‘fatal’, ‘injury’, and ‘PDO’. This feature makes it easy for transportation planners and practitioners to assess safety consequences for freeway segments. However, it should be pointed out that the number of crashes presented in each node of the classification tree for this analysis is the total of three years. To predict the number of crashes for individual segment, one should use corresponding crash rates from a crash-rate table.

It was found for the classification tree models for non-interchange segments of North Carolina and non-interchange segments of Tennessee that both of them did not classify segments with ‘AADT’ less than 55,000 and 54,490 vehicles per day into four-lane, six-lane, or eight-lane segments. The number of lanes was used for classifying non-interchange segments with ‘AADT’ higher than 55,000 and 54,490 vehicles per day, for North Carolina and Tennessee respectively. However, the same pattern was not found for the classification tree models for interchange segments of North Carolina and interchange segments of Tennessee. In addition, the highest rates for different crash types were found for different groups. The highest total crash rate for non-interchange segments of North Carolina was found on eight-lane segments with traffic volume higher than 112,500 vehicles per day. The highest total crash rate for interchange segments of North Carolina was found on segments with traffic volume in range of 69,835 to 81,500 vehicles per day. The highest total crash rate for non-interchange segments of Tennessee was found on four-lane segments with traffic volume higher than 118,260 vehicles per day. And the

highest total crash rate for interchange segments of Tennessee was found on segments with traffic volume higher than 146,021 vehicles per day.

An assessment of classification tree models and the prediction performance will be compared to those for ANOVA models and regression models and the results will be presented in the following chapter.

8. ASSESSMENT AND COMPARISON OF CRASH PREDICTION MODELS FOR FREEWAYS OF NORTH CAROLINA AND TENNESSEE

Different forms of crash prediction models for freeway segments based on different analytical approaches were developed and presented in preceding chapters. In order to see how well the models can perform, comparisons of model predictions with actual crashes were made. It should be pointed out that the predictions of ANOVA models and classification tree models are represented in terms of ‘crash rate’. Therefore, Equation 4-1 was used to convert the rates to ‘number of crashes’. However, the prediction of regression models is in terms of ‘number of crashes’ and so no conversion is needed for these models.

The difference between predicted crashes and actual crashes is assessed in terms of the root mean squared error (RMSE). The performance of crash prediction models is measured by their ability to replicate actual crashes, and the lower the value of RMSE the better. The formula for RMSE is presented in Equation 8-1.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=0}^N (\text{Actual Crashes}_i - \text{Predicted Crashes}_i)^2} \quad (8-1)$$

i = An individual segment

N = Numbers of segments

RMSE was calculated for four separate groups of freeway segments, one for non-interchange segments with four lanes, one for non-interchange segments with more than four lanes, one for interchange segments with four lanes, and another for interchange segments with more than four lanes. For each group of segments, the RMSE was calculated for each of three different crash types. The three crash types based on severity are ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes.

The RMSE for each group of segments was further converted to ‘percent error’, which represents an average percent error based on a ‘number of crashes occurring on an average segment’. The ‘percent error’ was calculated using Equation 8-2.

$$\text{Percent Error} = \frac{\text{RMSE}}{\text{Number of Crashes Occuring on an Average Segment}} \times 100 \quad (8-2)$$

The ‘number of crashes occurring on an average segment’ was calculated by dividing all actual crashes by the number of segments within the group. For example, the number of ‘injury’ crashes occurring on an average non-interchange segment with four lanes was calculated by dividing all actual ‘injury’ crashes occurring on non-interchange segments with four lanes by the number of those segments. The purpose of calculating ‘percent error’ is to give a ‘relative measure’ of the error, which is easy to comprehend.

The crash prediction models for segments with four lanes are labeled as “models for four-lane”. These models are ‘ANOVA model for four-lane’ and ‘regression model for four-lane’. The other group of crash prediction models for freeways with more than four lanes is labeled as “models for more-than-four-lane”. These models are ‘ANOVA model for more-than-four-lane’ and ‘regression model for more-than-four-lane’. The classification tree model separated segments into sub-categories – four-lane, six-lane, and eight-lane.

Assessment and Comparison of Crash Prediction Models Developed for North Carolina Freeways

Each model was used to estimate crashes for each individual segment, and RMSE was calculated for each case. Next, the percent error for each model was calculated using Equation 8-2. Table 8-1 presents the number of actual crashes by severity type and the number of segments for North Carolina data that were used for calculating the ‘number of crashes occurring on an average segment’, which was used for calculating ‘percent error’.

**Table 8-1 Number of Actual Crashes by Severity and Number of Segments for
North Carolina Freeways**

Type of Segments	Number of Actual Crashes			Number of Segments
	Injury	PDO	Fatal and Injury	
Non-Interchange Segments				
Four-Lane	647	1,192	661	36
More-than-Four-Lane	708	1,596	724	37
Interchange Segments				
Four-Lane	574	1,250	583	33
More-than-Four-Lane	764	1,840	772	32

For example, there are 647 ‘injury’ crashes and 36 non-interchange segments with four lanes as shown in Table 8-1. Thus, the number of injury crashes occurring on an average non-interchange segment with four lanes equals to 18 crashes.

Assessment and Comparison of Models for Non-Interchange Segments with Four Lanes

Percent errors for different prediction models developed for non-interchange segments with four lanes in North Carolina are presented in Figure 8-1. It can be seen that classification tree model performs better than others for predicting ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes with percent error varying between 36.46% and 39.42%. Meanwhile, the ANOVA model is found to have the highest percent error varying between 46.77% and 66.06%. For this sub-group, the regression model is in the middle of the three models with regard to its performance for estimating crashes.

Further, it is noticed for each modeling approach that the models for estimating ‘PDO’ crashes performed better than those for estimating ‘injury’ crashes and ‘fatal and injury’ crashes.

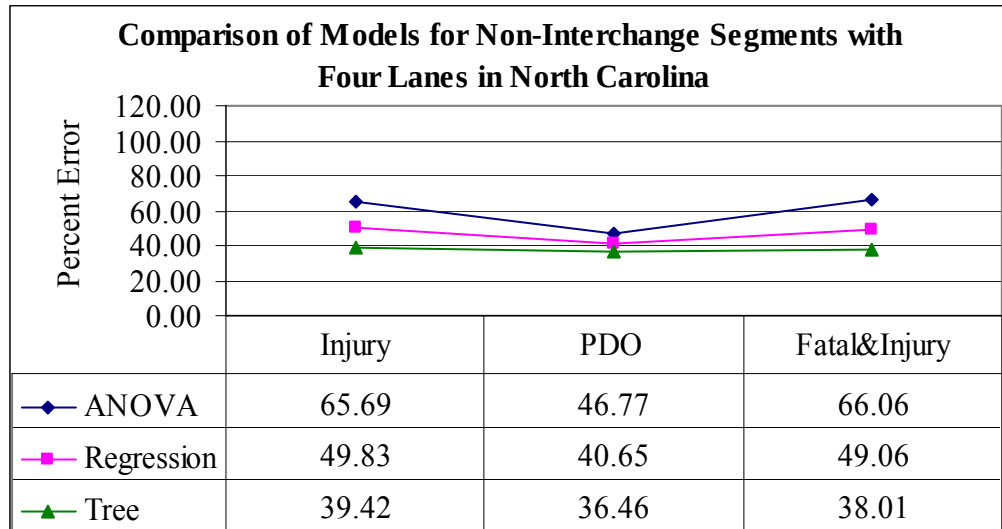


Figure 8-1 Comparison of the Models for Non-Interchange Segments with Four Lanes of North Carolina

Assessment and Comparison of Models for Non-Interchange Segments with More than Four Lanes

The data used for developing ANOVA and regression models for segments with more than four lanes include both six-lane and eight-lane segments, and these models did not distinguish among them. However, the classification tree model separated the segments with more than four lanes into six-lane and eight-lane segments respectively, and for this model, crash rates were developed separately for six-lane and eight-lane segments. For calculating percent error for the classification tree model, the crash rates for six-lane segments were used to estimate the number of crashes on six-lane segments and the crash rates for eight-lane segments were used to estimate the number of crashes on eight-lane segments.

The percent errors calculated for the models developed for the group of segments with more than four lanes are presented in Figure 8-2. It is apparent that the classification tree model has the lowest percent error for estimating three crash types.

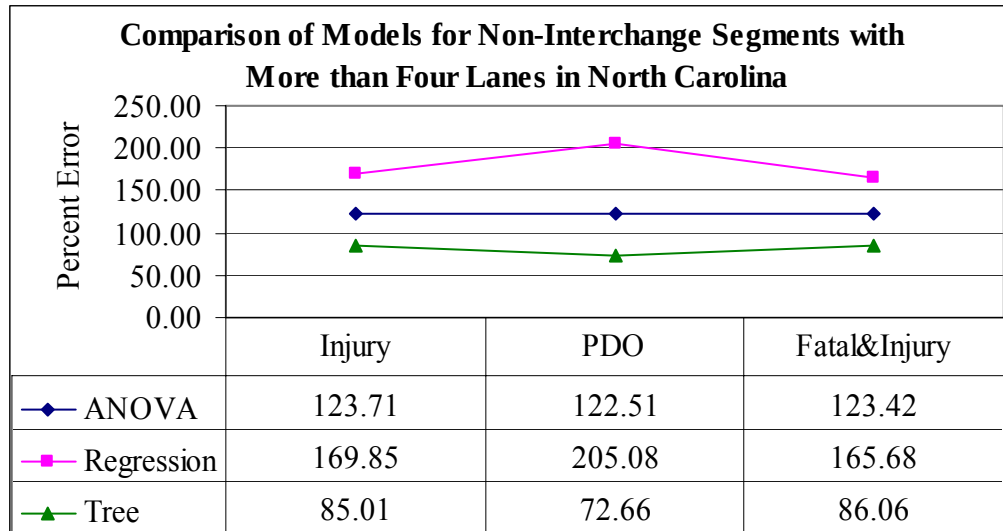


Figure 8-2 Comparison of the Models for Non-Interchange Segments with More than Four Lanes of North Carolina

Further, the regression model is found to have worst performance with its percent error higher than those for the other two models for estimating all three crash types. In addition, it is found that the models for estimating ‘injury’ crashes and ‘fatal and injury’ crashes have percent error close to each other.

Assessment and Comparison of Models for Interchange Segments with Four Lanes

The percent error for different crash prediction models for interchange segments with four lanes is presented in Figure 8-3. All three models are found to have similar performance for estimating all crash types with the percent error varying between 52.90% and 59.13%.

Although the models developed from different approaches perform similarly for predicting all crash types, it is noticed that, for the same model approach, the models for estimating ‘PDO’ crashes perform slightly better than those for estimating ‘injury’ crashes and ‘fatal and injury’ crashes.

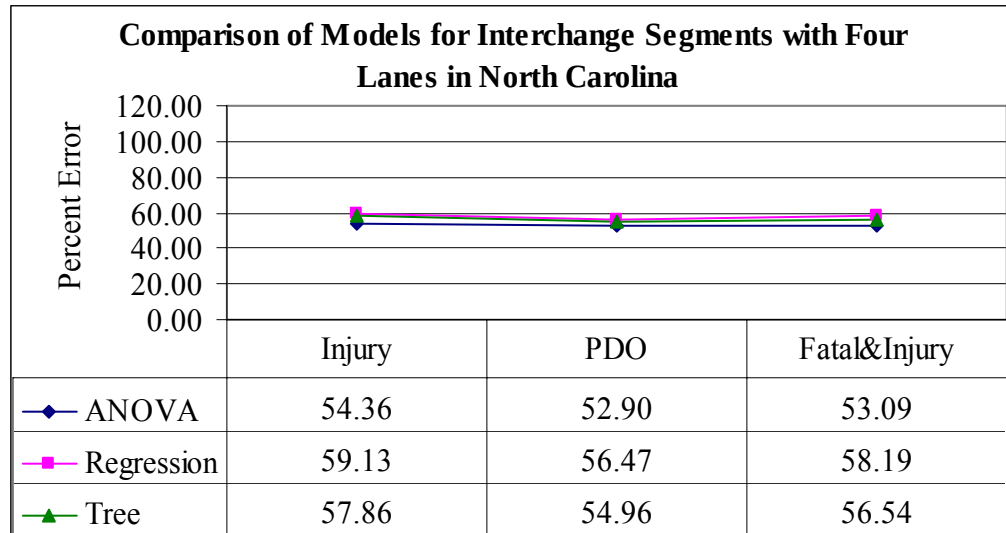


Figure 8-3 Comparison of the Models for Interchange Segments with Four Lanes of North Carolina

Further, the prediction performance of the models for estimating ‘injury’ crashes and that for the models for estimating ‘fatal and injury’ crashes are found to be similar. The percent error for the models for ‘PDO’ crashes varies between 52.90% and 56.47% while that for the models for ‘injury’ crashes and ‘fatal and injury’ crashes varies between 53.09% and 59.13%.

Assessment and Comparison of Models for Interchange Segments with More than Four Lanes

The percent error for crash prediction models for interchange segments with more than four lanes are presented in Figure 8-4. It can be seen that all the three models have similar performance for predicting all crash types for this group of segments for North Carolina freeways. The percent error for these models varies between 53.87% and 56.71% for estimating ‘injury’ crashes, between 53.95% and 56.58% for estimating ‘PDO’ crashes, and between 53.34% and 56.11% for estimating ‘fatal and injury’ crashes.

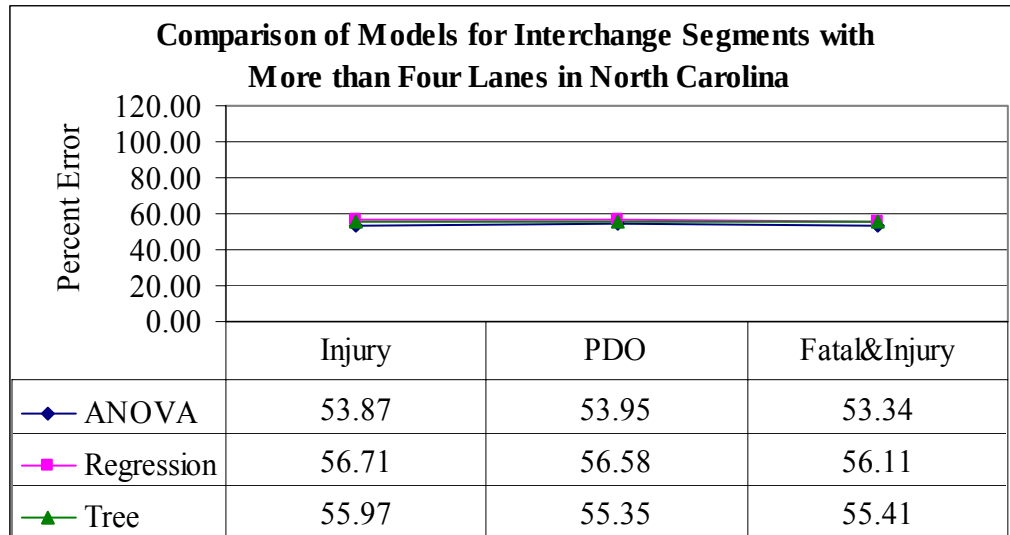


Figure 8-4 Comparison of the Models for Interchange Segments with More than Four Lanes of North Carolina

Further, for each modeling approach, the models for ‘injury’ crashes, ‘PDO’ crashes, and ‘fatal and injury’ crashes respectively are found to have almost identical performance.

Assessment and Comparison of Crash Prediction Models Developed for Tennessee Freeways

The assessment of crash prediction models for Tennessee freeways was performed in a similar manner to those for North Carolina. Similar groups of freeway segments were used for each state.

Each model was used to estimate crashes for each individual segments and RMSE was calculated, which was further used to calculate ‘percent error’ using Equation 8-2. Table 8-2 presents the number of actual crashes by severity and the number of segments for Tennessee data that were used for calculating ‘number of crashes occurring on an average segment’, which was used in Equation 8-2.

Table 8-2 Number of Actual Crashes by Severity and Number of Segments for Tennessee Freeways

Type of Segments	Number of Actual Crashes			Number of Segments
	Injury	PDO	Fatal and Injury	
Non-Interchange Segments				
Four-Lane	585	1,516	599	36
More-than-Four-Lane	2,436	6,070	2,477	106
Interchange Segments				
Four-Lane	389	1,035	395	15
More-than-Four-Lane	2,531	6,374	2,572	50

Assessment and Comparison of Models for Non-Interchange Segment with Four Lanes

The percent error for different crash prediction models developed for non-interchange segments with four lanes in Tennessee is presented in Figure 8-5. Regression models and classification tree models are found to have fairly similar performance for estimating three crash types and these models perform better than ANOVA models. The percent error for regression and classification tree models varies between 70.83% and 85.07% while that for ANOVA models varies between 86.91% and 98.50%.

Further, for each approach, it is found that the models for estimating ‘PDO’ crashes perform better than those for other crash types. Meanwhile, the models for ‘injury’ crashes and the models for ‘fatal and injury’ crashes are found to have similar performance. The percent errors for models for ‘injury’ crashes and models for ‘fatal and injury’ crashes vary between 82.67% and 98.50% while those for models for ‘PDO’ crashes vary between 70.83% and 86.91%.

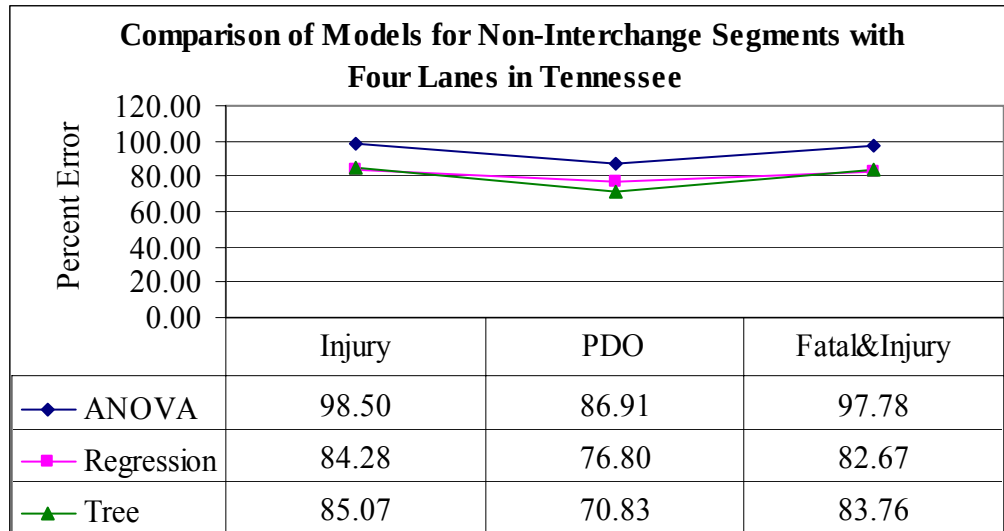


Figure 8-5 Comparison of the Models for Non-Interchange Segments with Four Lanes of Tennessee

Assessment and Comparison of Models for Non-Interchange Segments with More than Four Lanes

The percent error for different models developed for non-interchange segments with more than four lanes in Tennessee is presented in Figure 8-6. Overall, the models developed for this group of segments have fairly similar performance for estimating three crash types with percent errors varying between 79.57% and 101.14%. It is found that the ANOVA model has the lowest percent error for estimating ‘injury’ crashes and ‘fatal and injury’ crashes while the classification tree model has the lowest percent error for estimating ‘PDO’ crashes. In addition, it is found for each modeling approach that the models for ‘injury’ crashes and those for ‘fatal and injury’ crashes have the percent error close to each other, and these values are lower than those for the models for ‘PDO’ crashes.

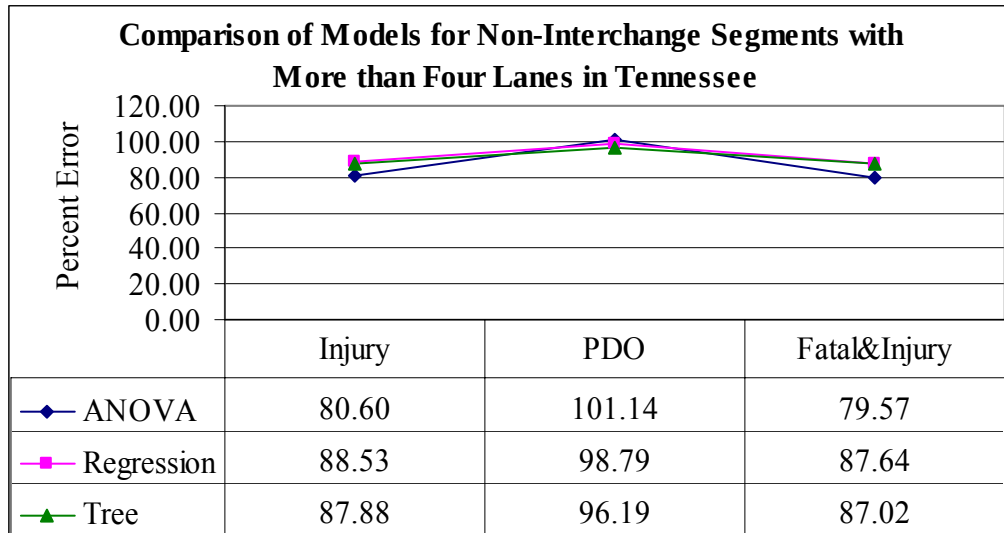


Figure 8-6 Comparison of the Models for Non-Interchange Segments with More than Four Lanes of Tennessee

Assessment and Comparison of Models for Interchange Segments with Four Lanes

Comparison of the model performance in terms of the percent error is presented in Figure 8-7 for the case of four-lane interchange segments. The ANOVA model apparently has the lowest performance while the regression model is found to have the highest performance with percent error varying between 45.00% and 56.96%. Further, for each modeling approach, it is found that the models for estimating ‘injury’ crashes and those for estimating ‘fatal and injury’ crashes have the percent error close to each other, and these models are found to perform better than those for estimating ‘PDO’ crashes.

Assessment and Comparison of Models for Interchange Segments with More than Four Lanes

The assessment of the models developed for interchange segments with more than four lanes is presented in Figure 8-8. The models are found to have percent error varying between 54.23% and 60.97%.

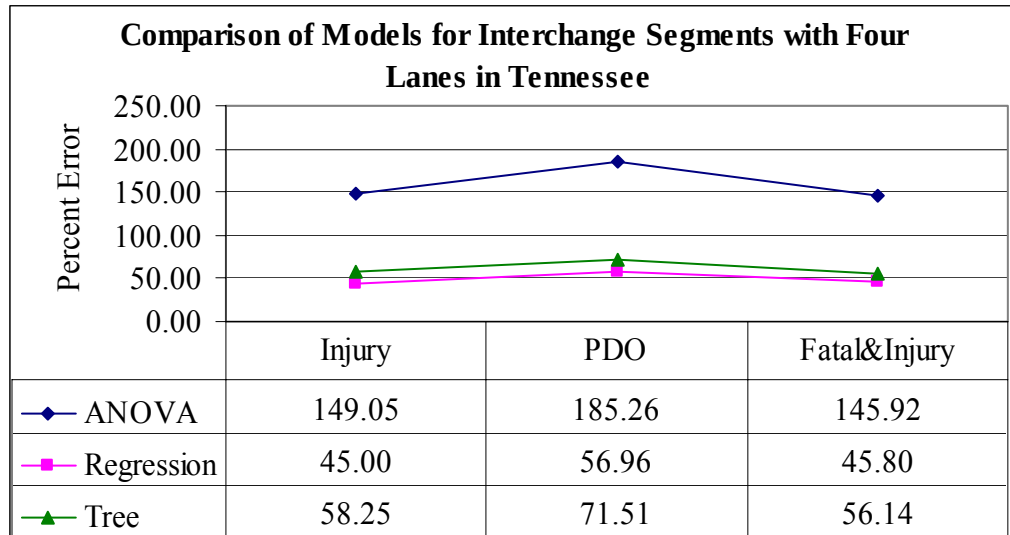


Figure 8-7 Comparison of the Models for Interchange Segments with Four Lanes of Tennessee

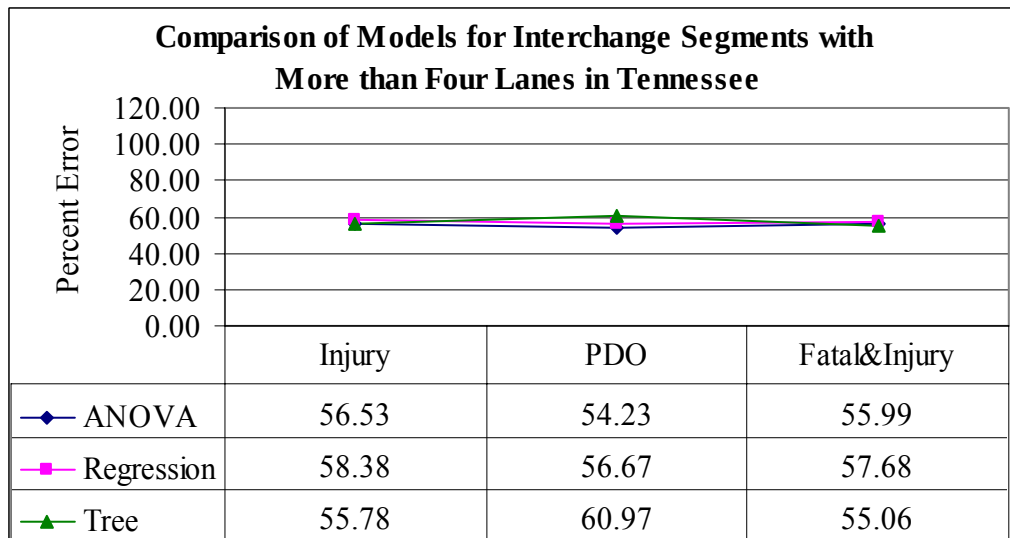


Figure 8-8 Comparison of the Models for Interchange Segments with More Than Four Lanes of Tennessee

The classification tree model has the lowest percent error for estimating ‘injury’ crashes and also ‘fatal and injury’ crashes. The ANOVA model has the lowest percent error for estimating ‘PDO’ crashes. However, the performance of all the models is fairly similar.

Conclusions

The assessment of how well each model can predict different types of crashes was performed using the measure of ‘percent error’, which was derived from ‘root mean squared error’. Overall, the different models were found to have fairly close performance in predicting different types of crashes although the approach and formulation of these models are quite different. However, for the groups of non-interchange segments with four lanes and non-interchange segments with more than four lanes in North Carolina, classification tree models were found to have better performance than others for estimating different types of crashes.

For the group of interchange segments with four lanes in Tennessee, regression model was found to have the best performance. Further, it was found that the models developed for interchange segments performed better than those developed for non-interchange segments except those for non-interchange segments with four lanes in North Carolina.

For each modeling approach, the models for estimating ‘injury’ crashes and those for estimating ‘fatal and injury’ crashes were found to have fairly similar performance in all cases.

9. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

General Conclusions

This dissertation was motivated by the recent emphasis on including safety consideration in long-range transportation planning process. Safety has been included explicitly in the seven planning factors specified for metropolitan planning by the Transportation Equity Act for Twenty First Century (TEA-21). The literature review indicated that majority of researchers in the highway safety area have been interested in developing crash prediction tools for non-interstate highways, and most of these crash prediction models are suitable for corridor and localized applications, which require detailed data on roadway characteristics and geometrics. Since detailed data are not available for long-range highway alternatives to be evaluated, most of these existing models are difficult to use for long-range planning.

There were two main objectives of this dissertation. The first objective was to understand the pattern of crash occurrences for urban freeways as well as the influence of interchanges on these incidences. The second objective was to develop and test different techniques/approaches for developing simple forms of crash prediction models for estimating crashes by crash severity for future urban freeways based on variables that can be predicted for long-range highway alternatives.

There were three main tasks for this research. The first task was to prepare and link crash and freeway characteristics data into formats needed for developing models. The second task involved the development of the models based on different techniques – ANOVA models, regression models, and classification tree models. From previous studies for developing crash prediction models for freeways, it was found that none of them developed separate models for two different areas on freeways – areas near interchanges and areas away from interchanges. This finding led to one of the main hypothesis that

non-interchange segments and interchange segments should have different levels of crash risk. This hypothesis was found to be true and it led to the task that developed separate crash prediction models for non-interchange segments and interchange segments. An investigation on the influence of interchanges on crash occurrences on freeways was performed during the process of developing ANOVA models. The third task was to assess the prediction performance of each model by comparing model generated results with actual crashes.

The crash and freeway traffic data used in this study were collected during the period between the year 2000 and the year 2002 by North Carolina Department of Transportation (NCDOT) and Tennessee Department of Transportation (TDOT). Data from six counties in North Carolina and four counties in Tennessee that include urbanized areas were used in this study. The proportions of crashes by severity type for freeways were found to be similar in all areas of both states. These proportions were 70% for PDO crashes, 30% for injury crashes, and less than 1% for fatal crashes. However, the average rates for non-fatal crashes for non-interchange segments of Tennessee were found to be at least twice as high as those for North Carolina whereas fatal crash rates for non-interchange segments were equal for both states. Further, fatal crash rates and PDO crash rates for interchange segments of Tennessee were significantly higher than those for North Carolina while there was no significant difference for injury crash rates for interchange segments. This indicated that crash risk is likely to be unique for each state, and therefore, separate crash prediction models for each should be developed.

It was found that the presence of interchanges usually affects crash potential for traffic on freeways. Crash rates for interchange segments were found to be at least twice as high as those for non-interchange segments. This finding supported the approach for treating non-interchange segments and interchange segments as different types and developing separate models for each of them. For each type of freeway segments, the number of

freeway lanes was used as a criterion for further separating them into a group with four lanes and another with more than four lanes.

The development of accurate crash prediction models for long-range transportation planning was difficult due to the limited availability of data that can be used as independent variables. Difficulty for making forecasts of independent variables was a major concern. Typically, information on traffic volumes, number of freeway lanes, and vehicle miles traveled (VMT) are available for long-range highway plans. Only these variables were used in all forms of models in this research.

Three different approaches for developing crash prediction models were pursued in this research. These are ANOVA models, regression models, and classification tree models. In addition, the prediction performance was assessed and measured in terms of the percent error.

For the first modeling approach, ANOVA models were developed. The factor freeway traffic volume was classified into ranges increasing at 25,000 vehicles per day. The factor type of segments was classified into non-interchange segments and interchange segments. And the factor number of freeway lanes was classified into four-lane and more-than-four-lane. There was a difficulty with developing ANOVA models for estimating ‘fatal’ crash rates due to small numbers of fatal crashes on individual segment, which result in very small values of fatal crash rates. Models for ‘fatal and injury’ crashes combined were used to circumvent this difficulty.

For the second modeling approach, regression models with the negative binomial distribution assumption for crashes were developed. The variables ‘freeway segment length’ and ‘traffic volume’ were found significant in the models for non-fatal crashes and these results were consistent with the findings by Khan, Shanmugam, and Hoeschen (1999). However, developing regression models for estimating ‘fatal’ crashes was

difficult as was the case with ANOVA models. Therefore, regression models for estimating ‘fatal and injury’ crash combined were also developed to overcome this difficulty.

The third approach utilized the classification tree modeling approach. This study extended the conceptual framework presented by Lau, May, and Smith (1989) for utilizing a classification tree as a prediction model. The advantage of a classification tree model over other models was found that the tree structure can provide easy interpretation and one tree can be used to simultaneously estimate different types of crashes for the same group of freeway segments. This technique identifies significant threshold values of independent variables and does not rely on the judgment of the analyst.

An assessment of the crash prediction models was performed by comparing model generated estimates with actual numbers of crashes. Although the approach and formulation of these models are quite different, ANOVA, regression, and classification tree models were found to have fairly similar performance for the cases of interchange segments of North Carolina, more-than-four-lane non-interchange segments of Tennessee, and more-than-four-lane interchange segments of Tennessee. Further, it was found that most of the models developed for interchange segments performed better than those developed for non-interchange segments.

For the case of non-interchange segments with four lanes, North Carolina models performed better than Tennessee models, whereas Tennessee models performed better than North Carolina models for the case of non-interchange segments with more than four lanes. However, the models developed for both states were found to have fairly similar performance for the cases of interchange segments. In addition, the models for estimating ‘injury’ crashes and the models for estimating ‘fatal and injury’ crashes were found to have similar performance in all cases. This is because of the small number of fatal crashes in comparison to injury crashes.

Contributions of this Research

This research contributes to the subject area of safety conscious transportation planning in several ways. First, it explores the influence of interchanges on crash occurrences on urban freeways. Freeways represent a major component of urban highway networks as they carry a large proportion of the VMT in urban areas.

Second, this research suggests an appropriate freeway segmentation scheme by separating non-interchange segments from interchange segments. The segmentation scheme used by this research led to the development of separate models for non-interchange segments and interchange segments. No other study was found that developed separate crash prediction models for these two types of segments. Further, it should be pointed out that in long-range transportation planning for medium and large size urban areas, the alternative highway systems that are commonly examined involve different proportions of freeway lanes. The models developed by this research will provide transportation planners tools for assessing the safety consequences of different highway networks with varying proportions of four-lane and more-than-four-lane freeways.

Third, this research investigated alternative approaches that can be used for assessing a freeway network's safety impact. Typically, a regression model is the primary technique pursued by the majority of researchers. However, Chatterjee, et al. (2003) pointed out that the simplicity of prediction model forms should be considered in long-range planning applications considering the limited technical resources many MPOs have, and one model form they recommended is a cross-classification table. This research examined opportunities to provide alternative modeling approaches by developing ANOVA models as well as classification tree models. This research's classification tree model represented an improvement over the model presented by Lau, May, and Smith (1989) as it was developed using different schemes and a crash rate table was added to each tree model.

The classification tree models were found to perform as well as or better than the other two types of models.

Limitations and Difficulties of this Research

Several difficulties were experienced in this research with respect to data collection. This resulted in some limitations of research results. The first limitation is the lack of complete crash data and freeway traffic data. For North Carolina, the crash data available for this research were for only two-digit Interstate highways and the data did not include the three-digit sections, which represent loops and spurs commonly found in large urban areas. Further, North Carolina DOT had difficulty with matching all crashes with corresponding roadway segments because of incomplete location (mile post) information for certain data items. The matched data accounted for approximately 80% of the total number of crashes in the recorded database for six counties. Therefore, one should be aware that all prediction models for North Carolina were developed based on 80% of crash data. An adjustment factor of 1.25 may be used to convert model predicted number of crashes to represent 100% values. It should be pointed out that four counties in North Carolina – Forsyth, Guilford, Mecklenburg, and Orange – could not be included in the analysis because the matching process captured less than 80% of data in those counties.

The second limitation is related to the difficulty of developing models for ‘fatal’ crashes. This difficulty was also encountered by other researchers. As an alternative approach, Chatterjee, et al. (2003) suggested the use of fatal crashes and injury crashes combined in the models for planning level applications. There was evidence in this research that prediction models for these two severity types combined could overcome the difficulty. It should be pointed out that, for certain types of assessment such as an economic analysis of highway improvements, separate estimates are needed for fatal crashes. To meet such needs, the combined prediction of fatal and injury crashes can be separated by using proportions for each type of crashes based on past experience.

The last limitation is related to the application of the crash prediction models developed in this research. It should be pointed out that these prediction models were developed for long-range planning applications at a macroscopic level and these are not meant to substitute for the detailed models needed for evaluating safety impacts of projects at the design stage or for corridor planning.

Further Extensions of the Research

This research can be extended in several ways. The first one is to include more crash and freeway data from other states. Since this study revealed that crash patterns are different in the two states of North Carolina and Tennessee, further work should be done to examine crash frequencies and patterns of crashes in other states in a comparative manner. Further, the knowledge about crash occurrences in different states will provide an opportunity to develop model forms that can be shared across states as well as to develop a set of coefficients and calibration factors that are applicable to specific states.

The second extension is to develop prediction models for interchange crashes recognizing crashes on ramps and crashes on cross-roads within interchange influence areas. Modeling interchange related crashes will require sound strategies for classifying interchanges based on geometric information, as well as strategies for identifying the location of crashes within these interchange areas. Procedures for collecting and coding crash data for interchange areas should be carefully determined.

The third extension is to expand the conceptual framework for developing crash prediction models at long-range planning level to other types of roadway facilities especially non-freeway principal arterials. Crash prediction models for freeways and non-freeway arterials will provide a comprehensive set of tools for long-range transportation planning and will provide an opportunity for transportation planners to assess the trade-offs involving different proportions of freeways and non-freeway arterials.

LIST OF REFERENCES

LIST OF REFERENCES

- Abdel-Aty, M. A. and Radwan, A. E. 1999. *Developing crash predictive models for a principal arterial*, Proceedings of the Conference: Traffic Safety on two Continents, VTI konferens, Swedish National Road and Transport Research Institute, Volume 13a, Part 1, pp. 181-194.
- Abdelwahab, H. T. and Abdel-Aty, M. A. 2001. *Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections*, Transportation Research Record 1746, TRB, National Research Council, Washington DC, pp. 6-13.
- Bonneson, J. A. and McCoy P. T. 1993. *Estimation of Safety at Two-Way Stop-Controlled Intersections on Rural Highways*, Transportation Research Record 1401, TRB, National Research Council, Washington DC, pp. 83-89.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. I. 1984. *Classification and Regression Trees*, Wadsworth, Belmont, CA.
- Chatterjee, A., Seaver, W. L., Everett, J. D., and Wegmann, F. J. 2003, *Tools for Assessing Safety Impact of Long-range Highway Plans in Urban Areas*, University of Tennessee.
- Fazio, J., Holden, J., and Rouphail, N. M. 1993. *Use of Freeway Conflict Rates as an Alternative to Crash Rates in Weaving Section Safety Analyses*, Transportation Research Record 1401, TRB, National Research Council, Washington DC, pp. 61-69.
- Frantzeskakis, J. M., Assimakopoulos, V., and Kindinis G. 1994. *Interurban Accident Prediction by Administrative Area Application in Greece*, ITE Journal, Institute of Transportation, Volume 64, Number 1, pp. 35-42.

- Giudici, P. 2003. *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley & Sons Ltd., England.
- Golob, T. F., Recker, W. W., and Alvarez, V. M. 2004. *Freeway safety as a function of traffic flow*, Accident Analysis and Prevention, Volume 36, Issue 6, pp. 933-946.
- Greibe, P. 2003. *Accident prediction models for urban roads*, Accident Analysis and Prevention, Volume 35, Issue 2, pp. 273-285.
- Guevara, F. L., Washington, S. P., and Oh, J. 2004. *Forecasting Crashes at the Planning Level: A Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona*, Proceedings of the 83rd Transportation Research Board Annual Meeting, Washington DC.
- Hanneman, R. *Generalized Linear Models*,
faculty.ucr.edu/~hanneman/linear_models/c10.html
- Harwood, D. W., Council, F. M., Hauer, E., Hughes, W. E., and Vogt A. 2000. *Prediction of the Expected Safety Performance of Rural Two-lane Highways*, FHWA-RD-99-207, US Department of Transportation, Washington DC.
- Hauer, E., Council, F. M., and Mohammedshah, Y. 2004. *Safety Models for Urban Four-lane Undivided Road Segments*, Proceedings of the 83rd Transportation Research Board Annual Meeting, Washington DC.
- Hauer, E., Ng, J. C.N., and Lovell, J. 1988. *Estimation of Safety at Signalized Intersections*, Transportation Research Record 1185, TRB, National Research Council, Washington DC, pp. 48-61.
- Hummer, J. E., Hultgren, C. A., Khattak, A. J., Hao, T., Stamatiadis, N., Jones, S., Hall, L. A., and Hill, M. 1998. *Accidents on Secondary Highways and Counter-Measures*, Phase I Final Report, Department of Civil Engineering, North Carolina State University.

- Hummer, J. E., Hultgren, C. A., Khattak, A. J., Renski, H. 1999. *Identification of Promising Sites on Rural Two-lane Highways using Inventory Data*, Phase II Final Report, Department of Civil Engineering, North Carolina State University.
- Ivan, J. N. 2004. *A New Approach for Including Traffic Volumes in Crash Rate Analysis and Forecasting*, Proceedings of the 83rd Transportation Research Board Annual Meeting, Washington DC.
- Ivan, J. N., Wang, C., and Bernardo, N. R. 2000. *Explaining two-lane highway crash rates using land use and hourly exposure*, Accident Analysis and Prevention, Volume 32, Issue 6, pp. 787-795.
- Jovanis, P. P. and Chang, H. 1986. *Modeling the Relationship of Accidents to Miles Traveled*, Transportation Research Record 1068, TRB, National Research Council, Washington DC, pp. 42-51.
- Khan, S., Shanmugam, R., and Hoeschen, B. 1999. *Injury, Fatal, and Property Damage Accident Models for Highway Corridors*, Transportation Research Record 1665, TRB, National Research Council, Washington DC, pp. 84-92.
- Konduri, S. and Sinha, K. C. 2002. *Statistical Models for Prediction of Freeway Incidents*, Proceedings of the Seventh International Conference on Applications of Advances Technology in Transportation Engineering, Cambridge, MA, pp. 167-174.
- Kononov, J. and Allery, B. K. 2004. *Explicit Consideration of Safety in Transportation Planning and Project Scoping*, Proceedings of the 83rd Transportation Research Board Annual Meeting, Washington DC.
- Kraus, J. F., Anderson, C. L., Arzemanian, S., Salatka, M., Hemyari, P., and Sun G. 1993. *Epidemiological Aspects of Fatal and Severe Injury Urban Freeway Crashes*, Accident Analysis and Prevention, Volume 25, Issue 3, pp. 229-239.

- Lau, M. Y.K., May, A. D., and Smith, R. N. 1989. *Applications of Accident Prediction Models*, Transportation Research Record 1238, TRB, National Research Council, Washington DC, pp. 20-30.
- Littell, R., Stroup, W. W., and Freund, R. 2002. *SAS for Linear Models*, 4th Edition, John Wiley & Sons Inc.
- Lord, D. and Persaud, B. N. 2000. *Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure*, Transportation Research Record 1717, TRB, National Research Council, Washington DC, pp. 102-108.
- Lord, D. and Persaud, B. N. 2004. *Estimating the Safety Performance of Urban Road Transportation Networks*, Accident Analysis and Prevention, Volume 36, Issue 4, pp. 609-620
- Maher, M. J. and Summersgill, I. 1996. *Comprehensive Methodology for the Fitting of Predictive Accident Models*, Accident Analysis and Prevention, Volume 28, Issue 3, pp. 281-296.
- Maycock, G. and Hall, R. D. 1984. *Accidents at 4-Arm Roundabouts*, TRRL Laboratory Report 1120, U.K. Transportation Research Laboratory.
- Maze, T. H., Preston, H., Storm, R. J., Hawkins, N., and Burchett, G. 2005. *Safety Performance of Divided Expressways*, ITE Journal, Institute of Transportation, Volume 75, Number 5, pp. 48-53.
- McCullagh, P. and Nelder, J. A. 1989. *Generalized Linear Models*, Second Edition, London: Chapman and Hall.
- Miaou, S. P. 1996. *Measuring the Goodness-of-Fit of Accident Prediction Models*, FHWA-RD-96-040, US Department of Transportation, Washington DC.

- Miller, J. S., Hoel, L. A., Kim, S., and Drummond, K. P. 2001. *Transferability of Models That Estimate Crashes as a Function of Access Management*, Transportation Research Record 1746, TRB, National Research Council, Washington DC, pp. 14-21.
- Mountain, L., Fawaz, B., and Jarrett, D. 1996. *Accident Prediction Models for Roads with Minor Junctions*, Accident Analysis and Prevention, Volume 26, Issue 6, pp. 695-707.
- Myers, R. H., Montgomery, D. G., and Vining, G. G. 2001. *Generalized Linear Models: With Applications in Engineering and the Sciences*, John Wiley & Sons Inc., New York.
- National Highway Traffic Safety Administration. 2002. *Traffic Safety Facts 2001: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System*, DOT HS 809 484, National Center for Statistics and Analysis, U.S. Department of Transportation, Washington DC, www-nrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/TSFAnn/TSF2001.pdf.
- Neville, P. G. 1999. *Decision Trees for Predictive Modeling*, SAS Institute Inc.
- Ott, L. R. and Longnecker, M. T. 2001. *An Introduction to Statistical Methods and Data Analysis*, Fifth Edition, Duxbury Press.
- Persaud, B. and Dzbik, L. 1993. *Accident Prediction Models for Freeways*, Transportation Research Record 1401, TRB, National Research Council, Washington DC, pp. 55-60.
- Persaud, B. N. 1991. *Estimating Accident Potential of Ontario Road Sections*, Transportation Research Record 1327, TRB, National Research Council, Washington DC, pp. 47-53.

- Persaud, B. N. 1994. *Accident prediction models for rural roads*, Canadian Journal of Civil Engineering, Volume 21, Number 4, pp. 547-554.
- Resende, P.T. V. and Benekohal, R. F. 1997. *Development of Volume-to-Capacity Based Accident Prediction Models*, Proceedings of Traffic Congestion and Traffic Safety in the 21st Century, Chicago, Illinois, pp. 215-221
- Robertson, H. D., Hummer, J. E., and Nelson, D. C. 2000. *Manual of Transportation Engineering Studies*, Prentice-Hall, Inc.
- Saccomanno, F. F. and Fu, L. 2003. *GIS-Based Integrated Model for Road Accident Analysis and Prediction*, www.civil.uwaterloo.ca/saccomanno/publications.htm.
- SAS Institute Inc. 1999. *SAS OnlineDoc V8*, Cary, North Carolina, v8doc.sas.com/sashtml/.
- SAS Institute Inc. 1999. *SAS/STAT User's Guide: The GENMOD Procedure*, Cary, North Carolina.
- Satterthwaite, S. P. 1981. *A Survey of Research into Relationship between Traffic Accidents and Traffic Volumes*, Supplementary Report 692, U.K. Transportation Research Laboratory.
- Sayed, T. and Rodriguez, F. 1999. *Accident Prediction Models for Urban Unsignalized Intersections in British Columbia*, Transportation Research Record 1665, TRB, National Research Council, Washington DC, pp. 93-99.
- Schmidhammer, J. L. 2004. *Lecture Notes: Data Mining*, Department of Statistics, University of Tennessee, Spring Semester.
- Schwetz, T. B., Reiff, B., and Chatterjee, A. 2004. *Assessing Safety Impacts of Long Range Plans in Small and Medium-Sized Communities*, Proceedings of the 83rd Transportation Research Board Annual Meeting, Washington DC.

- Shankar, V., Mannering, F., and Barfield, W. 1996. *Statistical Analysis of Accident Severity on Rural Freeways*, Accident Analysis and Prevention, Volume 28, Issue 3, pp. 391-401.
- Social Sciences Teaching and Research Statistics, University of Kentucky. *Poisson and Negative Binomial Regressions*,
www.uky.edu/ComputingCenter/SSTARS/Contents_P_NB.htm
- SPSS Inc. 1999. *SPSS White Paper: AnswerTree Algorithm Summary*.
- Traffic Safety Systems Management, North Carolina Department of Transportation.
2000-2002 Crash Rates,
www.doh.dot.state.nc.us/preconstruct/traffic/safety/ses/rates/rates.html
- Vogt, A. and Bared, J. G. 1998. *Accident Models for Two-Lane Rural Roads: Segments and Intersections*, Report Number FHWA-RD-98-133, Federal Highway Administration. www.tfhr.gov/safety/98133/
- Washington, S. P., Karlaftis, M. G., and Mannering F. L. 2003. *Statistical and Econometric Methods for Transportation Data Analysis*, Chapman & Hall/CRC.
- Wilkinson, L. 1992. *Tree Structured Data Analysis: AID, CHAID, and CART*, Sawtooth/SYSTAT Joint Software Conference, Sun Valley, ID.

VITA

Vasin Kiattikomol was born and raised in Bangkok, Thailand. He received Bachelor's degree in Civil Engineering from Thammasat University in his hometown in 1999. He worked as a civil engineer with Berger-CSC1 Consortium, Thailand, for one year. From there, he went to the University of Tennessee, Knoxville and received his Master's degree in Civil Engineering (Transportation) in 2002 and his Doctor of Philosophy's degree in Civil Engineering with a concentration in Transportation Engineering in 2005.

During his graduate studies, he worked with the department of Civil and Environmental Engineering and the Center for Transportation as a teaching and research assistant. He is a member of Institute of Transportation Engineers and a member of the honor society of Phi Kappa Phi.

Vasin is currently joining the transportation research group as a post doctoral scholar at the University of California, Davis, CA.