2005

Neural Network Trees And Simulation Databases: New Approaches For Signalized Intersection Crash Classification And Prediction

Piyush Nawathe
University of Central Florida

STARS Citation

Nawathe, Piyush, "Neural Network Trees And Simulation Databases: New Approaches For Signalized Intersection Crash Classification And Prediction" (2005). *Electronic Theses and Dissertations, 2004-2019*. 475.
https://stars.library.ucf.edu/etd/475

NEURAL NETWORK TREES AND SIMULATION DATABASES: NEW APPROACHES FOR SIGNALIZED INTERSECTION CRASH CLASSIFICATION AND PREDICTION

by

PIYUSH NAWATHE
B. Tech. & M.Tech, Indian Institute of Technology - Madras, Chennai, India 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
2005

ABSTRACT

Intersection related crashes form a significant proportion of the crashes occurring on roadways. Many organizations such as the Federal Highway Administration (FHWA) and American Association of State Highway and Transportation Officials (AASHTO) are considering intersection safety improvement as one of their top priority areas. This study contributes to the area of safety of signalized intersections by identifying the traffic and geometric characteristics that affect the different types of crashes.

The first phase of this thesis was to classify the crashes occurring at signalized intersections into rear-end, angle, turn and sideswipe crash types based on the traffic and geometric properties of the intersections and the conditions at the time of the crashes. This was achieved by using an innovative approach developed in this thesis "Neural Network Trees". The first neural network model built in the Neural Network tree classified the crashes either into rear end and sideswipe or into angle and turn crashes. The next models further classified the crashes into their individual types. Two different neural network methods (MLP and PNN) were used in classification, and the neural network with a better performance was selected for each model. For these models, the significant variables were identified using the forward sequential selection method. Then a large simulation database was built that contained all possible combinations of intersections subjected to various crash conditions. The collision type of crashes was predicted for this simulation database and the output obtained was plotted along with the input variables to obtain a relationship between the input and output variables. For example, the analysis showed that the number of rear end and sideswipe crashes increase relative to the angle and turn crashes when there is an increase in the major and minor roadways' AADT and speed limits, surface conditions, total left turning lanes, channelized right turning lanes for the major roadway and the protected left turning lanes for the minor roadway, but decrease when the light conditions are dark.

The next phase in this study was to predict the frequency of different types of crashes at signalized intersections by using the geometric and traffic characteristics of the intersections. A high accuracy in predicting the crash frequencies was obtained by using another innovative method where the intersections were first classified into two different types named the "safe" and "unsafe" intersections based on the total number of lanes at the intersections and then the frequency of crashes was predicted for each type of intersections separately. This method consisted of identifying the best neural network for each step of the analysis, selecting significant variables, using a different simulation database that contained all possible combinations of intersections and then plotting each input variable with the average output to obtain the pattern in which the frequency of crashes will vary based on the changes in the geometric and traffic characteristics of the intersections. The patterns indicated that an increase in the number of lanes of the major roadway, lanes of the minor roadway and the AADT on the major roadway leads to an increased crashes of all types, whereas an increase in protected left turning lanes on the major road increases the rear end and sideswipe crashes but decreases the angle, turning and overall crash frequencies.

The analyses performed in this thesis were possible due to a diligent data collection effort. Traffic and geometric characteristics were obtained from multiple sources for 1562 signalized intersections in Brevard, Hillsborough, Miami-Dade, Seminole and Orange counties and the city of Orlando in Florida. The crash database for these intersections contained 27,044 crashes.

This research sheds a light on the characteristics of different types of crashes. The method used in classifying crashes into their respective collision types provides a deeper insight on the characteristics of each type of crash and can be helpful in mitigating a particular type of crash at an intersection. The second analysis carried out has a three fold advantage. First, it identifies if an intersection can be considered safe for different crash types. Second, it accurately predicts the frequencies of total, rear end, angle, sideswipe and turn crashes. Lastly, it identifies the traffic and geometric characteristics of signalized intersections that affect each of these crash types. Thus the models developed in this thesis can be used to identify the specific problems at an intersection, and identify the factors that should be changed to improve its safety.

ACKNOWLEDGMENTS

I would firstly like to thank my advisor, Dr. Mohamed Abdel-Aty, for being helpful and supportive throughout my Masters study at UCF. I would also like to thank Dr. Amr A. Oloufa and Dr. Balaji Ramachandran for serving in my committee. I thank Dr. Balaji again for his help in making me understand GIS. Xuesong and Hari have been immensely helpful in all stages of my thesis work, and I thank them for this. I thank Anurag and Ravi for being ready to discuss any aspect of my thesis at any point of time. My friend circle at UCF, Swapnil, Ravi, Raja, Sai, Sandy, Albinder, Nezam and Jai, have been a continuous source of encouragement. Akshay and Narsi have guided me through tough times and I am thankful to them. Finally, I am indebted to my parents and sister for encouraging me to perform at the best of my potential and to never give up.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Background

One of the most complex situations faced by a driver on a roadway is an intersection. With many vehicles and pedestrians entering and leaving an intersection, there are greater possibilities of different types of crashes. According to Federal Highway Administration (FHWA), (National Agenda for Intersection Safety, May 2002) more than 2.8 million intersection-related crashes had occurred in the United States in the year 2000, which represented 44% of the total crashes reported. Around 8500 fatalities, representing 23% of the total fatalities, and almost one million injury crashes had occurred at intersections. FHWA also states that more than half of all rear end crashes occur at or near the intersections and more than one-third of all deaths to vehicle occupants occur in angle crashes. Both human and property damage losses from rear-end crashes cost the United States billions of dollars each year in medical expenses, lost productive time and numerous property insurance claims. The cost to society for intersection-related crashes is approximately $40 billion a year. National Highway Traffic Safety Administration (NHTSA) estimates that the injury costs alone for rear-end crashes exceed $5 billion per year. Thus there is a need to study the crash phenomenon and identify the factors that make such crashes more probable. This can be achieved by classifying crashes into their respective collision types.

Numerous highway safety organizations have identified intersection safety as a national priority. The FHWA has identified intersection safety as one of four safety priority areas in the agency's performance plan. The Am.erican Association of State Highway and Transportation Officials (AASHTO) Strategic Highway Safety Plan includes 22 key emphasis areas, one of which is improving the design and operation of highway intersections. Therefore, there is a need to study the crash characteristics at intersections and put forth appropriate solutions that can make intersections a safer place to travel. The solution to this is to build models to predict the number of crashes that can be expected to occur at signalized intersections and to identify the variables that affect each type of crash. If the model suggests that the intersection is expected to have a large number of crashes, the intersection characteristics that lead to an increased number of crashes can be controlled in order to decrease the crash rate.

1.2 Research Objectives

The main objective of this thesis is to analyze the crash characteristics at or near signalized intersections and to develop models that will be helpful in increasing safety at intersections. To accomplish this, the first step will be to review previous studies on intersection safety and determine the methodologies used by them. Then, data will be collected on intersection properties and characteristics of the crashes that occurred at these intersections. The next step would be to analyze the data in order to predict the frequency of crashes occurring at the intersections. Different neural network models will be utilized in this phase to accurately predict the crash frequencies and the best model will be identified that will give the least error in estimation. The frequencies of crashes with different collision types (e.g., rear end and angle crashes) will also be predicted using the neural network models. This will be followed by the identification of significant variables for each of these models and the manner in which they affect the crash frequencies.

Another objective of this thesis is to classify crashes into their respective collision type based on the traffic and geometric characteristics of the intersections and the conditions known at the time of the crash. By using different neural network methods, the best method that can be used in the classification of crashes with a high accuracy will be acknowledged. The significant variables will be determined for this classification and their effect on the classification will be determined. Thus, this analysis will indicate the type of crash that will be most likely to occur based on the given input variables.

1.3 Organization of Thesis

This thesis is organized as follows:

1. *Literature Review:* This chapter consists of a review of various studies performed in the area of traffic safety and the analysis methods used by them.
2. *Methodology:* The relevant models that can be used in the thesis in order to fulfill the objectives have been identified. The functioning of these models will be discussed in detail.
3. *Data Collection:* The data collection effort for the thesis has been described in this chapter. The data finally obtained in the study has been described in detail.
4. *Using Neural Networks to identify Unsafe Intersections:* This chapter describes a new technique in predicting crashes at signalized intersections that is also capable of identifying if an intersection in safe or unsafe. The formulation of such a

technique and the results obtained in estimating different types of crashes have also been discussed in the chapter.

5. *Classification of Crashes using Neural Network Trees:* The classification of crashes into their respective collision types using an innovative method called the "Neural Network Trees" has been described in this chapter. The chapter also discusses the significant variables identified in this analysis and their effects on the classification.
6. *Conclusions:* The final chapter contains a briefing on the work carried out in the thesis and discusses appropriate conclusions.

2 LITERATURE REVIEW

2.1 Introduction

According to the Bureau of Transportation Statistics, 6,328,000 crashes were estimated to have occurred in the year 2003, of which there were 42,643 fatalities and around 3 million were estimated to be injuries. Of these fatalities, 21% were reported as intersection fatalities, which is a high percentage. Hence there is a need to identify the intersection characteristics that lead to an increased rate of crashes at these locations. By controlling these factors, the intersections can be made a safer place to travel. In order to achieve this, some statistical, geographical or neural network methodologies have to be applied on crash data collected at intersections. This chapter explores some studies that have been carried out in the recent past, which have used such methods to bring forth the characteristics of highways and intersections that can alter their safety.

2.2 Poisson and Negative Binomial Modeling

Poch et al. (1996) presents a negative binomial analysis to study the relationship between road geometrics/traffic related elements and accident frequencies at intersections. Four different models were developed that predicted total accident frequency, rear-end accident frequency, angle accident frequency and approach turn accident frequency. A total of 63 intersections were studied in the analysis, for which the data was collected between the years 1988 and 1992. The variables used in the analysis were approach volumes, number of approach lanes, speed limits, highway grades, signal-control characteristics, presence of horizontal curves, sight distance restrictions and indicator variables for the calendar year for the data and the location of the intersection. The model is able to identify the factors that tend to increase/decrease the accident frequencies, for various collision types. Hence the authors conclude that the negative binomial regression model can be satisfactorily used in identifying the significant traffic and geometric elements that tend to increase or decrease the accident frequency.

The concept of random effect negative binomial model was used by Chin et al. (2003) to identify elements that affect intersection safety. This model can deal with the spatial and temporal effects in the data by treating the data in a time-series cross-section panel. A total of 52 four-legged intersections in the Southwestern part of Singapore were used, which accounted for 832 crashes between the periods of 1992 to 1999. The random effect negative binomial model was used to examine a total of 32 possible explanatory variables, which were classified into traffic volumes, geometric elements and regulatory control measures. The results showed that 11 variables significantly affected the safety at the intersections. The total approach volumes, the numbers of phases per cycle, the uncontrolled left-turn lane and the presence of a surveillance camera are among the variables that are the highly significant. On the other hand, the presence of an acceleration section and the provision of bus bays as well as the use of adaptive signal control tend to point to lower total crash occurrence. These findings might be limited by the relatively small sample size used.

Another study to formulate practicable accident prediction models that would describe the expected number of accidents at junctions and road links in urban areas was conducted by Greibe (2003). Poisson distribution model was used to identify factors affecting safety, geometry, land use, etc. The model incorporated accident data for five years, and also contained a plethora of variables like AADT counts, length of section, speed limit, one or two-way traffic, number of lanes, road width, speed reducing instruments, etc for roadway sections and traffic volumes, number of lanes, median, turning lanes, bicycle facilitation, signalized/non-signalized, and number of signal arms for intersections. The results for both roadway segments and intersections indicated that ADT contributed the most to crash frequency. Explanatory variables describing road design and road geometry proved to be significant for road link models but less important in junction models.

Vogt et al. (1998) used poison and negative binomial models to study the three-legged and four-legged intersections' crashes. The data were obtained from Highway Safety Information System (HSIS) files for the states of Minnesota and Washington for the time periods 1985 to 1989 and 1993 to 1995 respectively. Intersections in Minnesota were selected from a population of HSIS observations divided into four bins, with random selection from each bin. The bins were defined by median values of mainline traffic and minor road traffic. In case of Washington, no HSIS intersection file was available, but an intersection database was developed through combining video-log information with data provided by the state of Washington. The results showed that right-turn lanes on the mainline increase the likelihood of crashes at three-legged intersections. For the four-legged intersections, fewer crashes result at right-angled intersections.

Both Poisson and negative binomial regressions were also used by Oh et al. (2004) 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 the condition whether the intersection was lighted caused fewer crashes to occur.

Shankar et al. (1997) suggests that the accident frequencies can be considered to be belonging to two states: one in which the roadway section from which the accident data is collected is inherently safe, and the other is the accident state in which accident frequencies follow a known distribution. The former distribution case is the zero accident state in which no accidents will be observed. If the two processes are modeled as a single process that assumes that all sections are in accident state, the estimated models will be inherently biased because there will be an over representation of zero-accident observations in the data. Hence the paper explores the conditions under which the Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB) models are more appropriate than the simple Poisson and Negative Binomial models. Analysis was carried out with the data collected for highway sections. The data is limited to non-intersection roadway sections. The section defining information included changes to district number, urban or rural section, roadway type, number of lanes, roadway width, shoulder width, presence of curb or retaining wall, divided or undivided highway, speed, AADT, truck percentage, peak hour factors and vertical and horizontal curve characteristics. For the model estimation, 2-year summary of accident data was used.

The analysis shows that several variants of the ZIP/ZINB are plausible, and that roadway engineers can isolate design control factors that affect zero-accident processes and positive accident processes.

Persaud (2003) 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. 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.

Rodriguez et al. (1999) developed crash prediction models for estimating the safety performance of urban unsignalized intersections. The models are developed using the generalized linear modeling (GLIM) approach that addresses and overcomes the shortcomings associated with conventional linear regression. The safety predictions obtained from the models are refined using the empirical Bayes approach to provide more accurate, site-specific safety estimates. The study made use of sample crash and traffic volume data corresponding to unsignalized intersections located in urban areas of the British Columbia. Four applications of the models are described: identification of crash-prone locations, developing critical crash frequency curves, ranking the identified crash-prone location, and before and after safety evaluation. These applications showed the importance of using crash prediction models to reliably assess the safety of unsignalized intersections.

In 1998, Turner reviewed models used in practice to relate crashes to traffic flow, with particular emphasis on the appropriateness of the model form and the statistical analysis technique employed for parameter estimation. The development of generalized linear models for predicting individual crash types at intersections in New Zealand was then described. The use of covariate analysis to identify the effect of intersection location, an investigation of the effect of non-collision flows, and the use of the models for predicting intersection crashes in three networks were also described. It was concluded that generalized linear models for estimating different crash types (based on the conflicting flows) were better than models for estimating total crashes (based on the approach flows), especially when the cost of different crash types was known. It was also found that intersection location affects the number of different crash types. It was important to consider the interactions between turning flows (to take better account of the mechanisms of crash occurrence) as well as non-collision flows. Comparison of the predicted and observed numbers of crashes showed that there was poor agreement for individual intersections, but fairly good agreement for networks.

Mountain et al. (1996) developed and validated a method for predicting expected accidents on main roads with minor junctions where the traffic counts on minor arms are not available. The study was based on data for around 3800 km of highway in the U.K with more than 5000 minor junctions. Generalized linear modeling was used in this study to develop regression estimates of expected accidents for six highway categories and an empirical Bayes procedure was used to improve these estimates by combining them with accident counts. In the paper, accidents on highway sections have been shown to be a non-linear function of exposure and minor junction frequency. In addition, it has been shown that the presence of minor junctions has an important influence on link accident frequencies. The best results were obtained using the empirical Bayes method.

In 1998, Mountain et al. (1998)developed models to predict the accident rates at junctions by taking into consideration the change in accident trends over time due to traffic growth and local or national road safety policies and programs. The data used in this study comprised details of highway and junction characteristics, personal injury accidents and traffic flows on the road networks, collected for periods between 5 to 15 years. Of the 501 junctions used, 96 were signalized intersections. The junction characteristics included number of arms and method of control, major road carriageway type and speed limit. The relevant information of the accident was its date, location, severity, road surface condition and lighting condition. Generalized Linear Models were developed for estimating expected number of accidents per year as a function of the accident risk and the major and minor road inflows. The trend for the change of traffic flows and national road safety programs and policies was incorporated in a separate model. It was found that effect road safety policies and programs result in a decline in accident risk from year to year. Many more conclusions were drawn on the trend of the variables. For example, it was found that the ratio of the dark to night accidents depends on the minor road flow.

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, the probability of fatal crashes is affected by wet road surface, use of drugs of alcohol, nearby interchange ramp, crash type, and the traffic volume.

In 2003, Wang et al. (2003) presented a new mechanism for predicting rear end accidents based on a probabilistic approach. Using the data from 115 intersections and 589 rear-end crashes, the occurrence of rear end crashes was studied considering the probability of encountering an obstacle and the probability of a driver failing to react quickly enough to avoid a collision with the obstacle vehicle. The probability of encountering an obstacle vehicle is assumed to be a function of the frequency of disturbances that cause the driver of a leading vehicle in a vehicle pair to decelerate. The probability of the trailing vehicle's driver failing to respond is the probability that this drivers' needed reaction time is less than the available reaction time. Hence the effect of a variable could be found on both the probabilities, giving a dual impact of the variables.

2.3 Nested Logit, Ordered Probit and Regression Tree Models

In an exploratory study, Shankar et al. (1996) attempts to develop a multinomial logit model for predicting the motorcycle-rider accident severity. The model uses a 5-year statewide data on single-vehicle motorcycle accidents from the state of Washington, that considers environmental factors, roadway conditions, vehicle characteristics and rider attributes. The study shows that the multinomial logit formulation is a promising approach to evaluate the determinants of motorcycle accident severity.

Shankar et al. (1996) developed a nested logit formulation as a means for determining accident severity given that an accident has occurred. The study involved collection of the following six categories of data from 61 km of study section of I-90 in Washington State: 1. individual accident data (primary identified causes, most severe consequences of the accident, time and location of the accident), 2. weather data, 3. geometric data (radii of horizontal curves, number of horizontal and vertical curves per kilometer, percentage length of horizontal curves) , 4. pavement surface data, 5. vehicle data, and 6. driver-related data (drug/alcohol usage by driver, age and gender of the drivers). Accidents that had occurred within a five year period were considered for estimating the four levels of severities: property damage only, possible injury, evident injury and disabling of fatal injury.

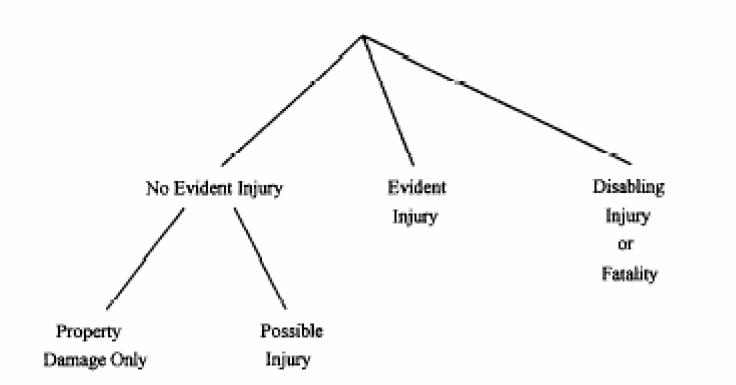

Figure 2.1 The most efficient Nested Logit structure developed by Shankar et al. (1996)

Among the various models tested statistically for the nested logit model, the model depicted in Figure 2.1 proved to be of the correct nested structure for accident severities. This diagram implies that the injury severity can be modeled in the form of two nests: an accident injury split up into no evident injury, evident injury and disabling/fatal injury; and the No evident injury split up into property damage only and possible injury. The variables were tested in both the nests and the effect of each variable on the injuries has been illustrated. For example, when the lower nest was tested, it was found that the overturn accident indicator played an important role and it had a greater probability of possible injury severity relative to the property damage only. Similarly, when this variable was tested for the upper nest, it was found that the variable had a greater probability of evident injury or disabling injury/fatality.

An ordered probit model was used by Quddus et al. (2002) to investigate how variations in various factors can lead to variations in the level of both injury severity and damage severity to motorcycles in motorcycle crashes. Crash data of 27570 accidents for the years 1992 to 2000 was collected in order to estimate the parameters in the ordered probit models. The results indicate that there are more severe injuries in the early morning periods and less severe injuries occurring during the day time. It was found that higher road design standards increase the probability of severe injuries and fatalities. But collision types of the accidents were included as input in the database that could have created a bias the database. For example, collision with pedestrians is almost always considered as a severe crash. Hence the database will be based on the collision type rather than any other variables.

Krull et al. (2000) used logit models to analyze driver injury severities for single-vehicle crashes. The authors analyzed three-year crash data from Michigan and Illinois in order to explore the effect of rollover, while controlling for roadway, vehicle, and driver factors. Results showed that driver injury severity increases with: (a) failure to use a seatbelt, (b) passenger cars as opposed to pick-up trucks, (c) alcohol use, (d) daylight, (e) rural roads as opposed to urban, (f) posted speed limit, and (g) dry pavement as opposed to slippery pavement.

Abdel-Aty (2003) analyzed driver injury severity levels for roadway sections, signalized intersections, and toll plazas in Central Florida using ordered probit models. The database used in the analysis consisted of variables related to the driver, vehicle, roadway and environmental conditions obtained from three counties in Central Florida. Results of the analysis showed that the older driver, male drivers and those not wearing a seat belt were most prone to severe injury crashes. The same was observed for drivers of passenger cars, vehicles struck at the driver's side, and drivers who speed. Variables related to the location of the crash like the roadway curves and dark lighting conditions were found to contribute to higher probabilities of injuries on roadway sections.

Nested models were also developed in this study to model injury severity. But ordered probit approach was found to produce better results than the multinomial logit approach, and was also considered simpler than the ordered probit model.

Although the Hierarchical Tree-Based Regression (HTBR) models have been used in many areas of transportation engineering like traffic planning to forecast trip generation (Washington and Wolf (1997)), they were used in traffic safety by Karlaftis et al. (2002) to quantitatively assess the effects of various rural road geometric characteristics on accident rates, and provide a mathematically sound way of predicting accident rates. The advantages of HTBR are that it allows for the quick estimation of predicted accident rates for a given rural road section and that it is easily amenable to 'if-then' statements for incorporation in expert systems.

The data used in the analysis is a combination of two databases: first consisting of road sections and their various traffic and geometric characteristics and the second containing the description of the location and type of accidents that occurred at these sections. This data was used in the HTBR model to predict the crash rate. The output comprised of tree shaped diagrams that can be transformed into a set of 'if-then' statements. The output for two-lane roads indicated that AADT was the most significant variable, lane width, serviceability index, pavement type and friction ratio were the other important variables affecting the crash rates. While the importance of lane width seemed to increase with higher flows, the importance of pavement condition factors seemed to increase with lower flows. For the multilane roads, the important factors were: AADT, median width, access control and pavement condition.

Although the paper demonstrates that HTBR can be used to find the most important factors in the crash rate prediction, the paper fails to mention if this model has been used on a test data to predict the crash rate so as to test the performance of the model on new data. The paper also does not mention the accuracy of the prediction of the crash rate.

More recently, HTBR model was used by Abdel-Aty et al. (2005) to determine the significant factors for different collision types and to determine if there was a difference between models based on complete and restricted datasets. Complete dataset is one in which all crashes are taken into account including the minor crashes with property damage only, whereas restricted dataset contains only major crashes reported as long forms in the state of Florida.

The authors chose the HTBR model primarily because the model does not need any assumptions or knowledge of the true functional form in advance. Also the model was used to take advantage of its robustness against multicollinearity between variables, handling missing values in the model and its ability to easily identify outliers.

The HTBR model developed to predict the frequency of crashes in each collision type. These models clearly indicated the factors that lead to increased accidents at signalized intersections. For example, the paper shows that for a complete dataset the factors affecting angle crashes are: number of left turn protected lanes on the major road, number of lanes on the minor roads, number of right turn channelized lanes on the major road, the traffic volume on the minor road, etc in that sequence. The consistency of these results for the complete and restricted datasets was compared in the study.

This study also had a testing phase where the number of crashes expected in 2002 was calculated for City of Orlando and Brevard County. In conclusion, the authors feel a need to develop models for predicting the frequency of crashes for each collision type instead of aggregating crash types to predict the total number of crashes. Also, the crashes reported on short-forms were found important while modeling the number of expected crashes.

2.4 Neural Network Models

A lot of papers have been published in the 1990's that deal with the application of neural networks in various areas of transportation. Neural networks have been used to predict driver behavior, pavement maintenance, vehicle detection/classification, freight operations, traffic pattern analysis, traffic forecasting, traffic operations, etc (Dougherty (1995)). Abdelwahab and Abdel-Aty (2001) discuss the classification of injury severities of accidents at signalized intersections into three levels (no injury, possible/evident injury, and disabling injury/fatality) using Artificial Neural Networks (ANNs). MLP Neural Network and Fuzzy ARTMAP Neural Network are the two ANN models have been used to classify the injury severities. These models have been compared to bring out a model that gives better classification accuracy. The 1997 accident data for the Central Florida area (Orange, Seminole and Osceola counties) was used in this study. The data consisted of accident characteristics and circumstances, information about the vehicles and vehicle maneuver before the accident, information on drivers and the condition or action of driver that contributed to the accident.

An MLP Neural Network was developed with nine input nodes, fifteen hidden nodes and three output nodes for the three injury levels. The number of hidden nodes was selected by running the model for 5 to 25 hidden nodes and selecting the number of nodes giving the best performance. All transfer functions used in the hidden and output layers were hyperbolic tangent sigmoid transfer functions. This model gave a classification accuracy of 65.6 and 60.4 percent for the training and testing phases, respectively. The model gave a classification accuracy of 63.7 percent for 1996 Central Florida accident data.

An Ordered Fuzzy ARTMAP Neural Network was also developed with 285 nodes in the ART_a module and 3 nodes in the ART_b module. It gave a generalized performance of 58.1 percent. Since the MLP Neural Network consisted of lesser number of nodes and gave a better performance, the authors concluded that MLP Neural Network has a better performance. The authors found that the MLP Neural Network performed better than the ordered logit model for the 1997 Central Florida accident data.

Hence the MLP Neural Network was found to have a promising potential in modeling injury severity.

The main objective of the work carried out by Abdelwahab and Abdel-Aty (2002) was to investigate the use of fuzzy Adaptive Resonance Theory MAP (ARTMAP) neural networks to analyze and predict injury severity of drivers involved in traffic accidents. Two accident databases have been used in this paper: one from the Florida Department of Highway Safety and Motor Vehicles (DHSMV) for the year 1996 through 1997, and the other from the Central Florida expressway system for the years 1999 and 2000. The latter database contained accidents that occurred in the vicinity of toll plazas. The authors developed a Fuzzy ARTMAP algorithm using a Visual C++ code. Since the order of pattern presentation affects the performance of the fuzzy ARTMAP training algorithm, the authors used three different orders of pattern presentation out of which two were random and one was an ordered pattern presentation. The data was ordered in the latter case using K-means clustering.

Three models were developed using Fuzzy ARTMAP. The first model was developed by training the 1997 accident data and testing over the 1996 accident data, for all accidents in the Central Florida region. Driver age, gender, alcohol, use of seat belt, vehicle type, point of impact, speed ratio, area type, lightning condition, and trafficway characteristics were found to be significant in predicting driver injury. The ordered version of fuzzy ARTMAP gave the best classification accuracy of 70.6%. The second model was developed for signalized intersections, with 1997 data used for the training phase and 1996 data used for the testing phase. The variables found significant were: driver age, gender, use of seatbelt, fault, vehicle type, point of impact, speed ratio and area type. The classification accuracy for this model was 58.1%. The third model was developed for the injury prediction in accidents around the vicinity of toll plazas. 2000 accident data was used in the training phase and 1999 accident data was used in the testing phase. The variables that were found significant were driver age, gender, payment method (electronic toll collection vs manual toll payment), plaza type (main vs ramp), use of seat belt, alcohol involvement, vehicle type, point of impact, number of impacts and weather condition. The model had a classification accuracy of 71.2%.

The authors also carried out a simulation experiment to extract knowledge from the trained network. Simulated input patterns were created using all possible combinations of input variables. The variables were plotted and the relationships between them were identified. Then these results were transformed into marginal effects to show the significance of an input variable on driver injury severity.

A more recent publication by Abdelwahab and Abdel-Aty (2004) compares the injury severity level prediction capability of a Multilayer Perceptron (MLP) Neural Network to the prediction capability of a fuzzy Adaptive Resonance Theory (ARTMAP) neural network and an Ordered Probit Model. The models are compared based on the 1996 and 1997 crash data of the Central Florida region consisting of Orange, Osceola and Seminole counties. The 1997 crash database was used in the training phase and 1996 crash database was used in the testing phase. 12 input variables were initially used in all the models. The number was reduced in each model based on the significance of the variables. For the neural networks, several runs were made on the training data set to prune the size of the inputs. One or more variables were removed at each run and the performance was compared to the complete model. Those variables were excluded that gave the best performance when excluded from the model. For the ordered probit model, the *t* tests and conditional likelihood tests were used to assess the goodness of fit of the reduced models against the full model.

The MLP neural network had a classification accuracy of 76.2 and 73.5% in the training and testing phases respectively. Peak period and weather were the insignificant variables in the model. The fuzzy ARTMAP had a classification accuracy of 70.6%. Peak period and weather were also found significant in this model. This was also true in the

Ordered Probit Model. Driving under influence was another factor that was insignificant in the Ordered Probit model, but its interaction with the seat belt factor was significant. The classification accuracy of the model was 62.6 and 61.7% in training and testing phases respectively.

Since the MLP neural network had a better classification accuracy and smaller network size compared to the fuzzy ARTMAP, it was concluded to be the better model for predicting the injury severity level. To compare the MLP neural network and Ordered Probit models, a test for the difference of two proportions was performed. MLP neural network showed better performance in this test and was hence declared to be the better of the two models. Hence the MLP neural network was found to be promising in modeling injury severity.

In 1999, Mussone et al. (1999) tried to identify the most significant parameters that determine the possibility of an accident occurring at an intersection by using an MLP neural network. The accident database consisted of 10 files containing information on location of the crash, vehicle information, driver information, injury severity, possible traffic violations of the driver of each vehicle, roadway conditions, visibility, weather and characteristics of vehicles and drivers, etc for all crashes that occurred in Milan from 1992 to 1995. An accident index was created for each intersection that was an indicative value for the degree of danger relative to the most dangerous intersection over the period of four years. The accident index was calculated as the ratio of the number of accidents at a particular intersection and the number of accidents at the most dangerous intersection. The authors selected intersections from a particular region and not from the entire city of Milan. 217 conflict points for the accidents on intersections were found out. The MLP model that was developed consisted of 10 input neurons and 4 neurons in the hidden layer. The transfer function in the input layer was linear and in the hidden layer was sigmoid. The Root Mean Square Error for this model was 18.24%. Multiple Linear and exponential regression techniques were also used to predict the accident index. These methods gave RMSE of 0.5 to 0.7.

Models have been developed using back propagation MLP neural networks to study the effect of intersection characteristics on numbers of intersection related traffic accidents, in a study conducted by Liu et al. (2004). A total of 28 different traffic and geometry related variables were collected for 62 signalized intersections for the years 2000 and 2001, that accounted for 1593 accidents during the 2 years. The crash details like the crash spot, date, time, illumination and weather condition of each crash were also recorded. The data was split based on the approaching directions of vehicles involved in the crashes, based on the Approaching Direction Combination (ADC).

The Back Propagation network developed in this study consisted of 53 input nodes, 22 hidden nodes, and 1 output node. The output obtained was very accurate for the test data. A sensitivity analysis was conducted to find the variables that had a greater influence on the crashes. A scheme for improvement of intersection deficiencies is proposed using the generated model. A case study is performed afterwards to examine the appropriateness of the proposed scheme.

Sayed et al. (1998) investigate the classification of road accidents using neural networks and fuzzy classification techniques. A feed forward back propagation neural network is used in the study to assign membership of accidents into three classes defined as the driver, the vehicle and the road. The database consisted of a detailed list of 900 accidents with each accident having around 21 variables associated with it, like the degree of curvature, road grade, speed limit, surface, weather and light condition, land use, accident time, location and type, severity, contributing causes of the accident, etc. The data was standardized to a (0, 1) range. The comparison of fuzzy classification technique to the neural network classifier showed that the neural network performed better. The neural network techniques have been compared with the fuzzy classification technique.

2.5 Geographical Information Systems (GIS) in Traffic Safety

Transportation professionals the world over have discovered and embraced GIS as an important tool in managing, planning, evaluating, and maintaining transportation systems. GIS has been used for diverse purposes from modeling travel demand 20 years in the future to tracking a snowplow; from analyzing the annual capital improvement plans to identifying noise regulation violations around airports; from improving transit service throughout rejuvenated urban centers to planning scenic byways in recreational areas. In transportation safety, the analytical capabilities of GIS support a variety of tasks, like crash location and reporting, incident and response management, accident analysis and "hot spot" identification, safety engineering and capital improvement, etc. Research is being carried out to study transportation safety from a geographic viewpoint, so as to relate the safety aspects with locational details. For example, Pawlovich (1998) presents a concept typology to organize the use of GIS, along with statistical techniques, to explore the relationship between crash incidence and underlying demographic, socioeconomic, and land use data.

To estimate the number of traffic accidents and assess the risk of traffic accidents in a study area, Ng et al. (2002) developed an algorithm that involves a combination of GIS techniques and statistical methods. The algorithm is developed as a four stage process: 1. GIS is used to locate accidents on a digital map; 2. Cluster analysis is used to group the homogeneous data together; 3. Regression analysis is performed to identify the relation between the number of accident events and the potential causal factors; 4. Accident risk is computed using the Empirical Bayes approach. A case study illustrates that the algorithm improves the accident risk estimation when compared to estimated risk based on only on the historical accident records. The algorithm is found to be more efficient, especially in the case of fatality and pedestrian-related crashes.

Kam (2002) presents a disaggregate approach to crash rate analysis. The approach involves combining two disparate datasets on a geographic information systems (GIS) platform by matching accident records to a defined travel corridor. As an illustration of the methodology, travel information from the Victorian Activity and Travel Survey (VATS) and accident records contained in CrashStat were used to estimate the crash rates of Melbourne residents in different age–sex groups according to time of the day and day of the week. The results show a polynomial function of a cubic order when crash rates are plotted against age group, which contrasts distinctly with the U-shape curve generated by using the conventional aggregate quotient approach. Owing to the validity of the many assumptions adopted in the computation, this study does not claim that the results obtained are conclusive.

The project carried out by Mistry et al. (2003) involves developing a new Geographic Information System (GIS) application for the display and analysis of crash data. Multi-year crash data from Tuscaloosa County are mapped on a commercially available base map, and these crash locations are compared with existing roadway features. After geocoding the base map with nodes, links, and route-milepost data, spatial analysis and "hot spot" identification is done using thematic mapping, buffering, and route impedance.

Using GIS, a methodology was developed by Abdel-Aty et al. (2000) to examine the association between driver characteristics and traffic crash involvement. The aims of the study were to identify areas in the state of Florida that have high crash rates and provide drivers there with suitable educational programs to improve their safety behaviors and enhance their knowledge of traffic safety problems. Two conventional driver characteristics were investigated in this research: driver's age and gender. Income level was also investigated. Data and variables from the 1995 Florida crash database and census data were used in the analysis. Results showed a strong relationship between income level and crash involvement while under the influence of alcohol/drugs, and crash involvement when seat belts were not used. Male drivers had higher crash rates than females, and teenagers are riskier drivers than the elderly.

2.6 Summary

The methodologies used in the studies described in this chapter have proved to be an invaluable tool in predicting and modeling the frequency of crashes. Although a lot of research has been performed in improving the safety of the highways in general, not many studies have concentrated on the safety of signalized intersections. There have been a plethora of studies carried out using the statistical negative binomial and Poisson models. But the applications of the recent tools of neural networks, regression trees and GIS has been limited in this field. Hence the present study will try to utilize these methodologies to predict the severity and collision types of crashes at signalized intersections in Florida.

3 METHODOLOGY

3.1 Introduction

The main objectives of this thesis are to classify and predict crashes at signalized intersections using the data on traffic and geometric properties of intersections and the properties of the crashes. The studies by Abdelwahab and Abdel-Aty (2002), Abdelwahab and Abdel-Aty (2004) and Mussone et al. (1999) indicate that the classification and prediction of crash parameters can be achieved efficiently using Artificial Neural Networks (ANN). Therefore, this chapter discusses the various neural network methods that will be used in the analysis phase of this thesis.

3.2 Artificial Neural Networks

According to According to Nigrin (1993): “A neural network is a circuit composed of a very large number of simple processing elements that are neurally based. Each element operates only on local information.” One of the advantages of using ANNs, as described by Haykin (1999) is that it can perform massive computations through its massively parallel distributed structure and its ability to learn and generalize. Neural networks also possess the ability to produce reasonable results by adapting themselves to the inputs that are not encountered during its training. ANNs can adapt themselves to changes in the input variables by adjusting their weights. Thus they can perform well even under a variation of input variables for which they haven’t been trained. Nonlinearity is an important characteristic of an ANN as it can nonlinearly map input variables to output variables. The neural networks are also considered to be fault tolerant because their performance falls gracefully under adverse operating conditions.

The Multi Layer Perceptron (MLP) and Probabilistic Neural Networks (PNN) were used for the classification analysis, and MLP and Generalized Regression Neural Networks (GRNN) were used for the prediction of crash frequencies.

3.3 Multi Layer Perceptron (MLP) Neural Network Architecture

MLP neural networks are an important class of neural networks and are very widely used. Typically, an MLP neural network consists of a set of source nodes that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computation nodes. A descriptive diagram of the MLP neural network is given in

Figure 3.1. The input signal propagates through the network in a forward direction on a layer-by-layer basis. By adding one or more hidden layers, the network is enabled to extract higher ordered statistics. The number of output nodes depends on whether the MLP is being used for classification or prediction. In the classification phase, the number of output nodes is typically equal to the number of classes the data is split into, whereas only one output node is required in the prediction phase.

The MLP shown in Figure 3.1 has K input nodes, J hidden nodes and I output nodes, and ***w*** represents the weight functions. The nodes at the input layer of the MLP supply the respective inputs of the activation pattern to the hidden layer. The output of the hidden layer is again transferred to the output layer as input, and the activation pattern of this forms the output. The nodes in the MLP neural network transform their input by using a scalar-to-scalar function called the activation function. The commonly used activation functions are the sigmoid (or the tanh function), logistic (1/(1+exp(-x)))) and the linear functions.

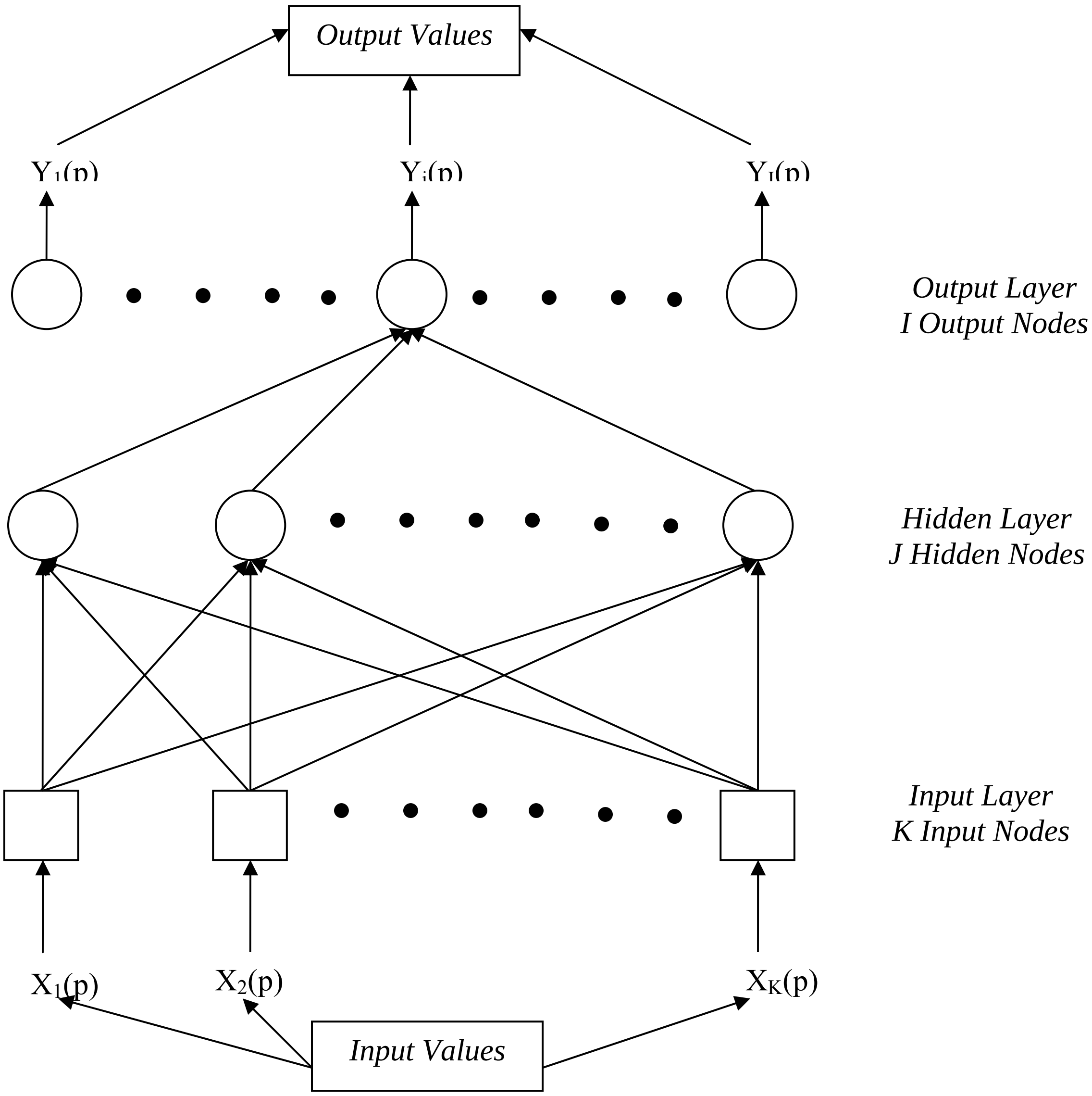

Figure 3.1 Multi Layer Perceptron Feedforward Network

The MLP neural network has been used to solve complex problems using the back-propagation algorithm. This algorithm consists of two passes through the different layers of the neural networks: a forward pass in which the input is applied at the input layer and the output is produced as the actual response of the network, and a backward pass in which all weights are adjusted according to the error correction rule. According to this criterion, the actual response is subtracted from the target outputs to obtain the error signal. The error signal is propagated backwards so that the weights can be adjusted accordingly. Hence the algorithm gets its name of back-propagation algorithm.

The aim of the training phase of the MLP neural network is to map a given set of inputs in the training data, say x(1), x(2)…. x(PT), to the output values in the training data, say d(1), d(2),… d(PT) respectively. Hence the input x(p) has to be mapped to the output d(p). For this purpose, the following error function is constructed:

$$E^{p}(w) = \frac{1}{2}\sum_{i=1}^{I}[d_i^2(p) - y_i^2(p)]^2$$

The objective is to change the weights ***w*** so that the error function is minimized, which means that the actual output is being made as close as possible to the desired output. The error is back-propagated through the neural network to adjust for the weights between the layers. The error function is minimized using the gradient descent procedure that changes the weight vector ***w*** by an amount proportional to the negative gradient of the function *E(w).* Using detailed calculations, Georgiopoulos and Christodoulou (2001) determine the amount by which the weights in each layer can be changed so as to minimize the weight functions.

The error is minimized until a stopping criterion is met. The stopping conditions usually set are that the number of epochs (or presentations of the inputs) does not exceed a certain value, or the error function becomes sufficiently small (Georgiopoulos and Christodoulou, 2001).

3.4 Probabilistic Neural Networks (PNN)

The probabilistic neural network (PNN) was developed by Donald Specht. This network provides a general solution to pattern classification problems by following an approach of the Bayesian classifiers. The network paradigm also uses Parzen Estimators which were developed to construct the probability density functions required by Bayes theory.

Chen (1996) states that in order to classify a variable into one of two classes based on a set of measurements represented by a p-dimensional vector $\mathbf{X}^{\mathbf{t}}$ the two category decision surface Baye's criteria can be arbitrarily complex. This is true even with a multi-category classification. The key to using the Bayes classifiers is the ability to estimate PDFs based on training patterns. Parzen showed that a class of PDF estimators asymptotically approaches the underlying parent density provided it is continuous. Therefore the accuracy of the decision boundaries depends on the accuracy with which the underlying PDFs are estimated. Parzen showed that a family of estimates of f(x) can be constructed using the formula:

$$f_n(x) = \frac{1}{n\sigma}\sum_{i=1}^{n} W\left(\frac{x - x_i}{\sigma}\right)$$

which is consistent at all points *X* at which the PDF is continuous. This was extended to the multivariate case where the multivariate estimates can be expressed as:

$$f_k(X) = \frac{1}{(2\pi)^{p/2}\sigma^p}\frac{1}{m}\sum_{i=1}^{m}\exp\left[\frac{-(X - X_{ki})^T(X - X_{ki})}{2\sigma^2}\right]$$

where k = category

i = pattern

m = total number of training patterns

X_{ki} = *i*th training pattern from category *k*

σ = smoothing parameter or spread

p = dimensionality

The smoothing parameter σ defines the width of the bell curve that surrounds each sample point. The only parameter that has to be adjusted for is the spread.

The PNN uses a supervised training set to develop distribution functions within a pattern layer. These functions are used to estimate the likelihood of an input feature vector being part of a learned category or class. The learned patterns can also be combined with the a priori probability of each category to determine the most likely class for a given input vector.

The structure of the PNN has been shown in
Figure 3.2. The input nodes provide the same input values to the nodes in the pattern layer. Each pattern unit forms a dot product of the input vector X with the weight vector Wi: Zi = X * Wi, and then performs a nonlinear operation on Zi before outputting its activation level to the summation unit (Chen, 1996). Instead of a sigmoid function commonly used for backpropagation, the nonlinear operation used in PNN is $\exp[(Zi - 1)/\sigma^2]$. Both X and Wi are normalized to unit length which is equivalent to using the probability density function:

$$F(X) = \exp(-(\mathbf{W}_i - \mathbf{X})^t(\mathbf{W}_i - \mathbf{X})/2\sigma^2)$$

Where *i* is the pattern number, **X** is the training pattern and σ is the smoothing parameter or the spread. The network is trained by setting the Wi weight vector in one of the pattern units equal to each of the X patterns in the training set and then connecting the pattern unit's output to the appropriate summation unit. A separate neuron (also called pattern unit) is required for every training pattern. The same pattern units can be grouped by different summation units to provide additional pairs of categories and additional bits of information to form the output vector.

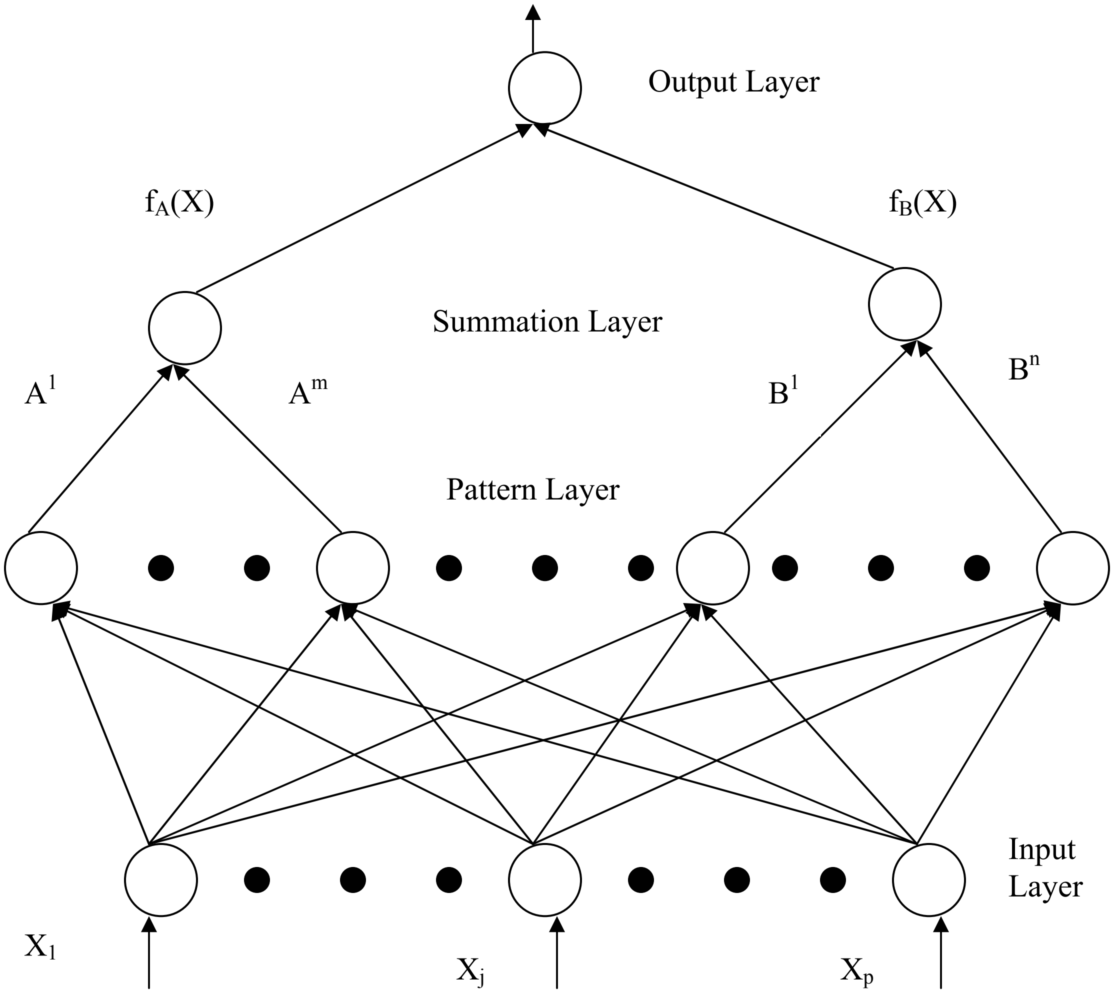

Figure 3.2 Structure of a PNN

3.5 Generalized Regression Neural Network (GRNN)

A GRNN provides estimates of continuous variables and converges smoothly to underlying linear or nonlinear regression surface. Like PNN, a GRNN features instant learning and a highly parallel structure. GRNN provides smooth transition from one observed value to another even with sparse data in multidimensional measurement space. The GRNN can also be used for regression problems where an assumption of linearity is not justified.

GRNN uses Parzen's estimators along with a joint continuous probability density function. The conditional mean of y given $\mathbf{X}$ is given by

$$E[y \mid \mathbf{X}] = \frac{\int_{-\infty}^{\infty} y f(\mathbf{X}, y) dy}{\int_{-\infty}^{\infty} f(\mathbf{X}, y) dy}$$

For nonparametric estimate of $f(\mathbf{x}, \mathrm{y})$, the Parzen's estimator can be used. This leads to the equation (Chen, 1996):

$$\hat{Y}(\mathbf{X}) = \frac{\sum_{i=1}^{n} Y^i \exp[-\frac{(\mathbf{X}-\mathbf{X}^i)^t(\mathbf{X}-\mathbf{X}^i)}{2\sigma^2}]}{\sum_{i=1}^{n} \exp[-\frac{(\mathbf{X}-\mathbf{X}^i)^t(\mathbf{X}-\mathbf{X}^i)}{2\sigma^2}]}$$

The estimate $\hat{Y}(\mathbf{X})$ can be visualized as a weighted average of all the observed values Y^{i}, where each observed value is weighted according to its Euclidean distance from **X.** When σ becomes large, $\hat{Y}(\mathbf{X})$ assumes the value of the sample mean of the observed Y^{i}, and as σ tends to 0, $\hat{Y}(\mathbf{X})$ assumes the value of Y^{i} associated with the observation closest to **X**. For intermediate values of σ, all values of Y^{i} are taken into account, but those corresponding to points closer to **X** are given heavier weight.

The structure of the GRNN is shown in **Error! Reference source not found.**. This network estimates vector **Y** from measurement vector **X**.

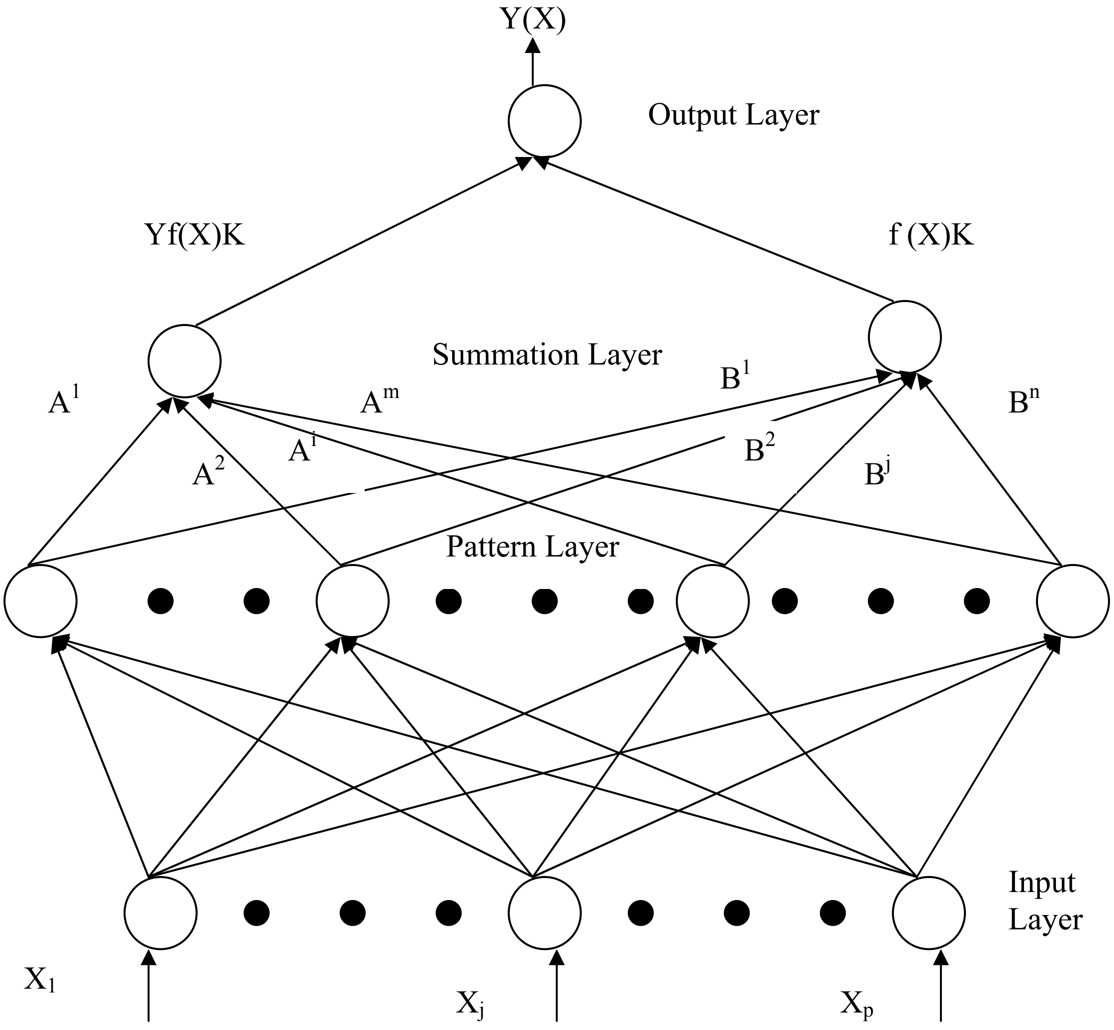

Figure 3.3 Structure of GRNN

The first two layers are identical to the PNN. The summation node performs a dot product between a weight vector and a vector composed of the activations from the pattern layer. The summation node generates the estimate of *f*(**X**)K that sums the outputs of the pattern layer weighted by the number of observations each cluster center represents. The summation node that estimates $\hat{Y} f(\mathbf{X})K$ multiplies each value from a pattern node by the sum of the samples Y^j associated with the cluster center X^i. The output unit divides $\hat{Y}\ f(\mathbf{X})K$ by $f(\mathbf{X})$K to yield the desired estimate of Y.

3.6 Summary

This chapter has briefly described the methodologies that will be used in the classification of crashes and the prediction of crash frequencies. Artificial Neural Networks (ANNs) will be used in the study as they possess a lot of advantages over other methods like their ability to efficiently handle non-linear problems, their adaptivity to new data, their efficiency in performing massive calculations, and their fault tolerance. The theory and the working mechanism of the MLP, PNN and GRNN neural networks have been discussed.

4 DATA COLLECTION AND CLASSIFICATION

4.1 Introduction

The analysis and results of any project are a reflection of the type of the data used in the project. The data collected should be appropriate and abundant so as to meet both the qualitative and quantitative requirements of a project. This means that efforts have to be made to collect as much quality data as possible, and this data should be useful in a variety of ways to the project. This has been carefully considered while collecting data for the present project, and this chapter describes the various types of data collected and the efforts put in to collect the data.

4.2 Collecting Data for Six Counties

Data was collected from six counties: Brevard, City of Orlando, Hillsborough, Miami-Dade, Orange and Seminole. Data pertaining to various intersections and crashes occurring at these intersections were collected in different formats from each county.

Data collected for the counties was divided into two parts: the geometry database containing all the intersection characteristics, and the crash database containing the details about crashes. To develop the geometry database, CAD files or aerial pictures of the intersections were obtained from each County so as to identify the intersections' configuration. Not all of these files were clear, and so a field visit was needed in many cases to identify their configuration. The data collected pertaining to the geometric characteristics of the intersections includes number of through, left, and right lanes for each approach, presence of channelization at each approach and the presence of median for each approach. Also, the data on the speed limit, traffic volume (AADT) and K-factors for each approach was incorporated in this database.

Different sources were used for developing the crash database, namely the county mailed/handed files, county websites, Department of Highway Safety & Motor Vehicles (DHSMV) data, photocopied crash reports, and F-DOT websites that included the SSO Online Document Retrieval System and the Crash Analysis and Reporting (CAR) database in the FDOT Mainframe. It is important to note that every county saves their data in different ways. There is inconsistency among counties in the way they keep data, which posed a challenge to obtain complete data from each county and maintain uniformity among counties as much as possible. The contents of the crash database have been listed in Table 4-1. Most of these data was available for all counties.

In the crash database, the crashes were sometimes labeled as occurring at an intersection when they actually occurred up to a mile away from the intersection. To be consistent with the FDOT's definition of an intersection related crash, only the crashes occurring at a radius of 250ft around the intersection were selected as intersection related crashes. Therefore, any crash listed as occurring over 250 feet from an intersection was not included in the crash database.

Although efforts were made to collect the maximum amount of data possible for signalized intersections from all counties being considered in this study, not all of the data could be collected for all of the counties. The data collection efforts from each of the counties have been described in the following subsections.

Table 4-1 Format of Crash Database

	Field #	Field Caption
	1	crash report number
Intersection Data	2	node number
	3	intersection (routes names)
	4	AADT
	5	type
	6	category
	7	Speed Limit
	8	K-Factor
Crash Data	9	crash date
	10	time of crash
	11	county code
	12	city code
	13	number of lanes
	14	divided/undivided highway
	15	total property damage
	16	investigating department
	17	fist harmful event
	18	subsequent harmful event
	19	road system identifier
	20	location type
	21	lighting condition
	22	road surface condition
	23	weather
	24	road surface type
	25	1st contributing cause-road
	26	2nd contributing cause-road
	27	1st contributing cause-environment
	28	2nd contributing cause-environment
	29	1st traffic control
	30	2nd traffic control
	31	site location
	32	trafficway character
	33	type of shoulder
	34	state road crash
	35	day of week
	36	rural/urban
	37	crash injury severity
	38	alcohol/drugs
	39	total number of vehicles
	40	total number of fatalities
	41	total number of injuries

4.2.1 Orange County

Data was first collected for Orange County. Signalized intersection drawings were obtained from the county's traffic engineering department. From these drawings, a geometry database was created that contained intersection characteristics. In addition, several other geometric characteristics were collected from these drawings and input into the database. Due to the fact that the drawings were not always consistent, Orange County was contacted again for more information. Through their help, complete geometric characteristics were obtained for the signalized intersections in Orange County. Information received included intersection drawings and several signal time sheets and turning volumes.

While continuing the efforts on building the geometry database, new intersections were identified based on the level of service report published by the county in an effort to collect AADT volumes and k-factors for the new intersections. Next, available turning volumes and signal timings were associated with the appropriate intersections. Finally, a new geometry database was created reflecting the most complete data.

As a next step, efforts were made to identify all intersections that underwent construction during the years 1999 and 2000. If an intersection was under construction during a year it would be excluded from analysis for that particular year.

Based on the available information, the intersections were classified based on the number of lanes on the major and minor road (i.e. 2x2, 4x2, 4x4, 6x2, 6x4, and 6x6). Some intersections contained Two Way Left Turning Lanes (TWLTL), and were represented as 3x2, 4x3, 5x2, 5x3, 5x5, etc. These intersections were considered in lane configurations without the TWLTL, i.e., 3x2 was considered in 2x2, 4x3 was considered in 4x2, 5x5 was considered in 4x4, etc. Since there were a significant number of T-intersections, they were further divided as per lane configuration into 2xT2, 4xT2, 4xT4, 6xT2, etc.

The crash database for the Orange County was developed for the years 1999 and 2000. It was not possible to retrieve records from 2001 onwards because the county began coding their records in a manner different from that of FDOT crash database, while using a numbering system different from the crash report numbering. As Orange County does not keep a record of the short form crash reports, only the long form crashes were collected.

In the crash database developed for the years 1999 and 2000, several crash records were found missing. In order to remedy this problem, our team visited the Orange County Public Works department for a total of four days and was able to make photocopies of about 500 crash records from 1999 and 2000. This ensured that the database was complete.

Another database was then created in Access to input the data from the crash reports as well as all roadway geometry from the previous database. An Access program was written to collect the required information from the crash reports. In an effort to account for all crashes and to ensure that the final crash database was as accurate and complete as possible, the county, FDOT and DHMSV databases were cross-checked. This ensured the completeness of our data as each of the databases was found to be missing some crash reports.

A SAS program was written to match the crash report number in the crash database to the crash report number in the DHSMV database, and then to extract the information on the hour of the day, day on the week, month, light conditions, surface conditions, severity and collision type of the crash. Collision type was categorized into rear end, head on, angle, left turn, right turn, sideswipe, pedestrian, fixed object and other collisions. Injury severity was subcategorized into property damage, possible injury, non-incapacitating injury, incapacitating injury and fatal injury. Light condition was branched off to daylight, dusk, dawn, dark (with street lights), dark (without street lights) and unknown. Weather condition was sub divided into dry, cloudy, rain, fog, others and unknown. The fifth category, Surface conditions, was separated into dry, wet, other and unknown. Months of the year consisted of months January through December. Each day of the week was a separate category and time of the day was divided into seven groups. The groups consists of 00:00-06:00, 06:01-09:00, 09:01-11:00, 11:01-13:00, 13:01-15:00, 15:01-18:00, and 18:01-24:00. Using this method, a large amount of the data related to the crash was collected. Similar methods were adopted to extract these variables for the crashes in the other five counties.

After collecting the data, the traffic and geometric characteristics of every intersection were combined along with the information of all the crashes that had occurred at or influenced by that intersection. The following steps were followed during this process:

1. The database containing the crashes contained the Crash Report Numbers (CRN) of the long form crashes for the years 1999 and 2000. The names of the intersecting roads are available, no node number is provided for the intersection.

2. The DHSMV databases for the years 1999 and 2000 were used to extract the information of the above crashes. The CRN was used to link the Orange county Excel spreadsheet with the DHSMV databases.

3. All these missing CRNs were photocopied from the original crash reports to complete the DHSMV databases for Orange County.

4. The crash data developed in this phase was crosschecked with the FDOT Mainframe’s CAR database and the missing crashes were added.

5. A unique node number for every intersection was generated for further use.

6. Using CAD drawings for every intersection, the research team developed a database that has the geometric characteristics of each intersection and its unique node number. This job was done manually for each intersection.

7. Using the Orange county traffic reports posted on their website, a database was developed that had the traffic characteristics of each intersection and its unique node number.

8. A SAS code was written to read the above databases and combine them in the master database of Orange county;

    a. The CRN was used to link Orange county Excel spreadsheet with DHSMV databases to produce a dummy database.

    b. The intersecting street names were used to link the dummy database to the geometric and traffic databases to produce the final master database of Orange County.

4.2.2 Seminole County

The website of Seminole County was first reviewed for information such as traffic counts based on the type of roadways. The county was then contacted directly to get additional information. The County provided a list of signalized intersections as well as a CD containing partial intersection geometry and signal details. A geometry database was built for Seminole County where each intersection was classified based upon the number of through lanes. Other geometric information was also available and input into the database as well. Using electronic drawings on Excel spreadsheets for the intersections, a database was developed containing all the geometric characteristics of each intersection and its unique node number. This job was done manually for each intersection.

A unique node number was assigned to every intersection for further use. Using the Seminole county traffic reports posted on their website, a database was developed that contained the traffic characteristics of each intersection and its unique node number. Approach speed limits at the intersections were obtained from the CD. For the intersections for which these values were unavailable, they were obtained by driving on the roadways and noting the speed limits manually.

To make Seminole County more compatible to Orange County for a more accurate comparison, the roadway k-factor values were searched for Seminole County because this information was readily available for our Orange County intersections. Seminole County k-factors were found on the Florida Department of Transportation's website for state roads only and this information was then input into the geometry database.

Next, crash records were obtained for the county. The database contained the following data:

1. The Crash Report Number (CRN) for both long- and short-form crashes for the years 1999, 2000, and 2001.
2. The crash information, similar to DHSMV data and format.
3. The names of the intersecting roads.
4. No node number was provided for the intersection.

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.

A SAS code was written to read the above databases and fuse them in the master database of Seminole county. The intersecting street names were used to link the Access database to the geometric and traffic databases to produce the final master database of Seminole county.

4.2.3 Hillsborough County

Hillsborough County officials provided a CD containing aerial photographs and field drawings for some of the signalized intersections in the county. Again, the intersections were classified by lanes and included any other information that could be gathered in the geometry database. During the process of collecting the county's information, several items were found missing and it became necessary to meet with the county officials directly. One member of the research team was sent to the main office in Tampa for two days in an effort to retrieve all the possible data.

Hillsborough County did not provide any AADT counts, so they were located on the website in the form of a spreadsheet. The format made the extrapolation of the necessary information very difficult. In order to use this spreadsheet, all intersections had to be located on a map of the county and their location was found relative to the locations where AADT counts were measured. It took several weeks to complete this process. When finished, these AADT values were compared to the AADT values at comparable intersections in both Orange and Seminole counties. It was then evident that the AADT values reported on the Hillsborough spreadsheet were an inaccurate representation of the actual street volumes because the Hillsborough AADTs were considerably lower than those in Orange and Seminole counties. It was decided that a more credible source was needed for these counts. After searching the Internet, an up-to-date level of service report was found that not only included the AADT and level of service but also the number of lanes on the roadway as well as whether it was divided or not. Upon comparison of the previously used spreadsheet, these numbers were found to be more accurate especially since roadways with relatively low AADT were graded with a better level of service. In addition to replacing the erroneous AADT values in the geometry database, all of the streets were checked to ensure that they were consistent in the number of lanes and roadway division with the official level of service report.

Another task for Hillsborough County included identifying all intersections that went under construction during our data period, 1999, 2000, and/or 2001. If an intersection was under construction during a year it would be excluded from analysis for that particular year. Modification information was received in spreadsheet form from the county. There were a total of 12 intersections that were to be excluded from at least one year's analysis.

The crash data was then downloaded from the county's ftp site, which included both long and short forms for years 1999 to 2001. Another code was written to extract each crash individually into an Excel spreadsheet, which would allow for much easier manipulations. When this task was finished, each intersection listed in the crash file was manually reviewed and the unique county number was attached to the ones that had been included in the geometry database.

The next step taken was to associate the available intersections with their respective crashes occurring between 1999 and 2001. Crash information was downloaded from the county's FTP site and included both long and short forms with the type clearly stated. To link these crashes, an Excel spreadsheet was created with the different spellings of each intersection as well as the intersection's unique county number that was assigned to the intersection. Then a SAS code was written to perform two tasks; first, to associate the crashes with the link from the Excel spreadsheet and, second, to use the link again to associate crashes with their respective geometry information. Upon completion, a master database was created for Hillsborough County and was crosschecked with the FDOT Mainframe and DHSMV database to ensure completeness.

4.2.4 City of Orlando

Two CD-ROMs were obtained from the City of Orlando, one containing intersection geometry and signal timing details for 355 intersections in the City of Orlando, and the other containing the crash specifications at every intersection. About one-third of the intersections in the database consisted of one-way streets.

Of the 355 intersections in City of Orlando that were received in the drawings from city officials, geometry characteristics could be collected for most of them. The speed limit values for the approach roadways were collected from Internet sources. However, of those the AADT values were known for only 171 intersections. Due to the fact that most of City of Orlando's intersections are nearby at least one other intersection, AADTs for intersections missing this information was interpolated using the two nearest intersections. This was done by locating each intersection on a city map and then locating the next two closest intersections. If the nearby intersections both had AADT counts, then the missing intersection's AADT would be the average of these actual AADTs. This process turned out to be particularly tedious but worthwhile because AADT could be identified for 124 more intersections, increasing the number of intersections for City of Orlando to 295.

A geometry spreadsheet was created for the city and the intersections were classified in the same way as for the aforementioned counties.

The crash details for City of Orlando were obtained in the form of an Access database, in a similar format as for the Seminole County. This database contained crashes for the years 2000, 2001 and 2002. The crash list included both long and short forms for the years 2000 to 2002. A SAS code was written to match crashes with the intersection's characteristics by way of a unique number that was assigned to each intersection. This database generated was crosschecked for completeness and accuracy.

4.2.5 Brevard County

Brevard County was originally contacted for cooperation and was able to provide hand-drawings for a lot of intersections. Each drawing was categorized and information was recorded into a geometry database showing each intersection. When this was complete, intersection AADT information was found from the Internet and the database was updated.

After completing the geometry database for the intersections, Brevard County was contacted again for a crash list. The county provided an Excel spreadsheet listing each long and short-form crash for the years 2000 to 2002. A code was written to extract crashes and the unique county numbers were attached to all of their locations. Some additional crashes were added to the database obtained from the FDOT Mainframe's CAR database. The next step was to use the county numbers attached to the crashes to match them individually to the intersection the crash occurred at based on the geometry database and create another master county database as was done in other counties.

4.2.6 Miami-Dade County

Several CDs were obtained from the county containing geometric information for a total of 3200 intersections. Upon looking into these, it was found that many intersections were not signalized, some were signalized pedestrian crosswalks, and others were mechanical bridges. Also, crash records could not be retrieved from the county. Therefore all crashes had to be downloaded from the FDOT Mainframe's CAR database. The FDOT database reports long-form crashes from state roads only. Hence 1501 state road drawings were identified from the 3200 that the county had sent. Of these intersections the geometric information was recorded for 580 state-road intersections. This information only included size of the intersection, e.g., number of left turn lanes, roadway median type and whether there the right turn was channelized. The database contained no information on AADT, k-factors or speed limits. The county was unable to provide any more information. Since the roads in the database were only state road intersections, the AADT and k-factors were extracted from the FTI2003 CD-ROM from FDOT.

The crash list was obtained from the FDOT mainframe database by specifying the intersection in the mainframe and specifying the time period for which the crashes had to be downloaded. This procedure proved to be time consuming but worthwhile because 28,380 crashes were downloaded for 413 intersections in Miami-Dade County for the years 1999, 2000, 2001 and 2002. These crashes were extracted from the FDOT crash database were crosschecked with the DHSMV database to make sure that the database was consistent.

The final step was to link the crashes to their respective intersections and geometry information. This was done by writing a SAS program to join the intersections to the crashes by the intersection ID common to both the geometry and crash databases. The method was similar to the one used in the joining the geometry and crash files in the other counties. Thus the master database developed included crash and geometric information from 413 intersections in Dade County.

4.2.7 Summary of the Databases

The data collected for each of the six counties consisted of geometry and crash databases in different formats. These databases were combined to form a Master Database that contained all the characteristics of a particular crash as found in the crash database, and also the intersection characteristics as found in the geometry database. This step provided one final database for each county.

One important aspect that came up while building the Master Database was whether to include short form crashes in the database. By the best of efforts, a database of both long and short form crashes was developed for City of Orlando, Brevard, Hillsborough and Seminole counties. Although the Orange County crash databases were obtained from the county, Orange County does not keep a record of short forms. As for Miami-Dade, the county was unable to provide the crash database and thus the crash database was downloaded from the FDOT sources that contained only long form crash records for intersections with at least one road being a State Road. Therefore, except for Orange and Miami-Dade Counties, all other counties contained crash databases consisting of both long and short forms. It was decided upon to include these crashes, because, there will be a consistent under reporting some types of crashes (such as PDO crashes, which tend to be rear-end in many cases) if they are not accounted for. Hence the Master Databases contained both long and short form crashes for the four counties: Brevard, City of Orlando, Hillsborough and Seminole, while they contained only long form crashes for Miami-Dade and Orange counties. Since FDOT was only interested in long-form reported crashes, the focus of the project has been on these crashes. But detailed records of all types of crashes were included.

The complete summary of the Master Databases of all six counties has been tabulated in Table 4-2.

Table 4-2 Summary of Data in all Six Counties

	Intersection Type	Includes Types:	Brevard	Hillsborough	Orange	City of Orlando	Seminole	Miami-Dade	Sub-total	Total
Classification	6 x 4	8 x 6		1					**1**	**90**
		8 x 4		5				2	**7**	
		6 x 6		1	4	1	1	5	**12**	
		6 x 5				1		1	**2**	
		6 x 4	2	8	24	5	6	23	**68**	
	6 x 2	6 x 3			5	3		9	**17**	**158**
		7 x 3					1		**1**	
		8 x 2		6			2	2	**10**	
		6 x 2	7	16	27	14	19	47	**130**	
	4 x 4	5 x 4			1	1	1	6	**9**	**158**
		4 x 4	11	16	40	23	18	41	**149**	
	4 x 2	4 x 3	2	1	14	6		6	**29**	**541**
		5 x 2			1	3		4	**8**	
		4 x 2	76	50	109	90	60	119	**504**	
	2 x 2	3 x 3				1			**1**	**175**
		3 x 2				4	1		**5**	
		2 x 2	17	40	36	33	30	13	**169**	
	SubTotal		115	144	261	185	139	278	**1122**	
	3-Legged		24	32	36	41	55	61	**249**	**249**
	One Ways/Ramps		12	15	15	69	6	74	**191**	**191**
	Total		**151**	**191**	**312**	**295**	**200**	**413**	**1562**	
	AADT		Yes	Yes	Yes	Yes	Yes	Yes		
Data Availability and sources	k Factor		No	Yes	Yes	No	No	Yes		
	Speed Limit		Yes	Yes	Yes	Yes	Yes	Yes		
	Turning Volumes		23 sites only	No	Yes	No	No	No		
	Signal Timings		No	No	Yes	No	No	No		
	Modification dates		No	Yes	Yes	No	Yes	No		
	Crash Years		00,01,02	99,00,01	99,00	00,01,02	99,00,01	99,00,01,02		
	Crash Source(s)		Excel file from County	County FTP Site	County, FDOT Site & Copies	CD from the city for the 3 years	Access file from County	FDOT Site		
	Number of crashes		1486	4651	3616	5764	2527	28380		
	Master-Database		Done	Done	Done	Done	Done	Done		

4.3 Classification of Intersections

The intersections were classified into various groups in order to study the crash patterns. In order to identify the best AADT values to classify intersections of a particular configuration, the AADT/major-lane values were tabulated for the intersections of Orange County. Orange County was chosen for the analysis because it was the first county to have a complete database. These tabulated AADT/lane values were plotted as frequency and cumulative graphs, as shown in Figure 4.1 and Figure 4.2 for 4 x 2 intersections. Similarly, frequency and cumulative frequency plots were plotted for each type of intersection for AADT/lane for minor roads and entire intersection. Figure 4.3 and Figure 4.4 show such frequency plots for 4 x 2 intersections.

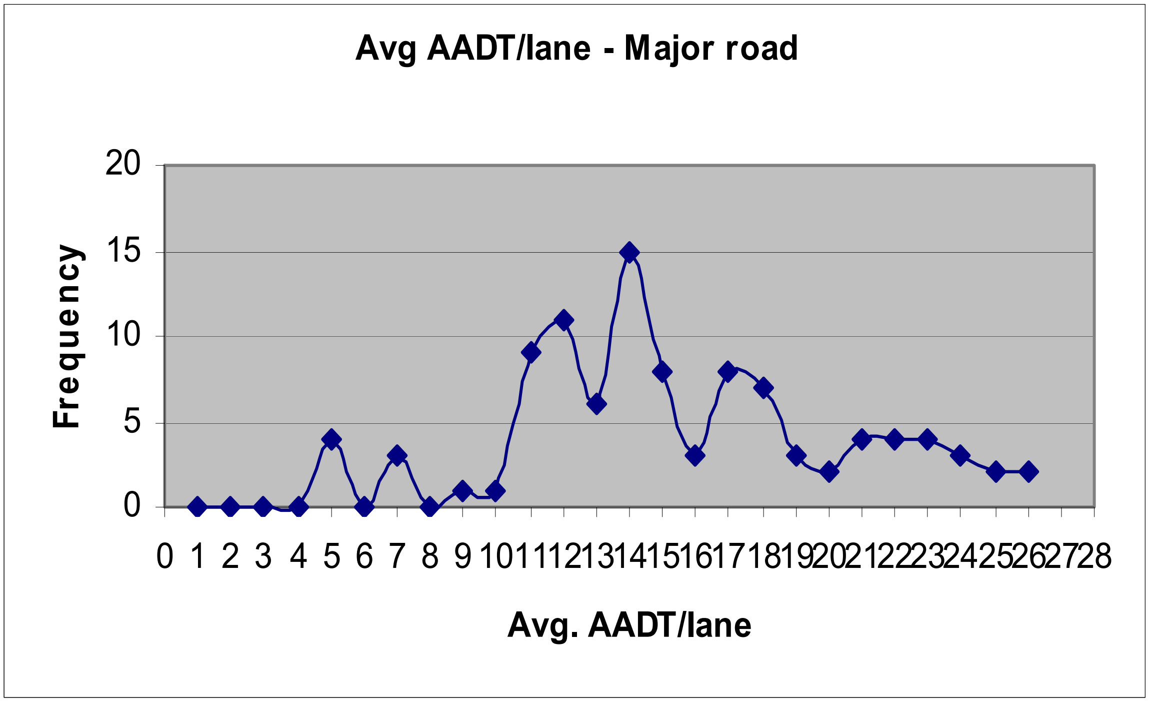

Figure 4.1 Frequency plot for Avg. AADT/Major-lane for 4 x 2 intersections

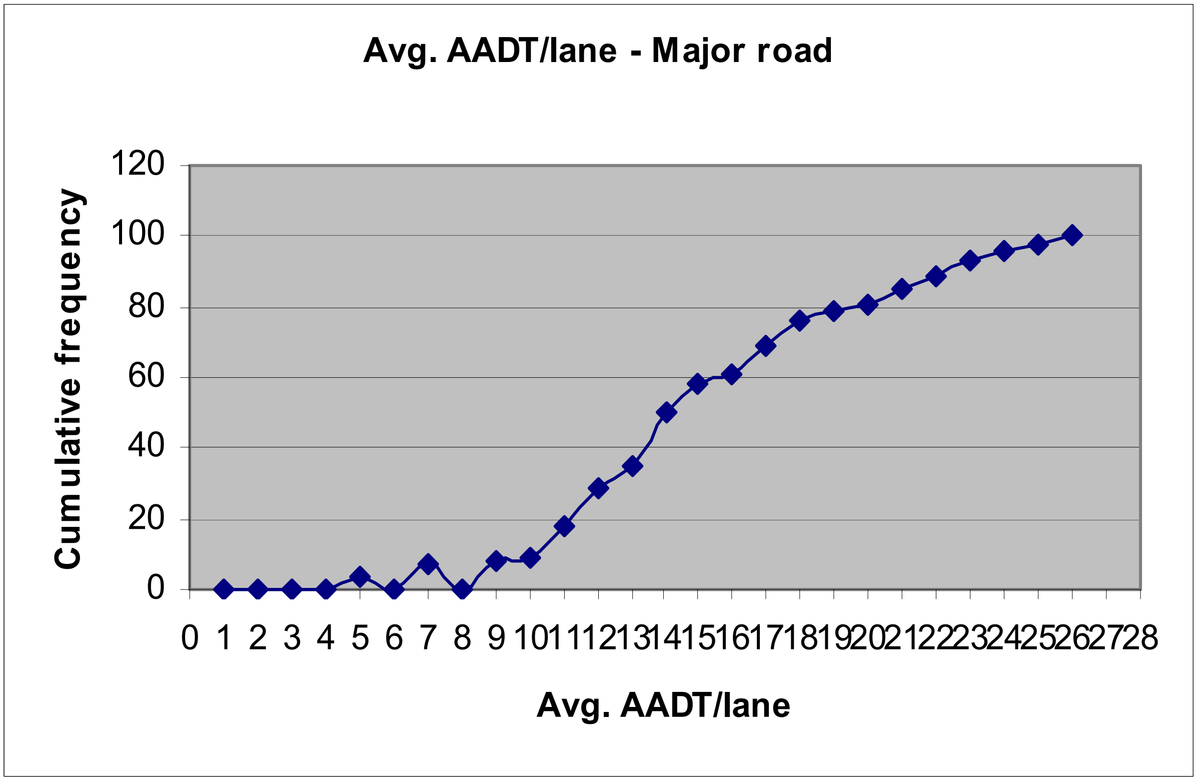

Figure 4.2 Cumulative frequency plot for Avg. AADT/Major-lane for 4 x 2 intersections

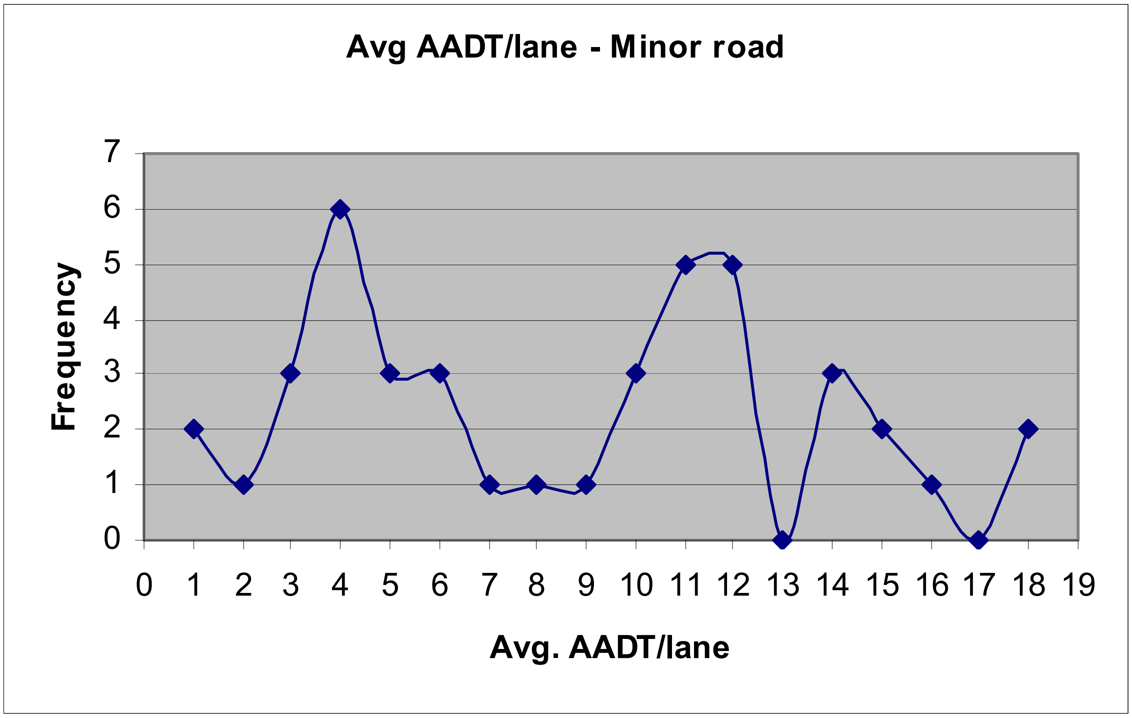

Figure 4.3 Frequency plot for Avg. AADT/Minor-lane for 4 x 2 intersections

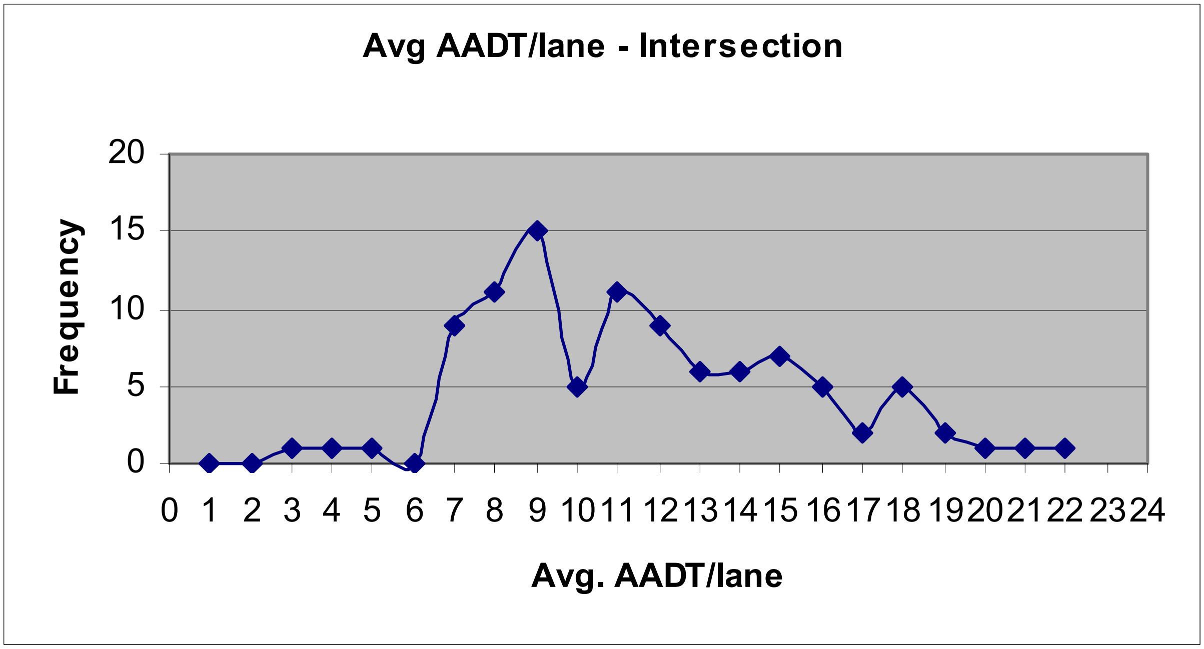

Figure 4.4 Frequency plot for Avg. AADT/(through lanes at intersection) for 4 x 2 intersections

Looking at the frequency plots, it was fairly reasonable to deduce that AADT/lane for the major road followed a somewhat normal distribution and therefore it was decided to use it to classify intersections based on traffic volume. After deciding on using AADT/lane for major road for classification of the intersections, the intersections were further categorized based on AADT/lane. In case of classifying each intersection configuration into high/low traffic volume, a cut-off ADT/lane had to be identified. The cumulative frequency plots for each type of intersection were carefully analyzed and the 50th percentile volumes were estimated for this purpose. It was checked if a balance was maintained after each type of intersection was classified as per average AADT per lane, i.e. if more or less, equal number of intersections fell in the below and above cut-off points. For example, the cut-off point for 2x2 intersections was set at 5,000 as this resulted in the distribution of intersections below and above 5,000 equally. 4x2 intersections were further classified based on number of left turning lanes (i.e. <= 2 and > 2). The complete summary of the categories have been listed in Table 4-3. Table 4-3 indicates the category of intersections present in each county.

Table 4-3 Classification of intersections into 19 categories

Size	MJ AADT/ MJ Lane	Category
2 x 2	≥5,000	1
	<5,000	2
4 x 2	≥7,000 (Total LTL ≤ 2)	3
	≥7,000 (Total LTL > 2)	4
	<7,000 (Total LTL ≤ 2)	5
	<7,000 (Total LTL > 2)	6
4 x 4	≥7,500	7
	<7,500	8
6 x 2	≥7,500	9
	<7,500	10
6 x 4 and 6 x 6	-	11
3-Legged (T-intersections)	≥7,500	12
	<7,500	13
One Way Major	-	14
One Way Minor	-	15
Both Major and Minor One-Way	-	16
Ramp Intersections	≥7,500	17
	<7,500	18
3-Legged Intersection with at least one One_Way Street		19

Table 4-4 Categories of intersections present in each county's master database

County	Categories Present
Brevard	1 to13, 17, 18
City of Orlando	1 to 16
Hillsborough	1 to 14, 17, 18
Miami-Dade	1 to 19
Orange	1 to 13
Seminole	1 to 13

Once the intersections were broken down according to type, the means, standard deviations and percentiles were determined for each category. Tables were made to incorporate all the data related to eight (8) different divisions, which consisted of collision type, severity class, light conditions, weather, surface conditions, month of the year, day of the week and hour of the day. A versatile code was written in SAS to compute crash statistics like mean, standard deviation and the 85th, 90th and 95th percentiles for all the nineteen classification tables that contained the above mentioned categories and their respective crash summary. Table 4-5 gives a sample of the table developed. The top header of the table indicates the category (10) and configuration (6x2) of the intersections used to develop the table, and the number of intersections (16) present in this category. The numbers in the first column indicate the total number of crashes pertaining to their respective crash criteria (Collision type, Severity etc.) average over the years 1999, 2000, 2001 and 2002 for Miami-Dade County. The numbers in the second column represent the average crashes per year. The rest of the columns indicate the mean crashes per intersection per year, the standard deviations for every category, and the 85th, 90th and 95th percentile of crashes.

Similar tables were developed for the 19 categories of intersections in all six counties.

Table 4-5 A sample of a classification table for Dade County

EXPECTED ANNUAL ACCIDENT TABLE - DADE COUNTY

TYPE 10 - 6 LANE x 2 LANE SIGNALIZED INTERSECTION, AADT PER LANE ON MAJOR ROAD < 7,500

TOTAL NUMBER OF INTERSECTIONS - 16

		Average Number Crashes Per Year*	Mean Crashes Per Year Per Intersection	Standard Deviation**	85th Percentile	90th Percentile	95th Percentile
Collision Type	Head On	2	0.14	0.22	0	0	1
	Left Turn	33	2.05	1.84	4	5	5
	Pedestrian/Bicycle	5	0.30	0.34	1	1	1
	Rear End	79	4.95	3.93	8	9	11
	Angle	53	3.33	2.69	6	7	8
	Sideswipe	22	1.38	1.02	2	3	3
	Right Turn	4	0.27	0.31	1	1	1
	Other/Unknown	26	1.61	1.38	2	3	4
Severity	No Injury	148	9.27	5.47	14	17	20
	Possible Injury	44	2.75	2.15	5	6	6
	Non-Incapacitating Injury	24	1.48	1.34	3	4	4
	Incapacitating Injury	7	0.45	0.59	1	1	2
	Fatal Injury	1	0.06	0.11	0	0	0
Light Conditions	Daylight	162	10.11	6.33	18	19	21
	Dusk	5	0.33	0.31	1	1	1
	Dawn	3	0.17	0.31	0	1	1
	Dark (w/street lights)	53	3.33	2.25	6	6	7
	Dark (w/o street lights)	1	0.08	0.12	0	0	0
Surface Conditions	Dry	189	11.80	7.56	21	23	26
	Wet	33	2.03	1.40	4	4	4
	Others	3	0.19	0.23	0	1	1
Month of Year	January	24	1.50	0.98	3	3	3
	February	19	1.17	0.83	2	2	3
	March	22	1.34	0.85	2	3	3
	April	18	1.14	0.91	2	2	3
	May	22	1.36	1.02	2	3	3
	June	17	1.08	0.66	2	2	2
	July	21	1.30	1.33	3	4	4
	August	21	1.30	1.12	2	3	3
	September	17	1.05	0.63	2	2	2
	October	17	1.05	0.73	2	2	2
	November	13	0.83	0.58	1	2	2
	December	15	0.91	0.69	2	2	2
Day of Week	Monday	27	1.70	1.23	3	4	4
	Tuesday	30	1.86	1.06	3	4	4
	Wednesday	36	2.27	1.57	3	4	6
	Thursday	35	2.16	1.40	4	4	4
	Friday	38	2.38	1.53	4	5	5
	Saturday	35	2.16	1.70	4	5	5
	Sunday	24	1.50	0.91	3	3	3
Hour of Day* **	00:00 - 06:00	16	0.97	0.81	2	2	2
	06:01 - 09:00	17	1.08	0.97	2	3	3
	09:01 - 11:00	13	0.80	0.61	2	2	2
	11:01 - 13:00	14	0.89	0.82	2	2	2
	13:01 - 15:00	18	1.14	0.88	2	3	3
	15:01 - 18:00	38	2.39	1.56	4	5	5
	18:01 - 24:00	42	2.64	1.70	4	5	6

* Crashes extracted for years 1999, 2000, 2001and 2002 for long forms only.

** Standard Deviation column represents the standard deviation for mean crashes per year per intersection.

*** Hour of Day statistics are based upon a portion of the crashes with time information available.

4.4 Building a Combined Database

The previous sections have described the process of building the crash databases of each of the six counties. These crash databases were combined to form a complete database in order to study the crash characteristics for all counties. The database was developed for the years 2000 and 2001 for all counties, except for Orange County for which the 1999 and 2000 year database was used because the 2001 database was not available. This database consisted of 27230 crashes for 1562 intersections for two years.

Initially, both the long and short form crashes from all six counties were used to make a complete database. Then the long form crashes were filtered out to develop a separate long only crash database for all six counties. Another database was developed by filtering out the crashes from the four counties Brevard, City of Orlando, Hillsborough and Seminole containing both long and short form crashes. Tables for the expected number of crashes for the 19 categories were developed for both the databases: the long-form-only crash database and the four county long-short crash database.

4.5 Tests to compare each County to the Combined Database

Since the tables for the expected number of crashes on long forms for each county as well as for the combined database were prepared for the 19 categories, tested were conducted to find out if there was a difference between the tables of each county and the tables for the combined six counties. This could be used in finding out if the tables for the combined database can be referred for finding the crash characteristics of a county, rather than referring to each county table. For example, this analysis would enable us to see if the mean number of sideswipe crashes for a 6 x 2 intersection in Brevard County is any different from the sideswipe crashes for a 6 x 2 intersection for the combined six counties. If they are the same, the characteristics for the sideswipe crashes for the 6 x 2 intersections in Brevard County tables for expected number of crashes would be similar to those in the combined database tables. Hence the tables for the combined database can be used in such a case instead of referring to each of the county tables.

This analysis was carried out by conducting a Student's t-test to compare the mean number of crashes of each county to the means of the combined database. The results show whether the means are equal or not. The results were tabulated and a sample is shown in Table 4-6. The mark "√" in the table indicates that the mean number of crashes for a particular county is similar to the mean number of crashes in the combined database, indicating that the data from combined database can be used for these counties and categories. BC denotes Brevard County, CO denotes City of Orlando, HC denotes Hillsborough County, OC denotes Orange County, SC represents Seminole County and DC denotes Dade County. Category 19 was not included because this category has been assigned only in Dade County.

Table 4-6 Comparison of means of each of six counties to the means of the combined six counties

Type 4
4 Lane x 2 Lane Intersection, Signalized, AADT/lane for Major Road ≥ 7000 (LT lanes > 2)

		BC	CO	HC	OC	SC	DC
Collision Type	Rear End	√					√
	Head On		√				
	Angle	√	√			√	√
	Left Turn	√	√	√			√
	Right Turn		√				
	Sideswipe	√					√
	Pedestrian/Bicycle						
	Other		√		√	√	√
Severity	No Injury	√	√	√	√	√	√
	Possible Injury		√			√	√
	Non-Incapacitating Injury			√		√	√
	Capacitating Injury			√			√
	Fatal Crashes						
Light Conditions	Daylight	√	√		√	√	√
	Dusk						
	Dawn						
	Dark (w/street lights)	√	√		√	√	√
	Dark (wo/street lights)				√		√
Surface Condition	Dry	√	√	√	√	√	√
	Wet				√		√
	Others		√		√		
Month of year	January		√				√
	February	√					√
	March		√				√
	April						√
	May						√
	June						√
	July	√					√
	August	√	√				√
	September	√					√
	October						√
	November						√
	December						√
Day of week	Monday	√	√			√	√
	Tuesday	√			√		√
	Wednesday	√			√	√	√
	Thursday	√			√		√
	Friday	√	√	√	√		√
	Saturday	√					√
	Sunday	√			√		√
Hour of day	00:00 - 06:00	√					√
	06:01 - 09:00				√		√
	09:01 - 11:00						√
	11:01 - 13:00						√
	13:01 - 15:00				√		√
	15:01 - 18:00			√	√		√
	18:01 - 24:00	√	√	√	√		√

√ represents the similarity in the county mean and the mean of the combined database

4.6 Classifying the Combined Database

As the combined database had larger number of intersections, they could be divided into a larger number of categories. Hence an analysis was conducted to increase the number of categories of intersections in the combined database. All intersections from the six counties were categorized into various types so that all intersections in each category had similar crash characteristics. The first step involved was to combine the geometry files of all six counties. The geometry files were sorted based on the field “int_id”, which is the unique ID assigned to each intersection. Intersections were filtered out from this database based on the lane configuration of each intersection (2x2, 4x2 etc). Separate tables were made for intersections of the same type. As AADT is one of the most important factors affecting the crash frequency, the crashes were categorized based on AADT/number of approach lanes on the major road. For each type of intersection the median value of AADT was noted. The number of intersections for every 1000 AADT values for each configuration of intersection was listed. A sample of such a list has been shown for the 2x2 intersections in Table 4-7.

In order to form categories for a particular type, the range of AADT was widened and the number of intersections under each range was noted. Table 4-8 shows this method for 2x2 intersections. The table first shows the initial splits made in AADT/lane, and the intersections present in each of the splits (indicated in brackets). Then the range of AADT/lane was widened to make six categories of intersections of the type 2x2. This range was further increased to form three, and later two categories. It was decided to split the intersection into three categories (shown in bold) because the split of the intersections was even and each category had sufficient number of intersections.

Table 4-7 Initial sampling of 2 x 2 intersections based on the AADT/Major Lane values

2 x 2

Total number of intersections = 175

Median = 5874 = approximate by 5900 or 6000

Split Number	AADT Range	Number of Intersections
1	=< 3000	11
2	> 3000 and =< 4000	12
3	> 4000 and =< 5000	38
4	> 5000 and =< 6000	31
5	> 6000 and =< 7000	20
6	> 7000 and =< 8000	14
7	> 8000 and =< 9000	15
8	> 9000 and =< 10000	7
9	> 10000 and =< 11000	10
10	> 11000	17

At the end of this process, there were various combinations of categories for each type of intersection. The optimum number of categories for each type was obtained by making sure that (a) the number of intersections in each category was almost the same, (b) adequate sample size is achieved, and, (c) the cutoff AADT/lane values were similar. Categories were formed based on this range and the idea that the number of intersections in all categories was as close as possible. Various combinations of categories were formed for each type of intersection with different range of AADT values.

Table 4-8 Categorizing the intersections based on different AADT/lane for 2x2 intersections

Total number of intersections = 175		
Median = 5874 = 6000 (approx)		
splits:	≤ 3000 (11)	
	> 3000 and ≤ 4000 (12)	
	> 4000 and ≤ 5000 (38)	
	> 5000 and ≤ 6000 (31)	
	> 6000 and ≤ 7000 (20)	
	> 7000 and ≤ 8000 (14)	
	> 8000 and ≤ 9000 (15)	
	> 9000 and ≤ 10000 (7)	
	> 10000 and ≤ 11000 (10)	
	> 11000 (17)	
6 Categories:	≤ 4000 (23)	
	> 4000 and ≤ 5000 (38)	
	> 5000 and ≤ 6000 (31)	
	> 6000 and ≤ 8000 (34)	
	> 8000 and ≤ 11000 (32)	
	> 11000 (17)	
3 Categories:	**≤ 5000 (61)**	
	> 5000 and ≤ 9000 (80)	
	> 9000 (34)	
2 Categories:	≤ 6000 (92)	> 6000 (113)

All the types were categorized based on the AADT values. Since the intersections of type 4x2 were very large in number (541), they were subcategorized based on a new variable. First the intersections were subcategorized based on the number of left-turning lanes. But a majority of the intersections had 4 left-turning lanes. Hence the sub-classification of intersections based on left-turning lanes was not considered appropriate. Thus this variable was discarded for the purpose of sub-classification. Next, the intersections were subcategorized based on the speed limit on the major road. The median speed limit was 40mph, and sub-classification based on this speed produced satisfactory results. Therefore the 4 x 2 intersections were classified first by AADT and then by the speed limit on the major road.

This method was adapted to develop classifications for all types of intersections. After completing the classification, it was found that 38 categories of intersections were developed. These have been tabulated in Table 4-9. Then a summary of intersections was developed indicating the number of categories formed for each type of intersection. This has been shown in Table 4.10.

Table 4-9 Classification of intersections into 38 types

S.No	Type	Condition for AADT/Lane of Major Road	# Intersections
1	2 x 2	=< 5000	61
2		> 5000 and =< 9000	80
3		> 9000	34
4	4 x 2	=< 5000 and MJ speed =< 40	41
5		=< 5000 and MJ speed > 40	37
6		> 5000 and =< 7000 and MJ speed =< 40	48
7		> 5000 and =< 7000 and MJ speed > 40	65
8		> 7000 and =< 9000 and MJ speed =< 40	99
9		> 7000 and =< 9000 and MJ speed > 40	63
10		> 9000 and =< 11000 and MJ speed =< 40	41
11		> 9000 and =< 11000 and MJ speed > 40	42
12		> 11000 and MJ speed =< 40	32
13		> 11000 and MJ speed > 40	73
14	4 x 4	=< 5000	21
15		> 5000 and =< 7000	36
16		> 7000 and =< 9000	35
17		> 9000 and =< 11000	35
18		> 11000	31
19	6 x 2	=< 7000	44
20		> 7000 and =< 9000	49
21		> 9000 and =< 11000	37

22		> 11000	27
23	6 x 4	=< 9000	50
24		> 9000	40
25	2 x T2	=< 8000	26
26		> 8000	20
27	4 x T2	=< 7000	44
28		> 7000	69
29	4 x T4		28
30	6 x T2		42
31	6 x T4		14
32	One Way Major	=< 7000	45
33		> 7000	40
34	One Way Minor		36
35	Both One way		13
36	One way and T		14
37	Ramps	=< 7000	24
38		> 7000	26

Table 4-10 Summary of Classifications

S.No	**Type**	**Number of Categories**
1	2 x 2	3
2	4 x 2	10
3	4 x 4	5
4	6 x 2	4
5	6 x 4	2
6	2 x T2	2
7	4 x T2	2
8	4 x T4	1
9	6 x T2	1
10	6 x T4	1
11	One Way Major	2
12	One Way Minor	1
13	Both One Ways	1
14	One way and T	1
15	Ramps	2
	Total	**38**

After developing the 38 categories, tables were developed to predict the expected number of crashes at each category of intersections. These tables were developed for the database containing only the long form crashes, and were represented in the same way as the 19 category tables.

4.7 Summary

This chapter has discussed the methodology used in data collection, classifying intersections, developing tables for the expected number of crashes in all categories and the method used in combining the databases of all counties and reclassifying the intersections into 38 categories. The tables developed for the expected number of crashes can be used to estimate the average number of crashes occurring at any particular configuration of intersection in any of the six counties. The tables for the combined database can also be used to estimate the mean number of crashes at these intersections as they have been finely classified in 38 categories, instead of the 19 categories in the tables of each county. These tables can also be used to estimate the number of crashes in other counties in Florida. The combined six counties could be used to represent other counties in the state that are not represented in the database. Also, counties with similar characteristics or at proximity with one of the six counties can use the tables for the respective county.

5 USING NEURAL NETWORKS TO IDENTIFY UNSAFE INTERSECTIONS

5.1 Predicting Frequency of Crashes at Intersections

The objective of this study is to predict the frequency of crashes at various intersections using different neural network models and identifying the geometric and traffic characteristics at intersections that affect particular types of crashes. These characteristics were evaluated to identify the manner in which they affect the crash frequency at intersections. If the models predict that an intersection has a lot of crashes, the characteristics of the intersection can be changed so as to make the intersection safer.

To predict the crash frequency, a database of intersections was first developed. This database contained the geometric and traffic variables using which the crash frequency would be predicted, that is, it contained the input variables for the models that were going to predict the crash frequency. The database consisted of 1563 intersections from all six counties, as found in the combined database. The intersection database contained the following input variables:

1. Number of through lanes on the major road
2. Number of through lanes on the minor road
3. Total Left Turning Lanes at the intersection
4. Number of Protected Left Turning Lanes on the major road
5. Number of Protected Left Turning Lanes on the minor road
6. Number of channelized right turning lanes on the major road
7. Number of channelized right turning lanes on the minor road
8. Speed Limit on the major road

9. AADT on the major road

Since the data for the speed limit and AADT on the minor roadway was not known for more than half the intersections in the database, these variables could not be used.

The number of crashes that had occurred at these intersections during the years 2000 and 2001 were identified. Since these crashes amounted to two years, the number was halved to obtain the crash frequency for each year.

Then the neural network models were used to predict the crash frequency at these intersections. The Multi Layer Perceptron (MLP) Neural Network, Probabilistic Neural Network (PNN) and the Generalized Regression Neural Network (GRNN) models have been used in this study. The MLP models have been used frequently in many traffic safety studies, and have often been found to be very effective in analyzing the crash frequencies. The GRNN model has hardly been used in traffic safety analysis. The comparison of the models will prove if the MLP model is in fact the best neural network model available to be used in traffic safety studies, as has been found by Abdelwahab and Abdel-Aty (2001, 2004).

5.2 Crash Frequency Prediction using MLP Neural Network

A program was written in MATLAB to build the MLP neural network. The program performed the following functions:

1. The input variables in the database were normalized. This was carried out because the contribution of an input will depend heavily on its variability relative to other inputs. If one input has a range of 0 to 1, while another input has a range of 0 to 1,000,000, then the contribution of the first input to the distance will be overruled

by the second input. So it is essential to rescale the inputs so that their variability reflects their importance. It is common to standardize each input to the same range or the same standard deviation. Hence the database was normalized for a unit variance.

2. Take an input of the crash frequency data for the 1563 intersections.
3. Shuffle the input data and take the first 75% of the data for training and the rest 25% for testing.
4. Use 1 hidden nodes for training the data. The Resilient Back Propagation neural network was used in the training. The activation functions that proved to be the best for the hidden and output layers were hyperbolic tangent sigmoid and pure linear respectively. The maximum number of epochs used was 3000. The learning rate was 0.05.
5. Calculate the root mean squared error (RMSE) by adding the squares of the difference of the actual value and the predicted value of the crash frequencies, averaging them over the intersections used in the testing phase, and taking a square root of this value.
6. Vary the number of hidden nodes from 1-15.
7. Repeat the whole procedure five times and take the average of the results (Root Mean Squared Error - RMSE) for each value of the number of hidden nodes.

The results obtained are shown in Table 5-1. The results are arranged with ascending order of RMSE. The lowest RMSE obtained is 9.44. The MAPE (Mean Absolute Percentage Error) for this model was around 80%. This is a large value considering that the error in predicting crash frequencies for each intersection can have an error of 10 crashes. An error of 10 crashes implies a possible misinterpretation of the safety at the intersection. Hence this model was not considered suitable for predicting the crash frequencies at signalized intersections.

Table 5-1 Results obtained for predicting long form crash frequencies using MLP NN

No Hidden Nodes	Average RMSE
4	9.44
2	9.51
3	9.53
1	9.57
5	9.61
8	9.68
6	9.76
9	9.97
7	10.02
12	10.15
13	10.19
11	10.20
14	10.28
10	10.37
15	10.38

5.3 Crash Frequency Prediction using GRNN

A program was written in MATLAB to develop the GRNN. This program was similar to the MLP program, except that instead of the hidden nodes, learning rate and number of epochs, the spread was varied from 0.01 to 5.0 with increments on 0.02. A lot of spread values were used to make certain that the results obtained are accurate. The results obtained from GRNN are tabulated in Table 5-2.

Table 5-2 Results obtained for predicting long form crash frequencies using GRNN

Spread	Average RMSE
1.19	9.03
1.17	9.03
1.21	9.03
1.15	9.03
1.23	9.03
1.13	9.03
1.25	9.04
1.11	9.04
1.27	9.04
1.09	9.05

Although the results obtained using GRNN were better than the MLP NN, they were not satisfactory. The MAPE value was similar to that of the MLP model. This model cannot be used for predicting crash frequency with such a RMSE.

The possible reason because of which the errors were so large was that the crash frequency per intersection ranged from 0 to 113, and the models were unable to perform well when the intersections had high crash frequencies. A very small percentage of intersections have a very high number of crashes, and the models developed cannot predict these crash frequencies correctly leading to a large error. Hence an appropriate method was sought after that could accurately predict crashes for all range of crash frequencies.

5.4 Predicting Total Crash Frequency Based on Number of Lanes

A new methodology was devised to predict the crash frequencies more precisely at signalized intersections. First, the total number of lanes at each intersection was calculated by summing up the number of through lanes, exclusive left turning lanes and channelized right turning lanes on the major and minor roads. Since this number indicates the total number of lanes at the intersection, it is a representation of the size of the intersection. It could also implicitly indicate the magnitude of AADT at the intersection. The greater the number of lanes at the intersection, the bigger it is. Then a graph was drawn to observe the variation of the average number of crashes per intersection with the total number of lanes at the intersections, as shown in Figure 5.1.

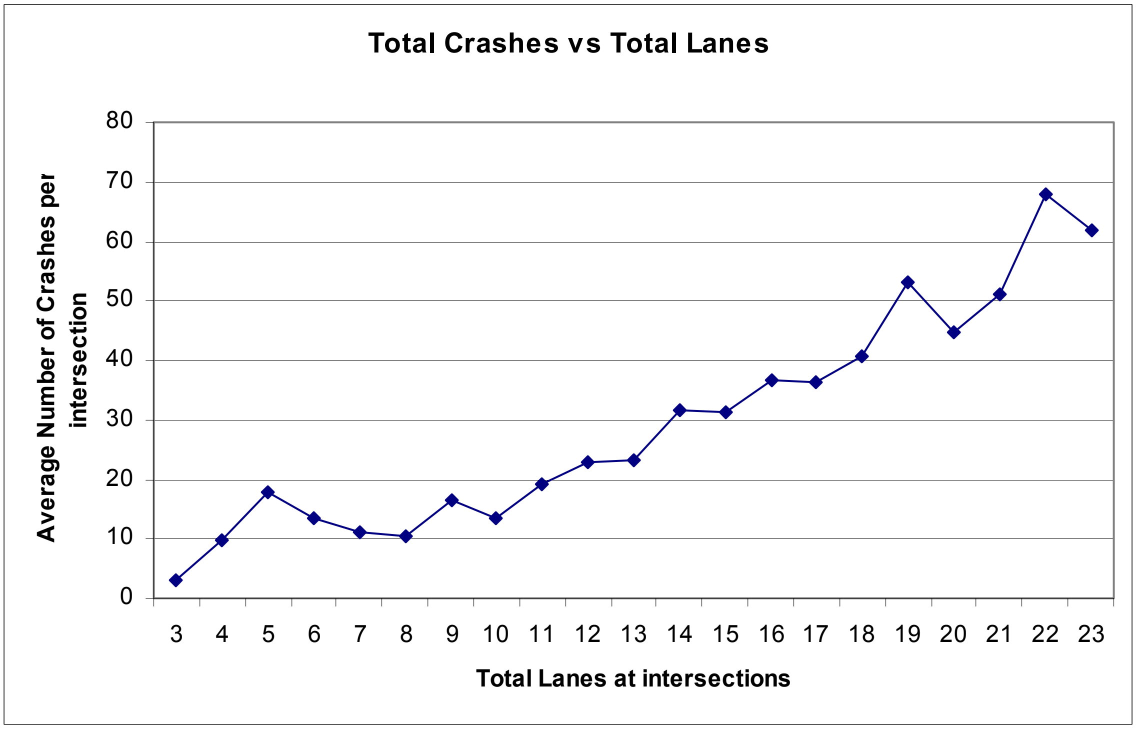

Figure 5.1 Variation of total crashes per intersection with total lanes per intersection

Clearly, the graph shows an increasing trend of total crashes per intersection as the total lanes at the intersections increase. Thus it can be concluded that the number of crashes at an intersection increase as the size of the intersection increases. Therefore, any intersection can be classified into one of the following types: (a) the intersection has more crashes than the average number of crashes for intersections with the same number of total lanes; (b) the intersection has less than or equal number of crashes than the average number of crashes for the intersections with the same number of total lanes. The intersections in the former category can be considered as “unsafe intersections” while the rest can be considered as “safe intersections”.

Therefore, intersections can be categorized into safe or unsafe intersections based on the total crashes it has incurred and the total number of lanes it has. In order to predict the crash frequencies, a model can first be developed that easily and efficiently classifies intersections into safe and unsafe categories. Then the frequency of crashes can be predicted for the safe and unsafe intersections by developing separate models for the two types. This method develops models for separate data ranges, and is thus expected to reduce the error in crash frequency prediction.

Thus, neural network models were built to classify intersections into safe and unsafe intersections first. To accomplish this, the intersection database was divided into the two categories. First, the intersections with total lanes between 3 to 5, 6 to 10, 11 to 15, and 16 and above were grouped together. The average crashes per intersection were found for these groups. These values have been shown in Table 5.3. If an intersection incurred more crashes than the average number of crashes obtained from Table 5.3, the intersection was categorized as an unsafe intersection. If not, it was categorized as a safe intersection. The neural network models used for this classification were the MLP and PNN models.

Table 5-3 Average number of crashes for different groups of intersections

Total Lanes at an Intersection	Average Number of Crashes
3 to 5	13.17
6 to 10	12.71
11 to 15	23.93

16 and above	40.69

Separate PNN and MLP neural network models were developed to predict the number of crashes at safe and unsafe intersections. These crashes were predicted using the MLP and GRNN models. These models were compared to the previous model that predicted the crash frequency for all intersections. The models that worked the best were used as the final models for predicting the frequency of crashes at the intersections.

5.4.1 Classification of Intersections

As was mentioned earlier, MLP and PNN models were utilized to classify the intersections into safe and unsafe categories. The following steps were carried out to classify the intersections:

1. The database was classified into the two categories: safe and unsafe, using the method mentioned in the previous section. 65% of the intersections were categorized as safe intersections.
2. The input variables were normalized as was described in section 5.2.
3. The number of input and output nodes was decided. The number of input nodes is equal to the number of input variables, which are 9. The number of output nodes is one, indicating an output of 0 or 1.
4. The database was randomized. Out of this randomized database, 75% of unsafe intersections and an equal number of safe intersections were selected for training. This ensured that an equal proportion of safe and unsafe intersections were trained so that there was no bias in the estimation of results. The rest of the intersections are used for testing the neural network model developed.

5. For MLP neural network:
    a. The learning rate was set to 0.05, the maximum number of epochs was set to 3000 and the Resilient backpropagation (rprop) algorithm was used to develop the neural network. The Resilient backpropagation algorithm leads a transparent and powerful adaptation process that is straightforward and very efficiently computed with respect to both time and storage consumption (Riedmiller and Braun, 1993). Thus, the rprop was used as they were considered more advantageous compared to the ordinary backpropagation algorithms.
    b. One hidden layer was used. The number of neurons in the hidden layers was increased from 5 to 50. The performance of the neural network is evaluated for different number of hidden nodes.
    c. The activation functions for the hidden and output layers were tan sigmoid and pure-linear. This combination of activation functions gave the best results when tested with other combinations.
    d. The MLP neural network was trained using the training data selected in step 4.
    e. The training data was used to simulate the network and predict the output. This output was compared to the predicted output and the accuracy in prediction is calculated. The accuracy with which the total database is classified is calculated; the accuracy in predicting the safe and unsafe intersections is also calculated.

    f. The MLP neural network model was used to classify the intersections in the test database into safe and unsafe intersections. The test output was compared to the actual output and the accuracies were determined. The test accuracy was determined that represents the percentage of intersections that were correctly classified. The accuracies with which the safe and unsafe intersections were predicted were also determined.

6. For the PNN model:
    a. Spread of the neural network was varied from 0.01 to 2.0 with increments of 0.02. The PNN model with a spread value greater than 2.0 did not perform well.
    b. The PNN model was trained using the training data selected in step 4.
    c. The test data was used to predict if the intersections in the database were safe or unsafe. The accuracies were determined in the same manner they were calculated for the MLP neural network model.
7. This process was repeated five times and the results of the MLP and PNN models were stored in separate files.

Since the training and test databases were randomly chosen, the results were averaged. These results have been tabulated in Table 5.4 and Table 5.5 for MLP and PNN models, respectively. The numbers in the table represent the percentage accuracy. For example, a test accuracy of 64.66 indicates that 64.66% of the test database was classified correctly. Accuracy of safe intersection being 58.48% indicates that this percentage of safe intersections was classified correctly. The results have been tabulated in a decreasing order of test accuracies. Thus the best accuracy of 64.66% for the MLP neural network is obtained using 5 hidden nodes. The highest accuracy attained by the PNN model is 65.00%, which is almost equal the accuracy of the MLP neural network model. An interesting point to note is that the PNN classified the safe and unsafe intersections with almost similar accuracies, whereas the MLP model classified the unsafe intersections with a higher accuracy compared to safe intersections.

Table 5-4 Results of the testing phase of MLP neural network for classifying intersections into safe and unsafe categories

# Hidden Nodes	Test Accuracy	Accuracy of Safe Intersections	Accuracy of Unsafe Intersections
5	64.66	58.48	70.83
25	64.07	61.72	66.42
35	63.83	62.22	65.44
30	63.70	62.93	64.46
50	63.41	61.61	65.20
20	63.34	62.22	64.46
55	63.33	63.92	62.75
60	63.15	61.34	64.95
10	63.01	63.76	62.25
40	62.82	62.66	62.99
15	62.22	61.45	62.99
45	60.53	61.01	60.05

Table 5-5 Results of the testing phase of PNN for classifying intersections into safe and unsafe categories

Spread	Test Accuracy	Accuracy for Safe Intersections	Accuracy for Unsafe Intersections
1.25	65.00	64.80	65.20
1.01	64.85	64.25	65.44
1.23	64.73	66.72	62.75
1.03	64.61	64.03	65.20
1.29	64.59	65.46	63.73
1.31	64.58	65.68	63.48
1.27	64.51	65.79	63.24
0.87	64.48	65.24	63.73
0.99	64.35	66.94	61.76
1.19	64.34	64.96	63.73

Both the MLP and PNN models can be considered to be equally good in classifying intersections into safe and unsafe intersections.

5.4.2 Determining the Significant Variables in Classifying Intersections

The models developed above can give a good prediction about the classification of an intersection as safe or unsafe intersection. These models take into account the entire nine variables used in the input phase. While some variables might play a significant role in governing if an intersection is of safe or unsafe type, some variables might not be affecting the output at all. Hence there is a need to determine the significant variables that govern the model. If a variable is found to be significant, it can be controlled to make an intersection a safer place to travel.

The significant variables were identified using a Forward Sequential Selection method. According to this method, just one input variable is used at a time to train and test the databases. Once all the inputs have been used individually, the test accuracies are compared. The variable that gives the maximum accuracy is chosen as the most significant variable. Then training and testing of databases is carried out by using this variable along with the other input variables one at a time. The variable whose combination with the first significant variable gives the highest accuracy is chosen as the next significant variable. This process is repeated till there is no further increase in accuracy by addition of any of the variables. All the variables selected in this process are determined to be the significant variables in the model.

As the PNN model gave a slightly better performance in classifying the intersections into Safe and Unsafe categories, the significance of variables was tested using PNN. The network was developed in the same method as described in section 5.4.1.

The only difference was that the number of input variables changed. The significant variables were identified using the forward sequential selection method. Table 5.6 lists the significant variables along with the accuracy and the spread used in each run.

Table 5-6 Significant Variables identified in classifying intersections into safe and unsafe intersections

Run#	Spread	Order of Significant Variables	Test Accuracy	Accuracy of Safe Intersections	Accuracy of Unsafe Intersections
1	0.06	Major AADT	63.69	68.7	58.68
2	0.31	Major Speed	65.69	59.97	71.81
3	1.06	Major LTP	67.31	66.72	67.89
4	0.81	Total Left Turning Lanes	67.35	68.75	65.89
5	1.61	Minor RTC	67.6	70.73	64.46

A combination of these variables with any other variable did not show any significant increase in the test accuracy. Hence these variables govern whether an intersection can be classified as a safe intersection or an unsafe intersection. The results show that the Major AADT, major speed limit and total left turning lanes are important factors in classifying the intersections. This is a reasonable result because an increase in these factors can be expected to increase the crash frequencies at intersections, thus making them unsafe.

5.4.3 Predicting the Crash Frequency for Safe Intersections

From the intersection database, the intersections classified as safe intersections were filtered out. The database consisted of 1017 safe intersections. The total crashes (per year) occurring at these intersections were predicted in this step using the MLP neural network and the GRNN. The methods used in developing these models were similar to those described in sections 5.2 and 5.3. The root mean square errors and the mean absolute percentage square errors for the test phase of the MLP and GRNN models have been listed in Tables 5.7 and 5.8 respectively.

Table 5-7 Errors in predicting the frequency of total crashes for safe intersections using the MLP neural network model

# Hidden Nodes	Test RMSE	Test MAPE
1	2.75	60.76
3	2.79	64.08
4	2.80	62.99
5	2.82	66.72
7	2.87	64.52
8	2.93	69.00
6	2.98	64.28
9	3	72.11
2	3.02	69.48
10	3.10	66.52

Table 5-8 Errors in predicting the frequency of total crashes for safe intersections using the GRNN model

Spread	Test RMSE	Test MAPE
1.95	2.785	63.31
1.85	2.786	62.79
1.65	2.787	62.00
1.6	2.787	61.83
1.9	2.787	63.02
1.55	2.787	61.80
1.5	2.788	61.14
1.8	2.789	63.02
1.45	2.791	60.94

Both the MLP and GRNN models performed equally well. The test MAPE obtained was also similar for both the models. The MAPE values seem large because this database consists of safe intersections having small number of crashes. Hence when the crashes are predicted for an intersection having a very small number of crashes, even a small absolute error will be portrayed as a large percentage error. Therefore the RMSE was used to judge the model.

5.4.4 Significant Variables for Predicting Crashes at Safe Intersections

The significant variables were identified for the MLP neural network model using the forward sequential selection method. The criterion for choosing the significant variables was that the RMSE was the least for the significant set of variables. The significant variables have been listed in Table 5.9. This table indicates that safe intersections are affected only by the three variables listed. For example, if an intersection is identified as a safe intersection, an increase or decrease in AADT can significantly affect the number of crashes occurring at the intersection. The number of hidden nodes was varied from 1 to 15. Again, the results obtained are reasonable because an increase in the major AADT and minor LTP can be expected to increase the frequency of total crashes.

Table 5-9 Significant variables identified in predicting the frequency of crashes at safe intersections

Run #	Significant Variables	# Hidden Nodes	Test RMSE
1	Minor LTP	1	2.90
2	Major AADT	2	2.82
3	Minor RTC	1	2.62

5.4.5 Predicting the Crash Frequency for Unsafe Intersections

The intersections classified as unsafe intersections in section 5.4.1 were used in this analysis for predicting the crash frequency of unsafe intersections. 75% of the intersections were used in the training phase and 25% were used in the testing phase. The models were developed in a method similar to the ones adapted in sections 5.2 and 5.3. The results of the MLP and GRNN models have been tabulated in Tables 5.10 and 5.11 respectively. The MLP neural network model performed better compared to the GRNN model.

Table 5-10 Errors in predicting the frequency of total crashes for unsafe intersections using the MLP neural network model

# Hidden Nodes	Test RMSE	Test MAPE
3	5.76	31.94
1	6.07	31.83
2	6.18	32.41
5	6.27	32.10
4	6.57	32.79
8	6.58	34.10
6	6.59	34.51
7	6.67	34.25
10	6.80	33.77
9	6.82	35.16

Table 5-11 Errors in predicting the frequency of total crashes for unsafe intersections using the GRNN model

Spread	Test RMSE	Test MAPE
1.33	5.88	31.33
1.35	5.88	31.48
1.37	5.89	31.37
1.39	5.89	31.26
1.41	5.90	31.77
1.43	5.91	31.22
1.45	5.92	31.84
1.47	5.93	31.93
1.49	5.94	31.33
1.51	5.95	32.03

As can be seen in the table, the MLP model performed slightly better when compared to the GRNN model. The MAPE of the testing phase was 31.94% for the PNN model, which can be considered low. This value is lower than the value found for the safe intersections because the MAPE depends on the values of the output. Since the crash frequency is large for unsafe intersections, the denominator of the MAPE is larger resulting in lesser value of the error. Therefore, RMSE is considered a better option for evaluating the errors. To evaluate the significant variables, RMSE has been used.

5.4.6 Significant Variables for Predicting Crashes at Unsafe Intersections

The MLP neural network model was used to identify the significant variables in predicting the frequency of crashes at unsafe intersections. These variables have been listed in Table 5.12.

Table 5-12 Significant variables identified in predicting the frequency of crashes at unsafe intersections

Run #	Significant Variables	# Hidden Nodes	Test RMSE
1	Minor Lanes	8	6.36
2	Total Left Turning Lanes	3	5.65
3	Major Lanes	8	5.35

Only three variables were found significant, and the RMSE obtained was lesser than the value obtained by using all input variables. Hence these variables were used for further analysis.

5.4.7 Estimating a Pattern in Significant Variables

Identifying the significant variables is the first step in predicting how the input variables affect the output. The affect of a change in input on the output has to be found. The factors that tend to increase the crash frequency at an intersection can be checked and controlled if an intersection is found to have a large number of crashes and hence is unsafe for travel.

To identify this, the pattern in which the significant input variables lead to an intersection being safe or unsafe is found out. For this purpose, a “simulation” database was created that contained all possible combinations of the original 9 input variables. This database basically contained all possible intersection types that could be generated with the input variables. These intersections were tested using the MLP neural network model that was used in classifying the intersections into safe and unsafe intersections. Only the significant variables were used to identify if the intersections could be classified as safe or unsafe.

The database was created by referring to the combined master database and observing the data patterns. The following steps were followed in developing the simulation database:

1. The AADT on the major road was used between 10000 and 80000 with increments of 10000. Since only 3 intersections in the master database had an AADT of over 80000, this number was set as the maximum limit for AADT.
2. The speed limit was varied from 30-55 mph with increments of 5 mph.
3. The number of major lanes was varied from 2 to 6 with increments of 2 lanes. Since no intersections with an AADT of over 30000 had 2 lanes on the major road in the master database, the minimum number of lanes used for AADT of 30000 and above was 4. Similarly the maximum number of lanes for an AADT of below 30000 was used as 4.
4. The number of minor lanes was also varied from 2 to 6 with increments of 2. Also, the minor lanes were always set to be equal to or less than the number of major lanes.
5. The left turning lanes were from 0 to 8 with increments of 2. The number of left turning lanes was always equal to or lower than the sum of major and minor lanes at the intersection, but never exceeding 8.
6. The number of protected left turning lanes was varied between 0 and 4. They were always lesser than or equal to the number of left turning lanes. Also, the sum

of protected lanes on both major and minor roads was less or equal to the number of left turning lanes.

7. Channelized right turning lanes (both major and minor) were varied between 0 and 2.

The test database was developed based on these combinations of the nine variables. The total number of intersections obtained using this method was 98928. These intersections were used as input in the PNN model that was used in classifying the intersections into safe or unsafe type. Only the significant variables were used in classifying the intersections.

The PNN neural model classified 49176 intersections as safe intersections and the rest as unsafe intersections. The safe intersections were separated out from the unsafe intersections and were saved in different files. These intersections were used in the MLP neural network models for predicting the frequency of crashes at safe and unsafe intersections. Separate programs were written that trained the safe and unsafe intersections using the complete databases used in sections 5.4.3 and 5.4.5. This was done so that the performance of the neural network models could be further enhanced.

The training of the MLP model for safe intersections was carried out by using only the significant variables shown in Table 5.9. The intersections identified as the safe intersections during the classification phase of the simulation database were chosen and the significant variables were extracted from these intersections (those shown in Table 5.9). The number of crashes occurring at these safe intersections was predicted using the trained MLP model. The frequency of crashes at unsafe intersections was predicted in a similar manner using the MLP neural network model developed in section 5.4.5.

The files containing the frequency of crashes at safe and unsafe intersections were merged together. From these files, the number of crashes occurring at intersections with each value of the input variable was found out. For example, the number of crashes occurring at different values of AADT were found and plotted. This plot establishes a trend of variation of the number of crashes occurring at the intersection with a change in AADT. Similar plots were drawn for all input variables and their affect on the output variable was established. The following points discuss the trends observed with different input variables:

1. *Major Lanes:* The number of crashes occurring at intersections with 2, 4 and 6 through lanes on the major road were determined. The number of intersections with each number of major lanes was different because of the constraint imposed in using the major limit while creating the database (the minimum number of lanes used for AADT of 30000 and above was 4, which was also the maximum number of lanes for an AADT of below 30000). Hence an average rate of crashes occurring per intersection was found out. As can be seen from the Figure 5.2, the number of crashes per intersection shows an increasing trend as the number of lanes increase. The average value of crashes per intersection has been indicated in the graph. This pattern is consistent with the findings of Keller (2004), who too finds that the trend increases.

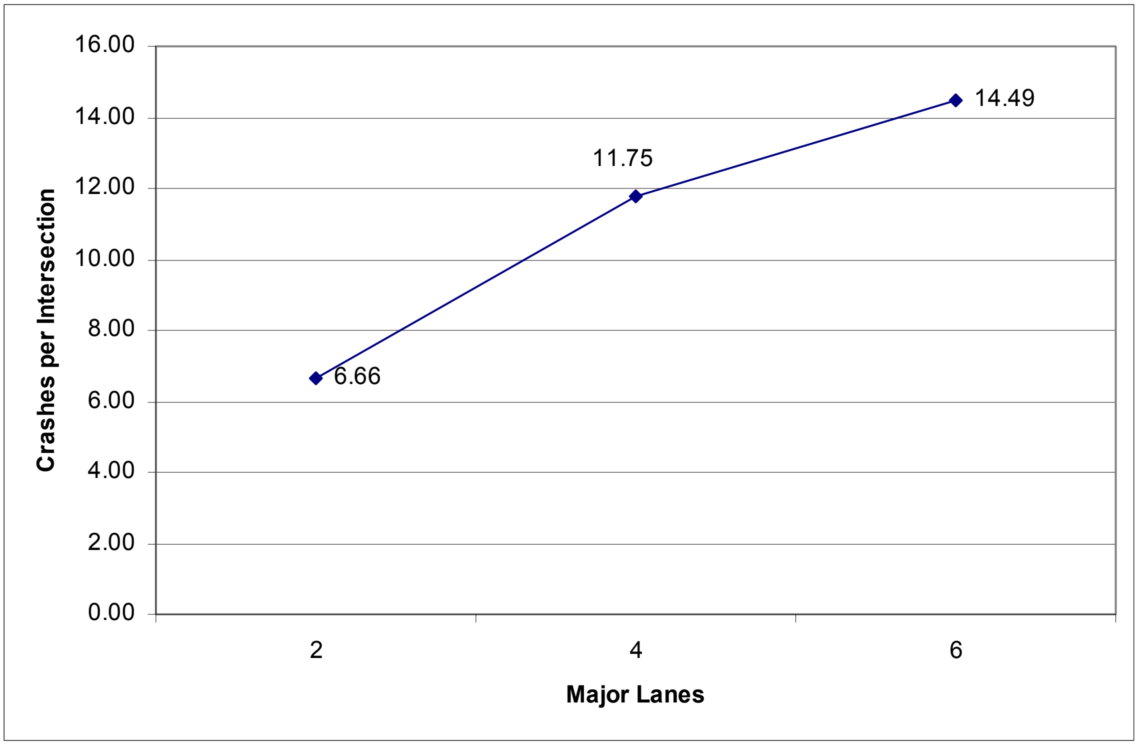

Figure 5.2 Average expected number of crashes per intersection per year for different values of through lanes on the major road

2. *Minor Lanes:* Since the number of intersections having 2, 4 and 6 through lanes on the minor road were different, average number of crashes occurring at intersections per year was determined and plotted in Figure 5.3. Keller(2004) observes a similar increasing trend, where the frequency of crashes increases drastically when the number of minor lanes are over 2. The same pattern is observed here when the frequency of crashes almost doubles when the number of minor lanes increase. Greibe (2003) also finds that an increase in minor roads increases accident risk.

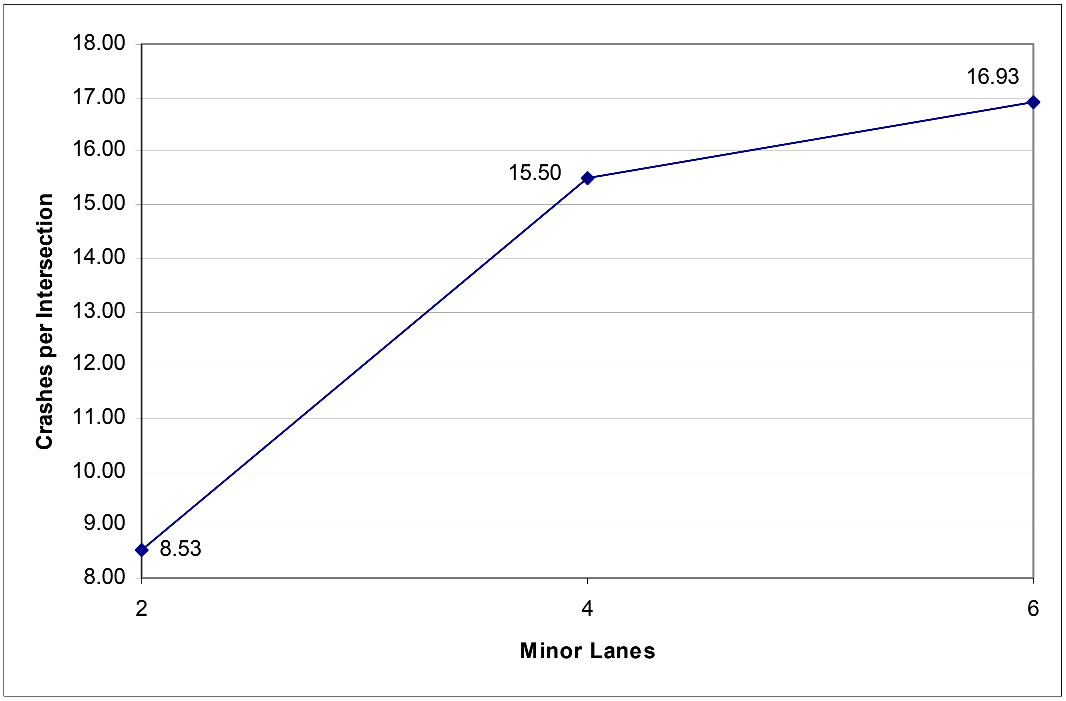

Figure 5.3 Average expected number of crashes per intersection per year for different values of through lanes on the minor road

3. *Total Left Turning Lanes (LTL):* To be consistent with the previous plots, all graphs were drawn for crashes per intersection, although the number of intersection per value of the input variable was the same. The graph for the variation of the crash frequency with the total left turning lanes has been shown in Figure 5.4. The graph shows that the frequency of crashes decreases slightly when the number of left turning lanes increases at an intersection, and then it starts increasing once the number of left turning lanes exceeds 4. This implies that the frequency of crashes can be decreased at an intersection by increasing the left turning lanes to a certain extent, but if the total LTL increases beyond 4 there is an increase in the crash frequency.

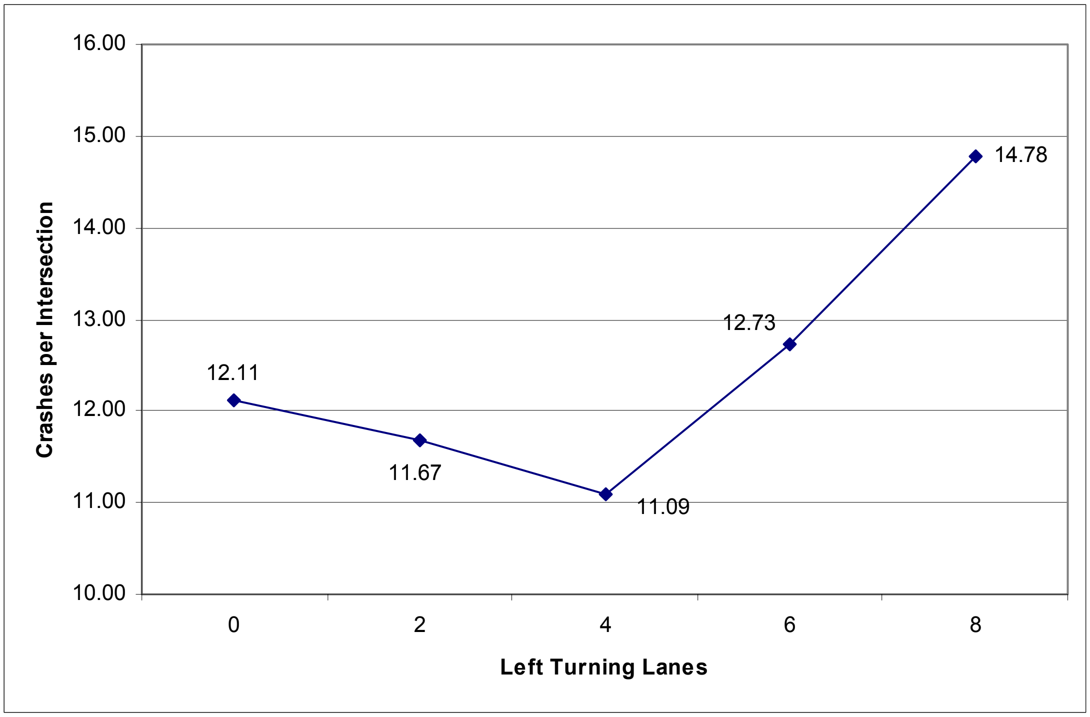

Figure 5.4 Average expected number of crashes per intersection per year for different values of total left turning lanes

4. *Major Left Turning Protected (LTP) Lanes:* From the Figure 5.5, it can be seen that an increase in the protected left turning lanes actually decreases the crash frequency. An almost linear trend can be established between the protected left turning lanes and the crash frequency. This decrease can be expected because an increase in the LTP lanes can decrease the number of left turning crashes. However, this variable has not been found significant by Keller (2004).

5. *Minor Protected Left Turning Lanes:* From the Figure 5.6, it is evident that the protected left turning lanes had a slight affect on the prediction of the crash frequency. Keller (2004) also reports an increase in the crash frequency with an increase in the protected left turning lanes on the minor road.

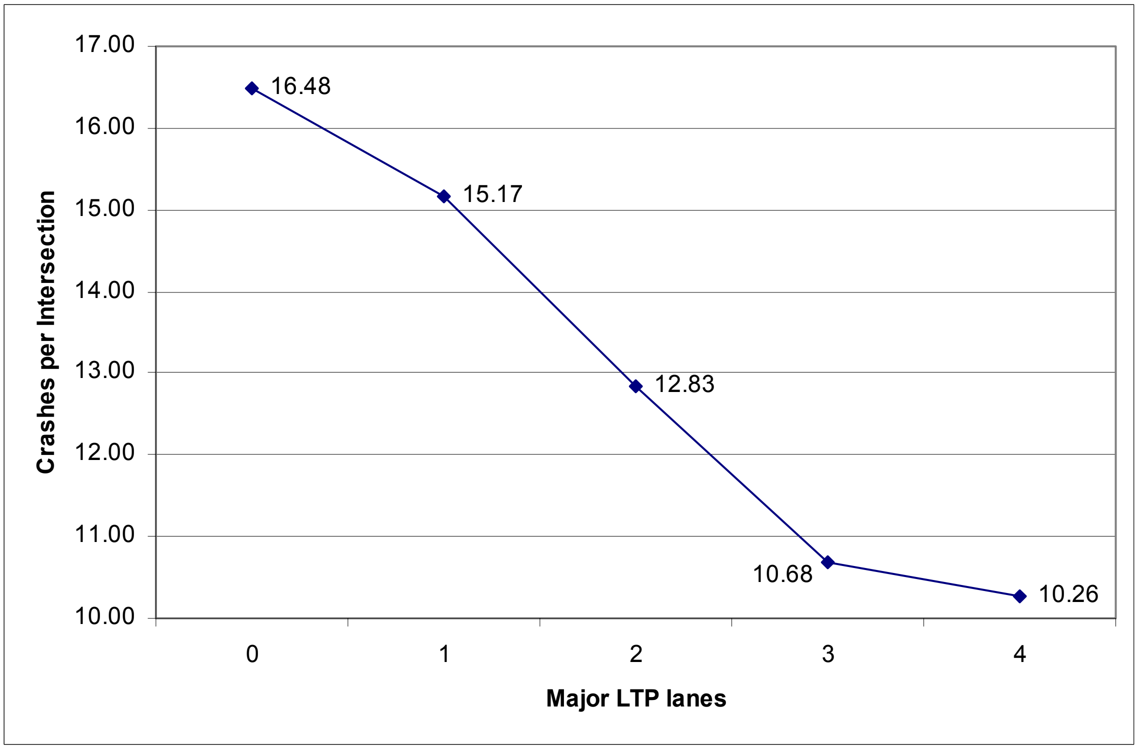

Figure 5.5 Average expected number of crashes per intersection per year for different values of protected left turning lanes on the major road

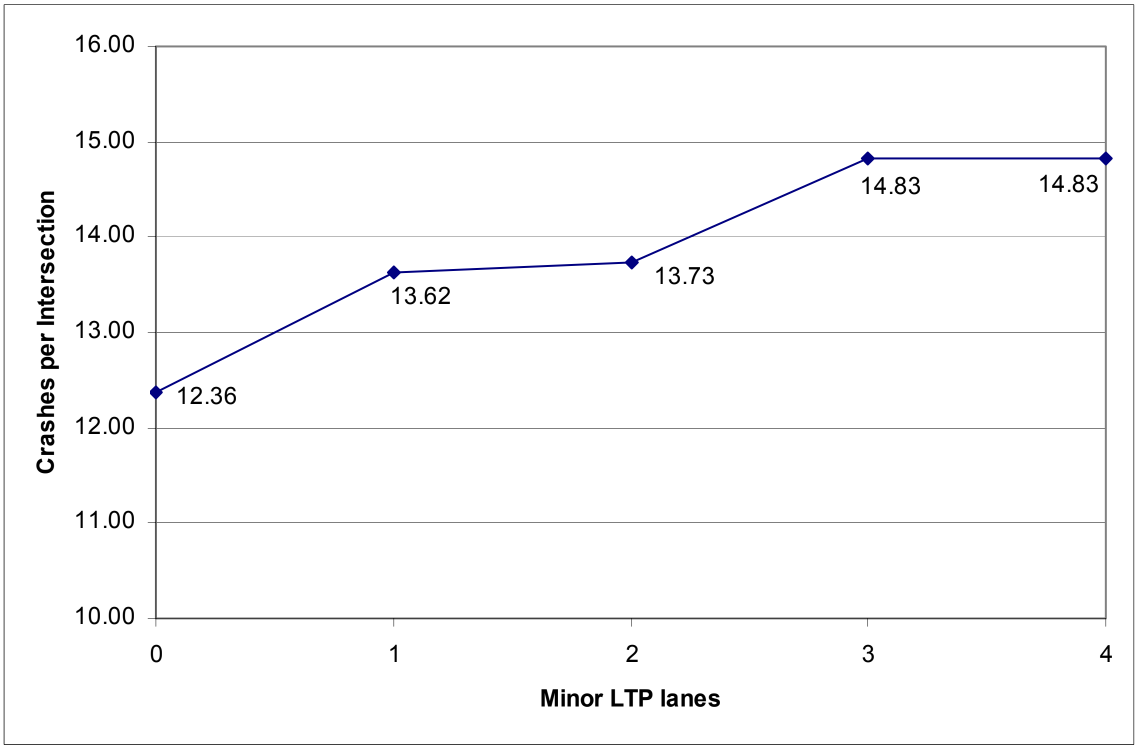

Figure 5.6 Average expected number of crashes per intersection per year for different values of protected left turning lanes on the minor road

6. *Major Channelized Right Turning Lanes:* This variable was neither found significant in classifying intersections into safe or unsafe intersections nor in predicting the crash frequency for any of the intersections. Hence it does not have any affect on the overall crash frequency at an intersection.
7. *Minor Channelized Right Turning Lanes:* As can be seen in Figure 5.7, an increase in the RTC lanes on the minor road tends to slightly decrease the crash frequency. The intersections with no RTC on the minor road can be expected to have a larger number of crashes because there is a greater possibility for the vehicle taking a right turn from the minor road on to the major road to get involved in a crash as the traffic on the major road is high. The same cannot be said about the channelization on the major road because the minor road will have a lesser traffic than the major road and hence a vehicle taking a right turn is less exposed to a crash.
8. *Major Speed Limit:* The speed limit on the major road does not have a very significant effect on the crash frequency. This can be observed in Figure 5.8. Keller (2004) finds that the crash frequency increases drastically when the speed limit is over 35.
9. *Major AADT:* As can be clearly observed in Figure 5.9, the frequency of crashes increases as the AADT on the major roadway increases. The increase is especially high between the values of 20000 and 60000. In the study by Keller (2004), the AADT was found to significantly increase the crash frequency. Poch and Mannering (1996) and Greibe (2003) also find that an increase in AADT increases the frequency of crashes at intersections.

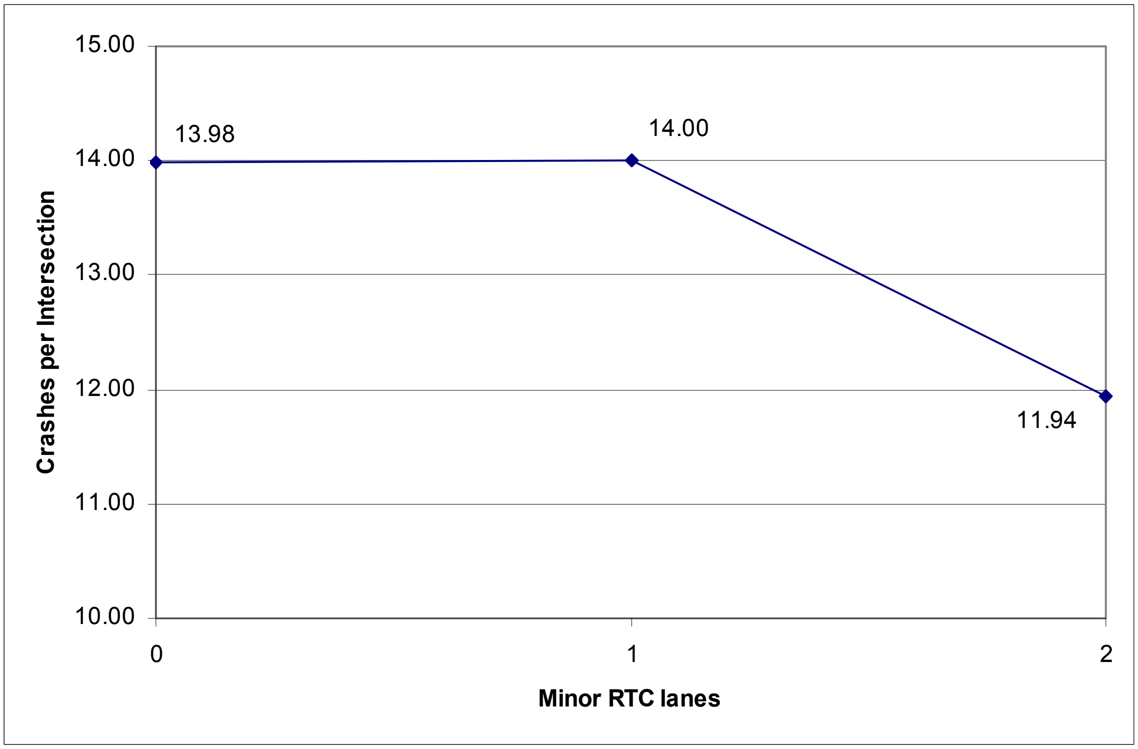

Figure 5.7 Average expected number of crashes per intersection per year for different values of channelized right turning lanes on the minor road

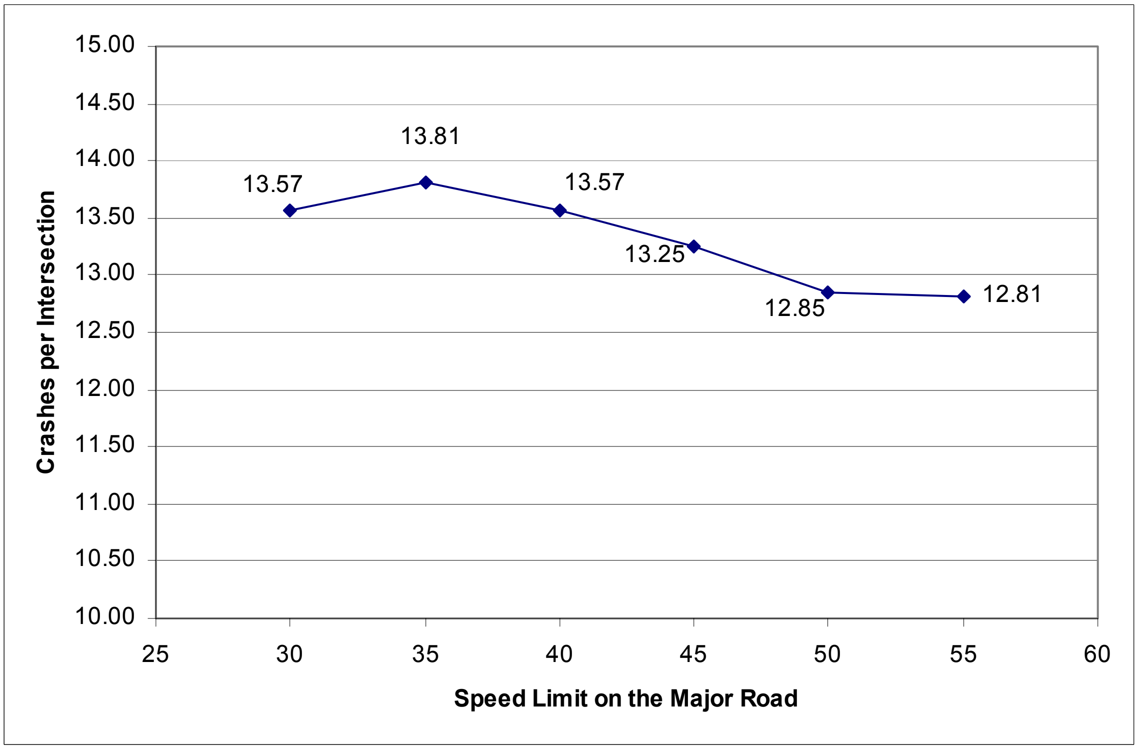

Figure 5.8 Average expected number of crashes per intersection per year for different values of Speed Limit on the major road

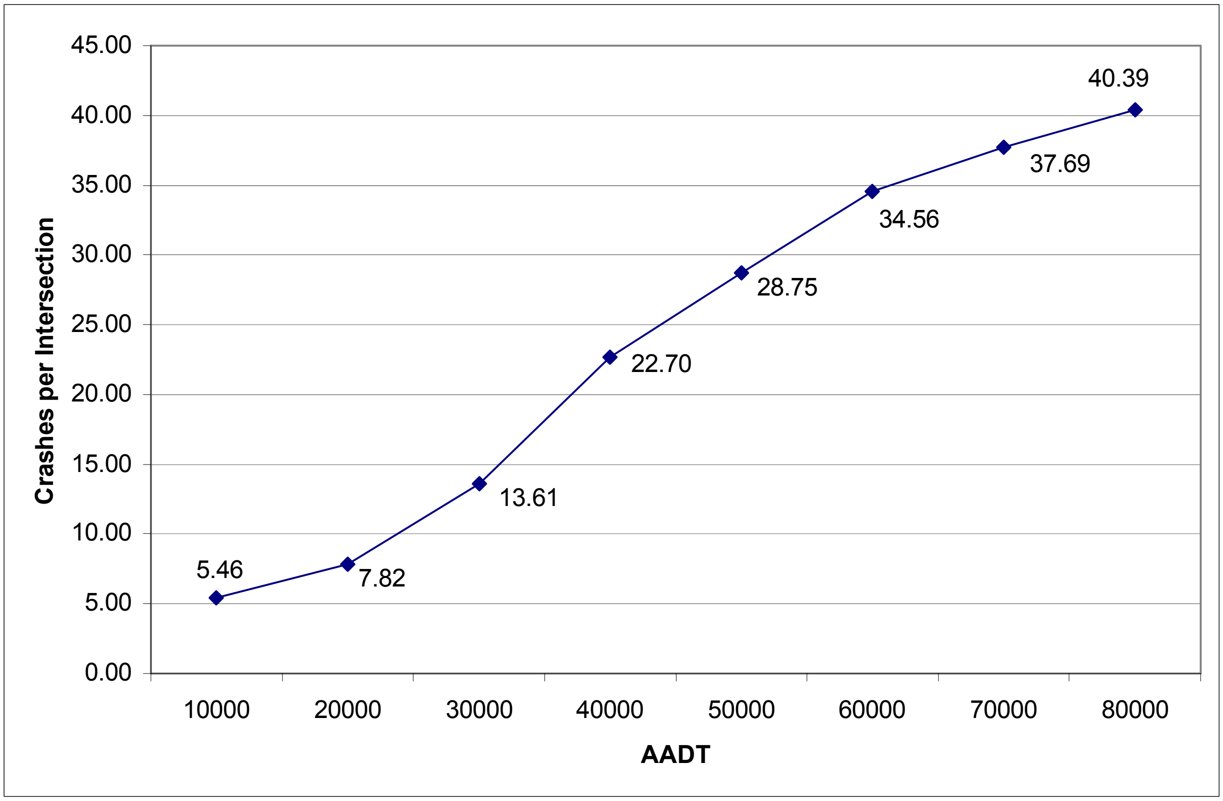

Figure 5.9 Average expected number of crashes per intersection per year for different values of AADT on the major road

Hence the results obtained from this method were very similar and comparable to those obtained by other studies. Therefore, it was decided to use the same methodology to predict the crashes of different collision types.

5.5 Predicting Rear End Crash Frequency Based on Number of Lanes

Rear end crashes form the majority of the crashes in the database. Almost half the crashes in the database are rear end crashes. Rear end crash frequencies were predicted by using the same methodology as described section 5.4. This section briefly describes the results obtained by using this method and compares these results to the results obtained in other studies.

As a first step, the average number of rear end crashes was determined for intersections with different number of total lanes. These have been shown in Table 5-13. The values clearly indicate an increasing crash frequency with an increase in the total number of lanes at the intersection, thus allowing us to use the model developed in section 5.4.

Table 5-13 Average number of rear end crashes for different groups of intersections

Total Lanes at an Intersection	Average Number of Rear End Crashes
3 to 5	3.11
6 to 10	4.65
11 to 15	9.98
16 and above	19.72

5.5.1 Classification of Intersections

The intersection database was first classified into safe and unsafe intersections based on Table 5-13. Both the MLP and PNN models were used to classify these intersections. The MLP neural network model gave a highest accuracy of 63.31% with 10 hidden nodes. The accuracies for the safe and unsafe intersections were 63.53% and 63.06% respectively. The PNN model had a highest accuracy of 65.222% with accuracies of 68.46% and 62% for the safe and unsafe intersections respectively. Therefore the PNN model was considered as the better model for classifying the intersections into safe or unsafe intersections.

The AADT, major through lanes and minor through lanes were identified as the factors in the classification process. A combination of these variables gave an average accuracy of 68.22%.

5.5.2 Predicting the frequency of rear end crashes for safe intersections

The GRNN model performed better with a least RMSE of 1.40 and a corresponding MAPE of 33.04%. The MLP model showed a minimum RMSE of 1.46 with a corresponding MAPE of 40.53%. Since the GRNN model performed better with both RMSE and MAPE criteria, it was used to identify the significant variables for safe intersections. The variables found significant in predicting the rear end crashes at safe intersections were major LTP lanes, major lanes, AADT and minor LTP lanes. The minimum RMSE obtained by predicting the rear crashes using these significant variables was 1.37, which is equal to the RMSE obtained by using all the input variables. This error is significantly less than the error in predicting the total crashes at safe intersections.

5.5.3 Predicting the frequency of rear end crashes for unsafe intersections

The rear end crash frequency at unsafe intersections was predicted with a lesser RMSE and MAPE by the MLP model. The values attained by the MLP model for RMSE and MAPE were 5.14 and 41.73% respectively, while those attained by the GRNN model were 5.35 and 43.13% respectively. Hence the MLP neural network model was used in identifying the significant variables and also in the simulation process.

The significant variables identified using the MLP model were: AADT, number of minor lanes, speed limit on the major road, minor RTC lanes, minor LTP lanes and major LTP lanes. The minimum RMSE obtained was 4.3. This was much better than the RMSE obtained using all nine input variables.

5.5.4 Estimating a Pattern in Significant Variables

The same simulation database as used in section 5.4.7 containing 98928 crashes was used to estimate the pattern in which the significant variables affect the frequency of the rear end crashes. The PNN model was used to classify the intersections into safe and unsafe categories. GRNN model predicted the frequency of safe intersections whereas MLP model estimated the frequency of unsafe intersections. The results were combined and averaged for each value of the input variable. The following points discuss the association of the significant input variables with the frequency of rear end crashes:

1. *Major Lanes:* As can be seen in Figure 5.10, the average number of crashes occurring on 4 and 6 lanes is almost double the number for 2 lanes. Keller (2004) and Poch and Mannering (1996) find an increasing trend in the rear end accidents with an increase in major lanes. An increase in number of lanes can be related to an increase in the traffic volume on the roadway. Greater volume implies that the vehicles move closely, thereby increasing the possibility of a rear end crash. Hence an increase in the number of lanes on the major road can be related to an increase in the rear end crashes.

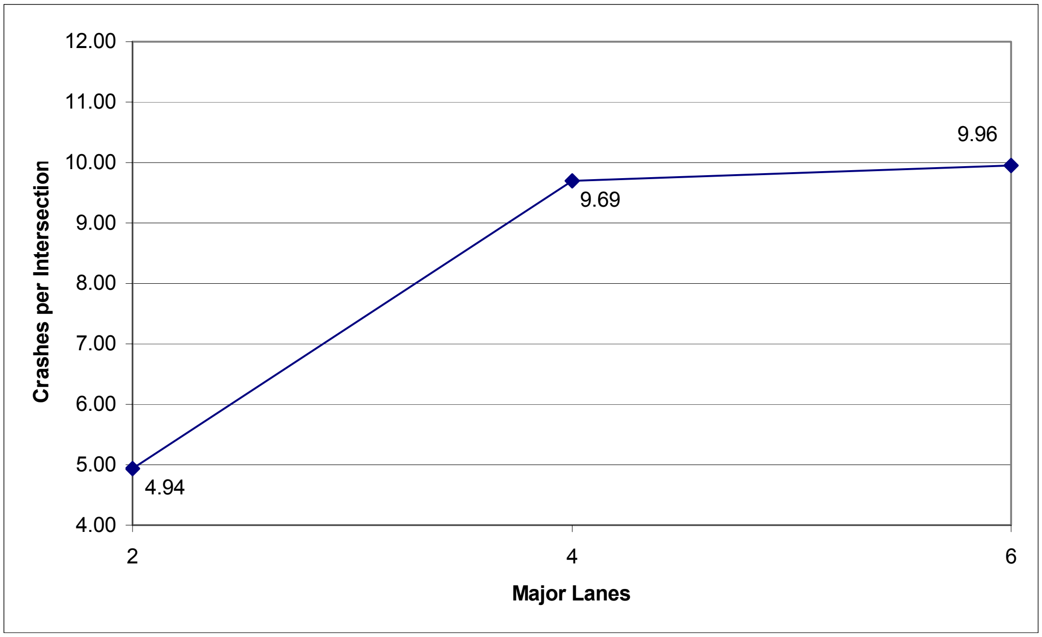

Figure 5.10 Average expected number of rear end crashes per intersection per year for different values of through lanes on the major road

2. *Minor Lanes:* The increase in minor lanes was also found to affect the rear end crashes. This has been depicted in Figure 5.11. An intersection with 6 minor lanes can be expected to have almost double the number of rear end crashes as that of an intersection with 2 minor lanes. The reasoning is same as followed for the major roads: an increase in minor roads usually means that the size of the intersection is increasing, which in turn can be due to the high traffic volume on minor roadway. This increases the possibility of rear end crashes.

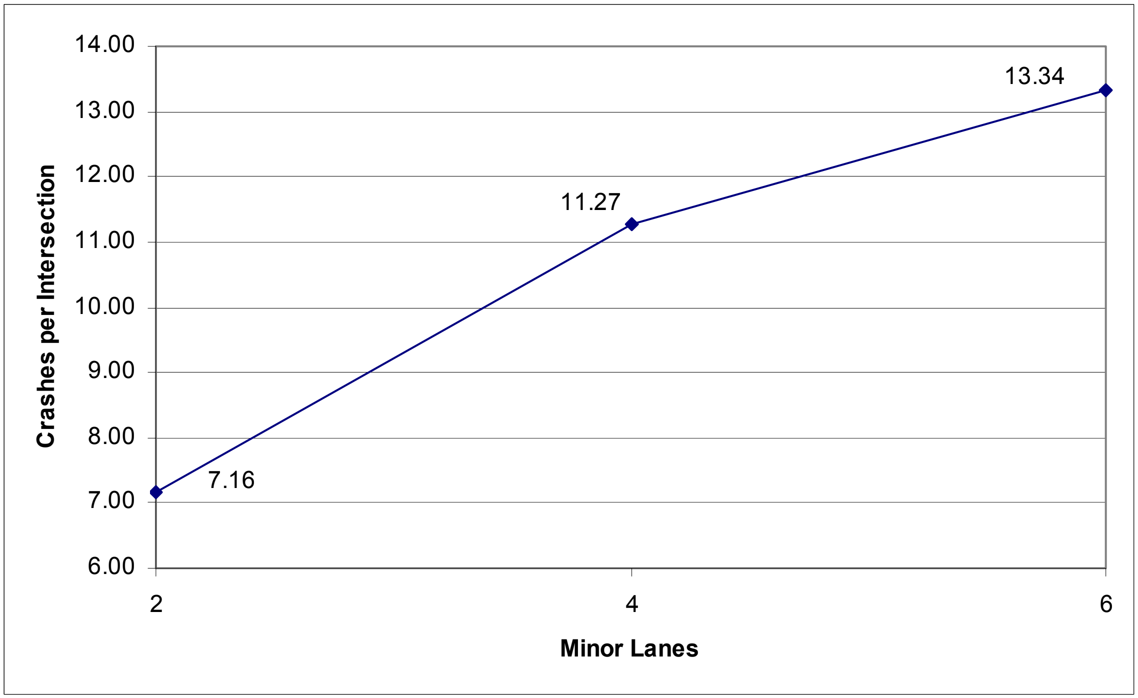

Figure 5.11 Average expected number of rear end crashes per intersection per year for different values of through lanes on the minor road

3. *Total Left Turning Lanes:* Since this variable was not found significant in either the classification of prediction models, it does not have any affect on the rear end crashes.

4. *Major LTP:* Figure 5.12 shows that an increase in protected left turning lanes on the major road also tends to increase the number of rear end crashes. Keller (2004) also reports an increasing trend in rear end crashes with an increase in the major LTP lanes. The possible reasoning for this phenomenon is that an intersection with a higher number of left turning lanes can be expected to have a higher left turning volume. An increased turning volume leads to an increased tendency in the driver to maneuver the left turn before the signal is red. This sometimes causes confusion among drivers wherein the driver of the lead vehicle decides not to take a left turn when the signal is turning red, but the driver in the

following vehicle intends to take a left turn but rear ends the vehicle in lead instead.

5. *Minor LTP:* The number of rear end crashes has been observed to increase when LTP lanes are present on the minor roadway, for reasons similar to the ones given for the major LTP lanes. Figure 5.13 illustrates such an occurrence. An increasing trend has also been observed by Keller (2004). Although the graph shows a decrease after 2 minor LTP lanes, the decrease is insignificant.
6. *Major RTC:* Since the RTC lanes were not significant in any of the models, they did not affect the rear end crashes in any way.

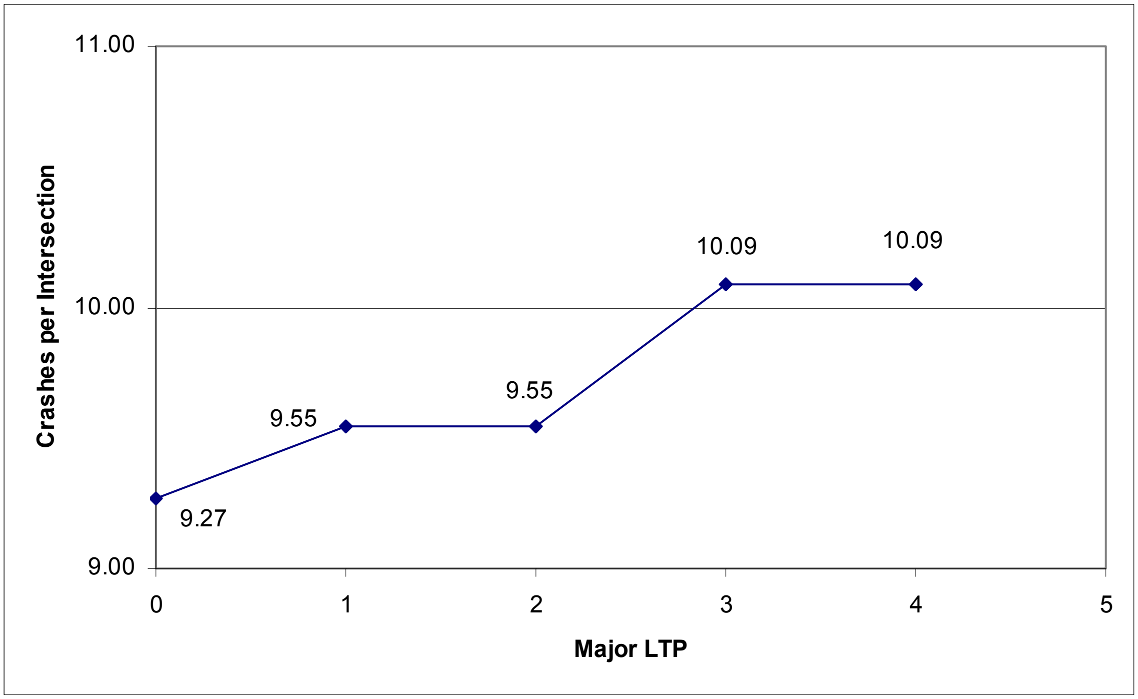

Figure 5.12 Average expected number of rear end crashes per intersection per year for different values of protected left turning lanes on the major road

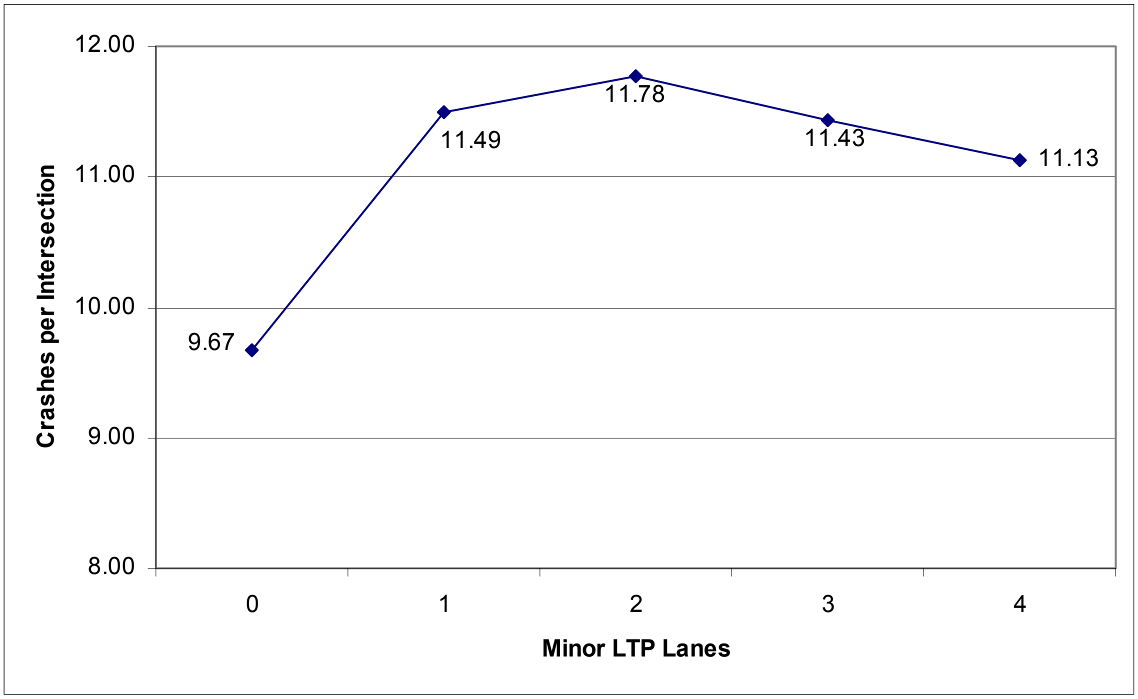

Figure 5.13 Average expected number of rear end crashes per intersection per year for different values of protected left turning lanes on the minor road

7. *Minor RTC:* From Figure 5.14, it can be seen that an increase in the RTC lanes on the minor road leads to an increase in the rear end crashes. The possible reasoning for this is similar to the rear end crashes during a left turn: the driver following a vehicle thinks that the vehicle is taking a right turn and accelerates, but the vehicle in lead slows down to stop. This leads to a rear end crash.

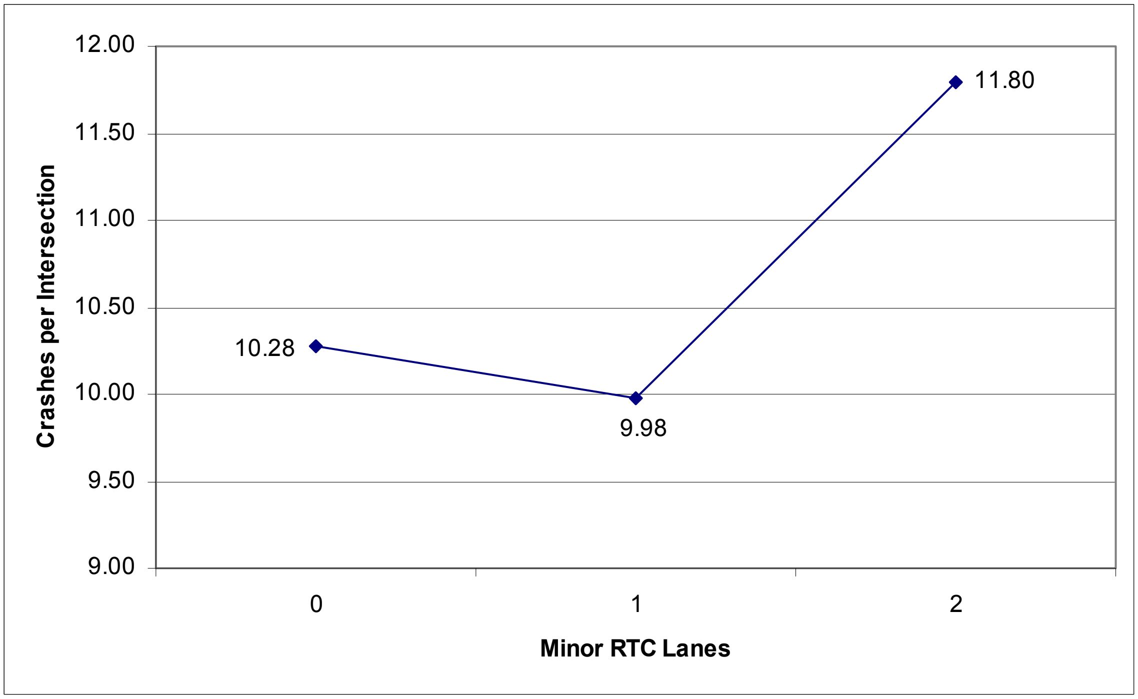

Figure 5.14 Average expected number of rear end crashes per intersection per year for different values of channelized right turning lanes on the minor road

8. *Speed Limit:* Rear end crashes are found to increase linearly with the increase in speed limit on the major road in Figure 5.15. The possible reason for this phenomenon is that higher speed limit implies a greater distance is required to stop the car, and if a minimum distance is not maintained between the cars a rear end crash is possible when the vehicles decelerate. An increasing trend is also observed by Keller (2004) and Poch and Mannering (1996).

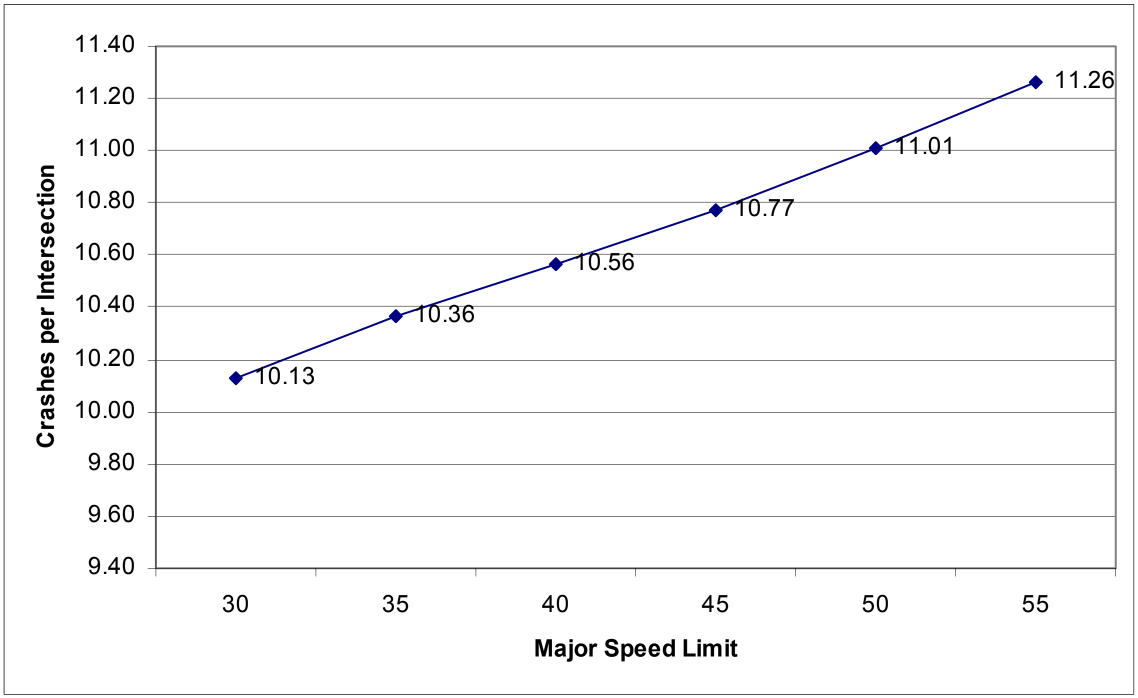

Figure 5.15 Average expected number of rear end crashes per intersection per year for different values of speed limit on the major road

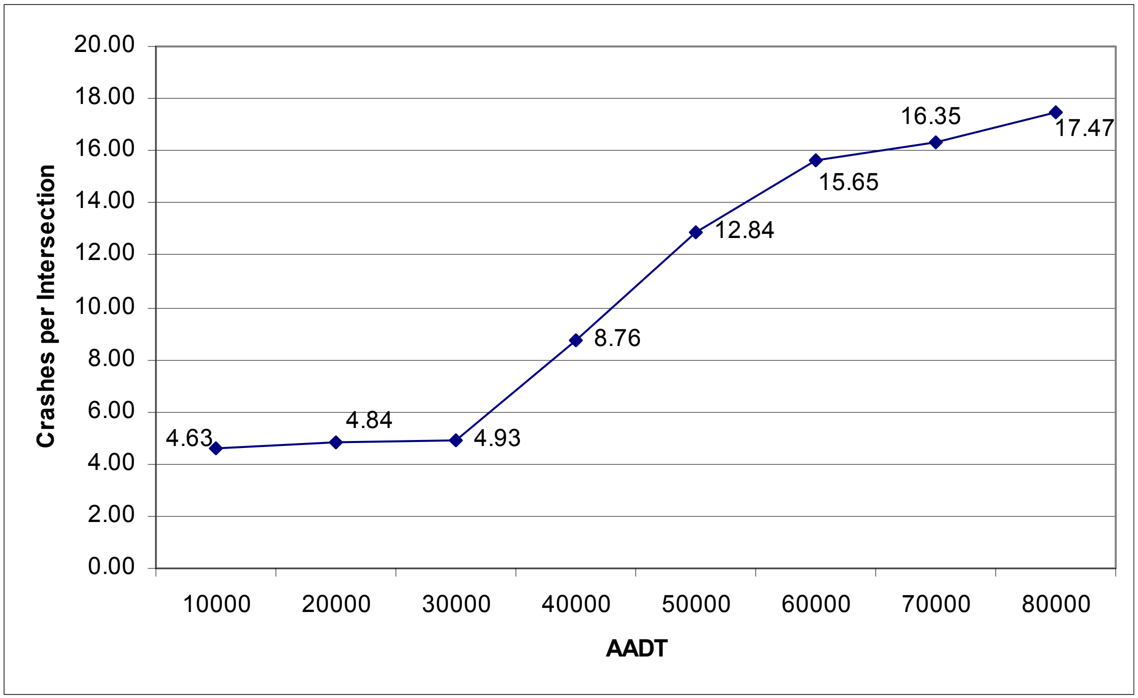

Figure 5.16 Average expected number of rear end crashes per intersection per year for different values of AADT on the major road

9. *AADT:* The frequency of rear end crashes shows a huge increasing trend when the AADT on the major roadway is increased, as can be seen in Figure 5.16. Keller (2004) and Greibe (2003) also find an increasing trend in the rear end crashes with an increase in the AADT.

The model performed well with the new model. The errors for both the prediction models are reasonably low. The output obtained seems reasonable. Hence this method can be used effectively for the prediction of the frequency of rear end crashes.

5.6 Predicting Frequency of Angle Crashes Based on Number of Lanes

Angle crashes formed the next highest percentage of crashes in the database after the rear end crashes. The first step was to identify the trend in the crash frequency with an increase in the total number of lanes at the intersection. As can be observed from Table 5.14, the angle crashes clearly show an increasing trend with an increase in the total number of lanes at the intersections.

Table 5-14 Average number of angle crashes for different groups of intersections

Total Lanes at an Intersection	Average Number of Angle Crashes
3 to 5	2.90
6 to 10	3.01
11 to 15	5.19
16 and above	6.54

5.6.1 Classification of Intersections

Based on Table 5.14, the intersections in the database were classified to safe and unsafe intersections for angle crashes. Using the MLP neural network, a test classification accuracy of 62.57% was obtained, whereas the PNN model gave a highest accuracy of 64.97%. The PNN model was judged as the better model with the classification of accuracies of safe and unsafe intersections as 66.1% and 63.75% respectively.

The factors found significant in this classification process were: number of protected LTP lanes on the major road, the number of lanes on the major road, the number of LTP lanes on the minor road and the number of lanes on the minor road. The accuracy for this combination of variables was 68.24%, which was significantly higher than the model built using all variables.

5.6.2 Predicting the frequency of angle crashes for safe intersections

Both the MLP and GRNN models performed equally well in predicting the number of angle crashes for safe intersections. The RMSE for both the models was around 0.8 and the MAPE was around 36%. The MLP NN model was used to identify the significant variables, which turned out to be the number of minor lanes, major lanes and AADT.

5.6.3 Predicting the frequency of angle crashes for unsafe intersections

The MLP NN model predicted the frequency of angle crashes for unsafe intersections with a RMSE of 3.5 and MAPE of 52%. The GRNN model predicted the same with RMSE and MAPE of 3.3 and 49% respectively. Based on a lower RMSE, the GRNN model was chosen for selecting the significant variables. The major lanes, minor lanes and the LTP lanes on the major road turned out to be the significant variables in the model. The RMSE for the restrained model turned out to be 2.65, which is significantly lower from the RMSE of the overall model.

5.6.4 Estimating a Pattern in Significant Variables

The pattern among the significant variables was identified by predicting the number of angle crashes occurring at the intersections in the simulation database. The

PNN model was used to classify the intersections into safe and unsafe intersections and the MLP and GRNN models were used to predict the number of crashes occurring at safe and unsafe intersections respectively. The results were plotted to establish the relationship between the angle crashes and the input variables. The following points explain the relationships obtained:

1. *Major Lanes:* An increase in the number of through lanes on the major roadway tremendously increases the chances of angle crashes, as can be observed in Figure 5.17. The possible reason for this trend is that an increase in the number of through lanes implies a greater amount of traffic flowing in the through lanes, thereby increasing the possibility of an angle crash.

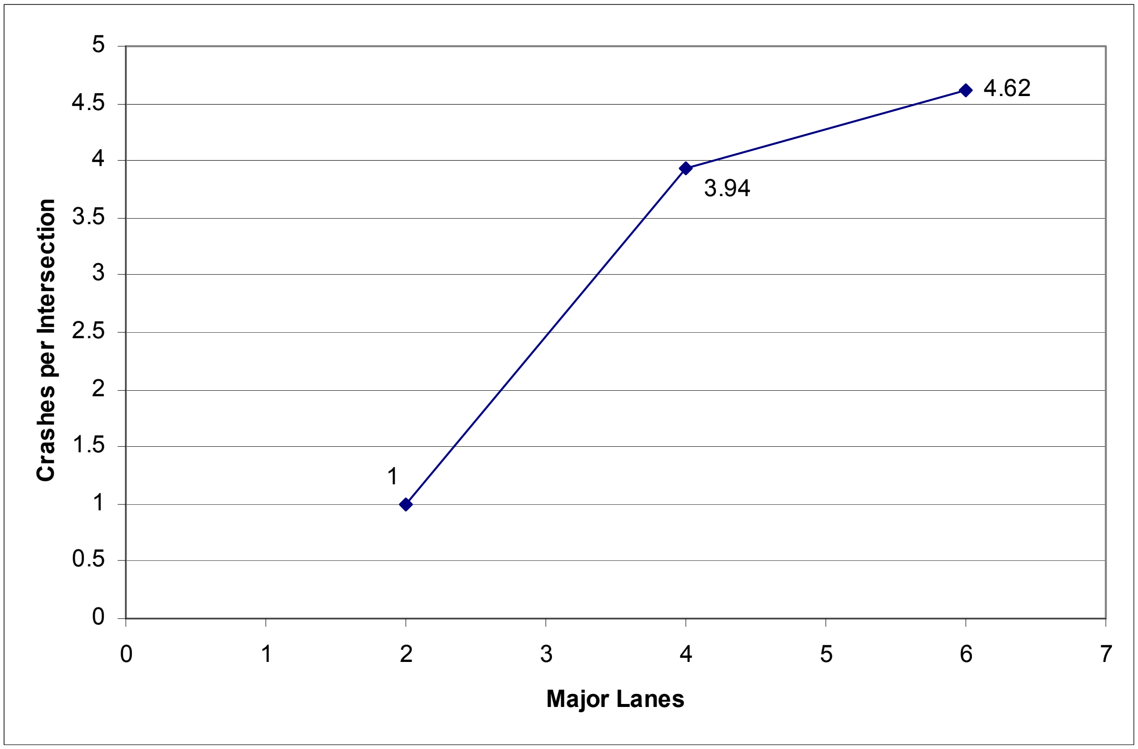

Figure 5.17 Average expected number of angle crashes per intersection per year for different values of through lanes on the major road

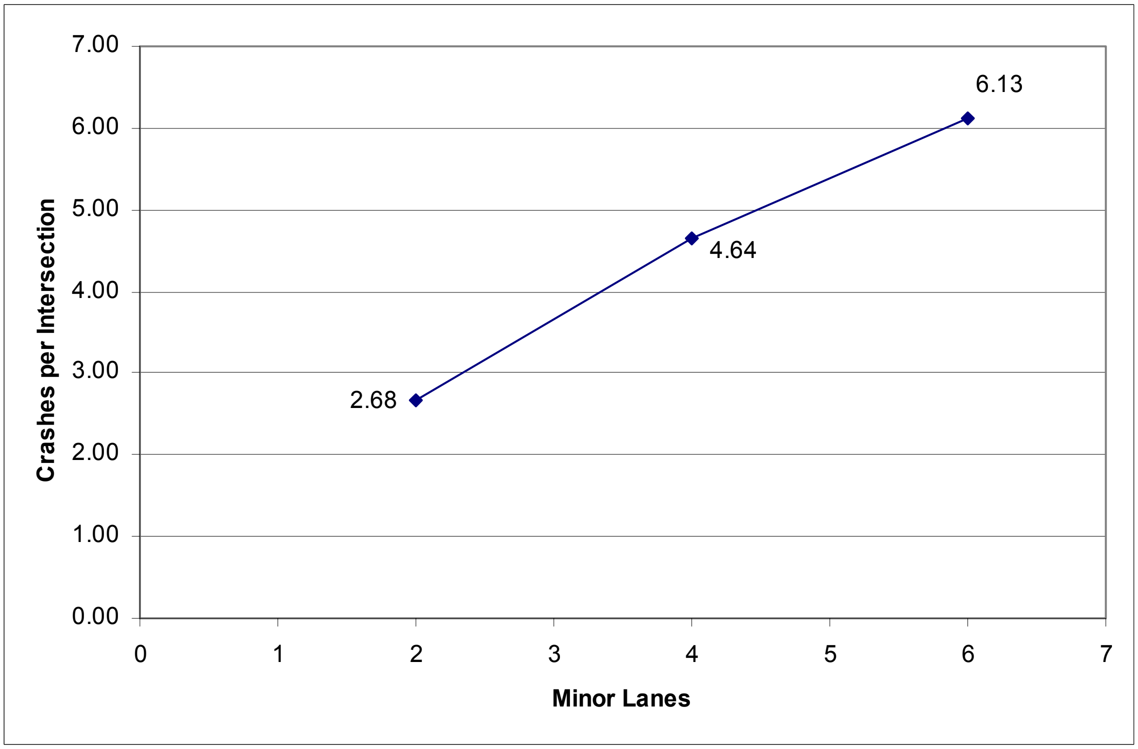

Figure 5.18 Average expected number of angle crashes per intersection per year for different values of through lanes on the minor road

2. *Minor Lanes:* As can be observed in Figure 5.18, the frequency of angle crashes also shows an increasing trend with an increase in the number of minor lanes. Keller(2004) found this variable to be the most important variable in predicting angle crashes and observed an increasing trend in the frequency of angle crashes with an increase in the minor lanes.

3. *Left Turing Lanes:* Since the left turning lanes were not found to be significant in any of the models, this variable does not affect the frequency of angle crashes.

4. *Major LTP Lanes:* The protected left turning lanes on the major road were found to decrease the number of angle crashes, as can be seen in Figure 5.19. Poch and Mannering (1996) also finds that presence of protected left turning lanes reduces the number of angle crashes.

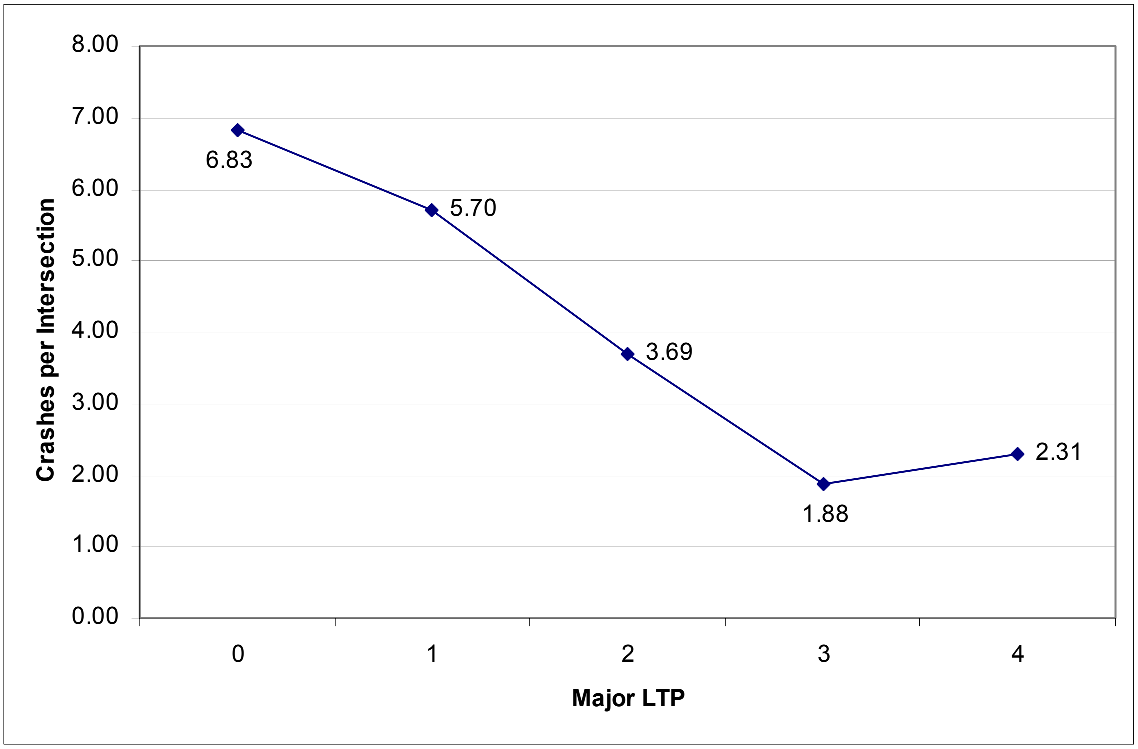

Figure 5.19 Average expected number of angle crashes per intersection per year for different values of LTP lanes on the major road

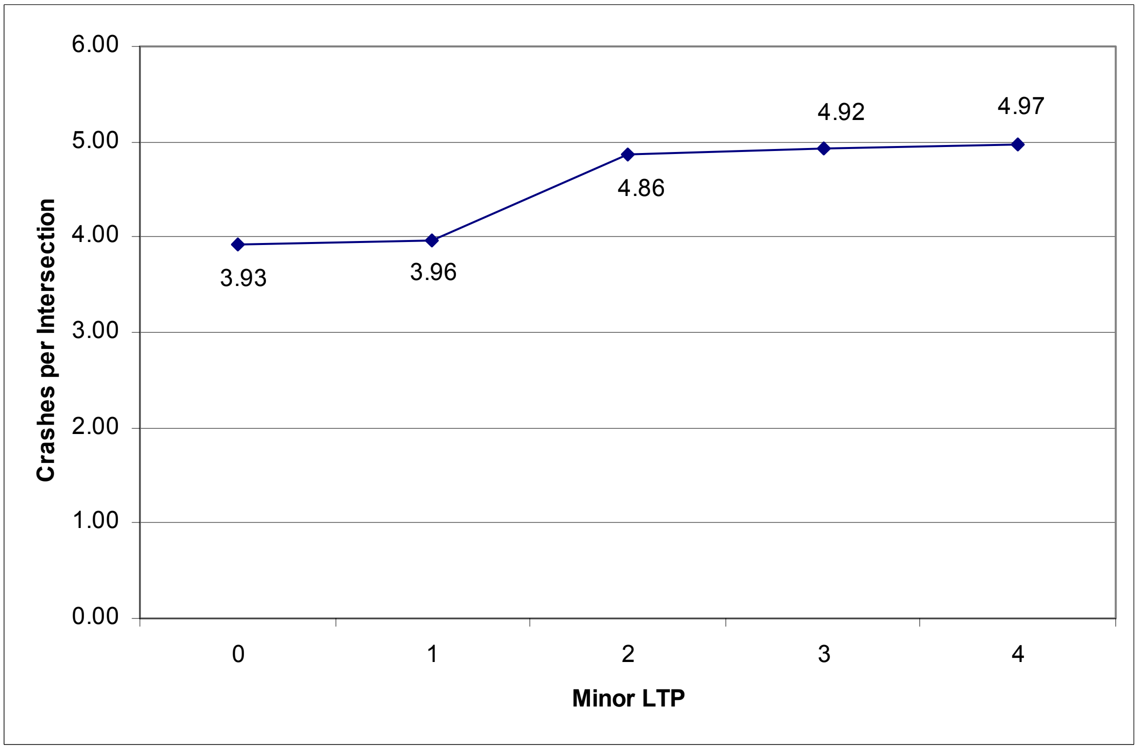

Figure 5.20 Average expected number of angle crashes per intersection per year for different values of LTP lanes on the minor road

5. *Minor LTP Lanes:* The protected left turning lanes on the minor road have been observed to increase the number of angle crashes, as can be seen in Figure 5.20. A similar phenomenon has been observed by Keller (2004).
6. *Major AND Minor RTC Lanes:* As this factor was not found significant in any of the models, it does not affect the frequency of angle crashes at the intersections.
7. *Major Speed Limit:* As this factor was not found significant in any of the models, it does not affect the frequency of angle crashes at the intersections.
8. *Major AADT:* As can be seen in Figure 5.21, an increase in the AADT on the major road results in a slight increase in the angle crashes. Keller (2004) finds the AADT on the minor roadway to be a very important factor in predicting the frequency of angle crashes.

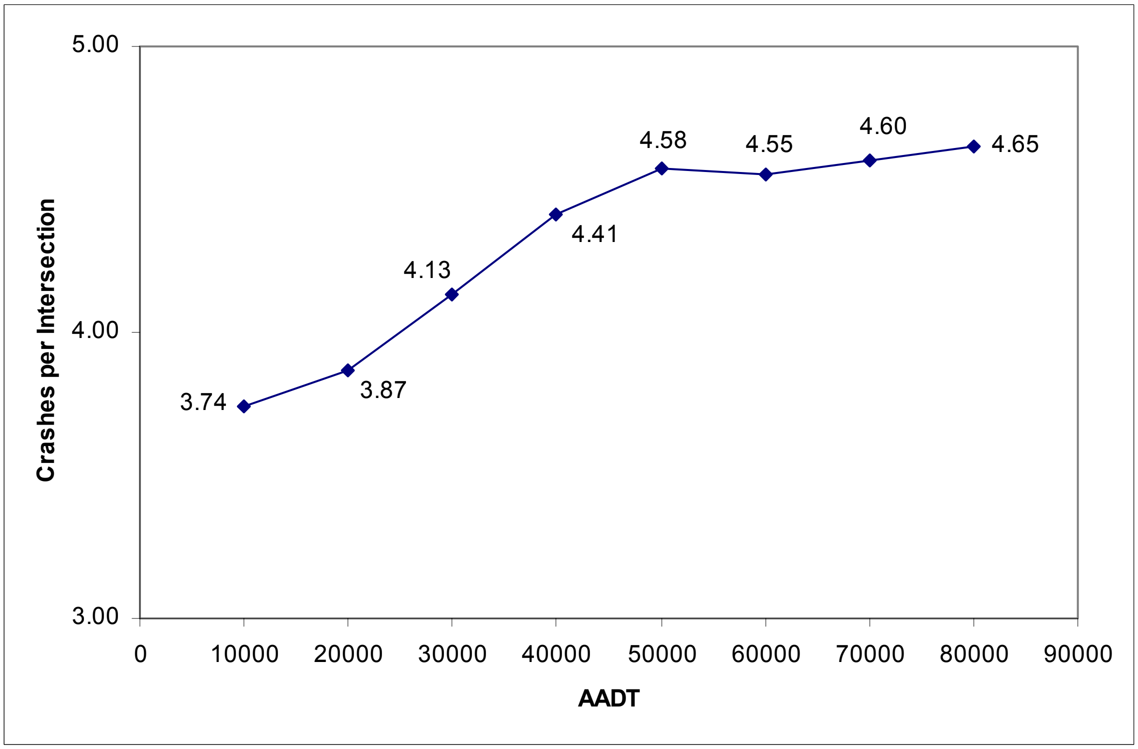

Figure 5.21 Average expected number of angle crashes per intersection per year for different values of AADT on the Major road

5.7 Predicting the Frequency of Turning Crashes Based on Number of Lanes

The turning crashes consisted of both left turning and right turning crashes. In order to predict the frequency of these crashes using the method described in section 5.4, the increasing trend of turning crashes with an increasing in the total number of lanes at the intersections had to be established. Hence the average number of turning crashes for all intersections was determined and tabulated in Table 5.15. The number of turning crashes definitely seems to increase with an increase in the size of the intersection. Therefore, the crash frequencies can be predicted using the method of classifying the intersections into safe and unsafe type and then predicting the crash frequencies for each type of intersection.

Table 5-15 Average number of turning crashes for different groups of intersections

Total Lanes at an Intersection	Average Number of Turning Crashes
3 to 5	1.69
6 to 10	1.97
11 to 15	3.83
16 and above	5.86

5.7.1 Classification of Intersections

The intersections in the database were classified into safe and unsafe intersections using the MLP and PNN models. The MLP neural network classified the intersections with a highest accuracy of 63.03%, whereas the PNN classified the intersections with an accuracy of 64%. Hence the PNN was judged as a better model and was used to identify the significant variables in the classification process. The following variables were found to be significant in the model: AADT, major LTP lanes, major RTC lanes, major lanes, and minor LTP lanes. The accuracy for the model with this combination of variables was 64.43%.

5.7.2 Predicting the frequency of turn crashes for safe intersections

Both MLP and GRNN models performed equally well in predicting the turning crashes, with RMSE of 0.622 and 0.615 respectively. The MAPE for both models were around 47%. The GRNN model was chosen to identify the significant variables as the GRNN can train and test the neural networks faster compared to the MLP neural network model. The following variables were found significant: left turning lanes, minor lanes, major lanes, major RTC lanes, speed limit and AADT.

5.7.3 Predicting the frequency of turn crashes for unsafe intersections

The MLP and GRNN models again performed equally well with RMSE of 2.3. The MAPE of both the models was close to 45%. The GRNN model was used to identify the significant variables, which were minor lanes, major lanes, major LTP lanes and AADT.

5.7.4 Estimating a Pattern in the Significant Variables

The models developed in the previous sections were used to predict the number of turning crashes occurring at the intersections in the simulation database. The final output of the three models was plotted with each of the significant variables to establish the pattern in which the input variables affect the frequency of the turning crashes. The following points give a detailed explanation of the relationships between the input variables and the frequency of the turning crashes:

1. *Major Lanes:* As can be seen in Figure 5.22, an increase in the number of lanes on the major road tends to increase the number of turning crashes. An increase in the through lanes means that a left (or right) turning vehicle has a greater exposure to the through traffic commuting from the opposite direction. Hence a turning crash is more likely with an increase in the number of lanes on the major road. This variable has been found to be significant in predicting the right turning crashes by Keller (2004).

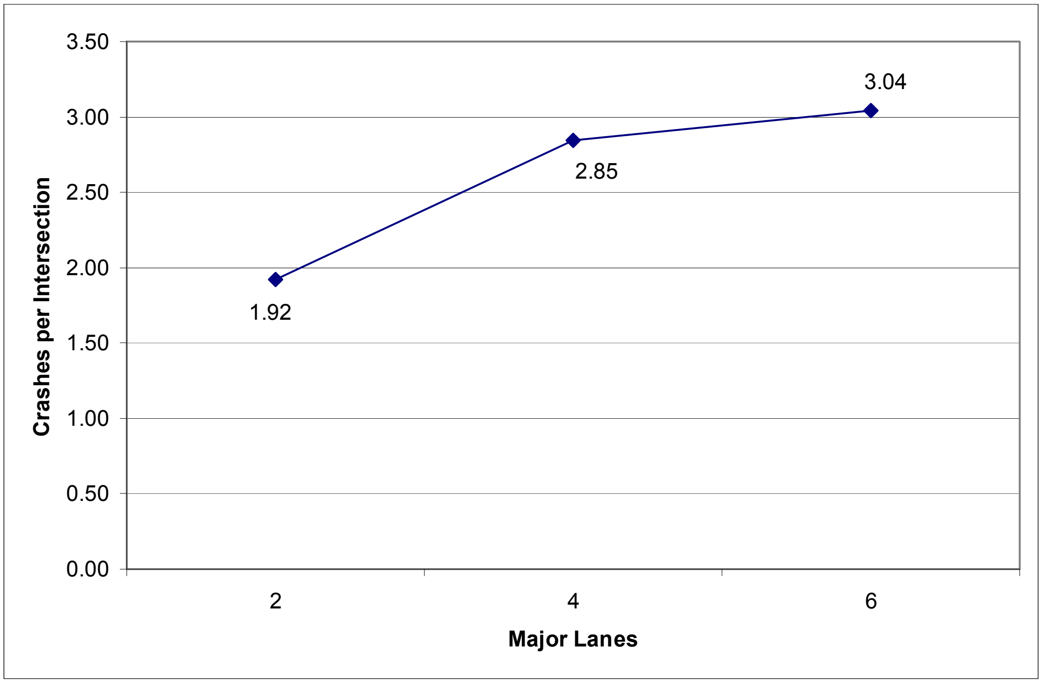

Figure 5.22 Average expected number of turning crashes per intersection per year for different number of through lanes on the Major road

1. *Minor Lanes:* An increase in the number of lanes on the minor road also tends to increase the number of turning crashes at intersections, as can be observed in Figure 5.23. This variable has been found to be significant in predicting the frequency of left turning crashes by Keller (2004).

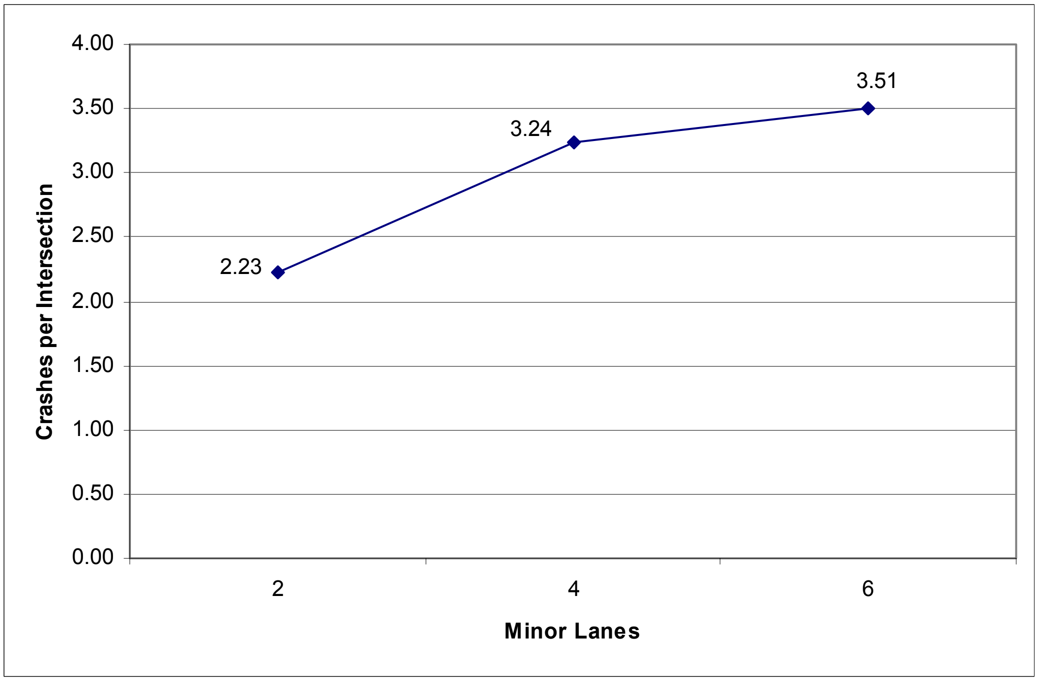

Figure 5.23 Average expected number of turning crashes per intersection per year for different number of through lanes on the Minor road

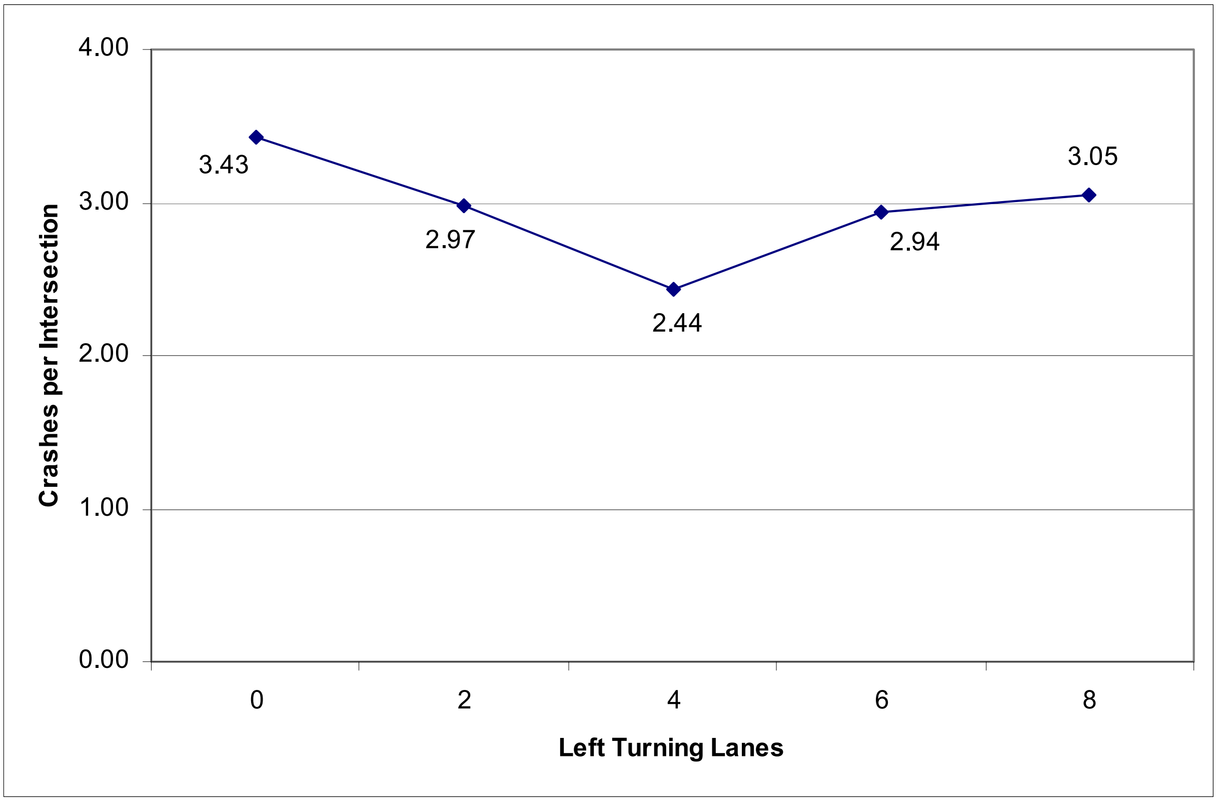

Figure 5.24 Average expected number of turning crashes per intersection per year for different number of left turning lanes

2. *Left Turning Lanes:* The turning crashes show a slightly decreasing trend with an increase in the left turning lanes, and then show an increasing trend when the number of left turning lanes is above 4. This suggests that left turning lanes decrease turning crashes, but more than four left turning lanes tend to increase the turning crashes.
3. *Major LTP Lanes:* It is clear from Figure 5.25 that an increase in LTP lanes on the major road leads to a decrease in the turning crashes. This illustrates that protected left turning lanes prevent the left turning crashes. Major LTP lanes has been found significant by Keller (2004) in predicting both left and right turning crashes.

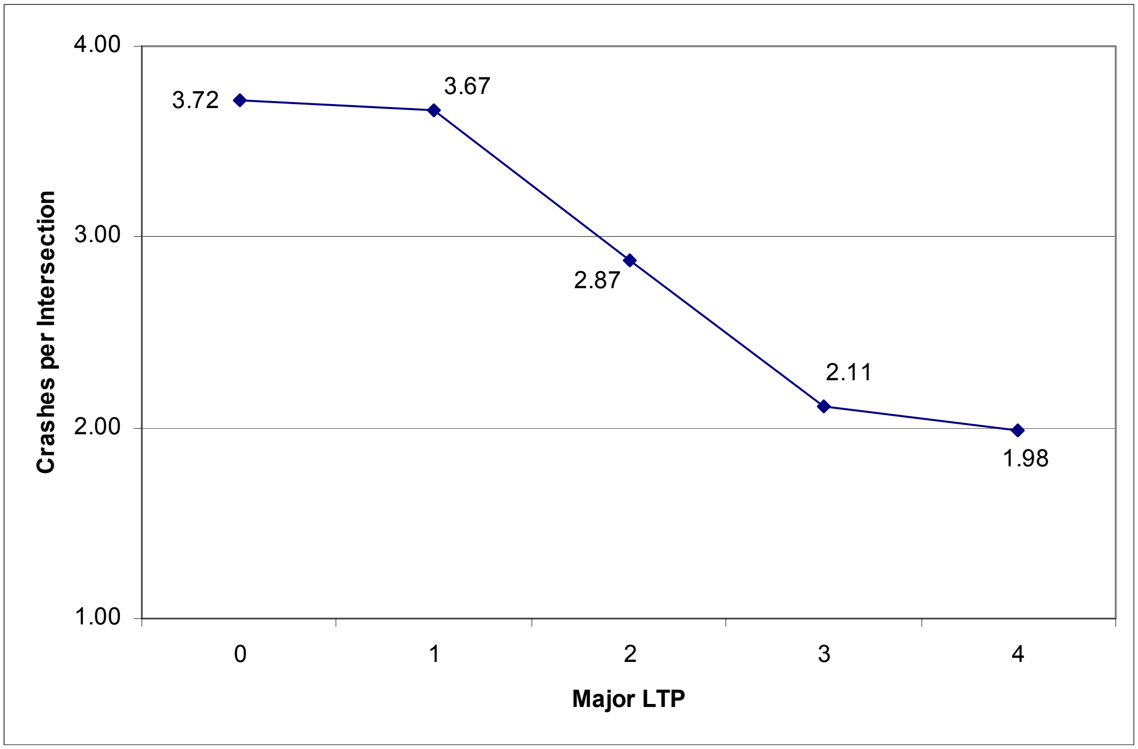

Figure 5.25 Average expected number of turning crashes per intersection per year for different number of major LTP lanes

4. *Minor LTP Lanes:* The turning crashes have been found to increase very slightly with an increase in the minor LTP lanes, as can be observed in Figure 5.26. Keller (2004) also finds that this factor tends to increase the left turning crashes, and finds this to be one of the most important factors in predicting the left turning crashes. But in the present study, this variable has been found to be significant only in judging if the intersections are safe or unsafe. Since the turning crashes show an increasing trend with an increase in the LTP lanes, it can be concluded that increasing LTP lanes on the minor roadway makes the intersection less safer with regard to turning crashes.
5. *Major RTC Lanes:* Channelized right turning lanes are usually provided at intersections when the right turning volume is large. Higher right turning volume indicates that more right turning vehicles are exposed to the traffic from other directions, and thus more right turning crashes can be expected. The same phenomenon has been observed in the simulation output, and can be observed in Figure 5.27.

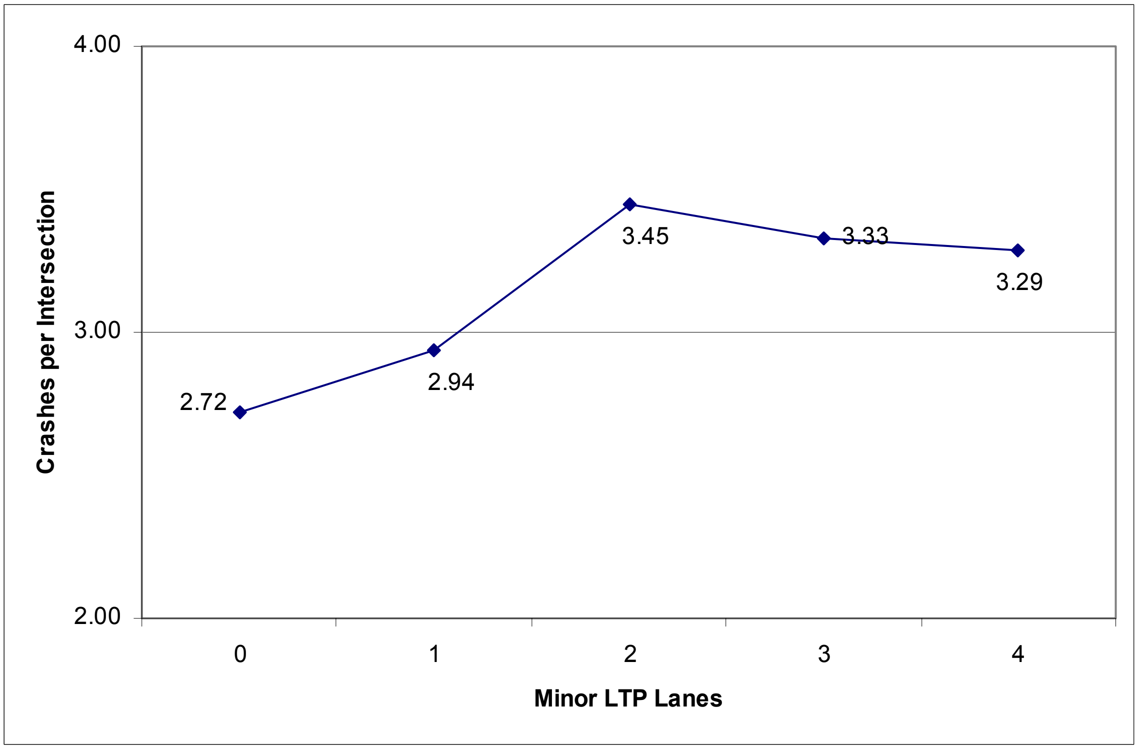

Figure 5.26 Average expected number of turning crashes per intersection per year for different number of minor LTP lanes

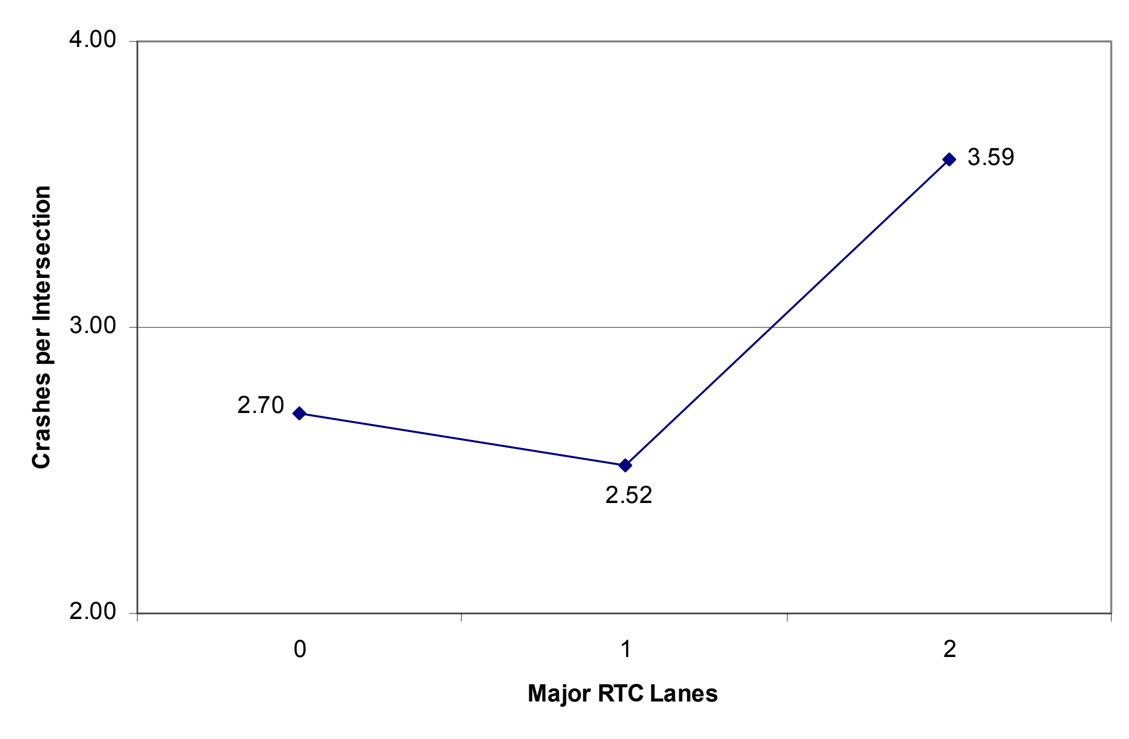

Figure 5.27 Average expected number of turning crashes per intersection per year for different number of major RTC lanes

6. *Minor RTC Lanes:* As this variable was not found to be significant in any of the models, it can be considered not to affect the frequency of the turning crashes.
7. *Major Speed Limit:* Although the speed limit was a significant factor in predicting the turning crashes on safe intersections, it was not found to be a very significant factor in the simulation phase. This was reflected in the simulation output shown in Figure 5.28, where the turning crashes vary only very slightly with an increase in the speed limit. Poch and Mannering (1996) found the speed limit for the opposing approach to be the least significant factor in predicting approach turning crashes. The major speed limit was also found to be significant by Keller (2004).

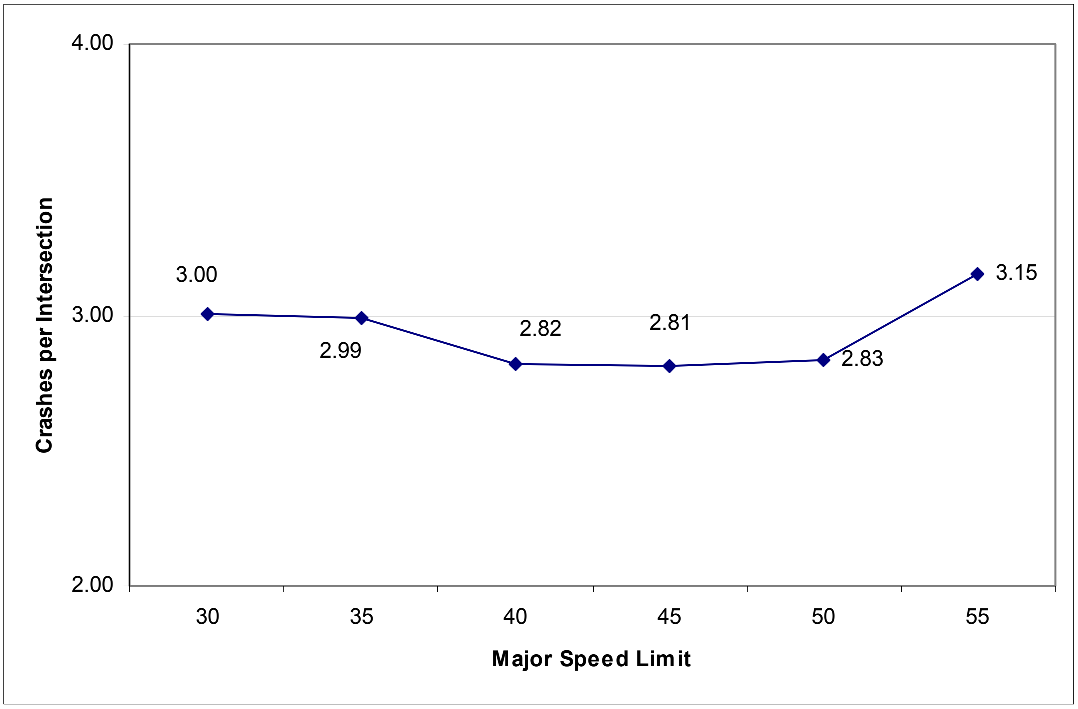

Figure 5.28 Average expected number of turning crashes per intersection per year for different values of speed limits on the major road

8. *Major AADT:* The turning crashes show a very slight increasing trend with the increase in AADT, as can be observed in Figure 5.29. Keller (2004) finds AADT to be one of the least significant factors in predicting the frequency of left turning

crashes. Hence it can be concluded that the AADT does not significantly affect the turning crashes.

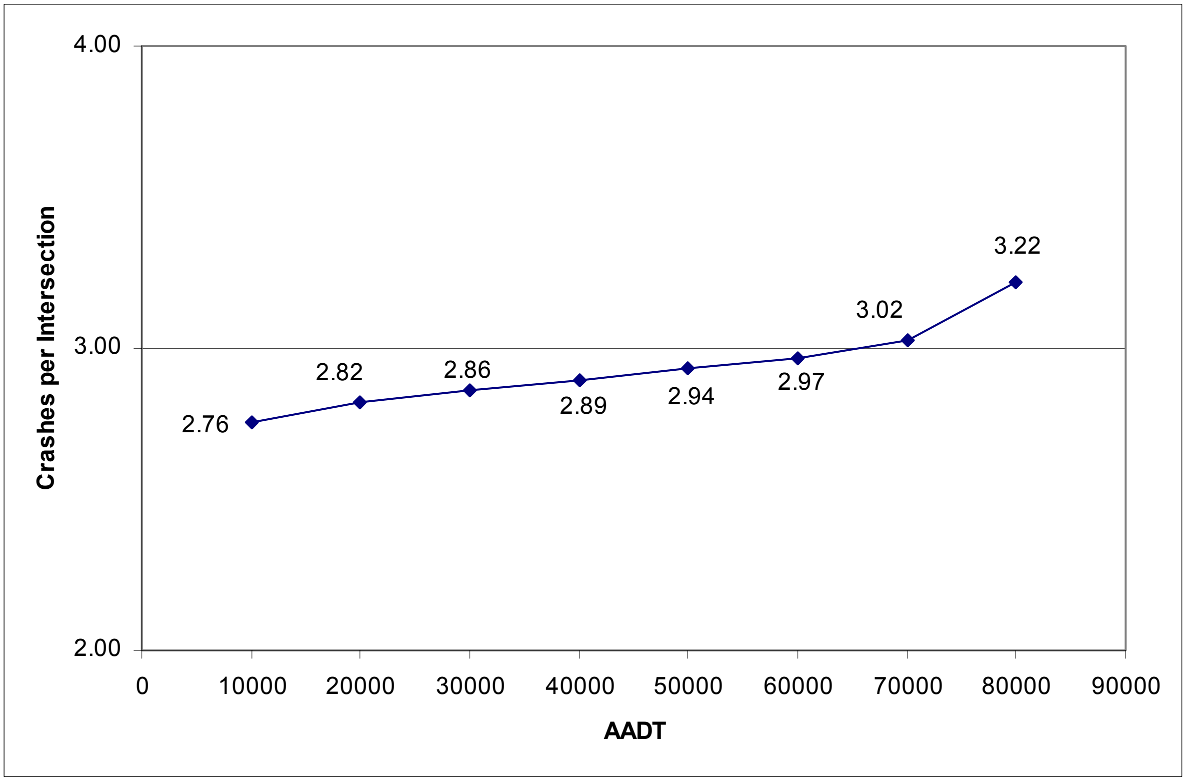

Figure 5.29 Average expected number of turning crashes per intersection per year for different values of AADT on the major road

The variables found to be significant in predicting the frequency of turning crashes have also been found to be significant in other studies. The error in predicting the turning crashes reduced significantly by using the new method. The RMSE in predicting the frequency of turning crashes for safe and unsafe intersections was 0.62 and 2.3 respectively. The error was significantly reduced by using this method, considering that the error in predicting the turning crashes for all intersections taken together was 3.35. Hence this method can be considered to be very efficient in predicting the turning crashes at signalized intersections.

5.8 Predicting the frequency of Sideswipe crashes based on the number of lanes

To predict the number of sideswipe crashes at signalized intersections, all intersections were taken and the 75% of the intersections were randomly selected and trained by the GRNN model. The rest of the intersections were tested using this neural network, and the RMSE and MAPE values obtained were 2.62 and 57.6% respectively. To develop a better model that can predict the sideswipe crashes more efficiently, the method described in section 5.4 was used. In order to use the method, it was first checked if the sideswipe crashes increase with an increase in the total lanes. From Table 5.16, it is clear that the sideswipe crashes do show an increasing trend. Therefore this method was used to predict the sideswipe crashes and also to check if the error can be reduced.

Table 5-16 Average number of angle crashes for different groups of intersections

Total Lanes at an Intersection	Average Number of Sideswipe Crashes
3 to 5	1.77
6 to 10	1.10
11 to 15	2.08
16 and above	3.08

5.8.1 Classification of Intersections

The intersections were classified into safe and unsafe types by using the values in Table 5.16. In the database, 375 intersections were categorized as safe intersections for sideswipe crashes. The classification was carried out using both the MLP and PNN models. The MLP neural network gave a best accuracy of 68.4%, whereas the PNN model demonstrated a better accuracy of 70.7%. Hence the PNN model was chosen to identify the significant variables. The following variables were identified to be significant using this model: major LTP lanes, AADT, major speed limit, minor RTC lanes and major lanes. The accuracy of the model increased to 71.6% upon using only the significant variables.

5.8.2 Predicting the frequency of sideswipe crashes for safe intersections

Both the MLP and GRNN models performed equally well in predicting the frequency of sideswipe crashes at signalized intersections. The RMSE values for both the models were around 0.46, and the MAPE values were around 71.6%. The GRNN model was chosen to identify the significant variables as GRNN is faster in training and testing data compared to the MLP neural network model. The significant variables identified in the model were as follows: major lanes, minor lanes, major LTP lanes, speed limit and AADT.

5.8.3 Predicting the frequency of sideswipe crashes for unsafe intersections

Compared to the MLP model, the GRNN model performed much better in predicting the sideswipe crashes. The RMSE of the GRNN model was 2.4 whereas for the MLP model was 2.96. The MAPE for both the models was around 49%. Hence the GRNN model can be considered as a better model for predicting the sideswipe crashes at unsafe intersections. The significant variables identified using the GRNN model were as follows: minor lanes, major RTC lanes and major LTP lanes. The RMSE of the model drastically reduced to 1.61 upon using these significant variables.

5.8.4 Estimating a Pattern in Significant Variables

To identify the relationship between the input variables and the frequency of sideswipe crashes, the method of testing the simulation data with the models developed in the previous sections was used. The PNN model was used to classify the intersections into safe and unsafe types, and the GRNN models were used to determine the frequency of sideswipe crashes at these intersections. The following points establish the relationship between the input variables and the sideswipe crashes:

1. *Major Road:* The number of through lanes on the major road was found to be a significant factor for classifying the intersections as well as for predicting the frequency of sideswipe crashes for safe intersections. From the output obtained for the simulation database, which is shown in Figure 5.30, it was found that their increase leads to an increase in the sideswipe crashes. This result is reasonable, because the increase in the through lanes implies that more lane changing maneuvers occur that increase the chances of a sideswipe crash. This factor has been found to be significant by Keller (2004).

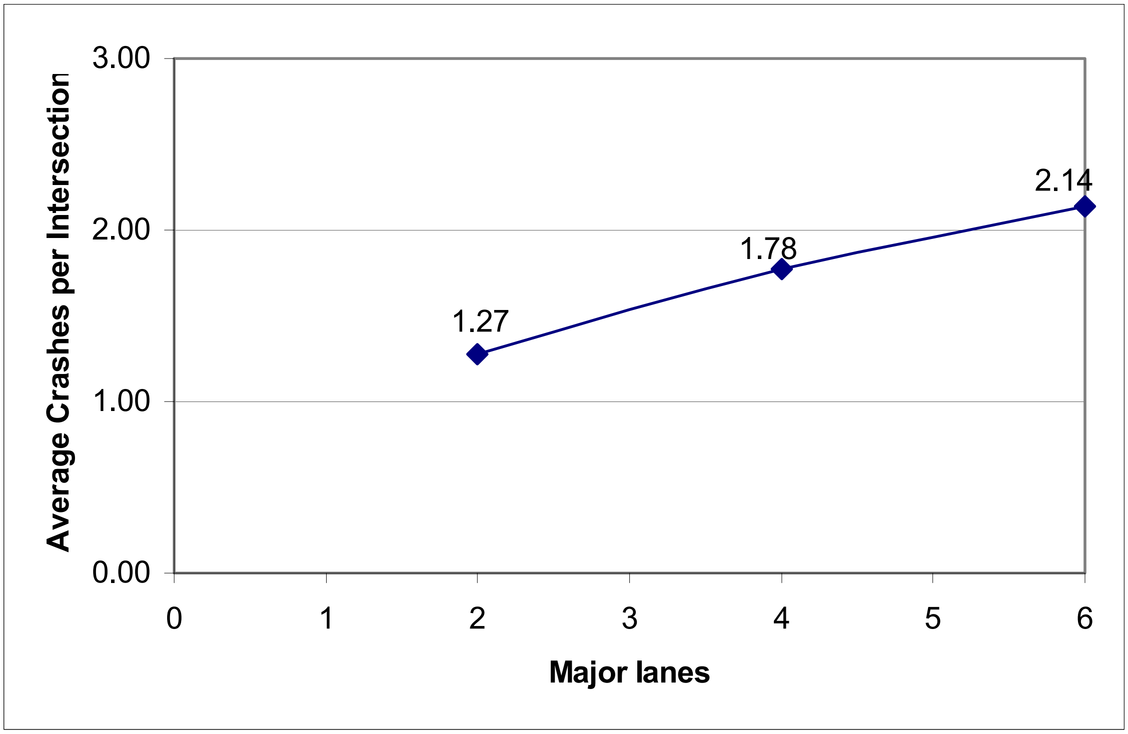

Figure 5.30 Average expected number of sideswipe crashes per intersection per year for different number of lanes on the major road

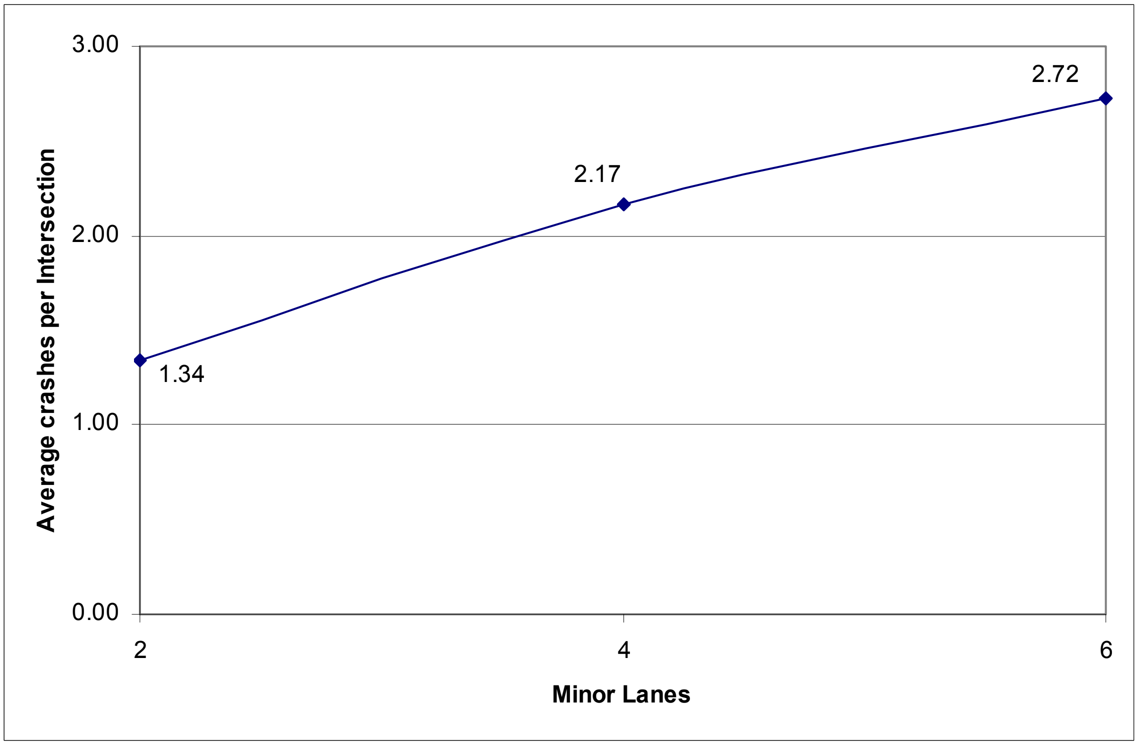

Figure 5.31 Average expected number of sideswipe crashes per intersection per year for different number of lanes on the minor road

2. *Minor Road:* Based on the reasoning given for the increase in sideswipe crashes with an increase in the major lanes, the sideswipe crashes can also be expected to increase when the number of lanes on the minor road increases. This has been observed in the simulation output, and has been shown in Figure 5.31.
3. *Total Left Turning Lanes:* Since this variable was not found to be significant in any of the models, it does not influence the frequency of sideswipe crashes.
4. *Major LTP Lanes:* The LTP lanes on the major road tend to increase the frequency of sideswipe crashes, as can be seen in Figure 5.32. One possible reason for this increase is that many vehicles taking a left turn have a sideswipe crash with a vehicle taking a right turn onto that road, and an increase in protected left turns means that vehicles in the rightmost left lane are very susceptive to sideswipe crashes. Another possible reason is that many drivers in the left turning

bay decide to change-over to the through lane, and become a part of a sideswipe crash in this maneuver. Thus the result is reasonable, and has also been found to be significant by Keller (2004).

5. *Minor LTP Lanes:* Since this variable was not found to be significant in any of the models, it does not influence the frequency of sideswipe crashes.
6. *Major RTC Lanes:* During a right turning maneuver, a vehicle is subjected to a sideswipe crash either from a vehicle going through or taking a left turn onto the roadway in which the right turning vehicle is heading to. Thus the presence of channelized right turning lanes can be expected to increase the sideswipe crashes. This has been demonstrated in Figure 5.33.

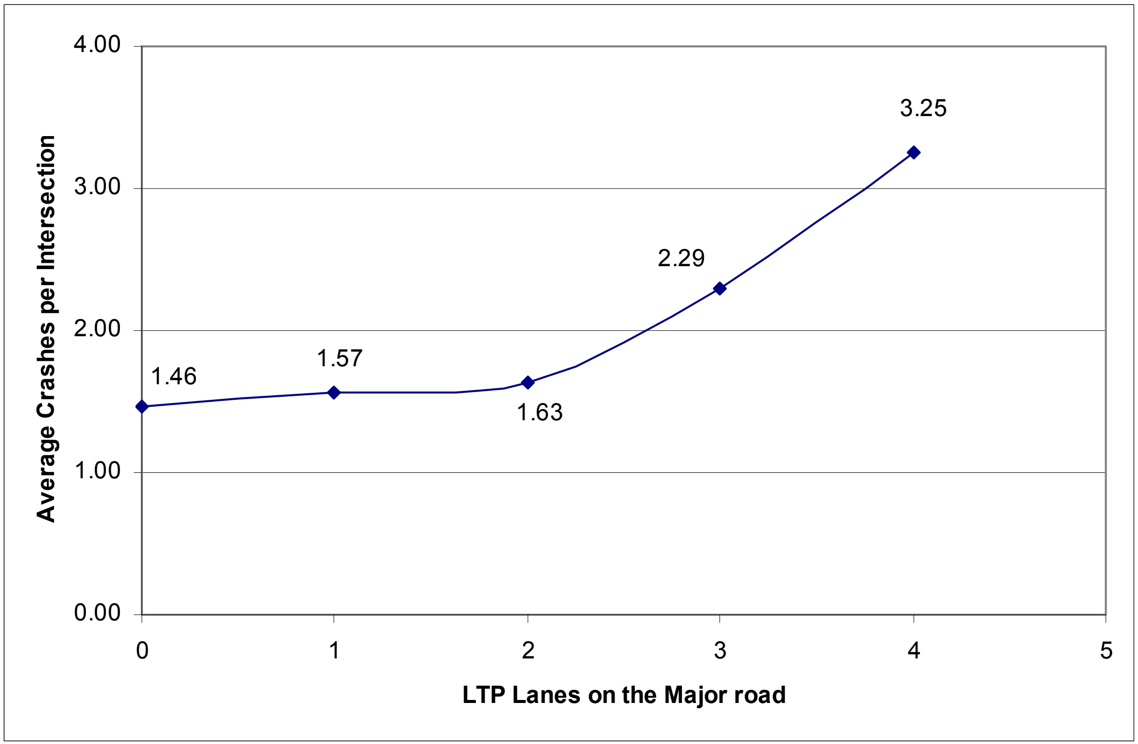

Figure 5.32 Average expected number of sideswipe crashes per intersection per year for different number of LTP lanes on the major road

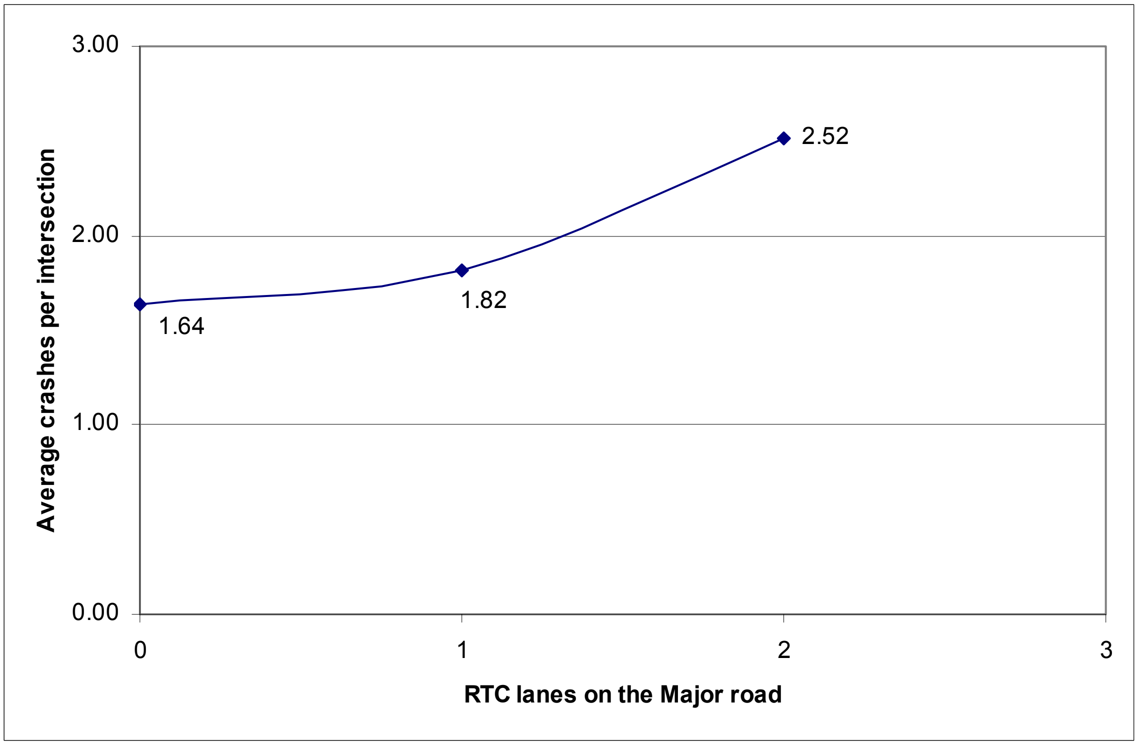

Figure 5.33 Average expected number of sideswipe crashes per intersection per year for different number of RTC lanes on the major road

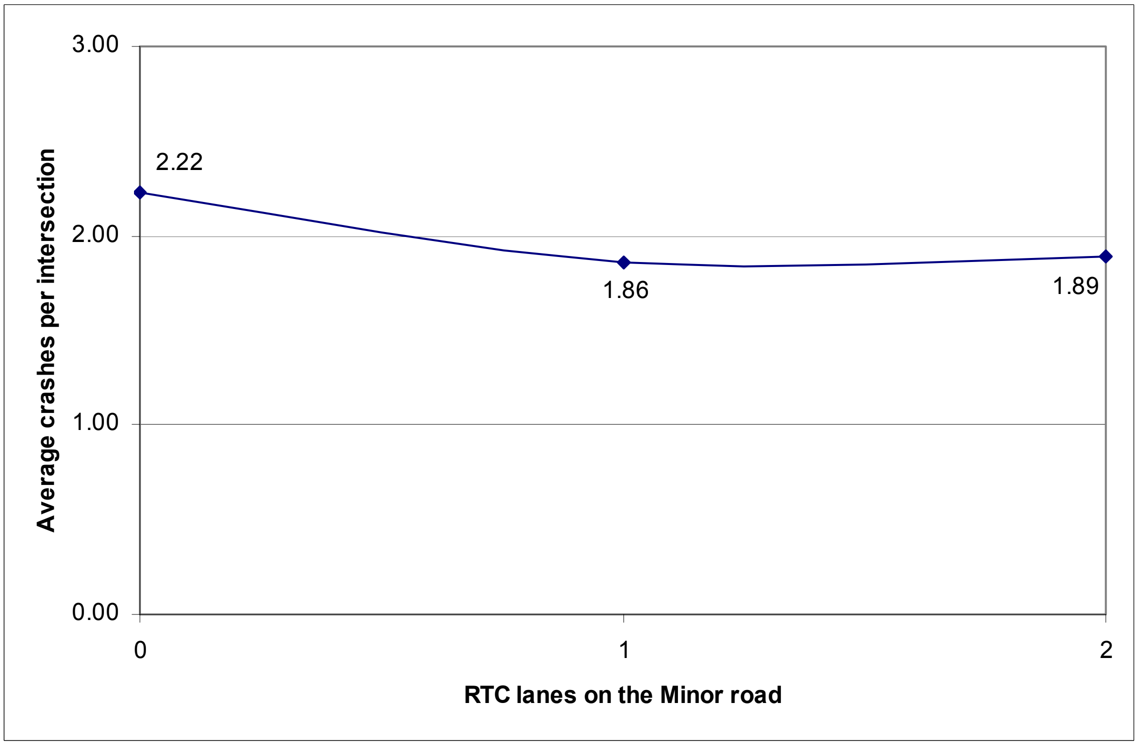

Figure 5.34 Average expected number of sideswipe crashes per intersection per year for different number of RTC lanes on the minor road

7. *Minor RTC Lanes:* This variable has been found significant only in classifying intersections into safe and unsafe types. From Figure 5.34, it can be clearly seen that it does not have much impact on the prediction of sideswipe crashes.
8. *Speed Limit:* Speed limit shows an uncommon trend in the Figure 5.35. Although the graph seems to be wavering, the difference in values is small. Since the speed limit was found to slightly significant in predicting crashes at safe intersections, referring to Figure 5.35 it can be considered not to affect the sideswipe crash frequency by much.
9. *Major AADT:* As the traffic volume increases, the spacing between vehicles decreases, increasing the chances of a sideswipe crash. Thus more sideswipe crashes can be expected at higher traffic volumes, as can be seen in Figure 5.36. AADT has also been found significant by Keller (2004).

Thus the model for sideswipe crashes gave reasonable results. The accuracies of the models were better compared to the model that used all the intersections in the database for predicting the frequency of crashes. Therefore, the models can efficiently predict the frequency of sideswipe crashes at signalized intersections.

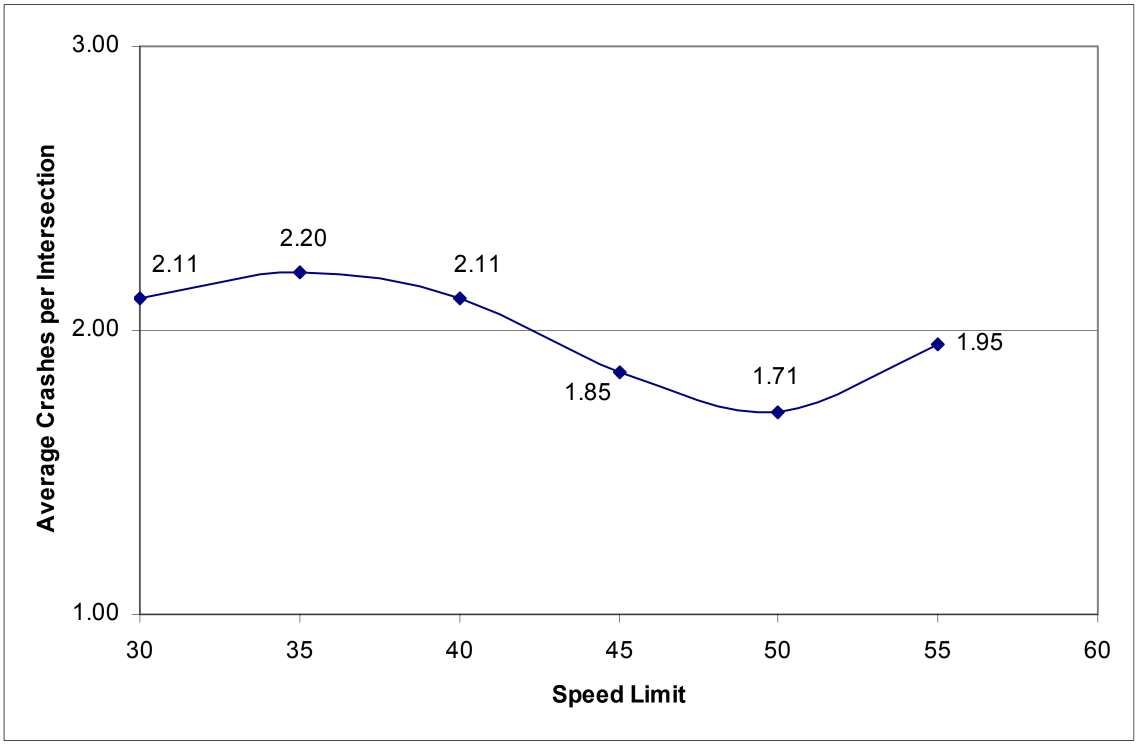

Figure 5.35 Average expected number of sideswipe crashes per intersection per year for different values of speed limit

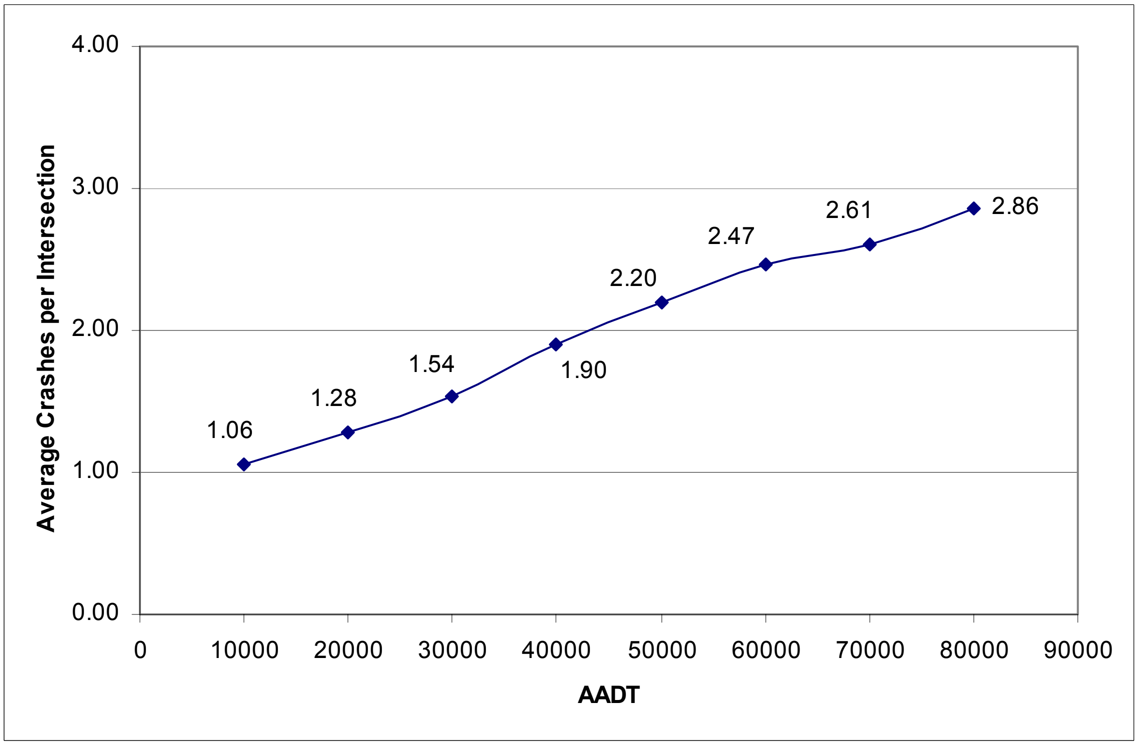

Figure 5.36 Average expected number of sideswipe crashes per intersection per year for different values of AADT

5.9 Summary

This chapter illustrates the method of predicting the frequency of various types of crashes based on simple intersection characteristics that include the lane configuration, speed limit and traffic volume. At first, all intersections from the long form crash database were utilized to develop models for determining their crash frequencies. Two models: GRNN and MLP neural network, were used and their outputs were compared. But the error obtained in these models was large and unsatisfactory. Thus a new method was developed that demonstrated an efficient way of determining the crash frequencies. Firstly, it was established that an increase in the total number of lanes at an intersection leads to an increase in crashes. Secondly, the average number of crashes per intersection was determined for intersections with different number of total lanes. If the number of crashes at an intersection was lesser than the average number of crashes at intersections with the same number of total lanes, the intersection was classified as a "safe" intersection. If the value was greater, it was classified as an "unsafe" intersection. Then, models were developed to classify the intersections in into safe and unsafe categories using PNN and MLP neural networks. The intersections in the safe category were separated and models were developed to predict the frequency of crashes using MLP and GRNN methods. Similar models were developed to predict the frequency of crashes for unsafe intersections. For each of the models, the best neural network model was identified and the significant variables were identified using the Forward Sequential Selection method. A simulation database was then built containing 98928 intersections with all possible combinations of the input variables. The frequency of crashes was predicted for these intersections using the above models. Lastly, the output of the model was plotted with the input variables to establish relationships between the input and output variables. This method was followed to predict the frequency of crashes of each collision type.

Table 5.17 illustrates the different neural network models that were used and the accuracies of each model. The neural networks that were either most accurate or were more suitable for the models have been shown. The PNN model always performed better in the classification phase compared to the MLP neural network model. Hence it can be safely concluded that PNN is more efficient in classifying when compared to the MLP neural network. In the prediction phase, the MLP and GRNN have both performed well. In some cases the performance of both the models was the same. But GRNN model was preferred in such a case because of it's capability of training large amounts of data in a short time.

Table 5-17 Significant neural network models and their accuracies in predicting different types of crashes

	Total Crashes	**Rear End**	**Angle**	**Turn**	**Sideswipe**
Classification *Model/Accuracy*	PNN / 67.6%	PNN / 68.22%	PNN / 68.24%	PNN / 64.43%	PNN / 71.6%
Predicting Crashes: Safe Int *Model/RMSE*	MLP / 2.62	GRNN / 1.37	MLP / 0.78	GRNN / 0.61	GRNN / 0.45
Predicting Crashes: Unsafe Int *Model/RMSE*	MLP / 5.35	MLP / 4.3	GRNN / 2.65	GRNN / 2.3	GRNN / 1.61

Table 5.18 illustrates the results of testing the models on the simulation database. The cells show the type of pattern each input variable shows for predicting the frequency

of different types of crashes. “Increase” means that crashes increase with an increase in the input variable. “-“ means that the variable was not found to be significant in predicting the frequency of crashes. Most of the variations obtained are reasonable.

Table 5-18 Relationship between the input variables and the frequency of various types of crashes

	Total Crashes	Rear End	Angle	Turn	Sideswipe
MJ Lanes	Increase	Increase	Increase	Increase	Increase
MN Lanes	Increase	Increase	Increase	Increase	Increase
Total LTL	Increase	-	-	-	-
MJ LTP	Decrease	Increase	Decrease	Decrease	Increase
MN LTP	Increase	Increase	Increase	Increase	-
MJ RTC	-	-	-	Increase	Increase
MN RTC	Decrease	Increase	-	-	-
MJ Speed	-	Increase	-	-	-
MJ AADT	Increase	Increase	Increase	Increase	Increase

Therefore, this chapter establishes a strong method in accurately predicting the frequency of crashes using simple traffic and geometric characteristics of signalized intersections. This method can prove very useful in crash prediction at various stages of an intersection. For example, before the construction of an intersection, this method can be used to predict the expected frequency of crashes using the variables that have been proposed for the intersection. If the model suggests that the crash frequency for this intersection is high, the input variables can be altered to design a safer intersection.

Similarly, the intersection characteristics of an operational signalized intersection can be used as an input to the model, and these input variables can be changed so as to obtain optimum characteristics of the intersection that can make it a safer place to travel on.

6 CLASSIFICATION OF CRASHES USING NEURAL NETWORK TREES

6.1 Introduction

An analysis was conducted to estimate the collision type of a crash based on the intersection properties, traffic characteristics and conditions prevalent at the time of the crash. Given any of these characteristics and given the criterion that a crash will occur, the models formed in the analysis would predict the type of collision the crash will be subjected to. This will be helpful in studying the factors that lead to a particular type of crashes. For example, the model will be able to specify the intersection properties that can lead to increased rear end crashes, and therefore it will help us design countermeasures directed to reducing this particular crash type if the intersection experiences above normal rates of this type.

6.2 Database Used

To analyze the data for predicting the collision type, the database was first prepared for analysis. The database with crashes reported on long forms was chosen for the analysis because all the counties contained information on long form crashes, but the crashes reported as short forms were not present in the database of two counties (Orange and Miami-Dade). The following variables were selected and encoded:

1) Light Conditions:

1. Daylight
2. Dusk
3. Dawn
4. Dark with street lights
5. Dark without street lights
6. Others/unknown

2) Surface Conditions

1. Dry
2. Wet/Slippery
3. Others/unknown

3) Month

4) Day of the week

5) Time of the day:

1. 12:00 am – 6 am
2. 6am –9 am
3. 9am – 11 am
4. 11 am – 1 pm
5. 1 pm – 3 pm
6. 3 pm – 6 pm
7. 6 pm – 12pm

6) Number of Major Lanes

7) Number of Minor Lanes

8) Total Left Turning Lanes

9) Total Left Turning Protected Lanes on the Major Road

10) Total Left Turning Protected Lanes on the Minor Road

11) Total Right Turn Channelized Lanes on the Major Road

12) Total Right Turn Channelized Lanes on the Minor Road

13) Speed Limit on the Major Road

14) Speed Limit on the Minor Road

15) AADT on the Major Road

16) AADT on the Minor Road

17) Collision Type:

1. Rear End
2. Angle
3. Turn crashes
4. Sideswipe
5. Pedestrian crashes
6. Head On
7. Other crashes/unknown

This complete database consisted of 27,044 crashes that had occurred in the years 2000 and 2001 for all counties, except for Orange County for which the crash data was available only for the years 1999 and 2000. Not all the crashes contained the information on the speed limit and AADT on the minor roadways. Of these crashes, the pedestrian and bike crashes were only 2% in number and 1.3% were Head-on crashes. Since the percentage of the crashes was too low, most of the crashes would not be predicted correctly. Hence these crashes were deleted from the analysis database. Although the percentage of crashes whose collision type was unknown were 9% of the total crashes, they were not used in the database as no specific properties of the intersection/crash conditions can be underlined for the occurrence of such crashes.

Now the analysis database contained 23216 crashes. Since some of the crashes were missing the speed and traffic volume data for the minor roadway, the complete database was considered for the analysis by initially not considering these variables in the analysis.

The training and test databases were developed by using the data for the year 2000 for training, and 2001 for testing, except in the case of Orange County for which the data for the year 1999 was used in testing as the data for 2001 was unavailable. As the frequency of right turning crashes was very low, they were combined with the left turning crashes to form a category of "turning crashes". Therefore, the test database contains 11726 crashes (50.6% rear end crashes, 22.56% angle crashes, 16.5% turn crashes and 10.3% sideswipe crashes) and the training database contains 11490 crashes (50.36% rear end, 21.93% angle, 17.2% turn and 10.6% sideswipe crashes).

The Multi-Layer Perceptron (MLP) and Probablistic Neural Networks (PNN) have been used in this study. The MLP neural network has been used several times in traffic safety by Abdelwahab and Abdel-Aty (2001; 2004), Mussone et al. (1999) and Sayed and Abdelwahab (1998), and these studies have found MLP to be one of the best neural network models. Not many studies in traffic safety have used the PNN models.

The neural networks were developed using the Neural Network Toolbox in the MATLAB software. The procedure of developing the models is as follows:

1. The training and test databases were developed as explained earlier.
2. These datasets were normalized to unit standard deviation, as was carried out in the previous chapter.
3. The number of input and output nodes was decided. The number of input nodes is equal to the number of input variables being considered in the model. The number of output nodes depends on the number of categories of the output. If the output has four categories, the number of output nodes can be chosen to be four.
4. The training database was randomized so that there is a randomized presentation of inputs.
5. A neural network is built based on the input and output values of the training database. This network depends on the number of hidden nodes and the number of epochs in case of the MLP and on the value of spread in case of PNN.

6. The test database is tested using the network developed to get the predicted output.
7. This predicted output is compared to the actual output to check how well the network performs. From the comparison, the accuracy of prediction for different categories of outputs is calculated.
8. This process is repeated for different values of hidden nodes for MLP and spread for PNN.
9. The model that gives the highest accuracy is declared as the better model.

6.3 Predicting Collision Type ignoring the AADT and speed for the Minor road

6.3.1 Multi Layer Perceptron (MLP) Neural Network

The MLP neural network consisted of 14 input nodes and 4 output nodes for predicting the four types of collisions in the form of (1, 0, 0, 0) for rear end, (0,1,0,0) for angle, (0,0,1,0) for turn and (0,0,0,1) for sideswipe crashes.

At first, only one hidden layer was used in the analysis. The number of neurons in this hidden layer was increased from 5 to 60 with increments of 5 neurons. The number of neurons with which the accuracy was highest was chosen as the optimal number of neurons to be used in the model. The maximum number of epochs was set to 1000. The activation function for the hidden layer was tan sigmoid, and for the output layer was pure linear. The Resilient backpropagation (rprop) algorithm was used for the MLP model. The model was trained with the training data and then tested on a testing data to find the percentage accuracy for the complete database as well as for each individual type of collision types.

Tables 6.1 and 6.2 illustrate the results of training and testing phases of the MLP NN. The numbers in the table represent the correct percentage of the particular collision type predicted. For example, the value of 6.0824 in Table 6.1 for the Angle Accuracy represents that 6.08% (194 crashes out of 3202 angle crashes) accuracy in predicting angle crashes. In the tables, RE indicates a rear end crash and SS indicates a sideswipe crash.

As can be clearly seen in the tables, the accuracy of the model was low. Attempts to increase the number of hidden neurons and adding a hidden layer could not make the results any better. All the crashes were being classified as Rear End crashes as they formed the majority in the database.

Table 6-1 Training Accuracy of MLP for predicting Collision Types

Number of Nodes	Overall Accuracy	RE accuracy	Angle accuracy	Turn Accuracy	SS accuracy
5	50.3011	97.2246	6.0824	0	0
10	50.4875	98.5198	3.6625	0.20894	0.27155
15	50.4946	97.5519	5.1014	1.0029	0.67889
20	50.4946	97.922	4.3492	1.2119	0.13578
25	50.5018	96.9826	6.6056	1.2119	0
30	50.3513	96.0717	7.881	1.2119	0.27155
35	50.0215	93.7518	10.3663	2.9252	0.27155
40	50.3082	95.3886	8.0772	2.2984	0.95044
45	50.0717	93.6664	8.6658	5.0982	1.1541
50	48.7957	89.1261	11.2819	7.0205	2.1724
55	49.2975	90.0085	11.6743	7.0623	1.833
60	49.7276	91.1756	11.9032	5.8504	1.833

Table 6-2 Testing Accuracy of MLP for predicting Collision Types

Number of Nodes	Overall Testing Accuracy	RE accuracy	AN accuracy	Turn Accuracy	SS accuracy
5	50.8771	97.3717	6.8729	0	0
10	51.0743	98.5816	4.9047	0.12804	0.068493
15	50.701	97.4273	5.3733	0.72557	0.13699
20	51.0391	97.8445	5.4983	1.4085	0
25	50.9123	96.8989	7.3102	1.0243	0.068493
30	50.9264	96.8572	7.5914	0.89629	0
35	50.1021	93.6031	10.0594	2.4755	0.068493
40	50.7503	95.4109	8.8722	2.2194	0.47945
45	50.074	93.5475	8.6223	3.7132	1.2329
50	48.2212	88.0962	10.9028	5.4204	2.3288
55	49.0102	89.6815	11.184	5.8045	0.9589
60	49.4893	90.9053	11.8713	3.9693	1.0274

To avoid this situation, the database was balanced by first selecting all the crash data for the collision type that has a minimum number in the database and then randomly selecting an equal number of crash data for the other collision types. Thus all the categories will have an equal representation in the training method. This method was expected to increase the prediction accuracy for the categories whose representation was very small. Since the data was chosen in a random fashion, some of the input values would be excluded from the database. By repeating the process for a few times, different input values can be used in each cycle. The average of the results would indicate the average result of using this process. Thus this process was repeated five times and an average was taken for the accuracies to find the average accuracy of the prediction. Although this process increased the accuracy of the collision types other than rear-end, the results were not encouraging.

6.3.2 Probabilistic Neural Network (PNN)

PNN was developed similar to the MLP neural network. The spread was varied from 0.1 to 2.0 with increments of 0.05. The PNN develops the model in such a manner that the output of training dataset is predicted accurately. Hence the training accuracy was not checked for the PNN model. Only the test data was checked for its accuracy. The results have been indicated in Table 6.3.

Although PNN was able to produce better results for the angle, turn and sideswipe crashes, the prediction accuracy of the rear end crashes and the overall model decreased. Similar to the MLP, the PNN failed to produce any significant results to predict the collision type using the available variables. The spread was varied on a larger scale, but no fruitful result availed.

Table 6-3 Predicting Collision Type using a PNN

SPREAD	OVERALL ACCURACY	RE ACCURACY	ANGLE ACCURACY	TURN ACCURACY	SIDESWIPE ACCURACY
0.05	39.4196	58.6567	22.4617	19.8805	13.2192
0.1	38.4334	54.7768	25.2109	21.3737	14.3151
0.2	38.6095	55.0549	25.0547	21.843	14.2466
0.3	39.2012	56.5568	24.8985	21.2031	13.9726
0.4	40.4128	59.5606	24.1799	20.5631	13.5616
0.5	41.7019	63.8993	22.6179	18.6007	11.3014
0.6	43.5193	69.5453	20.9622	16.2116	8.6301
0.7	45.7382	75.97	19.1503	13.5239	6.8493
0.8	47.1893	81.2961	16.4324	10.7509	5.137
0.9	48.4996	86.2328	13.527	8.4044	3.6986
1	49.4435	90.3769	10.7779	5.5887	3.0137
1.1	50.1761	93.9091	8.0912	3.413	2.1233
1.2	50.4931	96.287	5.4983	1.9625	1.5068
1.3	50.6622	97.8584	3.5301	1.0666	1.1644
1.4	50.7185	98.7206	2.343	0.72526	0.61644
1.5	50.8171	99.3047	1.687	0.63993	0.27397
1.6	50.8312	99.7497	1.0622	0.29863	0.13699
1.7	50.7256	99.9166	0.37488	0.17065	0
1.8	50.6692	100	0.06248	0	0
1.9	50.6622	100	0.03124	0	0
2	50.6551	100	0	0	0

6.4 Using data for the Minor Roadway

The next step in the analysis was to use the data based on the minor roadway to predict the collision type. It was expected that adding the speed and traffic data for the minor roadway would produce better results.

As the data for the flow and speeds of the minor roadway was limited, the database shrunk to 9801 crashes, which was almost one-thirds of the original database. Because of this drastic decrease, the format of some variables had to be changed. The light conditions were originally classified into 6 categories. But in the new database, most of the crashes had occurred either in daylight or dark lighting conditions. Hence the classes in light conditions were brought down to two. All categories except for daylight conditions were combined into the second category, as the lighting conditions will be almost be dark during the other cases. Now there were 7202 crashes in daylight conditions and 2599 crashes in the dark conditions.

In the new database, rear end crashes formed 54.5%, angle crashes formed 19.4%, and there were 13.6% of turn crashes and 12.6% of sideswipe crashes. This is only slightly different from the original database. A Chi-squared test was performed to prove that this data is not different from the full dataset.

MLP neural network was developed using the same method used in the previous case, except that there were 2 extra input nodes of the speed limit and AADT on the minor roadway. The same algorithm was used, but the results hardly improved. The same was tried with the PNN, but there was no significant improvement in the results.

6.5 Neural Network Tree for Predicting Collision Type

Since the MLP and PNN performed below expectation to predict the collision type of the crashes, they cannot be used to satisfy our objectives. The neural networks were not able to perform well with four output types. Therefore a new strategy had to be used that could deal with this problem and also make the model significantly better. We developed a new idea to use a Neural Network Tree.

The concept used in developing the Neural Network Tree is similar to the modeling used for developing a Nested Logit structure. In such a Tree shaped Neural Network, the classes of collision type could be wisely combined together to obtain two classes instead of four. It was perceived that more often than not, rear-end and sideswipe crashes occur along the same direction. Hence they usually have the same characteristics. On the other hand, angle and turn crashes usually occur because of the interference of traffic from one direction with the other. Therefore, they have a similar pattern. This resulted in a method in which the rear-end and sideswipe crashes together were combined into one category and angle and turn crashes into another category. Thus a neural network model was first developed to classify a crash into these two categories based on the 16 variables identified in the earlier sections. This classification would form the first branch of the neural network tree. The next branch would classify rear-end and sideswipe crashes and the third branch would classify the angle and turn crashes. The Neural network tree is depicted in Figure 6.1. Then the models could be used to identify the significant variables and identify their effect on the crashes.

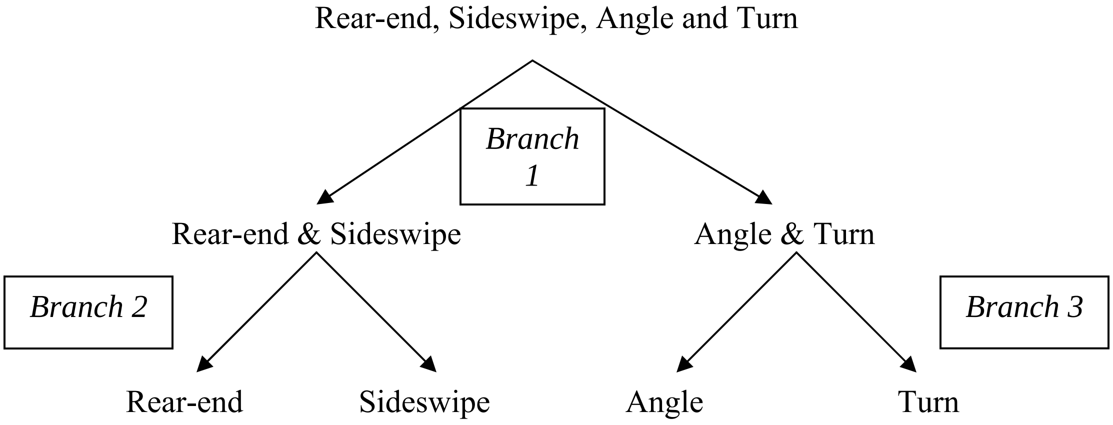

Figure 6.1 Proposed structure of the Neural Network Tree

Overall, the tree structure will be constructed in the following pattern:

1. In the database used for prediction of the collision types, the rear end and sideswipe crashes will be combined to form category 1 and angle and turning crashes will be combined to form category 2.
2. MLP and PNN models will be used to classify the two categories.
3. The model with higher classification accuracy will be identified.
4. Significant variables will be identified for the models.
5. This model will be used on a test database to check how the variation of input affects the output.
6. The previous steps will be repeated to develop the other two branches of the neural network tree.
7. The Neural network tree will be formed with a neural network model at each node.

The following sections discuss developing the neural network tree by taking one neural network model at a time.

6.6 Distinguishing Rear End and Sideswipe crashes from Angle and Turn crashes

6.6.1 MLP Neural Network

The database used in the previous neural network model was used. It contained 9801 crashes for the years 2000 and 2001. The rear end and sideswipe crashes formed category 1 and angle and turn crashes formed category 2. The 16 input variables were normalized as discussed earlier. As the database contained 67% of category 1 crashes, the MLP program was developed so that it would randomly extract category 1 crashes from the training database to make them equal to the number of category 2 crashes. Hence the proportion of category 1 and category 2 crashes became equal in the training database. The model was predicted using this data and this method was repeated five times so that different proportions of random category 1 crashes could be chosen in each run. Then the average of the 5 runs was taken to find the actual output of the model.

As a first step, a complete model was developed to predict the collision type categories. The model consisted of 16 input nodes, hidden nodes varying from 5-60 with increments of 5, and one output node (indicating 0 or 1). The Resilient backpropagation (rprop) algorithm was used in the study. The maximum number of epochs used were 1000. The results of the algorithm have been summarized in Table 6.4.

As can be observed from Table 6.4, the model performed satisfactorily. The best accuracy obtained in the testing phase is for 57.81% for 5 hidden nodes, whereas for the training phase is 66.38% for 40 hidden nodes. Since the test data is common for all models, the best model was selected based on the test results. Hence the best model was the MLP with 5 nodes in the hidden layer. The increase in the number of hidden nodes does not have a significant effect on the test accuracy.

Table 6-4 Summary of results of the Testing phase of MLP model for classifying crashes into rear-end and sideswipe crashes (cat 1) or into angle and turn (cat 2) crashes

# Hidden Nodes	Train Accuracy	Cat 1 Accuracy	Cat 2 Accuracy	Test Accuracy	Cat 1 Accuracy	Cat 2 Accuracy
5	63.04	64.79	61.30	57.81	60.59	55.02
10	62.79	64.42	61.16	57.10	61.12	53.07
15	64.36	65.24	63.48	57.16	60.53	53.79
20	64.19	65.22	63.16	56.93	59.07	54.79
25	65.19	65.63	64.75	57.55	59.57	55.52
30	65.73	67.05	64.40	57.55	60.77	54.33
35	64.79	66.44	63.14	57.24	59.97	54.52
40	66.38	67.81	64.95	57.31	59.95	54.66
45	65.08	66.71	63.46	56.72	58.91	54.54
50	65.99	66.81	65.18	57.18	58.98	55.37
55	65.97	67.10	64.85	56.77	59.32	54.22
60	65.77	67.14	64.40	56.29	58.65	53.93

6.6.2 PNN

A model was developed using PNN to distinguish rear end and sideswipe crashes from angle and turn crashes. This model consisted of the same data used in the MLP model. As discussed in the previous PNN models, the spread was varied from 0.05 to 2 with increments of 0.1. The best accuracy obtained, as can be seen in Table 6.5, is 57.75%, which is almost the same as the MLP model. Hence the MLP and PNN models gave the same accuracies for the first branch. But the runtime of the PNN model was far more than the MLP model. Hence the MLP model was chosen to find the significant variables.

Table 6-5 Summary of results of the Testing phase of PNN model for classifying crashes into rear-end and sideswipe crashes (cat 1) or into angle and turn (cat 2) crashes

Spread	Test Accuracy	Cat 1 Accuracy	Cat 2 Accuracy
1.05	57.75	59.98	55.52
1.15	57.42	61.39	53.45
1.35	57.13	63.19	51.07
0.95	57.11	58.36	55.87
1.45	57.07	64.30	49.85
1.55	57.00	65.25	48.75
1.25	56.99	61.93	52.04
1.65	56.92	66.31	47.52
1.75	56.69	66.98	46.41
0.85	56.54	56.71	56.37
1.85	56.47	67.80	45.14
1.95	56.46	68.48	44.45
0.75	56.16	56.12	56.20
0.65	55.47	55.27	55.67
0.55	55.41	55.33	55.49
0.45	54.67	53.76	55.59
0.35	54.45	53.50	55.40
0.25	54.23	53.47	54.99
0.15	54.17	53.44	54.90
0.05	52.17	61.47	42.88

6.7 Distinguishing Rear End crashes from Sideswipe crashes

From the database used in the first branch of the neural network tree, the rear end and sideswipe crashes were filtered out. These crashes were separated into categories 1 and 2, based on whether the crashes were rear end or sideswipe. Thus the training database consisted of 3331 crashes and the test database contained 3243 crashes. As in the previous methods, the training database represented the crashes that had occurred in 2000 while the test database consisted of crashes in 2001. In both the databases, rear end crashes constituted to around 80% of the crashes. So the number of rear end cases were matched with the number of sideswipe cases, as was done in the previous method.

6.7.1 MLP Neural Network

Based on the 5 runs, 16 inputs and the number of hidden nodes, which varied from 5 to 60, the test results showed a highest accuracy to be 55.47%. The results can be seen in Table 6.6. Since the number of hidden nodes were 5, another program was written to check if better results could be obtained by varying the number of hidden nodes from 2 to 10. But the test accuracy was not significantly different from 55.47%. Hence the MLP neural network gives a classification accuracy of 55.47% to distinguish between rear end and sideswipe crashes.

Table 6-6 Summary of results of the Testing phase of MLP model for classifying crashes into rear-end or sideswipe crashes

# Hidden Nodes	Test Accuracy	RE Accuracy	SS Accuracy
5	55.47	55.46	55.49
20	55.01	53.84	56.18
15	54.76	53.19	56.32
10	54.70	53.99	55.42
45	54.64	53.99	55.28
55	54.39	54.10	54.68
35	54.28	52.55	56.01
25	54.26	52.83	55.69
40	53.98	53.35	54.62
60	53.95	53.88	54.02
50	53.92	53.81	54.03
30	53.76	52.56	54.97

6.7.2 PNN

This model consisted of the same data used in the MLP model. The model had the highest test accuracy of 57.97%, which is better than the MLP test accuracy. Hence the PNN model was found to perform better in classifying rear end and sideswipe crashes.

Table 6-7 Summary of results of the Testing phase of PNN model for classifying crashes into rear-end (cat 1) or sideswipe crashes

Spread	Test Accuracy	RE Accuracy	SS Accuracy
1.45	57.97	48.52	67.42
1.55	57.94	47.31	68.58
1.65	57.86	46.16	69.56
1.75	57.69	44.49	70.89
1.95	57.41	41.21	73.61
1.85	57.39	42.79	71.99
1.35	57.20	49.47	64.93
1.25	56.75	50.71	62.79
1.15	56.51	51.62	61.40
1.05	56.29	52.46	60.13
0.95	55.82	53.13	58.51
0.85	55.54	53.61	57.47
0.75	54.87	54.12	55.61
0.65	54.64	54.18	55.09
0.05	54.55	68.08	41.03
0.55	54.22	54.09	54.34
0.45	54.11	54.39	53.82
0.25	53.96	54.91	53.01
0.15	53.92	55.01	52.84
0.35	53.85	54.74	52.95

6.8 Distinguishing Angle crashes from Turn crashes

The angle and turn crashes were filtered out from the database used in the branch 1 of the neural network tree. These crashes were categorized into category 1 and 2 based on whether the crashes were angle or turn crashes. The training database consisted of 1633 crashes, out of which 932 were angle crashes. The test database consisted of 1594 crashes out of which 60% were angle crashes.

6.8.1 MLP Neural Network

The best accuracy for distinguishing between angle and turn crashes was developed was 55.05% for a MLP neural network model with 60 hidden nodes. The results are shown in Table 6.8.

Table 6-8 Summary of results of the Testing phase of MLP model for classifying crashes into angle or turn crashes

# Hidden Nodes	Test Accuracy	Angle Accuracy	Turn Accuracy
60	55.05	55.37	54.73
40	54.37	53.90	54.83
45	54.20	53.26	55.13
35	54.18	53.33	55.03
55	54.00	55.04	52.96
15	53.84	53.83	55.84
10	53.83	52.58	55.08
50	53.39	53.51	53.26
20	53.08	54.76	51.39
25	53.00	53.55	52.45
30	52.85	55.12	50.58
5	52.16	53.69	50.63

6.8.2 PNN

The PNN model was developed based on the same database with which the MLP neural network was developed. As can be seen in Table 6.9, the MLP model showed mildly higher results compared to the PNN model. Hence the MLP model was chosen as the best model for classifying between angle and turn crashes.

Table 6-9 Summary of results of the Testing phase of PNN model for classifying crashes into angle or turn crashes

Spread	Test Accuracy	Angle Accuracy	Turn Accuracy
1.15	54.60	55.44	53.77
1.25	54.37	55.22	53.52
1.55	54.34	55.72	52.96
1.35	54.30	55.19	53.41
1.45	54.25	55.90	52.60
1.05	54.12	55.44	52.81
1.65	54.09	55.22	52.96
1.85	54.08	55.51	52.66
0.95	54.02	55.90	52.15
1.95	54.02	55.44	52.60
1.75	53.93	55.04	52.81
0.85	53.61	55.72	51.49
0.75	53.54	56.19	50.89
0.45	53.42	56.11	50.73
0.55	53.42	56.15	50.68
0.65	53.09	55.76	50.43
0.35	52.95	55.58	50.33
0.05	52.62	71.41	33.84
0.25	52.56	55.40	49.72
0.15	52.37	55.22	49.52

6.9 Summary of the Neural Network Tree

The above analysis can be summarized in the Table 6.10. The table illustrates the results of both MLP and PNN models with all three branches and the models selected for further analysis.

Table 6-10 Summary of the model comparisons and the models selected for the neural network tree

	Model	Accuracy	Model selected
Rear end and Sideswipe vs Angle and Turn	MLP	57.81%	MLP
	PNN	57.75%	
Rear end vs Sideswipe	MLP	55.47%	PNN
	PNN	57.97%	
Angle vs Turn	MLP	55.05%	MLP
	PNN	54.60%	

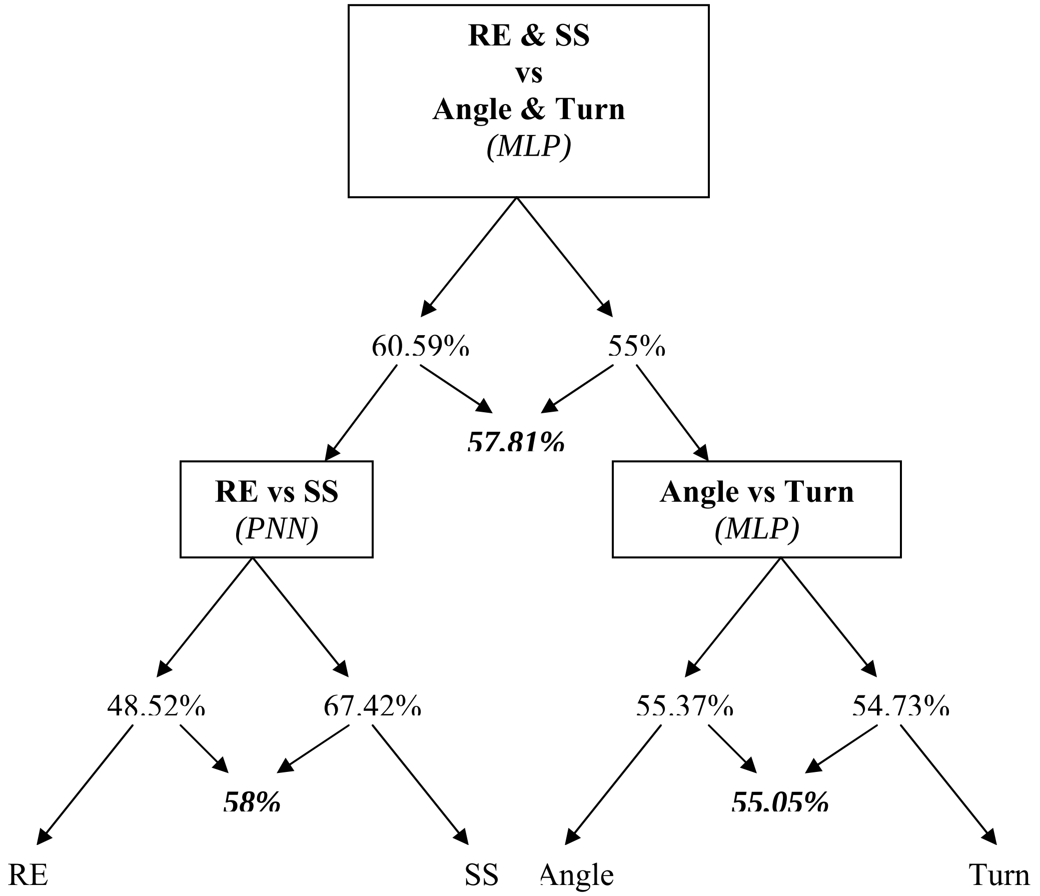

Figure 6.2 Diagram indicating the prediction accuracies of the Neural Network Tree

The tree can be represented in branches as shown in Figure 6.2. The collision type that the model is predicting has been written in the rectangular boxes. The type of model used has been listed in Italics. These boxes are branched off into the two categories that are being predicted. The accuracy with which the categories are predicted on the test data has been written between the model type and category being predicted. The overall model accuracy has been listed in Italics by joining the two individual collision type accuracies. For example, for distinguishing between angle end and turn crashes, the accuracy with which the angle crashes are predicted is 55.37% and the accuracy with which the turn crashes are predicted is 54.73%. The overall model accuracy is 55.05%.

6.10 Determining Significant Variables for the models

The models developed above can give a good prediction about the collision type of a crash. These models take into account the complete list of variables collected in the data collection phase. But an important phase of the analysis is to determine the variables that lead to an increased/decreased number of crashes with a particular collision type. For example, it is important to know which of the 16 input variables effects the angle crashes, and in what manner: whether an increase in these variables will lead to an increase or a decrease in angle crashes. This will illustrate the trends of the significant variables for each collision type.

The identification of significant variables was carried out by starting the modeling procedure by using just one variable to carry out the classification. Each of the 16 input variables were used individually to classify the crashes. The variable that gave the highest accuracy was selected as the most significant variable. Then this variable was used with the other 15 variables one at a time and the model accuracies were compared. The variable that gave the highest accuracy was chosen as the next most significant variable. The procedure was repeated until there was no significant change in the model accuracies when any of the remaining variables were added to the significant variables. These set of variables were chosen to be the significant variables.

The significant variables were found for the first branch of the neural network tree, i.e. to distinguish the rear end and sideswipe crashes from the angle and turn crashes. An MLP was written to test each variable individually with number of hidden nodes varying from 2-25 and the activation functions in the hidden and output layer being tangential sigmoid and pure linear functions respectively. The range of the number of hidden nodes was changed because the MLP requires lesser number of hidden nodes for lesser input variables. The hidden neurons were increased to 35 when the number of input variables were more than 5. When the first significant variable was found, the second run of the program was conducted to use this variable individually with the rest of the variables. The significant variable was selected from the second run and further runs of the program were conducted in a similar manner.

For the first branch of the neural network tree, the test of significance showed that the AADT on the Major road was the most important variable in distinguishing rear end and sideswipe crashes from angle and turn crashes. The rest of the significant variables have been listed in Table 6.11. The table shows the significant variables found in each run, and the increase in accuracy with an increase in significant variables. The accuracy of the model increased to 59.12% when all the eleven significant variables were used. This is slightly more than the accuracy obtained when all variables were used. Since all variables are not always available, using these results can help in determining the needed variables.

Table 6-11 List of significant variables for distinguishing rear end and sideswipe crashes from the angle and turn crashes

Run#	Variables	Hidden Nodes	Test Accuracy	RE-SS	Angle-Turn
Run 1	AADT Major	12	55.50	65.12	46.86
Run 2	AADT Minor	9	56.00	65.12	46.86
Run 3	Speed Limit Minor	12	56.62	61.43	51.90
Run 4	Surface Conditions	14	56.82	56.61	57.59
Run 5	Light Conditions	14	57.38	63.37	51.63
Run 6	Major Lanes	13	57.83	57.31	58.42
Run 7	Speed Limit Major	25	58.14	59.94	56.34
Run 8	Left Turning Lanes	32	58.72	61.33	55.46
Run 9	Right Turn Channelized MN	29	58.82	56.52	61.86
Run 10	Left Turn Protected Minor	23	58.92	63.12	55.46
Run 11	Right Turn Channelized MJ	30	59.12	58.25	60.67

Since the PNN model performed better in distinguishing between rear end and sideswipe crashes, it was used to find the significant variables for the second branch. The spread was increased from 0.05 to 2 with increments of 0.05. The results obtained have been tabulated in Table 6.12. In the table, LTP indicates Left Turn Protected lanes and SL indicates Speed Limit. The table indicates that the number of through lanes on the minor roadway was determined as the first significant variable, followed by minor LTP lanes, major through lanes, major LTP lanes and the major speed limit. The accuracy of the final model is 58.9%.

Table 6-12 List of significant variables for distinguishing between rear end and sideswipe crashes

Run#	Variables	Spread	Test Accuracy	Rear End	Sideswipe
Run 1	MN Lanes	0.25	53.94	54.93	52.95
Run 2	MN LTP	0.35	57.09	52.19	61.98
Run 3	MJ Lanes	0.25	57.44	50.66	65.10
Run 4	LTP MJ	0.45	58.02	58.57	59.02
Run 5	MJ SL	0.55	58.90	49.31	66.84

The MLP neural network was used to distinguish between angle and turn crashes. The neural network was built similar to the first branch of the neural network tree. The complete list of variables is shown in Table 6.13. The most significant variable turned out to be the AADT on the minor roadway. The accuracy of this model was 57%, which is also slightly greater than the model developed considering all input variables. LTL indicates the total number of left turning lanes at the intersection.

Table 6-13 List of significant variables for distinguishing between angle and turn crashes

Run#	Variables	Hidden Nodes	Test Accuracy	Angle Crashes	Turn Crashes
Run 1	MNAADT	32	54.33	63.8503	42.7921
Run 2	LTP MN	12	54.85	52.0856	58.2701
Run 3	Surface Conditions	18	55.4	58.5027	54.173
Run 4	LTP MJ	7	55.87	40.4278	71.0167
Run 5	MJ AADT	18	56.47	58.7166	51.7451
Run 6	LTL	18	57	44.8128	67.9818

6.11 Estimating a Trend in the Significant Variables

The identification of significant variables was the first step in determining how the input variables affect the type of collision. The effect of change in input variables has to be found on the type of collision. This will indicate the type of crash that is likely to occur when the geometric or traffic characteristics change.

To accomplish this, a simulation database was built with all possible combinations of significant input variables. This database represented crashes that had occurred at various crash conditions and at intersections with different values of AADT, number of lanes and speed limits. The simulation database was used in the relevant neural network to determine the expected type of collision, given the crash conditions and the traffic and geometric characteristics of the intersections. This output was grouped according to the input variables, and the trend of the output with a change in the input variables was determined. This procedure is explained in greater detail in the following sections which discuss this “simulation” process for each branch of the neural network tree.

6.11.1 Simulation for Branch 1 of the Neural Network Tree

Since 11 variables were found significant in distinguishing the rear end and sideswipe crashes from angle and turn crashes, a simulation database was built using these variables. A program was written in MATLAB to build the database. The database was developed using the variables listed below:

1. The AADT on the major roadway was considered between 10000 and 80000 with increments of 10000.

2. Since the AADT on the minor roadway was below 60000, it was varied from 1000 to 60000 with increments of 6000.
3. Speed limit for the minor roadway was varied from 25 to 55 mph with increments of 10 mph.
4. Surface and light conditions were categorized as was done while developing the models.
5. The number of through lanes on the Major roadway was varied from 2 to 6 with increments of 2.
6. Speed limit on the major roadway was varied from 25 to 55 mph with increments of 10 mph.
7. The number of left turning lanes was increased from 0 to 8 with increments of 2.
8. The right turn channelized lanes were varied from 0 to 2 for both the major and minor roadways.
9. The number of protected left turning lanes on the minor roadway was varied from 0 to 4.

A database was built using these crash characteristics, but with the following constraints, as was observed in the database collected from the six counties:

1. The AADT on the minor roadway was either equal to or lower than the AADT on the major roadway.
2. It was observed from the county databases that the minimum speed limit on the minor roadway was 25 mph when the AADT was less than 20000, and 30 mph when more than 20000. Hence this criterion was implemented.

3. The number of major lanes for roadways having the major AADT lesser than 30000 was set to either 2 or 4. When the AADT was greater than 30000, the number of major lanes were either 4 or 6.
4. The number of right turn channelized (RTC) lanes on the minor roadway were set to be either equal to or lesser than the RTC lanes on the major roadway.

The simulated database obtained with this procedure consisted of 615600 different crashes. This database was used in the MLP neural network model, developed using the crash database for the year 2000 (the training database used in the model development phase). The network developed using this database was used to predict the collision type for the simulated database. Thus results were obtained to indicate the collision type of each crash given the input variables.

To establish a relationship between the input variables and the type of collision, the number of rear end and sideswipe crashes were determined for each value of the input variables. This was divided by the total number of crashes in the simulation database with the respective values of input variables to obtain the percentage of rear end and sideswipe crashes. For example, the number of rear end and sideswipe crashes classified by the MLP neural network model for the value of Major AADT = 10000 was found. This value was divided by the total crashes in the simulation database having Major AADT of 10000 to obtain the percentage of rear end and sideswipe crashes for this value of Major AADT. The percentages were similarly found for the other Major AADT values (20000 to 80000). These values were plotted on a graph to obtain the variation of rear end and sideswipe crashes with Major AADT. The graph was not plotted for the total number of rear end and sideswipe crashes because the number of crashes for each value of input variable was different in the simulation database due to the constraints used while developing the data (that is, the number of crashes in the simulation database for Major AADT = 10000, 20000, etc were different). Converting the number of crashes into percentages gives an unbiased estimate of the relationship between the significant input variables and the type of collision. This percentage signifies a value where if 100 crashes have occurred at different intersections with AADT of 10000 (or any other input value), this percentage of crashes will be of the rear end or sideswipe collision type. Graphs were plotted for angle and turn crashes in a similar way.

The graphs for the significant variables have been shown below along with a possible explanation for the type of trends observed. The results have been compared to other studies, like Abdel-Aty and Keller (2005).

1. *Major AADT:* The graph for the Major AADT has been shown in Figure 6.3. Clearly, the number of rear end and sideswipe crashes increase relative to the angle and turn crashes when there is an increase in AADT on the major roadway. This relationship can be understood by the mechanism of the type of crashes. As the traffic volume increases, the roadway gets more and more congested, thus reducing the spacing between the following vehicles. When a vehicle stops or decelerates, there is a higher chance of a rear end crash compared to a situation where the spacing between vehicles is more. Similarly, a person trying to change lanes is more likely to have a sideswipe crash when the traffic is more and the spacing between vehicles is less. Hence more of rear end and sideswipe crashes can be expected when the traffic volume increases. This result can be deduced from the analysis conducted in the previous chapter. For the rear end and

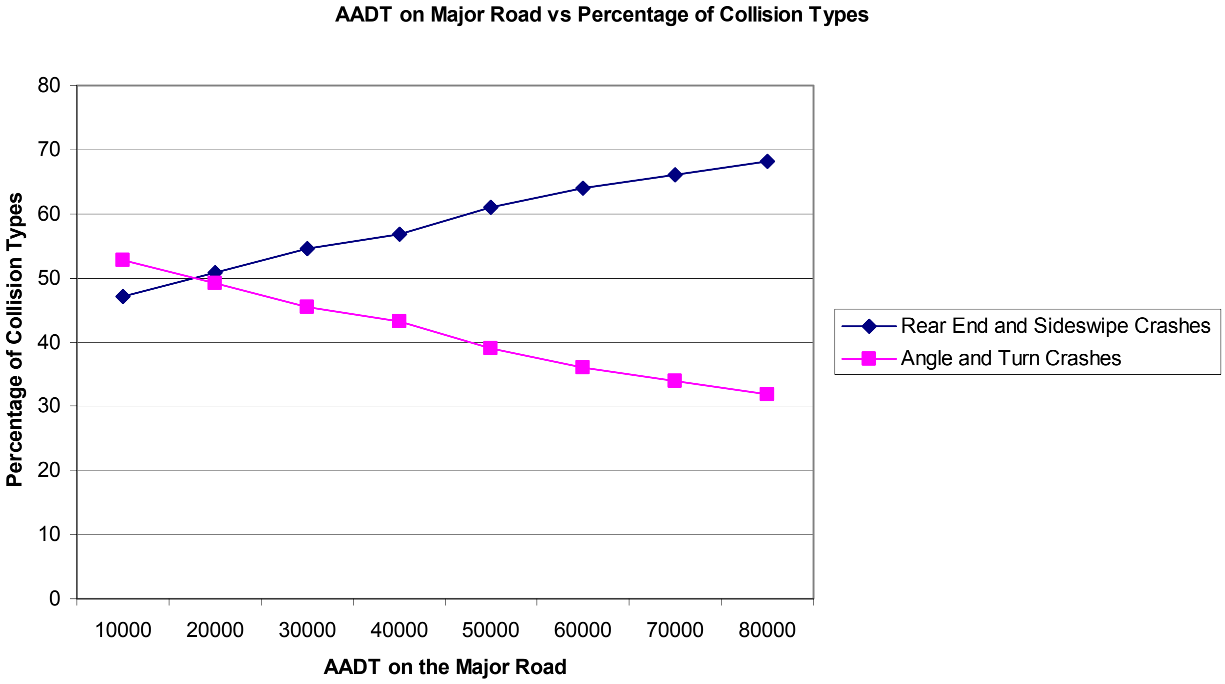

Figure 6.3 Variation of the collision types with Major AADT

sideswipe crashes in Figures 4.9 and 4.36, the increase in crash rate is higher compared to the angle and turn crashes as shown in Figures 4.16 and 4.21. This result is also supported by Abdel-Aty and Keller (2005), who observe that AADT on the Major road is significant for rear end, sideswipe and right turning crashes, but is not significant at all for angle and left turning crashes. But the interesting point to note is that for a very low AADT, the number of angle and turn crashes are more than rear end and sideswipe crashes. This is possible because the spacing between vehicles is large in low traffic conditions, and there is a lesser chance of a rear end crash. Higher spacing between vehicles also implies that the lane changing maneuvers will be safer, thus reducing the possibility of a sideswipe crash. Thus, if a crash occurs when the AADT is low, it will have a higher chance of being an angle or a turn crash. AADT was the most important factor observed by Greibe (2003) and Chin and Quddus (2003).

2. *Minor AADT:* The pattern observed for the AADT on the minor roadway is similar to that of the major AADT, as has been shown in Figure 6.4. Abdel-Aty and Keller (2005) find that the angle, rear end and sideswipe crashes are most affected by the minor AADT, and turn crashes are not affected at all.
3. *Minor Speed Limit:* According to Poch and Mannering (1996), total and rear end crashes increase and the other crash types are unaffected with an increase in approach speed limit. Abdel-Aty and Keller (2005) find that the speed limit on the minor road is significant in predicting angle, left turn and rear end crashes. But the increase in rear end crashes is found to be the highest due to the increase in the speed limit. Therefore, the increase in speed limit can be expected to increase the rear end crash and sideswipe crashes more than the angle and turn crashes. This has been found in Figure 6.5.

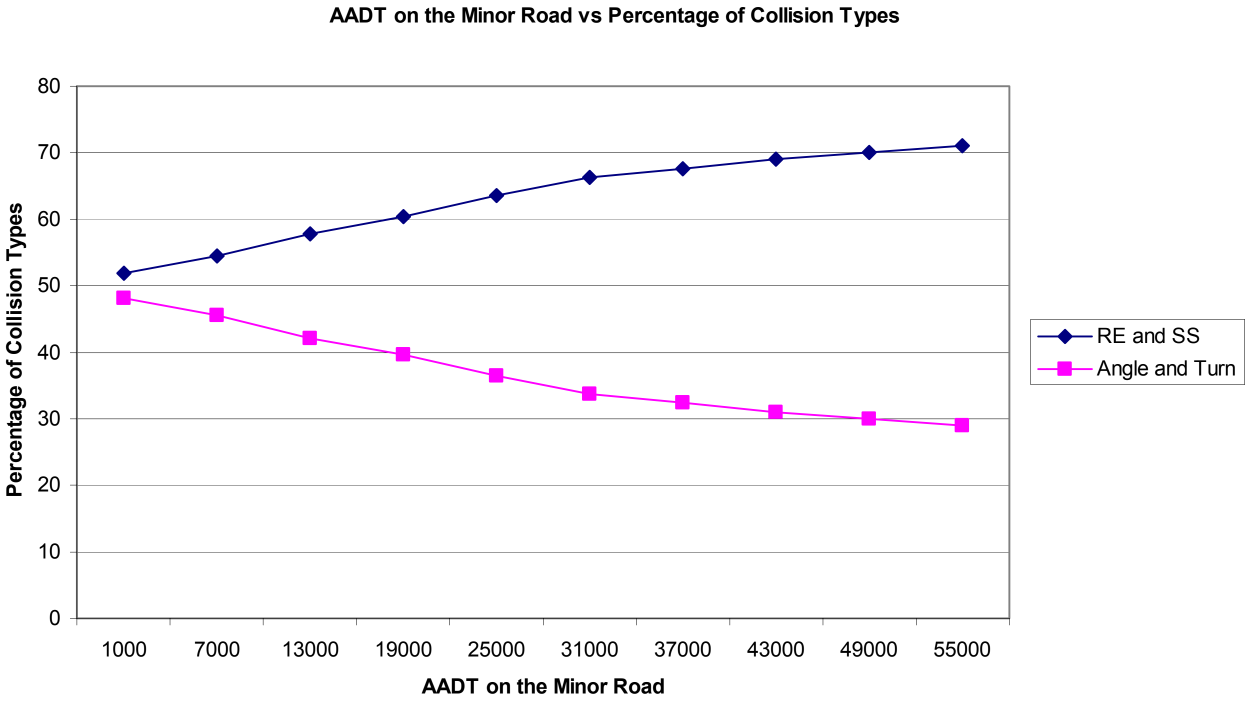

Figure 6.4 Variation of the collision types with Major AADT

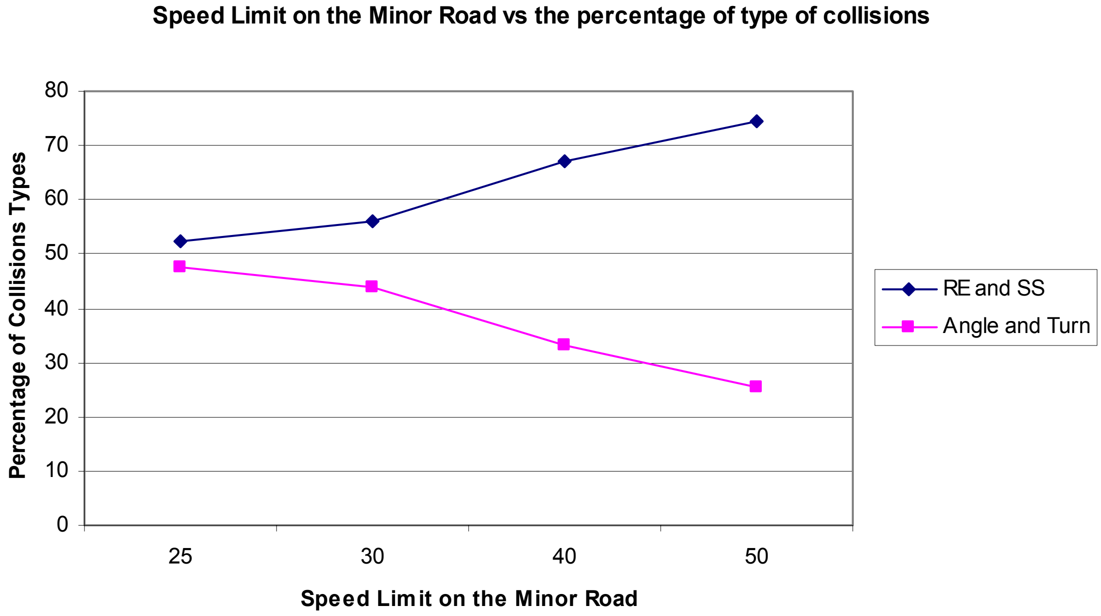

Figure 6.5 Variation of the collision types with Minor Speed Limit

4. *Surface Conditions:* According to Figure 6.6, more rear end and sideswipe crashes can be expected in wet surface conditions compared to dry surface conditions. This is true because when the surface is wet or slippery, it is takes a longer time to stop the vehicle, thus increasing the chances of colliding with the vehicle in lead. A crash is more likely to be a rear end crashes compared to an angle or a turn crash in such conditions.
5. *Light Conditions:* When the light conditions are dark, drivers can see the vehicles going along in their direction clearly. Angle or turn crashes are more likely to happen at such conditions because the vehicle coming in the other roadway is difficult to spot, and hence there is a greater chance of such collisions. This has been observed in Figure 6.7.

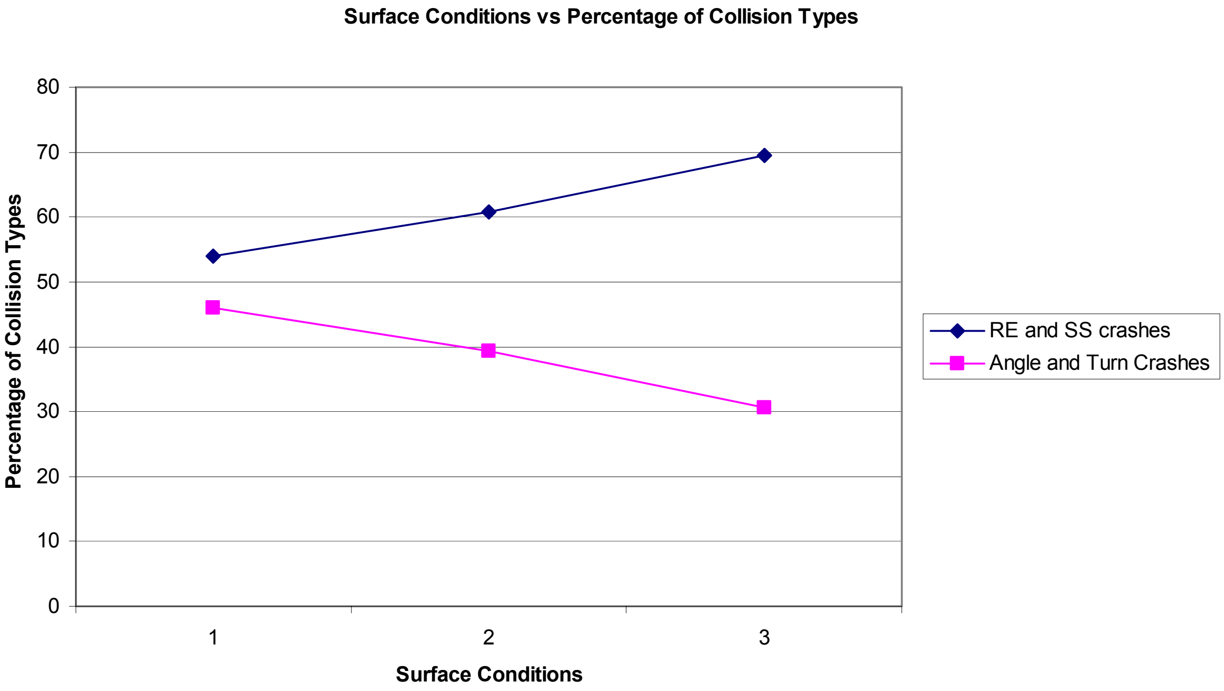

Figure 6.6 Variation of the collision types with Surface conditions

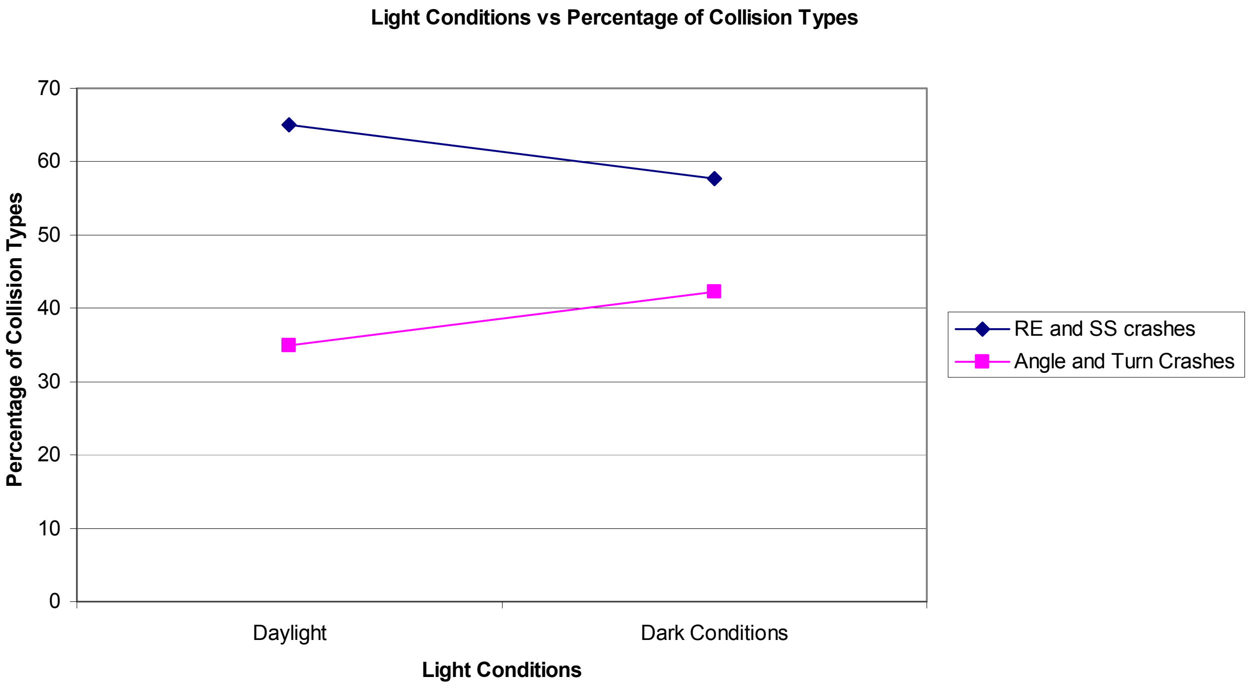

Figure 6.7 Variation of the collision types with Light conditions

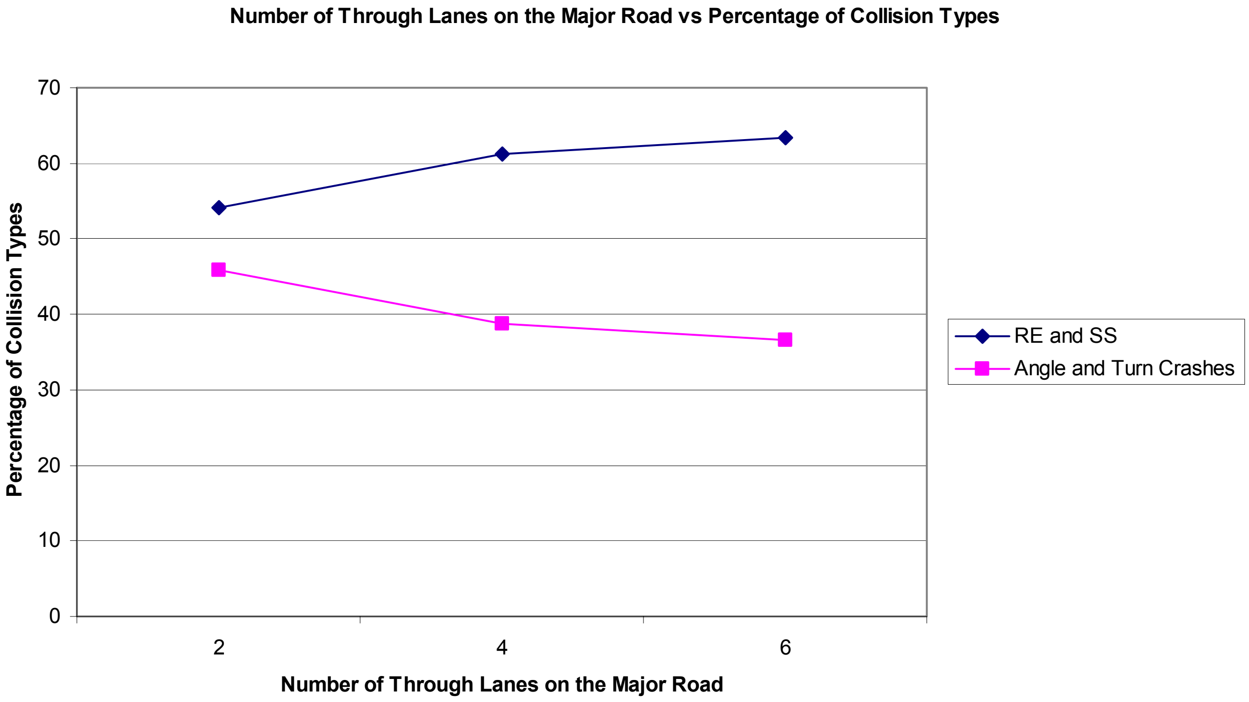

Figure 6.8 Variation of the collision types with Major Lanes

6. *Number of Major Lanes:* In the analysis conducted in the previous chapter, it was found that the increase in the number of lanes on the major road increases all types of crashes. Abdel-Aty and Keller (2005) find that the number of major lanes is not significant in predicting only left turning crashes. Figure 6.8 gives a deeper insight showing that if the number of lanes on the major road increases, the crashes is more likely to be a rear end or a sideswipe crash. Therefore, if an intersection is already subjected to high rear end and sideswipe crashes, an increase in number of through lanes on the major road of that intersection will only make the intersection more dangerous for such crashes.
7. *Major Speed Limit:* The results produced in the analysis showed that the rear end and sideswipe crashes are more likely to happen as the speed limit on the major road increases, as can be seen in Figure 6.9. But at lower speed limits, angle and turning crashes are more likely to occur. Abdel-Aty and Keller (2005) find that the speed limit on the major road is only significant in predicting angle and rear end crashes. The study in the previous chapter found the major speed limit to be significant in predicting the rear end crashes. Therefore, the increase in speed limit can be expected to increase the rear end crash and sideswipe crashes more than the angle and turn crashes. However, at lower speed limits, the rear end and sideswipe crashes are less likely to occur because a vehicle traveling at such speeds can stop easily to prevent such crashes. Therefore, if there is a crash at lower speed limits, it is more likely to be an angle or turn crash.

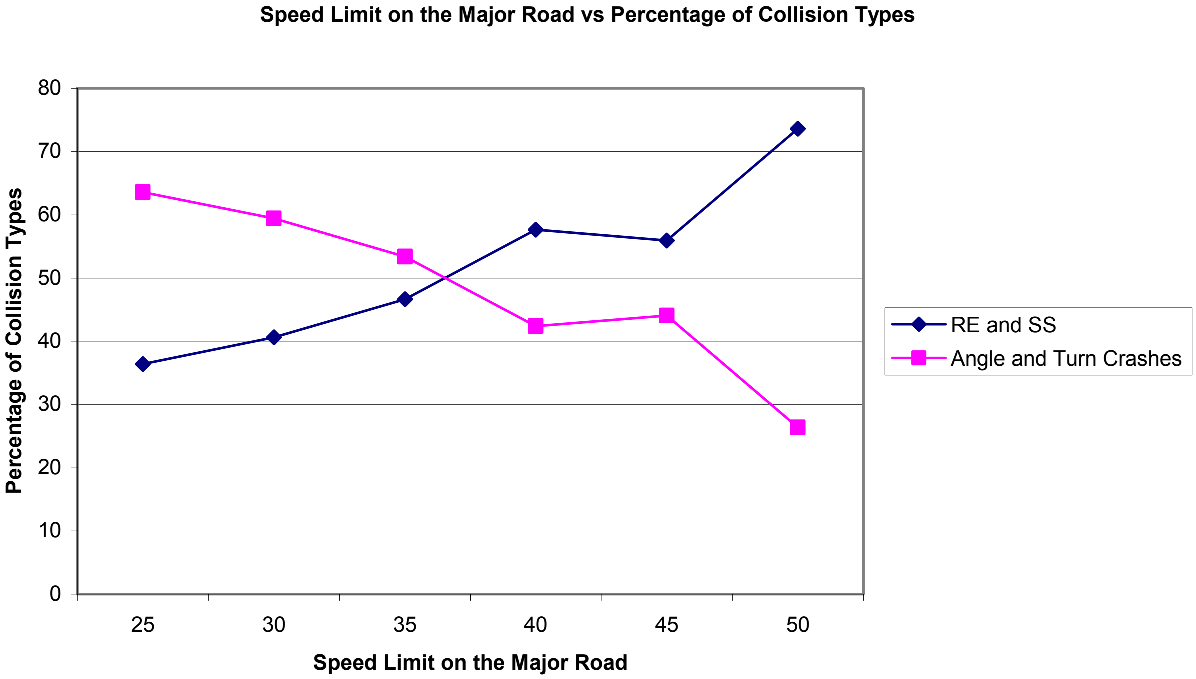

Figure 6.9 Variation of the collision types with Major Speed Limit

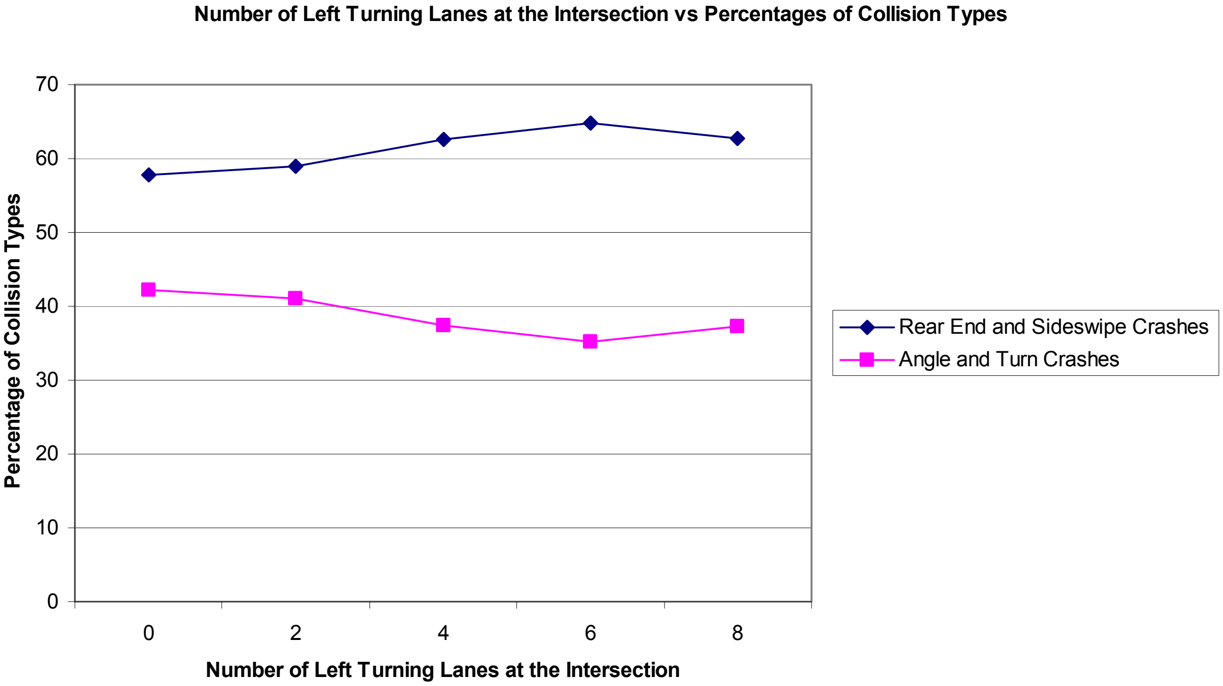

Figure 6.10 Variation of the collision types with Left Turning Lanes

8. *Number of Left Turning Lanes:* An increase in the left turning lanes at the intersection result in crashes more likely to be rear end and sideswipe crashes compared to angle and turn crashes, as can be seen in Figure 6.10. The possible reason for this phenomenon is that a greater number of left turning lanes indicates greater amount of left turning traffic trying to get through the protected phase. When the protected phase is about to end, the vehicles in the left turning bay try to finish the left turning movement. If a driver slows down in such a case, it leads to a rear end crash wherein the driver of the vehicle behind the slowing vehicle does not slow down (in order to finish the left turning maneuver) and rear ends the slowing ahead. Thus rear end crash is more likely in such a case. Also, greater number of left turning lanes can be expected to decrease the number of left turning crashes.
9. *Minor RTC Lanes:* Minor RTC lanes have been found to increase the number of rear end crashes in the analysis conducted in the previous chapter. The pattern of crashes observed on increasing the number of channelized right turning lanes on the minor road is shown in Figure 6.11.
10. *Minor LTP Lanes:* An increase in the LTP lanes on the minor roadway can increase the possibility of a rear end or sideswipe crash compared to an angle or turn crash, as can be seen in Figure 6.12. The possible explanation for this is similar to the theory given for rear end and sideswipe crashes being more likely with an increase in the total left turning lanes.

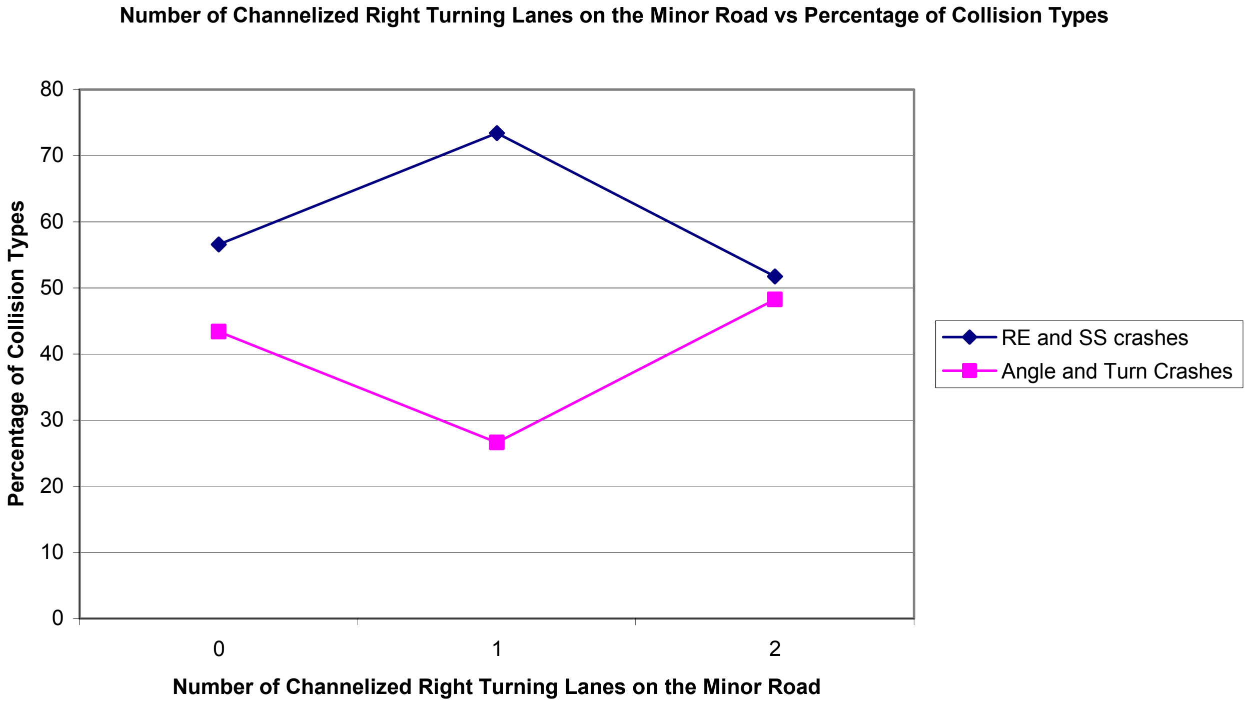

Figure 6.11 Variation of the collision types with Minor RTC Lanes

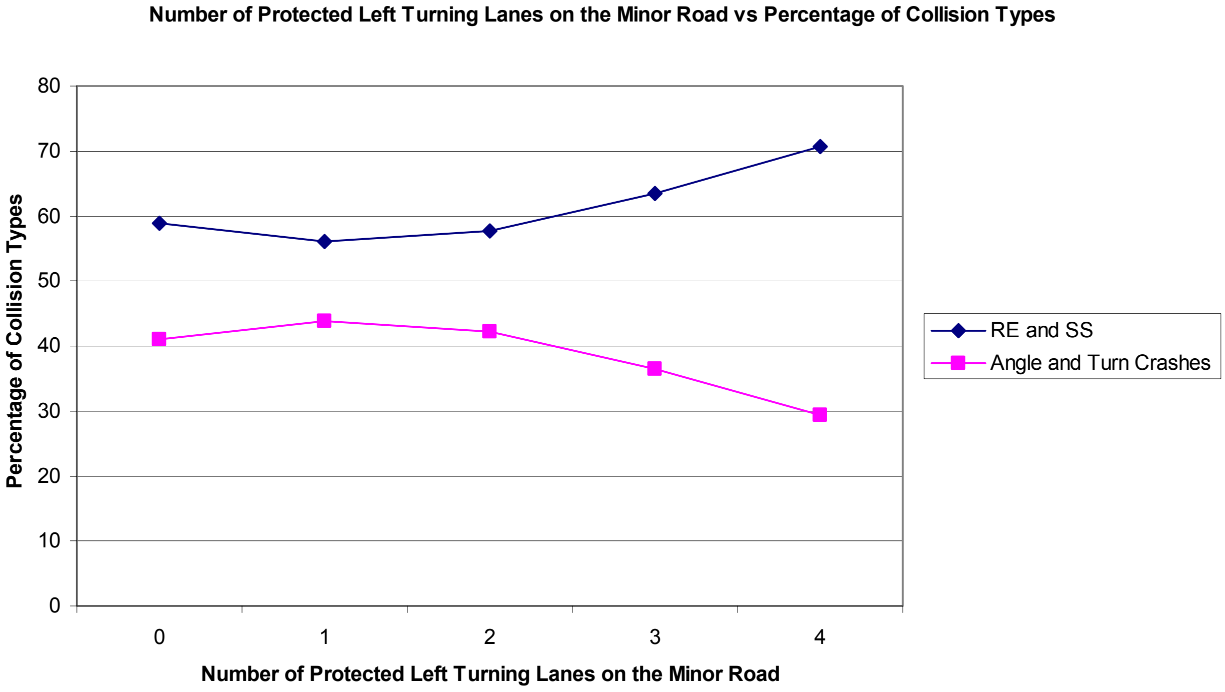

Figure 6.12 Variation of the collision types with Minor LTP Lanes

11. *Major RTC Lanes:* According to Abdel-Aty and Keller (2005), only rear end and sideswipe crashes are affected by a change in RTC lanes on the major road. In the study conducted in the previous chapter, turning and sideswipe crashes are affected by this variable. But the increase in sideswipe crashes is larger when the RTC lanes increase. Hence a crash can more likely be either a rear end or a sideswipe crash when the number of RTC lanes on the major increase, as is seen in Figure 6.13.

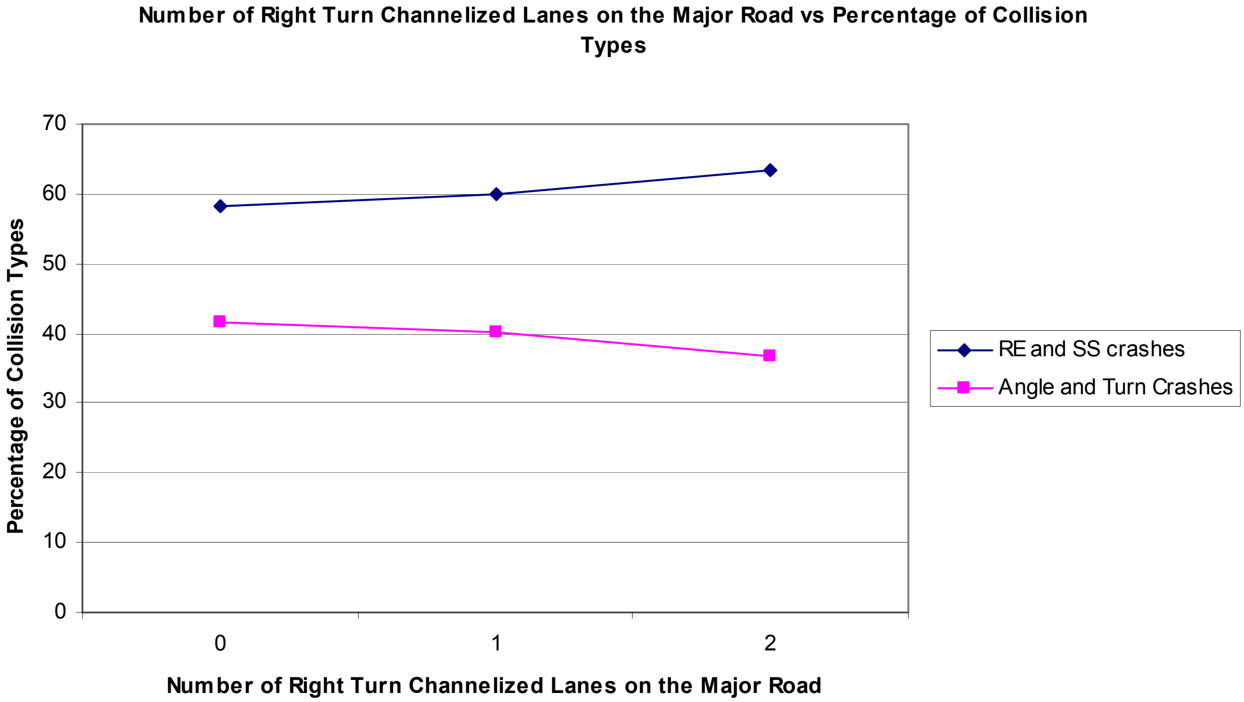

Figure 6.13 Variation of the collision types with Major RTC Lanes

6.11.2 Simulation for the Second Branch of the Neural Network Tree

A simulation database was developed in a method similar to the database built for branch 1. This database consisted of all possible combinations of the five significant input variables. The limits of the input variables were same as those used in the database created for the first branch of the neural network tree. This simulated database consisted of 2625 crashes. A PNN model was developed using the training data of the year 2000 for rear end and sideswipe crashes. The simulation database was used to predict the output and this output was summarized and plotted. The variations observed are as follows:

1. *Minor Lanes:* Although the Table 4.18 suggests that both rear end and sideswipe crashes increase with the increase in number of lanes on the minor road, Figure 6.14 shows that the chances of a crash being a sideswipe crash increase with an increase in the number of lanes. The possible reason for this is that the number of lane changing maneuvers increase with an increase in the number of lanes, thus leading to a greater probability of a crash being sideswipe. Abdel-Aty and Keller (2005) find that the number of lanes on the minor roadway does not affect either of the crash types. But the present analysis gives an appropriate result that has a good reasoning associated with it.
2. *Minor LTP Lanes:* As stated in the previous section, an increase in this variable leads to an increase in rear end crashes. Also, the analysis for predicting the crash frequencies suggests that the minor LTP lanes only affect the rear end crash frequency. As can be seen in Figure 6.15, a crash is more likely to be a rear end crash when the number of LTP lanes on the minor road increase.

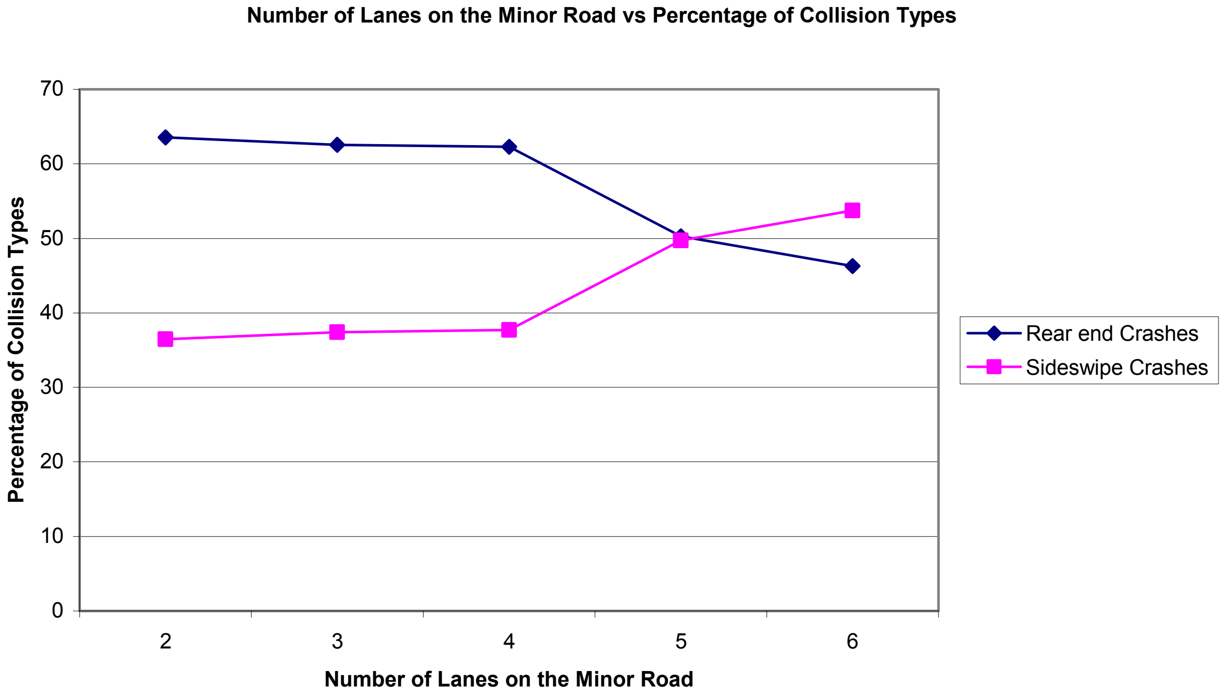

Figure 6.14 Graph indicating the variation of rear end and sideswipe crashes with the through lanes on the Minor road

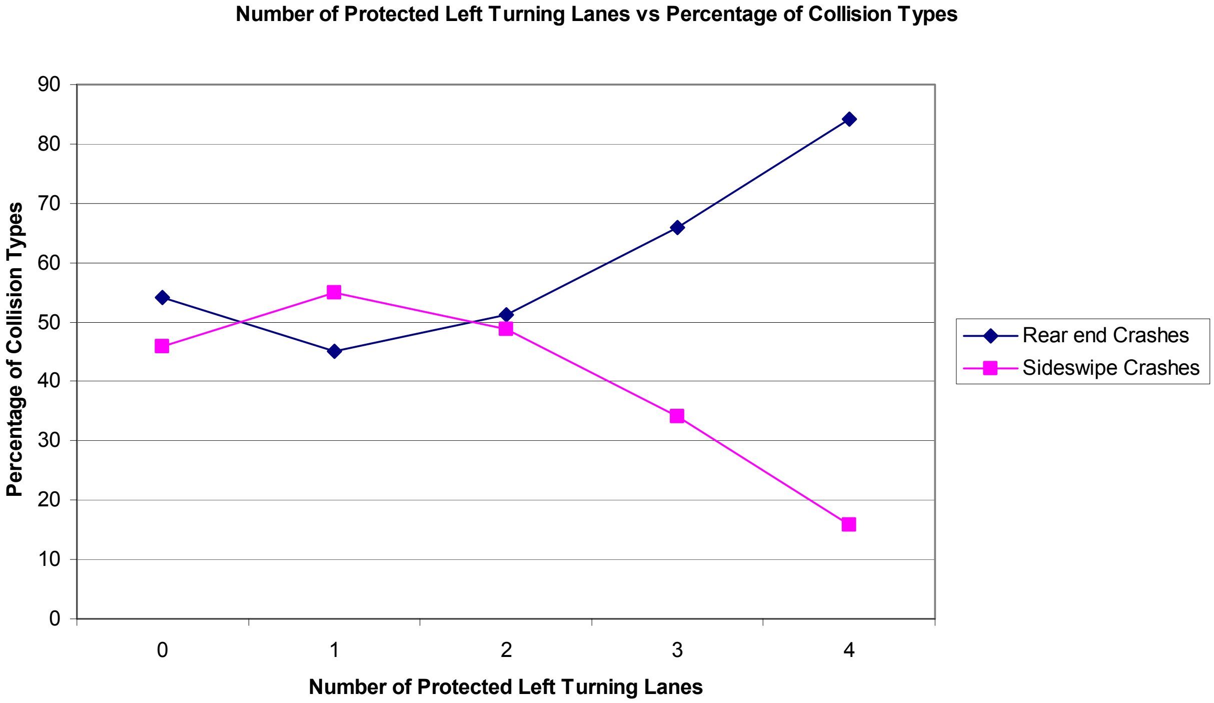

Figure 6.15 Graph indicating the variation of rear end and sideswipe crashes with the protected left turning lanes on the Minor road

3. *Major Lanes:* Figure 6.16 suggests that the chances of a crash being sideswipe increase very slightly when the number of lanes on the major road increase.

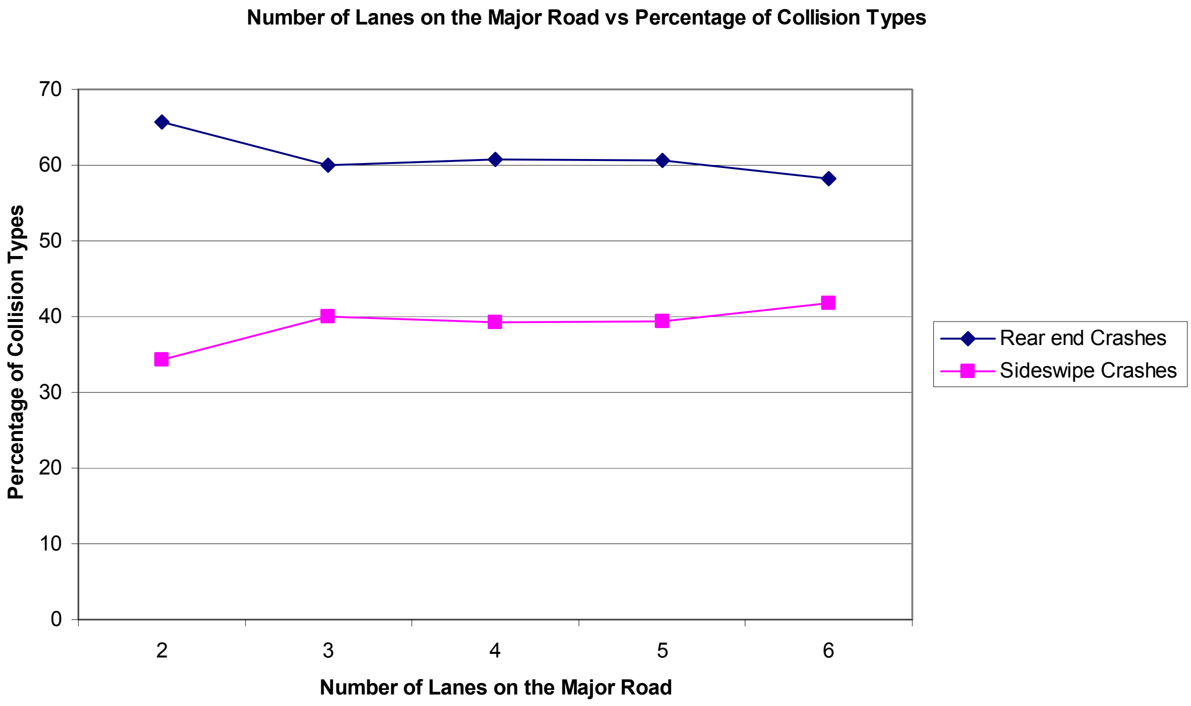

Figure 6.16 Graph indicating the variation of rear end and sideswipe crashes with the through lanes on the Major road

4. *Major LTP Lanes:* The trend obtained for an increase in major LTP lanes can be seen in Figure 6.17. As seen in the previous chapter, the sideswipe crashes double when the major LTP lanes increase, whereas the rear end crashes increase only slightly. This shows that as major LTP lanes increase, the probability of crash being sideswipe is higher.

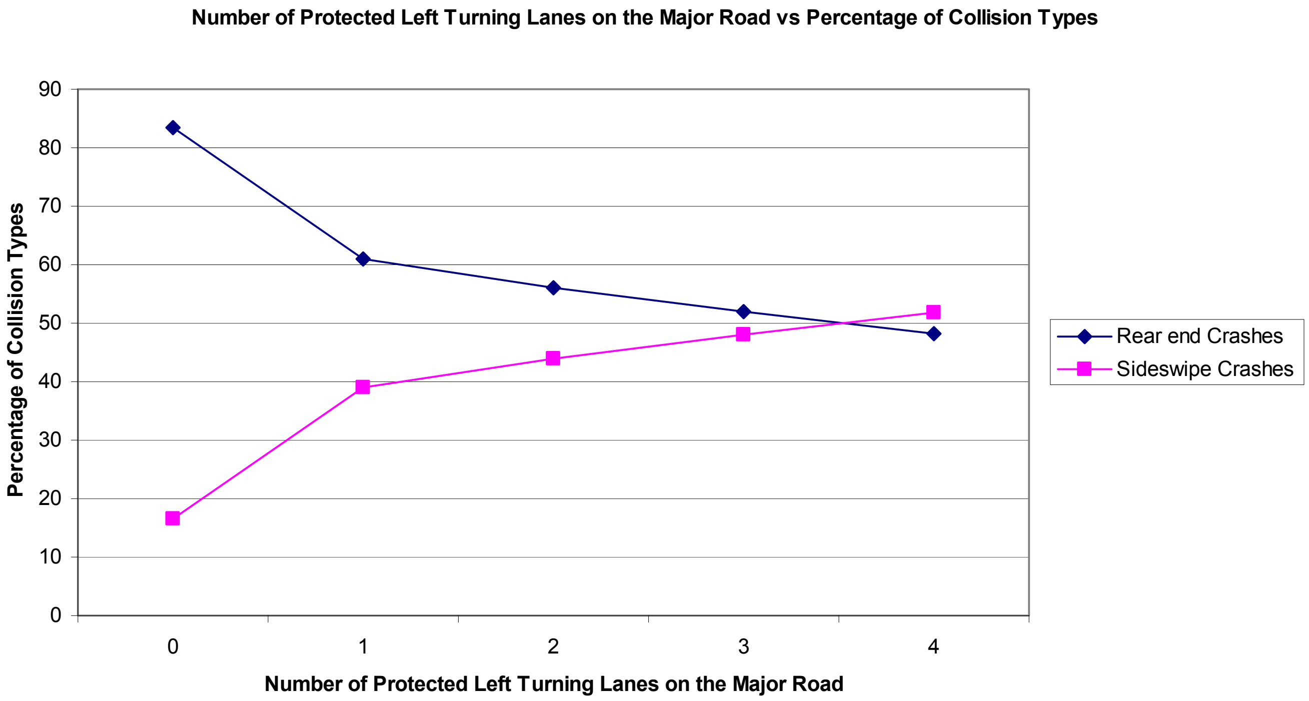

Figure 6.17 Graph indicating the variation of rear end and sideswipe crashes with the protected left turning lanes on the Major road

5. *Major Speed Limit:* In the studies conducted for predicting the frequency of rear end and sideswipe crashes in the previous chapter, by Poch and Mannering (1996) as well as by Abdel-Aty and Keller (2005), the rear end crashes show an increasing trend with an increase in the speed limit on the major road. But the sideswipe crashes do not show any variation with speed limit. Hence the rear end crashes are more likely to occur as the speed limit on the major road increases.

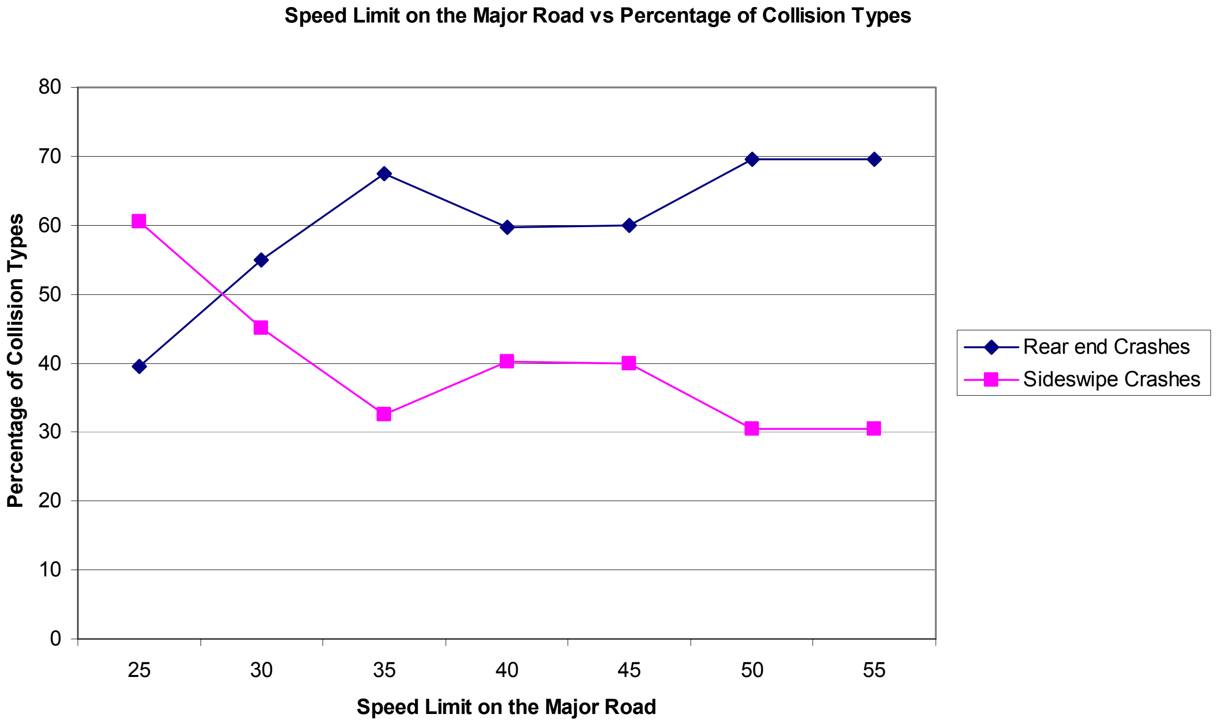

Figure 6.18 Graph indicating the variation of rear end and sideswipe crashes with the speed limit on the Major road

6.11.3 Simulation for the Third Branch of the Neural Network Tree

The simulation database for the third branch of the neural network tree consisted of 13275 crashes generated by using the six significant variables identified in the model. The results of this analysis have been listed below.

1. *Minor AADT:* In the study conducted by Abdel-Aty and Keller (2005), only the angle crashes are affected by the minor AADT and they show an increasing trend with the increase in this variable. The result obtained in the present analysis is consistent with this result, and has been illustrated in Figure 6.19.

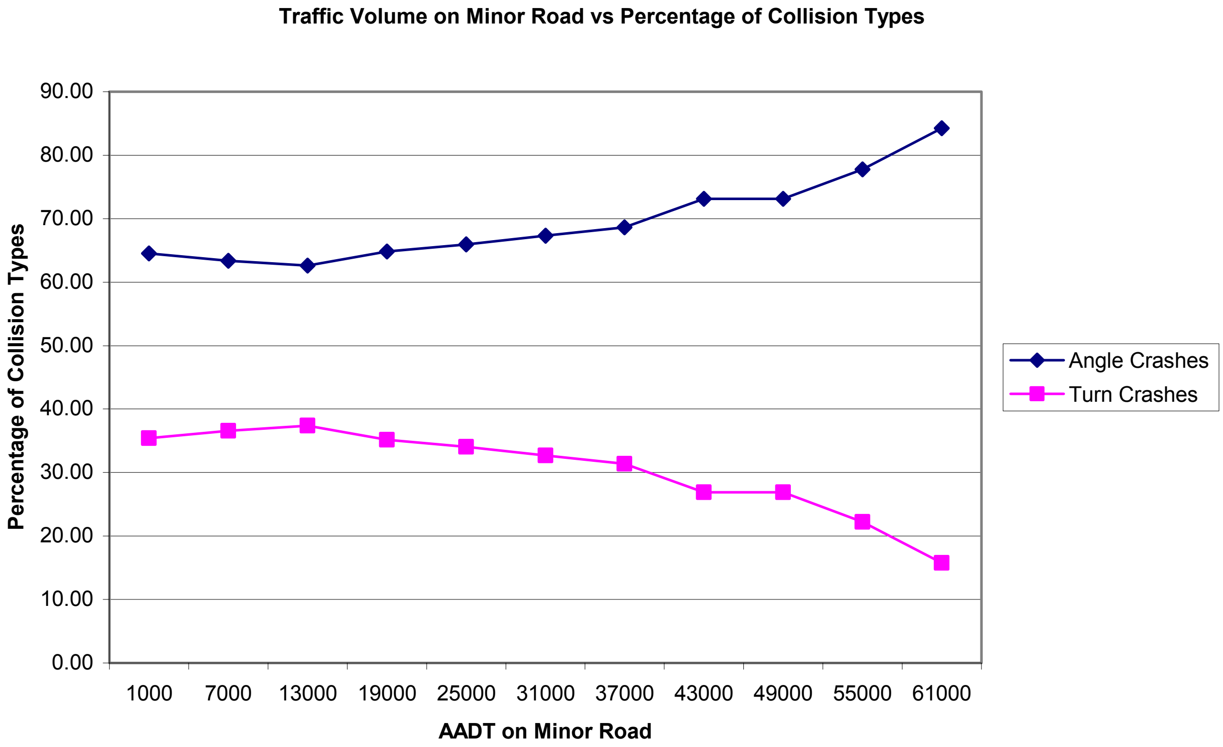

Figure 6.19 Graph indicating the variation of angle and turn crashes with the AADT on the Minor road

2. *Minor LTP Lanes:* According to Figure 6.20, the chances of a crash being a turn crash increase with an increase in LTP lanes on the minor road. But the chances of a crash being an angle crash are always higher. The probable reason for this is that an increase in the LTP lanes implies that the traffic on the roadway is large. As the traffic increases, a crash is more likely to be an angle crash as can be seen in Figure 6.19. Hence, with an increase in the minor LTP lanes a crash is more likely to be an angle crash.

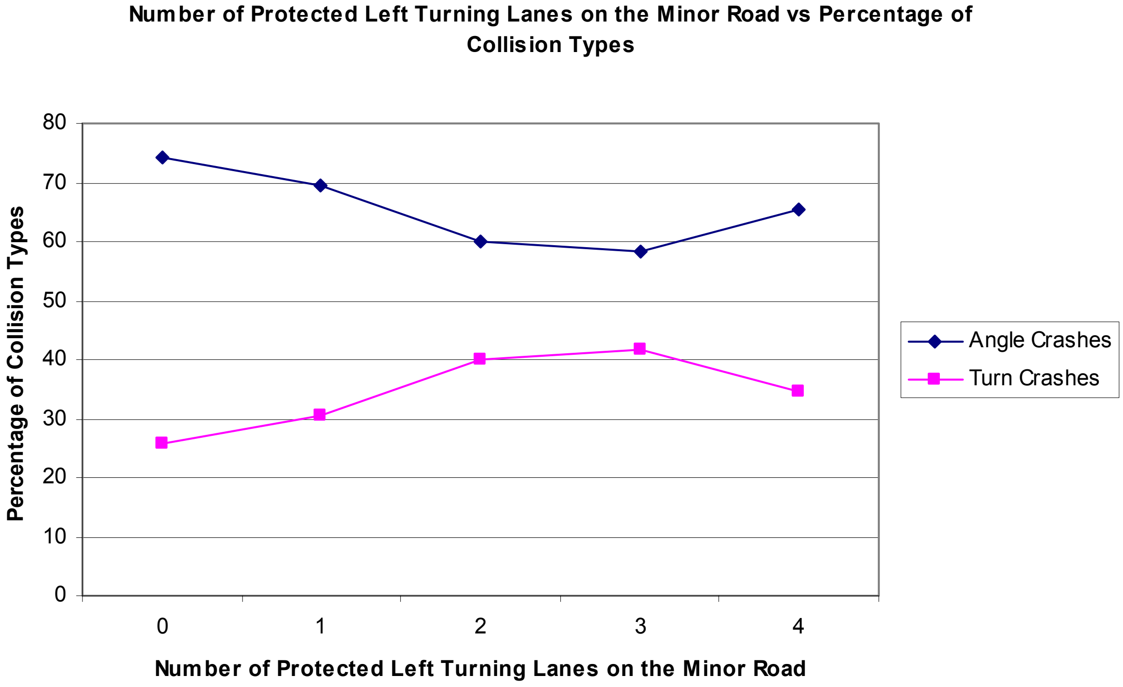

Figure 6.20 Graph indicating the variation of angle and turn crashes with the protected left turning lanes on the minor road

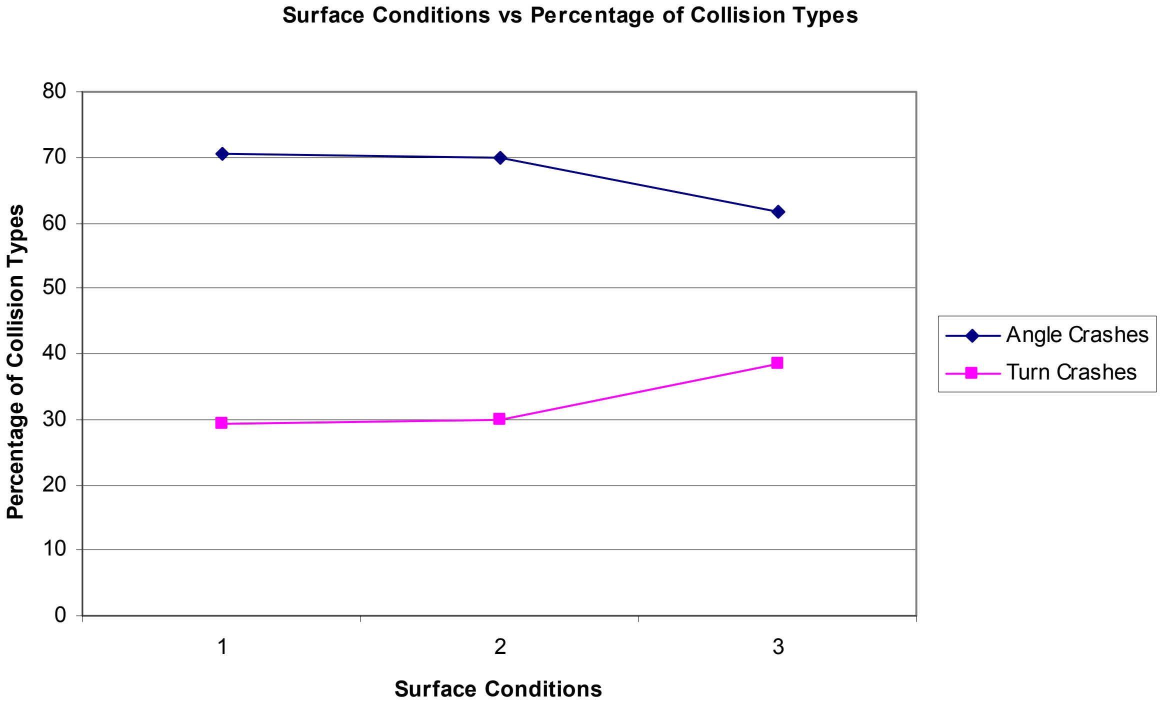

Figure 6.21 Graph indicating the variation of angle and turn crashes with the surface conditions

3. *Surface Conditions:* Figure 6.21 shows that a crash is more likely to be an angle crash when the surface conditions are dry, wet or slippery.

4. *Major LTP Lanes:* Figure 6.22 shows that a crash is more likely to be an angle crash for any number of LTP lanes on the major road. But the chances of a turn crash increase with an increase in the variable. The reasoning is similar to that given for Minor LTP lanes.

5. *Major AADT:* As the AADT increases, turn crashes are more likely to occur compared to angle crashes, as can be seen in Figure 6.23. According to the crash frequency prediction in the previous chapter as well as in the study by Poch and Mannering (1996), both the collision types show an increasing trend with the approach volume. But Abdel-Aty and Keller (2005) report that only left turning crashes increase with an increased AADT on the major road. Thus this result is justified. But it is the exact opposite to the trend observed for minor AADT because as the minor AADT increases, the chances of interactions between the vehicles traveling on major and minor road increases that lead to a higher likelihood of angle crashes.

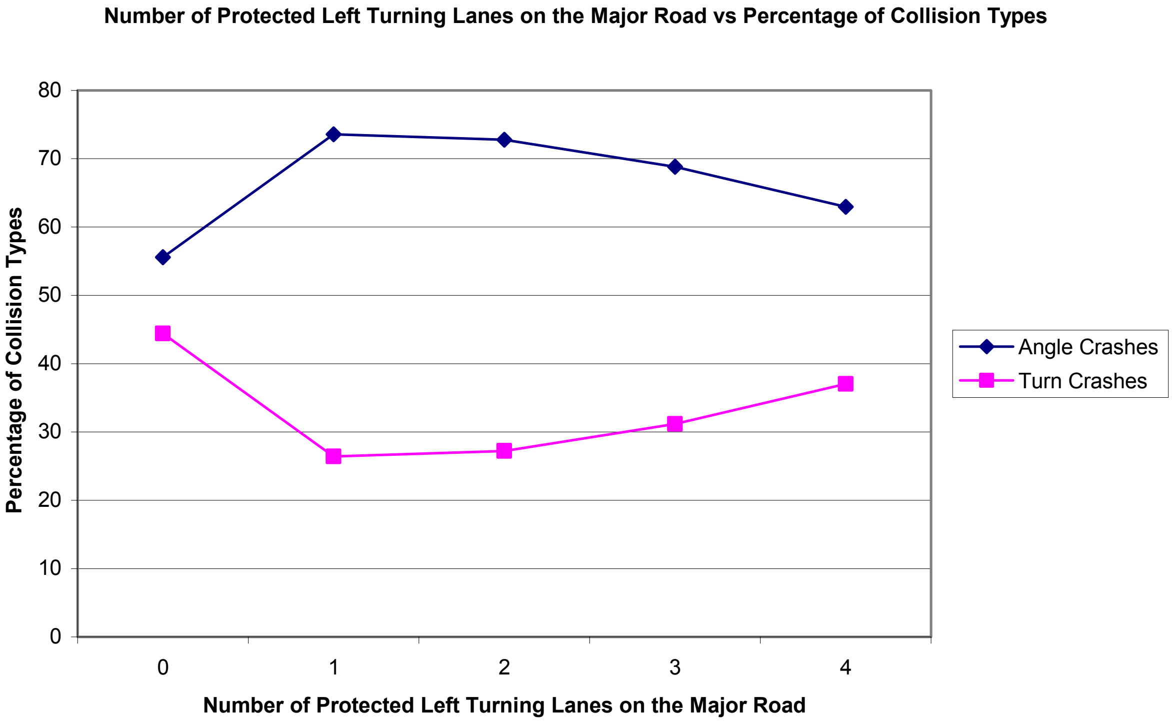

Figure 6.22 Graph indicating the variation of angle and turn crashes with the protected left turning lanes on the major road

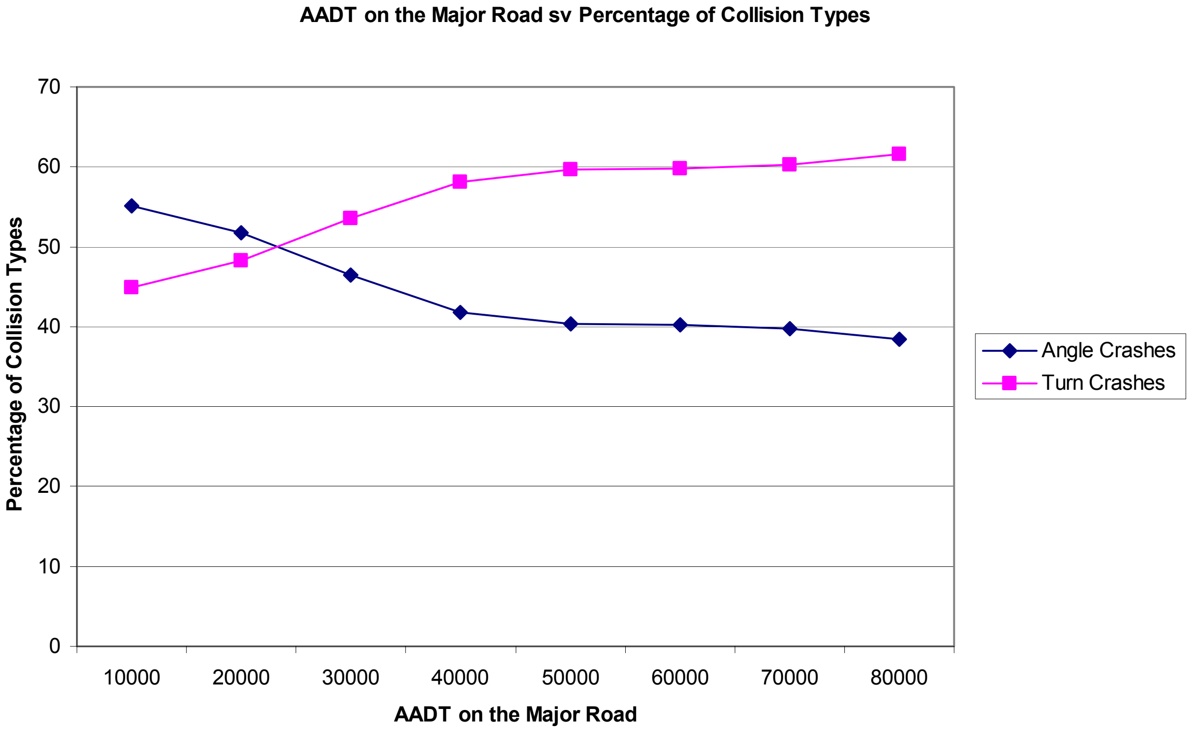

Figure 6.23 Graph indicating the variation of collision types with the AADT on the major road

6. *Left Turning Lanes:* The result shown in Figure 6.24 is a direct reflection of the results obtained for LTP lanes on the major and minor roadway.

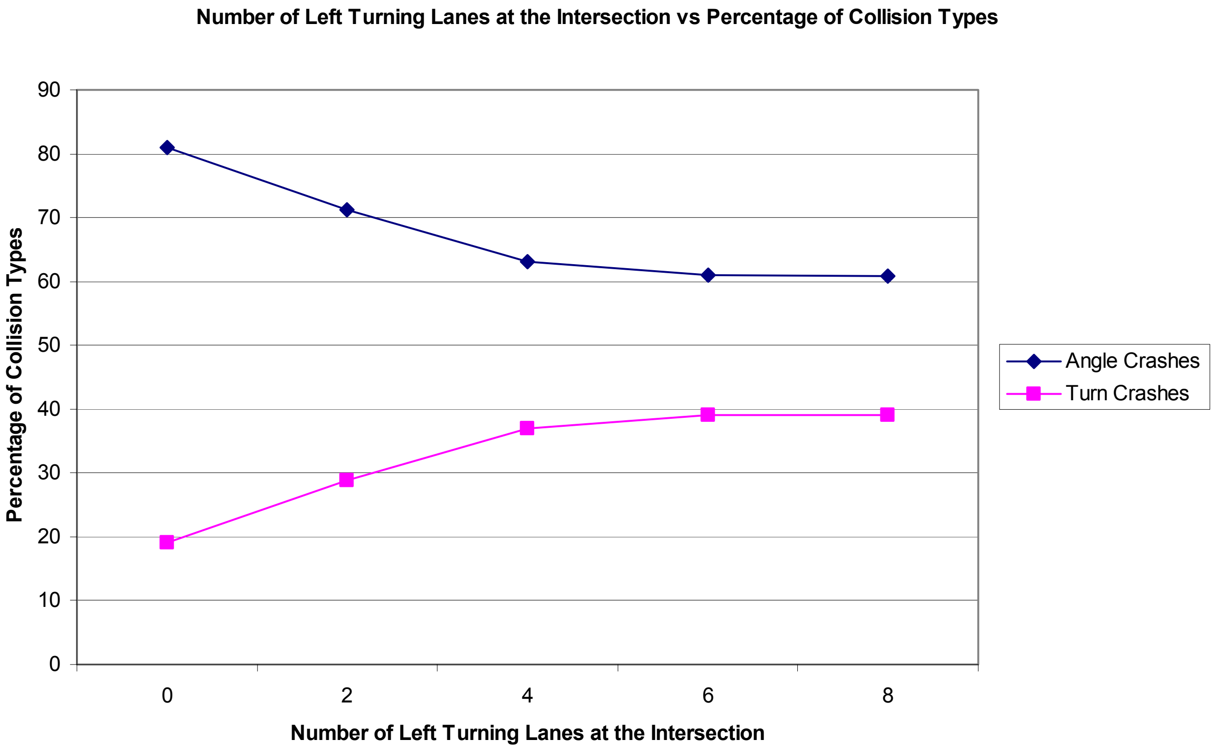

Figure 6.24 Graph indicating the variation of angle and turn crashes with the total left turning lanes

6.12 Summary

This study explores several methods used to identify the collision type of a crash given the crash conditions and the geometric and traffic characteristics of the intersection at which the crash has occurred. Two neural network models: Multi Layer Perceptron (MLP) Neural Network and Probabilistic Neural Network (PNN) have been used to develop the models. At first, the crashes were classified into rear end, angle, turn or sideswipe crashes by using these models. But the results of this analysis were not very encouraging. Hence a new method was developed, the Neural Network Tree, to classify the crashes into their respective collision types. The tree would first classify all crashes into either rear end and sideswipe crashes or angle and turn crashes. Then the crashes would be classified to rear end and sideswipe crashes separately, and angle and turn crashes separately. The MLP and PNN models built performed well. They were compared and the best model was used for predicting the corresponding collision types. The significant variables were then identified for each of these models using a forward sequential method. This was followed by the building of simulation databases for each model using all possible combinations of the significant input variables. The output of the simulation database was used to study the influence of the input variables on the collision type classified. The relation between the input variables and the collision types was studied, and was compared to other studies. The results obtained were found to match very well with previous studies. For instance, for distinguishing between rear end and sideswipe from angle and turn crashes, Abdel-Aty and Keller (2005) observe that the major and minor AADT affects the rear end and sideswipe crashes more than angle and turn crashes, which is the same observation made is the present study. Therefore it can be concluded that using Neural Network Trees results in reasonably accurate results. Thus the Neural Network Tree can be used as an effective method in classifying various collision types.

7 SUMMARY AND CONCLUSIONS

Intersections generally experience high crash rates. Due to the vehicles arriving and leaving in different directions, there are a large number of conflict points at intersections. This implies a greater chance of a crash occurring at these places. The intersections can be made safer by studying the characteristics of the intersections that affect different types of crashes. These properties can be controlled during the design of a roadway, thereby designing intersections that are less prone to crashes.

This research delves into the safety of signalized intersections. The first objective was to predict the frequency of crashes at signalized intersections and to identify the traffic and geometric aspects of the intersections that most affect the crash frequencies. The second objective was to classify the crashes into their respective collision types based on the conditions at the time of the crash and the traffic and geometric characteristics of the intersection at which the crash occurred.

The first task was to extensively review the work carried out in this field and to study the various techniques employed in these studies. The usage of Negative Binomial, Poisson, Nested Logit, Ordered Probit, Regression Trees, Neural Network Models and GIS techniques in some of the studies that were reviewed. The benefits of these methods were analyzed and it was decided to use the neural network models because of their various advantage, such as their ability to perform non-linear operations very efficiently, their capability of learning and generalizing, and their ability to produce reasonable results by adapting to new inputs not encountered during training. The Multilayer Perceptron (MLP), Probabilistic Neural Networks (PNN), and Generalized Regression Neural networks (GRNN) were utilized to perform the analyses.

The next task consisted of collecting the data necessary to perform the analysis. Various counties in Florida were contacted to obtain data on the signalized intersections and on the crashes occurring at these locations. The data was collected from Brevard, City of Orlando, Hillsborough, Miami-Dade, Orange and Seminole Counties. First, a geometry database was developed that contained all the geometric and traffic characteristics of the intersections. This database totally consisted of 1562 intersections from the six counties. Secondly, a crash database was built that consisted of the characteristics of all the crashes occurring at a distance of 250 ft from the intersection. Thirdly, the two databases were merged to form a Master Database. Finally, these databases were combined for all the six counties to obtain a Combined Database. This database was built for the years 2000-2001 as they were the common years for which the crash data was available for the all counties. This was followed by classifying intersections in each county into 19 categories such that each category represented a set of similar geometric and traffic characteristics. The means of each category of crashes were compared to the means of the respective categories in the combined database to identify the counties whose mean number of crashes differed considerably from the mean crashes occurring in the six counties combined. Then the intersections in the combined database were finely classified into 38 categories so that each category represented intersections with similar traffic and geometric aspects more accurately.

In order to predict the frequency of crashes occurring at signalized intersections, the MLP and GRNN models were developed and tested using the data in the combined database. The crash data for the year 2000 was used for the training phase of the model and the data for 2001 was used for testing the accuracy in the prediction of crash frequencies. The Root Mean Squared Error (RMSE) for the models was checked, and this turned out to be very high. To minimize the error in the prediction values, a new method was devised. This method first classifies intersections into safe and unsafe categories depending on the size of intersection (that is, the total number of lanes at the intersection), and then predicts the crash frequencies for the two categories separately using different neural network models. The MLP and PNN models were developed for the classification phase and the PNN model was found to perform marginally better. MLP and GRNN models were developed for the predicting the crash frequencies for the two categories, and the MLP neural network was found to perform marginally better than the GRNN model for both categories. The error in predicting the frequency of crashes using the new technique was considerably lesser than the error obtained in the previous models developed. This was followed by the identification of significant variables for each model using the forward sequential method. Then a simulation database was developed that consisted of 98928 intersections. The crash frequencies were predicted for these intersections using the models developed using the significant variables. In order to determine the manner in which the significant variables affect the output, the average number of crashes per intersection was determined for each value of the significant input variables and these were plotted. The graphs show if the significant variables have an increasing or decreasing effect on the frequency of crashes. Such models were developed for rear end, angle, turn (left and right turn crashes) and angle crashes. Table 5.17 summarizes the neural network models that were considered to perform well in each phase of the analysis, along with the accuracy of the models. PNN was found to perform better than MLP in all of the classification phases, whereas MLP and GRNN performed equally well in the prediction phase. Table 5.18 summarizes the effects of input variables on the prediction of the frequencies of different types of crashes. An increase in the number of through lanes on the major and minor roadway and the AADT on the major roadway tends to increase all types of crashes. Increase in the speed limit on the major roadway did not have a considerable effect on the variation of any of the crash types other than the rear end crashes. An increase in the channelized right turning lanes on the major roadway tends to increase the turn and sideswipe crashes. All crash types except for the sideswipe crashes increase with an increase in the protected left turning lanes on the minor road. Rear end and sideswipe crashes increase with an increase in the major LTP lanes, but all other crash types show a decreasing trend. Increase in the total number of left turning lanes increases total crashes at intersections. Thus the new technique was not only able to predict the frequencies of different types of crashes accurately, it was also able to identify the manner in which the geometric and traffic characteristics of the intersections influence the crash frequencies.

The next phase of the research was to classify crashes intro rear end, angle, turn (left and right turn) and sideswipe crash types. At first the MLP and PNN models were used to achieve this, but the performance of the models was not satisfactory. An innovative method called the Neural Network Trees was developed that classifies the crashes either into a category of rear end and sideswipe crashes or into a category of angle and turn crashes. The crashes are further classified by separate neural network models into their respective collision types. This has been shown graphically in Figure 6.1. The MLP and PNN models were used in each classification phase and the better model was chosen for identifying the significant variables, and also in the simulation phase for identifying the manner in which the input variables affect the classification. Figure 6.2 depicts the models that were considered to perform better and the accuracy attained by the models on a test dataset. The PNN performed better in classifying the rear end and sideswipe crashes and the MLP neural network performed better in the other two models. The accuracies obtained in these models were considerably better than the accuracies obtained in classifying crashes into the four collision types. Eleven variables were found to be significant in distinguishing the rear end and sideswipe crashes from angle and turn crashes, as shown in Table 6.11. These included the AADT and speed limits on the major and minor roadways, surface and light conditions at the time of the crash, number of through lanes on the major roadway, total left turning lanes, RTC lanes on both the major and minor roadways and LTP lanes on the minor roadway. The significant variables for the other two models have been listed in Tables 6.12 and 6.13. Upon using these models on the simulation datasets, the effect of the significant input variables on the classification of the crash types was known. For example, Figures 6.3 and 6.4 show that an increase in the AADT on the major and minor roadways considerably increases the chances of a crash being a rear end or a sideswipe crash. These trends have been plotted in Figures 6.3 to 6.24 and have been compared to other studies in order to verify the results. It was found that the trends obtained were comparable to the outputs of other studies, thereby verifying the validity of the Neural Network Trees.

Thus, this thesis shows the use of innovative neural network techniques in prediction and classification of crashes at signalized intersections. The neural network techniques have produced results that are comparable to other studies. These models can be used to accurately predict the crash types an intersection will be most prone to. If an intersection is found to have a high number of crashes, the intersection can be used in the model with possible improvements to check if the crash rate at the intersection decreases. Therefore, an optimum improvement plan for an intersection can be determined that can lower the crash rate. If an intersection is in its design phase, its characteristics can be used as an input to the models to determine the crash rate at the intersection. If it is found to be too high, the design can be altered to make the intersection safer.

Since the simulation phase of the analysis conducted for predicting the frequency of crashes estimated the frequency of crashes at a large combination of possible intersections, a program can be developed that takes an input of the traffic and geometric characteristics of an intersection from a user, refers to the simulation output to obtain the frequency of crashes at such intersections, and shows this output to the user. This eliminates the process of developing a neural network and training it to predict the frequency of crashes at different intersections.

The neural networks showed a satisfactory performance. The analysis shows that even if the neural network models are not able to perform well, they can be modified to obtain better results. This demonstrates the flexibility of the neural networks. On comparison of the MLP and PNN neural networks, both were found to perform better in different cases. But for classifying intersections into safe or unsafe with respect to different collision types, PNN always performed better. PNN was faster in training databases compared to MLP. PNN also demonstrated its advantages by not being trapped in local minima and has only one parameter that has to be varied in order to obtain optimum results. The only disadvantage found for PNN was that it takes a long time and consumes a lot of memory in simulating the results of a test database. Therefore large databases had to be split up into parts in order to make the process faster and less taxing on the computer. But on the whole, PNN can be considered as a better method for classification. The GRNN and MLP neural networks showed similar performances and hence can be considered equally efficient in predicting values.

Further studies can be carried out to extend the techniques demonstrated in this research. Models can be developed to classify the crash injury types using the Neural Network Trees and studying the effects of traffic, geometric and driver characteristics on the injury types. The results obtained can be compared to the results of other statistical models such as the nested-logit and ordered-probit models. Crash frequency prediction models can be developed to estimate the frequency of fatal and severe-injury crashes at signalized intersections. Other statistical models such as the Negative Binomial and Poisson models can be developed using the same dataset and the results can be compared to check the performance the neural network models. Additional parameters like the signal timing can also be used to further enhance the models.

REFERENCES

Abdel-Aty, M. "Analysis of driver injury severity levels at multiple locations using ordered probit Models." *Journal of Safety Research* Vol. 34 Iss. 5 pp. 597-603 2003

Abdel-Aty, M. A. and A. H. As-Saidi. Using GIS to Locate High Risk Driver Population. *Traffic Safety on Two Continents*, Malmo, Sweden. 2000

Abdel-Aty, M. A., Keller, J. and P. A. Brady. Analysis of the types of crashes at signalized intersections using complete crash data and tree based regression. *Presented in the 84th Annual Meeting of the Transportation Research Board*. Washington D. C. 2005

Abdelwahab, H. T. and M. A. Abdel-Aty "Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections." *Transportation Research Record 1746,* 2001

Abdelwahab, H. T. and M. A. Abdel-Aty "Investigating Driver Injury Severity in Traffic Accidents using Fuzzy ARTMAP." *Computer Aided Civil and Infrastructure Engineering* Vol. 17 Iss. 6 pp. 396-408 2002

Abdelwahab, H. T. and M. A. Abdel-Aty "Predicting Injury Severity Levels in Traffic Crashes: A Modeling Comparison." *Journal of Transportation Engineering* Vol. 130 Iss. 2 pp. 204-210 2004

Bureau of Transportation Statistics "State Transportation Statistics (STS) 2004." *http://www.bts.gov/publications/state_transportation_profiles/state_transportation_statistics_2004/* 2004

Chen, C. H. "Fuzzy Logic and Neural Network Handbook" McGraw-Hill,1996

Chin, H. C. and Quddus, M. A. "Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections." *Accident Analysis and Prevention* Vol. 35 Iss. 2 pp. 253-259 2003

Christodoulou, C. and Georgiopoulos, M. "Applications of Neural Networks in Electromagnetics" *Artech House*, Boston, 2001

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.

Dougherty, M. "A review of Neural Network applied to Transport." *Transportation Reseach C* Vol. 3 Iss. 4 pp. 247-260 1995

ESRI. “GIS in Transportation.” http://www.esri.com/industries/transport/

Greibe, P. "Accident Prediction Models for Urban Roads." *Accident Analysis and Prediction* Vol. 35 Iss. 2 pp. 273-285 2003

Kam, B. H. "A disaggregate approach to crash rate analysis." *Accident Analysis and Prevention* Vol. 35 Iss. 5 pp. 693-709 2002

Karlaftis, M. G. and I. Golias "Effects of road geometry and traffic volumes on rural roadway accident rates." *Accident Analysis and Prediction* Vol. 34 Iss. 3 pp. 357-365 2002

Keller, J. M. (2004). Analysis Of Type And Severity Of Traffic Crashes At Signalized Intersections Using Tree-Based Regression And Ordered Probit Models, University of Central Florida**:** 142.

Krull, K., A. Khattak and F. Council. Injury effects of rollovers and events sequence in single-vehicle crashes. *Presented in the 80th Annual Meeting of the Transportation Research Board*. 2000

Liu, P. and H.-G. Young. 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

Mistry, G., A. Graettinger, J. Lindly and S. Ullah. GIS Analysis and Display of Crash Data. *29th International Forum on Traffic Records & Highway Information Systems*. 2003

Mountain, L., B. Fawaz and D. Jarrett "Accident Prediction Models for roads with Minor junctions." *Accident Analysis and Prediction* Vol. 28 Iss. 6 pp. 695-707 1996

Mountain, L., M. Maher and B. Fawaz "The influence of Trend on estimates of accidents at junctions." *Accident Analysis and Prevention* Vol. 30 Iss. 5 pp. 641-649 1998

Mussone, L., A. Ferrari and M. Oneta "An analysis of urban collisions using an artificial intelligence model." *Accident Analysis and Prediction* Vol. 31 Iss. 6 pp. 705-718 1999

Nigrin, A., *Neural Networks for Pattern Recognition, Cambridge*, MA: The MIT Press, 1993.

Ng, K.-s., W.-t. Hung and W.-g. Wong "An algorithm for assessing the risk of traffic accident." *Journal of Safety Research* Vol. 33 Iss. 3 pp. 387-410 2002

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.

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.

Pawlovich, M. D. A Method of Examining Dependence of Crashes on Demographic and Socioeconomic Data. *Mid-American Transportation Student Paper Contest*. http://www.ctre.iastate.edu/Research/gis-alas/gisalas1.htm 1998

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

Quddus, M. A., R. B. Noland and H. C. Chin "An analysis of motorcycle injury and vehicle damage severity using ordered probit models." *Journal of Safety Research* Vol. 33 Iss. 4 pp. 445-462 2002

Riedmiller, M. and Braun, H. "A Direct Adaptive Method for Faster Backpropagation Learning: The rprop Algorithm." *Proceedings of the IEEE International Conference on Neural Networks*, 1993

Rodriguez, F. and Sayed, T. "Accident Prediction Models for Urban Unsignalized Intersections in British Columbia." *Transportation Research Record* ,Vol. 1692, pp. 30-38. 1999

Sayed, T. and W. Abdelwahab "Comparison of Fuzzy and Neural Classifiers for Road Accidents Analysis." *Journal of Computing in Civil Engineering* Vol. 12 Iss. 1 pp. 42-47 1998

Shankar, V. and F. Mannering "An exploratory Multinomial Logit Analysis of Single-Vehicle Motorcycle Accident Severity." *Journal of Safety Research* Vol. 27 Iss. 3 pp. 183-194 1996

Shankar, V., F. Mannering and W. Barfield "Statistical Analysis of Accident Severity on Rural Freeways." *Accident Analysis and Prediction* Vol. 28 Iss. 3 pp. 391-401 1996

Shankar, V., J. Milton and F. Mannering "Modeling accident frequencies as zero altered processes: An empirical enquiry." *Accident Analysis and Prediction* Vol. 29 Iss. 6 pp. 829-837 1997

Turner, S. and Nicholson, A. "Intersection Accident Estimation: The Role of Intersection Location and Non-Collision Flows." *Accident Analysis and Prevention* Vol. 30 No. 4 pp. 505–517 1998. 1998

Vogt, A. and Bared, J. Accident models for two-lane rural segments and intersections. *Transportation Research Record* No. 1635 pp. 18-29. 1998

Wang, Y., H. Ieda and F. Mannering "Estimating Rear-End accident probabilities at Signalized Intersections: Occurrence Mechanism Approach." *Journal of Transportation Engineering* Vol. 129 Iss. 4 pp. 377-384 2003

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.