ANALYSIS OF TYPE AND SEVERITY OF TRAFFIC CRASHES AT SIGNALIZED INTERSECTIONS USING TREE-BASED REGRESSION AND ORDERED PROBIT MODELS

by

JOANNE MARIE KELLER
B.S. University of Central Florida, 2003

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Civil and
Environmental Engineering in the
College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term
2004

Major Professor
Dr. Mohamed A. Abdel-Aty, P.E

ABSTRACT

Many studies have shown that intersections are among the most dangerous locations of a roadway network. Therefore, there is a need to understand the factors that contribute to traffic crashes at such locations. One approach is to model crash occurrences based on configuration, geometric characteristics and traffic. Instead of combining all variables and crash types to create a single statistical model, this analysis created several models that address the different factors that affect crashes, by type of collision as well as injury level, at signalized intersections. The first objective was to determine if there is a difference between important variables for models based on individual crash types or severity levels and aggregated models. The second objective of this research was to investigate the quality and completeness of the crash data and the effect that incomplete data has on the final results.

A detailed and thorough data collection effort was necessary for this research to ensure the quality and completeness of this data. Multiple agencies were contacted and databases were crosschecked (i.e. state and local jurisdictions/agencies). Information (including geometry, configuration and traffic characteristics) was collected for a total of 832 intersections and over 33,500 crashes from Brevard, Hillsborough and Seminole Counties and the City of Orlando. Due to the abundance of data collected, a portion was used as a validation set for the tree-based regression.

Hierarchical tree-based regression (HTBR) and ordered probit models were used in the analyses. HTBR was used to create models for the expected number of crashes for collision type as well as injury level. Ordered probit models were only used to predict crash severity levels due to the ordinal nature of this dependent variable. Finally, both types of models were used to predict the expected number of crashes.

More specifically, tree-based regression was used to consider the difference in the relative importance of each variable between the different types of collisions. First, regressions were only based on crashes available from state agencies to make the results more comparable to other studies. The main finding was that the models created for angle and left turn crashes change the most compared to the model created from the total number of crashes reported on long forms (restricted data usually available at state agencies). This result shows that aggregating the different crash types by only estimating models based on the total number of crashes will not predict the number of expected crashes as accurately as models based on each type of crash separately. Then, complete datasets (full dataset based on crash reports collected from multiple sources) were used to calibrate the models. There was consistently a difference between models based on the restricted and complete datasets. The results in this section show that it is important to include minor crashes (usually reported on short forms and ignored) in the dataset when modeling the number of angle or head-on crashes and less important to include minor crashes when modeling rear-end, right turn or sideswipe crashes. This research presents in detail the significant geometric and traffic characteristics that affect each type of collision.

Ordered probit models were used to estimate crash injury severity levels for three different types of models; the first one based on collision type, the second one based on intersection characteristics and the last one based on a significant combination of factors in both models. Both the restricted and complete datasets were used to create the first two model types and the output was compared. It was determined that the models based on the complete dataset were more accurate. However, when compared to the tree-based regression results, the ordered probit model did not predict as well for the restricted dataset based on intersection characteristics. The final ordered probit model showed that crashes involving a

pedestrian/bicyclist have the highest probability of a severe injury. For motor vehicle crashes, left turn, angle, head-on and rear-end crashes cause higher injury severity levels. Division (a median) on the minor road, as well as a higher speed limit on the minor road, was found to lower the expected injury level.

This research has shed light on several important topics in crash modeling. First of all, this research demonstrated that variables found to be significant in aggregated crash models may not be the same as the significant variables found in models based on specific crash types. Furthermore, variables found to be significant in crash type models typically changed when minor crashes were added to complete the dataset. Thirdly, ordered probit models based on significant crash-type and intersection characteristic variables have greater crash severity prediction power, especially when based on the complete dataset. Lastly, upon comparison between tree-based regression and ordered probit models, it was found that the tree-based regression models better predicted the crash severity levels.

ACKNOWLEDGEMENTS

The author would like to thank Dr. M. Abdel-Aty for his continuous support and brilliant advice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 Background

Traffic crashes affect everyone. According to Cafiso et al. (2004), of the millions of crashes occurring each year in the United States, over 500,000 people are killed and more than 15 million people are injured. This corresponds to a crash-related death every minute. While crashes can be mostly attributed to human error, it is suggested that the design and characteristics of a roadway can also be responsible for causing crashes. For example, during 1999, there were 243,409 crashes recorded in the Florida Crash Database. Of these, 98,756 crashes occurred at or were influenced by a signalized intersection. To describe the seriousness of these numbers, the 98,756 crashes correspond to one crash every 5.5 minutes. Bhesania (1991) found that out of several thousand crashes in Kansas City, Missouri, signalized intersections experience the largest number of incidents. More specifically, Bhesania found that 9.6 crashes occur per year at signalized intersections per year compared to 2 per year where stop or yield signs provide traffic control. This further validates the point that roadway intersections are a common place for crashes, which may be due to the fact that there are several conflicting movements as well as a myriad of different intersection design characteristics. However, the factors affecting crashes are not well defined and this lack of knowledge may be the source of additional crashes. Therefore, there is a need to classify intersections and quantify the affects that certain geometric aspects have on the number of crashes at a specific intersection.

Furthermore, when a crash occurs and the local police department is notified, the responding police officer will determine whether to fill out a long or short crash form based upon several crash factors. For instance, if a crash involves an injury or a felony was committed, the

crash must be filed on a long crash form. If a crash involved only property damage (a minor crash), it will be identified on a short crash form. Crash forms are then sent to the respective counties, which choose whether or not to file short forms. From here, only the crashes reported on long forms are forwarded onto the Florida Department of Transportation (FDOT) and the Department of Highway Safety and Motor Vehicles (DHSMV), which maintain records based on only crashes reported on long forms. Since most crashes that occur involve only property damage and not a serious injury or a felony, it can be argued that the FDOT and DHSMV crash databases under-represent minor crashes as well as certain types of crashes that frequently involve property damage only. Moreover, by only keeping track of long forms, these agencies exaggerate the fraction of crashes that involve a serious injury, which make roadways appear less safe. Therefore, by excluding minor crashes, any models developed will under-represent the true number of crashes that occurred at a location and may cause a difference in the significance of the crash-related variables.

1.2 Plan of Action

Task 1: Collection of Data. Crash information was collected for four jurisdictions across the middle of the State of Florida: Brevard County, City of Orlando, Hillsborough County and Seminole County, for three of their most recent years of data. This information was obtained from either the county/city itself, Florida Department of Transportation or from the Department of Highway Safety and Motor Vehicles. Additionally, intersection information was also gathered from the individual counties/city. The information required for this analysis included the number of through lanes on each approach to the intersection, geometric configuration, speed limits, and daily traffic volumes. Most of this information was available from intersections

drawings and the other variables were identified from level of service reports, usually available on the counties' websites. In order to acquire the most accurate database for each county, crash records reported on long forms were checked and cross-referenced at the local and state government levels. Although this task was time-consuming, the results were that the databases created were as precise as possible. Finally, after obtaining records for crashes reported on short forms, these crashes were also cross-referenced against two governmental databases to ensure that none of the crashes were also reported on long forms.

Task 2: Organization of the Data. After the data had been collected it was organized into four separate master databases, one for each entity. These databases included each intersection that had complete information as well as the crashes that occurred at these intersections. Information for all four jurisdictions was then combined and, from this new spreadsheet, the data was prepared for analysis.

Task 3: Pre-Analysis and Data Exploration. Based upon past analyses on similar data, a pre-analysis was conducted in the effort to determine the most accurate and efficient way to analyze this specific assortment of data. Frequency tables were developed and distributions were graphed.

Task 4: Analysis Method. The analysis methods chosen were unique in that they involved a tremendous amount of information on several thousand crashes. Based upon the thorough literature review conducted, it was evident that studies such as this are rare. Models were built depending on variables found to be significant and in a way that has not been done previously for crashes occurring at signalized intersections.

Task 5: Reporting the Results. Finally, the results from the various models built were collected and the model interpretations are stated herein.

1.3 Research Objective

The rationale behind conducting this research is to further the understanding of the causes of traffic crashes at signalized intersections. This study explores the hypothesis that different types of collisions and different crash severity levels are affected by different independent variables. Furthermore, the author investigates the significant differences in the important crash-related factors between models based solely on crashes reported on long forms and models based on crashes reported on both long and short forms (i.e. models based on restricted and complete datasets). Several databases were crosschecked to ensure the completeness of our data. The chief intention was to create statistically significant models for two datasets: one including long-form-only crashes and the other a complete dataset including crashes reported on both long and short forms, and to then compare the results. The author anticipates that these results will provide a significant contribution to the area of safety at signalized intersections as well as consider the possible consequences of modeling restricted datasets.

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

Due to the fact that traffic crashes are both costly and a major inconvenience to anyone involved, crash prediction studies are of foremost importance. As a result of the many conflicting movements that occur at intersections, these locations usually have a higher crash rate than any other roadway locations. It is important to be able to understand and explain excessive crash locations as well as to be able to correct these problems with suitable solutions. One way to understand why an area is more prone to a crash is to collect and analyze data using various methods; frequency models, neural network applications and statistical methods.

2.2 Frequency Models and Statistical Methods

Storsteen (1999) identified intersections across the state of South Dakota and grouped them by several geometric types, control types and volumes. The mean, 90^{th} and 95^{th} percentiles were reported based upon the number of crashes per type and the project's output was used by the state's department of transportation to identify intersections with serious problems.

In 1998, Weerasuriya created tables based upon 3-legged intersection in Florida. The tables included values for the mean, variance, 90^{th} and 95^{th} percentile of crashes at each intersection. This study also identified the common types of 3-legged intersection as 2x2, 2x4, and 2x6. Data came from five counties across the state of Florida and formed 38 different intersection types. In 1996, Pietrzyk developed tables for the expected number of crashes based upon a study conducted for Urban Transportation Research. The purpose of this research was to identify common types of intersections and to create tables so that the expected number of

crashes at a certain location can be estimated. Intersections were divided into groups based upon signalization, number of approach legs and number of through lanes. A total of fifteen categories were distinguished from five counties in Florida. Linear regression was used to estimate the expected number of crashes.

Parsonson et al. collected information from 1,456 Atlanta intersections in 1993 for a highway safety project. The purpose of their study was to create tables with expected values for the number of accidents at various types of intersections based on whether the intersection was three- or four-legged, signalized or unsignalized, and the total number of entering vehicles per day. Crash values were determined for 7 categories; collision type, severity light conditions, surface conditions, season of the year, day of the week and hour of the day. Each category had between two and ten variables considered. Statistics calculated for each type included mean accidents per year, abnormally high accidents per year 90^{th} percentile and abnormally high accidents per year 95^{th} percentile. Results were tabulated into a total of 17 tables.

Similarly, PAB Consultants, Inc. (1997) conducted a project to development tables that displayed the average or expected number of crashes for different types of intersections. Four- and three-leg intersections that are both signalized and unsignalized were considered and expected number of crashes were determined for numerous different types of crashes including rear-end, head-on, angle, etc. Furthermore, the standard deviation of the expected number of accidents, the 90^{th} percentile and the 95^{th} percentile were also found for each type of intersection as well as each type of crash. The purpose of creating these tables was so that an engineer assessing the safety of a particular intersection would be able to refer the table corresponding to the size of that intersection to find the expected number of crashes. The engineer would then make a decision based on the actual number of crashes that occurs at this site as to whether or

not the intersection warrants safety improvements. This project used analyzed over a total of 350 intersections in 13 different category types. The authors' note that the statistics calculated are based on the assumption that the frequency of crashes can be approximated reasonably well by the normal distribution.

Pernia et al. investigated the before and after affects of newly signalized intersections at several locations throughout Florida. They collected information on over 518 intersections with a total of 4565 crashes. There were three phases to this project and the first phase was to calculate the percentage of crashes and the crash rates for each intersection for the before and after periods. The 50^{th} and 85^{th} percentile values were then calculated and used to compare the results. Paired t-tests were used to test if there was significance in the differences from values. The study concluded that the total number of crashes as well as crash rates increased after an intersection was signalized.

Thomas et al. (2002) researched the number of accidents and the benefit/cost (B/C) ratio to estimate Iowa traffic safety improvements' efficiency. Locations were grouped into one of seven categories based on the type of improvement; new traffic signal, new traffic signal and turn lane(s) addition, add turn phasing to existing signal, add turn phasing to existing signal and turn lane(s), replace pedestal mount signals with mast arm mount signals, add turn lane(s) only, and other geometric improvements. Two types of analysis were conducted for each of the seven categories; the first method was an estimation of the mean crash reduction and confidence interval, and the second method was calculation of the B/C ratio. The second method required the authors to assign monetary values to crashes involving fatality, major injury, minor injury, possible injury, and property damage only to be able to compare the cost benefits using net present worth analysis. In conclusion, when all data was taken into account, improvements with

new traffic signals only and pedestal mount signal replacement have the highest benefit-to-cost ratio. The authors claim the most important outcome to be the significance of traffic signal visibility. The research showed a large reduction in crashes and a high B/C ratio for the pedestal replacement projects. Additionally, analysis for category one projects, signal installation only, showed there might not be any improvement in safety for this type of change.

Lee et al. (2004) created zero-inflated accident frequency models to identify the factors that affect roadway and railway at-grade-intersection crash rates. This method was chosen over the standard Poisson or negative binomial methods because there were many situations where the number of accidents was zero. Several explanatory variables were used including those corresponding to location, roadway and grade crossing characteristics.

Steinman and Hines (2004) assessed safety at signalized intersections for pedestrians and bicyclists. They rated six intersection characteristics and found that a protected left turn phase with a pedestrian phase increased safety as well as a smaller intersection radius and prohibited right-turns-on-red. Additionally, a lower speed limit was found to make crossing conditions safer for pedestrians and bicyclists.

Oh et al. (2004) used Poisson and negative binomial regressions to create crash prediction models for three-legged, four-legged and signalized intersections for both the total number of crashes and the number of injury crashes. For the total crash model at signalized intersections, the traffic volume on both the major and minor road, the posted speed limit on the major and commercial driveways in the vicinity of the intersection caused more crashes. The higher the average degree of curvature for the intersection and whether the intersection was lighted caused fewer crashes to occur. Wang and Nihan (2001) used signalized intersections in Tokyo to create

negative binomial crash models to predict the number of angle crashes. Variables entered into the model included specific traffic volumes and geometric characteristics.

During the second phase of an FDOT project, conducted by Pernia et al., crash predictive models were developed to estimate the average number of crashes, and also the average number of crashes for four different types; angle crashes, left turn crashes, rear-end crashes, all other crashes for intersections that were recently signalized. Poisson regression was used initially but a negative binomial model was used in all cases where over-dispersion was detected. When this task was completed, Pernia et al. used the results obtained to test a validation set of 30 newly signalized intersections that were not used in building the model. While it was found that a higher average daily traffic (ADT) causes more crashes, business areas have more crashes, intersections with more than four lanes on the major roadway have a higher crash frequency, sites with a speed higher than 45mph on the major roadway have less crashes, sites that are divided have less crashes except in the case of rear-end crashes and intersections with paved shoulders have less crashes, it was also found that ADT was the only significant variable at the 5% significance level.

Greibe (2003) presented results from two accident prediction models on intersections and segments where the models were simple and feasible. The models were based on 1036 intersections and 142km of roadway in Denmark. Crashes were reviewed for a five-year period. For roadway segments, the following variables were recorded; traffic volume, length of section, speed limit, one or two-way traffic, number of lanes, road width, speed reducing instruments, number of minor intersections, bicyclist facilitation, footway, median, parking facilitation, bus stops, and land use. For intersections, the following variables were measured; traffic volumes, number of lanes, median, turning lanes, bicycle facilitation, signalized/non-signalized, and

number of signal arms. Initial statistical analysis revealed strong correlations between many of the independent variables and therefore the safety effect that one geometric feature has on the intersection or segment is immeasurable. The author assumed a Poisson distribution for the data mainly because of the simplicity of this distribution; the variance is equal to the mean. In conclusion for roadway segments, ADT was found to contribute the most to crashes followed by surrounding land use, number of minor intersections, parking facilities, speed limit, road width, number of access points and number of lanes. For intersections, it was again determined that traffic volumes contribute the most to crash frequency. Problems encountered include strong correlation between variables and that the number of crashes per site may not follow the assumed distribution.

Persaud et al. (2002) illustrated the difficulty with developing accident prediction models and then transferring them to other regions of interest. The data used in this research was from Toronto between the years of 1990-1995 and it was classified into four categories; Signalized 4-legged, Signalized 3-legged, Unsignalized 4-legged, Unsignalized 3-legged. There were a total of 1454 intersections used in this project. Using a method referred to as the "ID" method, the data distribution was determined to be a function of gamma. Both γ and R^2_α were calculated as goodness of fit measures for each model calibrated where high values of γ and R^2_α are desirable. For the first phase of this project, a model was fit to estimate average accident frequency. A total of 8 models were calibrated; two types, injury accidents only and all accidents, for each of the four intersection categories mentioned above. After the models were calibrated, each variable in every model was then tested for statistical significance using the t-statistic. Finally, a cumulative residual (CURE) plot was used to measure the fit for each of the eight models. A good fit was indicated by a CURE plot that fluctuates around the value of zero. The second phase of the

project included transferring and comparing the output from phase one to two other regions and it was concluded that transferring a model performs best when the two area's have similar traffic volumes.

Hauer (2002) found it necessary to specify a different type of statistical analysis because many methods of safety estimation are based only on accident counts with their precision in terms of standard deviation. These measurements are fairly accurate when there is a large occurrence of accidents; however, they become imprecise when only a few accidents occur over a long period of time. The other inadequacy of these types of estimates is that they are subject to a common bias. This type of bias is labeled 'regression-to-mean' bias and arises when one chooses to look at a particular entity because of the number of accidents, too many or very few, that have occurred in a specified time period. The fix to this problem includes a regression-to-mean correction. Using the Empirical Bayes (EB) method the two aforementioned problems are avoided. The EB method takes into account other 'clues' to determine a more precise estimate of the number of accidents to be expected. This method employs information from other similar entities during estimation and uses weights so that some observations are worth more to the model than others. As an extension of this research, Qin et al. (2003) described a new way of relating volumes to crash rates using a hierarchical Bayesian framework to fit zero-inflated-Poisson (ZIP) regression model.

For a highway safety project, Harwood et al. used three different types of statistical analyses to test the effects of certain geometric improvements. Three different before-after evaluation methods were employed by this project; yoked comparisons (YC), comparison groups (CG), and the Empirical Bayes (EB). The first method, YG, used a one-to-one matching of the improved intersections to the similar, non-improved sites. This procedure is intended to account

for the effects of time trends. Factors studied in this approach include the number of accidents for the before and the after periods for both the improved and non-improved sites, the expected number of accidents on the improved site and the observed accident reduction effectiveness. The second approach was the comparison groups. This method makes the same main assumption as the previous method but now takes the comparison sites as a whole instead of looking at them on a one-to-one basis. Again, the same factors were considered here as in the previous method. Additionally, a negative binomial model was created for the basis of traffic volume and state effects. This model was developed because it would also be useful in the last analysis method. However, this method is not capable of giving results when accident frequencies are zero and it is not capable of accounting for the "regression to mean" bias. The third method, EB, was chosen because of its three benefits; it accounts for the "regression to mean" bias, it accounts for changes in factors during the before and after periods, and it uses several years of data which is helpful a site has few or no accidents in a particular year. This method makes use of two sets of information: the number of accidents on the improved sites and the number of accidents on the comparison sites, to make an estimate of the expected number of accidents. EB method uses the comparison group to create relationships between site characteristics such as volume and accident experience. The major strength of this approach is that the method accounts for the "regression to the mean" bias. It also requires a smaller reference group than that of the CG method. Finally, it is capable of providing results even when the accident frequency is equal to zero. After applying three different methods of analysis and comparing their results, the authors concluded that the Empirical Bayes method had advantages over the other two mainly because it accounts for the "regression to the mean" bias.

Persaud (2003) again used the Empirical Bayes method to estimate the change in expected accident frequency after the installation of a signal and to use safety impact knowledge to determine where to place a signal. Accident counts and traffic volumes were used to estimate the expected accident rates if an intersection was not signalized. When developing the models, variables like area type, volumes, sight distance, and turn lanes were used. Additionally, for this research, the software package GENSTAT was used to create a general linear model assuming a negative binomial error distribution. The only variables that proved to be significant were the flows on the intersecting roadways. After the models were created, a before-after Bayesian analysis was performed to account for the regression-to-mean bias encountered. The results from this research were the development of a step-by-step procedure to determine whether a signal should be placed at a particular site.

In 1996, Al-Turk et al. created a series of negative binomial models capable of predicting the change in the number of accidents with changes in the volume-to-capacity (v/c) ratio. The main idea behind this research was to determine the relationship between the degree of volume saturation and the safety level at signalized intersections. The variables included in the final model were crash frequency, crash type, v/c ratio, and time of day.

Bonneson and Jun Son (2003) developed a model to predict the expected number of vehicles that will run a red light at urban intersections. To create their model they assumed a logistic distribution to define the probability that a vehicle will stop. To create an equation for the expected number of red-light-runners the probability function was multiplied to flow, integrated with two different limits and finally multiplied by the number of cycles per hour. Authors also collected real world data and used SAS's nonlinear regression and generalized modeling procedures to create models.

Sawalha and Sayed (2003) described a common method, generalized linear regression modeling (GLM), of estimating accident prediction models (APM). APM serve several purposes such as estimation of potentially hazardous situations, identification and ordering of accident-prone entities, assessment of safety improvements, and safety planning. To determine which model structure to use, the authors suggest the use of the Poisson distribution and to also calculate a dispersion value. If the value of this variable is much greater than one, the data is said to have a greater dispersion than can be explained by the Poisson distribution and, therefore, the error structure is fitted to the negative binomial distribution. The authors chose to include variables based on two criteria: if the t-ratio of the estimated parameter is significant at the specified α-level and if the addition of the specific variable causes a significant decrease in the scaled deviance at the specified α-level. The main purpose of this research was to demonstrate how to properly fit a model by selecting only relevant variables and by conducing outlier analysis.

Chin and Quddus (2003) used data from signalized intersections to create a random effect negative binomial model in their research to account for the shortcomings in the negative binomial distribution. Several variables were found to cause collinearity and were excluded from the model. The variables found to be positively associated with the number of crashes included approach and right turn volume, intersection sight distance and median width. The negatively associated variables were acceleration section on left turn lane, number of bus bays and signal control type.

In an effort to create crash severity models based on roadway medians, Donnell and Mason (2004) utilized logistic regression to find the probability of various types of injury levels based on geometric and environmental characteristics as well as traffic operations. Results

suggested that for interstate median crashes, all of the following affect the probability of a fatal crash; wet road surface, use of drugs of alcohol, nearby interchange ramp, crash type, and the traffic volume.

Ladron de Guevara et al. (2004) used crash statistics from Arizona Department of Transportation to create several negative binomial models. The objective of their study was to create models for safety at the planning level based on variables available through geographic information systems (GIS) such as population density, percent of population under the age of seventeen, number of employees in an area, intersection density and so on. Models were created for three types of crashes: fatal, injury, and property damage only. Due to that fact that errors between these three types are related, additional and more accurate information can be obtained by creating these models simultaneously. Initially, models were created on an individual basis for comparison purposes. Then models were created simultaneously and contrasted to the preliminary models. For the simultaneous models between fatal and injury crashes, a large correlation was found, suggesting that the simultaneous estimation was warranted. The conclusion drawn from this research was that simultaneous estimation was required in order to have "unbiased and efficient parameter estimates." For the reason that general models do not include all possible explanatory variables and many of the variables have inherent measurement errors, simultaneous model estimation was found to be an effective method of controlling for coincident correlation.

2.3 Other Types of Models

Liu and Young (2004) looked into 1,593 accidents spread over 62 signalized intersections between a two-year period in an effort to build a neural network model to accurately predict the

number of intersection accidents based on characteristics of the intersection. They were able to create a model with a high correlation coefficient, a mean square error of 3.38×10^{-6} and a misjudgment rate of 16.4%. The most important factors affecting crash rates were found to be the number of fast traffic lanes, width of the fast traffic lanes, median type between fast and slow traffic lanes, left turn signal timing and type of central median. Using a case study, the researchers claim that this procedure produces reliable results.

Barceló (2003) employed microscopic traffic simulators to analyze traffic safety instead of its usual application of the evaluation of traffic systems. The authors discuss a new crash prediction method that has a binary response to estimate the dynamic probability of each type of incident. The approach developed by the authors was called EIP-HLOGIT. The estimated probability of a crash is defined as the ratio of two exponential functions of a linear combination of coefficients and independent variables. After data was collected, the model was selected, calibrated and then fine-tuned. The authors conclude that their proposed method yields 'promising results' using a flexible statistical regression model, which is then applied to dynamic data in an effort to provide real-time results. Regardless of the numerous problems encountered when applying this approach to a test site, the results of the EIP-HLOGIT method to associate relationships between traffic information and crash rates were found to be very useful.

O'Connell and Kreis (2003) focused on the development of a new method to analyze the adequacy of highways that is an improvement over the previous HPMS-AP method used by the FHWA. The authors claim that their new method makes a clearer separation between roads that need repair and those that are adequate. The new method makes use of a weighting system where a road's specific characteristics are given a value. All of the road's values are added together where a resulting score of 100 would reflect a perfect road. The authors conclude by

claiming the new method measures safety more heavily, uses more objective values, more clearly separates roads needing repair from those that do not, and leaves room for change.

2.4 Tree-Based Regression and Ordered Probit Models

Of particular interest to this research, Karlaftis and Golias (2002) used software known as CART 1995 to develop a crash prediction model based on rural roadway geometry and crash rates by using methodology known as hierarchical tree-based regression (HTBR). It was thought that this method would provide an uncomplicated way of predicting crash frequency. The data used for this research was from rural roads in Indiana from 1991 to 1995, which was grouped into two categories; rural two-lane and rural multi-lane and ADT was found to be the most significant variable overall.

While there are few, if any, examples of safety analysis being conducted by means of hierarchical tree-based regression, several other studies have utilized this method for research. For example, Washington and Wolf (1997) used HTBR to forecast trip generation and the results were compared to the traditional ordinary least squares (OLS) method. Also in 1997, Washington et al. considered using HTBR to determine modal correction factors for motor vehicle emissions. It was reported that, while the theory behind HTBR is less developed than that of more traditional models, there are several advantages to using HTBR methods. Hallmark et al. (2002) used HTBR to identify geometric and operational roadway characteristics that influenced vehicle activity. Finally, Washington (2000) discussed the theory behind HTBR and presented an example that combines OLS and HTBR that can be used to forecast trip generation.

Abdel-Aty (2003) used ordered probit models to predict crash injury severity on roadway sections, signalized intersections and toll plazas. For roadway sections, it was found that females

and older drivers have an increased risk of severe injury. The results for signalized intersections were similar. For the toll plaza, similar variables were found to be significant including driver's age, gender, seat belt usage, collision type and type of vehicle.

Duncan et al. (1999) used ordered probit models to determine the factors that influence injury severity in truck/passenger car rear-end crashes. Research showed that higher speeds, crashes involving women, crashes at night, crashes with alcohol involved and differential speeds when a car hits a truck all result in a higher probability of a more severe crash. Klop (1998) used ordered probit models to determine the influence that certain factors had in the severity level of crashes involving bicyclists. The conclusion was that straight grades, curved grades, darkness and fog all increased the risk of severe injury during a crash.

O'Donnell and Connor (1996) created two ordered probit models to predict the injury levels for crashes in Australia. Increases in both the age of injured person and the speed of vehicle caused a greater injury level. Furthermore, seat location inside vehicle, vehicle type and make, alcohol involvement and collision type were also found to have significant impacts on the crash severity level.

2.5 Summary

Crash prediction and modeling have proven to be invaluable tools for engineers to estimate the safety both during planning and operational phases. While there are many methods of analysis that have been used in the past, there has not been much research to identify differences in crash-influencing factors at signalized intersections. Furthermore, little or no research has been conducted to address the issue that including crashes reported on short forms, to make the dataset complete, may produce different results and important factors. To explore

these concerns, initial modeling methods were similar to research conducted by Karlaftis and Golias (2002) that employed regression trees to determine significant factors affecting crash occurrences. The second type of model used in this research was the ordered probit model. From the literature it was evident that both of these methods are rarely used in the safety analysis of signalized intersections.

CHAPTER 3. DATA COLLECTION

3.1 Collection Plan

Data collection for this project began in early 2003 when several counties across the midsection of the State of Florida were contacted for cooperation. Of particular interest were those counties that maintained records on crashes reported on both long and short forms. Being that the University of Central Florida's main campus was located in Orange County, it was the first county contacted. However, as more information was gathered from Orange County, it became evident that the county only maintained long form crash records and, therefore, would not be relevant to this research. Fortunately, the next four counties contacted proved to be useful sources of information and the following sections described the data collection efforts in each of the entities.

3.1.1 Seminole County

The second county contacted was Seminole County and the county was able to provide several hundred AutoCAD intersection drawings on a CD-ROM. Each of the drawing files were opened and the geometry of each intersection was recorded into a geometry database for Seminole County. Information collected included the intersecting roadway's names, number of through lanes on each the major and minor roadway, the number of left turn lanes and whether they were exclusive for each approach, whether the roadways were divided for each approach, whether any of the right turns were channelized on each approach and the speed limits when available. Average daily traffic volume and non-state road speed limit information was obtained from their county website. However, state road speed limits were not available. Due to the fact

that Seminole County was nearby, two days were spent driving all of the counties state roads in an effort to collect the missing speed limits. When Seminole County's classification was complete, there were a total of 195 intersections completely identified. Once the classification on Seminole County was complete, the intersections were then alphabetized and then number-coded to indicate the county from which they belong.

Crash records were plentiful in Seminole County where crashes reported on both long and short forms were obtained for three years: 1999, 2000 and 2001. For crashes reported on long forms, a program was written to extract the necessary records from the FDOT and DHSMV databases and input them into a database for Seminole County to serve as a crosscheck for the records provided by the county. The final database contained crash information, geometry and AADT volumes, most of which were entered by hand. Crashes were found from most intersections, however, there were a handful of intersections that had no crashes over the three year period but were included because zero crashes is a valid number. In order to retrieve information for crashes reported on short forms, the county was contacted again was able to provide records from 1999-2001 of all crashes reported on short forms. By making the extra effort to obtain records for crashes reported on short forms, this dataset became complete.

3.1.2 City of Orlando

The next governmental entity contacted for information was the City of Orlando, which is located within Orange County. The city provided two CD-ROMs, one containing crashes reported on both long and short forms for the years 2000, 2001, and 2002, and the other containing an Excel spreadsheet for each intersection in the city. The spreadsheet included a line

diagram of the intersection as well as the speed limits on the two roads. While the information was organized, the files did not always contain the same information that had been gathered for Seminole County intersections. For instance, City of Orlando diagrams were not detailed enough to show whether right turn lanes were channelized or whether any of the intersection approaches were divided. Even without this information, a new geometry database was created for the city and all possible sites were included. A total of 296 intersections were classified from City of Orlando. When the geometry database was complete, traffic volumes were recorded from the city's transportation website.

3.1.3 Hillsborough County

The next county that cooperated for this analysis was Hillsborough County on the west coast of Florida. County officials provided a CD-ROM containing several hundred county intersection files. Within each file was either a line-diagram or an aerial photograph of the intersection. Some of the files also included speed limit information and annual average daily traffic (AADTs). Intersection geometry was again recorded into a separate geometry database. Since the CD provided by the county did not consistently have AADT, the missing information was sought elsewhere. A report published on the county's website provided initial AADT volumes. However, a more thorough look at the numbers revealed that they were erroneous. A second source of information was found from a level of service report published on the FDOT website. It provided the AADT volumes as well as the level of service, number of through lanes and whether the road was divided. This information served as a check to make sure that streets with relatively low volumes had higher level of services ratings, indicating that the low volumes

are indeed accurate. The other available information served as a safeguard to make sure that the roads were classified correctly. In the end, geometry was recorded for 190 intersections.

The crash records proved to be more difficult to obtain. An excursion to the county became necessary and two days were spent there to gather the required information. Several thousand records were examined and necessary information was noted. Crash records on both long and short forms for Hillsborough County intersections for the years of 1999, 2000 and 2001 were retrieved from this trip.

3.1.4 Brevard County

Brevard County, on Florida's East Coast, was next contacted for relevant information. After several weeks of communication, the county mailed a package containing paper drawings of several hundred county intersections. These drawings were similar to those from other counties and also displayed the same type of information. Again intersection geometry was recorded into a database coded for Brevard County. Most of the drawings were clear and a total of 151 intersections were recorded. Unfortunately, drawings did not include information on crash records, AADT volumes or speed limits. The county was again contacted and was able to send information on crash records for both long and short forms for the years 2000, 2001 and 2002. Unfortunately, the County was unable to supply any further roadway characteristics information. Therefore, AADT volumes were again located on an FDOT level of service report for Brevard County and then entered into a database by hand.

3.2 Summary

Information obtained from each drawing included the number of through lanes on each roadway, the number of left turn lanes and whether they were exclusive for each approach, the presence of medians on each approach, whether any of the right turns were channelized and the speed limits. Each county also provided a database of crashes reported on both long- and short forms for three recent years. In the meanwhile, crashes reported on long forms were also downloaded from FDOT and DHSMV databases and cross-referenced against the crashes reported on long forms provided by the counties. This process served as a check to ensure that each county's database was accurate. It was found that no database by itself was complete and each was missing crashes that another database included. Finally, a complete list of crashes reported on long forms from county, FDOT and DHSMV databases were combined with crashes reported on short forms from the counties' databases to ensure that the dataset for this analysis was complete as much as possible. After the intersection characteristics and crashes had been collected into separate files for each of the four jurisdictions: Seminole County, City of Orlando, Hillsborough County and Brevard County, it became necessary to merge the files. The master database created for this analysis includes 33,592 crashes from 832 intersections. Table 3-1 summarizes the data that was collected. Only data from years 2000 and 2001 would be used in model estimation (those were the only consistent years). Table 3-1 shows that, with the exception of Hillsborough County, all others have more crashes reported on short forms than long forms. This shows that Hillsborough County is reporting more crashes to state agencies. Table 3-2 is a summary of the final data set in all model estimations. Table 3-3 is a summary of the data used to validate the tree-based regression models.

Table 3-1. Summary of Data Collected

County	Number of Intersections	1999		2000		2001		2002		Total
		Long Forms	Short Forms	Long Forms	Short Forms	Long Forms	Short Forms	Long Forms	Short Forms	
Brevard County	151	-	-	490	1009	506	1015	561	1090	4671
City of Orlando	296	-	-	1793	2789	1745	2636	1690	2485	13138
Hillsborough Co.	190	1531	1052	1554	1262	1585	1333	-	-	8317
Seminole County	195	879	1556	799	1706	905	1621	-	-	7466
Combined Total	832	2410	2608	4636	6766	4741	6605	2251	3575	33592

Table 3-2. Summary of Final Data Set

County	Number of Intersections	2000		2001		Total
		Long Forms	Short Forms	Long Forms	Short Forms	
Brevard County	151	490	1009	506	1015	3020
City of Orlando	296	1793	2789	1745	2636	8963
Hillsborough County	190	1554	1262	1585	1333	5734
Seminole County	195	799	1706	905	1621	5031
Combined Total	832	4636	6766	4741	6605	22748

Table 3-3. Summary of Data Used for Tree-Based Regression Validation and Prediction

County	Number of Intersections	2002		Total
		Long Forms	Short Forms	
Brevard County	151	561	1090	1651
City of Orlando	296	1690	2485	4175
Combined Total	447	2251	3575	5826

CHAPTER 4. DATA EXPLORATION

4.1 Initial Consideration

Before organizing the final two-year database, it was necessary to explore the available data. Frequency graphs were created to display the relative amounts of crashes reported on both long and short forms for each type of collision and injury level. These graphs are presented in Figure 4-1 and Figure 4-2. Figure 4-1 shows the percentage of crashes reported on short forms, in gray, for each type of collision, which are excluded in the FDOT and DHSMV databases. These crashes were also excluded in the restricted dataset for this research to show the consequences of modeling an incomplete dataset. In Figure 4-1, it is shown that over 73% of right turn crashes are excluded from state databases because they are reported on short forms. Whereas, over 85% of vehicle crashes involving pedestrians or bicyclists are reported on long forms indicating that most of these crashes involve an injury or a felony. While all crashes reported on short forms are non-injury crashes, there are still some non-injury crashes as well as non-felony crashes that are reported on long forms and the final report form depends on the reporting police officer/police agency. Therefore, there is an inconsistency between municipalities and this discrepancy one of the reasons that it is useful to obtain data from multiple counties. Finally, Figure 4-1 shows that almost 65% of minor, no-injury crashes were unreported to the state agencies because they were reported on short forms. In brief, Figure 4-1 shows that by excluding crashes reported on short forms, not only do the databases exaggerate the average injury level, but they also under-estimate the true number of certain types of crashes such as right turn, rear-end and sideswipe crashes.

Figure 4-2 shows that crashes reported on short forms are primarily minor crashes because they consist solely of non-injury crashes. On the other hand, crashes can be recorded on a long form for any type of crash or severity level and the decision is ultimately up to the reporting police officer.

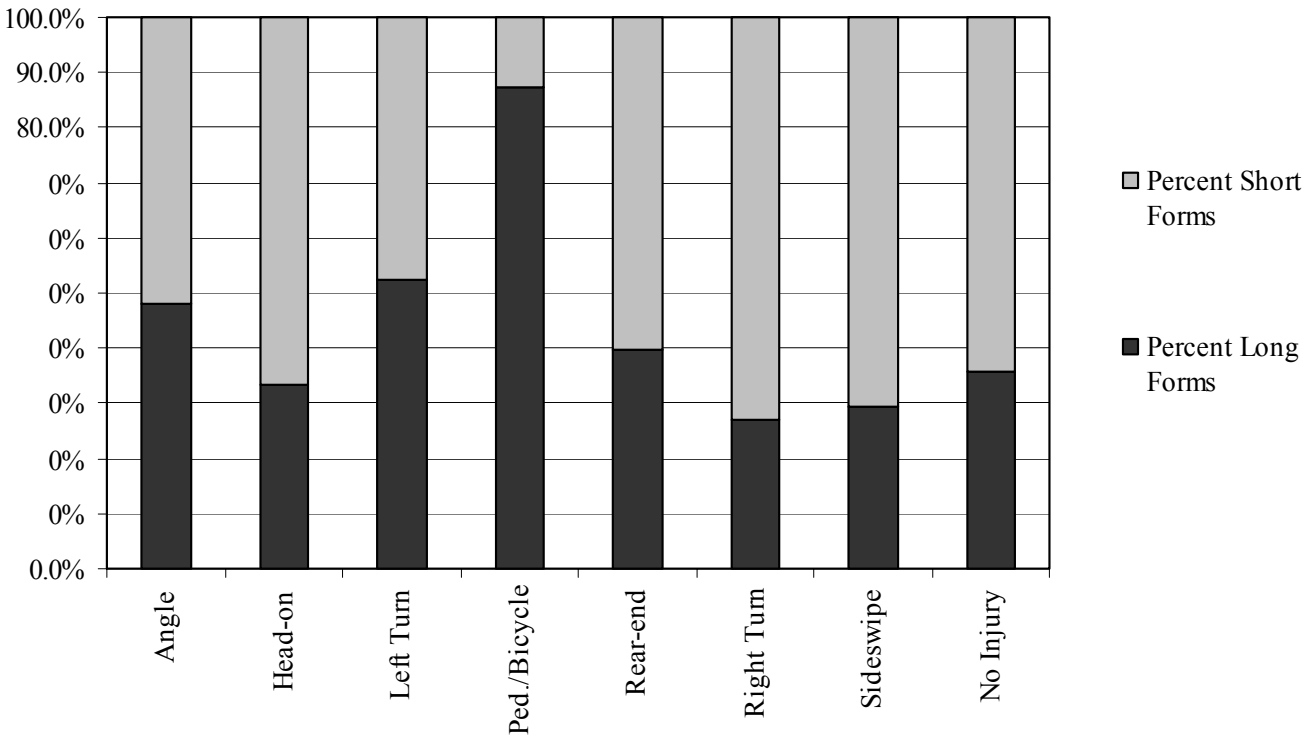

Figure 4-1. Frequency of Crash Types for Crashes Reported on Short and Long Forms for the Combined Four Entities over Four Years

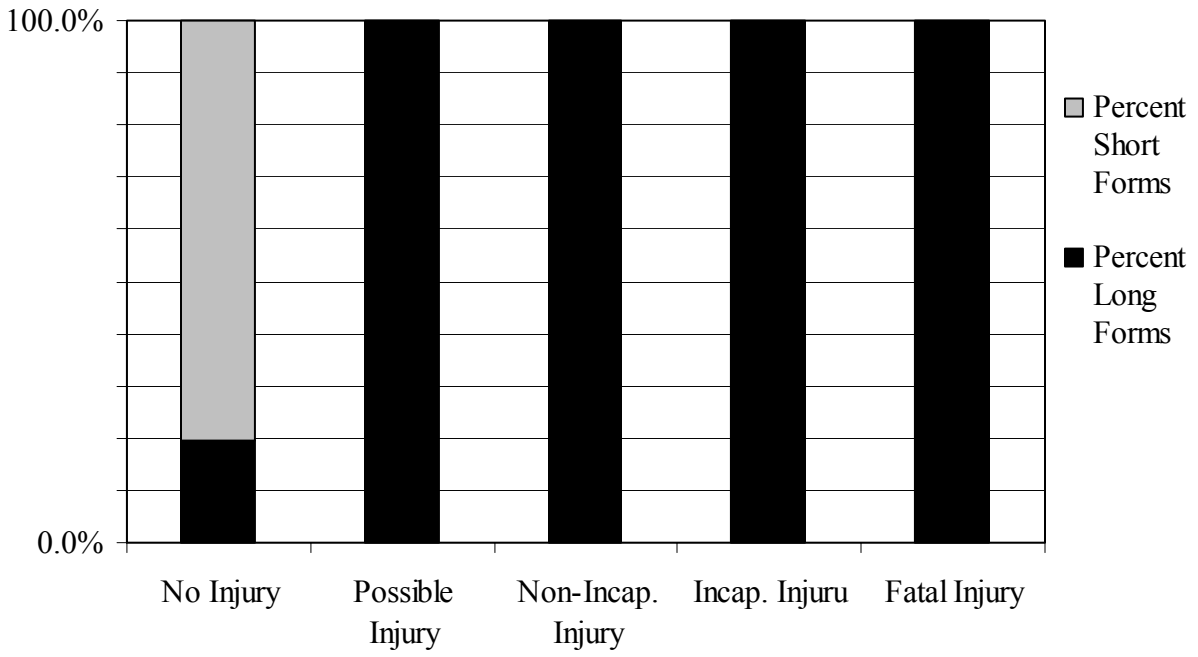

Figure 4-2. Frequency of Injury Levels for Crashes Reported on Short and Long Forms for the Combined Four Entities over Four Years

4.2 Data Structuring

Before conducting any analyses, it was necessary to explore the distributions of the variables. As previously mentioned, the combined data set includes 832 intersections and 22,748 crashes from years 2000 and 2001. The initial database reflected each of the 22,748 crashes, several crash characteristics and the geometric information about the specific intersection where each crash occurred. This database was used to create ordered probit models presented in Chapter 6, which required dummy variables to be created for crash type, severity level and county location. For the tree-based regression, a separate database was created such that each observation was a different intersection where the types of crashes were summed for each observation. This database included 26 possible dependent variables and 14 independent variables. Table 4-1 lists each of the variable names, their definitions and whether each was a dependent or independent variable in the tree-based regression models.

Table 4-1. Variables Included in the Tree-Based Regression Database

Variable	Variable Definition	Variable Role
int_id	Intersection Identification Number	Identification
MJ_Ln	Number of Through Lanes on Major Road	Independent
MN_Ln	Number of Through Lanes on Minor Road	Independent
Tot_LTMJ1 or 2	Number of Left Turn Lanes (LTL's) on Major Road Approach #1 or 2, respectively	Independent
Tot_LTMN1 or 2	Number of Left Turn Lanes (LTL's) on Minor Road Approach #1 or 2, respectively	Independent
Tot_LTLMJ	Total Number of LTL's on Major Road	Independent
Tot_LTLMN	Total Number of LTL's on Minor Road	Independent
LTProt_MJ1 or 2	Total Number of Exclusive LTL's on Major Road Approach #1 or 2, respectively	Independent
LTProt_MN1 or 2	Total Number of Exclusive LTL's on Minor Road Approach #1 or 2, respectively	Independent
LTProt_MJ	Total Number of Exclusive LTL's on Major Road	Independent
LTProt_MN	Total Number of Exclusive LTL's on Minor Road	Independent
RTChMJ1 or 2	Whether Right Turn Lanes are Channelized on Major Road Approach #1or 2, respectively	Independent
RTChMN1 or 2	Whether Right Turn Lanes are Channelized on Minor Road Approach #1 or 2, respectively	Independent
RTChMJ	Whether any or all Right Turns are Channelized on Major, Yes = 1 and No = 0	Independent
RTChMN	Whether any or all Right Turns are Channelized on Minor, Yes = 1 and No = 0	Independent
DivMJ1 or 2	Whether Major Road is Divided (by median or two-way LTL) on Approach #1 or 2, respectively	Independent
DivMN1 or 2	Whether Minor Road is Divided (by median or two-way LTL) on Approach #1 or 2, respectively	Independent
DivMJ	Whether any or both Approaches on Major Road are Divided, Yes = 1 and No = 0	Independent
DivMN	Whether any or both Approaches on Minor Road are Divided, Yes = 1 and No = 0	Independent
SL_MJ	Speed Limit on Major Road	Independent
SL_MN	Speed Limit on Minor Road	Independent
ADT_MJ	Average Daily Traffic on Major Road	Independent
ADT_MN	Average Daily Traffic on Minor Road	Independent
L_Angle	Number of Angle Crashes Reported on Long Forms	Dependent
L_S_Angle	Number of Angle Crashes Reported on Short and Long Forms	Dependent
L_Head	Number of Head-on Crashes Reported on Long Forms	Dependent
L_S_Head	Number of Head-on Crashes Reported on Short and Long Forms	Dependent
L_Left	Number of Left Turn Crashes Reported on Long Forms	Dependent
L_S_Left	Number of Left Turn Crashes Reported on Short and Long Forms	Dependent
L_Ped	Number of Pedestrain/Bicycle Crashes Reported on Long Forms	Dependent
L_S_Ped	Number of Pedestrian/Bicycle Crashes Reported on Short and Long Forms	Dependent
L_Rear	Number of Rear-end Crashes Reported on Long Forms	Dependent
L_S_Rear	Number of Rear-end Crashes Reported on Short and Long Forms	Dependent
L_RT	Number of Right Turn Crashes Reported on Long Forms	Dependent
L_S_RT	Number of Right Turn Crashes Reported on Short and Long Forms	Dependent
L_Side	Number of Sideswipe Crashes Reported on Long Forms	Dependent
L_S_Side	Number of Sideswipe Crashes Reported on Short and Long Forms	Dependent

After formatting the data into the layout mentioned previously, the distributions for each of the dependent variables from Table 4-1 was drawn using the statistical analysis software package (SAS). A total of 16 distributions were drawn for each of the dependent variables listed in the tree-based regression database. Figure 4-3 is an example of these graphs and is for the distribution of the total number of crashes reported on long forms per intersection for two years. In Figure 4-3, the column color is darker for higher percentages of intersections with crashes corresponding to a particular category. The tallest column in the graph corresponds to the number of crashes that is most common in each type, in other words, the mean. For instance, in Figure 4-3, the leftmost column shows that about 53% of the intersections analyzed had less than 7.625 crashes reported on long forms per intersection over the two-year period. Furthermore, only about 0.12% of the intersections had more than 114 crashes reported on long forms over two years. All of the distributions displayed similar trends and the remaining graphs are located in Appendix A.

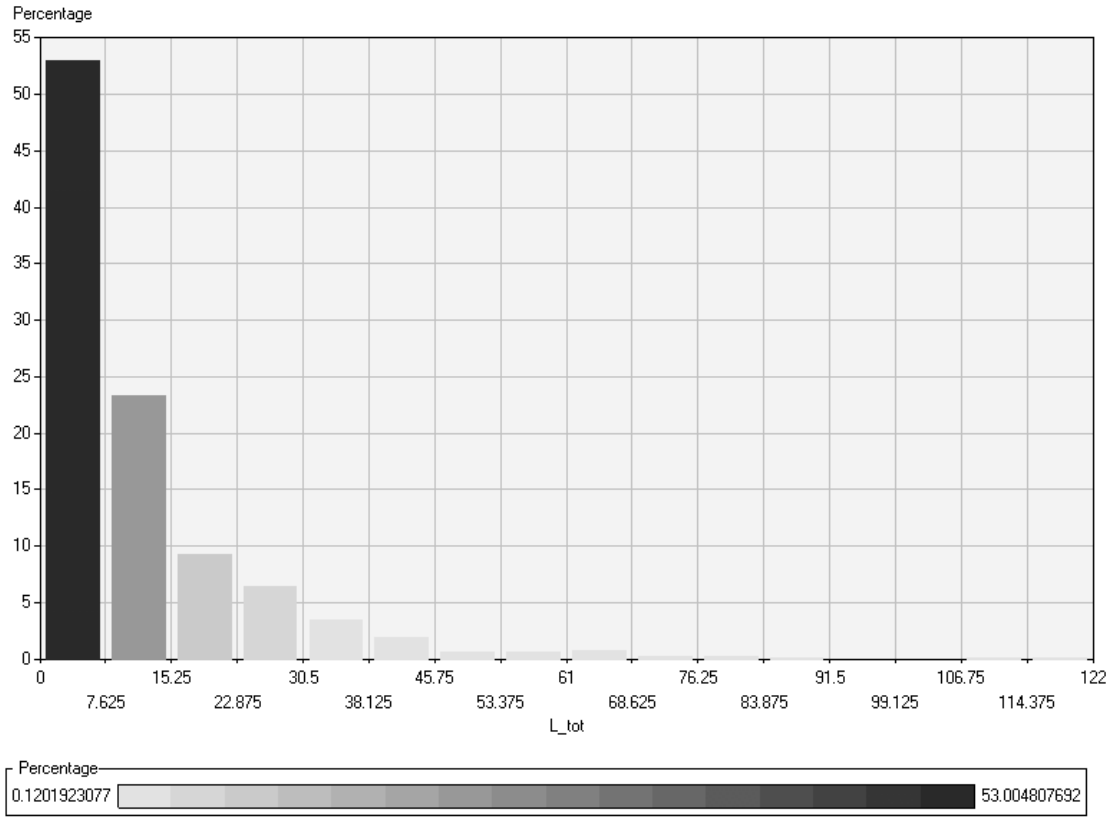

Figure 4-3. Distribution of the Total Number of Crashes Reported on Long Forms

4.3 Summary

It can be seen from the graphed distributions that the dependent variables used for the tree-based regression models do not follow a normal distribution. Poisson and negative binomial distributions are commonly implemented in traffic studies but require making assumptions. For instance, the Poisson distribution assumes that the mean is equal to the variance. Making this assumption would be inaccurate for this particular data set and, as such, the model applied to this data should be either non-parametric such as the hierarchical tree-based regression approach or based upon the probability of occurrence.

CHAPTER 5. NON-PARAMETRIC MODELING

5.1 Model Definition

In order to model this data in a logical way without knowledge of the true model form, hierarchical tree-based regression (HTBR) was used to predict the expected number of crashes reported on both long- and short forms for each type of crash. This method involves splitting the data into branches on a tree diagram based upon the given information and the average or expected value at each node. One of the most important benefits to this type of model is that, since it is based on crash frequencies under different conditions, the model does not require any assumptions or knowledge of the population's functional form in advance. HTBR is also robust against multicollinearity between the variables, which is commonly a problem in crash studies. Additionally, the model is capable of handling missing observations by treating a missing value as a valid response. This was advantageous considering City of Orlando was unable to provide information on whether the right turns on some intersections were channelized and several fields were left blank. In this instance, these missing values did not cause any changes in the output unless the statistical software recognized a specific pattern for all observations with the channelization information missing. Finally, outliers can easily be detected using tree-based regression because if an observation is an outlier, it will be on a branch alone. The model is essentially binary because the tree begins with one parent node that can split into exactly two child nodes. From here, each child node can either split into zero or two more child nodes. Nodes are split based upon the deviance of the sample and the splitting value is chosen such that the deviance in each of the two child nodes is minimized. Karlaftis and Golias (2002) defined the deviance as

$$D = \sum_{i=1}^{L} (Y_{i\,a} - X_a)^2$$

where D is the deviance (also the sum of squared error) of y at node *a*, Y_{ia} is the observation at node *a*, and X_a is the average of L observations in node *a*. The observations in Y_{ia} are divided into two sub-samples, which together contain the original L observations from node *a*. Supposing that the deviance for the original dataset is D_a and that the two new groups are labeled *b* and *c*, with deviances D_b and D_c, respectively, then the reduction is deviance (Karlaftis and Golias 2002) can be defined as

$$\Delta = D_a - D_b - D_c$$

Tree diagrams are split based upon the variables that maximize this reduction in deviance and variables that do not cause a reduction in deviance are insignificant in the model creation. The analysis was conducted using SAS where stepwise variable selection as well as a splitting criterion based on an F-test was engaged.

5.2 Methodology

Tree-based regressions were conducted for each type of collision and for the total number of crashes, for both restricted and complete datasets resulting in a total of 16 regressions. To visually identify the difference between models based on the restricted and complete datasets, a total of 16 tree diagrams were produced (refer to Table 4-1 for variable identification). Figure 5-1 was chosen as an example of the regression trees from this research. The top box in this figure contains the model name, which shows that Figure 5-1 was created for the prediction of the number of angle crashes reported on long forms per intersection for two years. The second box from the top reflects the number of observations in the dataset as well as the average number of

angle crashes reported on long forms per intersection for the dataset. The third box contains the name of the independent variable that the data was split by to cause the largest decrease in deviance. For Figure 5-1, the number of through lanes on the minor road was found to minimize the deviance most. The result is that the tree diagram breaks into two branches and then has the opportunity to branch again. In this case, the left branch is further divided by number of exclusive left turn lanes on the minor roadway and the right branch is divided by whether the major roadway is divided. Each of the new branches again splits one more time before stating the final expectation for the number of crashes. As an example, consider an intersection where the minor roadway has 4 through lanes, whether the major road is divided is not available (as is the case for some City of Orlando data), and the ADT on the minor road is less than 25387.5. From this tree it can be found that a total of 4.491 crashes are expected over a two-year period for this hypothetical intersection, based upon 55 observations on similar intersections. This prediction capability also serve as a planning tool since crashes can be forecasted for alternative signalized intersection designs to determine the safest intersection design for a specific application. The remaining tree diagrams created for the crash-type analysis are located in Appendix B.

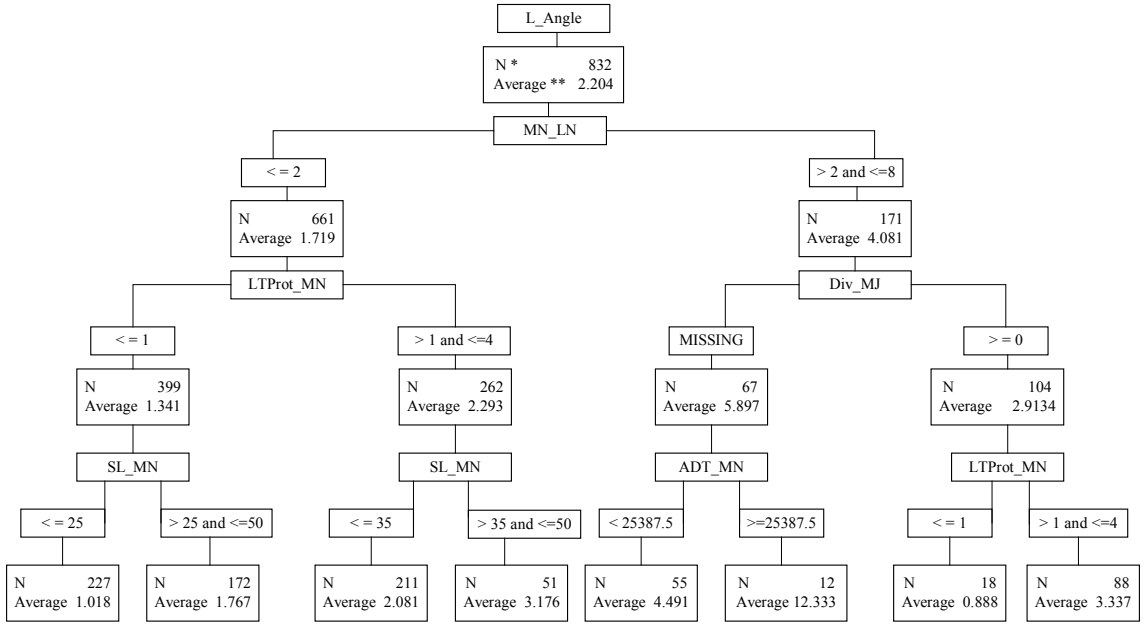

Figure 5-1. Regression Tree for the Expected Number of Angle Crashes Reported on Long forms Per Intersection for Two Years

5.3 Importance of Factors

In addition to creating tree diagrams to visually describe the difference in models, lists of variables that entered into each model and their relative importance were also produced. Variables found to be significant were identified as "input" variables, and "rejected" variables were those that did not enter the particular model. According Karlaftis and Golias (2002), to determine each variable's importance, the improvement in the reduction of deviance that can be attributed to each variable for the first split in the tree is rated. These values are then summed and scaled, showing the variable that reduced the deviation most to have an importance value of 1.00. A total of 16 important-factors tables were created. In order to illustrate the importance of these tables, Table 5-1 for angle crashes is included as an example. In Table 5-1, the model

based on crashes reported on long forms (indicated as the L_Angle model) shows that the most important factor for determining the number of angle crashes is the number of through lanes on the minor road. However, when the crashes reported on short forms (indicated as L_S_Angle) are added to the model, the relative importance of the number of through lanes on the minor road falls to 0.9114 and the most important factor becomes the number of exclusive left turn lanes on the major roadway. Additionally, for the first model in Table 5-1, the ADT for the major road was found to be insignificant; however, upon the addition of the crashes reported on short forms, this volume is now significant and has a relative importance of 0.2928. The remaining important-factors tables are located in Appendix C.

Table 5-1. List of Variables that Entered the Models based upon Angle Crashes and their Relative Importance

L_Angle		
Name	**Importance**	**Role**
MN_LN	1.0000	Input
ADT_MN	0.8936	Input
DIV_MJ	0.6911	Input
LTPROT_MN	0.5994	Input
SL_MN	0.3708	Input
MJ_LN	0.1842	Input
ADT_MJ	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

L_S_Angle		
Name	**Importance**	**Role**
LTPROT_MJ	1.0000	Input
MN_LN	0.9114	Input
RTCHMJ	0.5929	Input
ADT_MN	0.5518	Input
RTCHMN	0.3551	Input
LTPROT_MN	0.3504	Input
ADT_MJ	0.2928	Input
SL_MJ	0.2454	Input
DIV_MN	0.2209	Input
TOT_LTLMN	0.1244	Input
DIV_MJ	0.0000	Rejected
MJ_LN	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected

5.3.1 Contrast Among Types of Collisions

The main objective of this chapter was to explore the changes in important factors between different collision types. Due to the fact that most research conducted is based solely on crashes reported on long forms and in an effort to make the results more comparable to other studies, this section analyzes several crash characteristics using only crashes reported on long forms. Table 5-2 was created to illustrate the consequences of aggregating the types of collisions and creating only one model based on the total number of crashes. The most important variable in each model is identified with a relative importance value of 1.00 and a value of 0.00 indicates that the variable was found to be insignificant. The table is arranged by the variables and their significance level in decreasing order for the model based upon the total number of crashes reported on long forms. The other columns represent a different type of crash. For example, the most important factor for determining the total number of crashes in the restricted dataset was found to be the number of lanes on the minor roadway (indicated by an importance value of 1.000). This factor was also found to be the most important in angle crashes and its importance dropped to 0.4421 for left turn crashes. Furthermore, for each of the other types of crashes, the number of lanes on the minor road was found to be insignificant. From Table 5-2 it can be seen that the models created for head-on and left turn crashes change the most from the model created from the total number of crashes reported on long forms. The total number of left-turning lanes on the major road was found to be insignificant in all models created. In contrast, the most important factor in determining the number of left turn crashes in the restricted dataset was found to be the number of exclusive left turn lanes on the minor road. The most important factor for predicting the number of head-on crashes was whether there is a median on the minor road possibly because the presence of a median prevents vehicles from crossing into the path of

oncoming traffic. The most important factor in the expected number of right turn crashes is the traffic volume along the major roadway and the volume on the minor roadway was most significant for the number of sideswipe crashes.

Table 5-2. Relative Importance of Independent Variables for each Type of Crash for the Restricted Datasets

Variables	Total Crashes Restricted Dataset	Angle Crashes Restricted Dataset	Left Turn Crashes Restricted Dataset	Head-on Crashes Restricted Dataset	Ped/Bike Crashes Restricted Dataset	Rear-end Crashes Restricted Dataset	Right Turn Crashes Restricted Dataset	Sideswipe Crashes Restricted Dataset
Number of Lanes on Minor Road	1.0000	1.0000	0.4421	0.0000	0.0000	0.0000	0.0000	0.0000
Exclusive Left Turn Lanes on Minor Road	0.8551	0.5994	1.0000	0.3849	0.6265	0.5058	0.0000	0.0000
Right Turns Channelized on Major Road	0.7616	0.0000	0.0000	0.0000	1.0000	0.7260	0.0000	0.0988
Speed Limit on Major Road	0.5429	0.0000	0.4256	0.6152	0.0000	0.4503	0.0000	0.0000
Number of Lanes on Major Road	0.5335	0.1842	0.0000	0.0000	0.0000	0.3837	0.7186	0.2902
Daily Traffic Volume on Major Road	0.4281	0.0000	0.2753	0.0000	0.5122	0.7074	1.0000	0.4118
Daily Traffic Volume on Minor Road	0.3317	0.8936	0.0000	0.8482	0.0000	0.4618	0.0000	1.0000
Speed Limit on Minor Road	0.1624	0.3708	0.7894	0.0000	0.3257	0.3926	0.0000	0.0000
Median Present on Major Road	0.0000	0.6911	0.0000	0.5179	0.0000	0.4058	0.0000	0.9679
Median Present on Minor Road	0.0000	0.0000	0.0000	1.0000	0.0000	0.2830	0.9039	0.0000
Exclusive Left Turn Lanes on Major Road	0.0000	0.0000	0.5840	0.7268	0.0000	1.0000	0.5040	0.2299
Right Turns Channelized on Minor Road	0.0000	0.0000	0.6238	0.0000	0.0000	0.0000	0.0000	0.0000
Total Left Turn Lanes on Major Road	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Total Left Turn Lanes on Minor Road	0.0000	0.0000	0.3535	0.8162	0.2246	0.0000	0.0000	0.0000

5.3.2 Contrast Between Complete and Restricted Datasets

In addition to determining the differences of the important factors between collision types, Table 5-3 was created to show the relative importance of the independent variables in each of the models created for both the complete and restricted datasets. This table was created using the same data as in the previous section. By organizing the relative importance values in an array, it became clear that there is a difference between models based on the restricted and complete datasets. The complete dataset includes all crashes whereas the restricted dataset includes only crashes reported on long forms. For rear-end, right turn and sideswipe crashes, the important factors are fairly consistent between the models created by complete and restricted datasets. For example, the traffic volumes on both the intersecting roadways are important variables for explaining the number of rear-end crashes for both the complete and restricted datasets. Furthermore, the traffic volume on the major roadway is consistently an important factor for the number of right turn crashes. Finally, for sideswipe crashes, the important factors do not change drastically as the minor road's traffic volume and whether the major road is divided are the two most important factors for both models. These results show that factors in crashes causing mostly injuries or involving felony crimes for rear-end, right turn and sideswipe crashes are generally the same as factors causing non-injury or minor crashes. This indicates that models based on complete and restricted datasets for these types of crashes are roughly equivalent. On the other hand, important factors for angle and head-on crashes changed the most between the models because these types of crashes are unstable and different factors result in non-injury or minor crashes. The crash-causing factors were found to be significantly different when crashes reported on short forms were added to the dataset. For example, the volume on the major road was insignificant for angle crashes in the restricted dataset; however, when minor

crashes are added to the model, this variable becomes significant. Similarly, the total number of left turn lanes on the minor roadway was insignificant for the restricted model and significant in the complete model. The number of exclusive left turn lanes on the minor road was consistently the most important factor for left turn crashes; however, the number of exclusive left turn lanes on the major road did not become significant until crashes reported on short forms were added to the model. The number of exclusive left turn lanes was also the most important factor for rear-end crashes from the complete dataset. The presence of a median had an effect on the number of head-on crashes for both the complete and restricted models. The speed limit on the major road was found to be significant in 9 of the 16 models created; it was consistently insignificant for the right turn and sideswipe models. The speed limit on the minor road was insignificant in the restricted models for total, angle, left turn, head-on, rear-end, right turn and sideswipe but became significant in several of these models when crashes reported on short forms were added. From Table 5-3 it can be seen that the relative change of importance of factors between complete and restricted models is the greatest for angle crashes. The next type of crash that is most affected by the addition of crashes reported on short forms is head-on crashes. The most significant variable in head-on crashes for the restricted dataset is whether the minor road was divided, which became insignificant when non-injury crashes were added to the dataset.

Table 5-3. Relative Importance of Independent Variables for each Type of Crash for the Complete and Restricted Datasets

Variables	Total Crashes		Angle Crashes		Left Turn Crashes		Head-on Crashes		Ped/Bike Crashes		Rear-end Crashes		Right Turn Crashes		Sideswipe Crashes	
	Compl.	Restr'd	Compl.	Restr'd	Compl.	Restr'd	Compl.	Restr'd	Compl.	Restr'd	Compl.	Restr'd	Compl.	Restr'd	Compl.	Restr'd
Number of Lanes on MN	1.0000	1.0000	1.0000	0.9114	0.4421	0.6851	0.0000	0.0000	0.0000	0.0000	0.0000	0.5742	0.0000	0.4791	0.0000	0.0846
Exclusive Left Turn Lanes on MN	0.8551	0.8917	0.5994	0.3504	1.0000	1.0000	0.3849	0.6990	0.6265	0.5407	0.5058	0.8056	0.0000	0.0000	0.0000	0.0803
Right Turns Channelized on MJ	0.7616	0.7527	0.0000	0.5929	0.0000	0.6268	0.0000	0.1044	1.0000	1.0000	0.7260	0.4154	0.0000	0.0000	0.0988	0.0000
Speed Limit on MJ	0.5429	0.2466	0.0000	0.2454	0.4256	0.2532	0.6152	0.7237	0.0000	0.0000	0.4503	0.4631	0.0000	0.0000	0.0000	0.0000
Number of Lanes on MJ	0.5335	0.4664	0.1842	0.0000	0.0000	0.4337	0.0000	0.1294	0.0000	0.0000	0.3837	0.0000	0.7186	0.1270	0.2902	0.2287
Daily Traffic Volume on MJ	0.4281	0.6866	0.0000	0.2928	0.2753	0.4810	0.0000	0.7068	0.5122	0.7048	0.7074	0.8116	1.0000	0.7969	0.4118	0.4596
Daily Traffic Volume on MN	0.3317	0.4085	0.8936	0.5518	0.0000	0.0000	0.8482	0.9465	0.0000	0.0000	0.4618	1.0000	0.0000	0.0000	1.0000	1.0000
Speed Limit on MN	0.1624	0.0000	0.3708	0.0000	0.7894	0.0000	0.0000	0.0000	0.3257	0.1058	0.3926	0.0000	0.0000	0.0000	0.0000	0.0000
Median Present on MJ	0.0000	0.1183	0.6911	0.0000	0.0000	0.0000	0.5179	1.0000	0.0000	0.1890	0.4058	0.7815	0.0000	1.0000	0.9679	0.8654
Median Present on MN	0.0000	0.2721	0.0000	0.2209	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.2830	0.2997	0.9039	0.0000	0.0000	0.0000
Exclusive Left Turn Lanes on MJ	0.0000	0.0000	0.0000	1.0000	0.5840	0.0000	0.7268	0.0000	0.0000	0.0000	1.0000	0.0000	0.5040	0.6134	0.2299	0.7293
Right Turns Channelized on MN	0.0000	0.0000	0.0000	0.3551	0.6238	0.4783	0.0000	0.3179	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Total Left Turn Lanes on MJ	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2819	0.0000	0.0000
Total Left Turn Lanes on MN	0.0000	0.0000	0.0000	0.1244	0.3535	0.3176	0.8162	0.0000	0.2246	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

5.4 Validation of Models

In order to validate the regression models created, expected crash values were projected for year 2002 in Brevard County and City of Orlando and then compared to the actual number of crashes to determine the error rate. The results showed that the discrepancies were slight and, for the most part, the predicted number of crashes was reasonably close to the actual number of crashes. Crashes were predicted in terms of crashes per intersection over two years and in order to predict the number of crashes for only one year, the prediction values were halved. These values are located in the column labeled "Predicted Number of Crashes" in Table 5-4. In particular, for pedestrian and bicycle crashes reported on long forms, the model predicted the actual number of crashes that occurred showing a 0.0% error.

It can be seen in Table 5-4 that including minor crashes in the model causes the expected number of crashes to increase. This indicates that if the objective were to estimate the total number of a certain type of crashes, excluding crashes reported on short forms would make the model inaccurate. Furthermore, shown Table 5-4, the model accurately predicted the total number of crashes in both datasets with error rates of 5.9% and 11.1%. The models are most successful in predicting the number of crashes reported on long forms, particularly for angle, pedestrian/bicycle, rear-end and sideswipe crashes with percent errors of 2.9%, 0.0%, 7.3% and 6.3%, respectively.

Table 5-4. Prediction Errors Between the Actual and Predicted Number of Crashes in Brevard County and City of Orlando for Year 2002

Type of Collision	Actual Number of Crashes	Predicted Number of Crashes	Prediction Errors
Restricted Dataset- Angle Crashes	500	515	2.9%
Complete Dataset- Angle Crashes	1048	1229	14.7%
Restricted Dataset- Head-on Crashes	9	15	40.0%
Complete Dataset- Head-on Crashes	29	50	42.0%
Restricted Dataset- Left Turn Crashes	260	326	20.2%
Complete Dataset- Left Turn Crashes	596	753	20.8%
Restricted Dataset- Ped/Bike Crashes	61	61	0.0%
Complete Dataset- Ped/Bike Crashes	67	76	11.8%
Restricted Dataset- Rear-end Crashes	1001	1080	7.3%
Complete Dataset- Rear-end Crashes	2701	3412	20.8%
Restricted Dataset- Right Turn Crashes	7	12	41.7%
Complete Dataset- Right Turn Crashes	34	50	32.0%
Restricted Dataset- Sideswipe Crashes	225	240	6.3%
Complete Dataset- Sideswipe Crashes	804	832	3.4%
Restricted Dataset- Total Crashes	2251	2391	5.9%
Complete Dataset- Total Crashes	5826	6554	11.1%

5.5 Summary

Hierarchical tree-based regression was the analysis method chosen for this chapter for several reasons. Most importantly, since it is based on crash frequencies under different conditions, the model does not require any assumptions or knowledge of the true functional form in advance. This type of regression is also robust against multicollinearity between the variables, which was pointed out as a problem in the literature. Additionally, the model is capable of handling missing observations by treating a missing value as a valid response, which was useful since some of the variables required for the regression were not available from all sources. Considering missing values as valid was essential because these missing values often caused the statistical software to recognize a specific pattern and tree diagrams were split accordingly. Finally, outliers could have been easily identified using tree-based regression because if an observation is a severe outlier, it will be on a branch alone. For this study, there were no outliers detected.

The main objective of this research was to determine the factors found to be significant for different collision types. Results from Chin and Quddus (2003) and Liu and Young (2004) were that traffic volumes were the most important factor in predicting crashes while results presented in this paper show that the traffic volume along the major roadway was the most important factor only for predicting right turn crashes in the restricted dataset. On the other hand, since roadway volume is often related to other geometric characteristics, the importance of roadway volume in determining the number of crashes may have been captured by other variables. Similar to Oh et al. (2004), speed limits were found to be important for the total number of crashes as well as angle, left turn, head-on, pedestrian/bicycle and rear-end crashes.

Meanwhile, the speed limits on both the intersecting streets were insignificant for right turn and sideswipe crashes most likely because these crash types generally involve vehicles traveling in the same direction and adhering to the same posted speed limits causing a smaller speed differential between the colliding vehicles. Steinman and Hines (2004) found that protected left turns and speed limits both affected the number of crashes involving pedestrians and bicyclists, which were similar to the results presented herein. Table 5-3 shows that the most important factor for determining the number of pedestrian/bicycle crashes is whether the right turn lanes are channelized on the major road. Right turn channelization was also found to be significant in the models for total crashes as well as left turn, rear-end and sideswipe crashes. The main conclusion found is that different collision types often rely on different variables to predict the number of expected crashes and crash predictions from aggregated models may be inaccurate.

The second objective of this research was to determine if there was a difference between models based on restricted and complete datasets. Crashes reported on short forms, mostly property-damage-only crashes (PDO), are often ignored by state agencies that maintain records based only on crashes involving an injury or a felony and only some PDO crashes. Figure 4-1 was created to show the relative amount of crashes reported on short forms that are not included in state crash databases. By creating regression models for each type of collision and comparing the changes in the relative importance of each variable, it was found that angle and head-on crash models are most affected by the addition of crashes reported on short forms. The important factors for rear-end, right turn and sideswipe crashes remained consistent between complete and restricted datasets because the factors that cause injury or felony crashes are roughly the same as the factors that cause non-injury crashes.

Finally, to assess the precision of the models, the number of crashes expected in 2002 was calculated for City of Orlando and Brevard County. It was found that the models predict most accurately the restricted datasets for the individual crash types, however, both models based on the total number of crashes reasonably predicted the actual number of crashes in the complete and restricted datasets.

In summary, the results of this research showed that when attempting to forecast the number of expected crashes, it is imperative that models are developed for each type of collision instead of aggregating crash types to predict the total number of crashes. Furthermore, results showed that crashes reported on short forms are important when modeling the number of expected crashes and should therefore be documented in every crash database maintained by state agencies.

CHAPTER 6. CRASH SEVERITY ANALSYSIS

6.1 Model Definition

Due to the fact that some variables are naturally ordered, such as the severity level in a motor-vehicle crash, various types of models can be specified for these types of data. The data for this research included crash-specific information such as the injury type, which was categorized into one of five groups: no-injury, possible injury, non-incapacitating injury, incapacitating injury and fatal injury. These groups were then ranked from 0 to 4 with no-injury corresponding to the lowest level. Ordered probit models have gained popularity for this type of data mainly because they can account for the dependent variable's ordinal nature. The ordered multiple-choice model is as follows:

$$\sum_{j=1}^{j} P_n(j) = F(\alpha_j - \beta_j X_n, \theta),\ j = 1,\ldots,J-1$$

$$P_n(J) = 1 - \sum_{j=1}^{J-1} P_n(j)$$

where $P_n(j)$ is the probability that subject n belongs to category j, α_j is the alternative specific constant, X_n is a vector of measurable characteristics, β_j is a vector of estimable coefficients and θ is a parameter that controls the shape of the probability distribution F (Abdel-Aty, 2003). By assuming a standard normal distribution for F, the ordered probit model has the following form:

$$P_n(1) = \phi(\alpha_1 - \beta_j X_n)$$

$$P_n(j) = \phi(\alpha_j - \beta_j X_n) - \phi(\alpha_{j-1} - \beta_j X_n),\ j = 2,\ldots,j-1$$

$$P_n(J) = 1 - \sum_{j=1}^{J-1} P_n(j)$$

where ϕ is the cumulative standard normal distribution function. The predicted outcome is the j value with the largest probability (Abdel-Aty, 2003). Ordered probit models were created for this analysis using the econometric software LIMDEP. The data used in this section was the same as in the previous section with the exception that crashes missing severity information were excluded. The ordered probit models created in this chapter were based on crashes from the four counties/city for the years 2000 and 2001 as seen in Table 3-2. Only crashes where the exact injury severity was known were used for analysis. This same data was also used to estimate the prediction power of the models created. There were 7,833 crashes reported on long forms and 13,371 crashes reported on short forms, making a total of 21,204 crashes used for these ordered probit models. Figure 6-1 shows the frequency of each type of injury for crashes reported on both long and short forms. The main objectives for this analysis were to determine the factors affecting crash severity as well as determine if there is a difference when models are based on the completeness of the data.

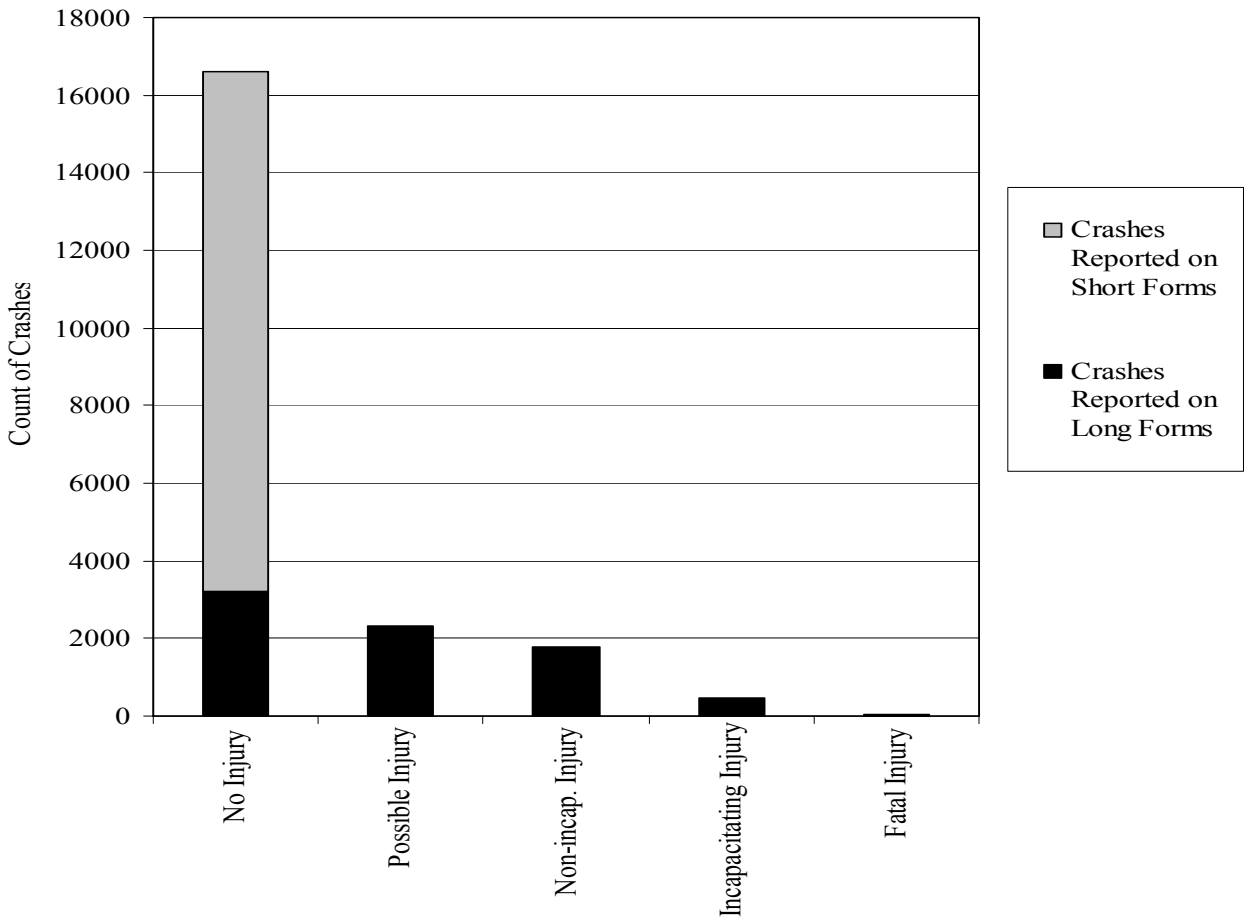

Figure 6-1. Frequency of Injury Severity Level for Crashes

6.2 Severity Models for Crash Types

The first ordered probit models created were arranged so that seven of the independent variables were dummies each representing different crash types: angle, head-on, left turn, pedestrian/bicycle, rear-end, right turn, sideswipe and other/unknown crashes (although there are eight types of crashes, only seven appeared in the model to prevent the dummy-variable trap). The last three independent variables were dummy variables for the location of the crash to account for any county-specific factors. Again, there are four counties/city but only three were

introduced into the model to prevent the dummy variable trap. Two models were created, one for the restricted dataset (based only on crashes reported on long forms) and the other for the complete dataset (based on crashes reported on both long and short forms). Table 6-1 shows the variables used in each model as well as the coefficients, t-statistics, p-values, Chi-squared test statistics, log likelihood functions and restricted log likelihood values.

Table 6-1. Variable Coefficients for Crash-Type Models

Restricted Dataset			
Variable	**Coefficient**	**t-Statistic**	**P-Value**
Constant	-0.5130	-11.552	0.0000
Angle	0.4596	10.358	0.0000
Head-on	0.6677	5.532	0.0000
Left Turn	0.5983	12.879	0.0000
Ped/Bike	1.1905	14.819	0.0000
Rear-end	0.2456	6.031	0.0000
Sideswipe	-0.3507	-5.475	0.0000
Brevard Co.	0.6755	13.582	0.0000
City of Orlando	0.5946	15.067	0.0000
Hillsborough Co.	0.4717	14.007	0.0000
α_1	0.8189	54.793	0.0000
α_2	1.8428	72.067	0.0000
α_3	3.0419	45.530	0.0000
Sample Size		7833	
Degrees of Freedom		9	
Chi-squared		783.450	
Log Likelihood Function		-9423.603	
Restricted Log Likelihood		-9815.328	

Complete Dataset			
Variable	**Coefficient**	**t-Statistic**	**P-Value**
Constant	-1.3490	-35.815	0.0000
Angle	0.3846	10.220	0.0000
Head-on	0.2469	2.891	0.0038
Left Turn	0.5617	14.148	0.0000
Ped/Bike	1.4982	19.123	0.0000
Rear-end	0.1480	4.267	0.0000
Right Turn	-0.3702	-4.110	0.0000
Sideswipe	-0.4006	-7.946	0.0000
Brevard Co.	0.4050	11.918	0.0000
City of Orlando	0.2995	10.555	0.0000
Hillsborough Co.	0.6477	23.282	0.0000
α_1	0.4883	51.150	0.0000
α_2	1.3039	61.542	0.0000
α_3	2.4079	40.256	0.0000
Sample Size		21204	
Degrees of Freedom		10	
Chi-squared		1656.981	
Log Likelihood Function		-14783.55	
Restricted Log Likelihood		-15612.04	

In Table 6-1, the values in the first model were calculated using only crashes reported on long forms, the second set of values were calculated using crashes reported on both forms. In the restricted model, right turn crashes were found to be insignificant in the prediction of injury severity, however, in the complete model right turn crashes were found cause lower injury crashes. The restricted dataset, which included 7,833 crashes reported on long forms, had a fairly

low prediction rate of 43.6%. However, when crashes reported on short forms were added to the dataset, the prediction rate increased to 78.4%. Table 6-2 shows the number of actual and predicted severity levels for both models.

Table 6-2. Predicted Crash Severity Levels for Crash Type Models

Restricted Dataset						
	Predicted					
Actual	0	1	2	3	4	Total
0	2759	354	110	0	0	3223
1	1724	376	214	0	0	2314
2	1106	411	279	0	0	1796
3	299	109	63	0	0	471
4	12	5	12	0	0	29
Total	5900	1255	678	0	0	7833

Complete Dataset						
	Predicted					
Actual	0	1	2	3	4	Total
0	16570	0	24	0	0	16594
1	2282	0	32	0	0	2314
2	1747	0	49	0	0	1796
3	443	0	28	0	0	471
4	27	0	2	0	0	29
Total	21069	0	135	0	0	21204

Since the prediction power of the complete dataset was much higher, interpretations were based on this model. These coefficients show that the crash type likely to have the highest injury level is a crash involving a pedestrian or bicyclist ($\beta = 1.4982$). Of the motor-vehicle crashes, left turn, angle and head-on crashes cause the most severe injury levels. On the other hand, it was found that right turn and sideswipe crashes tend to result in a lower crash injury level due to the negative coefficients in the model. Also, the model shows that of Brevard County, City of Orlando and Hillsborough County, the latter is more likely to be the location of a severe crash. Finally, the marginal effects for these models were calculated and are in Table 6-3.

Table 6-3. Marginal Effects for Crash Type Models for Severity Level

Marginal Effects - Restricted Dataset				
Variable	No Injury	Possible Injury	Non-incap. Injury	Incap. Injury
Constant	0.1992	-0.0268	-0.1164	-0.0520
Angle	-0.1785	0.0240	0.1043	0.0466
Head-on	-0.2592	0.0349	0.1515	0.0677
Left Turn	-0.2323	0.0313	0.1357	0.0607
Ped/Bike	-0.4622	0.0622	0.2701	0.1208
Rear-end	-0.0954	0.0128	0.0557	0.0249
Sideswipe	0.1362	-0.0183	-0.0796	-0.0356
Brevard Co.	-0.2623	0.0353	0.1532	0.0685
City of Orlando	-0.2309	0.0311	0.1349	0.0603
Hillsborough Co.	-0.1828	0.0246	0.1068	0.0478

Marginal Effects - Complete Dataset				
Variable	No Injury	Possible Injury	Non-incap. Injury	Incap. Injury
Constant	0.3797	-0.1556	-0.1696	-0.0518
Angle	-0.1083	0.0444	0.0483	0.0148
Head-on	-0.0695	0.0285	0.0310	0.0095
Left Turn	-0.1581	0.0648	0.0706	0.0216
Ped/Bike	-0.4217	0.1728	0.1883	0.0576
Rear-end	-0.0417	0.0171	0.0186	0.0057
Right Turn	0.1042	-0.0427	-0.0465	-0.0142
Sideswipe	0.1128	-0.0462	-0.0503	-0.0154
Brevard Co.	-0.1140	0.0467	0.0509	0.0156
City of Orlando	-0.0843	0.0345	0.0376	0.0115
Hillsborough Co.	-0.1823	0.0747	0.0814	0.0249

One issue with this type of analysis is that a crash must first occur in order to predict the severity level. Therefore, this analysis is only useful when trying to estimate the types of crashes that cause more severe injuries. To cope with this dilemma, models involving only intersection characteristics were created next.

6.3 Severity Models for Intersection Characteristics

In an effort to predict the expected crash severity levels for, say, a newly constructed intersection where no crashes have previously occurred, there was a need to create models based on other measurable variables such as intersection characteristics. The independent variables available for these models are the variables listed in Table 4-1 that relate to the actual intersections. Again, two models were created, one for the restricted dataset and the other for the complete dataset. For these models, backward selection was utilized for simplification so that only variables found to be significant were included in the model. Table 6-4 shows the results of these ordered probit models.

Table 6-4. Variable Coefficients for Intersection Characteristic Models

Restricted Dataset			
Variable	**Coefficient**	**t-Statistic**	**P-Value**
Constant	-0.0177	-5.950	0.0000
Speed Limit on MN	0.0001	3.866	0.0001
Brevard Co.	0.7835	14.297	0.0000
City of Orlando	0.5202	13.016	0.0000
Hillsborough Co.	0.4904	14.675	0.0000
α_1	0.7890	54.781	0.0000
α_2	1.7768	71.405	0.0000
α_3	2.9327	46.333	0.0000
Sample Size		7833	
Degrees of Freedom		4	
Chi-squared		315.1828	
Log Likelihood Function		-9657.736	
Restricted Log Likelihood		-9815.328	

Complete Dataset			
Variable	**Coefficient**	**t-Statistic**	**P-Value**
Constant	0.082980	2.608	0.0091
Major No. of Lanes	-0.009710	-2.978	0.0029
MJ Left Turn Lanes	0.022178	2.528	0.0115
RT Channel. On MJ	-0.075410	-2.881	0.0040
Division on MN	-0.006740	-6.647	0.0000
Speed Limit on MN	-0.000013	-9.361	0.0000
ADT on MJ	0.000001	2.050	0.0404
Brevard Co.	0.481081	11.164	0.0000
City of Orlando	0.516818	15.321	0.0000
Hillsborough Co.	0.207549	9.127	0.0000
α_1	0.4551	50.100	0.0000
α_2	1.2045	59.925	0.0000
α_3	2.2156	29.555	0.0000
Sample Size		21204	
Degrees of Freedom		-	
Chi-squared		-	
Log Likelihood Function		-21280.62	
Restricted Log Likelihood		-	

The model based on the restricted dataset had a relatively low prediction rate of 41.1% while the model based on the complete dataset maintained the prediction rate of 78.4%. Table 6-5 shows the number of actual and predicted severity levels for both models.

Table 6-5. Predicted Crash Severity Levels for Characteristics Models

Restricted Dataset						
	Predicted					
Actual	0	1	2	3	4	Total
0	3223	0	0	0	0	3223
1	2314	0	0	0	0	2314
2	1796	0	0	0	0	1796
3	471	0	0	0	0	471
4	29	0	0	0	0	29
Total	7833	0	0	0	0	7833

Complete Dataset						
	Predicted					
Actual	0	1	2	3	4	Total
0	15704	0	890	0	0	16594
1	1280	0	1034	0	0	2314
2	876	0	920	0	0	1796
3	412	0	59	0	0	471
4	13	0	16	0	0	29
Total	18285	0	2919	0	0	21204

In addition to the variables shown in Table 6-4, a log transformation on the variable for ADT on the major road was tested. It was found that the variable Log(ADT) was insignificant and caused a decrease in the prediction power of the model. Therefore, no transformations were used in the model.

Since the complete model was proven to be more accurate in Table 6-5, interpretations were again based on the model calculated from the complete dataset. Increases in the number of lanes and speed limit on the minor road, right turn channelization on the major road and division on the minor road were found to decrease the expected level of injury. Meanwhile, increases in the number of left-turning lanes as well as traffic volume on the major road were found to increase the crash severity level. For this model, City of Orlando intersections were found to have a higher crash severity risk than Brevard County and Hillsborough County because of the relatively larger coefficient associated with City of Orlando. The marginal effects for this model are in Table 6-6.

Table 6-6. Marginal Effects for Characteristics Models for Severity Level

Marginal Effects - Restricted Dataset				
Variable	No Injury	Possible Injury	Non-incap. Injury	Incap. Injury
Constant	0.0690	-0.0087	-0.0391	-0.0194
Speed Limit on MN	0.0000	0.0000	0.0000	0.0000
Brevard Co.	-0.3048	0.0384	0.1728	0.0856
City of Orlando	-0.2024	0.0255	0.1148	0.0569
Hillsborough Co.	-0.1908	0.0240	0.1082	0.0536

Marginal Effects - Complete Dataset				
Variable	No Injury	Possible Injury	Non-incap. Injury	Incap. Injury
Constant	-0.0323	0.0001	0.0118	0.0158
Major No. of Lanes	0.0038	0.0000	-0.0014	-0.0019
MJ Left Turn Lanes	-0.0086	0.0000	0.0032	0.0042
RT Channel. On MJ	0.0294	-0.0001	-0.0108	-0.0144
Division on MN	0.0026	0.0000	-0.0010	-0.0013
Speed Limit on MN	0.0000	0.0000	0.0000	0.0000
ADT on MJ	0.0000	0.0000	0.0000	0.0000
Brevard Co.	-0.1874	0.0009	0.0687	0.0918
City of Orlando	-0.2014	0.0009	0.0738	0.0987
Hillsborough Co.	-0.0809	0.0004	0.0296	0.0396

In addition to comparing models based on restricted and complete datasets, these ordered probit models were also compared to injury models created through hierarchical tree-based regression. A total of six tree-based regressions were run for the restricted dataset: one for each type of injury in the complete dataset with the same independent variables as used in the ordered probit models. For the non-injury crashes, two regressions were run because non-injury crashes can be reported on both long and short forms. For all other injury levels, only one regression was run because the other levels have complete data.

Similar to Chapter 5, the tree-based regression output was in two forms: a tree-diagram and a list of important variables. Figure 6-2 is an example of the tree diagrams created for crash severity level. In particular, this diagram is for predicting the number of crashes involving a possible injury. Other tree diagrams created for severity levels are in Appendix D. Table 6-7 shows the relative importance of each intersection characteristic variable for the six tree-based regression models based on injury severity arranged in decreasing order for the important variables in the complete dataset for non-injury crashes. Again, the other injury levels are not

divided into complete and restricted datasets since these injury levels do not include any crashes reported on short forms.

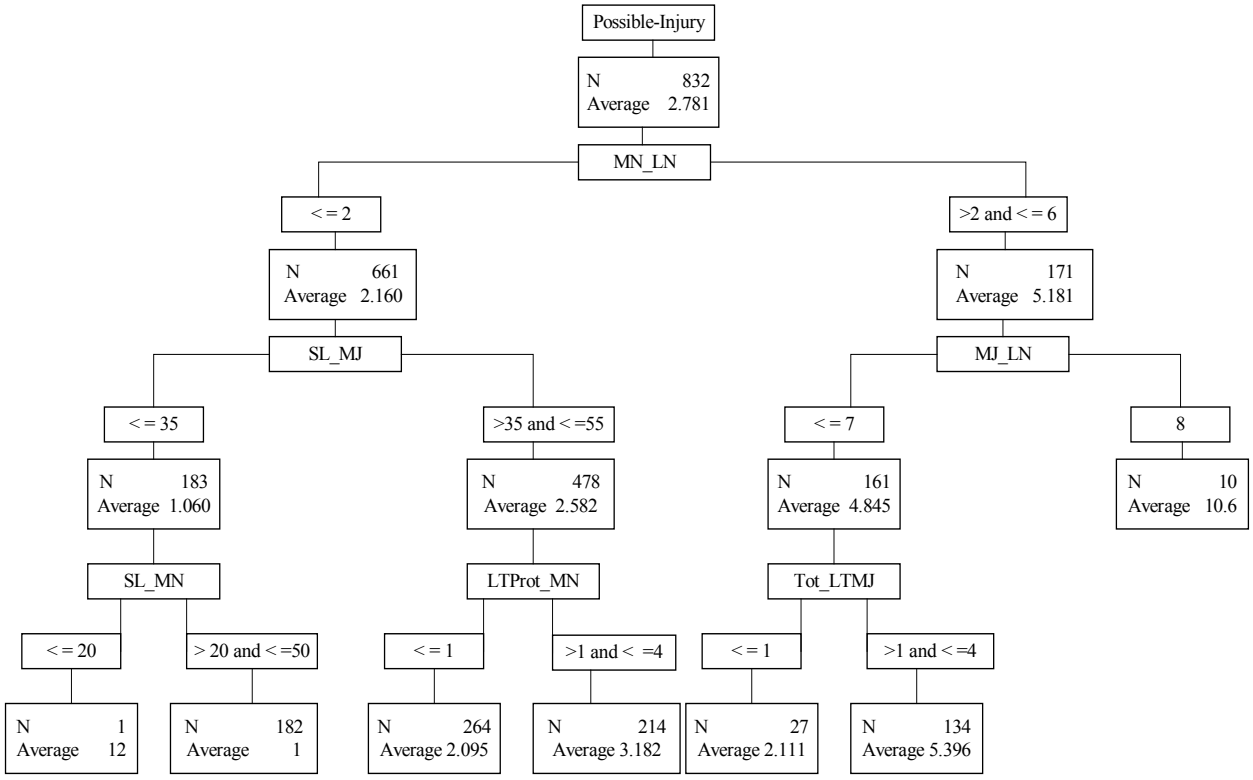

Figure 6-2. Regression Tree for the Expected Number of Possible-Injury Crashes Per Intersection for Two Years

Table 6-7. Relative Importance of Factors for Severity Models Based on Intersection Characteristics

Variables	No Injury		Possible Injury	Non-Incap. Injury	Incap. Injury	Fatal Injury
	Compl.	Restr'd				
Daily Traffic Volume on MJ	1.0000	1.0000	0.6042	0.1022	0.7861	0.0000
Speed Limit on MJ	0.5296	0.6133	0.5259	0.0000	0.4493	0.0000
Median Present on MJ	0.5290	0.5000	0.0000	0.3132	0.3735	1.0000
Exclusive Left Turn Lanes on MN	0.3856	0.6240	0.4134	1.0000	0.8114	0.0000
Number of Lanes on MN	0.3225	0.5318	1.0000	0.1567	0.0000	0.0000
Speed Limit on MN	0.1520	0.2756	0.3115	0.2204	0.3685	0.0000
Exclusive Left Turn Lanes on MJ	0.1457	0.3018	0.0000	0.0000	0.0000	0.4858
Daily Traffic Volume on MN	0.1266	0.5961	0.1976	0.3916	1.0000	0.0000
Number of Lanes on MJ	0.0853	0.5270	0.5512	0.4867	0.3417	0.7687
Right Turns Channelized on MJ	0.0000	0.0166	0.0000	0.6193	0.2076	0.0000
Median Present on MN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Right Turns Channelized on MN	0.0000	0.1016	0.0000	0.1526	0.1044	0.0000
Total Left Turn Lanes on MJ	0.0000	0.0000	0.4422	0.0000	0.0000	0.0000
Total Left Turn Lanes on MN	0.0000	0.1800	0.0000	0.0000	0.0000	0.0000

Table 6-7 shows that the most important factor in predicting a non-injury crash is the daily traffic volume on the major road, which was found to be insignificant in predicting fatal crashes. Quite contrary, in the ordered probit models in Table 6-4, the traffic volume on the major road was only significant in the complete model and it predicted higher severity levels for higher volumes. One similarity between the tree-based regression and the ordered probit complete model was that speed limits on the minor road significantly affected lower injury severity levels. The presence of a median caused lower crash severity levels in the complete ordered probit model but was found to be insignificant in the tree-based regression models. Only three variables (presence of a median on the major road, number of lanes on the major road and number of exclusive left turn lanes on the major road) were found to be important in the tree-based regression model for the number of fatalities. The complete ordered probit model did not show the presence of a median on the major road to be significant, however, the presence of a median on the minor road was found to lower the injury level. Finally, right turn channelization on the minor road was found to cause lower injury crashes in the complete ordered probit models but was essentially insignificant in the non-injury and possible injury models created from tree-based regressions.

In an effort to validate the tree-based regression models created severity levels, these results were used to predict the number injury and non-injury crashes for the year 2002 in Brevard County and City of Orlando. The predicted totals are presented in Table 6-8 where it can be seen that tree-based regression accurately predicts (1776 / 2057) = 86.3% of the total number of crashes.

Table 6-8. Tree-Based Regression Predictions for Severity Level in Brevard County and City of Orlando for Year 2002

	Actual Count	Predicted	Predicted Correctly
No Injury	549	581	549
Possible Injury	718	540	540
Non-incap. Injury	725	640	640
Incapacitating Injury	59	41	41
Fatal Injury	6	10	6
Total	2057	1812	1776

Predictions for the ordered probit models were based on the same data that was used to train the models. Predictions for the tree-based regression models were based on data outside the training set for year 2002 in Brevard County and City of Orlando. It might be assumed that the ordered probit predictions would be more accurate because they are predicting values from the training dataset, however, it is interesting to note that the tree-based regression has a higher percentage of correct predictions. The ordered probit model based on the restricted dataset with results displayed in Table 6-2 was only able to predict the actual injury severity level for 41.1% of the crashes. One reason that the tree-based regression model performs better is that it was trained with more data since missing values were treated as acceptable input, whereas missing values in the ordered probit model caused the entire observation to be ignored.

6.4 Severity Model Based on Combined Variables

Due to the fact that the models based on the complete dataset consistently had a higher prediction rate, it was hypothesized that a model created with both geometric and crash type variables for the complete dataset may have even higher prediction power. The variables available for this model were shown in Table 4-1 and backward selection was used to determine the most significant variables at a 5% level. Table 6-9 shows the final ordered probit model.

Table 6-9. Final Ordered Probit Model Based on the Complete Dataset and All Possible Variables

Complete Dataset			
Variable	**Coefficient**	**t-Statistic**	**P-Value**
Constant	-1.4780	-42.246	0.0000
Angle	0.5414	15.249	0.0000
Head-on	0.4859	5.310	0.0000
Left Turn	0.7590	20.175	0.0000
Ped/Bike	1.4757	20.352	0.0000
Rear-end	0.3150	10.062	0.0000
Division on MN	-0.0891	-3.731	0.0002
Speed Limit on MN	-0.0001	-6.728	0.0000
Brevard Co.	1.2950	25.224	0.0000
City of Orlando	1.3367	34.216	0.0000
Hillsborough Co.	0.6615	25.189	0.0000
α_1	0.6472	53.430	0.0000
α_2	1.5576	71.915	0.0000
α_3	2.6848	40.991	0.0000
Sample Size		21204	
Degrees of Freedom		10	
Chi-squared		7578.639	
Log Likelihood Function		-11822.72	
Restricted Log Likelihood		-15612.04	

The results indicate that crashes involving pedestrians or bicyclists have the greatest risk of a severe injury. Of the motor vehicle crashes, left turn, angle and head-on were most likely to result in a higher level of injury. Right turn and sideswipe crashes were found to be insignificant in the determination of severity and were dropped from the final model. With the exception of a median and the speed limit on the minor road, all other intersection characteristics were found to be insignificant at the 5% level. Specifically, the presence of a median on the minor road and a higher speed limit were found to lower the risk of a serious injury. Finally, the City of Orlando was found to have a higher expected level of injury for crashes at signalized intersections amongst Brevard County and Hillsborough County.

Table 6-10 shows the number of actual and predicted severity levels for the final model. With a prediction rate of 79.1%, the final model proved to have the best prediction power of the ordered probit models. Table 6-11 shows the marginal effects for the final model.

Table 6-10. Predicted Crash Severity Levels for Final Ordered Probit Model

Complete Dataset						
	Predicted					
Actual	0	1	2	3	4	Total
0	16506	0	88	0	0	16594
1	2124	0	179	11	0	2314
2	1537	0	253	6	0	1796
3	427	0	39	5	0	471
4	19	0	9	1	0	29
Total	20613	0	568	23	0	21204

Table 6-11. Marginal Effects for the Final Ordered Probit Model

Marginal Effects - Complete Dataset				
Variable	No Injury	Possible Injury	Non-incap. Injury	Incap. Injury
Constant	0.0338	-0.0280	-0.0056	-0.0002
Angle	-0.0124	0.0103	0.0021	0.0001
Head-on	-0.0111	0.0092	0.0018	0.0001
Left Turn	-0.0174	0.0144	0.0029	0.0001
Ped/Bike	-0.0338	0.0279	0.0056	0.0002
Rear-end	-0.0072	0.0060	0.0012	0.0001
Right Turn	0.0020	-0.0017	-0.0003	0.0000
Sideswipe	0.0000	0.0000	0.0000	0.0000
Brevard Co.	-0.0296	0.0245	0.0049	0.0002
City of Orlando	-0.0306	0.0253	0.0051	0.0002
Hillsborough Co.	-0.0151	0.0125	0.0025	0.0001

6.5 Summary

Crash severity level is an ordered variable and needs to be treated as such when used in statistical models. Therefore, ordered probit models were created in this section for three different types of models; one based on collision types, another for intersection characteristics and the last for a combination of significant variables. Both the restricted and complete datasets were used to create these models and the output was compared. Similar to the results in the previous chapter, it was determined that the models based upon the complete dataset were more accurate. However, when compared to the tree-based regression results, the ordered probit model did not predict as well for the restricted dataset based on intersection characteristics.

The final ordered probit model based on the complete dataset included ten significant variables and a constant term. The first five variables referred to the type of crash, the next two dealt with intersection characteristics on the minor roadway and the last three accounted for

county-specific factors. Division on the minor road, as well as a higher speed limit on the minor road, was found to lower the expected injury level. A median on the minor road may prevent more head-on crashes, which were found to be more severe crashes. A higher speed limit on the minor road may cause the differential speeds between vehicles on intersecting roads to be smaller, likely resulting in a decrease in the crash severity level. In this analysis, the vehicle crash type that has the highest risk for a severe injury is the left turn crash, followed by angle and head-on crashes. The fact that crash types were found to be significant for predicting injuries is similar to results of O'Donnell and Connor (1996) whose ordered probit models also showed that collision type to have a significant effect on crash injury levels.

CHAPTER 7. CONCLUSION

Approximately 4.8 times more crashes occur at signalized intersections than at intersections with other control types according to research conducted by Bhesania (1991). As such, it is important to quantify the effects that different intersection characteristics have on crashes in an effort to improve the level of safety. The main objective of this research was to determine if there is a difference in the significance of crash-causing factors for different collision types and injury levels. The second objective of this paper was to identify the consequences of modeling an incomplete dataset. An extensive literature review was conducted to gain knowledge of past research in this field and it was found that much research has focused on modeling crash data by the Poisson or the negative binomial distribution. Other types of analyses included frequency models, the Empirical Bayes method, linear regression, comparison groups, tree-based regressions and ordered probit models, where each author noted the benefits of their method chosen.

The data collection period for this analysis was very time-consuming because it was essential to obtain records for every crash that occurred to ensure the completeness of the data. The dataset used in this analysis not only accounted for crashes reported on long forms, but also included minor crashes that are unreported to state agencies. A total of four counties/city, which make up a significant portion of Central Florida, were contacted for cooperation with this research: Seminole County, City of Orlando, Hillsborough County and Brevard County. Information was collected for a total of 832 intersections and over 33,500 crashes. Crash information for these intersections was obtained from a combination of three sources; county databases, the FDOT database and the DHSMV database. The years 2000-2001 were used for

analysis because these were the only two consistent years throughout the counties/city. Due to the abundance of data collected, a portion was used as a validation set for the tree-based regression models to confirm the accuracy of models created.

The methods chosen to analyze the data collected for this project were hierarchical tree-based regression (HTBR) and ordered probit regression. HTBR is non-parametric, robust against multicollinearity between the variables and capable of handling missing observations. Ultimately, HTBR provided a direct and clear-cut method of analyzing this crash data. Ordered probit models were used to predict crash severity levels due to the ordinal nature of this dependent variable. These models were then compared to the HTBR models to determine if both models found the same variables to be significant. Finally, both types of models were used to predict the expected number of crashes and comparisons between the models were made based on their percentages of correct predictions.

Tree-based regression was used in Chapter 5 to consider the difference in the relative importance of each variable between the different types of collisions. In this section, only the regressions based upon crashes reported on long forms were taken into account to make the results more comparable to other studies. Table 5-2 showed the results from this analysis. It was found that the most important factor for determining the total number of crashes in the restricted dataset was the number of lanes on the minor road. This factor was also most significant in determining the number of angle crashes. However, it dropped in significance for determining the number of left turn crashes in the restricted dataset and was found to be insignificant in all other models. The number of exclusive left turn lanes was found to be significant in all models except those for predicting the number of right turn and sideswipe crashes in the restricted dataset. The total number of left turn lanes on the major road was found to be insignificant in all

models based on the restricted dataset. The total number of left turn lanes on the minor road was significant in predicting the number of left turn, head-on and pedestrian/bicyclist crashes in the restricted dataset. The traffic volume on the major road was significant in all models except for angle and left turn crash models. Speed limits on the major and/or minor roads were found to be important factors in all models for the restricted dataset except in those for right turn and sideswipe crashes. Channelization of right turn lanes on the minor road was only found to be significant for the number of left turn crashes while channelization on the major was important for the total number of crashes as well as pedestrian/bicyclist, rear-end and sideswipe crashes in the restricted dataset. The main finding was that the models created for angle and left turn crashes change the most from the model created from the total number of crashes reported on long forms. This result shows that aggregating the different crash types by only estimating models based on the total number of crashes will not predict the number of expected crashes as accurately as models based upon each type of crash individually.

To investigate the differences between modeling restricted and complete datasets, the relative importance of each factor was noted each of the tree-based regressions. Table 5-3 showed the relative importance for each type of crash when the different datasets were used to train the model. By organizing the relative importance values in an array, it became clear that there is consistently a difference between models based on the restricted and complete datasets. For rear-end, right turn and sideswipe crashes, the important factors are fairly consistent between the models created by complete and restricted datasets. These results show that factors in crashes causing mostly injuries or involving felony crimes for rear-end, right turn and sideswipe crashes are generally the same as factors causing non-injury or minor crashes. This indicates that models based on complete and restricted datasets for these types of crashes are roughly

equivalent. On the other hand, important factors for angle and head-on crashes changed the most between the models because these types of crashes are unstable and different factors result in non-injury crashes. In other words, the crash-causing factors were found to be significantly different when crashes reported on short forms where added to the dataset for angle and head-on crashes. The results in this section show that it is more important to include minor crashes in the dataset when modeling the number of angle or head-on crashes and less important to include minor crashes when modeling rear-end, right turn or sideswipe crashes.

Ordered probit and tree-based regression models were used in Chapter 6 to predict crash severity levels for three different types of models; the first one based on collision type, the second one based on intersection characteristics and the last one based on a significant combination of both models. Both the restricted and complete datasets were used to create the first two model types and the output was compared. Similar to the results in Chapter 5, it was determined that the models based upon the complete dataset were more accurate. However, when compared to the tree-based regression results, the ordered probit model did not predict as well for the restricted dataset based on intersection characteristics. The final ordered probit model based on the complete dataset included ten significant variables and a constant term. This model showed that crashes involving a pedestrian/bicyclist have the highest probability of a severe injury. For motor vehicle crashes, left turn, angle, head-on and rear-end crashes cause higher injury severity levels. Division on the minor road, as well as a higher speed limit on the minor road, was found to lower the expected injury level.

This research has shed light on several important topics in crash modeling. First of all, this research demonstrated that variables found to be significant in aggregated crash models may not be the same as the significant variables found in models based on individual crash types.

Furthermore, variables found to be significant in crash type models typically changed when minor crashes were added to complete the dataset. Thirdly, ordered probit models based on significant crash-type and intersection characteristic variables (instead of one variable set or the other) had better crash severity prediction power, especially when based on the complete dataset. Lastly, upon comparison between tree-based regression and ordered probit models, it was found that the tree-based regression models predicted the crash severity levels better. In conclusion, the results of this research showed it is imperative that models are developed for each type of collision and injury level instead of aggregating crash types to predict the total number of crashes. Furthermore, results showed that crashes reported on short forms are important when modeling the number of expected crashes and, contrary to current practice, should be documented in every crash database maintained by a state agency.

APPENDIX A

DEPENDENT-VARIABLE DISTRIBUTION GRAPHS

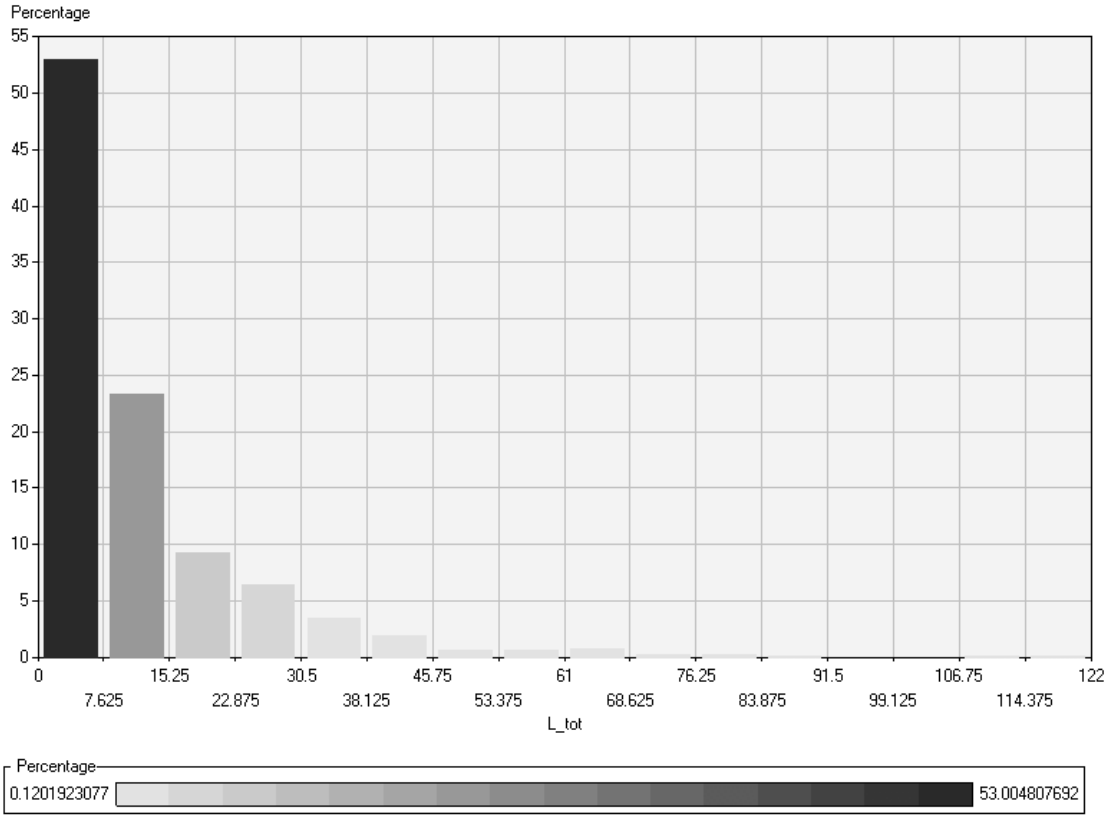

Figure A-1. Distribution of the Total Number of Crashes Reported on Long Forms

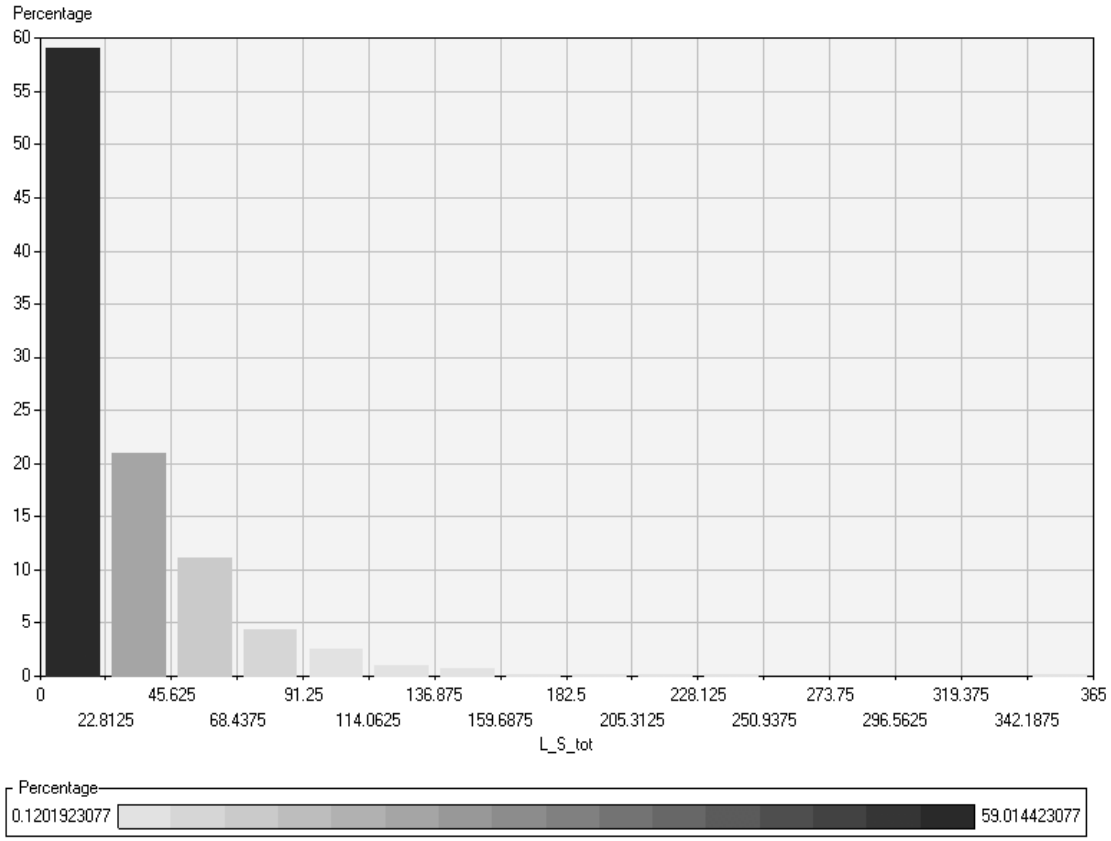

Figure A-2. Distribution of the Total Number of Crashes Reported on Long and Short Forms

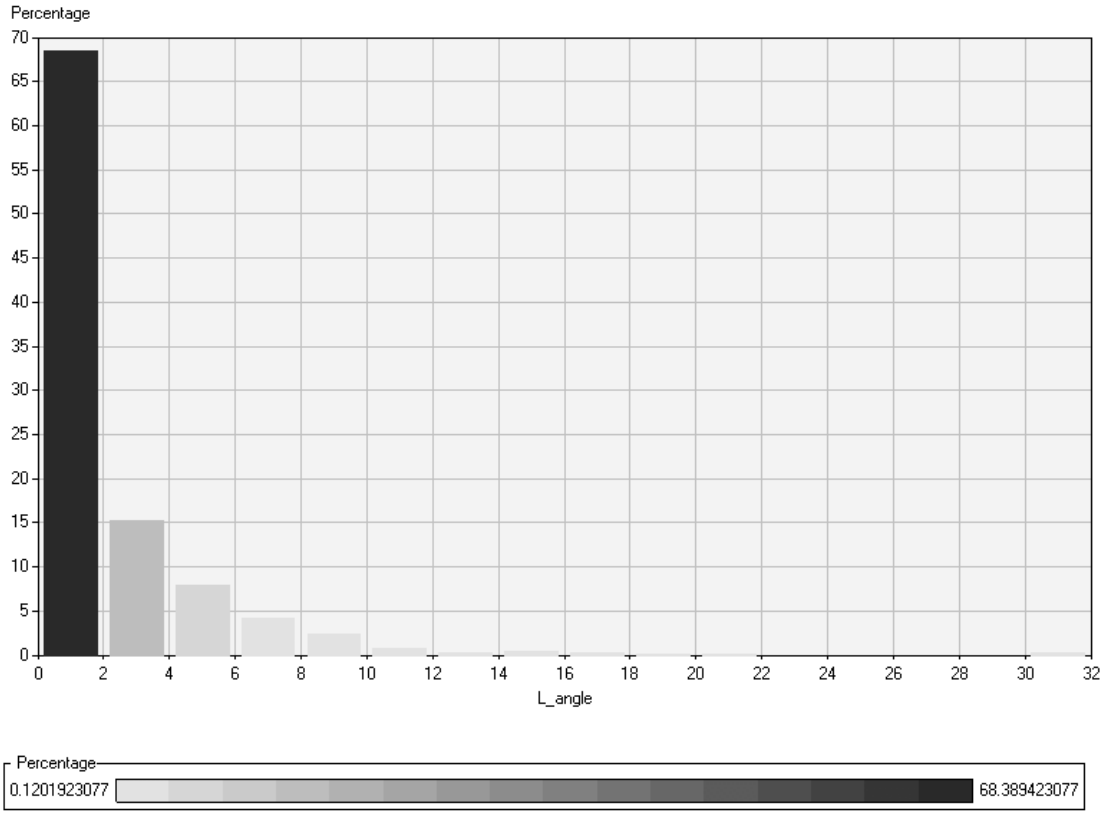

Figure A-3. Distribution of Angle Crashes Reported on Long Forms

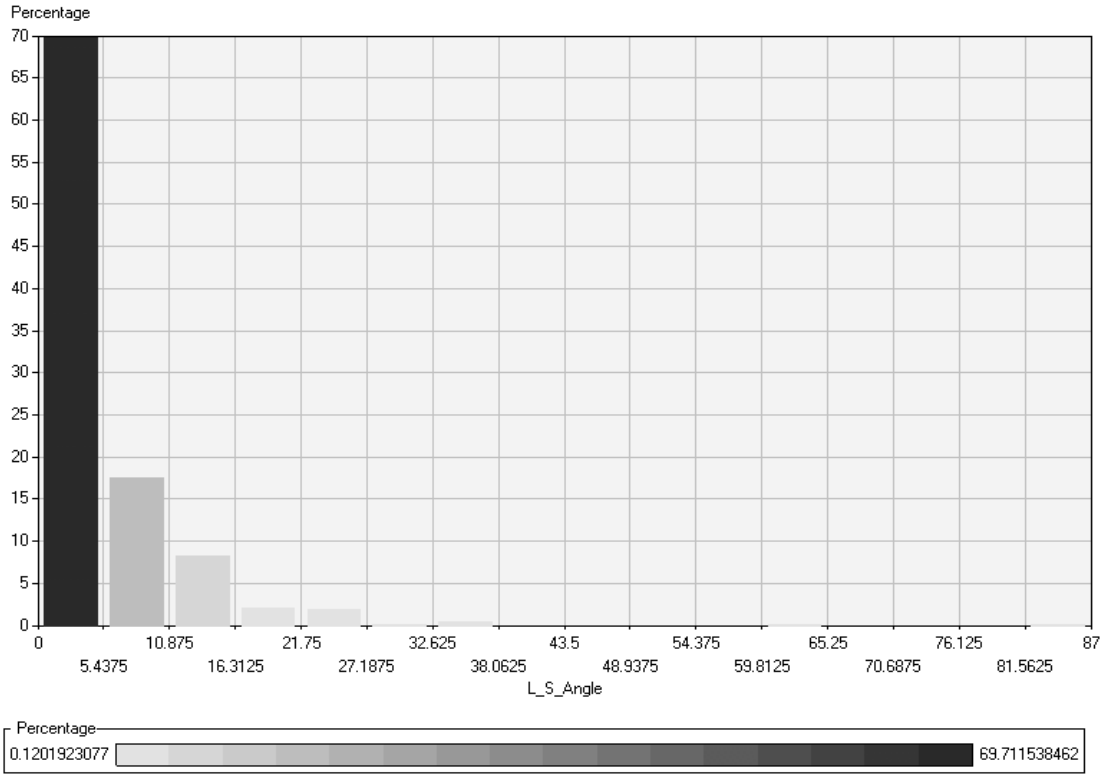

Figure A-4. Distribution of Angle Crashes Reported on Long and Short Forms

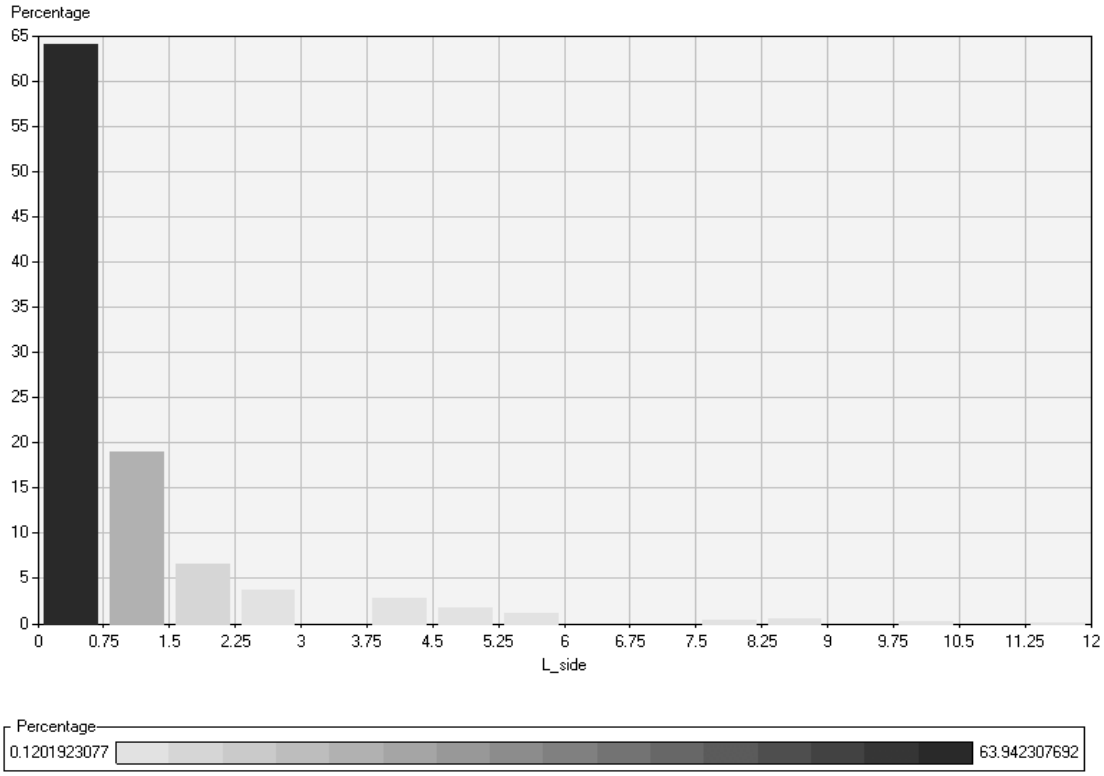

Figure A-5. Distribution of Sideswipe Crashes Reported on Long Forms

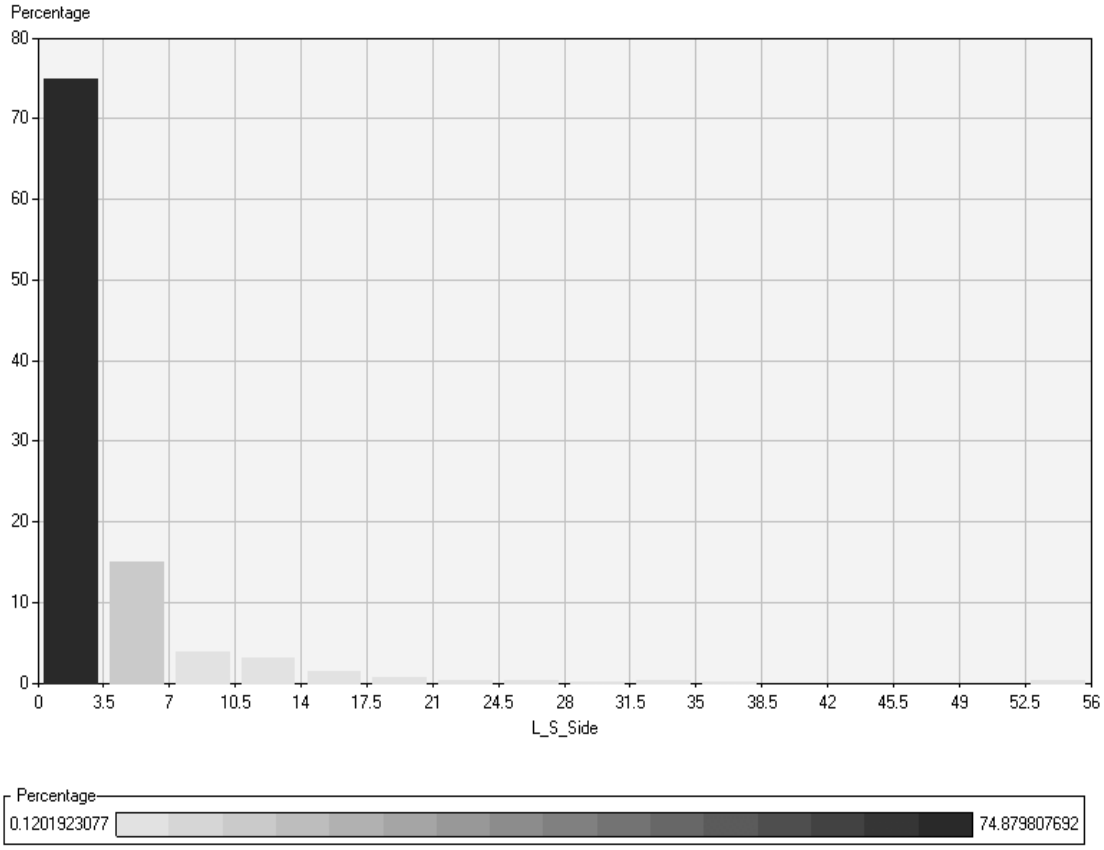

Figure A-6. Distribution of Sideswipe Crashes Reported on Long and Short Forms

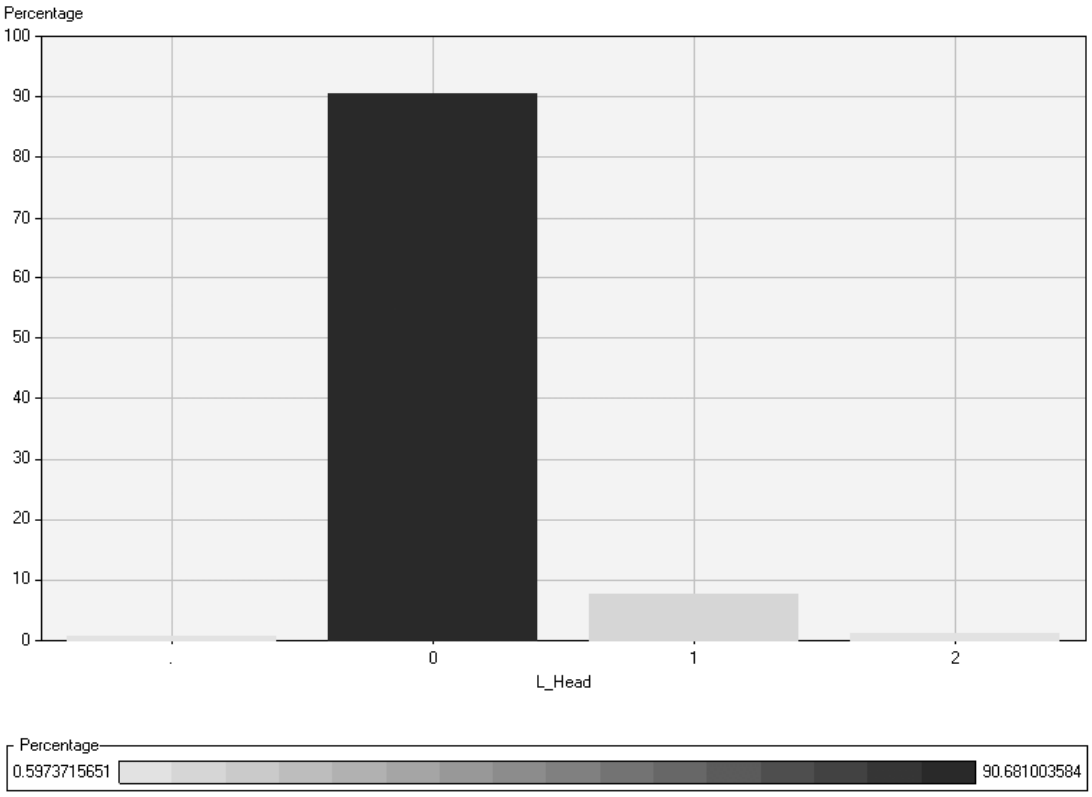

Figure A-7. Distribution of Head-on Crashes Reported on Long Forms

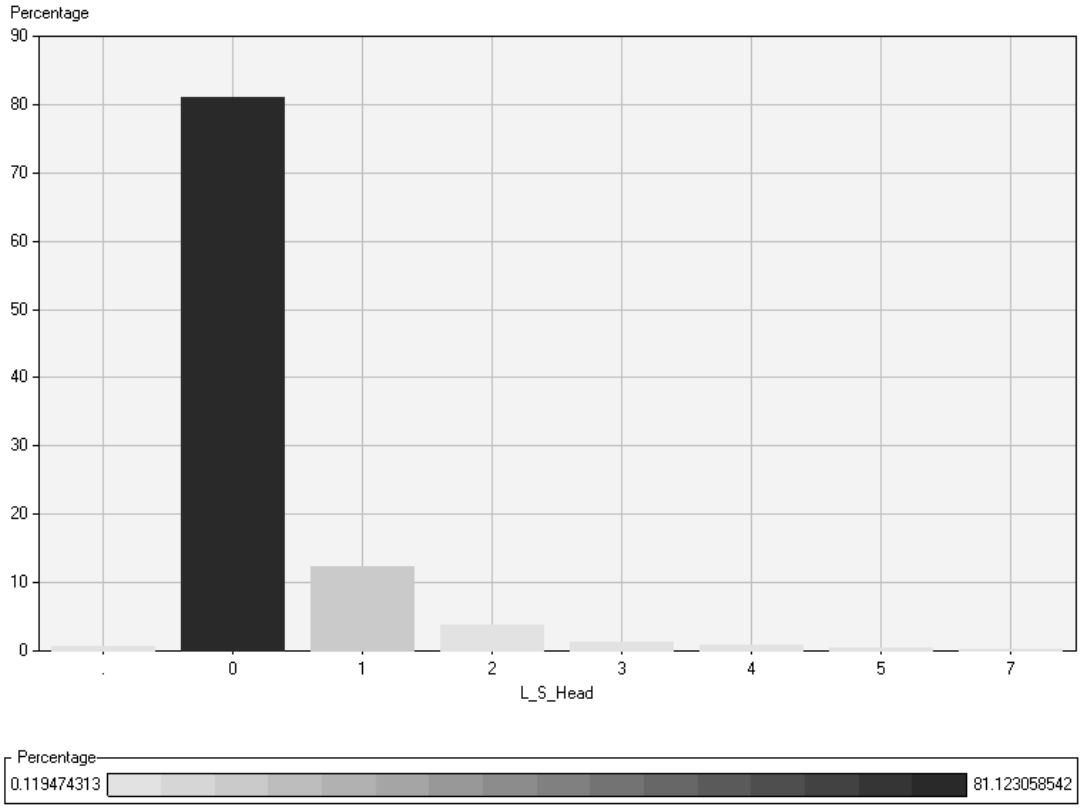

Figure A-8. Distribution of Head-on Crashes Reported on Long and Short Forms

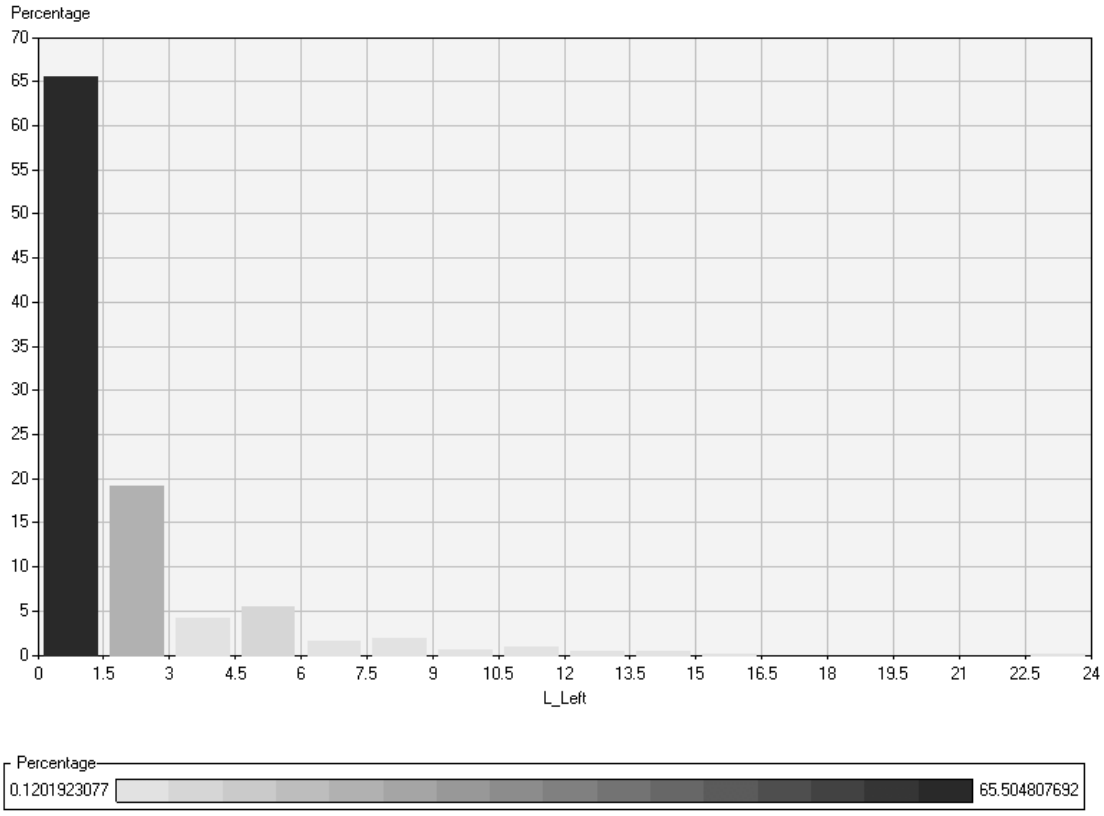

Figure A-9. Distribution of Left Turn Crashes Reported on Long Forms

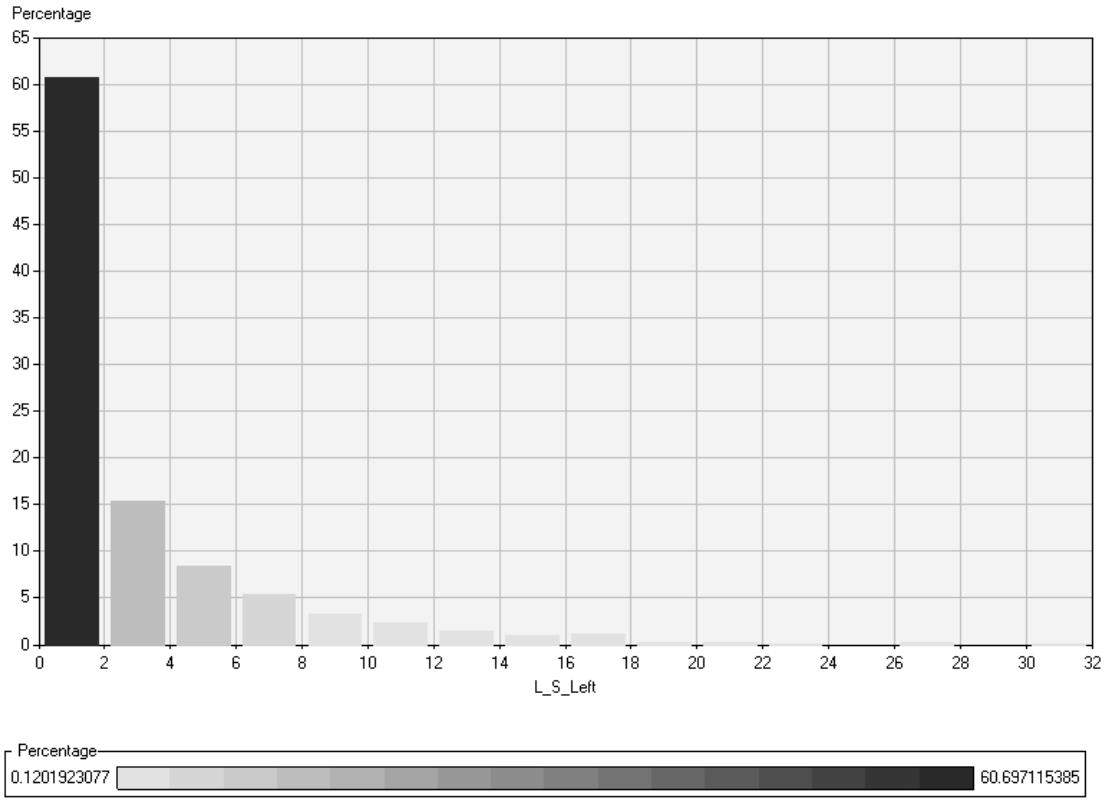

Figure A-10. Distribution of Left Turn Crashes Reported on Long and Short Forms

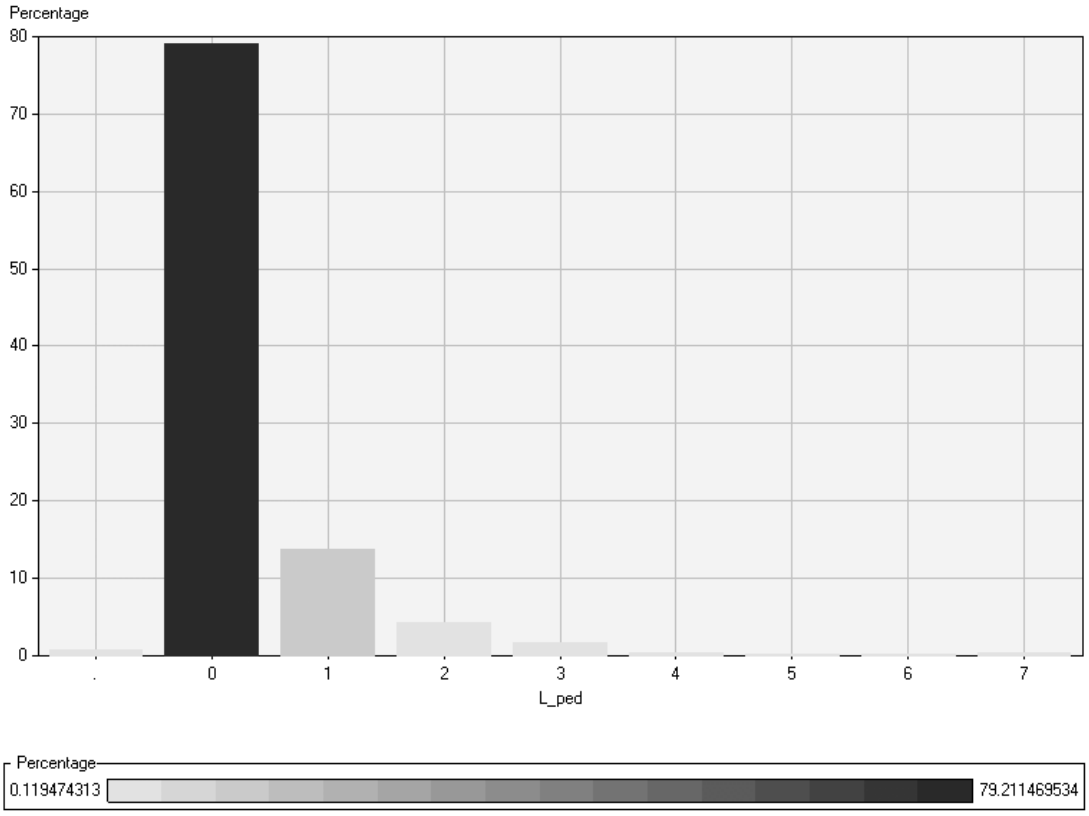

Figure A-11. Distribution of Pedestrian/Bicycle Crashes Reported on Long Forms

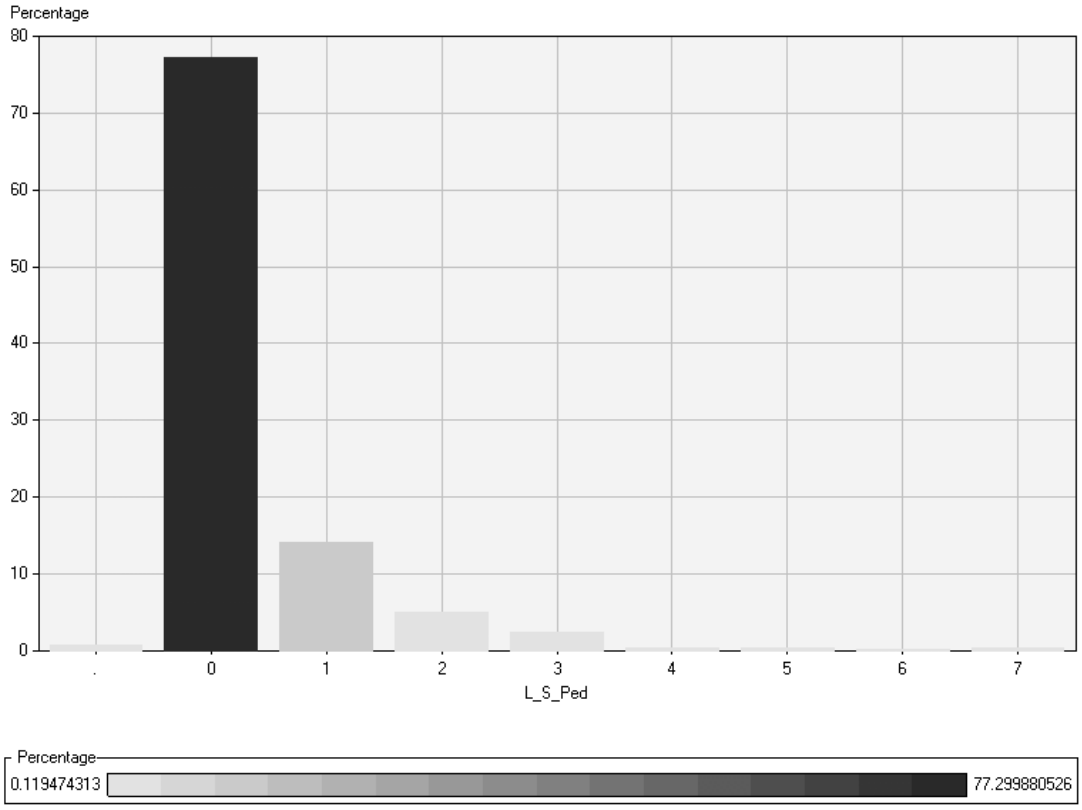

Figure A-12. Distribution of Pedestrian/Bicycle Crashes Reported on Long and Short Forms

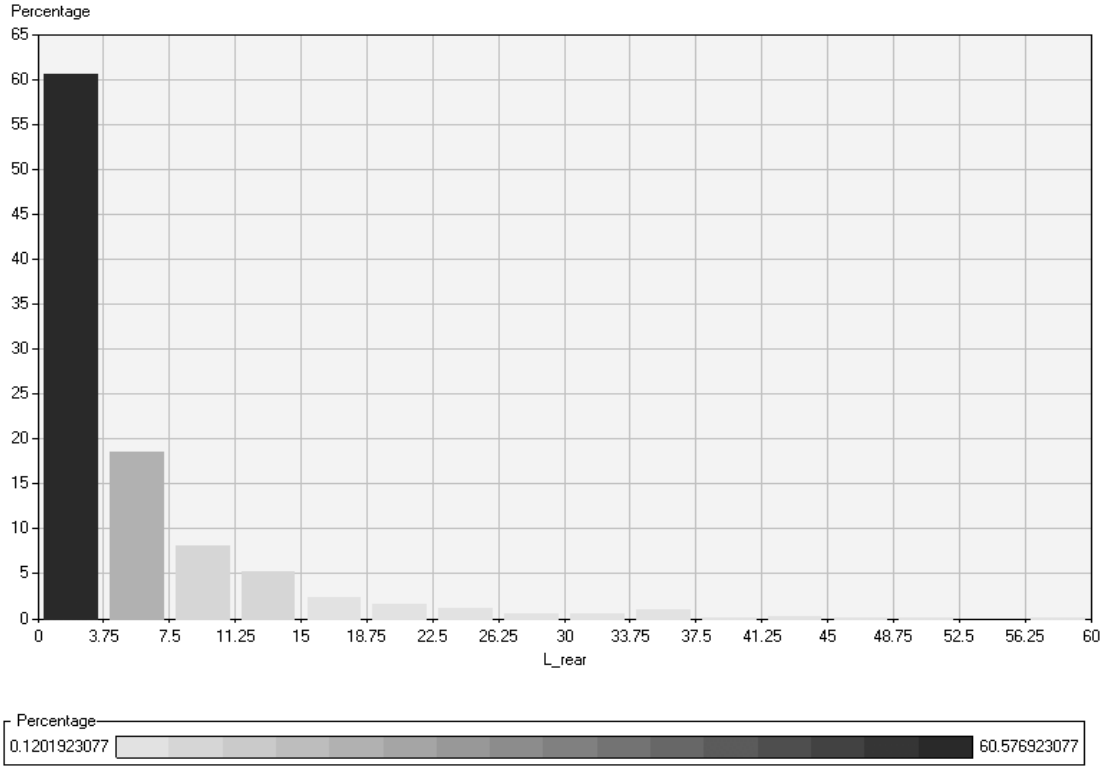

Figure A-13. Distribution of Rear-end Crashes Reported on Long Forms

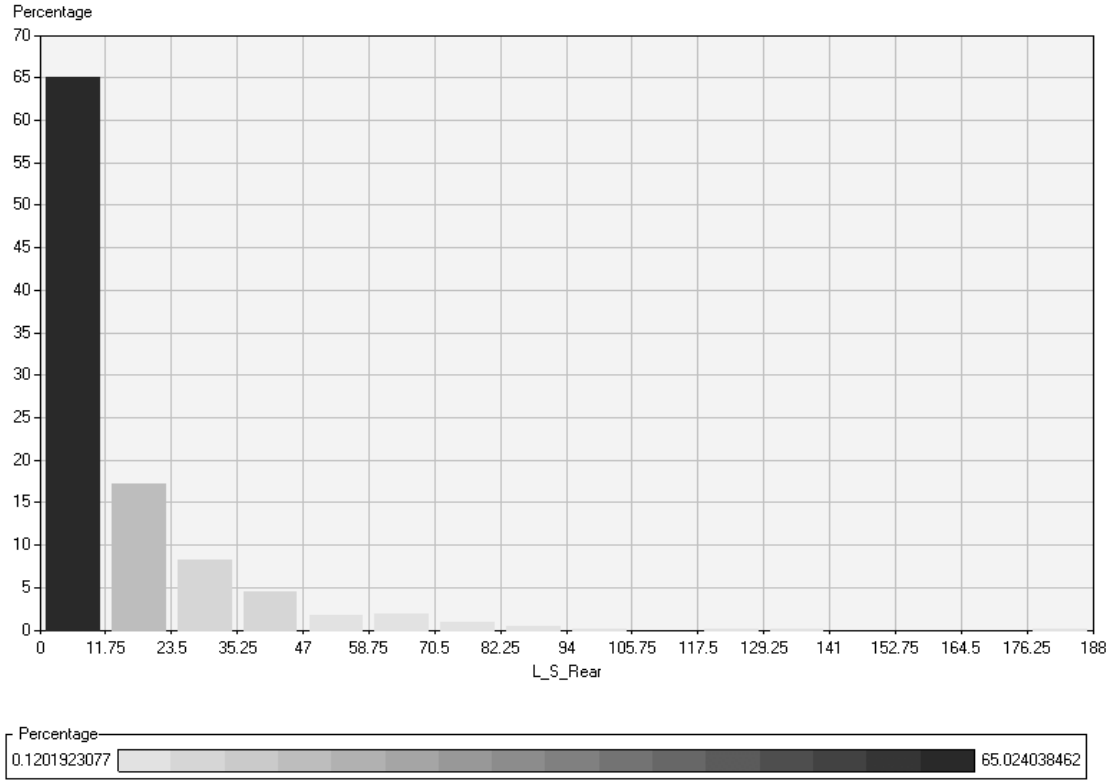

Figure A-14. Distribution of Rear-end Crashes Reported on Long and Short Forms

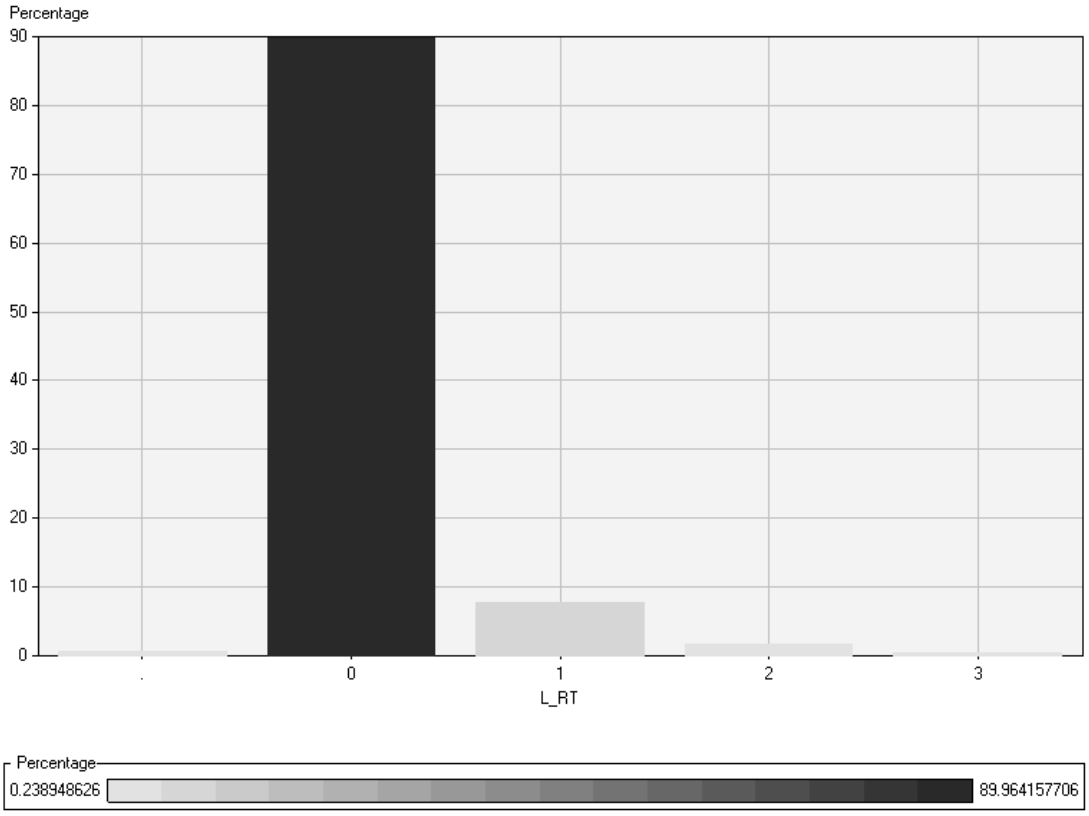

Figure A-15. Distribution of Right Turn Crashes Reported on Long Forms

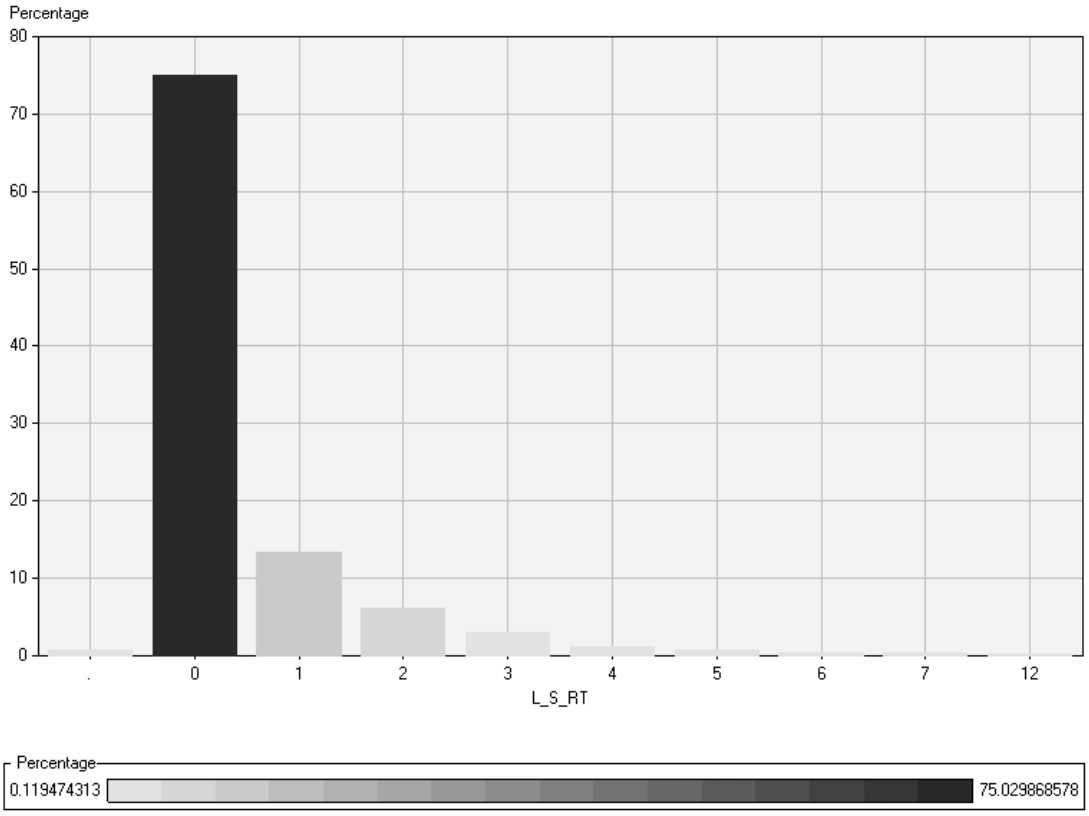

Figure A-16. Distribution of Right Turn Crashes Reported on Long and Short Forms

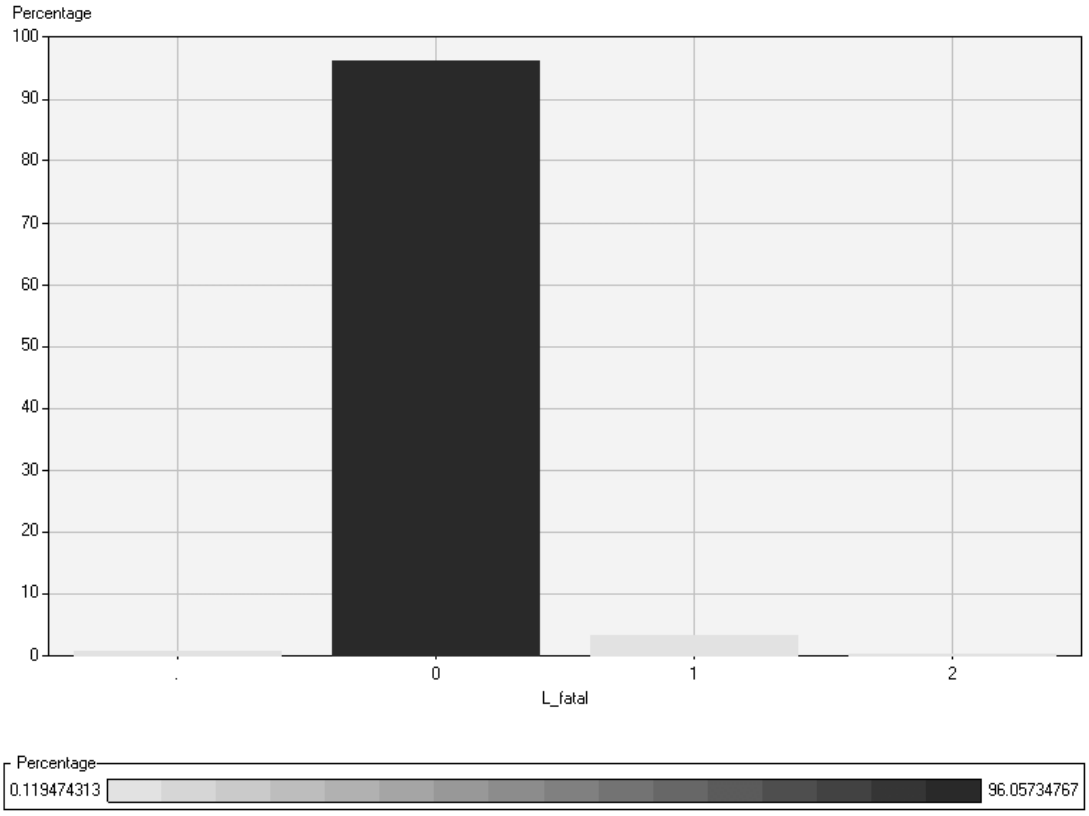

Figure A-17. Distribution of Fatal Crashes

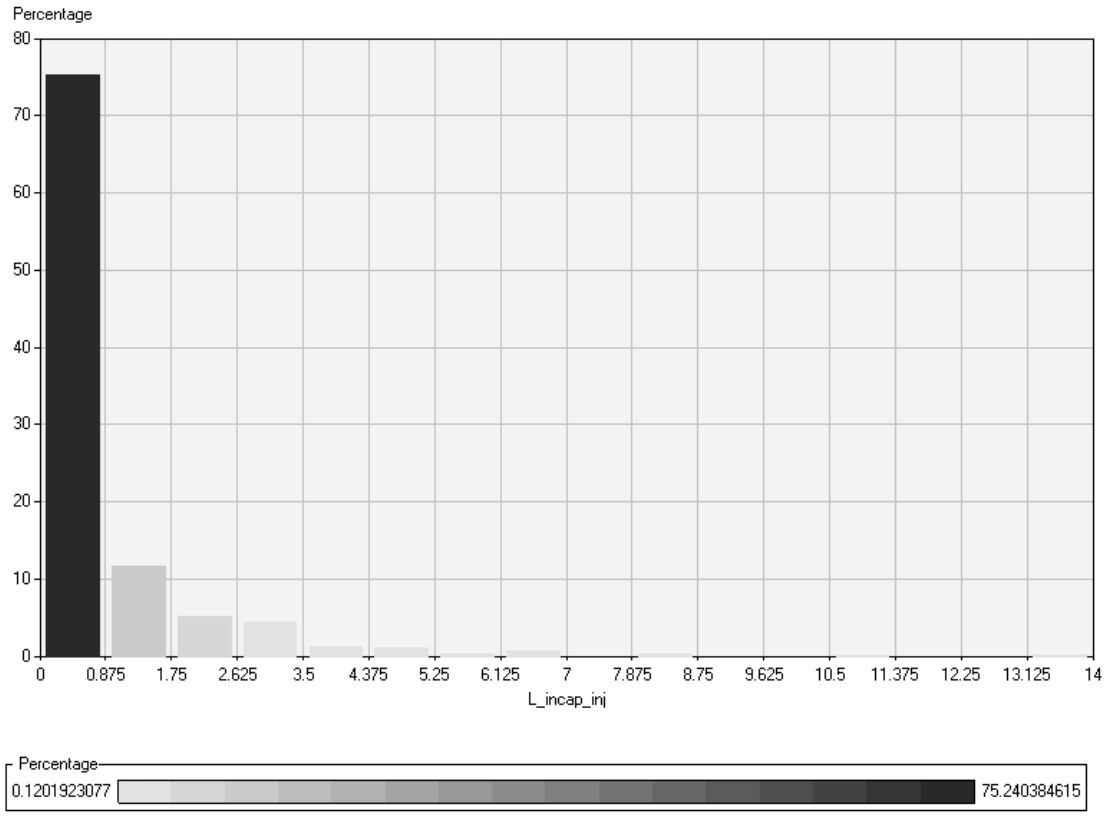

Figure A-18. Distribution of Incapacitating Injury Crashes

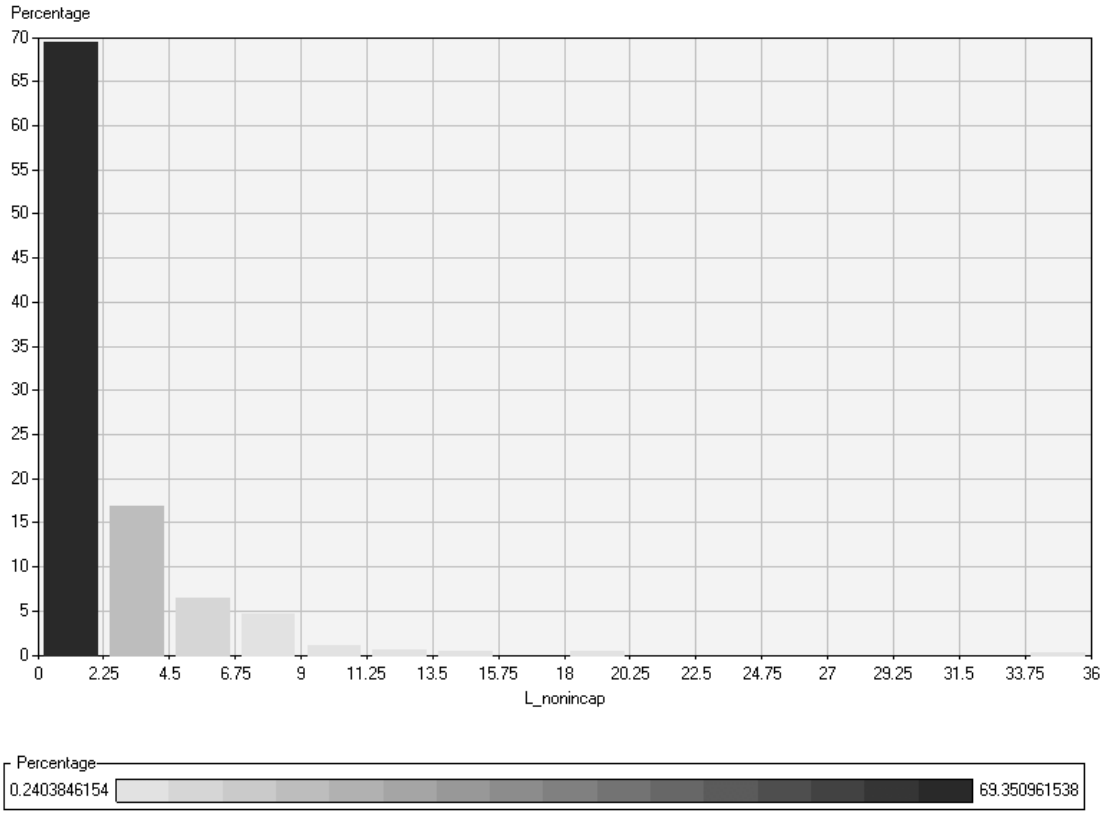

Figure A-19. Distribution of Non-incapacitating Crashes

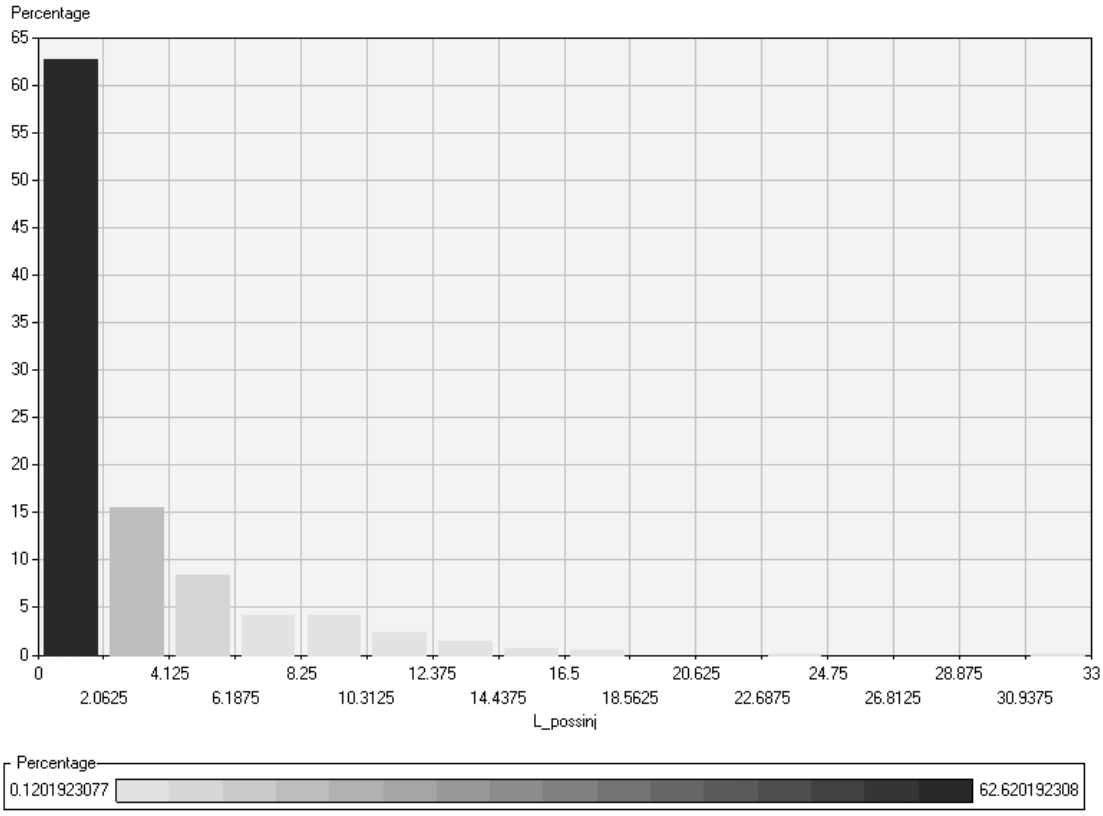

Figure A-20. Distribution of Possible Injury Crashes

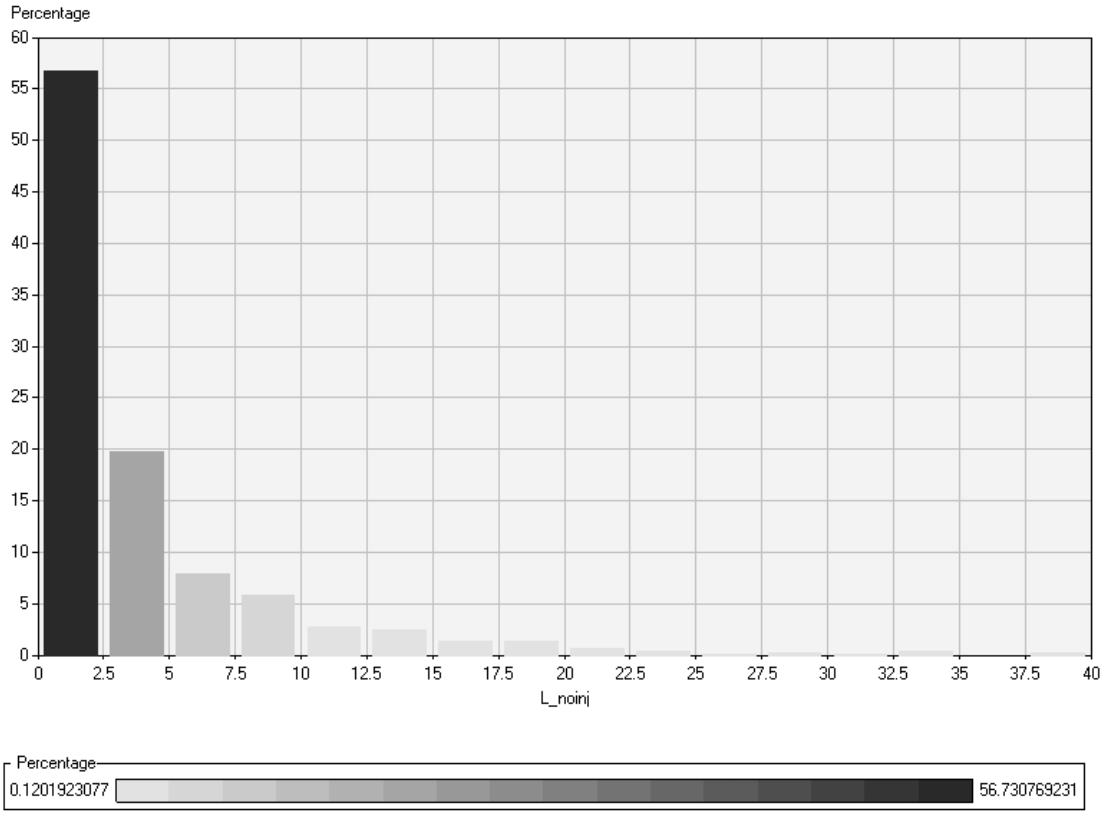

Figure A-21. Distribution of No-Injury Crashes Reported on Long Forms

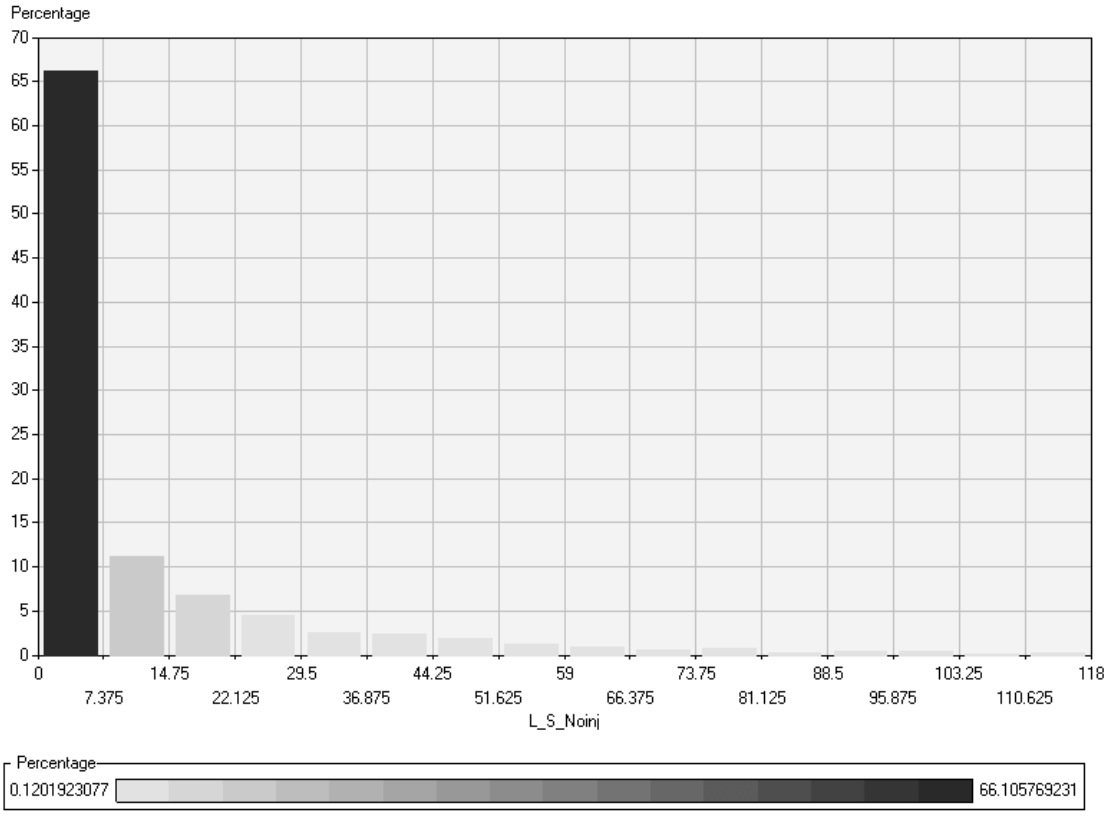

Figure A-22. Distribution of No-Injury Crashes Reported on Long and Short Forms

APPENDIX B

REGRESSION TREES FOR CRASH TYPES

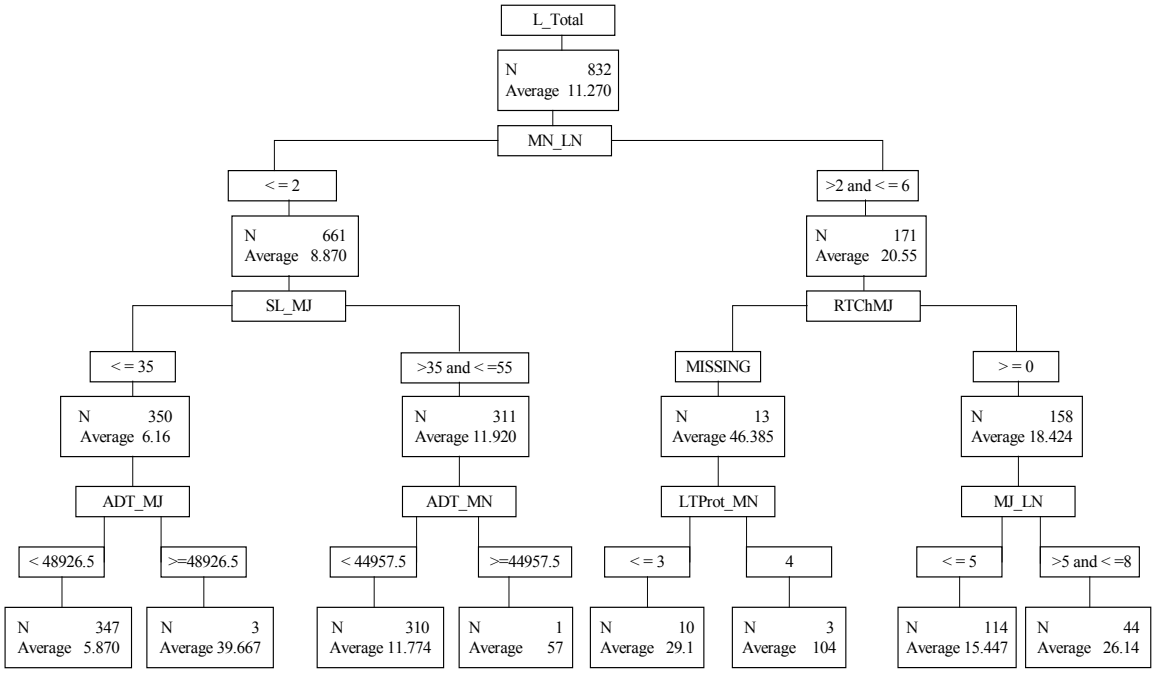

Figure B-1. Regression Tree for the Expected Total Number of Crashes Reported on Long Forms Per Intersection for Two Years

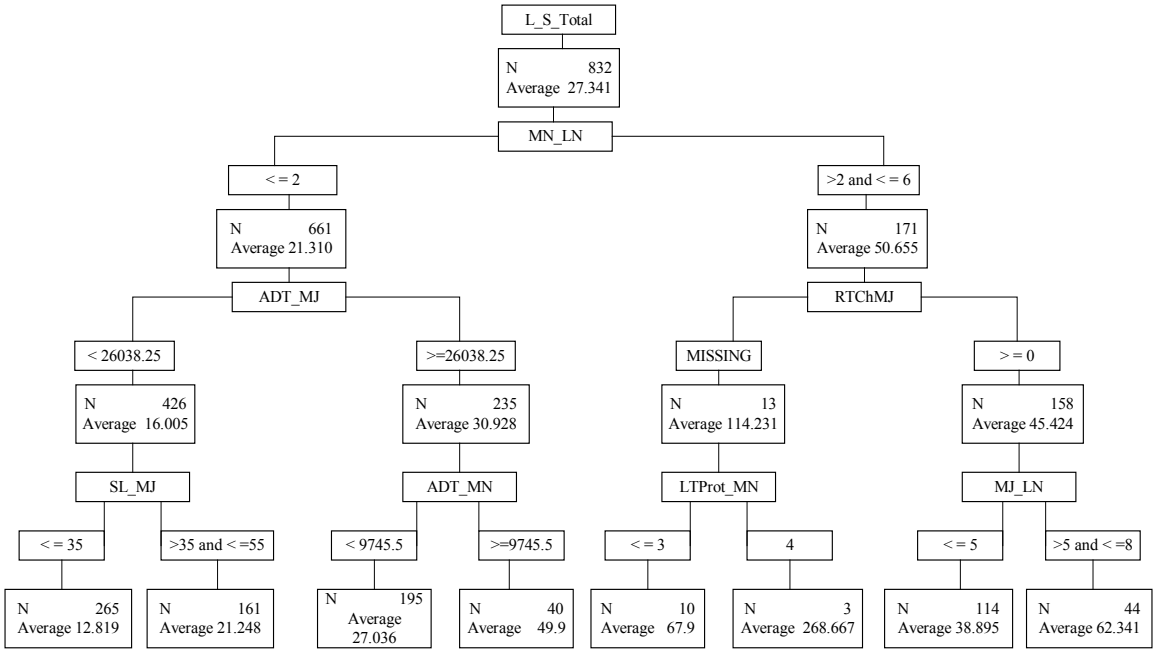

Figure B-2. Regression Tree for the Expected Total Number of Crashes Reported on Long and Short Forms Per Intersection for Two Years

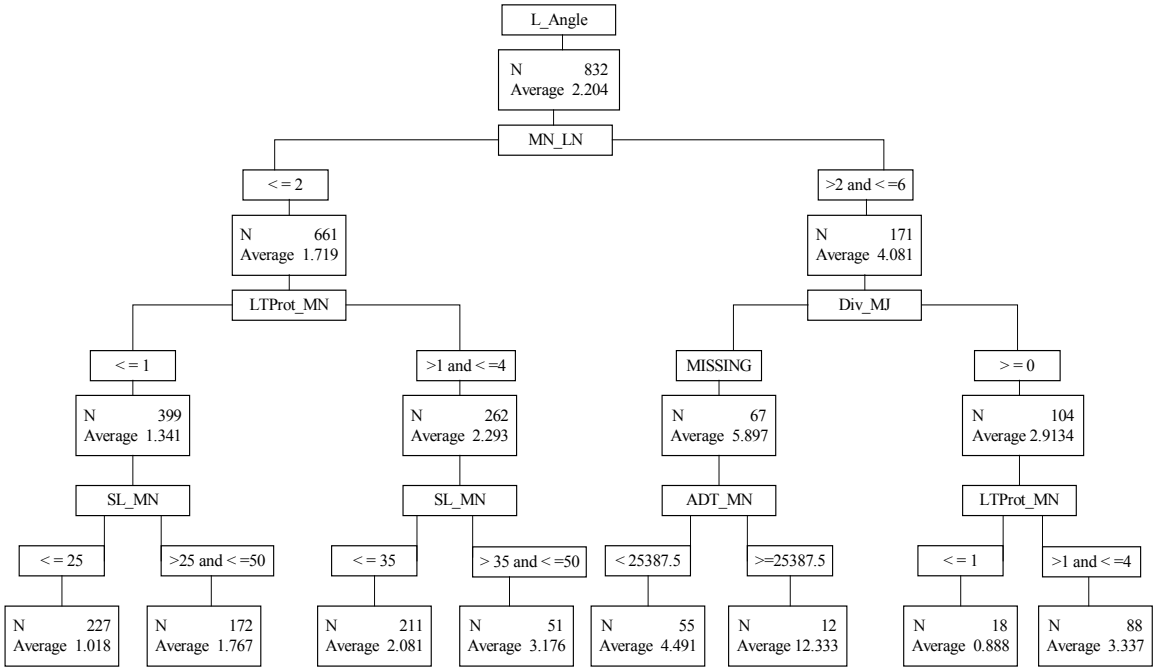

Figure B-3. Regression Tree for the Expected Number of Angle Crashes Reported on Long Forms Per Intersection for Two Years

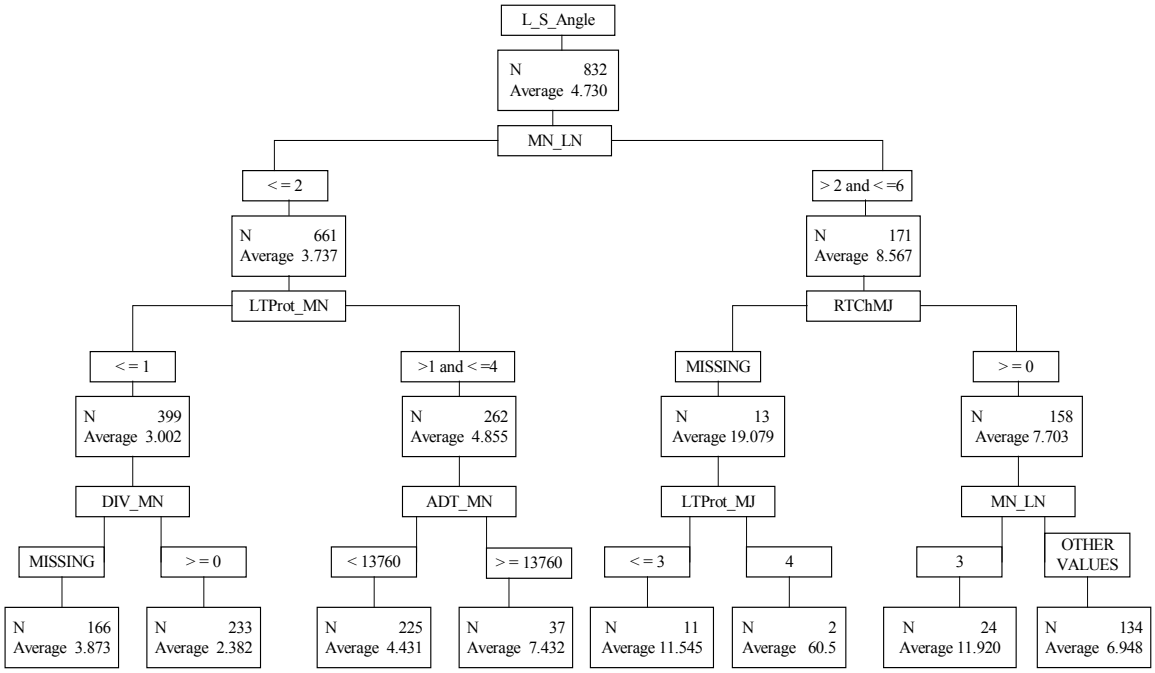

Figure B-4. Regression Tree for the Expected Number of Angle Crashes Reported on Long and Short Forms Per Intersection for Two Years

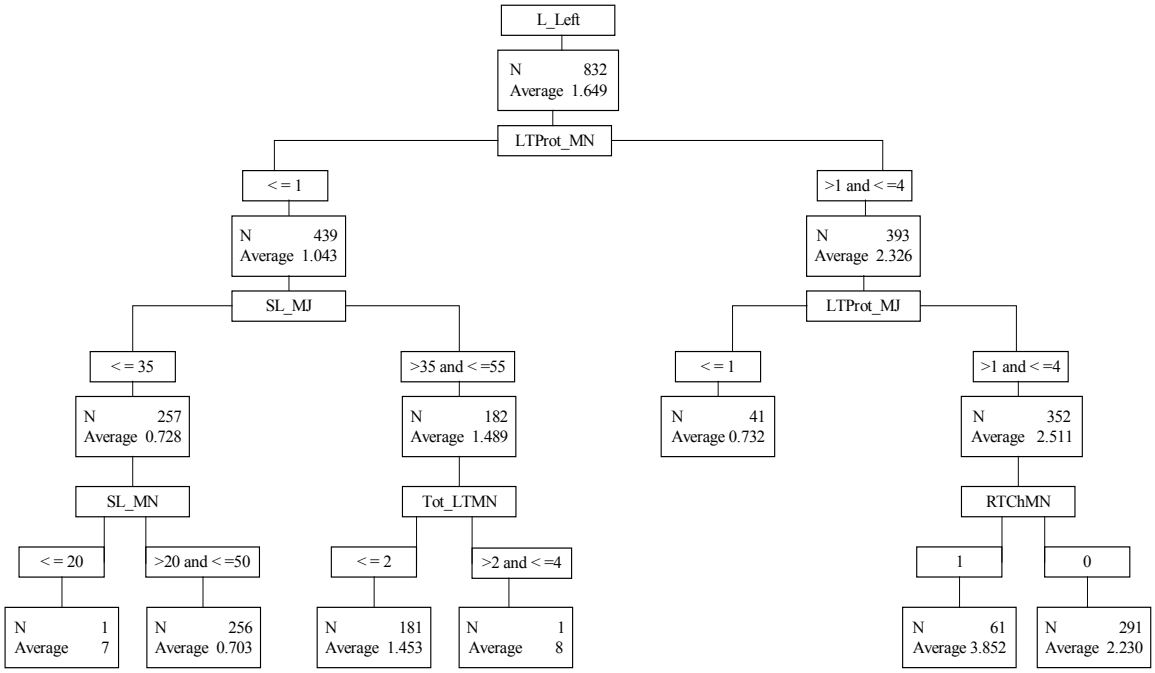

Figure B-5. Regression Tree for the Expected Number of Left Turn Crashes Reported on Long Forms Per Intersection for Two Years

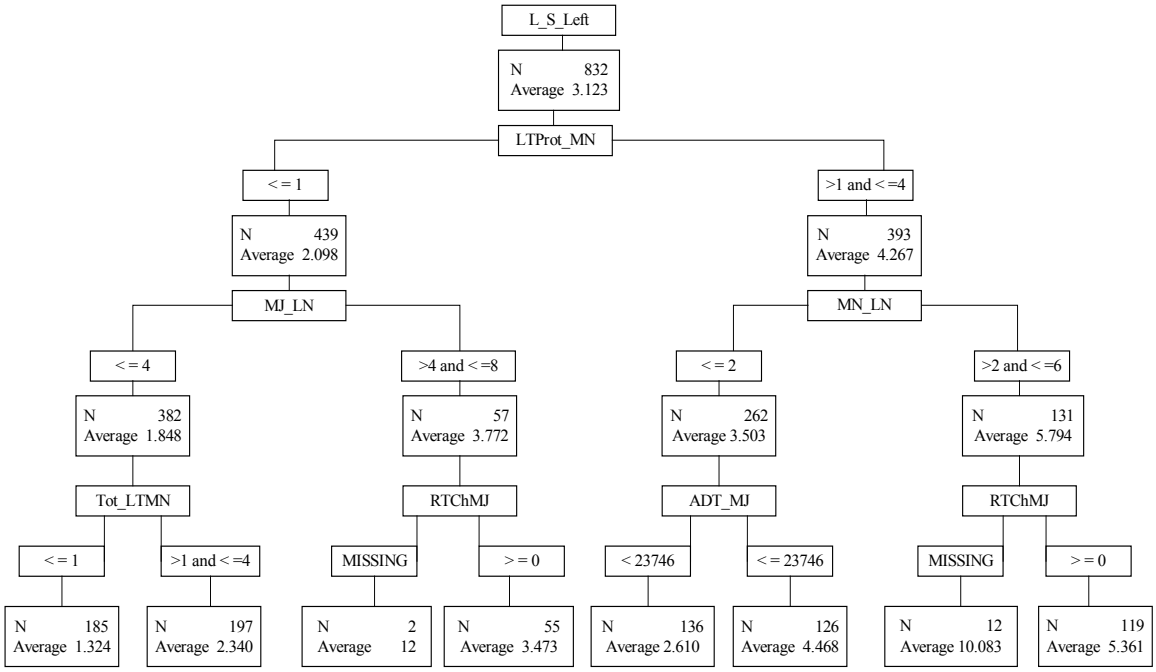

Figure B-6. Regression Tree for the Expected Number of Left Turn Crashes Reported on Long and Short Forms Per Intersection for Two Years

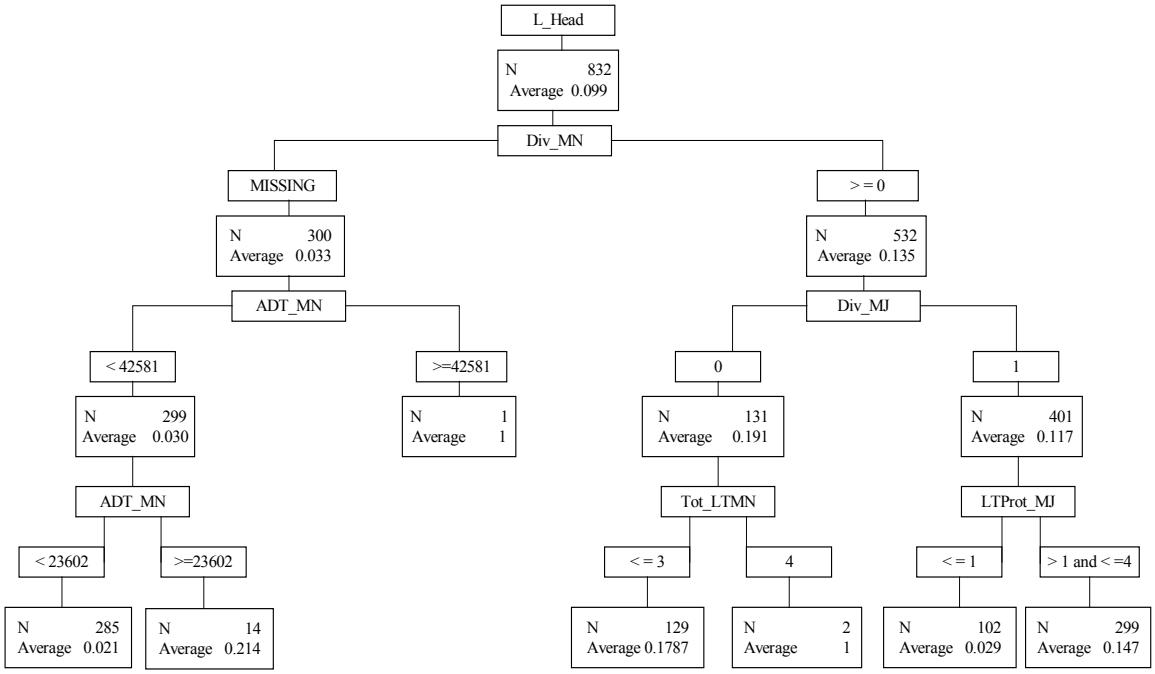

Figure B-7. Regression Tree for the Expected Number of Head-on Crashes Reported on Long Forms Per Intersection for Two Years

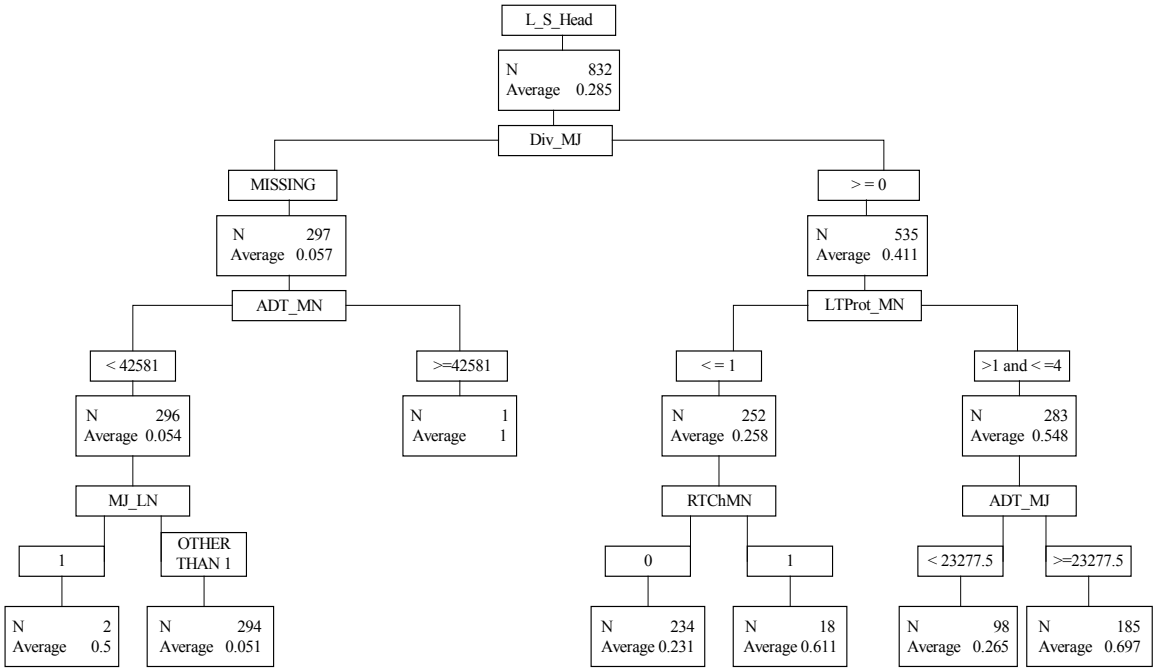

Figure B-8. Regression Tree for the Expected Number of Head-on Crashes Reported on Long and Short Forms Per Intersection for Two Years

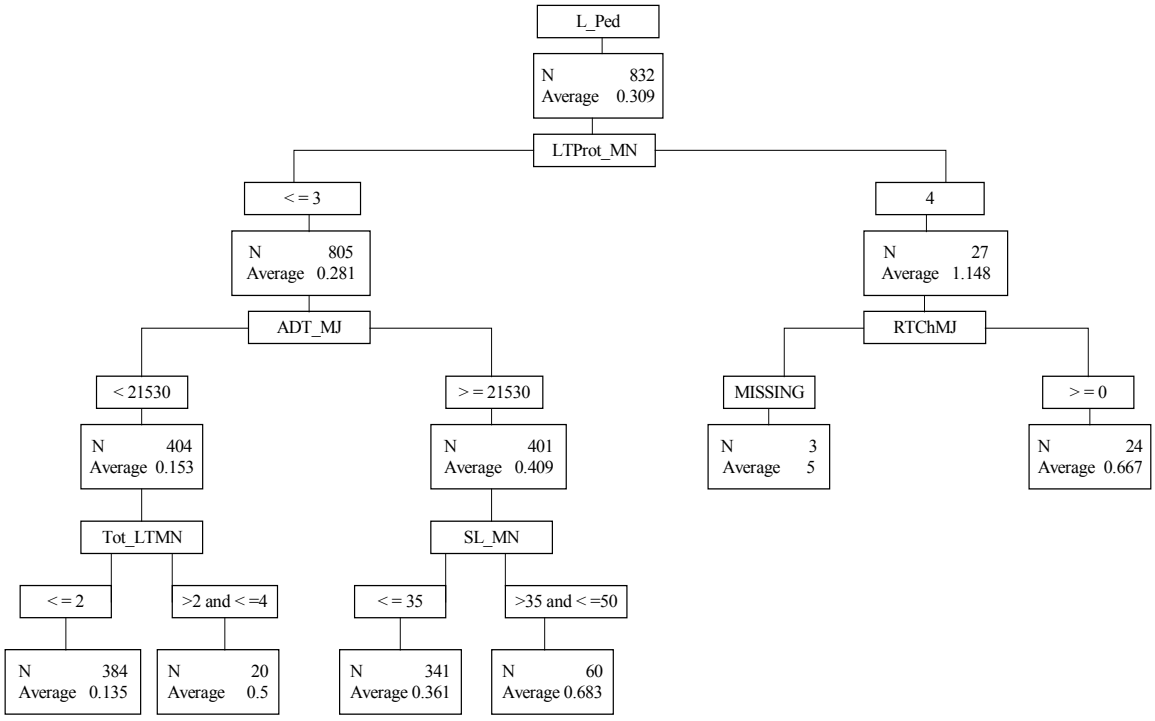

Figure B-9. Regression Tree for the Expected Number of Pedestrian/Bicycle Crashes Reported on Long Forms Per Intersection for Two Years

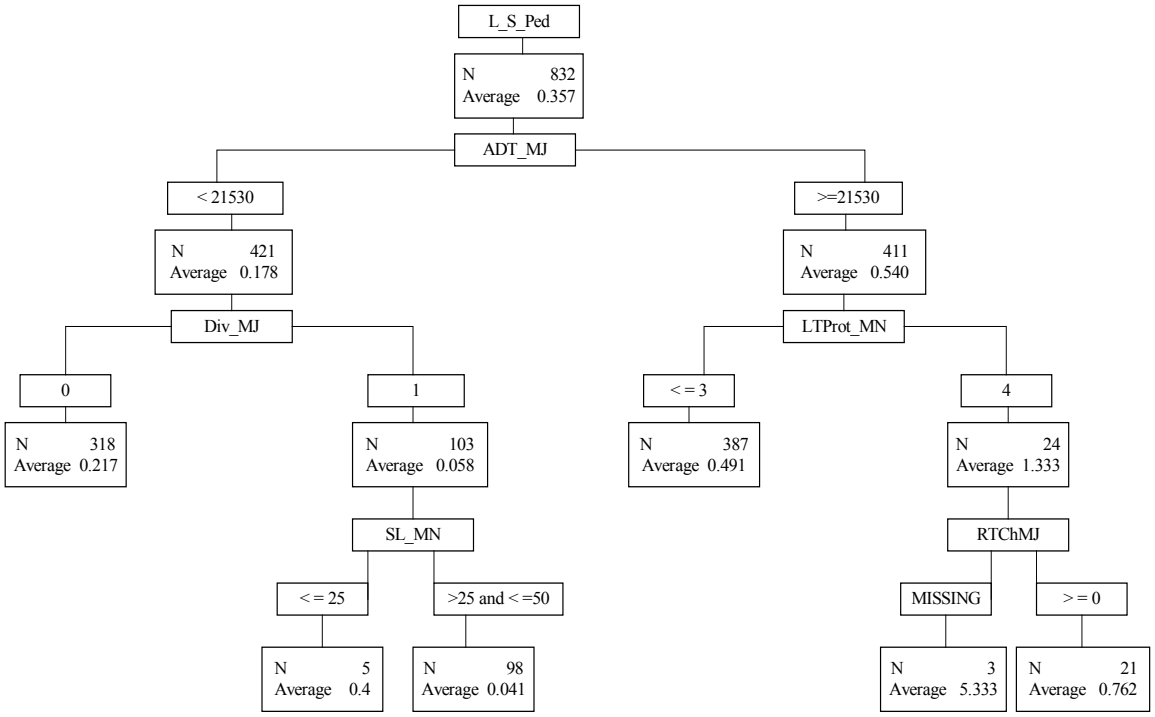

Figure B-10. Regression Tree for the Expected Number of Pedestrian/Bicycle Crashes Reported on Long and Short Forms Per Intersection for Two Years

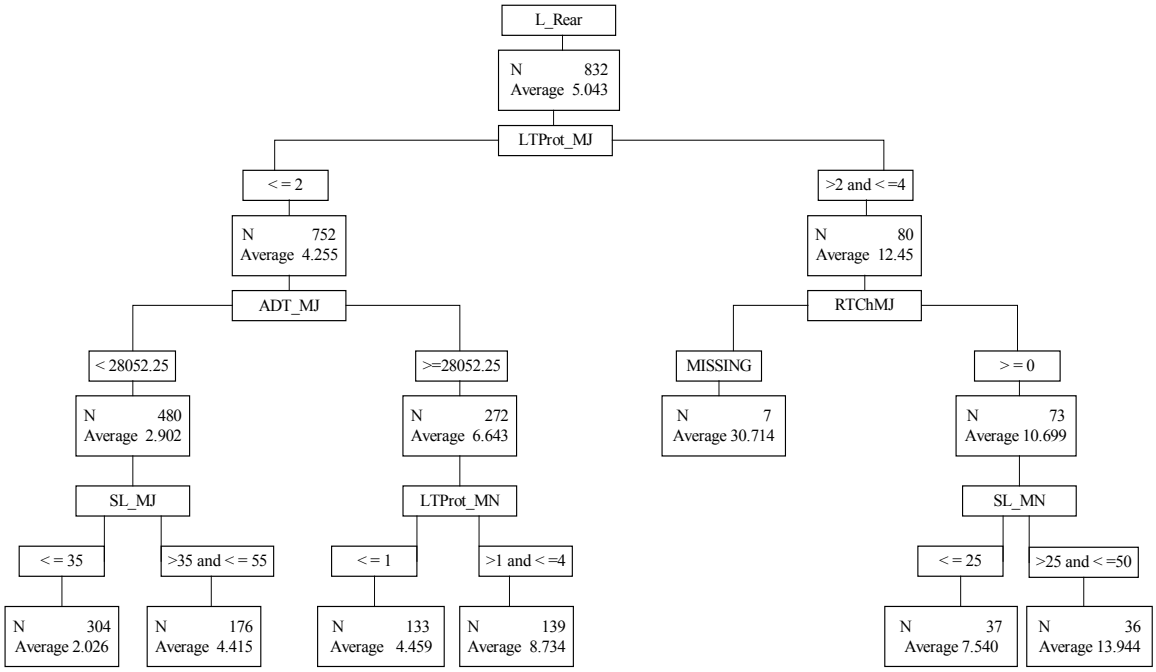

Figure B-11. Regression Tree for the Expected Number of Rear-end Crashes Reported on Long Forms Per Intersection for Two Years

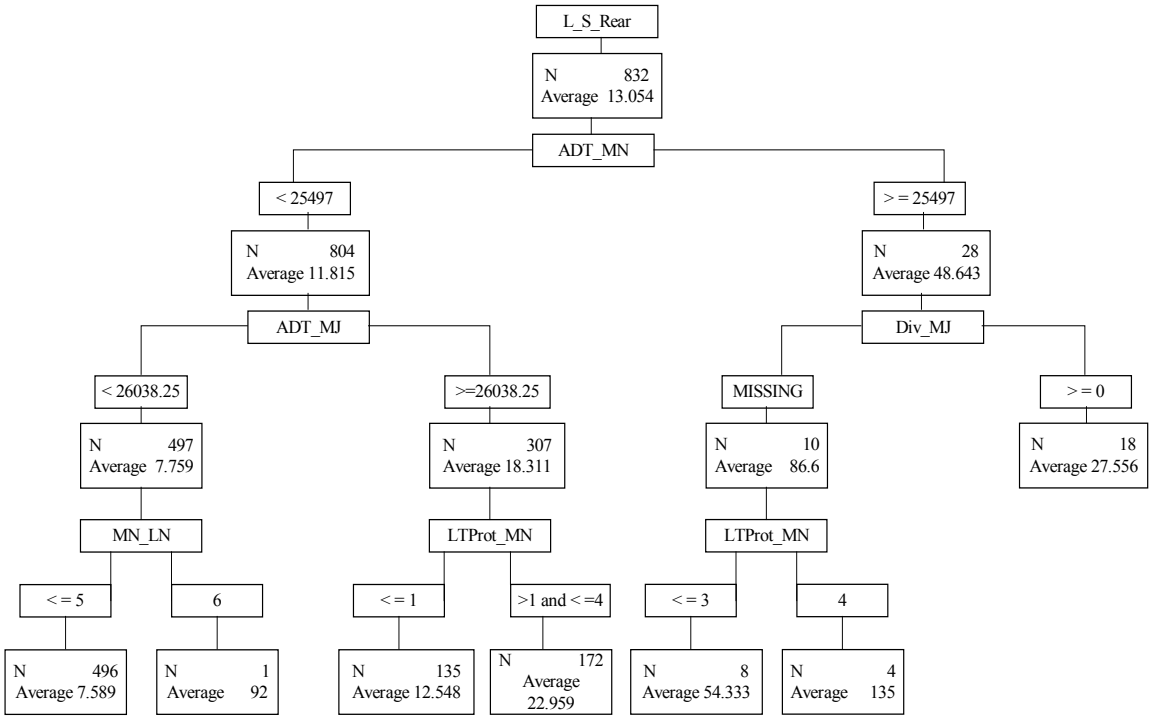

Figure B-12. Regression Tree for the Expected Number of Rear-end Crashes Reported on Long and Short Forms Per Intersection for Two Years

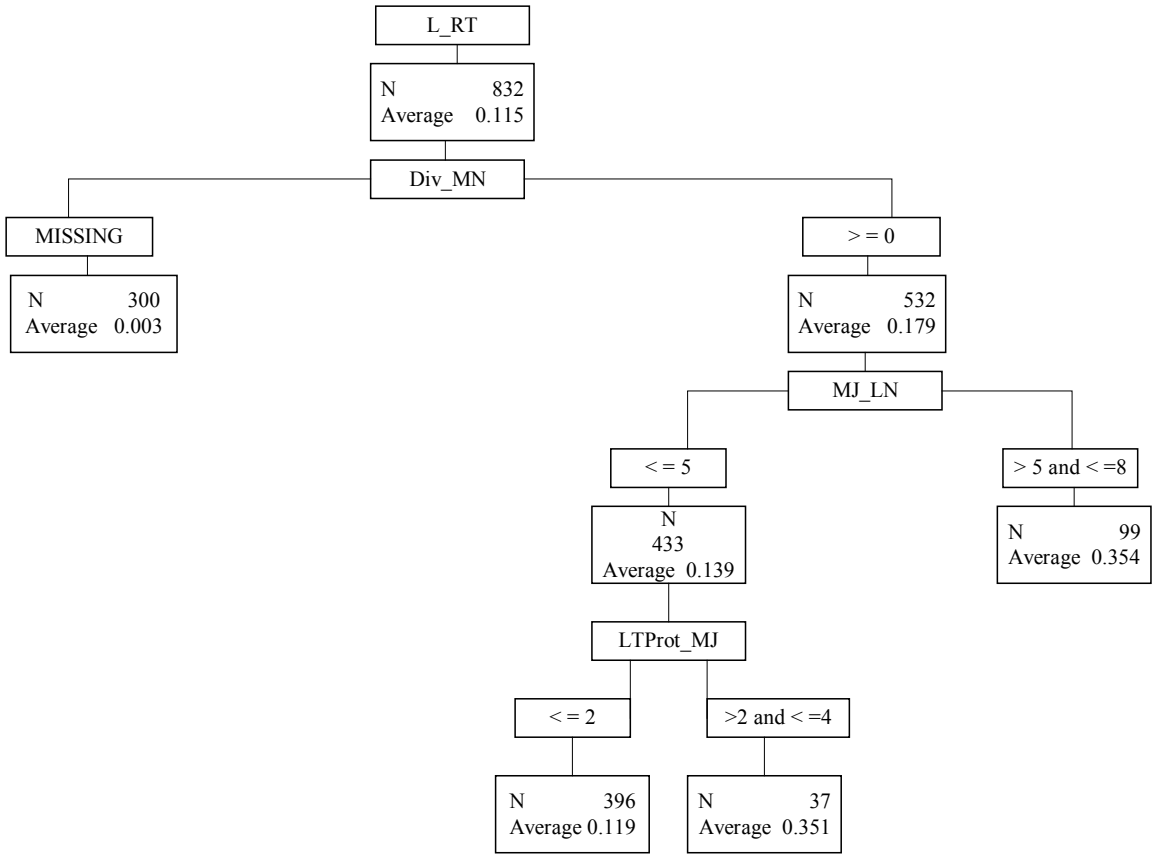

Figure B-13. Regression Tree for the Expected Number of Right Turn Crashes Reported on Long Forms Per Intersection for Two Years

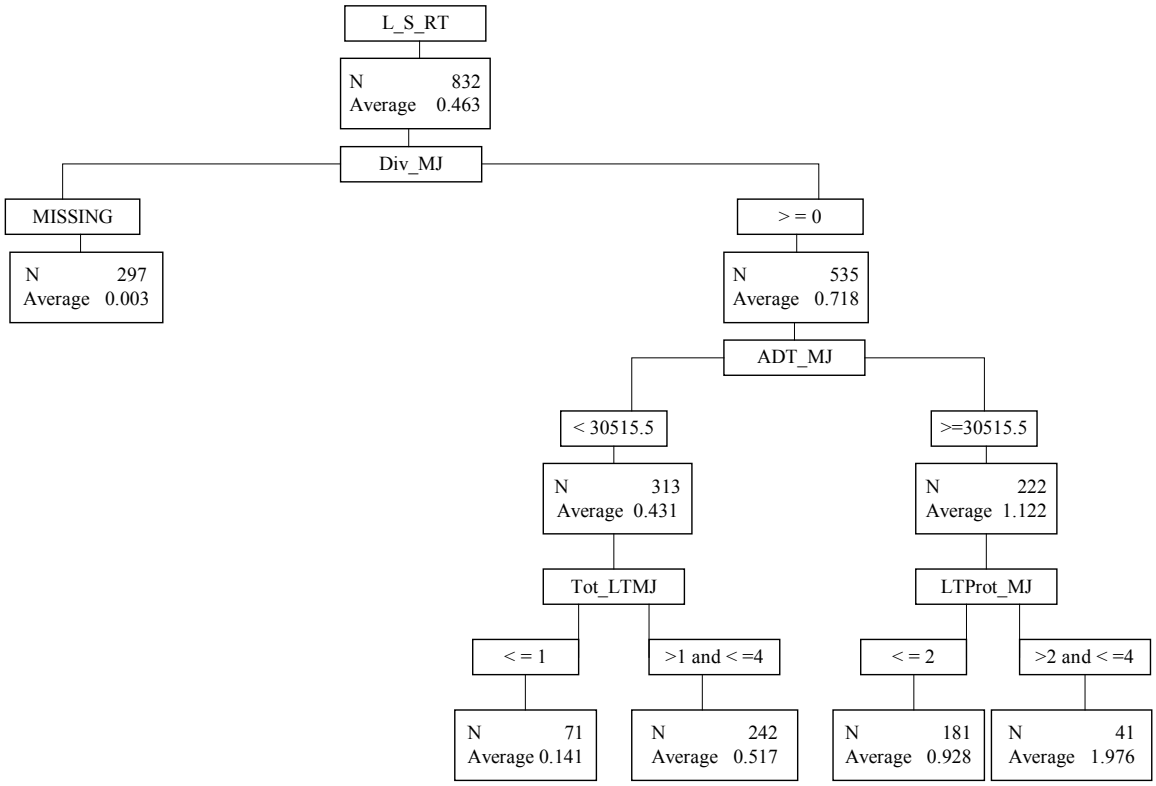

Figure B-14. Regression Tree for the Expected Number of Right Turn Crashes Reported on Long and Short Forms Per Intersection for Two Years

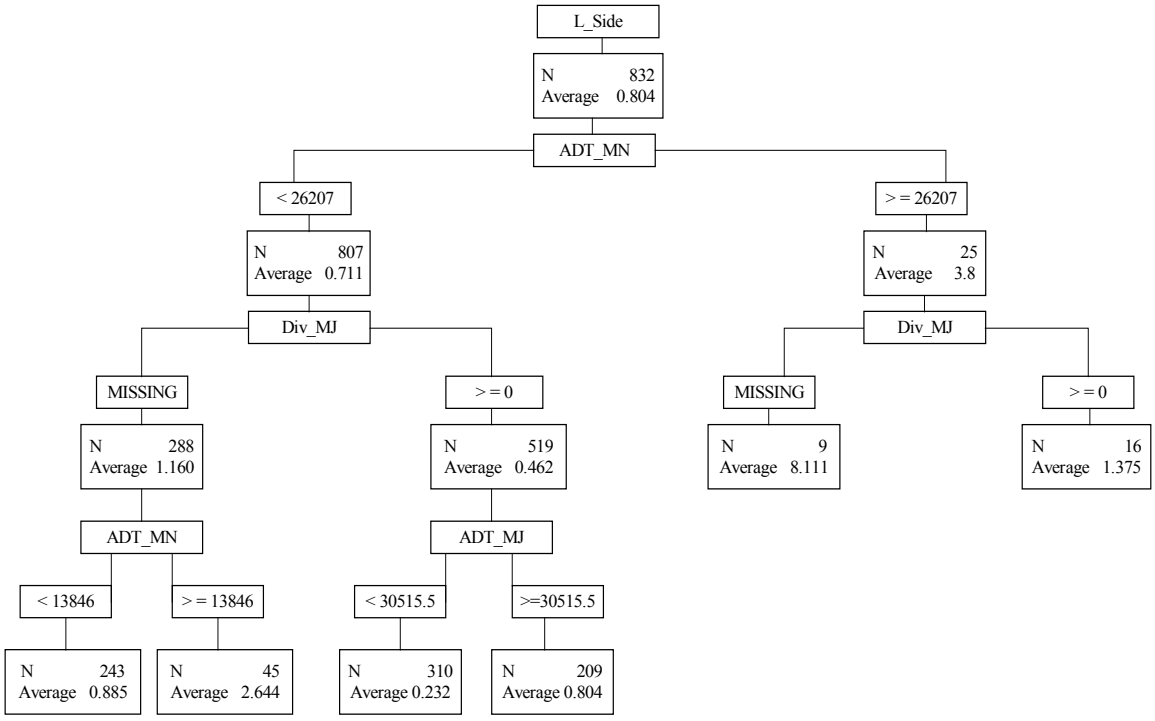

Figure B-15. Regression Tree for the Expected Number of Sideswipe Crashes Reported on Long Forms Per Intersection for Two Years

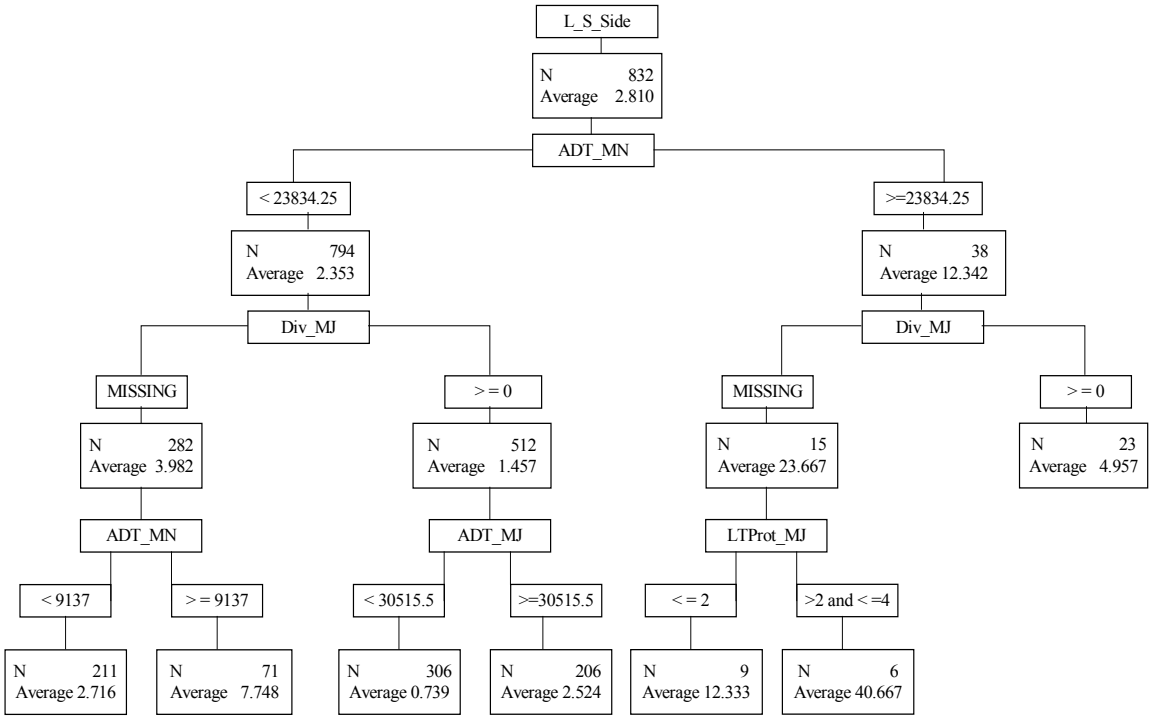

Figure B-16. Regression Tree for the Expected Number of Sideswipe Crashes Reported on Long and Short Forms Per Intersection for Two Years

APPENDIX C

RELATIVE IMPORTANCE OF FACTORS TABLES FOR CRASH TYPE AND SEVERITY LEVEL MODELS

Table C-1. List of Variables that Entered the Models based upon the Total Number of Crashes and their Relative Importance

Long Form Only Total Crashes		
Name	**Importance**	**Role**
MN_LN	1.0000	Input
LTPROT_MN	0.8551	Input
RTCHMJ	0.7616	Input
SL_MJ	0.5429	Input
MJ_LN	0.5335	Input
ADT_MJ	0.4281	Input
ADT_MN	0.3317	Input
SL_MN	0.1624	Input
DIV_MJ	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Long and Short Form Total Crashes		
Name	**Importance**	**Role**
MN_LN	1.0000	Input
LTPROT_MN	0.8917	Input
RTCHMJ	0.7527	Input
ADT_MJ	0.6866	Input
MJ_LN	0.4664	Input
ADT_MN	0.4085	Input
DIV_MN	0.2721	Input
SL_MJ	0.2466	Input
DIV_MJ	0.1183	Input
LTPROT_MJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-2. List of Variables that Entered the Models based upon Angle Crashes and their Relative Importance

Long Form Only Angle Crashes		
Name	**Importance**	**Role**
MN_LN	1.0000	Input
ADT_MN	0.8936	Input
DIV_MJ	0.6911	Input
LTPROT_MN	0.5994	Input
SL_MN	0.3708	Input
MJ_LN	0.1842	Input
ADT_MJ	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Long and Short Form Angle Crashes		
Name	**Importance**	**Role**
LTPROT_MJ	1.0000	Input
MN_LN	0.9114	Input
RTCHMJ	0.5929	Input
ADT_MN	0.5518	Input
RTCHMN	0.3551	Input
LTPROT_MN	0.3504	Input
ADT_MJ	0.2928	Input
SL_MJ	0.2454	Input
DIV_MN	0.2209	Input
TOT_LTLMN	0.1244	Input
DIV_MJ	0.0000	Rejected
MJ_LN	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected

Table C-3. List of Variables that Entered the Models based upon Left Turn Crashes and their Relative Importance

Long Form Only Left Turn Crashes		
Name	**Importance**	**Role**
LTPROT_MN	1.0000	Input
SL_MN	0.7894	Input
RTCHMN	0.6238	Input
LTPROT_MJ	0.5840	Input
MN_LN	0.4421	Input
SL_MJ	0.4256	Input
TOT_LTLMN	0.3535	Input
ADT_MJ	0.2753	Input
ADT_MN	0.0000	Rejected
DIV_MJ	0.0000	Rejected
DIV_MN	0.0000	Rejected
MJ_LN	0.0000	Rejected
RTCHMJ	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected

Long and Short Form Left Turn Crashes		
Name	**Importance**	**Role**
LTPROT_MN	1.0000	Input
MN_LN	0.6851	Input
RTCHMJ	0.6268	Input
ADT_MJ	0.4810	Input
RTCHMN	0.4783	Input
MJ_LN	0.4337	Input
TOT_LTLMN	0.3176	Input
SL_MJ	0.2532	Input
ADT_MN	0.0000	Rejected
DIV_MJ	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected

Table C-4. List of Variables that Entered the Models based upon Head-on Crashes and their Relative Importance

Long Form Only Head-on Crashes		
Name	**Importance**	**Role**
DIV_MN	1.0000	Input
ADT_MN	0.8482	Input
TOT_LTLMN	0.8162	Input
LTPROT_MJ	0.7268	Input
SL_MJ	0.6152	Input
DIV_MJ	0.5179	Input
LTPROT_MN	0.3849	Input
ADT_MJ	0.0000	Rejected
MJ_LN	0.0000	Rejected
MN_LN	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected

Long and Short Form Head-on Crashes		
Name	**Importance**	**Role**
DIV_MJ	1.0000	Input
ADT_MN	0.9465	Input
SL_MJ	0.7237	Input
ADT_MJ	0.7068	Input
LTPROT_MN	0.6990	Input
RTCHMN	0.3179	Input
MJ_LN	0.1294	Input
RTCHMJ	0.1044	Input
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
MN_LN	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-5. List of Variables that Entered the Models based upon Pedestrian and Bicycle Crashes and their Relative Importance

Long Form Only Ped/Bike Crashes		
Name	**Importance**	**Role**
RTCHMJ	1.0000	Input
LTPROT_MN	0.6265	Input
ADT_MJ	0.5122	Input
SL_MN	0.3257	Input
TOT_LTLMN	0.2246	Input
ADT_MN	0.0000	Rejected
DIV_MJ	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
MJ_LN	0.0000	Rejected
MN_LN	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected

Long and Short Form Ped/Bike Crashes		
Name	**Importance**	**Role**
RTCHMJ	1.0000	Input
ADT_MJ	0.7048	Input
LTPROT_MN	0.5407	Input
DIV_MJ	0.1890	Input
SL_MN	0.1058	Input
ADT_MN	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
MJ_LN	0.0000	Rejected
MN_LN	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-6. List of Variables that Entered the Models based upon Rear-end Crashes and their Relative Importance

Long Form Only Rear-end Crashes		
Name	**Importance**	**Role**
LTPROT_MJ	1.0000	Input
RTCHMJ	0.7260	Input
ADT_MJ	0.7074	Input
LTPROT_MN	0.5058	Input
ADT_MN	0.4618	Input
SL_MJ	0.4503	Input
DIV_MJ	0.4058	Input
SL_MN	0.3926	Input
MJ_LN	0.3837	Input
DIV_MN	0.2830	Input
MN_LN	0.0000	Rejected
RTCHMN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Long and Short Form Rear-end Crashes		
Name	**Importance**	**Role**
ADT_MN	1.0000	Input
ADT_MJ	0.8116	Input
LTPROT_MN	0.8056	Input
DIV_MJ	0.7815	Input
MN_LN	0.5742	Input
SL_MJ	0.4631	Input
RTCHMJ	0.4154	Input
DIV_MN	0.2997	Input
LTPROT_MJ	0.0000	Rejected
MJ_LN	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-7. List of Variables that Entered the Models based upon Right Turn Crashes and their Relative Importance

Long Form Only RT Crashes		
Name	**Importance**	**Role**
ADT_MJ	1.0000	Input
DIV_MN	0.9039	Input
MJ_LN	0.7186	Input
LTPROT_MJ	0.5040	Input
ADT_MN	0.0000	Rejected
DIV_MJ	0.0000	Rejected
LTPROT_MN	0.0000	Rejected
MN_LN	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Long and Short Form RT Crashes		
Name	**Importance**	**Role**
DIV_MJ	1.0000	Input
ADT_MJ	0.7969	Input
LTPROT_MJ	0.6134	Input
MN_LN	0.4791	Input
TOT_LTLMJ	0.2819	Input
MJ_LN	0.1270	Input
ADT_MN	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MN	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-8. List of Variables that Entered the Models based upon Sideswipe Crashes and their Relative Importance

Long Form Only Sideswipe Crashes		
Name	**Importance**	**Role**
ADT_MN	1.0000	Input
DIV_MJ	0.9679	Input
ADT_MJ	0.4118	Input
MJ_LN	0.2902	Input
LTPROT_MJ	0.2299	Input
RTCHMJ	0.0988	Input
DIV_MN	0.0000	Rejected
LTPROT_MN	0.0000	Rejected
MN_LN	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Long and Short Form Sideswipe Crashes		
Name	**Importance**	**Role**
ADT_MN	1.0000	Input
DIV_MJ	0.8654	Input
LTPROT_MJ	0.7293	Input
ADT_MJ	0.4596	Input
MJ_LN	0.2287	Input
MN_LN	0.0846	Input
LTPROT_MN	0.0803	Input
DIV_MN	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-9. List of Variables that Entered the Models based upon Fatal Injury Crashes and their Relative Importance

Long Form Only Fatal Crashes		
Name	**Importance**	**Role**
DIV_MJ	1.0000	Input
MJ_LN	0.7687	Input
LTPROT_MJ	0.4858	Input
ADT_MJ	0.0000	Rejected
ADT_MN	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MN	0.0000	Rejected
MN_LN	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Long and Short Form Fatal Crashes		
Name	**Importance**	**Role**
DIV_MJ	1.0000	Input
MJ_LN	0.7687	Input
LTPROT_MJ	0.4858	Input
ADT_MJ	0.0000	Rejected
ADT_MN	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MN	0.0000	Rejected
MN_LN	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
SL_MJ	0.0000	Rejected
SL_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-10. List of Variables that Entered the Models based upon Incapacitating Injury Crashes and their Relative Importance

Long Form Only Incapacitating Crashes		
Name	**Importance**	**Role**
ADT_MN	1.0000	Input
LTPROT_MN	0.8114	Input
ADT_MJ	0.7861	Input
SL_MJ	0.4493	Input
DIV_MJ	0.3735	Input
SL_MN	0.3685	Input
MJ_LN	0.3417	Input
RTCHMJ	0.2076	Input
RTCHMN	0.1044	Input
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
MN_LN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Long and Short Form Incapac. Crashes		
Name	**Importance**	**Role**
ADT_MN	1.0000	Input
LTPROT_MN	0.8114	Input
ADT_MJ	0.7861	Input
SL_MJ	0.4493	Input
DIV_MJ	0.3735	Input
SL_MN	0.3685	Input
MJ_LN	0.3417	Input
RTCHMJ	0.2076	Input
RTCHMN	0.1044	Input
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
MN_LN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-11. List of Variables that Entered the Models based upon Non-incapacitating Injury Crashes and their Relative Importance

Long Form Only Nonincap. Crashes		
Name	**Importance**	**Role**
LTPROT_MN	1.0000	Input
RTCHMJ	0.6193	Input
MJ_LN	0.4867	Input
ADT_MN	0.3916	Input
DIV_MJ	0.3132	Input
SL_MN	0.2204	Input
MN_LN	0.1567	Input
RTCHMN	0.1526	Input
ADT_MJ	0.1022	Input
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
SL_MJ	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Long and Short Form Nonincap. Crashes		
Name	**Importance**	**Role**
LTPROT_MN	1.0000	Input
RTCHMJ	0.6193	Input
MJ_LN	0.4867	Input
ADT_MN	0.3916	Input
DIV_MJ	0.3132	Input
SL_MN	0.2204	Input
MN_LN	0.1567	Input
RTCHMN	0.1526	Input
ADT_MJ	0.1022	Input
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
SL_MJ	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-12. List of Variables that Entered the Models based upon Possible Injury Crashes and their Relative Importance

Long Form Only Possible Injury Crashes		
Name	**Importance**	**Role**
MN_LN	1.0000	Input
ADT_MJ	0.6042	Input
MJ_LN	0.5512	Input
SL_MJ	0.5259	Input
TOT_LTLMJ	0.4422	Input
LTPROT_MN	0.4134	Input
SL_MN	0.3115	Input
ADT_MN	0.1976	Input
DIV_MJ	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Long and Short Possible Injury Crashes		
Name	**Importance**	**Role**
MN_LN	1.0000	Input
ADT_MJ	0.6042	Input
MJ_LN	0.5512	Input
SL_MJ	0.5259	Input
TOT_LTLMJ	0.4422	Input
LTPROT_MN	0.4134	Input
SL_MN	0.3115	Input
ADT_MN	0.1976	Input
DIV_MJ	0.0000	Rejected
DIV_MN	0.0000	Rejected
LTPROT_MJ	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

Table C-13. List of Variables that Entered the Models based upon No Injury Crashes and their Relative Importance

Long Form Only No Injury Crashes		
Name	**Importance**	**Role**
ADT_MJ	1.0000	Input
LTPROT_MN	0.6240	Input
SL_MJ	0.6133	Input
ADT_MN	0.5961	Input
MN_LN	0.5318	Input
MJ_LN	0.5270	Input
DIV_MJ	0.5000	Input
LTPROT_MJ	0.3018	Input
SL_MN	0.2756	Input
TOT_LTLMN	0.1800	Input
RTCHMN	0.1016	Input
RTCHMJ	0.0166	Input
DIV_MN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected

Long and Short Form No Injury Crashes		
Name	**Importance**	**Role**
ADT_MJ	1.0000	Input
SL_MJ	0.5296	Input
DIV_MJ	0.5290	Input
LTPROT_MN	0.3856	Input
MN_LN	0.3225	Input
SL_MN	0.1520	Input
LTPROT_MJ	0.1457	Input
ADT_MN	0.1266	Input
MJ_LN	0.0853	Input
DIV_MN	0.0000	Rejected
RTCHMJ	0.0000	Rejected
RTCHMN	0.0000	Rejected
TOT_LTLMJ	0.0000	Rejected
TOT_LTLMN	0.0000	Rejected

APPENDIX D

REGRESSIONS TREES FOR SEVERITY LEVELS

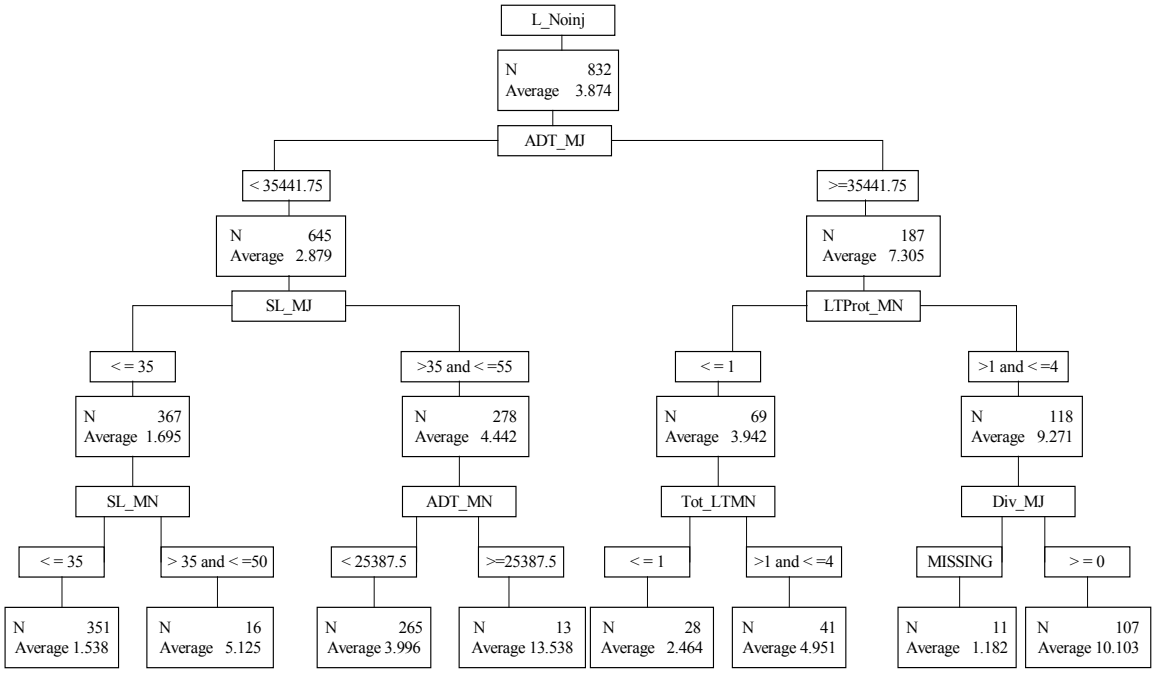

Figure D-1. Regression Tree for the Expected Number of No-Injury Crashes Reported on Long Forms Per Intersection for Two Years

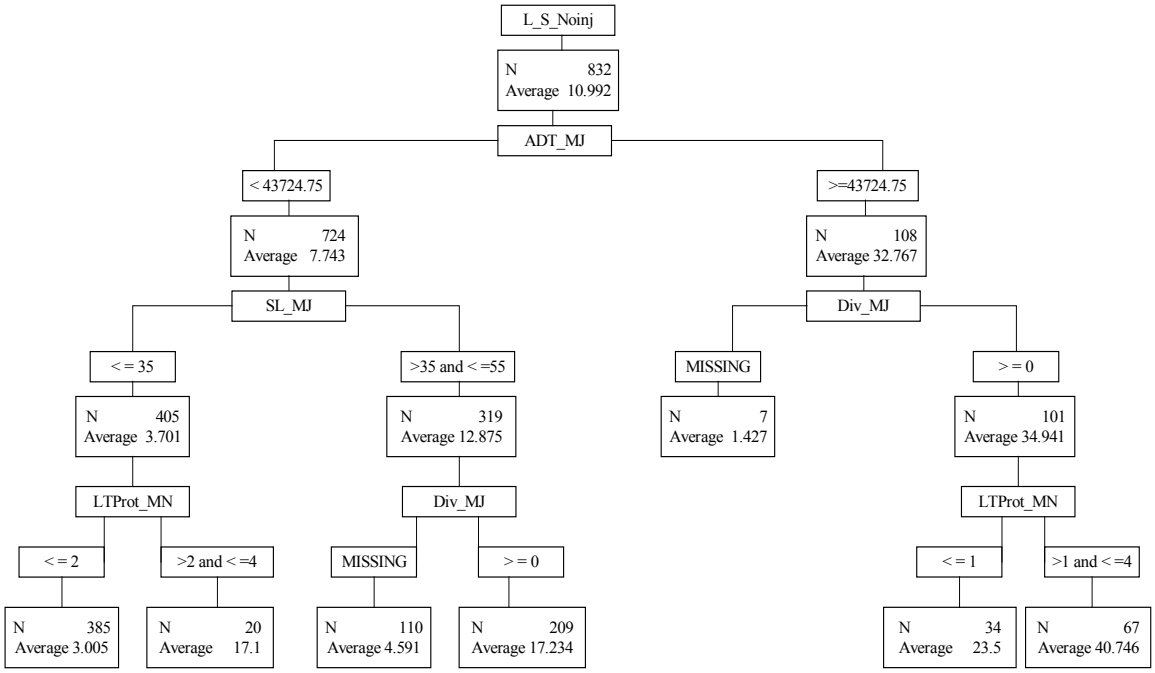

Figure D-2. Regression Tree for the Expected Number of No-Injury Crashes Reported on Long and Short Forms Per Intersection for Two Years

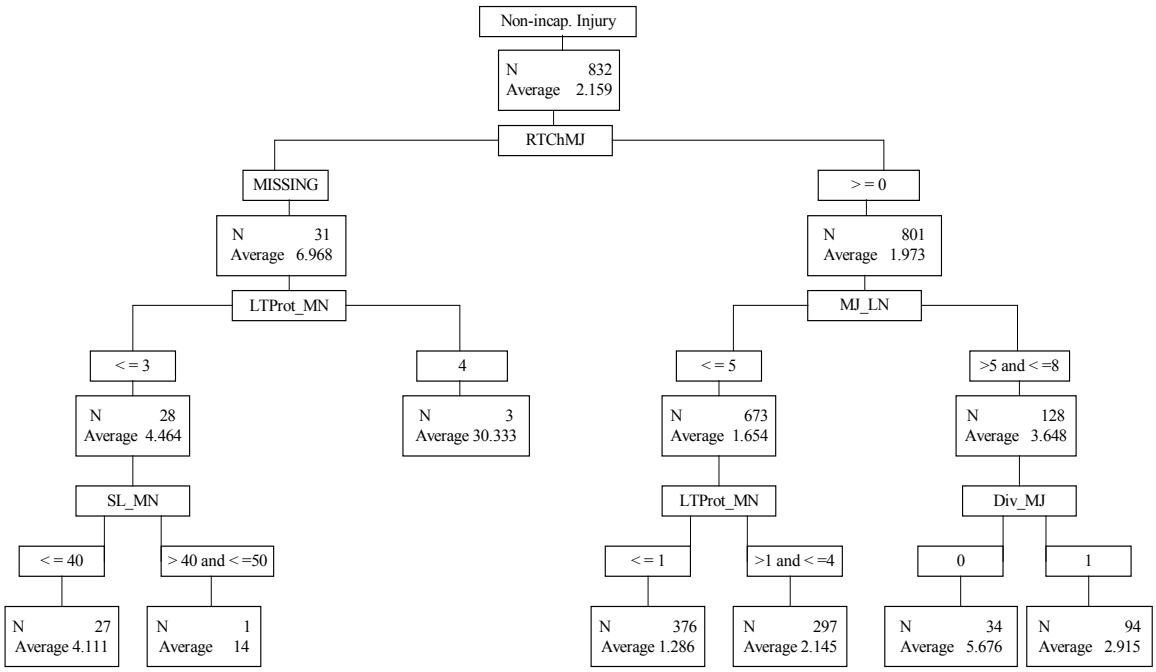

Figure D-3. Regression Tree for the Expected Number of Non-incapacitating Injury Crashes Per Intersection for Two Years

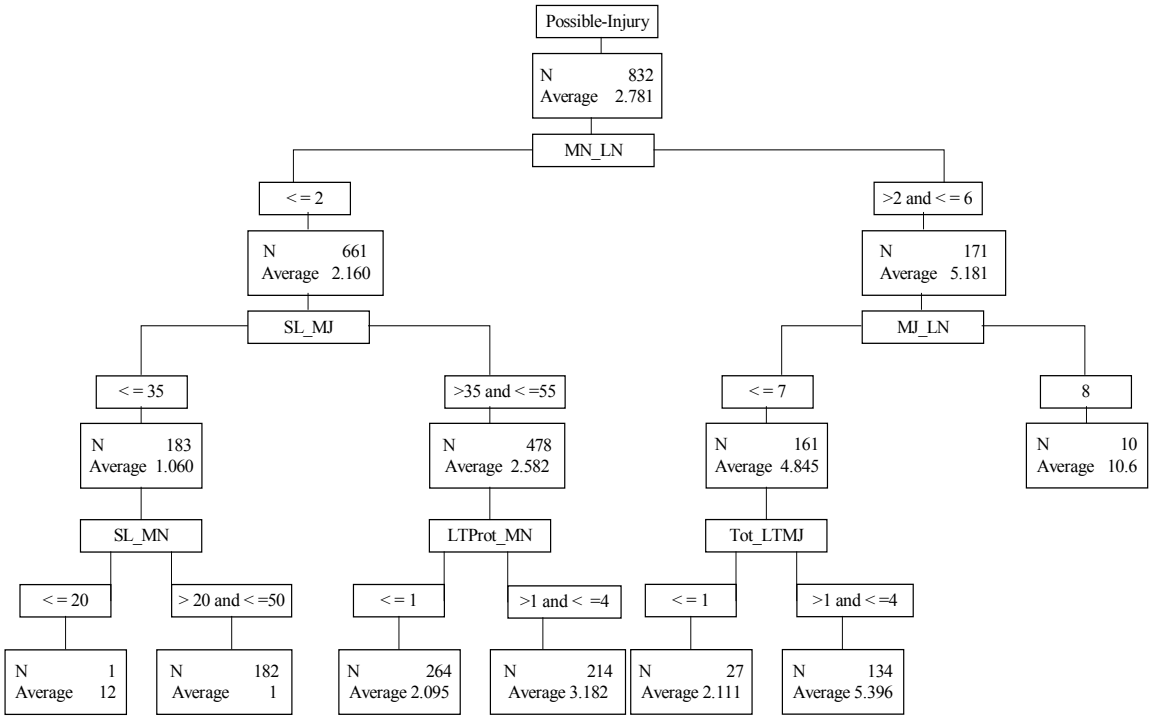

Figure D-4. Regression Tree for the Expected Number of Possible-Injury Crashes Per Intersection for Two Years

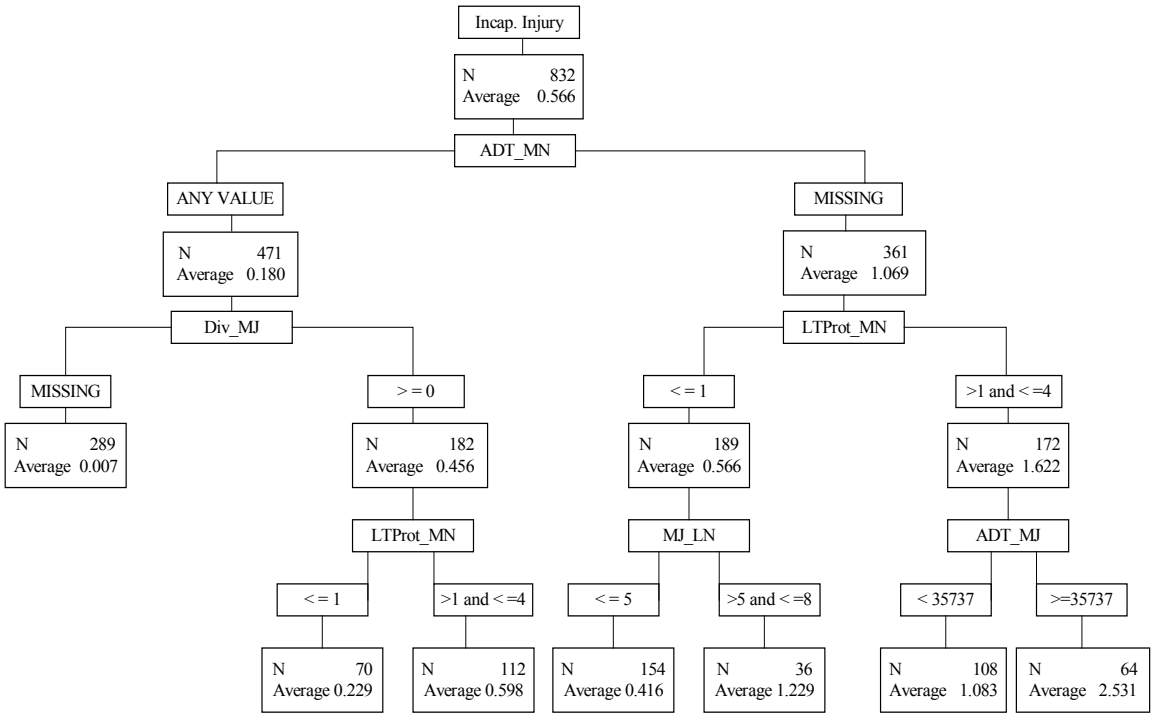

Figure D-5. Regression Tree for the Expected Number of Incapacitating Injury Crashes Per Intersection for Two Years

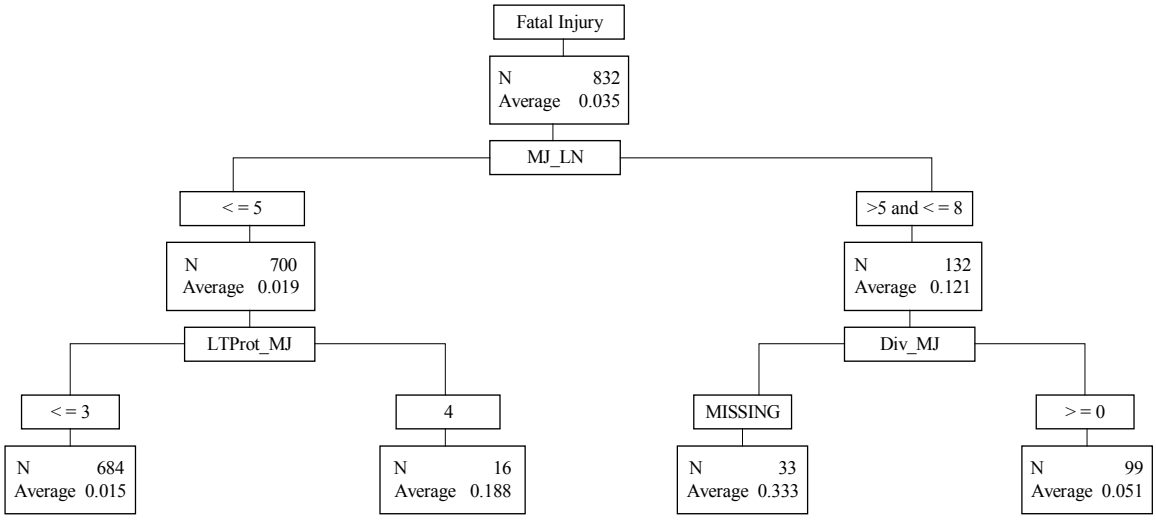

Figure D-6. Regression Tree for the Expected Number of Fatal Injury Crashes Per Intersection for Two Years

LIST OF REFERENCES

Abdel-Aty, M. "Analysis of Driver Injury Severity Levels at Multiple Locations using Ordered Probit Models." *Journal of Safety Research*. Vol 34. Issue 5 pp 597-603. 2003.

Al-Turk, M. and Moussavi, M. Accident Frequency as a Function of Traffic Congestion at Signalized Intersections. *Presented at the 75th Annual Meeting of the Transportation Research Board*, Washington, D.C. 1996.

Barceló, J. Dumont, A. Montero, L. Perarnau, J. and Torday, A. **"**Safety Indicators For Micro-simulation Based Assessments." *Presented at the 82nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2003.

Bhesania, R. P. "Using Crash Statistics and Characteristics to Improve Safety". *ITE Journal*, Vol. 61, No. 3, pp. 37-41. 1991.

Bonneson, J. and Jun Son, H. "Prediction of Expected Red-Light-Running Frequency at Urban Intersections." *Transportation Research Record No. 1830*. pp38-47. 2003.

Cafiso, S. Lamm, L. and La Cava, G. "A Fuzzy Model for Safety Evaluation Process of New and Old Roads." *Presented at the 83nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2004.

Chin, H. and Quddus, M. "Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections." *Accident Analysis and Prevention* Vol 35 Issue 2 pp. 153–159. 2003.

Donnell, E. and Mason, J. "Predicting The Severity Of Median-Related Crashes In Pennsylvania Using Logistic Regression." *Presented at the 83nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2004.

Duncan, C. Khattak, A. and Council, F. "Applying the Ordered Probit Model to Injury Severity in Truck-Passenger Car Rear-end Collisions." *Transportation Research Record No 1635*. pp 63-71. 1999.

Greibe, P. **"**Accident prediction models for urban roads." *Accident Analysis and Prevention* Vol 35 Issue 2 pp. 173–185. 2003.

Hallmark, S. Guensler, R. Fomunung, I. "Characterizing On-Road Variables That Affect Passenger Vehicle Modal Operation." *Transportation Research. Part D: Transport and Environment.* Vol 7 Issue 2 pp. 81- 98. 2002.

Harwood, D. Bauer, K. Potts, I. Torbic, D. Richard, K. Kohlman Rabbani, E. Hauer, E. and Elefteriadou, L. "Safety Effectiveness of Intersection Left- and Right-Turn Lanes." *Transportation Research Record No. 1840*. pp 131-139. 2003.

Hauer, E. Harwood, D. Council, F. and Griffith, M. "Estimating Safety by the Empirical Bayes Method: A Tutorial." *Transportation Research Record No. 1748*. pp 126-131. 2002.

Karlaftis, M. and Golias, I. **"**Effects of Road Geometry and Traffic Volumes on Rural Roadway Accident Rates." *Accident Analysis and Prevention* Vol 34 Issue 3 pp. 357-365. 2002.

Klop, J. "Factors influencing Bicycle Crash Severity on Two-Lane Undivided Roadways in North Carolina." *Presented at the 78th Annual Meeting of the Transportation Research Board, Washington, D.C.* 1998.

Ladron de Guevara, F. Washington, S. and Oh,J. "Forecasting Crashes at the Planning Level: A Simultaneous Negative Binomial Crash Model Applied in Tuscon, Arizona." *Presented at the 83nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2004.

Lee, J. Nam, D. and Moon, D. "A Zero-inflated Accident Frequency Model of Highway-Rail Grade Crossing." *Presented at the 83nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2004.

Liu, P. and Young, H. "A Neural Network Approach on Studying the Effect of Urban Signalized Intersection Characteristics on Occurrence of Traffic Accidents." *Presented at the 83nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2004.

O'Connell, L. and Kreis, D. "Accenting Safety And Objective Measures: Kentucky's New Method For Rating Highway Adequacy." *Presented at the 82nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2003.

O'Donnell, C. and Connor, D. "Predicting the Severity of Motor Vehicle Accident Injuries Using Models of Ordered Multiple Choices." *Accident Analysis and Prevention.* Vol 18 Issue 6 pp. 739-753. 1996.

Oh, J. Washington, P. and Choi, K. "Development of Accident Prediction Models for Rural Highway Intersections." *Presented at the 83nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2004.

PAB Consultants, Inc. "Expected Annual Accident Values Analysis for Dade County." *Florida Department of Transportation State Project No. 99006-1695.*

Parsonson, P. Doyle, J. Jia, X. Rumble, O. and Viera, J. "Expected Values for Accident Analyses at Atlanta Intersections." *Highway Safety Project No. FTE-93-06-105.* 1993.

Pernia, J. Lu, J. Weng, M. Xie, X. and Yu, Z. "Development of Models to Quantify the Impacts of Signalization on Intersection Crashes." *Florida Department of Transportation Contract BC353-5*. 2002. http://www11.myflorida.com/research-center/completed_te.htm.

Persaud, B. McGee, H. Lyon, C. and Lord, D. **"**Development of a Procedure for Estimating the Expected Safety Effects of a Contemplated Traffic Signal Installation." *Transportation Research Record No 1840*. pp 96-103. 2003.

Persaud, B. Lord, D. and Palmisano, J. "Calibration & Transferability of Accident Prediction Models for Urban Intersections." *Transportation Research Record No 1784*. 2002.

Pietrzyk, M. "Development of Expected Value Conflict Tables for Florida-Based Traffic Crashes." *USDOT WPI No. 0510711*, Washington, D.C. 1996.

Qin, X. Ravishanker, N. and Liu, J. "A Hierarchical Bayesian Estimation of Non-linear Safety Performance Functions for Two-Lane Highways Using MCMC Modeling." *Presented at the 82nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2003.

Sawalha, Z. and Sayed, T. "Statistical Issues in Traffic Accident Modeling." *Presented at the 82nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2003.

Steinman, N. and Hines, K. "A Methodology to Assess Design Features for Pedestrian and Bicyclist Crossings at Signalized Intersections." *Presented at the 83nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2004.

Storsteen, M. "Identification of Abnormal Accident Patterns at Intersections." *Transportation Research Board*, September 1999. South Dakota Department of Transportation, Pierre, SD.

Thomas, G. Smith, D. and Welch, T. **"**Effectiveness of Intersection Safety Improvements Using Crash Reduction and Benefit Cost Ratios**."** *Presented at the 81nd Annual Meeting of the Transportation Research Board,* Washington, D.C. 2002..

Wang, Y. and Nihan, N. "Quantitative Analysis On Angle-Accident Risk At Signalized Intersections." *Presented at 9th World Congress on Transport Research Seoul, Korea, July 2001.*

Washington, S. "Iteratively Specified Tree-Based Regression: Theory And Trip Generation Example." *Journal of Transportation Engineering.* Vol. 116 Issue 6 pp. 481-491. 2000.

Washington, S and J. Wolf. "Hierarchical Tree-Based Versus Ordinary Least Squares Linear Regression Models: Theory And Example Applied To Trip Generation." *Transportation Research Record No 1581*. pp 82-88. 1997.

Washington, S, Wolf, J. Guensler, R. “Binary Recursive Partitioning Method For Modeling Hot-Stabilized Emissions From Motor Vehicles.” *Transportation Research Record No 1587*. pp 96-105. 1997.

Weerasuriya, S. and Pietrzyk, M. “Development of Expected Conflict Value Tables for Unsignalized Three-legged Intersections.” *Transportation Research Record No. 1635*. pp 121-126. 1998.