ANALYSES OF CRASH OCCURENCE AND INURY SEVERITIES ON MULTI LANE HIGHWAYS USING MACHINE LEARNING ALGORITHMS

by

ABHISHEK DAS
M.S. (Transportation), University of Central Florida, Orlando, 2009
B. Tech. (Civil), Indian Institute of Technology, Delhi, 2005

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

Fall Term
2009

Major Professor
MOHAMED A. ABDEL-ATY, Ph.D., P.E.

ABSTRACT

Reduction of crash occurrence on the various roadway locations (mid-block segments; signalized intersections; un-signalized intersections) and the mitigation of injury severity in the event of a crash are the major concerns of transportation safety engineers. Multi lane arterial roadways (excluding freeways and expressways) account for forty-three percent of fatal crashes in the state of Florida. Significant contributing causes fall under the broad categories of aggressive driver behavior; adverse weather and environmental conditions; and roadway geometric and traffic factors. The objective of this research was the implementation of innovative, state-of-the-art analytical methods to identify the contributing factors for crashes and injury severity. Advances in computational methods render the use of modern statistical and machine learning algorithms. Even though most of the contributing factors are known a-priori, advanced methods unearth changing trends. Heuristic evolutionary processes such as genetic programming; sophisticated data mining methods like conditional inference tree; and mathematical treatments in the form of sensitivity analyses outline the major contributions in this research. Application of traditional statistical methods like simultaneous ordered probit models, identification and resolution of crash data problems are also key aspects of this study. In order to eliminate the use of unrealistic uniform intersection influence radius of 250 ft, heuristic rules were developed for assigning crashes to roadway segments, signalized intersection and access points using parameters, such as 'site location', 'traffic control' and node information. Use of Conditional Inference Forest instead of Classification and Regression Tree to identify variables of significance for injury severity analysis removed the bias towards the selection of continuous variable or variables with

large number of categories. For the injury severity analysis of crashes on highways, the corridors were clustered into four optimum groups. The optimum number of clusters was found using Partitioning around Medoids algorithm. Concepts of evolutionary biology like crossover and mutation were implemented to develop models for classification and regression analyses based on the highest hit rate and minimum error rate, respectively. Low crossover rate and higher mutation reduces the chances of genetic drift and brings in novelty to the model development process. Annual daily traffic; friction coefficient of pavements; on-street parking; curbed medians; surface and shoulder widths; alcohol / drug usage are some of the significant factors that played a role in both crash occurrence and injury severities. Relative sensitivity analyses were used to identify the effect of continuous variables on the variation of crash counts. This study improved the understanding of the significant factors that could play an important role in designing better safety countermeasures on multi lane highways, and hence enhance their safety by reducing the frequency of crashes and severity of injuries. Educating young people about the abuses of alcohol and drugs specifically at high schools and colleges could potentially lead to lower driver aggression. Removal of on-street parking from high speed arterials unilaterally could result in likely drop in the number of crashes. Widening of shoulders could give greater maneuvering space for the drivers. Improving pavement conditions for better friction coefficient will lead to improved crash recovery. Addition of lanes to alleviate problems arising out of increased ADT and restriction of trucks to the slower right lanes on the highways would not only reduce the crash occurrences but also resulted in lower injury severity levels.

ACKNOWLEDGMENTS

First, I would like to express my sincerest gratitude to my advisor Dr. Mohamed A. Abdel-Aty. His valuable guidance, constant support and constructive criticism helped shape this dissertation. It has been a privilege working with him. I am forever indebted to the love of my parents, grandparents and all the elders in my family with whose blessings, this goal in my life has been achieved. I would also like to acknowledge the support of my esteemed committee members, Dr's Essam Radwan, Haitham Al-Deek, Ni-Bin Chang and Nizam Uddin. I would like to extend special thanks to Dr's Anurag Pande and Ravi Chandra for their continuous support in all spheres of my life here in Orlando especially the insightful academic discussions late in the night. I would like to thank Dr. Ammarin Makkeasorn for his critical help in genetic programming research. Mr. Patrick Kerr has been the go-to guy for any sort of help. My stay in the beautiful state of Florida has been delightful because of friends like Piyush, Vinayak, Nezamuddin, Shankar, Albinder, Amit, Swapnil, Noor, Vikash, Cristina, Ryan, Rami, John, Chris, Kirolos, Parveen, Premchand, Sandesh, Colin, Sheetal, Ali, Ana, Olivia, Himansu, Rashmi, Sharad, Ashley, Aishwarya, Dipika, Bankim, Pallavi, Nisha, Pankaj, Mona, Akanksha and Divi. As I grow "old" at UCF, I have found new friendship in Zaidi, Hany, Mohamed, Turen, Sid, Jeremy and Taylor. As I continue this journey with the Grace of God, I know I have friends who would always look out for me. Swati, Kamalika, Vibhor, Ankur, Ravi, Iti, Sumit. Thank you all for being a part of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ACRONYMS/ABBREVIATIONS

AADT	Annual Average Daily Traffic
CAR	Crash Analysis and Reporting
DDHV	Directional Design Hourly Volume
DMV	Department of Motor Vehicle
FDOT	Florida Department of Transportation
DUI	Driving Under the Influence
FHWA	Federal Highway Administration
GP	Genetic Programming
RCI	Roadway Characteristics Index

CHAPTER 1. INTRODUCTION

1.1 Research Motivation

Improving the safety of arterials, by reducing fatalities and injuries, is one of the objectives of transportation safety researchers and engineers. Florida is one of the states with high number and rates of fatalities in the United States of America. In 2008, 2,978 fatalities occurred on roadways in Florida (NHTSA, 2008). Though this indicates a 7.3% decrease in the number of fatalities yet the number is alarmingly high. The state ranks third in the number of fatalities among all other states in the country. Among the different road types, principal and minor arterials account for the 57% of the total crashes in Florida (NHTSA, 2005). The proportion and the sheer number of fatal crashes on principal arterials (excluding freeways and expressways) in Florida were one of the highest in the nation in 2005. In particular, speeding-related fatalities on arterials with speed limits of 40 mph and above account for more than 72% of total speeding-related fatalities.

The statistics presented above indicate a need to improve the safety of Florida arterials, especially the high-speed, multi-lane arterials, by reducing fatalities and severe injuries. Fatal or severe crashes on arterials occur due to a combination of multiple factors. Hence to reduce fatalities and severe injuries on the arterials generally two approaches can be adopted. One is to look at the intersections and the roadway segments (excluding the intersections) separately and the other is to treat them together as a corridor. The former idea involves the use of the definition of influence distance of intersections to separate the intersection related crashes from the segment related crashes. Das et al. (2008), showed by the method of simultaneous estimation that

if the influence distance varied the crash characteristics associated with severe injuries also varies. This is due to the fact that the farther we move away from the center of an intersection, more crashes related to the connecting segment comes into play. Wang et al. (2008) used frequency modeling for crashes with fixed as well as varying influence distance and found different set of significant factors. These studies show that the concept of using influence distance for assigning crashes to the roadway elements could be erroneous. However it is believed that analyzing the crashes along a corridor will instead help us identify the significant factors more realistically and understand the interaction among the design elements and traffic characteristics better.

1.2 Research Objectives

The main objectives of this research are the following:

1. Critical review of the work done on safety analysis, specifically arterial or corridor safety studies carried out in the various states of the country.
2. Evaluate the futility of the 250 ft. influence radius of signalized intersections in the state of Florida
3. Enhancing the crash database with data from the roadway characteristics inventory and developing heuristic rules for assigning crashes to various roadway elements.
4. Analytical methods like data mining and machine learning algorithms for classification of injury severity models, identifying variables of importance and crash frequency modeling for broader understanding of the safety situation on the multi lane highways.

Mathematical treatment of crash count models for better evaluation of significant variables.

5. Design and probable policy recommendations to ameliorate safety on highways based on the research findings.

1.3 Organization

The literature review, following the introduction, includes arterial safety studies, crash prediction models, and corridor studies. In the third chapter the initial investigation of simultaneous ordered probit models is discussed where the topic of the fixed influence distance is taken up. The chapter following that describes the heuristic rules developed to assign crashes to the various roadway elements. The next chapter deals with the various databases available with Florida Department of Transportation (FDOT) namely: 1) Crash Analysis and Reporting (CAR) System; and 2) Roadway Characteristic Inventory (RCI). The clustering of corridors is also described in the chapter. The sixth chapter uses the innovative conditional inference trees and forests to understand the injury severity conditions and better the understanding of the significant factors. The seventh chapter introduces the concept of genetic programming for transportation safety study. It is introduced as an umbrella methodology for both classification and regression purpose. For the present work the author has used it for injury classification and crash count modeling. The eighth chapter is a graphical demonstration of the change in crash frequency as other continuous variables change. Relative sensitivity of the crash frequency model response towards the continuous input variables is also discussed.

CHAPTER 2. LITERATURE REVIEW

2.1 Previous Studies

This chapter summarizes some of the relevant studies. We review some of the past studies that are relevant to high speed multi lane arterials in general, and those looking into severe crashes in particular. The literature review is divided into 3 sections. Section 1 deals with the work done in the past related to arterials. The studies were not always focused on safety issues. Some of them e.g. dealt with median design guidelines. But they always required some crash studies to be undertaken. The section looks at selected and noteworthy work done from the late 1960's to mid 1990's. There were studies related to arterials, but not exclusively dealing with the issue of corridor safety. At the end of the section the overall results will be summarized.

Section 2 deals with selected and important work done on crash prediction models on roadway segments. The significant factors in the models will be discussed. In addition to it the research done to investigate contributing factors to severe crashes will also be discussed. Since we are interested in reducing fatalities and severe injury related crashes this discussion is essential. It will also address another important issue of how varied each researcher's point of view is on the definition of a roadway segment. The criteria to define a roadway segment differ. These discrepancies in the working definition of a segment can lead to confusing inferences. Again at the end of the section there will be a discussion on the results from the work and how vast is the problem of segment definition.

Section 3 discusses some chosen papers dealing with the issue of corridor safety especially from the point of view of access management. It will also take up recent research work on signalized intersections that has shown that there is a spatial correlation among them and they influence each other in many aspects. At the end of this section there will be a final discussion as to why it is important to address the safety aspect of the corridor as whole, both roadway segments and intersections included.

2.1.1 Section 1: Arterial Safety Research

Mulinazzi and Michael (1967) developed crash prediction models for urban arterials and Walton et al. (1978) built up a regression equation to predict crashes at two-way left turn (TWLT) median lanes in Texas. Average daily traffic (ADT), number of traffic signals per mile, were the common significant factors in both studies. The former also found number of high volume intersections per mile to be important. The latter specified that number of driveways per mile and area population were contributing factors too.

In his Virginia study for design guidelines for raised and transversable medians, Parker (1983) and also in an update in 1990, found that number of traffic signals per mile, number of driveways per mile, area population and ADT had a significant effect on crashes for raised median sections. In an update to his previous work Parker (1990) found the same results.

Squires and Parsonson (1989) in their study of crash comparison of raised median sections and TWLT median lanes in Georgia established that ADT and number of traffic signals per mile were important factors.

Bowman et al. (1995) found land use, median width, and number of driveways per mile, posted speed limit and crash reporting threshold in dollars to be significant factors in crash prediction models for urban or suburban arterials' roadway sections with homogeneity in median type. Though the study included arterial sections with signalized intersections, they did not find number of signalized intersections along the arterial section to be significant.

Mountain et al. (1996) developed crash prediction models for road network in seven counties of the U. K. Total two-way annual segment volume, length of the segment and number of minor intersections within the segment were significant factors.

From the results in the research papers discussed above it can be concluded that some design elements and certain traffic elements play a major role in crashes occurring along the arterials. The design elements significant in most of the work are number of traffic signals per mile, number of driveways per mile or driveway density, median width and length of the segment. The traffic characteristics that are important are ADT, speed limit and annual volume. In addition to the above mentioned factors some study also found land use and population of the area to be significant.

2.1.2 Section 2: Crash Prediction Models

Kim et al. (1995) investigated the predictors for crash and injury severity on roadways in Hawaii. Alcohol abuse and seat belt disregard were found to be important factors contributing to the cause of crashes and also result in more severe crashes.

O'Donnell and Connor (1996) in their work on predicting severity of motor vehicle crash injury had non-use of seat belt, head on collisions, and alcohol as significant factors. Female drivers were found to be more involved in severe crashes than male drivers.

Bonneson and McCoy (1997) investigated roadway segments for their study of the effect of median treatments on urban arterial safety. They defined the roadway segment as the section between two consecutive signalized intersections. In addition to that, for their work, they chose the segments with a minimum number of vehicles per day, speed limit, number of through lanes and length. ADT, segment length, driveway density, unsignalized public street approach density and land use were significant factors in their crash model which did not include crashes at intersections.

Milton and Mannering (1998) found section length, AADT, percentage of AADT occurring during peak hour, percentage of trucks, speed limit, number of lanes, shoulder width, horizontal curves, and tangent length as the significant factors contributing to crash frequencies on highway sections that excluded signalized intersections. Section or segment delimiters were number of

lanes, roadway width, shoulder width, state route number, road type, urban or rural location identifiers, speed, AADT, peak hour factors, and vertical and horizontal curve characteristics.

Chang and Mannering (1999) analyzed injury severity for truck- and non-truck involved crashes. For non-truck involved crashes driver ejection, driver restrained systems, alcohol impairment are responsible for fatalities and more severe injuries. Truck involved crashes are more serious.

Sawalha et al. (2001) examined safety of urban arterial roadway segments, which was defined as the part of the arterial between consecutive signalized intersections. Traffic volume, segment length, unsignalized intersection density, type of median, number of crosswalks, number of lanes and land use were important factors in the model developed by them.

Hanley et al. (2000) analyzed crash reduction factors on California State highways. The segments were chosen based on AADT and it is not very clear from the work whether intersection were included or not. Increases in shoulder width and curve correction with improved radius were found to be significant.

Zhang et al. (2000) in their study of the factors affecting severity of motor vehicles crashes in Ontario established that age, disobeying of traffic signs, non-use of seat belts, intersections without traffic control, speed, head on and turning collisions, overtaking maneuvers increased the risk of a fatal or severe injury crash. Alcohol and medical/physical condition of elderly drivers significantly increased the risk of fatalities.

Bedard et al. (2002) in their work on causes related to driver fatalities on roadways found that age, alcohol, point of impact, seat belt non-use and speed as significant factors. Older male drivers were more prone to fatal crashes than older female drivers.

Kockelman and Kweon (2002) examined driver injury severity and found that increased driver age, vehicle age, alcohol use, head on or rollover collision, numbers of vehicles involved were associated with more severe injuries. Female drivers and night time driving were related to increase in injury severity of two-vehicle crashes.

Martin (2002) sought to find the relationship between crash rate and traffic flow on French interurban motorways. Hourly traffic, day of the week and number of lanes were the contributing factors. Night time crashes and crashes occurring under light traffic conditions are found to result in more severe injuries. The roadway sections or segments were homogenous in terms of traffic between two motorway entry points. It is not apparent as to whether the entry points are signalized or unsignalized intersections.

Greibe (2003) built up crash prediction models for urban roads in Denmark where ADT, land use and speed limit were essential factors. Segments and intersections were treated independently. But intersections with low flow rate were included in the segments. And it is not obvious whether intersections are signalized or unsignalized.

Abdel-Aty (2003) analyzed driver injury safety levels at multiple locations and found driver's age, gender, seat belt use, point of impact, speed, vehicle type, weather condition and area type as major factors. His study also investigated segment and intersection crashes disjointedly. In his work he found seat belt disregard, age, gender, speed, point of impact and alcohol consumption to be important factors contributing to severe injury related crashes. Crashes occurring on curved segments had higher probability of resulting in severe injuries. Abdel-Aty and Abdelwahab (2004) also found similar results for injury severity levels in traffic crashes. Female drivers were more probable to be in a severe injury crash than male drivers. Older people were more likely to be involved in a severe injury crash than younger drivers.

Hiselius (2004) in his study of Swedish rural roads investigated roadway segments without intersections. His segment criteria were traffic flow, speed limit and road width.

In Illinois county-level data study by Noland and Oh (2004) roadway section categorization was based on location (urban or rural), divided or undivided cross section, number of lanes, average median width, average shoulder width, and horizontal and vertical curvature. Again it is not understandable as to whether intersections were included or not. Increase in number of lanes and increment in lane width was found to be associated with increase in fatalities and crashes. Increase in shoulder width resulted in fewer crashes.

Miaou and Song (2005) ranked sites for engineering safety improvements. They analyzed segments and intersections separately. The segments they considered had low traffic volume.

The design elements that were found to be most important in the above mentioned research works on roadway segments are segment length, driveway density, number of lanes, shoulder width. Other important design elements were road width, number of crosswalks, horizontal and vertical curves. The traffic elements of significance were ADT and speed limit. In addition to these some work also showed that standard deviation of traffic flow, percentage of different type of vehicle were also significant. Land use was also found noteworthy in some of the work.

As far as the factors contributing to fatal and severe injury crashes are concerned, it is observed that more driver related characteristics are responsible. Design and traffic parameters are not ruled out, but their contribution to specifically those crashes is less. Non-use of seat belt, older driver age, alcohol use, and speeding are found to be significant in most research work related to severity of crashes. Head-on and angle collisions result in more fatalities and severe injuries than any other type of crashes. Some research work show that crashes occurring at night and under light traffic conditions are more severe. Severity of crashes is also dependent on the point of impact of crash, especially the ones hitting from the side. Intersections without control witness more severe crashes.

Different authors have their own view point as to how to define the segment. A roadway is typically the section of the roadway between two consecutive signalized intersections. In some segment studies unsignalized intersections have been included. Some work mentioned the inclusion of low volume intersections but do not clearly specify whether those are signalized or unsignalized. The criteria to choose the segments are characteristically speed limit, number of

lanes, ADT, shoulder width and roadway width. Some researchers have in addition to the above criteria had vertical and horizontal characteristics, road type, urban or rural location as the segment defining criteria. So it can be clearly seen that the vast literature has a confusing definition of segments in crash modeling.

2.1.3 Section 3: Corridor Safety

Jernigan (1999) compared the various corridor safety improvement efforts by Pennsylvania, California and Virginia. He also provides model strategy for the development of these programs.

Levinson (1999) and Papayannoulis et al. (1999) developed a model for safety of corridors based on traffic volumes of corridors and access roads and access density. Increase in crashes was related to the increase in access density.

Brown and Tarko (1999) also found density of access points, proportion of signalized access points, outside shoulder, TWLT lanes and presence of medians with no openings between signals as significant factors for safety on urban arterials. They investigated the corridor as a whole.

Abdel-Aty and Radwan (2000) modeled traffic crash occurrence and involvement along SR-50 of Florida and found AADT, degree of horizontal curvature, lane shoulder, median width, urban or rural location, section length to be significant factors. Their section definition included intersections.

Drummond et al. (2002) used simulation approach to predict safety and operational impacts of increased traffic signal density along entire corridors. The major factors were main-line delay, speed limit and stops.

Rees (2003) in his corridor management studies investigated full corridors. His study also focused on applying access management treatments along corridors.

A very recent work Abdel-Aty and Wang (2006) have shown in their modeling work of signalized intersections that there is a spatial correlation between crash patterns of successive signalized intersections.

The work on the spatial correlation of crash patterns of successive signalized intersections show that there is a need to look at the sequence of signalized intersections along a corridor rather than treating each intersection as an isolated entity. Intersections are also access points. The access management studies for corridor safety illustrate that the roadway segments and intersections are integral part of the corridor. Therefore we should improve the corridor as a whole, both roadway segments and intersections, in order to achieve significant reduction in fatal and severe crashes.

2.2 Improvement strategies implemented by different States and the level of success

Corridor Safety Improvement Programs (CSIPs) were initiated on the fact that crashes are likely to occur along joined segments of highways. Some of these joined segments of highways or

corridors as they are commonly known as, have a relatively high crash rate. To reduce the fatality and injury rate on these corridors it may not be sufficient that only spot improvements are done (Jernigan, 1997). Therefore multidisciplinary cooperation is necessary to bring about major safety and traffic changes on these corridors. This report summarizes the work done on improving safety on high-speed multi-lane arterials by different states in the U.S. The first such improvement task was carried out by Pennsylvania Department of Transportation (PennDOT) on the U.S. Route 322 following a series of fatal crashes. The success of the program led to similar work throughout the state and the Federal Highway Administration (FHWA) encouraged for similar projects in other states. In 1991 the FHWA issued guidelines for developing a CSIP. The essence of the guidelines was to establish a leadership based program to oversee the work of improving safety along hazardous corridors. The guidelines had provision for involving various agencies, creating a multidisciplinary team, selecting corridors, creating an action plan, implement the recommendations and evaluate the effectiveness. The states who initiated corridor safety improvement program on their roads, more or less followed the FHWA guidelines.

The following discussion focuses on the work done in 10 different states across the nation. They have been selected for discussion here as substantial information could be gathered from various sources about their work. The states are: Pennsylvania, Washington, Virginia, California, Oregon, North Carolina, Kentucky, Arizona, Ohio and Florida.

2.2.1 Pennsylvania

The pilot project in the safety improvement program of Pennsylvania was the U.S. Route 322 in Delaware County which is a high-volume and high-speed highway. The route was chosen on the behest of the then Pennsylvania Governor Robert P. Casey following a crash on the stated corridor that resulted in multiple fatalities in 1988. The plan was successfully implemented in a period of 6 months. The typical corridor safety problems were identified along the designated corridor. Among the various countermeasures, highway design improvements, educational, media programs and enforcement drives to improve driver performance, and commercial truck safety inspections were the most important (Zogby et al., 1991). Emergency medical services were also improved along the designated corridor. The corridor had 40% less number of crashes in the 3 years following the improvements (Jernigan, 1999). Later on 55 corridors, totaling 880 miles of highway, were earmarked for the safety initiative. The sections not only accounted for 7% of the total fatalities but also had the maximum concentrations of severe crashes per mile. Three Pennsylvania agencies: PennDOT, Department of Health and the state and local police work in synergy. The improvements were applied over the entire length of the section and thus improved the overall safety along the length (Zogby et al., 1991). In 2002, Pennsylvania House Bill 2410 came into effect which allowed for fines to be doubled on the designated safety corridors. The safety effect of the bill has not been yet established.

2.2.2 Washington

Soon after the success of the Pennsylvania initiative the FHWA encouraged other states to follow similar programs. Washington was one of the first states to start such a statewide program. The program which started in 1992 is still on. Several projects have been successfully completed and others are on the way. The Washington State Corridor Safety Program is a joint program between Washington Traffic Safety Commission and the Washington Department of Transportation (WSDOT) and the goal is to reduce "fatal and disabling" crashes along the designated corridors. The corridors selected have to have a statistical evidence of crash problem and there must be local support for the undertaken project (Washington Traffic Safety Commission [WTSC], 2006). Some of the corridors like SR 14 which was one of the designated corridors had safety concerns like speeding, over the centerline crashes, driving under influence (DUI) and operating defective equipment (National Highway Traffic Safety Administration[NHTSA], 2004). The action plan primarily consisted of 3E's: enforcement, engineering and education (NHTSA, 1997). Till now 21 projects have been completed and 9 projects are still on. The number of crashes along 24 designated corridors has reduced by 6%; reduction in traffic injuries is by 11%; alcohol related crashes have gone down by 20%; most importantly fatality-disabling crashes has dipped by 34%. The fundamentals elements of the program are education, enforcement and engineering solutions to improve safety on the designated corridors (WTSC, 2006).

2.2.3 Virginia

Virginia also became active in the field of corridor safety in 1992 after the success of the Pennsylvania program. Virginia's program differed considerably from FHWA guidelines (Jernigan, 1997). The Virginia Department of Transportation (VDOT) and Virginia Department of Motor Vehicles (DMV) co-sponsored 2 pilot projects; one urban and one rural. Apart from safety the authorities wanted to find out the possible differences in the ability of the program to be effective (Jernigan 1999). The urban corridor was a 5.5 mile segment on U.S. Route 144 while the rural corridor was a 19 mile stretch on U.S. Route 24. The significant safety problems on the corridors were driver's inattention, speeding, defective vehicles, DUI, rear-end crashes, angle crashes, fixed object crashes (run off road), and sideswipes. The suggested improvements to check the safety issues were: lowering speed limit, enforcement, improving signage and sight distance, warning for DUI checkpoints along the corridor, installations of traffic signals, changes in approach to intersections, installing of guardrails, and addition of paved shoulder (Jernigan, 1997). After the improvements have been implemented there were 5% less number of injury crashes on the rural corridor and the injuries decreased by more than 10%. The situation was a bit different for the urban corridor. Though the injury crashes decreased by 10%, the injuries went up by 5%. Virginia has also developed a methodology for determining safety corridors for investigation and improvement (Fontaine and Read, 2006). The designated corridors should definitely have above average crash rate and densities (Virginia's Surface Transportation Safety Executive Committee, 2006).

2.2.4 California

In 1992 California also started a corridor safety program which was led by the California Highway Patrol and not by the state's DOT. A corridor of 21 mile length on State Route 1 in the Ventura County was chosen. This was done in collaboration with Caltrans and California's Office of Traffic Safety. The recommendation for safety improvement included enforcement, engineering solution, education, public information and emergency response. The number of injury crashes and injuries dropped by 25% on the corridor (Jernigan, 1999). The crash rate decreased by 11% to 37% within a 3 year analysis period and injury crash rate decreased by 13% to 47% (Fontaine and Read, 2006). SR 41 and SR 46 were designated safety corridors after a severe collision resulted in multiple fatalities in 1995. The safety problems identified were unsafe turning, unsafe speed, right of way violations, DUI and driver not at fault. The countermeasures implemented fell into the categories of 4E: education, enforcement, engineering solution and emergency response. The efforts paid off well. The fatalities were reduced by 10% and injury crashes decreased by 32% (Bichler-Robertson et al., 2001). In the recent past State Highways 25, 49, 65 have been designated as safety corridors. For SR 25 the goal is to reduce the fatal and injury crashes (California Department of Transportation, 2006).

2.2.5 Oregon

In 1993 Oregon jumped into the scene of corridor safety improvement programs. These were implemented along Oregon Route 34 and 22. The typical safety concerns on the corridor were speeding, variation in speed and access related crashes. Increased level of enforcements, dividing

the highway and limiting the number of access points, provision of acceleration and deceleration lanes at major access points, limited use of traffic signal and decreasing speed limit were some of the recommendations for improving safety. The program was a success as far as the phase 1 of the project was concerned (Hunter-Zaworski and Price, 1998). The safety corridors had less fatalities and crashes. In 2001, doubling fines were effective in the safety corridors of Oregon. An important conclusion that came out of that was drivers have a higher perception of accident risks, traffic citations and fines in work zones and school areas than safety corridors (Jones et al., 2002). For the new safety corridors' designation the following three criteria must be met: 1) the three-year average of the fatality and injury crash rate must be greater than or equal to 110% of the three-year statewide average for similar type of roadways; 2) if the state or the local law enforcement agencies commit to make a certain corridor "patrol priority"; 3) the designated team concur that the length is manageable from an enforcement and education point of view (Oregon Department of Transportation [ODOT], 2006). Oregon Routes 62, 22, 34, 11, 18, 99E, 140 and U.S. Routes 101, 199, 20, 26, 730 are the routes where the safety corridors in the state of Oregon currently located (ODOT, 2007).

2.2.6 North Carolina

In 1998 the highway safety program of North Carolina came into being in 21 counties across the state. Fatal truck related crashes were the major safety problem. As a result there were increased roadside inspections, more number of citations for commercial driving license (CDL) violations. Within a year of its implementation there were a 4.6% reduction in the crashes involving

commercial motor vehicles (CMVs) in the marked counties and 5.2% reduction in crashes involving CMVs in counties that had not being targeted. There was a decrease of fatalities by 17.7% from crashes involving CMVs in the targeted counties, where as there was an increase in fatalities by 7.6% for crashes involving CMVs in the non-targeted counties (Hughes, 1999). The overall crash rate however did not change substantially.

2.2.7 Kentucky

In 1997 the Kentucky Transportation Cabinet started the Safety Corridor Program in an attempt to reduce the number of crashes and the number of injuries and fatalities on the state highways. A methodology for selecting high crash corridors has been developed and also a crash analysis technique has been proposed (Green and Agent, 2002). The US Route 31W was the designated corridor. The rural section of the corridor had a higher percentage of the fatal/injury crashes at intersection resulting from angle crashes. There was also a high percentage of run off the road crashes in the rural section, while the urban section had a higher percentage of rear-end crashes and the urban section had more crashes on straight sections. From noon to 6 pm there was reported to be a high number of crashes. Business and industrial districts had a higher percentage of the crashes. Failure to yield, following too closely, driver's inattention were also major contributing factors for fatal/injury crashes (Green and Agent, 2002). The focus has been on enforcement and education to alleviate the safety problems.

2.2.8 Arizona and Ohio

A pilot study was conducted by Arizona Department of Transportation (ADOT) in 1995 to see how the corridor safety improvement program takes shape. It was concurred that the tools demonstrated for the pilot study could lead to progress in safety improvement identification and they could be used by agencies other than ADOT (Breyer and Joshua, 1999).

In 2005 Ohio's Highway Corridor Safety Program got started. Seven highways were identified: SR 37, 46, 49, 50, 60, 73 and 193 (Governor's Task Force on Highway Safety, 2005). The governor's task force on highway safety has issued a handbook of guidelines and procedures which includes process to select a safety corridor and also toolbox for safety study and countermeasure.

2.2.9 Florida

The goal of the project set up in 1992 by Florida's Safety Management System was to establish Corridor/Community Traffic Safety Program (C/CTSP) in the each of the 20 high crash counties across the state by 1996 to reduce the number of fatalities and injuries. The concept was pilot-tested in Lakeland, Florida in collaboration with Florida Department of Transportation (FDOT). The project was a success and a state-wide C/CTSP coalition has been formed (NHTSA, 1996). The chosen corridor was Florida Avenue. Speeding, DUI, no-use of seat belt were some of the safety concerns on the corridor and improvements were suggested accordingly. There was a reduction in number of crashes and injuries during the analysis period (Dummeldinger et al.,

1994). In a recent research work on the safety of six lanes divided highways it was recommended that reduction in horizontal curves, increase of median and shoulder width can reduce the rate of severe and fatal crashes (Petritsch et al., 2007). 94% of the fatal crashes are caused by human factors. 4E's is the recommended course of action to reduce sever/fatal crashes (Spainhour et al., 2005).

2.2.10 Overview of Typical Safety Issues on Corridors

The safety issues that the corridors experience can be broadly divided into 2 categories. One is the roadway design deficiencies and the other is drivers' performance failures. Roadway design deficiencies include too many access points, higher number of traffic signals than is actually required, inadequate shoulder, absence of or inadequate length of acceleration/deceleration lanes among others. Drivers' performance failures include speeding, DUI, CDL violations, over the centerline crashes, operating defective vehicles, right of way violations, and no-use of safety belts among others. The most common type of crashes observed were angle, rear-end, and fixed object (runoff road) crashes. Many safety corridors also had a high percentage of truck involved fatal/injury crashes.

The safety improvements on corridors under study are based on the observed safety issues. The typical implementation has been that of the 4E's: education, enforcement, engineering solutions and emergency response. Education and media information has helped to make the community aware of the hazardous corridors and urging people to proactively help in improving safety on

the roads. Enforcement activities in many states include increased patrolling, doubling fines on the designated corridors, increased number of citations for violations of traffic rules, booking drivers for DUI and increased roadside inspection of commercial vehicles. Changes in roadway design on section of the corridors, reducing or increasing traffic signals, access management, adding paved shoulders, modifying acceleration/deceleration lanes are among the recommended engineering changes required to alleviate safety. Improving emergency response to better the probability of survival for crash victims has been a top concern for state agencies. Table 2-1 provides a comparison of the work done in various states and a perceived success measure.

Table 2-1 Work done in various states on corridor improvement

	Initial Initiatives (Yes or No)	**Success Measure of Initial Initiatives**	**New Initiatives / Projects**	**Success Measure of New Initiatives**
Pennsylvania	Yes	High	Doubling fines	No data
Washington	Yes	High	New projects	No data yet
Virginia	Yes	Relatively good	New projects	No data yet
California	Yes	High	Doubling fines	High
Oregon	Yes	Relatively good	Doubling fines	No data yet
North Carolina	Yes	Relatively good	-	-
Kentucky	Yes	-	-	-
Arizona	Yes	Relatively good	-	-
Ohio	Yes	-	-	-
Florida	Yes	Relatively good	New projects	-

The ‘initial initiative’ column indicates whether the state had initiated any projects or enforcements to improve safety on the problematic corridors. The following column reflects on how successful the initiatives have been in due course of time. The ‘new initiatives/ projects’ column shows what type of initiatives the states are having for the future in terms of project implementation or policy changes. The column next to it again indicates how successful the new measures have been.

CHAPTER 3. URBAN ARTERIAL CRASH CHARACETRISTICS RELATED WITH PROXIMITY TO INTERSECTIONS AND INJURY SEVERITY

3.1 Introduction

As mentioned in the opening paragraph of the document that in spite of the lower prevailing speeds, compared to freeways/expressways, arterials experience a significant proportion of severe/fatal crashes. For example, arterials account for sites of 57% fatal crashes in Florida (NHTSA, 2005). Safety on an arterial corridor may be affected by crash patterns on two seemingly distinct roadway elements; intersections and the segments between the intersections. A study by Abdel-Aty and Wang (2006) demonstrated spatial correlation between crash patterns belonging to successive signalized intersections on an urban arterial. It indicates the need to look at sequence of signalized intersections along a corridor rather than analyzing each intersection as an isolated entity. For such an approach crashes on arterial segment(s) joining consecutive intersections would also be critical part of the analysis. There is a potential for achieving better understanding of crash patterns on arterials if the corridors are studied as a whole instead of as disjointed parts (i.e., intersections and segments separately).

An important issue to be addressed for understanding corridor safety as a whole is the difference between the intersection and segment crash patterns, especially as it relates to injury severity. There are significant variations in the injury severity patterns that may be partially explained by the separation of crash location from intersections. For example, Abdel-Aty et al. (2006) found

that the prevailing types of fatal or severe crashes at intersections are mostly angle and left-turn crashes while those on roadway segments farther from intersection are mostly fixed object collisions. Hence, if one observes crashes only at the physical area of intersections; crashes would involve higher proportion of angle and/or left turn crashes which tend to be more severe. However, as the definition of the intersection is changed to include some area around it (i.e., the influence area for an intersection is defined); rear-end and other groups of crashes would be included in the sample and the severity patterns may be altered.

The influence area for an intersection is characterized by the distance from the center of the intersection along either of the two legs belonging to the corridor under consideration. Crashes within this distance from any intersection (signalized or unsignalized) are categorized as intersection/intersection-related crashes while the crashes beyond are categorized as segment crashes. This study attempts to understand factors associated with crashes and their severity on a multilane arterial while accounting for the variations resulting from location of the crashes relative to intersections. It is accomplished by developing different models for different distance thresholds used to define the influence area for intersections. The methodology used in this study also accounts for the correlations between the factors explaining injury severity and the crash location (intersection vs. segment) at a particular threshold. The approach adopted herein provides a better understanding of relationship between crash location's relative proximity to intersections and severity outcome. It may also improve the understanding of how changes made to an intersection affect the neighboring segments of the arterial.

Crash data from SR-816 corridor in Broward County, Florida are used is this study. The crashes belonging to intersections are separated from crashes belonging to arterial segments by defining a binary variable whose definition changes based on the specified intersection influence distance. Injury severity of crashes is defined as an ordinal variable. Detailed characterization of these two variables is provided in the next section along with particulars of the solution approach and modeling methodology. The section providing details of the data used for analysis is then followed by the results and conclusions of this investigation.

3.2 Solution Approach and Modeling Methodology

Relationships between the following variables are of interest in this study:

1. A 3-level ordinal variable representing the injury severity. The variable is created from the injury severity information available from the Crash Analysis and Reporting (CAR) database of Florida Department of Transportation (FDOT).
2. A binary variable representing crash location; with its value being '1' for crashes which occur within the threshold influence distance of an intersection (intersection/intersection-related crashes) and '0' for crashes that occur outside this influence distance (segment crashes). In this study, the influence distance (taken from the center of the intersection) would be varied in 50 ft. increments on arterial corridors. Hence, there would be multiple binary variables that would distinguish between crashes based on their location (i.e., intersection and non-intersection crashes).

An ordered probit modeling framework would be used for the first variable since injury severity levels are naturally ordered. Ordered probit modeling has been applied to injury severity in several studies by Abdel-Aty (2003), O'Donnell and Connor (1996), and Duncan et al. (1999). However, none of these studies, except for Abdel-Aty (2003), compared the factors that affect injury severity at different roadway locations. Abdel-Aty (2003) used the ordered probit model to study severity of traffic crashes at roadway sections and at signalized intersections. The analyses for these roadway elements (segments and intersections), however, were carried out independent of each other.

In the preliminary analysis chi-square tests for association between injury severity and the binary variable(s) representing crash location suggested a possible association between them. Furthermore, the nature and strength of association changes as the definition of the variable representing crash location is varied. The results from these tests are later discussed in detail. The straight forward way to assess the impact of crash location (i.e., intersection) on injury severity would be to use the binary variable(s) representing crash location as an independent variable in the ordered probit model for injury severity. However, this binary variable would be related with the variables generally used in the model for the injury severity. For example, the crashes under rainy conditions are less likely to occur right at the intersection compared to the roadway segment influenced by intersections. Similarly, left-turn or angle crashes are more likely to occur within the physical area of the intersection (compared to segments) and they are also likely to be more severe. To avoid the confounding effects of other variables it was decided that the models for the crash location (binary dependent variable) and the injury severity (ordinal

dependent variable) would be estimated simultaneously. Since the location variable may be associated with certain variables included in the severity model; its inclusion (i.e., recursive specification) would have also led to problems of correlated independent variables, and biased and inefficient estimates for the coefficients.

Simultaneous estimation of the two models would improve the coefficient estimates by accounting for the correlations between the unmeasured factors. The difference between independent estimation and the simultaneous (bivariate) modeling procedure is that the later does not assume the errors for the two models to be uncorrelated. The simultaneous estimation procedure also provides the *p-value* for the statistical test on correlation with the null hypothesis being that the correlation coefficient $\rho=0$.

3.3 Model Formulation

According to Long (1997) logit and probit models provide very similar results in terms of resulting classification and standardized effects for independent variables. However, convergence is more likely for bivariate probit models, even though it may require more computational time (Indiana University (2007)). The model specification for the simultaneously estimated probit model equations is as follows (Green (2003)):

Equation 3-1

$$Y_1^* = X_1' \beta_1 + \varepsilon_1$$

Equation 3-2

$$Y_2^* = X_2' \beta_2 + \varepsilon_2$$

Where X_1= Vector of independent variables explaining the roadway location of the crash and

X_2 = Vector of independent variables explaining the crash injury severity.

Also, note that the disturbances ε_1 and ε_2 have the following specifications:

$E[\varepsilon_1 | X_1, X_2] = E[\varepsilon_2 | X_1, X_2] = 0$,

$Var[\varepsilon_1 | X_1, X_2] = Var[\varepsilon_2 | X_1, X_2] = 1$,

$Cov[\varepsilon_1, \varepsilon_2 | X_1, X_2] = \rho$

Y_1^* and Y_2^* are unobserved, latent, and continuous variables. The binary and ordinal scale dependent variables, Y_1 (Crash location) and Y_2 (Injury severity) are observed when the respective latent variables Y_1^* and Y_2^* fall in certain ranges. The two independent variables observed as discrete categories (i.e., Y_1 and Y_2) are specified below:

Equation 3-3

$$Y_1 = \begin{cases} 0 & if \quad Y_1^* \leq 0 \\ 1 & if \quad Y_1^* > 0 \end{cases}$$

Equation 3-4

$$Y_2 = \begin{cases} 0 & if \quad Y_2^* \leq 0 \\ 1 & if \quad 0 < Y_2^* \leq \mu \\ 2 & if \quad Y_2^* > \mu \end{cases}$$

Equation 3-3 (specified as a binary probit model) relates crash location with other crash characteristics; while Equation 3-4 (specified as an ordered probit model) relates injury severity with the independent variables. This formulation allows us to relate the crash location with the injury severity without confounding the effects of independent variables that relate to both crash location and injury severity. Detailed descriptions of the variables constituting the vectors X_1 and X_2 are provided in the next section (See Table 3-1).

The estimates for model coefficients may be obtained using maximum likelihood estimation. The likelihood function maximized to obtain the model coefficients incorporates the effect of correlation between the error terms. The coefficients for the models specified above (i.e., vectors β_1, β_2 along with $(\rho(u_1, u_2))$ were estimated using SAS (2007). The details of maximum likelihood estimation process may be found in by Greene (2003).

Multiple sets of simultaneous models (corresponding to different thresholds on influence distance) based on the above specification would be estimated for the corridor. The only difference between sets of simultaneous models would be the definition of Y_1 (i.e., Crash location variable). The definition of Y_1 would in turn depend on the threshold selected to separate intersection crashes from segment crashes. The details on these thresholds and the variables used in the analysis are provided in the next section.

3.4 Data Preparation

The crash data used in this study are from a 9.72-mile corridor of arterial SR-816 in Broward County, FL. Both signalized and non-signalized intersections are considered in this study. The intersection density (intersections per mile) for the corridor is 11.32. The total number of crashes involving at least a possible injury on this multilane arterial over the four year period (2002 through 2005) was 1575. 11.17% of these crashes were either fatal or involved incapacitating injury. The crash data for the above corridor were downloaded from FDOT's CAR database.

Table 3-1 Variable Description

Variable	Categories	Description
Independent Variables		
Traffic Condition (Based on time of day/day of week)	MPW	Morning peak traffic condition on weekday (7 a.m. – 9.30 a.m.)
	APW	Afternoon peak traffic condition on weekday (4 p.m. – 7 p.m.)
	FSN	Friday or Saturday night traffic condition (Friday 10 p.m. – Saturday 3.30 a.m.)
	OP	Off peak traffic condition
Sectional AADT	1*	Section AADT <= 52,000
	2*	52,000 < Section AADT <= 58,000
	3*	58,000 < Section AADT <= 64,500
	4*	Section AADT > 64,500
Road Surface		Binary (1 = dry surface, 0 = all other cases)
Lighting		Binary (1 = day time, 0 = night time)
Weather		Binary (1 = clear, 0 = all other cases)
Road Curvature		Binary (1 = straight, 0 = curve)
Road Surface Type		Binary (1 = blacktop, 0 = all other cases)
Road Condition at time of Crash		Binary (1 = No defects, 0 = all other cases)
Vision Obstruction		Binary (1 = no obstruction, 0 = all other cases)
Alcohol/Drug involvement		Binary (1 = No, 0 = Yes)
Pavement Surface Width		Width of the pavement (Continuous)
Shoulder Width1		Width of the shoulder closest to the travel lane (Continuous)
Shoulder Width2		Width of the shoulder farthest from the travel lane (Continuous)
Median Width		Width of the median (Continuous)
Speed Limit		Maximum posted speed limit (continuous)
Dependent Variables		
Crash Location ($\mathbf{Y_1}$; location_D)	1	Crashes within the 'D' ft. from the center of intersection
	0	Crashes beyond 'D' ft. from the center of intersection
Injury Severity ($\mathbf{Y_2}$)	2	Crashes resulting in incapacitating injuries or fatalities
	1	Crashes resulting in non-incapacitating injuries
	0	Crashes resulting in possible injuries

***The AADT values from various sections of the corridor have been split into four quartiles**

Before proceeding further some data issues require clarification. The issues mainly relate to the recorded crash location and the definition of influence distance. In the database used for this study each crash is assigned to an intersection (node) nearest to its location. The information on the distance of crash location from the node representing center of the intersection is also available in the database. Through a careful review of this information; it was noticed that a significant number of crashes are reported to have occurred at milepost associated with the

nodes. In other words, the distance between crash location and the center of the intersection is reported as 0 ft. It does not necessarily mean that all these crashes occurred at the mid-point of the intersection. Significantly large number of such crashes essentially implies that most crashes that occur inside the physical area of the intersection are reported to be 0 ft. from the center of intersection. Also, note that in the state of Florida physical area of the intersection is by default considered to be the area within 50 ft. from the center of the intersection. Hence, some of the crashes that are reported to be within 50 ft. of the node (representing the intersection) in the database may be very close to the Stop bar.

These crashes, while not strictly *at* intersection, would most likely be influenced by it. Therefore, the first two thresholds on influence distance (to separate intersection crashes from non-intersection crashes) were chosen to be 0 ft. and 50 ft., respectively. The threshold of 0 ft. means that the crashes within 50 ft. from the center of the intersection are classified as intersection crashes (i.e., only those crashes within the physical area of the intersection). For the model corresponding to D=50; the crashes that have occurred within the physical area of the intersection and those that have occurred within 50 ft. of the stop bar have been classified as intersection crashes. The successive thresholds were also chosen to be in 50 ft. increments, i.e., 100 ft., 150 ft., and so on. As mentioned earlier, this threshold defines one of the two simultaneously estimated dependent variables ($\mathbf{Y_1}$; See previous section).

It must be acknowledged that the selection of thresholds at 50 ft. increments is somewhat arbitrary. Therefore, the results from the sets of simultaneous models estimated using different

thresholds need to be interpreted in relative terms. For example, in case of the models with threshold at 100 ft.; crashes closer to intersection are treated as intersection/intersection related crashes compared to the set of models with threshold at 150 ft. Table 3-1 lists the independent (regressors) and dependent variables (responses) used in the study. The last row of Table 3-1 represents the crash location as binary variable "location_D" which would be '1' for crashes within 'D' ft. from the center of the intersections.

Crashes with fatalities and incapacitating injuries are combined into one category (of variable $\mathbf{Y_2}$ representing injury severity) for two reasons; first, the relatively small frequency of fatal crashes compared to other injury severity levels could create problems in the analysis. For example, the chi-square tests on contingency tables may not be valid due to low expected cell-frequency. The second reason is that the crashes that involve incapacitating injury could easily have been fatal and vice-versa depending on the vulnerability of the subjects involved. Also, note that the variables shown in Table 3-1 are gathered from the 'long-form' (complete crash reports) filled out by law enforcement authorities in Florida. The information on crashes involving no injury is likely to be incomplete for this set of crashes (Abde-Aty and Keller, 2005; Yamamoto et al., 2008). Therefore, only crashes that at least involve a possible injury are included in this study and the injury severity is categorized as a 3-level ordinal variable.

Note that some of the binary variables shown in Table 3-1 had in fact more levels in the original database. Some of the categories belonging to these variables were quite infrequent and were therefore combined with each other. Also, note that the AADT of the sections was divided into

four quartiles (such that they have close to 25% cases in each of the categories). In the analysis this variable is used as a nominal variable and not as an ordinal variable. The reason is that the categorization may not follow the natural order in terms of the relationship of AADT with injury severity ($\mathbf{Y_2}$). All the other variables shown in the table are self-explanatory.

3.5 Analysis and Results

As mentioned earlier, the association between the ordinal variable representing crash injury severity and the binary variable(s) representing the crash location was first examined with chi-square tests. To reliably assess the strength of this association using chi-square test; each cell of the contingency table is required to have a minimum expected frequency. With the increase in the influence distance (starting from 0 ft.) more crashes get assigned as intersection crashes and the number of crashes assigned as non-intersection or segment crashes is reduced. Beyond a certain influence distance the frequency of segment (or non-intersection) crashes becomes too low for the chi-square test statistic to be credible. Therefore, a maximum allowable influence distance was chosen such that at least 10% of all crashes were assigned as non-intersection crashes. The maximum allowable threshold influence distance for SR-816 was found to be 200 ft. using this criterion. Limiting the threshold distance to 200 ft. also helps in reducing the chances of having influence area of one intersection overlap with the other. The chi-square test statistics and corresponding p-values for testing associations between $\mathbf{Y_1}$ and $\mathbf{Y_2}$ (with definition of $\mathbf{Y_1}$ varying based on influence distance thresholds; D=0 ft. through D=200 ft.) are reported in Table 3-2.

Bivariate probit models, formulated earlier in the chapter, were then developed for the injury severity ($\mathbf{Y_2}$; ordered probit) and the location variable ($\mathbf{Y_1}$; binary probit). The bivariate formulation does not assume that the errors in the models being simultaneously estimated are uncorrelated. The significance of correlation coefficient (ρ) is tested and reported along with the estimated coefficients (and their significance) for independent variables included in the two models. The correlation essentially accounts for the common factors associated with both dependent variables that are not explicitly included in the models. Last column of Table 3-2 also provides the estimates for 'ρ' and its significance. Table 3-3 shows the detailed estimates of variables coefficients and their significance along with error correlation coefficient estimates shown in last column of Table 3-2.

Table 3-2 Chi-square statistics and error correlation coefficient estimates

Influence Distance (ft)	**Chi-Square** (p-value) (from Contingency tables)	**Correlation coefficient 'ρ'** (p-value) (from bivariate probit models shown in Table 3)
0	**4.369** (0.113)	**0.053** (0.172)
50	**1.354** (0.508)	**-0.046** (0.266)
100	**1.285** (0.526)	**-0.055** (0.201)
150	**7.889** (0.019)	**-0.135** (0.005)
200	**5.950** (0.051)	**-0.120** (0.016)

It can be observed in Table 3-2 that the significance trend for 'ρ' at various intersection influence distance is similar to the corresponding significance trend of the Chi-square statistic. In Table 3-2 and Table 3-3 cells with statistically significant parameters (at 90% confidence level) have been highlighted. It is worth mentioning that the values of μ (for converting the estimated latent continuous variable into the categorical injury severity) were also estimated for each of the five injury severity models and are provided in Table 3-3.

Table 3-3 Five simultaneous models for the crash location and injury severity levels on SR-816 (D=threshold influence distances in ft.)

		D = 0		D = 50		D = 100		D = 150		D = 200	
Parameter		Estimate	Approx p-value	Estimate	Approx p-value	Estimate	Approx p-value	Estimate	Approx p-value	Estimate	Approx p-value
Crash Location Model											
Traffic Condition	APW	-0.086	0.338	-0.166	0.074	-0.148	0.125	-0.157	0.152	-0.176	0.115
Traffic Condition	FSN	-0.090	0.504	-0.057	0.693	-0.076	0.618	-0.228	0.181	-0.222	0.211
Traffic Condition	MPW	-0.061	0.625	-0.055	0.671	-0.063	0.638	-0.062	0.685	-0.122	0.424
Traffic Condition	OP	0.000	.	0.000	.	0.000	.	0.000	.	0.000	.
Dry Road Surface		0.276	0.023	0.251	0.050	0.149	0.266	-0.010	0.948	-0.049	0.753
Daylight Condition		-0.150	0.039	-0.225	0.004	-0.246	0.003	-0.253	0.008	-0.246	0.013
Clear Weather		-0.110	0.298	-0.168	0.139	-0.167	0.157	-0.085	0.521	0.004	0.977
Straight Road Section		-0.483	0.159	-0.277	0.461	-0.402	0.339	0.016	0.970	0.071	0.870
Blacktop Road Surface		-0.112	0.245	0.007	0.944	0.064	0.537	0.201	0.080	0.223	0.056
No Vision Obstruction		-0.042	0.758	0.064	0.647	0.234	0.097	0.094	0.572	0.039	0.820
No Alcohol/Drug Use		-0.080	0.596	0.114	0.472	-0.099	0.570	-0.011	0.954	-0.189	0.394
Injury Severity Model											
Traffic Condition	APW	-0.216	0.017	-0.214	0.019	-0.215	0.018	-0.214	0.018	-0.214	0.019
Traffic Condition	FSN	0.154	0.229	0.154	0.228	0.154	0.229	0.154	0.229	0.154	0.231
Traffic Condition	MPW	0.045	0.710	0.042	0.729	0.040	0.737	0.039	0.745	0.040	0.741
Traffic Condition	OP	0.000	.	0.000	.	0.000	.	0.000	.	0.000	.
Dry Road Surface		0.095	0.425	0.094	0.433	0.093	0.435	0.093	0.433	0.094	0.430
Daylight Condition		0.044	0.535	0.044	0.527	0.045	0.520	0.047	0.507	0.047	0.508
Clear Weather		-0.021	0.837	-0.020	0.844	-0.020	0.845	-0.021	0.839	-0.021	0.836
Straight Road Section		-0.020	0.953	-0.019	0.956	-0.016	0.962	-0.018	0.958	-0.017	0.961
Blacktop Road Surface		-0.221	0.015	-0.223	0.014	-0.223	0.014	-0.224	0.014	-0.223	0.014
No Road Defects at time of Crash		0.029	0.881	0.046	0.812	0.047	0.806	0.072	0.708	0.064	0.742
No Vision Obstruction		-0.144	0.274	-0.150	0.256	-0.150	0.255	-0.157	0.233	-0.155	0.239
Pavement Surface Width		0.040	0.018	0.044	0.009	0.044	0.008	0.047	0.005	0.046	0.006
Closest Shoulder Width		-0.263	0.362	-0.258	0.371	-0.262	0.364	-0.259	0.368	-0.259	0.369
Farthest Shoulder Width		-0.181	0.370	-0.175	0.388	-0.171	0.399	-0.169	0.405	-0.171	0.399
Median Width		-0.013	0.016	-0.013	0.015	-0.013	0.015	-0.013	0.012	-0.013	0.012
Maximum Posted Speed Limit		0.023	0.023	0.021	0.033	0.021	0.033	0.020	0.040	0.020	0.037
AADT (1st Quartile)	1	0.329	0.003	0.333	0.002	0.335	0.002	0.326	0.003	0.323	0.003
AADT (2nd Quartile)	2	0.226	0.014	0.221	0.017	0.221	0.017	0.209	0.024	0.210	0.023
AADT (3rd Quartile)	3	0.044	0.638	0.045	0.631	0.043	0.648	0.028	0.762	0.028	0.762
AADT (4th Quartile)	4	0.000	.	0.000	.	0.000	.	0.000	.	0.000	.
No Alcohol/Drug Use		-0.322	0.021	-0.322	0.021	-0.322	0.021	-0.323	0.020	-0.324	0.020
μ (for classification)		0.916	<.0001	0.916	<.0001	0.916	<.0001	0.916	<.0001	0.916	<.0001
Correlation Coefficient											
Rho		0.053	0.172	-0.046	0.266	-0.055	0.201	-0.135	0.005	-0.120	0.016

The significance of 'ρ' changes as the influence distance for defining crashes on intersection and segment varies from 0 ft. through 200 ft. For the models developed for intersection influence distances 0, 50, and 100 ft. the 'ρ' values is insignificant. It indicates that error terms in the two models are not significantly correlated with each other. However, the correlation coefficient becomes significant beyond 100 ft. influence distance. Table 3-2 also depicted a similar trend for the significance of the Chi-square test statistic. This in effect means that on average when intersection crashes are defined such that they include a smaller influence area (within about 100 ft. of intersections for this corridor); severity on the arterial crashes may be modeled independent of crash location. It is worth mentioning again that the 100 ft. is the distance from the center of the intersection. Also, note that this distance may also vary from corridor to corridor depending on intersection density and traffic patterns. As mentioned earlier due to data constraints we have not been able to develop models for D > 200 ft. and beyond. It may be inferred that the correlation would probably have been significant.

Form this point forward the discussion would be about the factors that were found to be significant for the two simultaneously estimated models at various threshold distance. The crash location ($\mathbf{Y_1}$) model(s) for various threshold values (D) show what factors help discriminate between intersection crashes and segments crashes. The crash injury severity ($\mathbf{Y_2}$) model(s) for various threshold values (D) in Table 3-3 show the factors that relate significantly with the ordinal variable. Figure 1 depicts the significant parameters for the ordinal crash injury severity model in the form of a bubble plots. The size of the bubbles in the plots reflects the relative significance of these parameters with respect to each other. Also, note that the bubbles within a

plot may be compared horizontally but not vertically. The plot on the left shows the effect of the factors which decrease the severity of the crash (i.e., negative coefficients) and the plot on the right depicts those which increase the severity (i.e., positive coefficients).

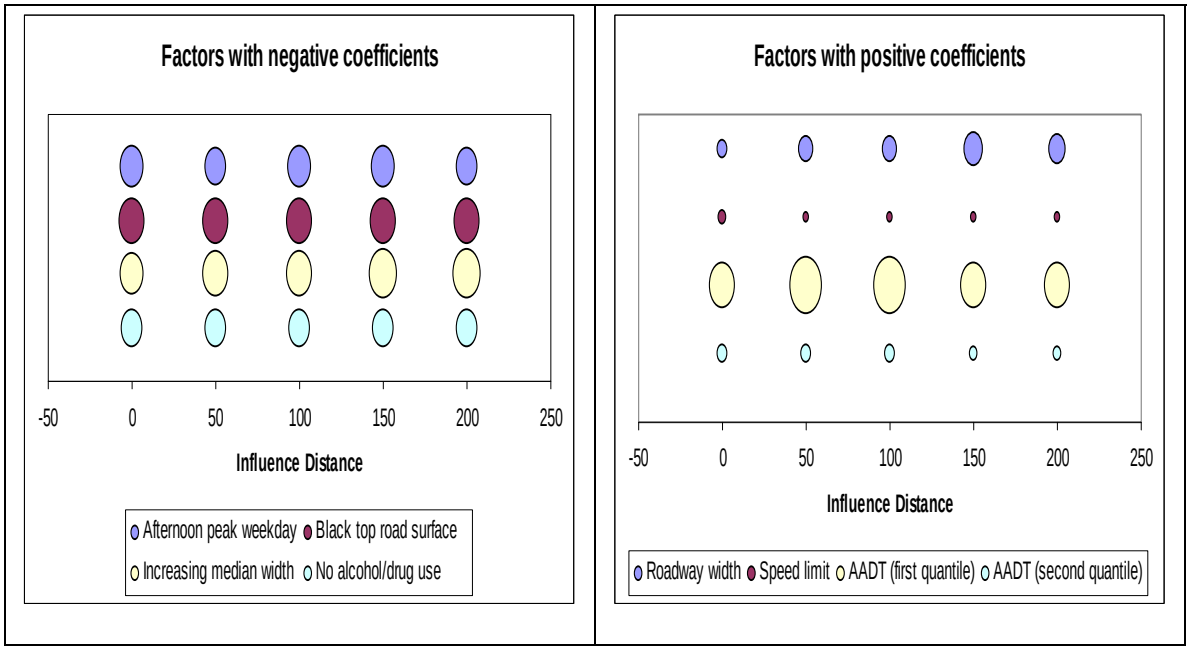


Figure 3-1 Significant parameters for crash injury severity model

Figure 3-1 and Table 3-3 illustrate that weekday afternoon peak period conditions (APW; see Table 3-1), blacktop pavement surface, and increase in median width decrease the severity of the crashes on SR-816 (LHS of Figure 3-1). No alcohol/drug use also has the same effect which essentially means that alcohol/drug involvement increases the severity of the crashes. During afternoon peak periods the speeds are generally lower due to congestion; therefore crashes are likely to be less severe. Likewise, higher median width may reduce the chances of severe crossover head-on collisions. It explains the significantly negative coefficient for median width.

Similar result for severity of peak-hour crashes at intersections was found by Abdel-Aty and Keller (2003). Presence of median was also found to reduce the severity of crashes in that study.

Blacktop surfaces are found to negatively affect severity in all five models (D=0 ft. through D=200 ft.). Note that this variable is also significant for separating intersection vs. segment crashes when intersection crashes include the crashes that occur within 150 ft. and 200 ft. of the intersections (Models for D=150 ft. and D=200 ft. in Table 3-2; Also see Figure 3-2). For other three values of 'D' (defining intersection crashes as only those that occur within 0, 50, and 100 ft. of intersections) this variable was not significant in the crash location model. These crashes are not only less severe (Abdel-Aty and Keller (2003), Jianming and Kockelman (2004)) but are also likely to be more frequent in the segment within 150-200 ft. from intersections. The findings also seem to corroborate with one of the studies that found that asphalt pavements may lead to higher frequency of peak period crashes (Abdel-Aty et al. (2006)). Since crashes on blacktop surfaces with asphalt base seem to have higher frequencies during peak periods and within 150-200 ft. of intersection; it indicates that these pavement surfaces *might* increase the odds of rear-end crashes. It may in turn be the reason for the negative coefficient of the variable representing blacktop surfaces with asphalt base in the injury severity model (since rear-end crashes tend to be less severe).

It was also found that increases in the roadway width and the speed limit increase the severity of the crashes. AADT below the median value (both 1st and 2nd Quartiles) are also positively associated with the severity (RHS of Figure 3-1). Among the factors positively influencing injury

severity, lower AADT is the most significant. It is also interesting that the effect of roadway width becomes more profound when the influence distance is greater than zero.

Figure 3-2 depicts significant parameters for five binary crash location models each estimated simultaneously with the corresponding injury severity model. The coefficients of the model were provided in Table 3-3. The size of the bubbles once again reflects the relative significance of the parameters. Note that the some parameters have no corresponding bubble at certain values of D (i.e., intersection influence distances). It represents that if crashes at intersections are defined based on this influence distance then the corresponding parameters does not contribute in discriminating the crash location. In Figure 3-2 the plot on the left shows the effect of the factors which decrease the likelihood of a crash being within a particular distance from intersection (negative coefficients) while the plot on the right depicts those which increase it (positive coefficients).

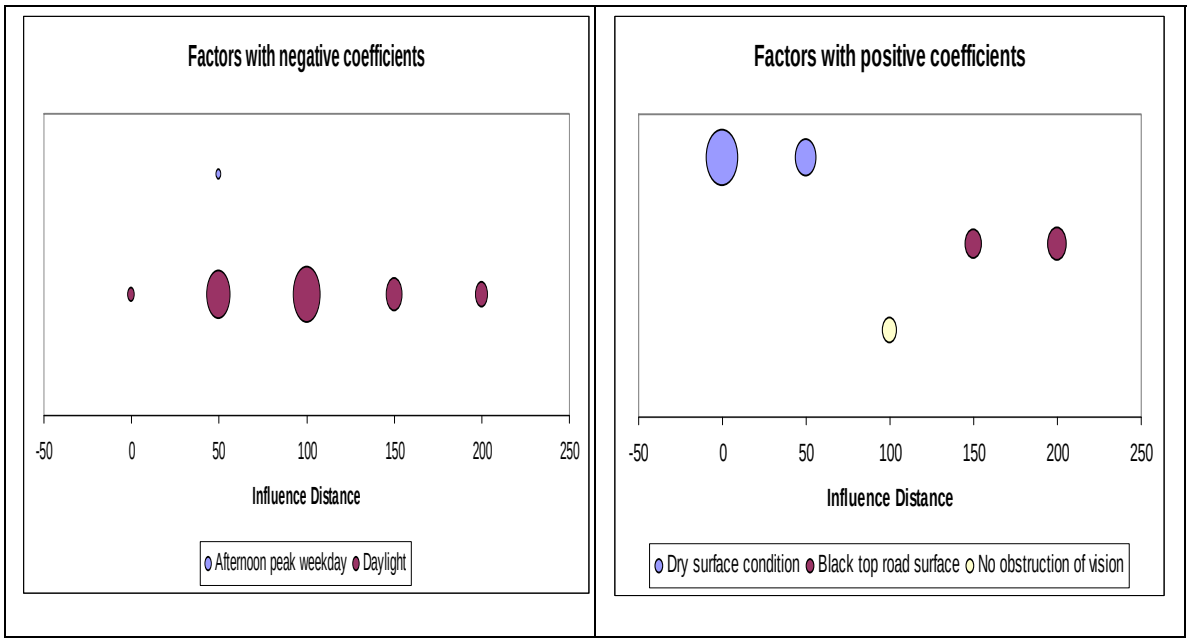


Figure 3-2 Significant parameters for crash location model

From Figure 3-2 and Table 3-3 it may be observed that during the afternoon peak period on weekdays the likelihood of a crash occurring within 50 ft. of the intersection is less compared to the off-peak traffic conditions. Note that while this difference is insignificant at influence distances of 100, 150, and 200 ft. (no corresponding bubble in LHS of Figure 3-2); the p-value is much closer to 0.10 (See Table 3-3). This difference between afternoon peak weekdays (APW; See Table 3-1) and off peak (OP; See Table 3-1) conditions is insignificant if one examines the relative likelihood of a crash occurring within the physical area of an intersection (influence distance=0 ft.). It is probably because during the afternoon peak hours, drivers expect congestion and expect to slow down/stop as they approach an intersection. It reduces the likelihood of crashes that prevail in the vicinity of intersections. The modeling results also show that the variable representing normal daylight is significant in separating crashes at intersection and segments regardless of the specified influence distance. It should be noted, however, that the significance is more profound if the influence area of intersections is defined as 50 and 100 ft. While it is hard to conclude definitively the smaller coefficient of this variable at influence distances D=150 ft. and D=200 ft. might be caused by the dilemma zone phenomenon.

Among the variables with positive coefficients (RHS of Figure 3-2) blacktop road surface is significant for separating intersection crashes from segment crashes if the influence distance is 150 ft. or 200 ft. The implications of this result were discussed earlier. Dry surface condition also augments the likelihood of a crash to occur at the physical area of an intersection or within 50 ft.

of it. It may also be observed that the significance is more profound for the physical area of the intersection compared to the case when the influence distance is 50 ft. It can essentially be interpreted as follows: if the influence area of the intersection is increased then the weather conditions' ability to discriminate between intersection and segment crashes diminishes. It might be due to wet weather crashes that are more prevalent on approaches to intersections. A result that was not clearly understood was the variable representing vision obstruction was found to be significant in identifying intersection crashes from segment crashes with the influence distance set at 100 ft. The variable was not significant at any other influence distance and p-values were not even on the margin. This might be a peculiar issue with the corridor under consideration such as a few intersections with vision obstruction problems along the corridor or the demographics of Broward County with a sizeable proportion of older drivers.

3.6 Concluding Remarks

Understanding safety on urban arterials is a complex problem since it is affected by interactions between traffic patterns on intersections and segments connecting them. Implementation of certain safety improvements at intersections may lead to unanticipated changes in safety/operation performance of nearby segments or vice versa. Hence, an improved understanding of safety may be achieved if consecutive intersections on arterial corridors are examined as a whole along with the segments connecting them instead of as isolated entities. Analysis presented in this chapter is an effort in that direction which focuses on injury outcomes of crashes.

The analysis is carried out by simultaneous estimation of models for crash location and injury severity at five different values of intersection influence distances. These values are varied from D=0 ft. through D=200 ft. at 50 ft. increments. The value of influence distance (D) essentially represents the distance from center of intersection along the corridor, up to which the crashes are categorized as intersection related. Simultaneous estimation of crash location and injury severity models allows us to account for correlation between errors of the two models. The correlation is likely the result of common unknown factors that affect both these variables but are not explicitly included in either model.

The model for the crash location variable indicated that during peak hours crashes are less likely to occur at or in the vicinity of intersections. It was also found that increase in the pavement surface width and speed limits expectedly increase the severity of the crashes. Lower AADT values are also positively associated with the severity. It may be inferred that certain conditions that make the task of driving easier (larger higher roadway width, low AADT) can lead to increased severity of crashes.

It should be noted that the results obtained in this study may be specific to the corridor under consideration. It may be expected, however, that similar results (for example, the influence distance threshold beyond which the error correlation coefficient becomes significant) would be obtained from corridors with comparable intersection density. The results also suggest that for corridors with higher intersection density (i.e., more closely spaced intersections) the errors may

not be correlated and hence crash location and injury severity may be modeled independent of each other. This inference is based on insignificant correlation between errors for the simultaneous models developed corresponding to D=0, 50, and 100 ft. On the other hand, arterials on which intersections are fewer and far between; injury severity models for the corridor need to account for crash location (i.e., intersection vs. segment crashes).

CHAPTER 4. RULES TO ASSIGN CRASHES

4.1 Background

Signalized and un-signalized intersections and the segments connecting them are the three basic elements of any given arterial. The common practice is that crashes have been assigned to these elements based on the crash location. For the present study signalized intersections will be referred to as intersections while as un-signalized intersections will be considered as a type of access points. An access point is any street that is intersecting the arterial and has a control other than a signal. It could be a county road or a private driveway. Most states in the country have in their jurisdiction an influence area for an intersection. For example in the state of Florida, crashes that occurred within 250 ft of any intersection are referred to as intersection related crashes (Abdel-Aty & Wang, 2006 and Wang et. al 2006). The problems of having an influence area to assign crashes could be:

1. Having an influence distance has its problem of wrongly classifying some segment crashes to intersection related.
2. Das et al. (2008), showed by the method of simultaneous estimation that if the influence distance varied the crash characteristics associated with severe injuries also varies. This is due to the fact that the farther we move away from the center of an intersection, more crashes related to the connecting segment comes into play. Wang et al. (2008) used frequency modeling for crashes with fixed as well as varying influence distance and found different set of significant factors. These very recent studies show that the concept of using influence distance for assigning crashes to the roadway elements could be erroneous.

Apart from the above problems associated with the influence distance, there are other problems that are related to the ways crashes are reported. Most of time the police officers do not do an actual measurement of the crash distance. The crash distance, which also decides the crash location, is the distance of the crash from the center of an intersection to the exact location. Also the distance is sometimes taken from the stop bar on the arterial. In addition to these, Florida State has a 50 ft default intersection size. Since not all intersections are of the same size, no matter how good the officer is at guessing the location indicated in the crash report is a very rough approximation. Hence using the influence distance to classify intersection related crashes is not recommended.

There is presently no standard guideline for un-signalized intersections such as influence distance in case of signalized intersections. If the "site location" is used to determine the location of a crash, the only access related crashes that could be identified are those with site location value of 'driveway access'.

For the present corridor level analysis it is critical to know how to assign a crash to its appropriate roadway element. The goal is to assign crashes to segments, intersections or access points. Police officers often report the crashes that have occurred at an un-signalized intersection as intersection related. This makes the site location parameter a weak indicator of assigning crashes. Using it alone to assign crashes could lead to erroneous results. This lead to an investigation to find out which other crash record parameters could be used to assign the crashes correctly. A closer study of crash reports revealed that traffic control in combination with the site location did a superior job in identifying the roadway element to be assigned to correctly. Hence the method of assigning a crash based on crash characteristics. However in certain cases the above two crash parameters may not be able distinguish crashes that are related to intersections or access points. In those cases it is necessary to check whether the particular node is signalized or un-signalized. Node generally refers to any type of intersection, both signalized as well as un-signalized.

Based on the detailed study of 377 crash reports certain rules, in the form of *if-then-else* statements, were developed to assign the crashes correctly. The rules had an overall accuracy of 93.63 % as compared to 57.82 % accuracy obtained when only site location is used.

In the sections that follow details for each rule will be given which will enable the reader to not only understand the rules but also learn as to how the rules were developed. The numeric representation of the parameters: 'site location' and 'traffic control' have been used. Table 4-1 and Table 4-2 provide the meaning of each numeric depiction for the above two parameters.

Table 4-1 Legend for 'Site Location'

Site Location	Numeric representation
Not at Intersection / RR / Bridge	1
At Intersection	2
Influenced by Intersection	3
Driveway Access	4
Railroad	5
Bridge	6
Entrance Ramp	7
Exit Ramp	8
Parking Lot – Public	9
Parking Lot – Private	10
Private Property	11
Toll Booth	12
Public Bus Stop Zone	13
All Other	77

Table 4-2 Legend for 'Traffic Control'

Traffic Control	Numeric representation
No Control	1
Special Speed Zone	2
Speed Control Sign	3
School Zone	4
Traffic Signal	5
Stop Sign	6
Yield Sign	7

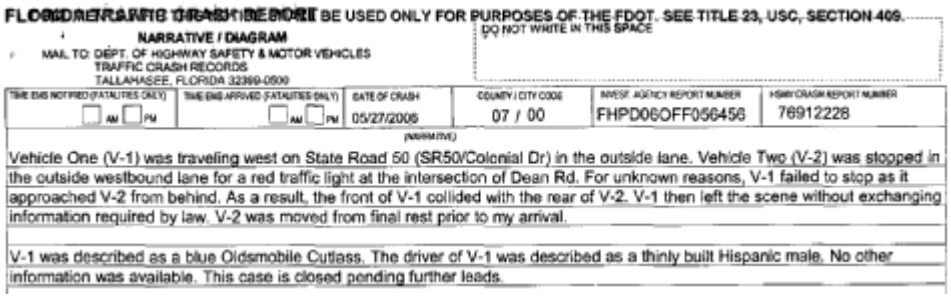
Flashing Light	8
Railroad Signal	9
Officer / Guard / Flag person	10
Posted No U-Turn	11
No Passing Zone	12
All Other	77

It is important to note several observations in Table 4-1. The author would like to bring to notice of the reader the site locations with values 1, 5 and 6. As can be observed the site location value of 1 relates to crashes not at intersection or railroad or bridges. However the practice is such that crashes at railroads and bridges are almost always have a site location value of 5 and 6 respectively which exclusively identifies crashes related to railroad and bridges. Apart from that crashes near railroad which have site location value of 1 will have traffic control value of 9. Similarly the author would also like to bring the discussion to site location values of 4, 9, 10 and 11. They all represent access related crashes. However the data will almost never have any crashes with site location values of 9, 10 or 11 since they are all driveway related and are included under site location value of 4. Site location value of 12 which represents toll booth is not a part of the present study.

4.2 Site location 1: Not at Intersection / RR Xing/ Bridge

Based on the site location value of 1 alone one would assign all the crashes to segments. It is true for crashes where the traffic control is 1, 2, 3, 4, 10 or 12. However when the traffic control is 5, 6, 7, 8, 9 or 11, then the crashes do not always occur due to segment characteristics. Given that site location is 1 and, for example, the traffic control is 5 then a closer look at the crash reports reveal that those crashes occurred due to signalized intersection related causes. Similarly when

the traffic control is 6 an investigation into the crash reports show that those crashes are related to un-signalized intersections with a stop sign ('access points' in our case). The above statements are exemplified in Figures 4-1 through 4-4. Figure 4-1 and Figure 4-2 are from the crash report #769122280 where the site location is 1 and the traffic control is 5. In this particular instance the 'at fault' driver rear-ended the stationary vehicle in front view. The stationary vehicle was stopped at an intersection and was waiting for the red light to turn green. Even though it has been classified as a 'not at intersection' crash, this crash definitely is related to the signalized intersection and need to be assigned as such.



FLORIDA TRAFFIC CRASH REPORT BE USED ONLY FOR PURPOSES OF THE FDOT. SEE TITLE 23, USC, SECTION 409.

NARRATIVE / DIAGRAM

MAIL TO: DEPT. OF HIGHWAY SAFETY & MOTOR VEHICLES
TRAFFIC CRASH RECORDS
TALLAHASSEE, FLORIDA 32399-0500

DO NOT WRITE IN THIS SPACE

TIME EMS NOTIFIED (FATALITIES ONLY)	TIME EMS ARRIVED (FATALITIES ONLY)	DATE OF CRASH	COUNTY / CITY CODE	INVEST. AGENCY REPORT NUMBER	HSMV CRASH REPORT NUMBER
AM PM	AM PM	05/27/2006	07 / 00	FHPD06OFF056456	76912228

(NARRATIVE)

Vehicle One (V-1) was traveling west on State Road 50 (SR50/Colonial Dr) in the outside lane. Vehicle Two (V-2) was stopped in the outside westbound lane for a red traffic light at the intersection of Dean Rd. For unknown reasons, V-1 failed to stop as it approached V-2 from behind. As a result, the front of V-1 collided with the rear of V-2. V-1 then left the scene without exchanging information required by law. V-2 was moved from final rest prior to my arrival.

V-1 was described as a blue Oldsmobile Cutlass. The driver of V-1 was described as a thinly built Hispanic male. No other information was available. This case is closed pending further leads.

Figure 4-1 Crash narrative by the police officer

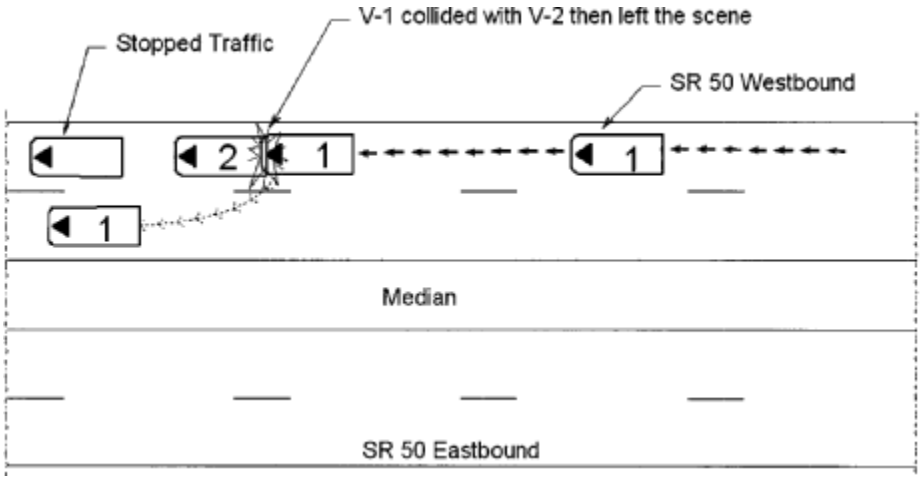
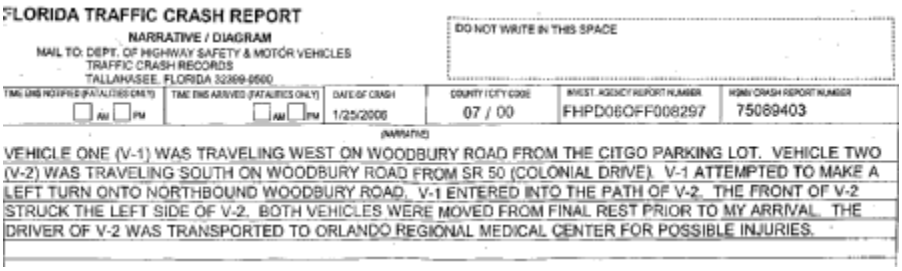


Figure 4-2 Graphical representation of how the crash had or may have occurred

Likewise Figure 4-3 and Figure 4-4 are from the crash report #750894030 where the site location is 1 and the traffic control is 6. The description and the illustration clearly indicate that the crash is related to the un-signalized intersection rather than the segment. The 'at fault' driver was getting out of a driveway and while attempting to make a left turn came in collision course of the other vehicle, resulting in an angle crash.



FLORIDA TRAFFIC CRASH REPORT

NARRATIVE / DIAGRAM

MAIL TO: DEPT. OF HIGHWAY SAFETY & MOTOR VEHICLES
TRAFFIC CRASH RECORDS
TALLAHASSEE, FLORIDA 32399-0500

DO NOT WRITE IN THIS SPACE

TIME EMS NOTIFIED (FATALITIES ONLY)	TIME EMS ARRIVED (FATALITIES ONLY)	DATE OF CRASH	COUNTY / CITY CODE	INVEST. AGENCY REPORT NUMBER	HSMV CRASH REPORT NUMBER
☐ AM ☐ PM	☐ AM ☐ PM	1/25/2008	07 / 00	FHPD08OFF008297	75089403

(NARRATIVE)

VEHICLE ONE (V-1) WAS TRAVELING WEST ON WOODBURY ROAD FROM THE CITGO PARKING LOT. VEHICLE TWO (V-2) WAS TRAVELING SOUTH ON WOODBURY ROAD FROM SR 50 (COLONIAL DRIVE). V-1 ATTEMPTED TO MAKE A LEFT TURN ONTO NORTHBOUND WOODBURY ROAD. V-1 ENTERED INTO THE PATH OF V-2. THE FRONT OF V-2 STRUCK THE LEFT SIDE OF V-2. BOTH VEHICLES WERE MOVED FROM FINAL REST PRIOR TO MY ARRIVAL. THE DRIVER OF V-2 WAS TRANSPORTED TO ORLANDO REGIONAL MEDICAL CENTER FOR POSSIBLE INJURIES.

Figure 4-3 Crash narrative by the police officer

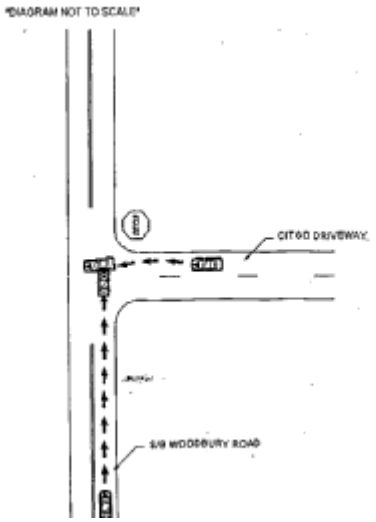


Figure 4-4 Graphical representation of how the crash had or may have occurred

Hence it is now clear that the site location only should not be used to assign crashes to the different roadway locations. At least a combination of site location and traffic control is required to correctly assign the crashes where the site location is 1.

Figure 4-5 is the flowchart of how a crash is to be appropriately assigned to one of the three roadway locations when the site location is 1. The flow chart is essentially a set of *if-then-else* statements which can conveniently be understood. After all the checks for the traffic control are made the crashes are assigned to the correct roadway component.

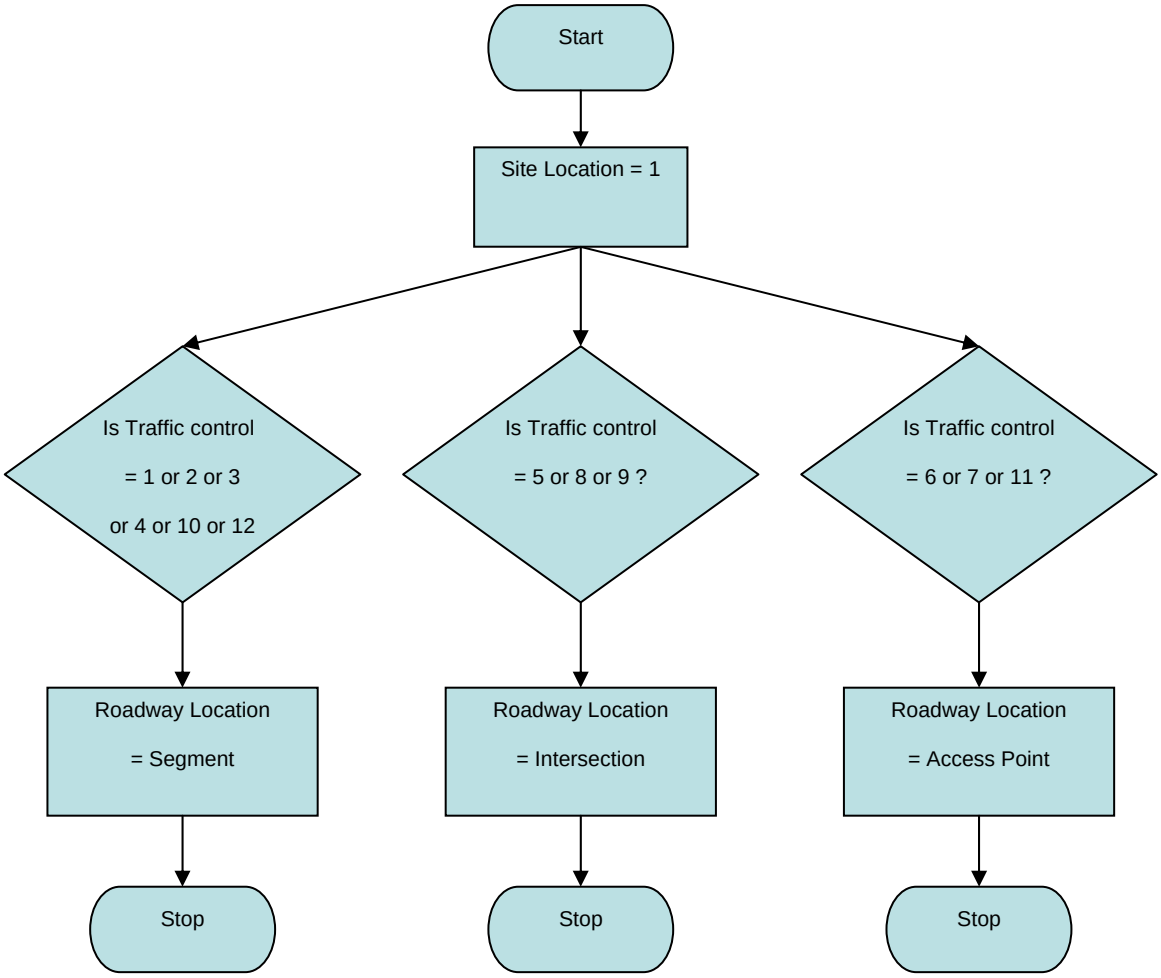


Figure 4-5 Rules to assign crashes to roadway elements based on Site Location = 1

4.3 Site location 2: At Intersection

The site location value of 2 essentially means that the crash has taken place inside a signalized intersection. However, as mentioned earlier, the way reporting is done a lot of crashes that have occurred in un-signalized intersections also are reported as intersection crashes. Therefore for these crashes some new parameter apart from site location and traffic control have to be taken into account to distinguish signalized intersection crash and un-signalized intersection crashes.

Here the node information, i.e. whether the crash is signalized or un-signalized is used to assign the crashes to intersections or access points. That particular variable is not necessary for traffic control values of 5, 6, 7, 8, 9 or 12 where there was found to be no conflict. Figures 4-6 through 4-9 will illustrate how the conflict may arise and thus support the use of the new binary variable. Figure 4-6 and Figure 4-7 of crash report #719651960 indicate that it is an access related crash. The combination of site location (= '2') and traffic control (= '1') alone cannot help resolve the misclassification. Hence there is a need to know the signal information of that particular node.

FLORIDA TRAFFIC CRASH REPORT
NARRATIVE/DIAGRAM

DO NOT WRITE IN THIS SPACE

DATE OF CRASH	COUNTY / CITY CODE	AGENCY REPORT NUMBER	CRASH REPORT NUMBER
12/05/2006	06/32	2006020603	71965196

(NARRATIVE)

V1 was exiting the parking lot of 6000 W. Glades Road and was proceeding to travel eastbound on W. Glades Road. V2 was traveling eastbound on W. Glades Road in the left thur lane approaching the intersection of Butts Road. D1 stated that the traffic in the right thru and center thru lane stopped so he could exit the parking lot and enter the left thru lane. D1 advised that he did not see V2 traveling in the left thru lane. D2 advised she could not avoid the collision, but applied her brakes but was unable to stop in time. As a result, the front of V2 collided with the left side of V1. Both vehicles were at final rest upon my arrival. No further information.

Figure 4-6 Crash narrative by the police officer

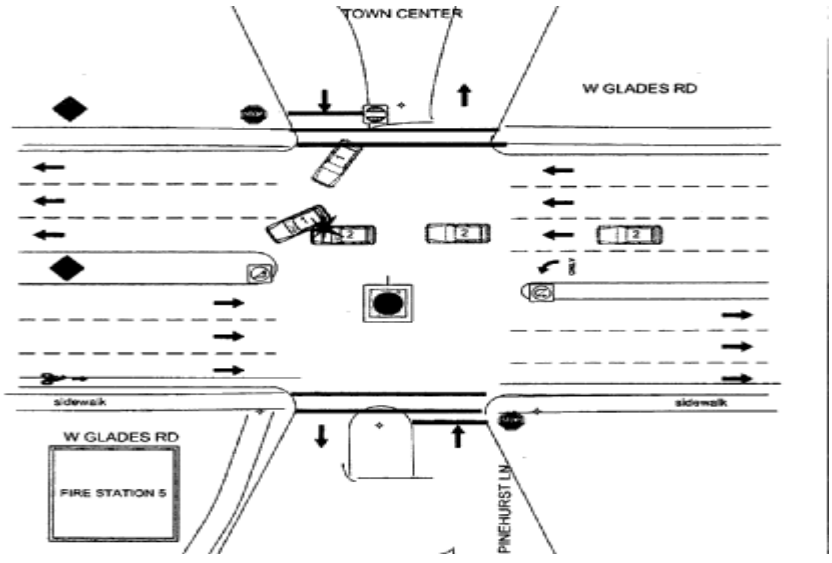


Figure 4-7 Graphical representation of the crash had or may have occurred

Figure 4-8 and Figure 4-9of the crash report #719651790 clearly point out that the crash is a signalized intersection crash. The site location is 2 and the traffic control is 1. Hence by just observing the site location or the simple combination of site location and traffic control variable we can correctly assign some of the crashes but not most of it. Hence the node check variable is important.

FLORIDA TRAFFIC CRASH REPORT
NARRATIVE/DIAGRAM

DO NOT WRITE IN THIS SPACE

DATE OF CRASH	COUNTY/CITY CODE	INVEST. AGENCY REPORT NUMBER	HSMV CRASH REPORT NUMBER
12/07/2006	06/32	2006020754	71965179

(NARRATIVE)

V1 was traveling Northbound on NW 15th Ave in the inside left turn lane to head westbound on W. Glades Rd. V2 was traveling Northbound on NW 15th Ave in the inside left turn lane to head westbound on W. Glades Rd. and stopped for traffic in the intersection with Glades Rd directly infront of V1. V1 failed to stop intime causing the front of her vehicle to collide into the rear of V2.

It should be noted, D1 stated a third vehicle was involved. The third vehicle was directly infront of V2. The license plate given to me of this vehicle (CW tag:131SVA) came back no record found in NCIC/FCIC.

Figure 4-8 Crash narrative by the police officer

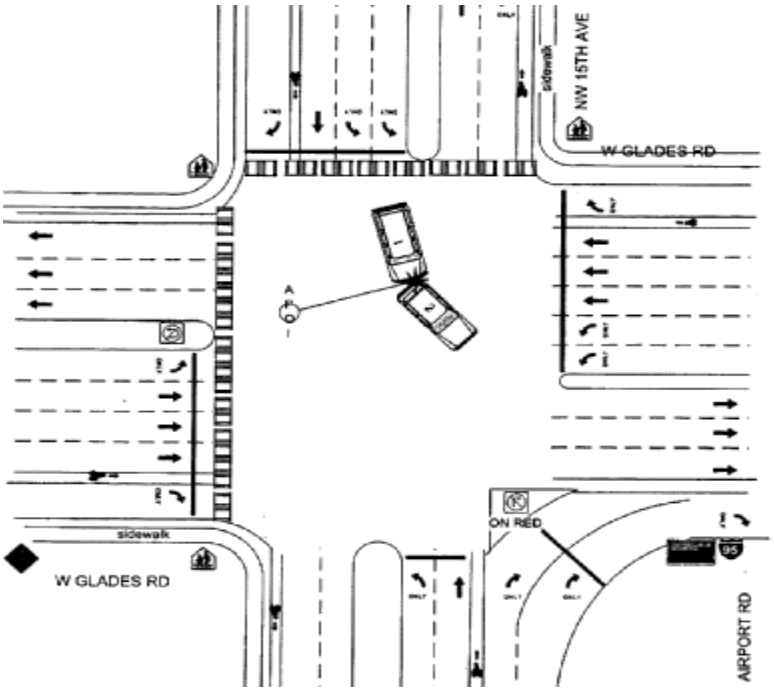


Figure 4-9 Graphical representation of how the crash had or may have occurred

Figure 4-10 and Figure 4-11 will provide an example for the site location value of 2 where the node information is not necessary and the simple rules may be applied. The crash report #754075840 has the traffic control value of 5, i.e. traffic signal and this is a clear example of a signalized intersection related crash.

FLORIDA TRAFFIC CRASH REPORT
NARRATIVE/DIAGRAM

DO NOT WRITE IN THIS SPACE

TIME DRS NOTIFIED (FATALITIES ONLY)	TIME EMS ARRIVED (FATALITIES ONLY)	DATE OF CRASH	COUNTY / CITY CODE	INVEST. AGENCY REPORT NUMBER	CRASH REPORT NUMBER
☐AM ☐PM	☐AM ☐PM	5/25/06	53	06010964	754075841

V2 WAS traveling WESTBound on S. Nccall ROAD in the left lane. V1 was turning RIGHT AT the RED light at the intersection of S. Nccall ROAD and PINE ST. V1 stopped and proceeded TO turn RIGHT ON TO S. Nccall ROAD, and drove Directly into the left lane of WESTBound traffic, sideswiping V2. V1 and V2 both pulled off the ROADWAY. The Driver of V2 complained of neck pain and EMS was notified and responded. V2's driver was examined and advised EMS that she will just visit her Chiropractor. The driver of V1 was cited for careless driving. citation #4662-DLD.

Figure 4-10 Crash narrative by the police officer

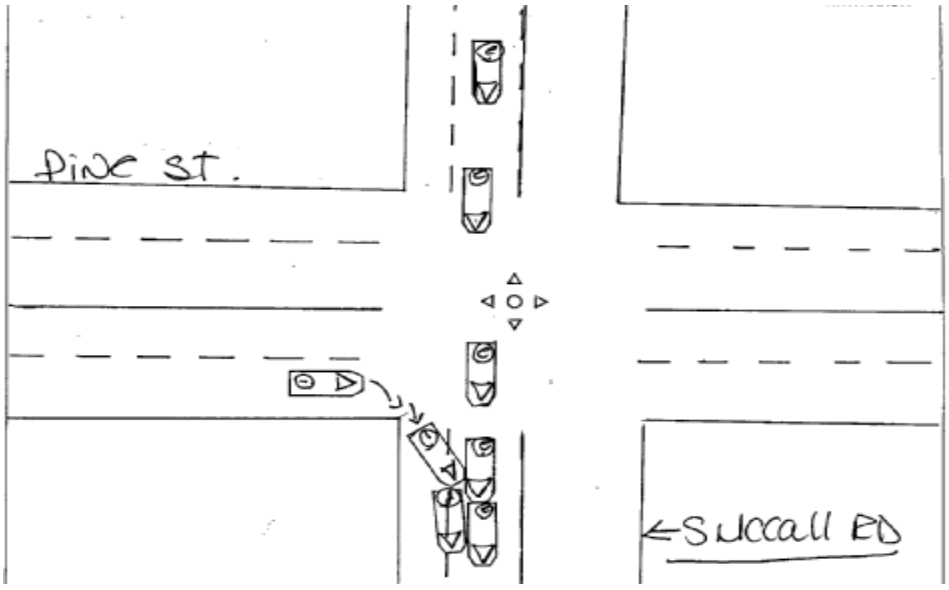


Figure 4-11 Graphical representation of how the crash had or may have occurred

Similarly when the traffic control is 6 i.e. stop sign the crashes are almost always access related. Therefore the traffic control value in some cases is sufficient to distinguish between signalized intersection crashes and access crashes. However some traffic control values are not discerning enough. Figure 4-12 is the flowchart of how a crash is to be appropriately assigned to one of the three roadway elements when the site location is 2. After all the checks for the traffic control and signalization of nodes are made the crashes are assigned to the correct roadway component.

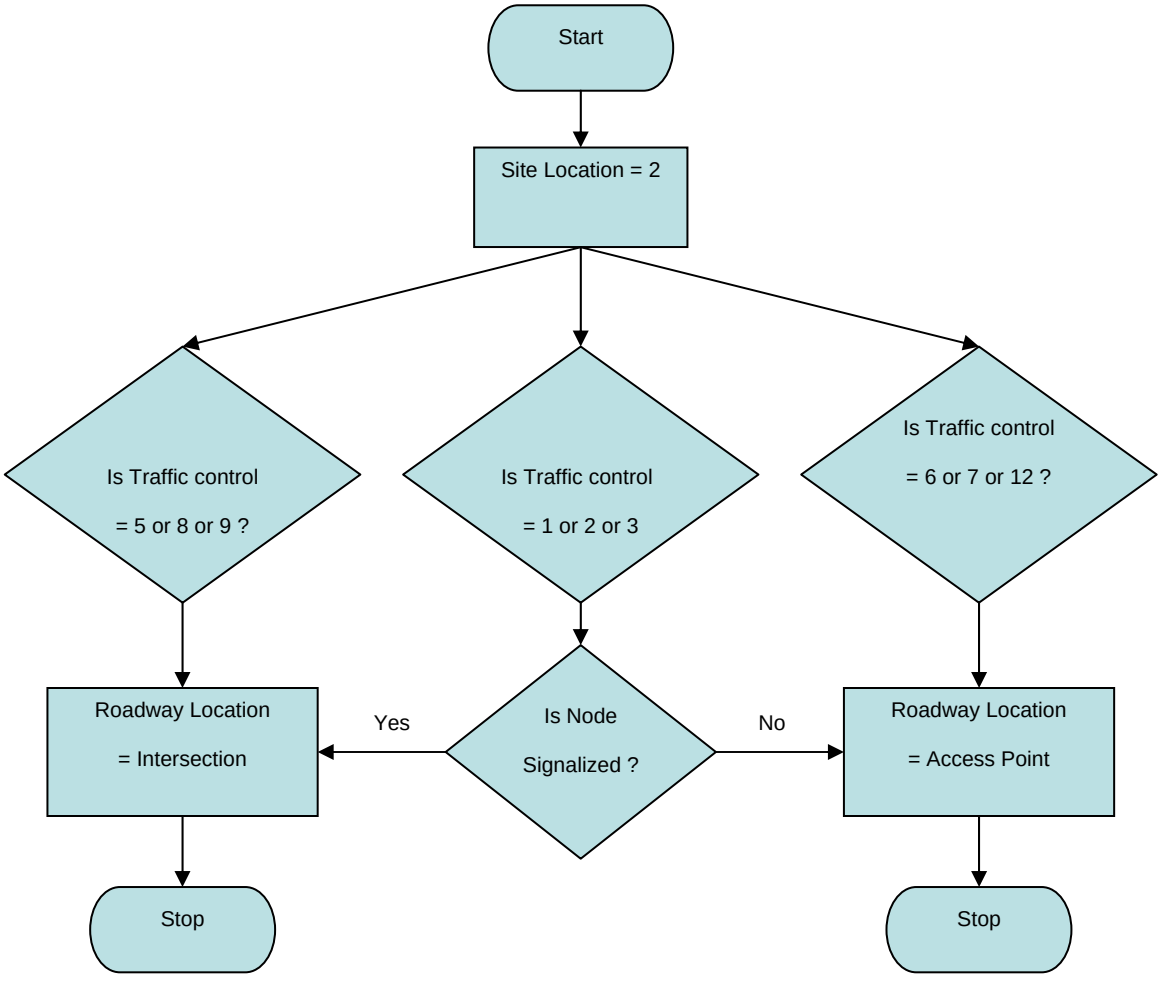


Figure 4-12 Rules to assign crashes to roadway elements based on Site Location = 2

4.4 Site location 3: Influenced by Intersection

The site location value of 3 which identifies crashes influenced by intersection site location has exactly the same issues as the site location value of 2. Certain values of traffic control are capable of correctly assigning the crashes while some others are not. Hence the rules are almost similar to those developed for the site location value of 2. Figure 4-13 show the rules for the present site location.

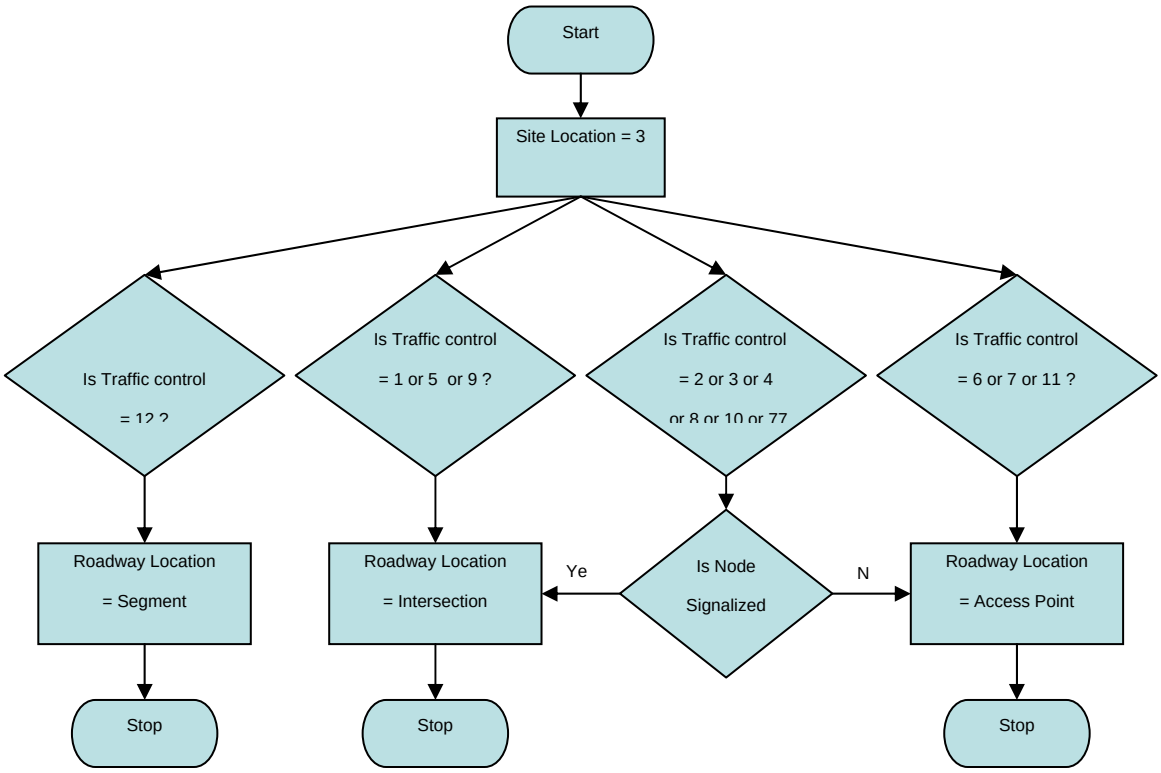


Figure 4-13 Rules to assign crashes to roadway elements based on Site Location = 3

4.5 Site location 4: Driveway Access

Driveway access related crashes have been assigned to as access related crashes. Earlier in the discussion it was mentioned that for the present analysis driveways along with other un-signalized intersections are access related. In this particular site location most of the crashes are related to access points. Except for the cases when the traffic control is 5 or 8 i.e. 'traffic signal' or 'flashing light'. The rules are given in Figure 4-14.

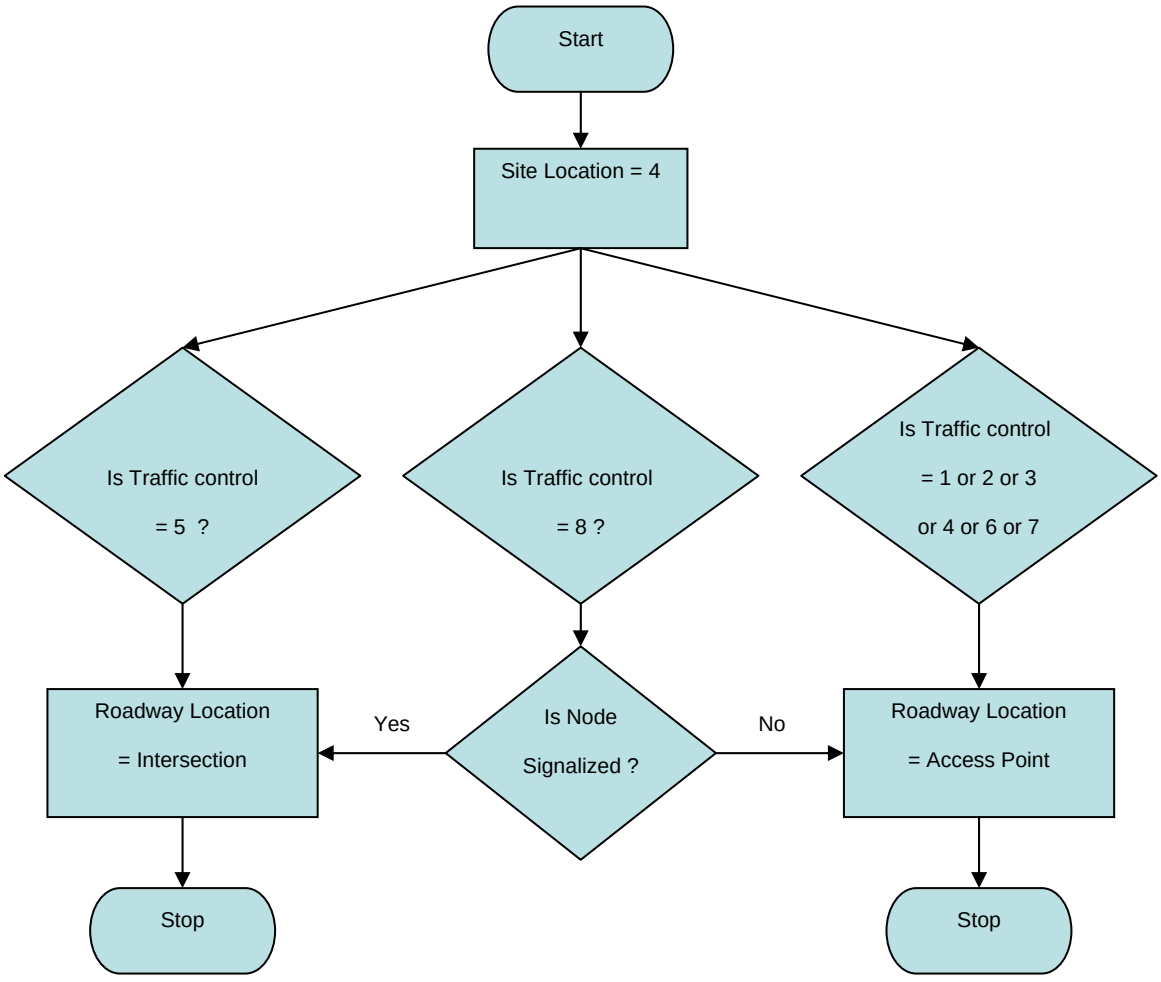


Figure 4-14 Rules to assign crashes to roadway elements based on Site Location = 4

4.6 Site location 5: Railroad

The site location value of 5 i.e. railroad helps in identifying those crashes which have occurred at or near a railroad intersection. The rules for node checking are required for certain traffic control values. The rules for correctly assigning the site location are given in Figure 4-15.

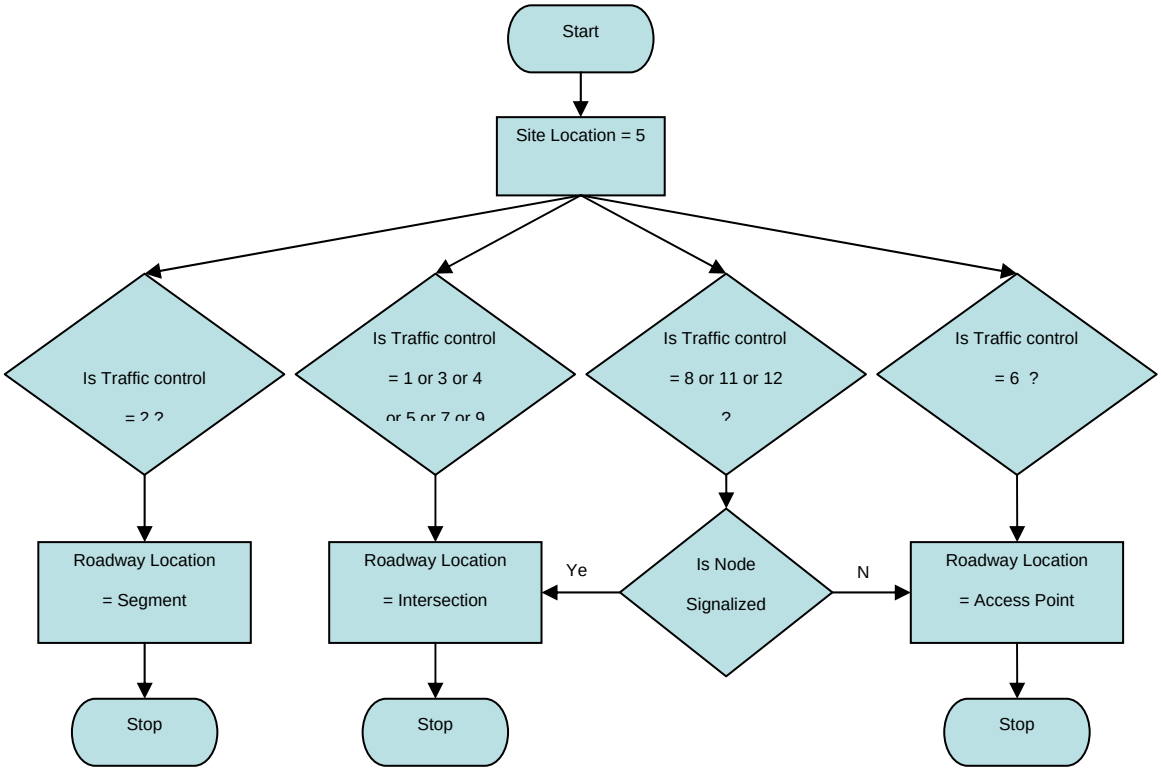


Figure 4-15 Rules to assign crashes to roadway elements based on Site Location = 5

4.7 Site location 6: Bridge

The site location value of 6 i.e. bridge helps in identifying those crashes which have occurred at or near a bridge. The rules' flowchart is given in Figure 4-16. Most of the crashes are segment

related, however the rules have to be made to correctly assign those crashes that could not have been due to the segment.

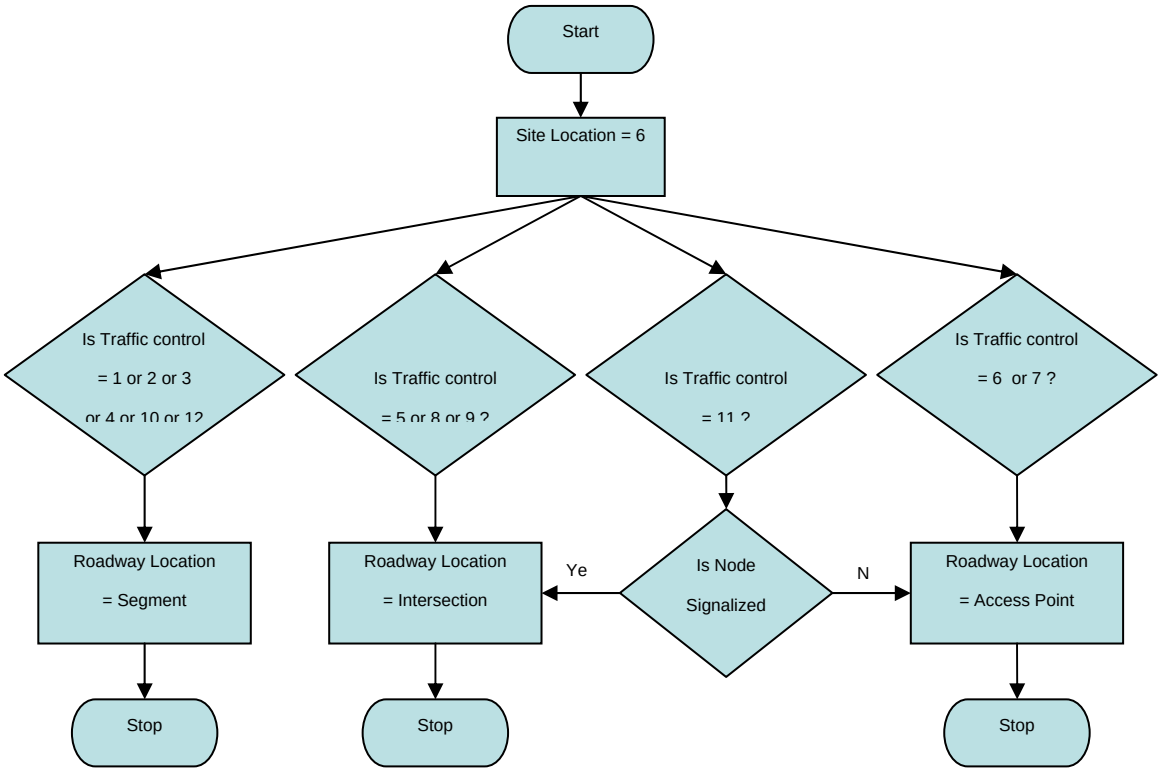


Figure 4-16 Rules to assign crashes to roadway elements based on Site Location = 6

4.8 Site location 7 / 8: Entrance / Exit Ramp

The site locations for entrance and exit ramps are essentially intersections which could be signalized or un-signalized. The traffic control will be used along with the node check procedure to assign the crashes correctly to being signalized intersection related or access related. Figure 4-17 shows the flowchart for the appropriate rules.

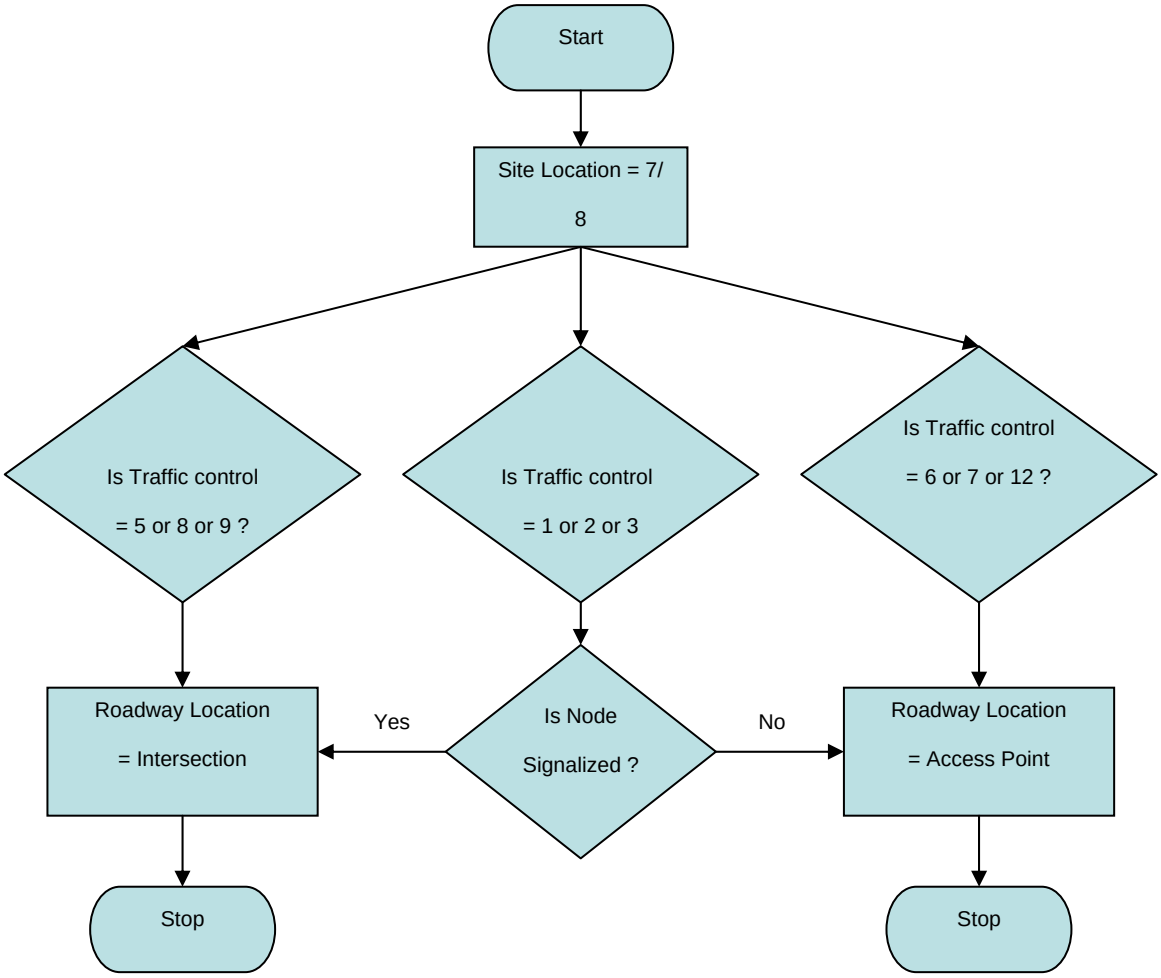


Figure 4-17 Rules to assign crashes to roadway elements based on Site Location = 7 or 8

4.9 Site location 13: Public Bus Stop Zone

The site location value of Public Bus Stop zone are segment related crashes with some being intersection or access related. The rules are given in Figure 4-18.

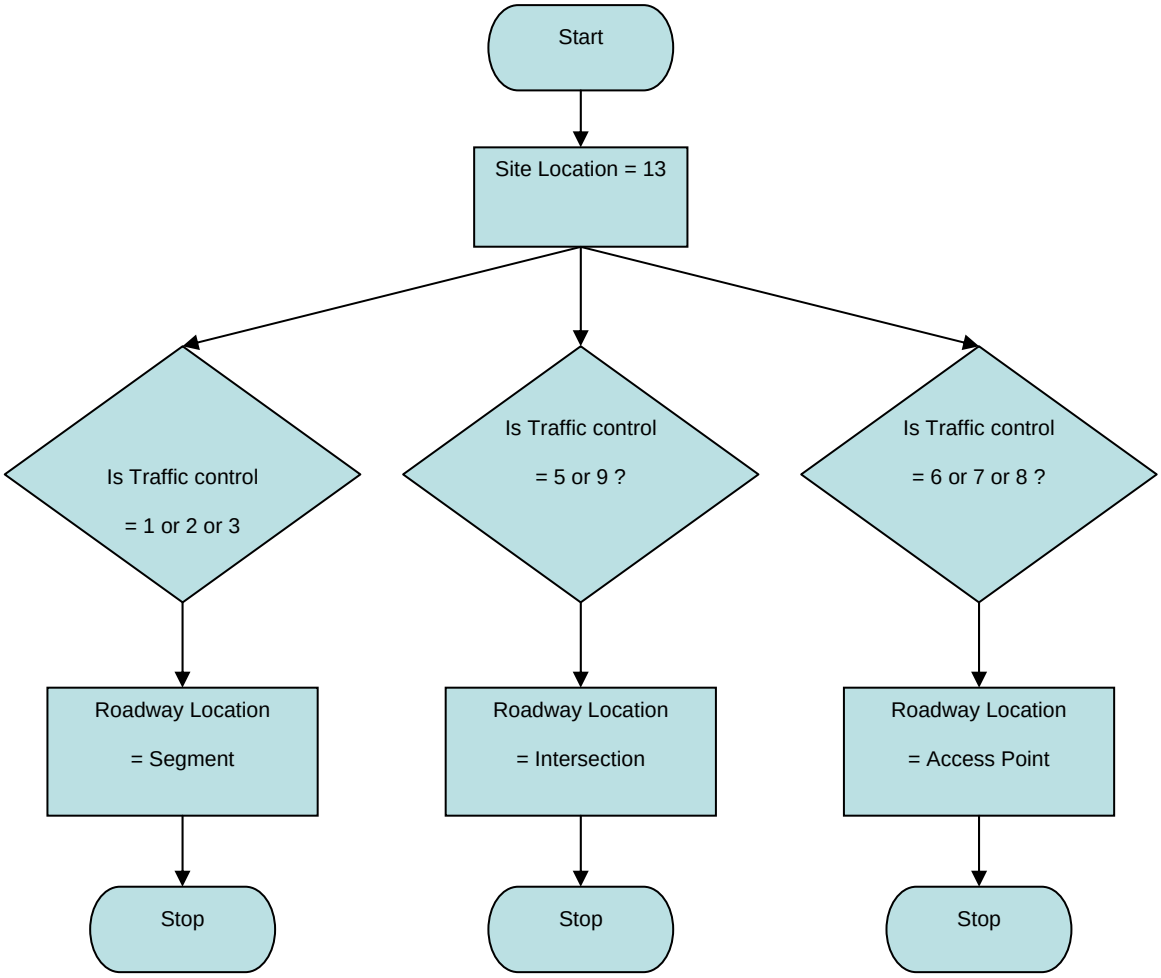


Figure 4-18 Rules to assign crashes to roadway elements based on Site Location = 13

4.10 Quantitative Validation of the Rules

As mentioned earlier the rules are based on careful observations of crash reports for different combinations of site location and traffic control. Though the rules have not been developed through any statistical process a quantitative validation is required so as to not only evaluate how good they perform but also to have an estimate as to how much better they are had only the site location been taken into consideration. Crash reports were thoroughly scrutinized and each crash

was assigned to one of the three roadway locations as defined earlier: segments, signalized intersections and access points. The rules were developed by analyzing 96 crash reports for different combinations of site location and traffic control. This essentially works as a training set of crash reports. The rules were then validated using 281 more crash reports. Hence a total of 377 crash reports were studied to come up with a complete set of rules.

Out of the first 96 crash reports the assigning accuracy without the rules i.e. by using only the site location was 53.13 % where as the accuracy using the rules was 87.5 %. Even at the training or development stage of the rules we can notice a considerable improvement in the assignment to the correct roadway element. The validation crash reports gave an assigning accuracy of 59.43 % without the rules and with the rules the accuracy improved to 95.73 %. The overall accuracy for the 377 crash reports was 57.82 % without the rules and 93.63 % with the rules. Hence it can be observed that approximately 36 % more crashes are assigned correctly by using the rules than by not using them.

CHAPTER 5. DATABASES

5.1 Existing databases

Florida Department of Transportation (FDOT) has two very comprehensive resources namely the Crash Analysis and Reporting (CAR) System and the Roadway Characteristics Inventory (RCI). Since we will be investigating severe injury / fatality crashes on Florida's State Roads we will be using the above mentioned databases for our research purpose as they provide all the necessary traffic and geometric variables.

5.1.1 CAR

The CAR has records for all crashes in the state of Florida that required a Florida Traffic Crash Report Long Form to be filled out. The crash records have information at levels of crash, vehicle, person and citation. This makes the CAR a very exhaustive resource. The records can be viewed online by authorized users and can also be downloaded in text format. The particular databases from CAR to be used were: 1) The Augmented Detail Extract and 2) Vehicle – Driver – Passenger Extract. The former has essentially the crash characteristics associated with roadway geometry and environmental conditions. The latter database has driver – passenger information for all the vehicles involved in the crash. Both of the databases have 86 variables each. Apart from these datasets the CAR also has statistical reports for 'high crash' roadway segments across the state of Florida. These crash locations (segments) are termed 'high crash', and are confirmed statistically problematic area, for certain confidence level and minimum number of crashes. The

default value for the confidence level is 99% and the minimum number of crashes is 8. The present study will require us to investigate corridors for the entire state. Therefore setting the values of confidence level and minimum number of crashes will require some assumptions, which may prove to be incorrect later on. For that reason it was decided to generate those reports with all the values set to zero. This also helps to study the crash information on all roadway segments across the state. Figure 5-1 given below is the snapshot of the report generated for roadway segments.

REPORT..CARPJ66- FLORIDA - DEPARTMENT OF TRANSPORTATION PAGE NO 1
DATE....07/19/2007 C A R - CRASH ANALYSIS REPORTING SYSTEM
TIME....16:17:23 REFERENCE CRASH ROADWAY SEGMENTS FOR 2005 USING: STATE AVERAGES AS OF: 08/11/2006
COMMENT: _____ USERID: SF945AP DISTRICT:

NUMB	COSECSUB	BMP	EMP	STROAD	LENGTH	CC	CRASHES	ADT	ACTUAL	AVERAGE	CONLV	FTL	INJ	PRTY	CL-1	CL-2	CL-3
4708	75000001	0.000	2.200	CR 436A	2.200	UNKN	0	N/A	*******	N/A	*****	0	0	0			
4709	75000012	0.000	7.844	CR OFF	7.844	UNKN	0	N/A	*******	N/A	*****	0	0	0			
4710	75000151	0.000	1.520	CR OFF	1.520	UNKN	0	N/A	*******	N/A	*****	0	0	0			
4711	75000154	0.000	2.821	CR OFF	2.821	UNKN	0	N/A	*******	N/A	*****	0	0	0			
4712	75000200	0.000	0.402	CR OFF	0.402	UNKN	0	N/A	*******	N/A	*****	0	0	0			
4713	75000227	0.000	6.350	CR OFF	6.350	UNKN	0	N/A	*******	N/A	*****	0	0	0			
4714	75000260	0.000	1.260	CR OFF	1.260	UNKN	0	N/A	*******	N/A	*****	0	0	0			
4715	75000267	0.000	0.835	CR OFF	0.835	UNKN	0	N/A	*******	N/A	*****	0	0	0			
4716	75001500	0.000	2.608	CR 532	2.608	UNKN	0	N/A	*******	N/A	*****	0	0	0			
25	75002000	0.000	0.220	SR 482	0.220	U-6DR	44	38,497	14.233	3.338	99.99	0	62	14			
233	75002000	0.220	0.520	SR 482	0.300	U-4DR	29	39,625	6.683	2.699	99.99	0	28	13			
642	75002000	0.520	0.682	SR 482	0.162	U-4DR	13	40,498	5.428	2.699	99.50	1	10	8			
2642	75002000	0.682	0.957	SR 482	0.275	S-4DR	3	40,498	0.738	1.461	25.24	0	2	1			
3223	75002000	0.957	1.367	SR 482	0.410	U-4DR	4	40,496	0.660	2.699	12.22	1	4	0			
4717	75002000	1.367	1.474	SR 482	0.107	S-4DR	0	40,499	0.000	1.461	0.00	0	0	0			
3089	75002000	1.474	2.110	SR 482	0.636	U-4DR	9	46,760	0.829	2.699	15.35	0	7	5			
1659	75002000	2.110	2.510	SR 482	0.400	S-4DR	11	47,500	1.586	1.461	50.00	0	5	6			
4718	75002000	2.510	2.810	SR 482	0.300	S-4DR	0	47,499	0.000	1.461	0.00	0	0	0			
54	75002000	2.810	3.280	SR 482	0.470	S-4DR	47	45,588	6.009	1.461	99.99	0	48	26			

Figure 5-1 Snapshot of the 'high crash' reference report for roadway segments

The column 'numb' in the report is a reference number that indicates segments from the highest to the lowest with specific criterion. This report also contains average daily traffic, the number of crashes, the actual crash rate, the average crash rate, fatalities, injuries, and property damage crashes on the particular length of the given segment The actual crash rate (crashes per million vehicle miles) is determined by dividing the number of crashes on the segment for the time spanned by the analysis by the total vehicle miles for the segment for the year span. It also gives information on the roadway design type (urban, sub-urban, and rural). This information will later be used to combine continuous roadway segments.

5.1.2 RCI

The RCI has all the essential traffic and geometric information pertaining to State maintained roadways. A detailed list of features and characteristics available for the roadways are given in the RCI Office Handbook (FDOT, 2007). Among the features that are useful for the present study are: functional classification, curvature of roadway segments, type of intersections, etc. These characteristics will be integrated with CAR variables for analyses. The RCI data can be downloaded from the FDOT mainframe. The RCI database has 107 roadway characteristics for each roadway segment.

Apart from these characteristics the RCI website also provides a plethora of reports with information that could be used in specific ways. For the present work, lists of roadway segments that are part of the multilane arterial segments were downloaded. In addition to it, the entire list of signalized intersections was also retrieved from the RCI website.

5.2 Data Preparation

At the outset it was critical to have a definition of corridor. The FDOT does not have an exact definition of corridor. Hence it was critical that we begin the analysis by defining a corridor. There were a lot of parameters on which we could have defined a corridor. But the defining parameter should be able to make the corridors homogenous in one way or the other. A representative state road is comprised of different roadway segments which are typically representation of administrative boundaries. Any change in the administrative boundary is bound

to affect the lengths of these corridor lengths. Hence the choice of using these managerial roadway segments is ruled out as the homogeneity will not be consistent. The other choices on which we could start off was the median type. There are essentially two types of physical medians: 1) Divided and 2) Un-Divided. Very large number corridors that resulted were less than 1 mile in length. This is essentially because as an arterial winds it way through the geographical area, cutting across various residential areas, the median type changes very frequently. Hence the very large number of smaller corridors. Though the method could provide a homogenous section, it was unacceptable because of the above mentioned reason. The other parameter that could be used was the design type. The roadway design of arterials is essentially of three types: 1) Urban; 2) Sub-Urban and 3) Rural. The features that distinguish these three types are the drainage type and the city limits. The urban roads have a curb and gutter design within city limits or urban residential areas. Roads with open drainage but within city limits are categorized as sub-urban. However roads with open drainage and outside the city limits are categorized as rural. The resulting corridors resulted in more number of longer homogenous sections. However a large number of roads were still less than a mile long. Hence a refinement was made based on the design and the city limits. The roadways with urban/ sub-urban design were combined together, thus giving rise to section within city limits. The rural roads, outside the city limits, were then combined together. Number of sections with length less than one mile reduced. These were later removed from further analysis. The reasons to drop these very short length sections were twofold: 1) the sectional characteristics will not change much for such short lengths; 2) the total number of severe crashes for most of those corridor sections was too few.

5.2.1 Clustering

With the corridors now grouped according to roadway design and city boundaries, the next task is analysis. The corridor lengths varied from 1 mile to 78 miles. The wide variation in the length justifies the clustering of the corridors based on the length itself. Corridors with similar length are more likely to have similar properties. The variations will be similar. Or in other words the heterogeneity will be minimized.

At the outset we need to the optimum number of clusters, one of the more difficult tasks in cluster analysis, to which the corridors have to be grouped into. In the present study we have used the partitioning around medoids (PAM) algorithm to find the optimum number of clusters. PAM algorithm operates on the average dissimilarity. According to Kaufman and Rousseeuw (1990) the 'medoid' is an object of the cluster whose average dissimilarity to all the objects in the cluster is minimal. Once the medoids are identified all the objects are assigned to the nearest medoid. The objective function is the sum of the dissimilarities of all the objects to the nearest medoid. The algorithm terminates when the interchange of an unselected object with an already selected object no longer minimizes the objective function. To find the optimum number of clusters and also to differentiate a bad cluster from a good one, a set of values called 'silhouettes' are computed (UNESCO, 2007). The following algorithm shows how one would calculate the silhouette value.

Consider any object 'k' in the data and let it be assigned to a cluster 'X'. Let 'x(k)' be the average dissimilarity of the object 'k' to all other objects in cluster X. For any other cluster 'Y'

different from ‘X’, ‘d(k, Y)’ be defined which is the average dissimilarity of object ‘k’ to all objects in ‘Y’. ‘d(k, Y)’ for all clusters ‘Y’ not equal to ‘X’ is computed and the smallest is computed. If the minimum is attained in cluster ‘Z’ then ‘d(k, Z) = z(k) and ‘Z’ is the neighbor of object ‘k’. The silhouette value, ‘s(k)’ is then defined as:

Equation 5-1

$$s(k) = (z(k) - x(k)) / \min(z(k), x(k))$$

A silhouette value close to 1 suggests that in-cluster dissimilarity is less than the between dissimilarity. A value of 0 suggests that the object could have belonged to either cluster. Negative silhouette values, especially those that are close to -1, suggest that the clustering has been poorly done. The silhouette values computed then can be used to find the optimal number of clusters. In the present study the optimal number of clusters found by using the PAM algorithm was 4. Once the optimal number of clusters has been defined, we move on to the actual clustering. The clusters found are given in Table 5-1.

Table 5-1 Cluster and respective Range

Cluster	Range (in Miles)
1	1.009 – 2.89
2	2.898 – 5.729
3	5.762 – 10.556
4	10.644 – 78.293

CHAPTER 6. USING CONDITIONAL INFERENCE FORESTS TO IDENTIFY THE FACTORS AFFECTING CRASH SEVERITY ON ARTERIAL CORRIDORS

6.1 Introduction

Approaches to safety on multilane corridors have traditionally been twofold. Brown and Tarko (1999), Abdel-Aty and Radwan (2000) and Rees (2007) treated the corridors in totality; while Milton and Mannering (1998) and Miaou and Song (2005) divided the corridors into segments and intersections. Abdel-Aty and Wang (2006) have shown a spatial correlation between crash patterns of successive signalized intersections, which may be attributed to the characteristics of the segments joining them.

Though both approaches have worked well for investigation purposes, the issue that still remains is how to assign crashes to the segments and the intersections. There is no uniformity in the influence area of an intersection among the states. For example, in Florida, all the crashes occurring within 250 ft. from the center of an intersection are categorized as intersection related crashes, as has been reported by Abdel-Aty and Wang (2006) and Wang et al. (2006). Recently Das et al. (2008) showed that proximity only is not the best way to assign crashes. Wang et al. (2008) used frequency modeling for crashes with fixed as well as varying influence distance and found different set of significant factors. Apart from the above research, it is also of common knowledge that the way the crashes are reported varies among different administrative units. The

author investigated several crash reports and came up with an innovative approach to assign crashes, the details of which are given in CHAPTER 4.

As previously mentioned, it is important not only to find the contributing factors but also to improve on the methodology adopted. Pande and Abdel-Aty (2008) in their work on association rules point out that data mining techniques remain underutilized for analysis of crash. The underutilization is especially noteworthy since most studies use observational data collected outside the purview of an experimental design. Simple data mining tools like classification and regression trees have traditionally been used to identify variables of importance in safety studies (Pande and Abdel-Aty (2008)). A decision tree, with all its simplicity and handling of missing values, can be very unstable. However, if instead of one tree, an ensemble of trees (commonly referred as forest) is used, the outputs become much more stable. The robustness of the forests makes them a better choice than the use of single trees. In this regard, Random Forests, developed using the Classification and Regression Trees (CART) algorithm, have been used by the authors (Abdel-Aty et al. (2008)) recently to identify variables of significance and then develop neural network classifiers. However, the method has been shown to have selection bias as shown by Strob et al. (2007). The selection bias is in favor of variables which are continuous or have higher number of categories. At the root of this selection bias is the application of 'Gini' index criterion to split a node (while building the tree) as well as for variable selection (generally based on the frequency a variable was chosen for the split). Details of the 'Gini' index criterion and the resulting bias have been provided in the 'Modeling Methodology' section. Hence, in this study conditional inference trees, developed by Hothorn et al. (2006), and their forests have been

used for the purpose of variable selection. The author is of the belief that the application of this new methodology will improve traffic safety research. Details of how this algorithm is different (and better suited for the application at hand) than the CART have been given in the methodology section.

The author included new variables like 'element', in this study, which assigns crashes to segments, intersections or access points based on the information from site location, traffic control and presence of signals. The author identified roadway locations where severe crashes tend to occur. Failures to use safety equipment by all passengers and presence of driver/passenger in the vulnerable age group (more than 55 years or less than 3 years) were also other new variables that were included in the data. The details of how the inclusion helped in a better understanding of the severity aspect has been discussed in the 'Analysis and Results' section later on in the chapter.

Crash data from the high-speed multilane arterials with partial access control in Florida have been collected. These arterials have been divided into groups based on their lengths and roadway design standards (urban/suburban and rural). The following section will focus on the details of the data collection and aggregation. It is followed by the methodology section where conditional inference trees and forests will be discussed. The results and analysis section will explain the results from the conditional inference trees and the forests. While the random forests provide a more robust set of variables associated with severe/fatal crashes; individual tree helps in making relevant inferences about the relationship.

6.2 Data Collection and Preparation

The crash data available were from the Crash Analysis and Reporting (CAR) system of the Florida Department of Transportation (FDOT). The Roadway Characteristics and Inventory (RCI) data was also made available to us through the FDOT. The data used are for the years 2004 through 2006 for all the state roads of Florida. The datasets have information regarding traffic, roadway geometric and driver related factors. The datasets were merged and the parameters were modified to suit the data mining methodology being implemented in the study. As mentioned in CHAPTER 5, the corridors were grouped in four clusters (see Table 5-1).

Different types of crashes occur on the corridors and the contributing causes for the different types also vary. Even though the overall safety of the corridor is being analyzed, the approach to investigate different crash types separately would shed more light. The crashes were grouped into 6 major types as follows: i) angle/ turning movement; ii) rear-end; iii) head-on; iv) sideswipe; and v) crashes involving single vehicles.

The conditional inference trees used in this study helps us in identifying the contributing factors associated with the severity of the crashes that occurred along a corridor. However too many parameters lessen the discriminating ability of the models as the overall degrees of freedom available for the model development decrease. Hence only a subset of the available factors should be chosen for model development. Milton et al. (2008) have also pointed out that event specific variables are least desirable in developing injury severity models. Hence for the analysis a few variables were chosen based on engineering judgment and taking into consideration that

event specific factors are not in use to a relatively large extent. The variables were broadly based on two different categories: 1) environmental and road geometric factors; 2) driver and vehicle related factors. The variables used in the study are described in Table 6-1. They have been derived directly from the datasets or a combination of parameters. Both these sets of parameters have their application values.

Table 6-1 Dependent / Independent Variables used for Conditional Inference Tree / Forest Analyses

Variable Name	Variable Description	Urban / Sub-urban
Target or Dependent Variable		
Sev	Severity	Binary (1 = incapacitating injuries/ fatalities; 2 = possible/ non-incapacitating injuries)
Environmental and Roadway Geometric Parameters		
pavecond	Pavement condition	4 levels (poor, fair, good and very good)
surf_type	Type of surface	Binary (1 = black top surface; 2 = other)
surface_width	Surface width	Continuous
shld_t	Type of shoulder	Binary (1 = paved; 2 = unpaved)
max_speed	Maximum posted speed limit	Continuous
park	Presence of parking	Binary (1 = no; 2 = yes)
skid_f	Friction resistance	Skid <= 34 34 < skid <= 38 Skid > 38
median	Types of median	9 levels (0 = no median; 1 = painted; 2 = median curb <= 6”; 3 = median curb > 6”; 4 = lawn; 5 = paved; 6 = curb <= 6” and lawn; 7 = curb> 6” and lawn; 8 = other)
ACMANCLS_num	Type of median openings	7 levels (0 = no median opening; 2 = restrictive opening w/ service roads; 3 = restrictive median; 4 = non restrictive median; 5 = restrictive median with shorter directional openings; 6 = non restrictive median with shorter signal connection; 7 = both restrictive and non-restrictive median types)
road_cond	Road condition at time of crash	Binary (1 = no defects; 2 = defects)
vision	Vision obstruction	Binary (1 = no; 2 = yes)
shld_side	Shoulder + sidewalk width	Continuous
curvclass	Horizontal degree of curvature	6 levels (curve < 4’; 4 <= curve <= 5’; 5 < curve <= 8’; 8 < curve <= 13’; 13 < curve <= 27’; curve > 27’
surf_cond	Surface condition	Binary (1 = dry; 2 = other)

light	Daylight condition	Binary (1 = daylight; 2 = other)
ADT	Annual daily traffic	ADT <= 31000 31000 < ADT <= 40000 40000 < ADT <= 52500 ADT > 52500
t_fact	Average truck factor	t_fact <= 4.05 4.05 < t_fact <= 5.895 t_fact > 5.895
k_fact	Average k - factor	k_fact <= 9.85 k_fact > 9.85
dayandtime	Combination of the day of week and time of day	Afternoon Peak Weekday Morning Peak Weekday Friday or Saturday Night Off-peak
trfcway	Vertical curvature	Binary (1 = level; 2 = upgrade/ downgrade)
element/ element 1	Assignment of crashes to roadway elements	Ternary (1 = segment; 2 = intersections; 3 = access points) / Binary (1 = segments/ access points; 2 = intersections)
LIGHTCDE	Street lighting	Ternary (Y = full lighting; N = no lighting; P = partial lighting)
Driver and Vehicle related Parameters		
age_gr	Age group of the at fault driver	Age < = 25; 25 < age <= 35; 35 < age <= 45; 45 < age <= 55; 55 < age <= 65; 65 < age <= 75; Age > 75
veh_type1	At-fault type of vehicle	4 levels (1 = automobiles; 2 = light trucks; 3 = heavy vehicles; 4 = light slow moving vehicles)
alcohol_use	Alcohol/ drug use of the at-fault driver	3 level (1 = no use; 2 = use; 3 = no information)
vuln_age	Presence of vulnerable age group passengers in the vehicle (age < 5 or age > 55)	Binary (1 = yes; 2 = no)
more	Presence of more than 5 passengers inside either of the involved vehicles	Binary (Y = yes; N = no)
sfty	Use of safety equipment in the vehicle by driver/passengers	Binary(1 = yes; 2 = no)
gender	Gender of the at-fault driver(s)	3 levels (1 = male; 2 = female; 3 = both)
veh_move1	Vehicle movement of the at-fault vehicle	4 levels (1 = straight ahead; 2 = turning movements; 3 = changing lanes; 4 = other)

The variables illustrated in Table 6-1 are mostly derived from the RCI database. Many variables have too many categories, in the raw form, to start off with. Hence, level reduction in variables is not only critical but also simplifies the model and makes them more readily explainable. For

example, vehicle movement, vehicle type, roadway conditions, vision obstruction, surface condition, surface type, and type of median are some of the variables with many categories. For example, the proposed methodology (conditional inference trees/forests) uses Chi-square test statistic to identify the relationship between a particular parameter and target variable. Each category of the variable should have sufficient number of observations in the contingency table for the Chi-square to be evaluated as discussed by Das et al. (2008). Continuous variables like ADT, Percentage of trucks, and K-factor (design hour volume as a percentage of ADT) and skid (friction resistance multiplied by a factor of 100) were also divided into categories. Their relationships with severe/fatal crash occurrence may not be monotonous in nature. Time of crash, along with day of week, were combined into one variable representing day of week *and* time of day. The weekend night times were not treated as off peak hours as there may be higher instances of alcohol impaired driving.

Traditionally the site location variable has been used by researchers to assign crashes to the three roadway elements (segments, intersections and access points). However a detailed review of several hundred crash reports, suggested that the 'site location' variable by itself was a weak indicator for the same. For example, it was observed that it is possible for a crash to be not attributed to a signalized intersection even if it may have occurred very close to one. In fact, 'traffic control' in combination with the 'site location' along with the information of the presence or absence of signal, did a superior job in attributing crashes to one of the three roadway elements (see CHAPTER 4). Based on these three independent parameters, a variable 'element' was created to assign the crashes to the three roadway elements, namely segments, intersections

and access points. However it was also observed and verified through the study of crash reports that distributing crashes to the three roadway elements works fine with all types of crashes except for the angle / turning related crashes. Most of such crashes occur at the signalized intersections. The crashes which occur on the segments were observed to have occurred mostly on auxiliary lanes (right / left turning lanes). Hence these could be either way attributed to the segment or access points. Therefore for angle / turning related crashes the ternary variable 'element' takes the form of binary 'element1' where the crashes either belong to the signalized intersection or to segment/ access points. This new variable appears in certain tree results (developed along with conditional inference forests for relevant inference) and also positively contributes to model development in the forests.

Zhang et al. (2000) found the non-use of seat belt to increase the risk of severe injuries. In this study, the parameter for safety equipment in use is for all the passengers. This is different from the traditional approach as it is more useful to look at the overall safety of all the passengers rather than just focusing on the safety equipment use of the driver. The importance lies in the fact that there a lot of crashes in which the drivers may not be injured at all. The vulnerable age group binary variable points out the presence of children or elderly passengers inside the vehicle. The physical fragility of the people belonging to these age groups described in Table 6-1 makes it an interesting variable and the results also show interesting pattern related to severity.

The median types were combined into 9 levels. It does the two fold job of not only giving a sense of the median obstruction imposed but also gives an idea as to how far apart the opposing

directional roads could be. The author observed that the median width was a variable that is really dependent on the median type. Hence the median width was sufficiently represented within the variable median type. A new variable called 'shld_side' has been created which simply represents the total width of the outside shoulder and the sidewalk. This variable gives more realistic idea of the side space available for the vehicles traveling in the outer lane, especially in the urban areas where the shoulder width sometimes is negligible as compared to those available in rural settings. Hence, the original information on shoulder width and the sidewalk width were replaced with this new variable.

The target variable of severity is binary. The first level represents fatalities and incapacitating injuries. They are combined into one level for two reasons; first, the relatively small frequency of fatal crashes compared to other injury severity levels. For example, the Chi-square tests may not be valid due to low expected cell-frequency. The second reason is that the crashes that involve incapacitating injury could easily have been fatal and vice-versa possibly due to vulnerability of the subjects involved (Das et al. (2008)). The second level includes crashes with possible injuries and non-incapacitating injuries. The crashes with no injuries were not included as these are likely expected to be incomplete. This issue has been well investigated and documented by Abdel-Aty and Keller (2005). Yamamoto et al. (2008) also have discussed the issue of possible under reporting of such crashes and the bias resulting from it. Hence, in the present study the authors have included those crashes with injury severity level of at least a possible injury or higher.

It should be noted here that the conditional inference forests, which have been used to calculate the variable importance score, do not accept missing values. Hence, the data set has no missing data. Hence the introduction of random parameters to account for missing data, as done by Milton et al. (2008), is not required in this study. As mentioned earlier the crashes have been grouped into 5 types, namely: i) angle/ turning movement; ii) rear-end; iii) head-on; iv) sideswipe; and v) crashes involving single vehicles. The numbers of crashes in each of the crash categories are 6231, 5532, 1261, 2204 and 2404, respectively, for the models developed for environmental and roadway geometric factors. Where as for the models developed for driver and vehicle related factors the number of crashes are 7759, 6775, 1583, 2612 and 2879, respectively. As no missing data record could be used, the records deleted for the environmental and roadway geometric factors' models are 31973. This accounts for 6.6% of the three years of Florida crash data used. Similarly for the driver and vehicle related factors' models the crash records that were deleted were 27997 which accounts for 5.8% of the three years of Florida crash data used.

6.3 Modeling Methodology

6.3.1 Conditional Inference Tree

Traditionally classification trees (Breiman et al. 1984) have been used to determine variable of importance in most transportation studies. Decision trees are tree-shaped structures representing sets of decision which self-generate (as opposed of being dictated) rules for the classification of a dataset (as opposed to a sample), in a hierarchical order, using algorithms such as ID3 and its

improvements C4.5 and C5.0, as well as CART and CHAID (Quinlan, 1986; Quinlan, 1993). No assumptions are made about the distribution of data.

The modeling approach adopted here in is the conditional inference trees and the forests developed there from. The focus of the study is to find out parameters that are related to the injury severity. The trees not only give the variables of importance but also help us in better interpretation of the results. Especially in severity analysis the advantage in using trees is that it helps us determine the values of parameters which contribute more to the severity of crashes. Hence from a safety aspect this is critical as it can help determine what changes need to be made in the design and/ or policies to improve the safety. Conventional classification and regression trees have always been used to select variables of importance. According to Strob et al. (2007), the CART trees have a variable selection bias towards variables which are continuous or with higher number of categories. The most common splitting criterion in the CART tree is the Gini Index to find a favorable split. The Gini Index checks for the purity of the resulting "daughter" nodes in the tree. According to Breiman et al. (1984), for a given node't' with estimated class probabilities 'p(j|t)', j = 1, ….., J, the node impurity 'i(t)' is given by:

Equation 6-1

$$i(t) = \Phi(p(1 \mid t), ..., p(J \mid t))$$

A search is made for the most favorable split, one that reduces the node or equivalently tree impurity. If the adopted form is Gini diversity index then 'i(t)' takes up the form:

Equation 6-2

$$i(t) = \sum_{j \neq i} p(j \mid t) p(i \mid t)$$

The Gini index considered as a function '$\Phi(p_1,\ldots\ldots,p_J)$' of the $p_1,\ldots\ldots,p_J$ is a quadratic polynomial with nonnegative coefficients. Therefore for any split 's', '$\delta(s, t) >= 0$'. Since the criteria looks for a favorable split, the chances to find a good split increases if the variable is continuous or has more categories. Therefore even if the variable is not informative, it could sit higher up on the tree's hierarchical structure. Hence, in this study conditional inference trees (Hothorn et al. (2006)) have been used, where the node split is selected based on how good the association is. The resulting node should have a higher association with the observed value of the dependent variable. The conditional inference tree uses a chi-square test statistic to test the association. Therefore, it not only removes the bias due to categories but also chooses those variables which are informative.

The key to this recent algorithm is the separation of variable selection and splitting procedure. The recursive binary partitioning which is the basis of the framework is given below.

The response '**Y**' comes from sample space 'Y', which may be multivariate. The m-dimensional covariate vector $\mathbf{X} = (X_1,\ldots,X_m)$ is taken from a sample space $\mathbf{X} = X_1,*\ldots\ldots*X_m$. Both the response variable and the dependent variables may be measured at any arbitrary scale. The conditional distribution of the response variable given the covariates depends on the function of the covariates.

Equation 6-3

$$D(Y \mid X) = D(Y \mid X_1,.....,X_m) = D(Y \mid f(X_1,.....,X_m)$$

For a given learning sample of 'n' iid observations a generic algorithm can be formulated using nonnegative integer valued case weights $\mathbf{w} = (w_1,\ldots,w_n)$. Each node of a tree is represented by a vector of case weights having nonzero elements when the corresponding observations are elements of the node and are zero otherwise. The generic algorithm is given below:

1. For case weights **w** the global null hypothesis of independence between any of the covariates and the response is tested. The step terminates if the hypothesis cannot be rejected at a pre-specified nominal level 'α'. Otherwise the j^{th} covariate X_j with the strongest association to the response variable is selected.
2. Set $A \subset X_j$, is chosen to split X_j, into two disjoint sets. The case weights $\mathbf{w}_{left}$ and $\mathbf{w}_{right}$ determine the two subgroups with $w_{left,i} = w_i I(X_{ji} \in A)$ and $w_{right,i} = w_i I(X_{ji} \notin A)$ for all $i = 1,\ldots,n$ and $I(\ )$ denotes the indicator function, which indicates the membership of an element in a subset.
3. Recursively repeat the steps 1 and 2 with modified case weights $\mathbf{w}_{left}$ and $\mathbf{w}_{right}$, respectively.

The separation of variable selection and splitting procedure is essential for the development of trees with no tendency towards covariates with many possible splits. For more details of the algorithm the readers are directed to the paper by Hothorn et al. (2006).

6.3.2 Conditional Inference Forest

Forests which are a collection of multiple tree classifiers are used for variable selection. A decision tree, with all its simplicity and handling of missing values, can be very unstable. In other words, small changes in the input variables might result in large changes in the output. In this regard, forests are more robust variable selection tool. Random Forests' algorithm was developed by Breiman (2001) which works in the framework of the classification and regression trees, but instead of having one tree, they have multiple trees. The forests are most important in calculating the variable importance measure. Recent works in transportation by Abdel-Aty et al. (2008) and Harb et al. (2009) used the random forests algorithm to determine the variables of importance. However Strobl et al. (2007) showed that the bootstrapping method (sampling with replacement) and the use of Gini index results in the biased selection of variables of importance. The Gini index shows a strong preference for variables with many categories or for the ones which are continuous. Variables with more potential cut off points are more likely to produce a good criterion value by chance. This variable selection bias which occurs in each individual tree also has an effect on the variable importance measure. In the previous sub section it was mentioned that the algorithm for recursive binary partitioning uses the association tests like chi-square test to select informative variables. Therefore bootstrap sampling with replacement induces bias because the cell counts in the contingency table are affected by observations that are either not included or are multiplied in the bootstrap sample. Hence the forests that we have used in this study comprise of the trees that have developed in the conditional inference framework. The next subsection describes the variable importance computation process.

6.3.3 Variable Importance

The basis of the variable importance in forests is as follows. By first randomly permuting the predictor variable X_j, the original association with the response variable Y is broken. When the permuted variable along with other non-permuted variables is used to predict the response for the out-of-bag observations the classification accuracy decreases substantially if the permuted variable is associated with the response. Hence the variable importance of a variable is the difference in the prediction accuracy before and after permutation of the variable X_j, averaged over all trees. Out-of-bag observations are those that the method excluded while developing the trees. They form an internal test data set and there is no need to allocate a test data set separately. Let $B^{(t)}$ be the out-of-bag sample for a tree't', with $t \in \{1,.....,ntree\}$. The variable importance of one tree is then given by the following:

Equation 6-4

$$VI^{(t)}(x_j) = \frac{\sum_{i \in B^{(t)}} I(y_i = \hat{y_i}^{(t)})}{|B^{(t)}|} - \frac{\sum_{i \in B^{(t)}} I(y_i = \hat{y}_{i,\pi_j}^{(t)})}{|B^{(t)}|}$$

Where $\hat{y}^{(t)} = f^{(t)}(x_i)$ is the predicted classes for observation 'i' before and

$\hat{y}_{i,\pi_j}^{(t)} = f^{(t)}(x_{i,\pi_j})$ is the predicted classes for observation 'i' after permuting its value of variable. The raw variable importance score for each variable is then computed as the mean importance over all trees and is given by:

Equation 6-5

$$VI(x_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(x_j)}{ntree}$$

Since the individual importance scores $VI^{(t)}(x_j)$ are computed from 'ntree' independent bootstrap samples, a simple test for the relevance of variable X_j can be constructed based on the central limit theorem for the mean importance of $VI^{(t)}(x_j)$. If individual importance has a standard deviation σ, then the mean importance from 'ntree' replications has a standard error of $\sigma / \sqrt{ntree}$.

The next section emphasizes on the results of the random forests results for the various severity models developed on the urban/sub-urban and rural corridors according to the various crash types.

6.4 Analyses and Results

6.4.1 Conditional Inference Forest Variable Importance Results

This section deals with the results of conditional inference forests which typically illustrate the variables of importance. In the present study the conditional inference forests generated for the models, with the binary severity variable as the target, gives the variable importance score for all the variables in the model. The sign (positive/ negative) of the importance score indicates whether the presence or absence of a variable in the model will improve or degrade the

efficiency of the model to exemplify the variable importance score the authors tabulate the results for a particular cluster (in this case Cluster 3 for angle/ turning movement crashes) in Table 6-2 and Table 6-3. As mentioned earlier in the section 6.2 Data Collection and Preparation, the variables have been categorized into two. Hence for each cluster and crash type two models had been developed, one for the environmental and roadway geometric and the other for driver and vehicle related characteristics. Results in Table 6-2 are for the model with only environmental and roadway geometric factors and those in Table 6-3 are for the driver and vehicle related characteristics' model. As a reminder to the readers, Table 6-1 has the explanation of the variables.

It should be noted that Table 6-2 and Table 6-3 are examples of the output of a condition inference forests. The variables with a positive variable importance score are the most important for the severity model developed here in the example. Their association with the target variable is the maximum and their absence would decrease the model performance. The variables with zero importance score are believed to have no effect on the model performance, while the ones with negative importance (as highlighted in Table 6-2) are the ones decreasing the model performance. Readers may note that the variable *LIGHTCDE* has not been included in the Table 6-2. The variable was not included in the particular model as it had only one level and can not be used for split during tree development. The same is the reason for *no information* on *LIGHTCDE* in some of the cells of Table 6-4 as well.

It is critical to distinguish the significant from non significant. As the dataset change, i.e. a new model is being developed, the importance score may also change. A particular variable may improve the model efficiency in one group where as it may decrease in another group, while being neutral in some other. All the conditional inference forests results were developed at 90% confidence level.

Table 6-2 Conditional Inference Forest sample result for environmental and roadway geometric factors

Variable Name	Variable Importance Score
Shoulder + Side	0.000358
Pavement condition	0.00026
Median Openings	0.000163
Median type	0.000163
Truck factor	0.00013
Vision obstruction	6.50E-05
Skid (friction resistance)	6.50E-05
Roadway condition	0
Horizontal Degree of Curvature	0
Surface condition	0
Parking type	0
Traffic-way character	0
Surface width	-9.76E-05
K factor	-6.50E-05
Day of the week and time of the day	-6.50E-05
Surface type	-3.25E-05
Daylight condition	-3.25E-05
Roadway element	-0.00013
Maximum posted speed limit	-0.00026
ADT	-0.00029
Shoulder type	-0.00036

Table 6-3 Conditional Inference Forest sample result for driver and vehicle related factors

Variable Name	Variable Importance Score
Alcohol usage	0.004544
Age group	0.004488
Vehicle movement	0.000309
Safety equipment use	0.00014
Vehicle type	5.61E-05
At fault driver gender	2.81E-05
Vulnerable age group	2.81E-05
Presence of more than 5 persons	0

Table 6-4 and Table 6-5 tabulate the conditional inference forests results developed for all severity models in the study. For certain types of crashes (namely: head-on; sideswipe; single vehicle involved; slow moving vehicles involved) the number of crashes in the urban clusters 1 and 2 were not sufficient for the trees to develop. Hence for these types of crashes the clusters 1 and 2 were combined. All the results were developed with the use of the statistical software package 'R'. The package 'party' developed by Hothorn et al. (2008) was used to generate the conditional trees and forests results. Key for Table 6-4 and Table 6-5 is:

'+': variables which increase the model efficiency,
'-': variables which decrease the model efficiency,
'0': variables which are neutral to model efficiency.

It should be noted that in Tables 3(a) and (b) there could be certain blank cells, i.e. they do not have any of the three symbols mentioned above. For example, the variable *LIGHTCDE* do not appear in the Table 3(a) in a number of cells. The reason for the exclusion is that the variable was not used for that particular model development, as it had only one level.

Table 6-4 Severity models' Conditional Inference Forest results for urban clusters with environmental and roadway geometric factors

	Cluster 1		Cluster 2		Cluster 1 & 2				Cluster 3						Cluster 4					
variable	angle	rearend	angle	rearend	headon	sideswipe	single	slow	angle	rearend	headon	sideswipe	single	slow	angle	rearend	headon	sideswipe	single	slow
surface_width	+	0	-	-	0	0	-	+	-	0	+	0	-	+	+	-	+	0	+	+
max_speed	-	0	+	-	-	0	-	-	-	+	-	0	-	-	+	0	0	0	-	+
LIGHTCDE	+	0			0	0	0	0							0	0	0	-	0	0
ACMANCLS_num	+	0	+	0	+	-	-	0	+	0	-	0	-	0	+	-	+	0	0	+
road_cond	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	0	0	0
vision	0	0	0	0	0	0	0	0	+	0	+	0	0	+	+	0	-	-	0	0
shld_side	+	0	+	0	0	+	-	-	+	0	0	0	0	+	+	-	0	0	+	+
curvclass	-	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0
surf_cond	+	0	-	0	-	-	-	-	0	0	-	0	-	+	+	0	0	0	0	-
light	-	0	0	0	0	0	+	0	-	0	0	0	-	+	0	0	0	+	0	+
ADT	+	0	+	+	-	0	-	+	-	+	+	0	-	+	+	-	0	-	-	+
t_fact	+	0	-	0	-	0	-	+	+	+	-	0	-	+	-	+	+	0	+	+
k_fact	+	0	+	0	-	-	0	0	-	0	-	0	+	-	+	0	0	0	-	+
dayandtime	-	0	+	0	0	-	+	0	-	-	-	0	-	0	+	-	0	0	+	+
trfcway	0	0	-	-	0	0	+	0	0	0	+	0	-	+	+	0	0	0	0	+
pavecond	+	0	-	0	+	0	-	+	+	0	0	0	+	0	+	-	+	-	+	-
park	+	0	0	0	0	0	0	0	0	0	0	0	0	-	0	-	0	0	0	+
surf_type	0	0	-	0	-	0	0	0	-	0	0	0	0	+	0	+	0	0	0	+
skid_f	+	0	+	0	+	0	+	+	+	-	+	0	0	+	+	0	+	0	-	+
median	+	0	+	-	+	0	0	+	+	0	+	0	-	+	+	+	+	-	+	+
element1	+	0	+	0	-	0	-	0	-	+	+	0	+	+	-	+	+	0	-	+
shld_t	+	0	-	0	-	0	+	0	-	0	-	0	-	+	+	+	0	-	-	+

Table 6-5 Severity models' Conditional Inference Forest results for urban clusters with driver and vehicle related factors

	Cluster 1		Cluster 2		Cluster 1 & 2				Cluster 3						Cluster 4					
variable	angle	rearend	angle	rearend	headon	sideswipe	single	slow	angle	rearend	headon	sideswipe	single	slow	angle	rearend	headon	sideswipe	single	slow
age_gr	-	0	+	0	-	0	-	+	+	+	+	-	-	+	+	0	+	+	-	+
veh_type1	+	-	+	0	0	0	+	-	+	0	+	-	+	+	+	0	-	-	+	+
alcohol_use	-	-	+	0	0	0	+	0	+	-	+	0	+	+	+	0	+	0	+	-
more	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sfty	+	0	+	0	0	0	+	+	+	0	-	0	+	+	-	0	0	0	-	+
gender	+	0	+	0	+	0	-	0	0	+	+	0	+	-	-	0	-	+	+	0
veh_move1	-	0	+	0	-	0	+	+	-	+	0	-	-	+	+	+	-	-	+	+
vuln_age	0	0	+	0	0	0	0	0	-	0	0	0	-	0	+	0	0	0	0	-

As mentioned earlier, the variables with "+" sign in the boxes are the variables with higher importance, i.e. they improve the model efficiency more than the other variables for the given model. The ones with "0" means they are neutral for the severity model. The variables with "-" are the ones with least effect on the corresponding model. It must however be understood that the "+" sign need not necessarily mean that the variable is positively associated with severity. For better interpretation of the variable's influence on the severity single conditional inference trees were developed for the models. And depending on how the variables split, the approach to severe/fatal crashes would be clearer. The next subsection deals with the individual conditional inference tree results.

6.4.2 Conditional Inference Tree Results

6.4.2.1 Example of Conditional Inference Tree and how to interpret them

The conditional inference trees are critical to observe which parameters are related more to severity and also how they are related. Before we move to the details of the results the authors would like to exemplify certain individual conditional inference tree results through Figure 6-1 and Figure 6-2. The trees shown in the figures are for angle/ turning movement crashes in Cluster 1. Figure 6-1 represents the tree model for environmental and roadway geometric factors, where as Figure 6-2 is the model for driver and vehicle related factors.

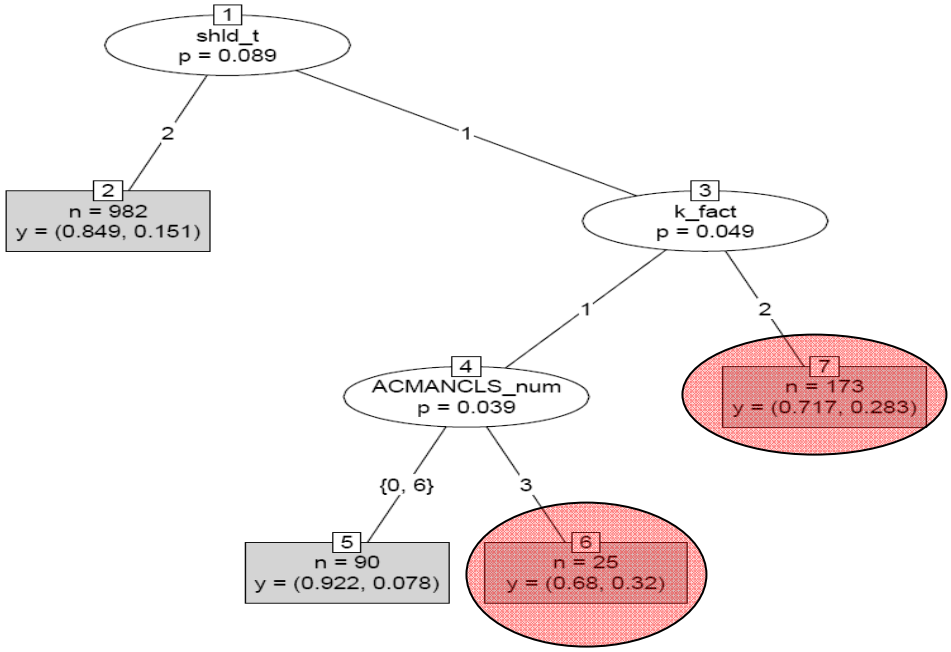


Figure 6-1 Conditional Inference Tree sample result for environmental and roadway geometric factors

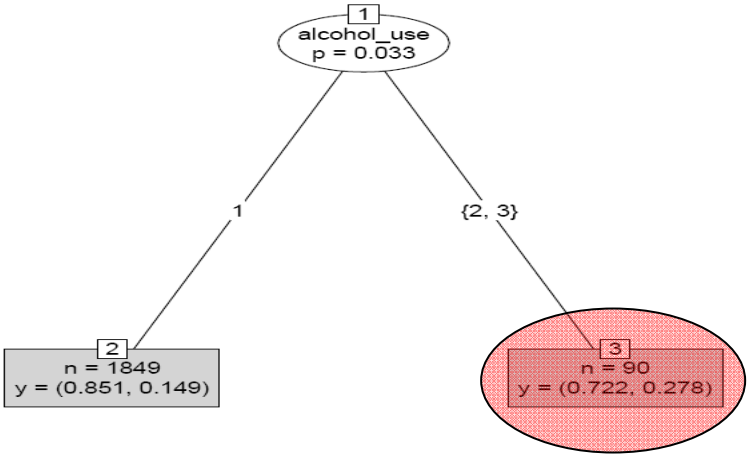


Figure 6-2 Conditional Inference Tree sample result for driver and vehicle related factors

All the trees were developed at 90% confidence level. The *p* value in the nodes of Figure 6-1 and Figure 6-2 denotes the actual significance level at which the split has taken place. All the nodes

are shown in white oval shape whereas all the terminal nodes (leaves) are shown in the rectangular boxes. The small square boxes with numbers on both the ovals and rectangles denote a unique numerical representation of the node or leaf. In the white oval shapes the variables mentioned is the split variable and the *p* value denotes the significance level. The numbers on the lines connecting the nodes to other nodes or leaves denotes the specific categories of variables or range of values of variables which lead to the extension of that particular branch of the tree. For example, in Figure 6-2 the variable *alcohol_use* splits the node and all the cases of the variable taking up the value 1 (denoting no alcohol/ drug use) leads to the leaf, which is uniquely numbered as '2'. For the other branch the variable either takes the value 2 or 3 (denoting alcohol use or pending test results) to reach the other leaf, uniquely numbered as '3'. The general direction of flow of the lines in any conditional inference tree is top to bottom. It goes from one node to other node/ leaf. As can be observed the leaf contains the information about the number of cases in the particular leaf, denoted by *n.* The proportion of non-severe and severe crashes is also shown in the leaf, through the numbers given by *y.* To exemplify, the authors again refer to Figure 6-2. The leaf, uniquely denoted by '2' has $n = 1849$ cases, where as the proportion of non-severe crashes was 0.851 while that of severe crashes was 0.149 (denoted by $y = (0.851, 0.149)$). As can be observed from Figure 6-1 and Figure 6-2, there are red ovals covering certain leaves. These leaves have higher proportion of severe crashes than the proportion of severe crashes in the particular dataset from which the model had been developed. The path taken from the original parent node to the particular leaf thus gives us the conditions which lead to higher severity. The variables on the path, on which the splits have been done, reflect which variables are associated with severity.

From here on the results will be based on crash type and the relevant results from different clusters will be grouped together. The explanation will include both the categories of models developed, namely: 1) environmental and roadway geometric and 2) driver and vehicle related. The order will be adhered to for most part of the explanation.

6.4.2.2 Angle / Turning Movement Crashes

As mentioned earlier the corridors in Cluster 1 (1.009 – 2.89 miles) are the smallest in length. According to the environmental and roadway geometric model for angle/ turning movement crashes occurring in this particular cluster's corridors, the severity is higher where the shoulders are paved and the k-factor is higher. Even though paved shoulders leading to higher severity seems counterintuitive, the only reason could be that better shoulders may be misused as additional lanes for dangerous maneuvers. The higher k-factor indicates that, higher the peak hour volume the higher risk it involves for angle/turning movement crashes. With lower k-factor but restrictive medians (with longer distance between openings), the severity of the crashes is found to be higher. Since angle/ turning movement crashes mostly occur at intersections, it is interesting to note that Levinson (2000) pointed that even though restrictive medians provided better separation of traffic and better pedestrian safety, however adequate provisions have to be made for left and U turns to avoid an overwhelming increase in movements at the intersections. Lack of adequate left or U turns could be one of the reasons why this result was observed. For the same cluster alcohol/drug use is also found to be associated with severe/fatal crashes in the model for driver and vehicle related factors. The authors in a previous study (Das et al., 2008)

found similar result for alcohol/ drug use. Wang and Abdel-Aty (2008) found an increasing effect of alcohol/ drug use in severity of crashes. In Cluster 2 (2.898 – 5.729 miles) for the environmental and roadway geometric model, posted speeds greater than 45 mph are found to be riskier. In a recent study by Malyshkina and Mannering (2008), they found higher posted speed limit to be associated with higher severity of injuries. For corridors where the posted speed limits are less than 45 mph and high k factor, conditions are suitable for crashes with higher injury severity. In the driver and vehicle related factors' model, failure to use safety equipment and alcohol/drug use also lead to severe/fatal crashes. Though much research highlights the seat belt use and its obvious benefits (Evans, 1996; Derrig et al., 2000; Eluru and Bhat, 2007), very few discuss the effects of other safety equipment in use inside the vehicle in general. Likewise for Cluster 3 (5.762 – 10.556 miles) corridors the model for environmental and roadway geometric reflects that posted speeds of greater than 50 mph leads to higher severity. While the model for driver and vehicle related factors show that the non-use of safety equipment and alcohol/drug use again lead to crashes which are more at threat to be severe. However, for Cluster 4 (10.644 – 78.293 miles) corridors, the two models (environmental and roadway geometric factors' model; and driver and vehicle related factors' model) were developed at only 70% and 75% levels of confidence, respectively. Hence, the results are not reported here. Summarizing the results reflects that angle/turning movement crashes are more severe under high speeds, no safety equipment in use and driving under the influence. The results are consistent with the common perception.

6.4.2.3 Rear-end Crashes

The environmental and roadway geometric factors' model for the rear-end crashes in Cluster 1 suggests that higher friction resistance (*skid* > 38) lead to higher severity of injuries given the crash has occurred. This is counterintuitive as higher friction should be better at preventing severe crashes. The result could provide insight to the phenomenon that when the friction is higher and the vehicles can brake within shorter distances, the internal movement could be sudden and any internal/secondary collision (i.e. passengers hitting something inside the vehicle) could lead to a severe injury. In the model for driver and vehicle related factors it was observed that the severe/fatal crashes are linked to light, slow moving vehicles like cycles, mopeds, etc. The higher severity level is intuitive, as any crash with light vehicles will generally be severe. Huang et al (2008) found similar result in their investigation of traffic crashes at intersections. The environmental and roadway geometric model for Cluster 2 corridors indicate that the posted speed limit of greater than 50 mph leads to severe rear-end crashes. Similarly, in a recent technical report developed for NHTSA by Liu and Chen (2009) it was observed that severe crashes are more likely to occur at corridors with posted speed limits of 50 mph or greater. On the other hand when the speeds are less than 50 mph, crashes will be severe/ fatal when the k-factor is high. For the same Cluster 2 corridors alcohol/ drug use leads to crashes which are severe/ fatal as shown by the driver and vehicle related factors' model. It is observed that when there is no alcohol/ drug use by the responsible driver the presence of a person in the vulnerable age group (> 55 yrs or < 3 yrs) makes the crash more severe in general. While alcohol/ drug use is a case of irresponsible driving behavior the presence of person in the vulnerable age group is a clear case of physical fragility. People in the vulnerable age group always tend to experience

severe injuries resulting out of a crash. The authors would like to bring a particular case reported by Batra and Kumar (2008) in which an 84 year old man succumbs to injuries resulted in a low velocity collision. In this particular case the injury was a subaxial cervical spinal cord injury which was triggered by the airbag deployment and interestingly the driver was not wearing a seat belt. The authors cite this particular example as it was observed that under relatively slower speeds (<50 mph) severe injuries can occur if the safety equipment is not properly used and it also confirms the observation that the presence of a person in the vulnerable age group will succumb to injuries that become more apparent due to physical fragility. On Cluster 3 corridors lower ADT leads to higher severity crashes while the driver and vehicle related factors' model indicate alcohol/ drug use leads to severe/fatal crashes. Lower ADT could mean higher speeds which more often lead to severe/ fatal crashes. For even longer corridor groups, i.e. Cluster 4, higher friction resistance (skid > 34) leads to severe rear-end crashes by the environmental and roadway geometric factors' model. The explanation has been given in the beginning of this subsection. For lower friction resistance, greater surface widths (corresponding to 3 or more lanes per direction) and the presence of median curb increase the severity level of crashes. The increase in surface width should traditionally reduce severity (Petritsch et al., 2007); however this result might seem counter intuitive. This could be explained in the following way. Higher surface width may result in higher speeds and more driver comfort which might cause some drivers to be less cautious. Hence the increase in speeds and less attention by the drivers could lead to crashes with severe injuries. The authors in one of their previous work (Das et al., 2008) had found similar result. On the same corridor group, older drivers (> 55 yrs) also are involved in severe rear-end crashes. The longer the corridors the more the exposure of the driver and the

older the driver the more prone is he/she to make an error. Marshall (2008) states that prevailing medical conditions and impairments associated with old age leads to deteriorating fitness and hence lead to higher crash risk for the older driver.

6.4.2.4 Head-on Crashes

For head-on type of crashes on corridors belonging to Clusters 1 and 2 combined, crashes on dry surface condition were found to be more severe/ fatal from the environmental and roadway geometric model. However, the model for driver and vehicle related factors was developed at lower confidence level of 70%; hence the results are not reported here. Dry surface conditions probably indicate fine weather and more vehicles on the road. Hence improper maneuvers could result in head on collisions, especially when the highways are undivided, resulting in severe crashes. In a related study by Yan et al. (2008) it is shown that slippery road conditions lead to a higher probability of crash avoidance maneuvers as drivers will drive more cautiously during unfavorable conditions. Hence, the results in this study indicate that drivers could be less attentive when driving in good weather and road conditions. In Clusters 3 and 4, alcohol/drug use is the primary reason for severe head-on crashes.

6.4.2.5 Sideswipe Crashes

In sideswipe crashes, restrictive medians are more threatening on shorter corridors (Cluster 1) as shown by the environmental and roadway geometric model. While on longer corridors (Cluster 3) straight ahead movement is crucial as observed from the driver and vehicle related factors'

model. For all other type of movements, severe sideswipe crashes occur when slow moving vehicle type and light trucks are involved. Research work by Anderson (2008) indicates that the increase in the light truck traffic increases the number of fatalities on the road. In the same work it was indicated that up to eighty percent of the increased deaths can be assigned to occupants in other vehicles and pedestrians. For severe/fatal sideswipe crashes involving slow moving vehicles, turning movements along with changing lanes are the significant parameters on Cluster 3 corridors. The more the lane changing maneuvers the higher the probability of crash severity as many of the maneuvers will be risky.

6.4.2.6 Single vehicle Crashes

For crashes involving single vehicles, higher friction factor leads to increased severity in crashes on shorter length corridors (Cluster 1 and 2 combined) according to the environmental and roadway geometric factors' model. On the other hand the driver and vehicle related factors' model for the same corridors indicate straight vehicle movement related crashes are found to be more severe. For the single vehicle type of crashes, occurring on Cluster 3 corridors, that are related to segments or access points the crashes tend to be more severe at stretches where the posted speed limits are 45 mph or greater. The driver and vehicle related factors' model show that failure to use safety equipment in slow moving vehicles also leads to severe injuries in crashes. In Cluster 4 the crashes are more at risk to be severe when the posted speed limit is greater than 50 mph. The driver and vehicle related factors' model for this cluster indicate that slow moving vehicles (e.g. cycles, mopeds, etc.) tend to be involved in severe crashes. This

could be explained by the fact that on corridors with 50 mph posted speed slow moving vehicles pose a risk as they will create speed variance on the roadways. Collisions with slow vehicles would likely be severe.

6.4.2.7 Results Summary

The results discussed in the preceding subsections are summarized in Table 6-6 through Table 6-10. The variables in the cell represent those which increase severity along with the range or categories. The blank cells indicate that the results could not be developed with the 90% confidence level. These tables will help the reader to have a comparative understanding of the variables entering a particular tree model and how they affect safety. Tabulating the results helps to better understand the results; particularly in this study where the results are brought together and compared across crash types and corridor clusters.

Table 6-6 Significant factors for Angle / Turning movement crashes

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Paved shoulders and k factor > 9.85; Paved shoulders and k factor < 9.85 and restrictive median	Posted speed limit > 45 mph; posted speed limit < 45 mph and k factor > 9.85	Posted speed limit > 50 mph	No significant results
Driver and Vehicle Related Factors	Alcohol/ drug use	Non-use of safety equipment and alcohol/ drug use	Alcohol/ drug use	No significant results

Table 6-7 Significant factors for Rear-end crashes

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Skid resistance >38	Posted speed limit > 50 mph; posted speed limit < 50 mph and k factor > 9.85	Lower ADT (<31,000)	Skid resistance >34; Skid resistance < 34 and surface width > 32 ft and presence of median curb
Driver and Vehicle Related Factors	Light slow moving vehicles	Alcohol/ drug use; No Alcohol/ drug use and presence of person in the vulnerable age group (> 55 yrs or < 3 yrs)	Alcohol/ drug use	Older drivers > 55 yrs

Table 6-8 Significant factors for Head-on crashes

	Clusters 1 and 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Dry surface condition	No significant results	No significant results
Driver and Vehicle Related Factors	No significant results	Alcohol/ drug use	Alcohol/ drug use

Table 6-9 Significant factors for Sideswipe crashes

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Restrictive medians	No significant results	No significant results	No significant results
Driver and Vehicle Related Factors	No significant results	No significant results	Straight ahead movement of the vehicle; turning movements along with changing lanes and slow moving vehicles	No significant results

Table 6-10 Significant factors for Single vehicle crashes

	Clusters 1 and 2	Cluster 3	Cluster 4
Environmental and Roadway Geometric Factors	Skid resistance >38	Crashes related to segments and/ or access points and posted speed limit > 45 mph	Posted speed limit > 50 mph
Driver and Vehicle Related Factors	Straight ahead movement of the vehicle	Non-use of safety equipment and slow moving vehicles	slow moving vehicles

6.5 Concluding Remarks

The application of conditional inference trees and forests leads to the identification of an unbiased set of variables significantly related with severity. The advantage of the new algorithm of tree/forest development over the traditional CART tree/forest is that it prevents the uninformative variables from being identified as significant just by the virtue of having higher number of categories or being continuous in nature. The novel way of separating the split criteria from the variable importance selection while developing a tree is what makes the conditional inference trees unique. The Chi-square test is used to determine the strength of association with the target variable, in the present application it is the binary severity variable. Once a variable is selected at a particular tree level for split, the split can then be decided based on any criteria, including those used in the CART algorithm. The conditional inference forests on the other hand calculates individual variable importance of each variable for every tree by first breaking the association with permutation and then testing the tree with out-of-bag estimates. In the forests, the variable importance is based on the result from multiple trees thus avoiding the instability of individual trees.

Among the results from the analysis, alcohol/ drug use is associated with increased severity of crashes irrespective of the length of the corridors or the type of crashes. Since the drivers are less likely to be in control; it invariably leads to severe crashes. Failure to use safety equipment has lead to increased severity of single vehicle as well as angle/turning movement related crashes. In this regard, conclusions drawn by Abdel-Aty and As-Saidi (2000), by analyzing the zip codes of the offenders for better targeting of the education programs, may be of renewed interest. Older at fault drivers are found to be more at risk of getting involved in a severe crash especially in a rear-end collision on longer corridors. On similar corridors, a crash is more likely to have a severe injury where there is a person in the vulnerable age group (more than 55 years or less than 3 years).

Slow moving vehicles like cycles and mopeds have been observed to be involved in severe injury crashes. Many of these severe crashes occur at signalized intersections. It indicates that the designs of the intersections need to improve with respect to the slow-moving vehicles and possibly even pedestrians. For shorter length corridors, higher k–factor is a significant parameter for increased severity crashes. Higher k-factor essentially means that the corridor is designed for handling higher volume during peak hour. It in turn has the potential not only to reduce rear-end crashes during the peak hour (due to improved congestion situation) but also to increase speeds due to better design during off-peak periods. Since rear-end crashes tend to be less severe, higher k-factor leads to increased likelihood of severe crashes. On longer corridors, like those in Cluster 3, severity of rear-end crashes increases when the posted speed limit is greater than 50 mph. Lowering the posted speed limit may not be the best strategy from an operations point of view

but it may lead to reduction in severity of crashes. Lower ADT also leads to severe rear-end crashes on Cluster 3 corridors, especially for rear-end crashes. Severe/fatal crashes involving single vehicles are more likely to be associated with access points on longer corridors. Reducing the number of access points may not always be feasible; however, design changes such as improved merging may be adopted for these issues.

Corridors of smaller lengths (generally less than 5 miles) have been observed to have problems of increased severity if crashes occur on corridors with high skid resistance values. Shorter corridors also have problems when the posted speed limit is greater than 45 mph. Since most of these small urban/ suburban corridors are located between longer stretches of rural corridors; they have lower speed limits compared to adjacent sections. However, since the congestion is not high on the rural sections, some drivers will tend to speed and thus create a larger variation in prevailing speeds. This variation could lead to more severe crashes on shorter length corridors. Restrictive median openings on shorter corridors have also been found to be problematic. The variable indicating the presence of vulnerable age group also came out significant on shorter corridors rather than on longer corridors. On longer length (greater than 5 miles) corridors, speed limit of greater than 50 mph is a cause of concern. Non-use of safety equipment is also more pronounced in contributing towards severity on longer corridors. In a recent paper by Eluru and Bhat (2007) the question of the endogenous relationship between the seat belt use and the injury severity is raised. There is possibility of intrinsically unsafe drivers not wearing the seat belt and are the ones to be likely involved in high injury severity crashes because of their unsafe driving habits. In the present study, however, the researchers observe the overall safety equipment in use

in the vehicle. Results also show that the failure to use the safety belt in single vehicle crashes and crashes involving slow vehicle lead to higher severity crashes. Thus the present study is not only in line with concurrent research but also goes a step further in identifying the type of crashes which are more likely to be affected by the underlying endogenous relationship.

Due to these observed differences, the decision to cluster the corridors has been justified. The subtle differences are highlighted when the groups are logically made. The clusters which were originally made based on the length actually shed light on the factors and a lot of new significant variables come into the picture.

The results from the forest and the trees are intuitive and their association with severity may be explained. Certain known results about severity of crashes have been confirmed while some new information is discovered about others. Alcohol/ drug use along with higher speed limits tend to result in more severe/fatal crashes. The new variable "element" which uses information from site location, signal type information and traffic control was also insightful in identifying locations which are more critical from the severity aspect. Drivers of vehicles with passengers in the vulnerable age group range must also be more careful while driving, as the physical fragility of these subjects, tends to make the injuries more severe. The authors also used the safety information for all passengers seated in the car. That particular variable also was significantly associated with severity of crashes. Hence, it is critical that internal safety should be a concern for the law enforcement agencies if they are intended to reduce the occurrences of severe/fatal crashes on the arterials of Florida.

CHAPTER 7. GENETIC PROGRAMMING FOR CLASSIFICATION AND FREQUENCY ANALYSES

7.1 Requirement for a common approach

Safety assessment of roadway elements such as mid-block segments, signalized intersections and un-signalized intersections (access points) includes investigations into the severity as well as the frequency of crashes. The objective of transportation safety engineers is not only to reduce the number of crashes but also to mitigate the injury severity in case of a crash. Hence, any research directed only towards the frequency or the severity analysis of crashes renders inadequate. Though this aspect of safety analysis is widely accepted, the existing body of knowledge however has very limited citations for a complete analysis involving both the crash counts and severity of the injuries resulting in the crashes. Recently Ma et al. (2008) used a multivariate Poisson-lognormal approach to model crash occurrence simultaneously at various levels of injury severity. However, the complex statistical structure of the study makes it less practical to implement.

Fundamental difference between the crash occurrence phenomenon and the injury severity levels is the response type. Crash occurrence is a continuous integer response while the severity is an ordinal target. Most statistical studies for the two phenomena are based on this difference. For crash count prediction, models such as negative binomial (Miaou, 1996; Harwood et al., 2000) and support vector machines (Li et al., 2008) are the norm. In case of injury severity, logistic regression (Huang et al., 2008; Sze and Wong, 2007), binary trees (Das et al., 2009; Chang and

Wang, 2006), ordered probit and logit models (Das et al., 2008, Obeng, 2008) and the innovative proportional odds model (Wang and Abdel-Aty, 2008) are the standard modeling practices.

In this study the authors investigate a generalized heuristic approach of Genetic Programming (GP) to model injury severity as well as the crash frequency. GP uses concepts from evolutionary biology, such as crossover and mutation, for the model development process and is the same for both regression and classification. The process of model evolution takes places, through generations, with decreasing mean error as the objective function for regression and increasing hit rate as the objective function for classification problems.

Presently the researchers investigate the frequency and severity analysis for crashes, specifically for urban arterials (*not* limited access facility) in this study. Though they are fundamentally different phenomena yet they have an overlapping set of contributing factors. It must be understood that crash occurrence and the injury severity is sequential in the reference frame of time, i.e. they are not simultaneous. First, a crash has to occur and then an injury may result. Hence, there is a one-way dependency between both events. The author suggests here independent approaches for building both the severity and frequency of crashes models under the broader umbrella of the heuristic GP. Since the crash occurrence and injury severity are fundamentally different phenomena it is not practical to have one model governing them. However, in this study a common heuristic model development process for both events has been proposed.

The following section explains the GP methodology and the overall model development algorithm. The results and analyses follow next, with the injury severity analysis preceding the crash count modeling for rear-end crashes on urban arterials. The data set preparation has been included in the respective analysis sub-sections. The crash frequency analysis includes graphical demonstration of the change in crash counts with the change in parameter values.

7.2 Genetic Programming (GP)

According to Deschaine and Francone (2004), genetic programming (GP) is observed to perform better than classification trees in terms of lower error rates and also outperforms neural networks in regression analysis. GP is a heuristic search technique that iteratively evolves better programs which could either be the best solutions or lead to the best solutions. The innovative evolutionary computation, GP, is based on the genetic algorithms (GA). In GA, the optimum solution is reached by using the well established techniques of evolutionary biology. In a recent work by Makkeasorn et al. (2006) in the field of water resources management, soil moisture estimation models were developed by the use of Discipulus^{TM} Genetic Programming (GP) software and were applied to the soil moisture distribution analysis. The work shows that GP, a type of GP, helps in the development of excellent nonlinear multivariate regression models. The work also compared the GP model developed with the linear regression and nonlinear regression models independently and the GP model was found to be the best for the data. The linear regression model overestimated the soil moisture while the nonlinear regression models tend to underestimate it. According to Chang and Chen (2000) the regression models generated by GP is

also independent of any model structure. Use of GA in transportation is not new. They have been used widely in traffic signal system optimization and network optimization (Park et al., 2000; Ceylan and Bell, 2004; Teklu et al., 2007). The use of GA or GP in transportation safety studies is relatively new and hence the authors intend to test the method and observe its potential.

7.2.1 Problems in Genetic Algorithm

GP, which is a class of evolutionary algorithms, has its roots in the GA. GA is a method to grow from one population to a new population through the process of evolution. For a detailed review of conventional GA the readers are directed to the classical work by Holland (1975) and Goldberg (1989). For the more advanced learners, typically, in GA the representation is generally fixed length representation of length 'l' and the alphabet size is 'k'. In the search space of a fixed length representation of length 'l' and alphabet size 'k' the available candidate solutions are k^l. The initial selection of string length limits the search space and puts restrictions on the learning process of the GA. Thus traditional GA sometimes converges on suboptimal solution. Suboptimal performance may also occur when there is no hill to climb, i.e. if there is a single fitness criterion. For example, in binary classification the criterion is to check whether it goes to the right bin or not. Hence the GA may fail during classification. This observation is critical for the choice of GP over GA in classification problems.

7.2.2 Genetic Programming

According to Koza (1992), “the most natural representation for a solution is a hierarchical computer program rather than a fixed-length character string”. The size and shape of the computer program, in other words the complexity of it, is not known apriori. The restrictions in the traditional genetic algorithms has led to the use of the more powerful and versatile genetic programming which takes into account the complexity of problem solving. They use other forms of representations like the tree structure or straight forward one line instructions to the machine. The author directs the inquisitive reader to the well documented work of Koza (1992) on GP.

In traditional GP, the programs, in the memory, are stored as tree structures. Every tree node has an operator and every leaf node is an operand. This makes the evolution as well as the evaluation of the tree much uncomplicated. The evolutionary biological operations like crossover and mutation are also fairly easy to implement. Typically during crossover there occurs an interchange of sections between two homologous chromosomes at a certain splice point. On the other hand mutation means the alteration of any particular point in a chromosome. Chromosome here refers to the program instructions. With a tree based structure replacing a node, which occurs during the crossover, the whole branch is replaced. The resultant individual is very much different from the parent. In mutation, either the node’s information is replaced or the node is removed.

However, in GP the crossover will occur between two or more instructions’ set whereas mutation will occur on a single instruction set. Figure 7-1 and Figure 7-2 show the crossover and mutation

occurring in GP. For example two functions, g(0) and h(0), be two instructions that has to be crossed over. The process of crossover between the two instructions is illustrated in Figure 7-1. The part of the instructions shown within the ovals will swap places.

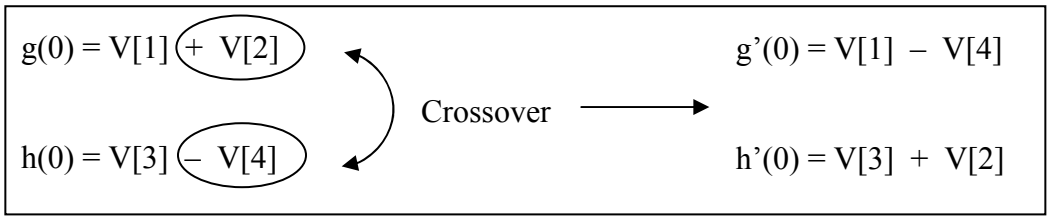


Figure 7-1 Crossover between two instructions in GP

As can be observed from Figure 7-1 that crossover takes place between the branches, along with the operand, resulting in two daughter instructions, g'(0) and h'(0). In Figure 7-2 the process of mutation in GP is illustrated. The operand, in this example the division sign, '/', has been circled. This operand can undergo mutation to any other mathematical operand. In this particular example it mutates to the multiplication symbol, '*'.

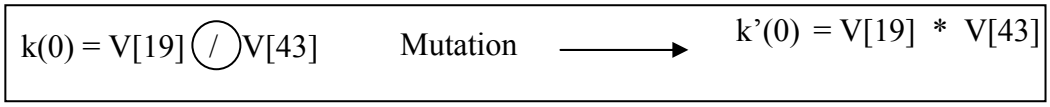


Figure 7-2 Mutation of an operand in an instruction in GP

Typically evolution or development occurs through generation and the fitness of the population, which is typically the evaluation criteria, is examined in every generation. Figure 7-3 represents the flowchart for a typical generation in GP. The fitness function in this study is the average of the squared errors, where error is the difference between the evolved output and the target output.

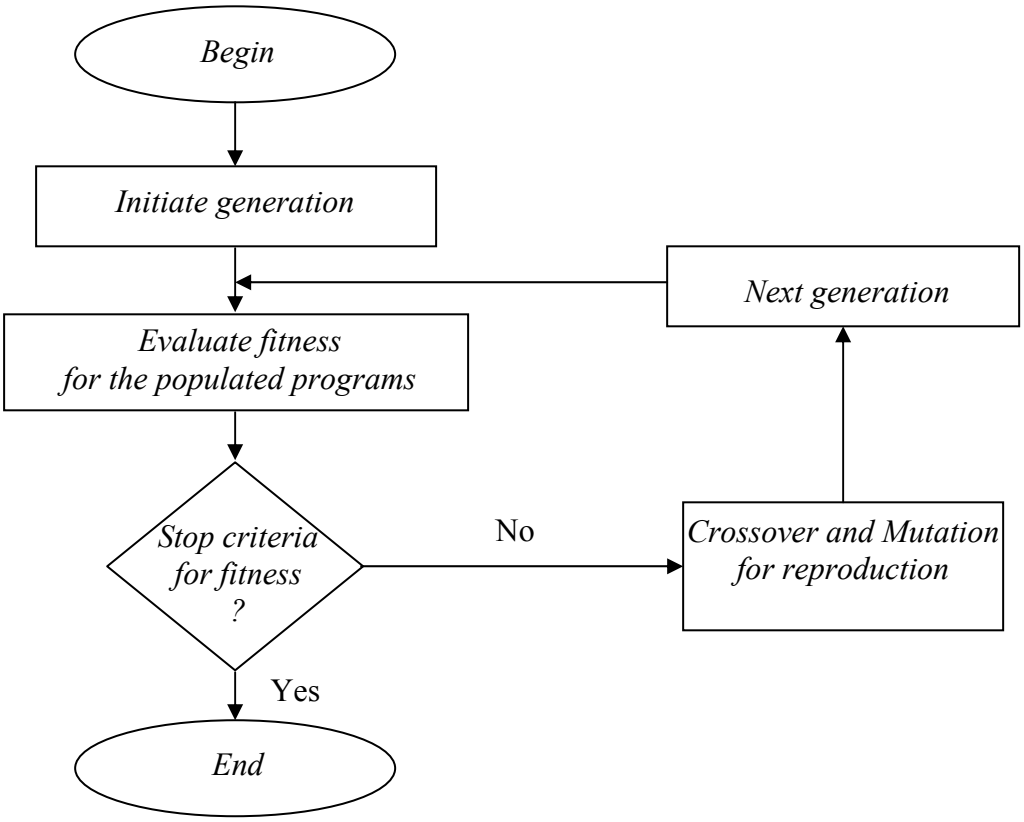


Figure 7-1 Typical steps in one generation in GP

However, in GP the representation of the computer programs is a set of instructions written in the machine language (Brameier and Banzhaf, 2007). The software Discipulus^{TM}, which has been used in this study, implements GP to develop best programs evolved for the problem at hand. Please note that from here on the authors will use GP term as that is the specific form of GP used. It must be noted that any reference to the term GP, in this particular study, always means the broader category of the heuristic approach.

7.2.3 Discipulus™

Since it is based on GP, the population is comprised of linear computer programs. From an initial pool of computer programs, a random "tournament" selection from a set of 4 randomly chosen programs is conducted. The tournament then chooses the two best programs based on the performance on the task designated. These programs are then copied and the standard crossover and mutation operators are applied. The new "child" programs replace the two loser programs and the process repeats till the GP finds the best program suited for the given task. The software is a multiple-run genetic-programming system. The fact that the genetic programming is a stochastic algorithm, running it multiple times yields a wide variety of results. In this particular study for every run 80 generations must pass without improvement for the run to be terminated. Figure 7-4 represents the process undergoing in a typical run in the GP. In each run, the population undergoes evolutionary changes through generations.

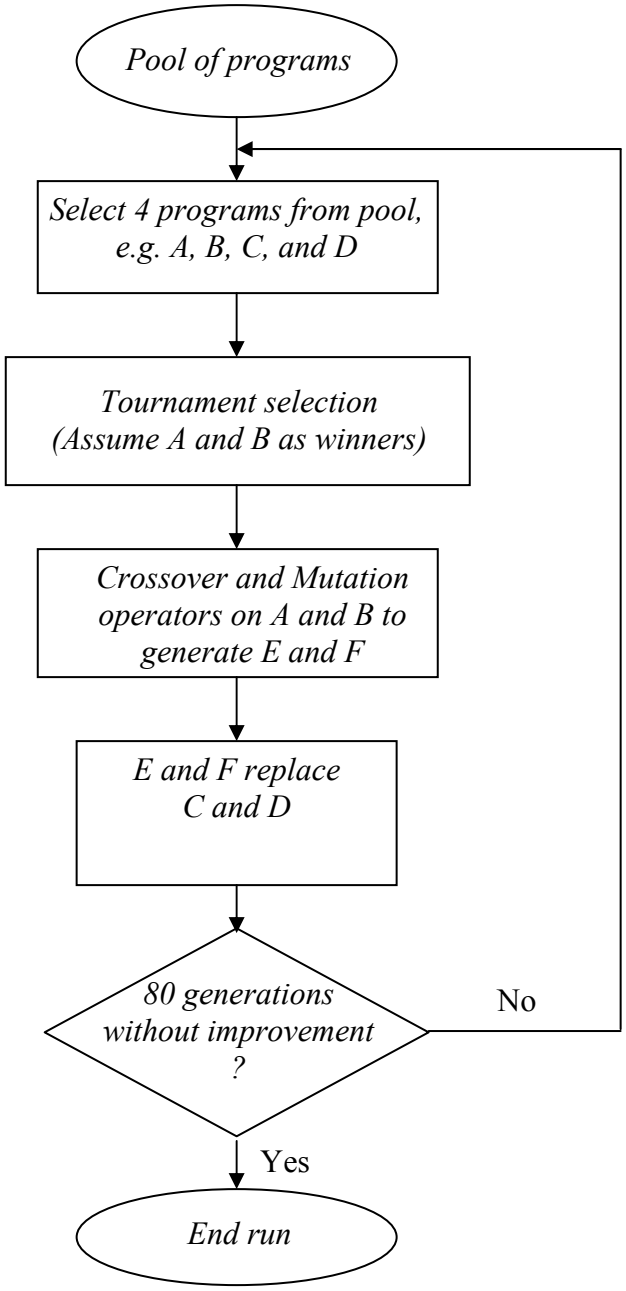


Figure 7-2 Flowchart for processes in a typical run

The software Discipulus^{TM} implements GP to develop best programs evolved for the problem at hand. Typically in this study a lower crossover rate (0.5) and a higher mutation rate (0.95) has

been implemented to avoid genetic drift. Genetic drift is the accumulation to a sub-optimal solution in the search space due to stochastic errors. The process of mutation always brings in novelty to the population of evolved generations. GP can also assemble teams of models than just individual models. As mentioned earlier minimization of the error rates is the criterion for selecting the best program (model) in the evolution process. Figure 7-5 illustrates the change of fitness (mean error) as the runs increase. Each run has a predetermined number of generations to evaluate.

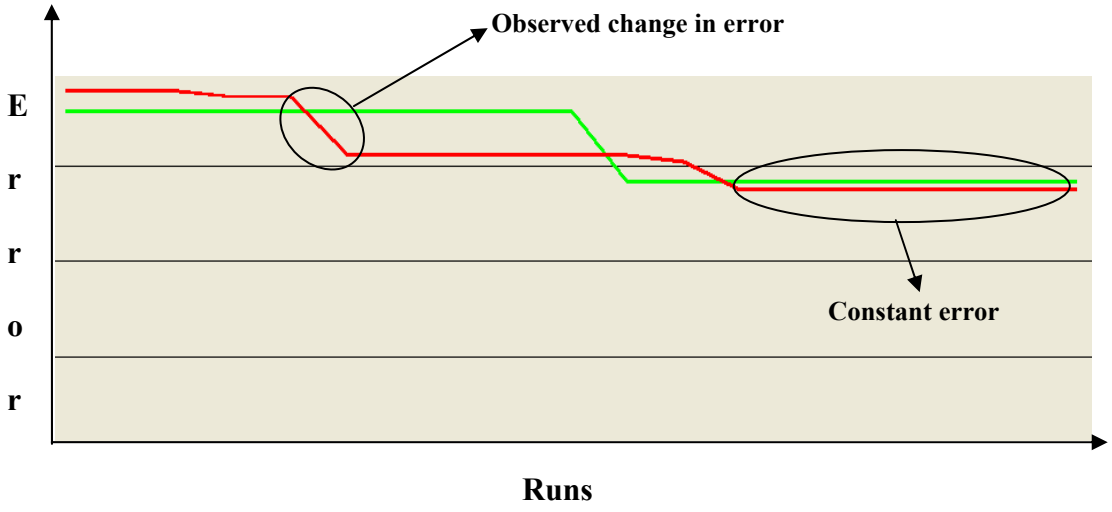


Figure 7-3 Decreasing mean error of the best individual program and the best team

The red line in the plot indicates the error rate of the best model at any given run; where as the green line indicates the error rate of the best team comprising of a fixed set of individual models. In this research the red line is of importance as the objective is to find the best individual models for the regression and the classification problem. The evolution process is externally observed by plots as illustrated in Figure 7-5. If the error remains constant over many runs, the model

development process can be terminated. It is not advisable to stop the process when a change in error is detected.

As stated earlier, the GP is the broader platform under which both classification and regression analyses can be implemented. Figure 7-6 shows the overall analytical approach adopted for the study.

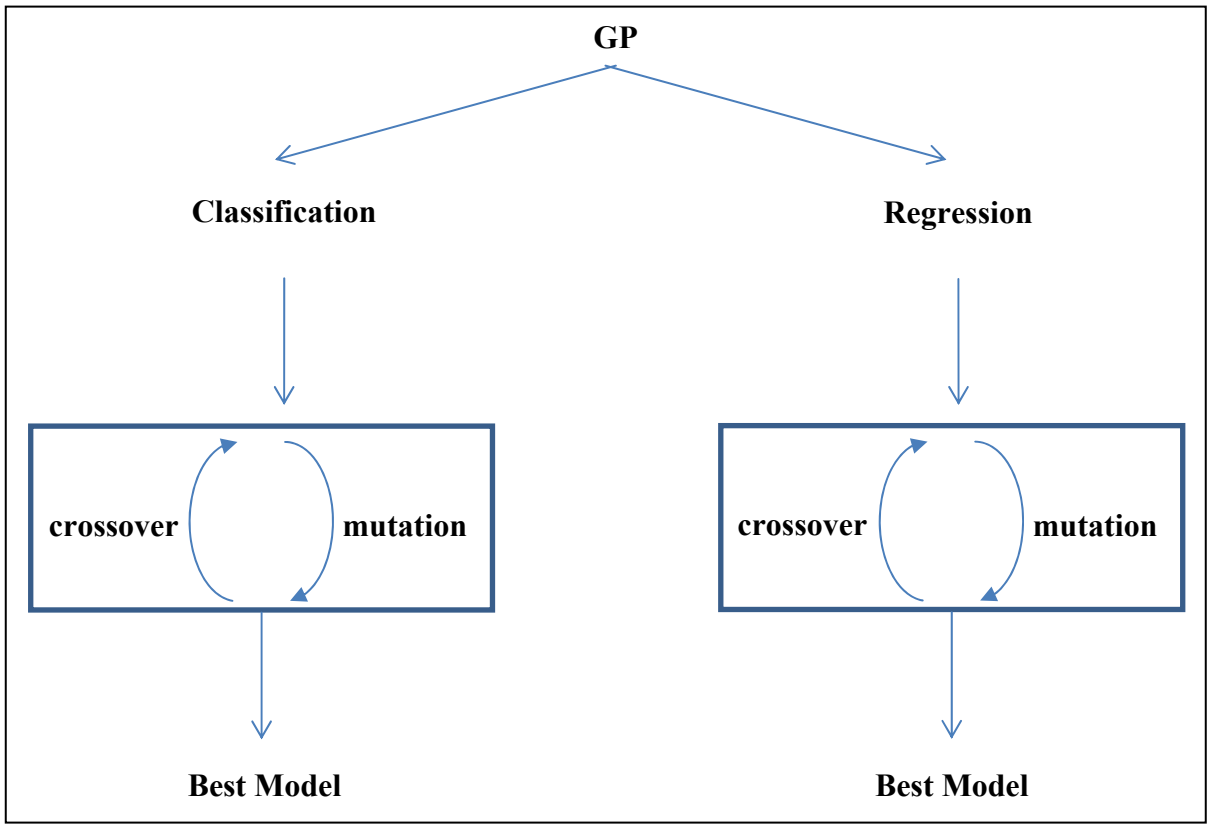


Figure 7-4 Overall analytical approach for model development

The objective achieved in this study is the adaptation of a uniform platform for model development for both classification and regression problems. The GP modeling approach is

particularly apt for this objective as the model building process is independent of the modeling intent.

During the model building process all the input variables are taken into consideration. However, all the variables are not included in all the programs as it searches for the programs best fit for the classification under study. The selection of variables is essentially analogous to any regression model where only the significant variables enter the final model from a host of input variables. In GP too, the various models (programs) have only a select subset of variables and each program has a different classification rule. The Discipulus™ software produces a series of 30 best programs evolved over the runs. The model development process continues till no further minimization of error or maximization of classification rate is observed through further runs (see Figure 7-5). The best model is chosen which has the least error (for regression problems) or the highest classification rate (for classification problems).

The program contains a series of effective register instructions along with introns (non-effective instructions). In order to find the simplest set of linear instructions, the researcher has to purge all the introns. Once the introns are removed the fitness of the program remains unchanged. The final set of instructions is read line by line to get the final form of the program. For development and evaluation of the models the primary dataset was split into training and validation datasets consisting of 70% and 30% of the data, respectively.

7.3 Analyses and Results

7.3.1 Injury Severity Modeling

7.3.1.1 Data Preparation

The crash data as well as the roadway characteristics data were made available through the CAR and the RCI system of FDOT respectively for the years 2004 through 2006. As mentioned in CHAPTER 5, the corridors were clustered into 4 groups (see Table 5-1). The types of crashes used in the study are: i) angle/ turning movement (44,088 crashes); ii) head-on (3709 crashes); and iii) rear-end (57,155 crashes). The other type of crashes could not be used as insufficient data failed to produce any classification rule for any cluster. Continuous variables like ADT, Percentage of trucks, and K-factor (design hour volume as a percentage of ADT) and skid (friction resistance multiplied by a factor of 100) were also divided into categories. Their relationships with severe/fatal crash occurrence may not be monotonic in nature. Nominal variables such as median types, access management, shoulder types, surface types, etc. were also used in the data set. In most statistical applications the nominal variables can be defined and the dummy variables are created internally. In the present study however, the researcher will have to create dummy variables for all the nominal variables with three or more categories. Otherwise the GP will treat it as an ordinal variable. A total of 58 variables have been used.

Table 7-1 presents all the independent variables along with the dependent severity variable. The present analysis deals with roadway geometric and design factors. The author would like to

reiterate that the objective of the study is to understand the classification of injury/no-injury crashes as well as severe/non-severe crashes. Apart from that, the researcher wanted to investigate the usefulness of using the heuristic GP methodology in the classification problem to identify significant variables and their relationship. As an initial approach the researcher has used specific roadway geometric and design factors in this particular study, information for which were completely available. A broad spectrum of variables is always available and open to investigation. However, in this study only certain variables (Table 7-1) have been included which broadly belongs to roadway geometric and design category. These variables are generally used in engineering studies to develop safety countermeasures. Many of these variables have been collected and are unique to this study. As discussed in the 'analysis and results' section, the results highlight intuitive observations and also help in discovering of interactions among variables. All other variables that have not been included are beyond the scope of the present study.

Table 7-1 Dependent / Independent variables used in crash classification

Variable Name	Variable Symbol	Description
Target or Dependent Variable		
Injury		Binary target variable
Severity		Binary target variable
Environmental and Roadway Geometric Parameters		
Surface_width	V_0	Width of the surface (Continuous)
Max_speed	V_1	Maximum posted speed limit (Continuous)
Road_cond	V_2	Road condition at time of crash (Binary (1 = no defects; 2 = defects))
Vision	V_3	Vision obstruction (Binary (1 = no; 2 = yes))
shld_side	V_4	Shoulder + sidewalk width (Continuous)
surf_cond	V_5	Surface condition (Binary (1 = dry; 2 = other))
light	V_6	Daylight condition (Binary (1 = daylight; 2 = other))
k_fact	V_7	Average k – factor (k_fact <= 9.85, k_fact > 9.85)
trfcway	V_8	Vertical curvature (Binary (1 = level; 2 = upgrade/ downgrade)
park	V_9	Presence of parking (Binary (1 = no; 2 = yes))
surf_type	V_{10}	Type of surface (Binary (1 = black top surface; 2 = any other surface))
shld_t	V_{11}	Type of shoulder (Binary (1 = paved; 2 = unpaved))
LIGHTCDE_1	V_{12}	No street light (Binary)
LIGHTCDE_2	V_{13}	Presence of street light (Binary)
LIGHTCDE_3	V_{14}	Partial lighting (Binary)
ACMANCLS_num_0	V_{15}	No median opening (Binary)
ACMANCLS_num_2	V_{16}	Presence of restrictive median with service roads (Binary)
ACMANCLS_num_3	V_{17}	Presence of restrictive median (Binary)
ACMANCLS_num_4	V_{18}	Presence of non-restrictive median (Binary)
ACMANCLS_num_5	V_{19}	Presence of restrictive median with shorter directional openings (Binary)
ACMANCLS_num_6	V_{20}	Presence of non restrictive median with shorter signal connection (Binary)
ACMANCLS_num_7	V_{21}	Presence of both restrictive and non-restrictive median types (Binary)
curvclass_1	V_{22}	Presence of curve < 4^o (Binary)
curvclass_2	V_{23}	Presence of 4^o <= curve <= 5^o (Binary)
curvclass_3	V_{24}	Presence of 5^o < curve <= 8^o (Binary)
curvclass_4	V_{25}	Presence of 8^o < curve <= 13^o (Binary)
curvclass_5	V_{26}	Presence of 13^o < curve <= 27^o (Binary)
curvclass_6	V_{27}	Presence of curve > 27^o (Binary)
ADT_1	V_{28}	ADT <= 31000 (Binary)
ADT_2	V_{29}	31000 < ADT <= 40000 (Binary)
ADT_3	V_{30}	40000 < ADT <= 52500 (Binary)
ADT_4	V_{31}	ADT > 52500 (Binary)
t_fact_1	V_{32}	t_fact <= 4.05 (Binary)
t_fact_2	V_{33}	4.05 < t_fact <= 5.895 (Binary)
t_fact_3	V_{34}	t_fact > 5.895 (Binary)
dayandtime_1	V_{35}	Afternoon Peak Weekday (Binary)
dayandtime_2	V_{36}	Morning Peak Weekday (Binary)

dayandtime_3	V_{37}	Friday or Saturday Night (Binary)
dayandtime_4	V_{38}	Off-peak (Binary)
pavecond_1	V_{39}	Poor condition (Binary)
pavecond_2	V_{40}	Fair condition (Binary)
pavecond_3	V_{41}	Good condition (Binary)
pavecond_4	V_{42}	Very Good condition (Binary)
skid_f_1	V_{43}	Skid <= 34
skid_f_2	V_{44}	34 < skid <= 38
skid_f_3	V_{45}	Skid > 38
median_0	V_{46}	No median (Binary)
median_1	V_{47}	Presence of painted (Binary)
median_2	V_{48}	Presence of median curb <= 6” (Binary)
median_3	V_{49}	Presence of median curb > 6” (Binary)
median_4	V_{50}	Presence of lawn (Binary)
median_5	V_{51}	Presence of paved median (Binary)
median_6	V_{52}	Presence of curb <= 6” and lawn (Binary)
median_7	V_{53}	Presence of curb> 6” and lawn (Binary)
median_8	V_{54}	Other median (Binary)
ele_1	V_{55}	Segment related crashes (Binary)
ele_2	V_{56}	Intersection related crashes (Binary)
ele_3	V_{57}	Access related crashes (Binary)

Most of the variables as can be observed are binary with a few continuous variables. Most of the binary variables are dummy variables which uniquely represent a particular aspect of the original nominal variable and hence, the results of the classification could be directly interpreted. The descriptions for the variables 16 through 20 described in Table 7-1 have *restricted median* or *non restrictive* median types. The *restrictive medians* are those medians which provide a physical barrier between the opposing traffic lanes; where as the *non restrictive medians* are those which are painted medians or center lines that do not provide a physical barrier. The variables 55 through 57 in Table 7-1 provide some new innovative variations to the traditional parameters. As explained earlier in CHAPTER 4, traditionally the site location variable has been used by researchers to assign crashes to the three roadway elements (segments, intersections and access points). However, ‘traffic control’ in combination with the ‘site location’ along with the information of the presence or absence of signal, did a superior job in attributing crashes to one

of the three roadway elements. Based on these three independent parameters, the variables *ele_1, ele_2, ele_3* were created to assign the crashes to the three roadway elements, namely segments, intersections and access points, respectively.

The author set up a classification problem for the injury occurrence as well as the severity of crashes. In a typical classification problem the algorithm develops a set of rules which when followed leads to a particular category of the target variable. For example, in crash severity analysis when the binary target variable represents severe/ non-severe crashes, the classification rule developed will lead to either severe crashes or non-severe crashes. The variables that enter the rule are significant and their directionality is critical for understanding the contribution of the variable in the analysis.

The first analysis that was carried out was a binary classification problem between injury crashes and non-injury crashes. Figure 7-7 shows the primary binary classification problem. It must again be noted that a major proportion of non-injury crashes are primarily PDO crashes which are known to be under-reported (Abdel-Aty and Keller, 2005; Yamamoto et al., 2008). A correction factor has not been included as that will over represent PDO crashes at many sites. It is not believed that this issue would affect the results and objectives of this study.

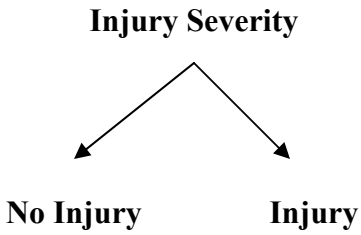


Figure 7-5 Binary Classification of Non-injury / Injury related crashes

However, this will be just a part of the analysis. Since the injury related crashes represent all types of injuries and the degree of severity ranges from possible injury to death, it should be further be split. Keeping in view the nature of the injury two possible grouping of the injury related crashes is possible. The crashes with fatalities and incapacitating injuries have been grouped together. They are put together into one level as the crashes that involve incapacitating injury could easily have been fatal and vice-versa possibly due to vulnerability of the subjects involved (Das et al., 2008). The other level includes the crashes with possible injuries and non-incapacitating injuries. A similar argument that a possible injury could easily have been a non-incapacitating injury and vice-versa depending on the subjects involved leads us to group the two categories together. Figure 7-8 shows the complete picture of the modeling concept adopted in the chapter. The first step in the analysis compares injury related crashes with no-injury. The second step (nested) analyzes the two broad groups of injury related crashes. This essentially carries out the classification of moderate injuries versus severe injuries.

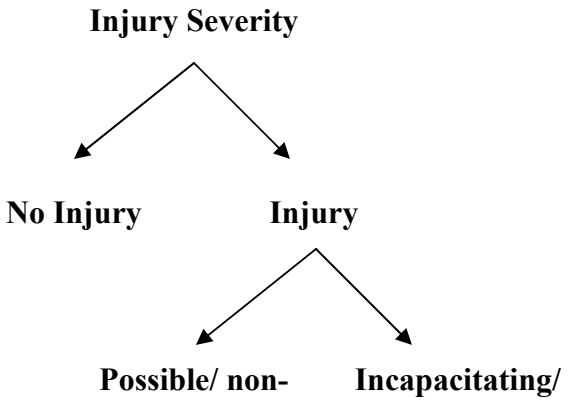


Figure 7-6 Nested Modeling concept

Each of the best programs chosen for the analysis in hand is a set of effective instructions which lead to the final classification rule. Typically for the classification problem the "Class 1 Hit Rate", "Class 0 Hit Rate" and the "Weighted Hit Rate (WHR)" for each of the best programs are provided. Once the criterion is chosen, the set of effective instructions (after the removal of introns) form the classification rule for that particular program. In the present study the WHR has been used as the criteria to select the classification model. The WHR reported is always for the validation dataset.

7.3.1.2 Angle / Turning Movement Crashes

This particular category of crashes includes all the angle crashes and also the left and the right turn crashes. As previously mentioned the corridors have been categorized into 4 clusters. Hence, the authors try to explain the results in light of the corridor clusters. This is critical to the

understanding of the results; especially the inclusion of the variables which enter the program's set of instructions.

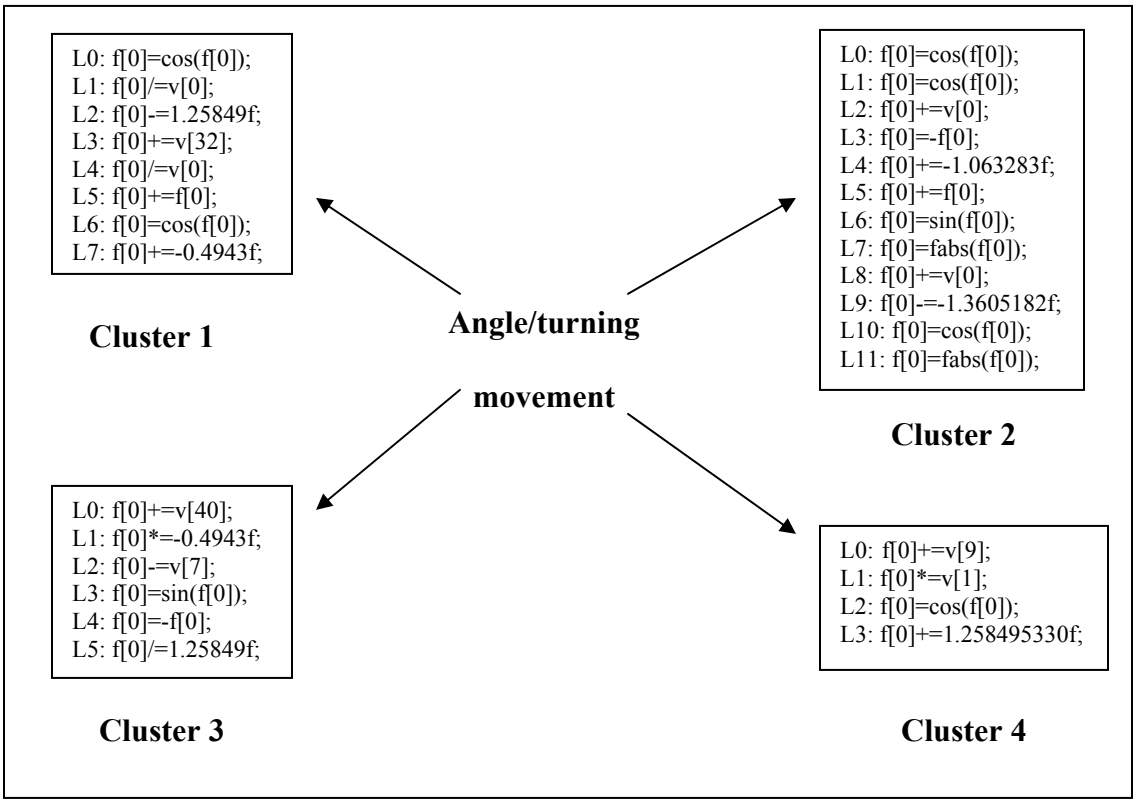


Figure 7-7 Non-injury / Injury classification rules for angle / turning movement crashes

The boxes in Figure 7-9 indicate the set of instructions (classification rules) that were developed for the particular cluster for the angle/ turning movement crashes for the injury and no-injury analysis. The classification rule (represented by 'f(0)' in the set of instructions) is developed line-by-line. The value of the function 'f(0)' is initialized to zero. At every step the information is updated through any arithmetic or trigonometric modification with either a variable (refer to Table 7-1 for all the variables appearing in the results) or a constant. The final value of f(0) is

then used to conduct the appropriate classification, based on a threshold value (in this case 0.5). The author also reports the WHR for all the programs mentioned here in the study.

To elaborate more, the authors explain one of the results, for example the result for Cluster 1, from Figure 7-9. The WHR for the program is 60.4106 which imply that 60.4106% of the cases were classified correctly. As mentioned earlier, at the start of the function the f(0) is initialized to zero. In the first line the cosine value of f(0) is computed. The resulting function is then divided by V_0 (surface width) followed by subtracting a constant and subsequently adding V_{32} (truck factor < 4.05) again to the function. The value of f(0) is thus calculated at every step and the final value is used for classification. In this study if the final value of f(0) is less than 0.5 then it is classified as a non-injury crash and as a injury crash, otherwise.

As mentioned earlier the corridors in Cluster 1 (1.009 – 2.89 miles) are the smallest in length. Crashes are most likely to be without injury if the surface width (V_0) is high. Higher surface width gives the driver more maneuvering space and thus more opportunity to take crash avoidance maneuver. Even if the crash does take place, it will mostly likely not to result in an injury. It is interesting to note that even low percentage of trucks on the corridors can result in injuries if a crash occurs. Seriousness of crashes with trucks and other vehicles has been well documented by Bjornstig et al. (2008). Interestingly in Cluster 2 (2.898 – 5.729 miles) (WHR = 57.8271) corridors higher surface width increases the likelihood of injury in a crash. In Cluster 3 (5.762 – 10.556 miles) (WHR = 57.7476) corridors, fair pavement condition (V_{40}) increases the possibility of injury. This indicates that pavement condition has to be good to excellent for safe

driving. Deteriorated pavements put the drivers at risk for a crash due to sudden unacceptable changes in the level and also due to poor traction. In Cluster 4 (10.644 – 78.293 miles) (WHR = 59.514) corridor injuries are more likely to occur when parking is available on higher speed limit segments (V_9* V_1). Emphasis on the restrictions of on-street parking has been highlighted in the work of Zegeer et al. (1994).

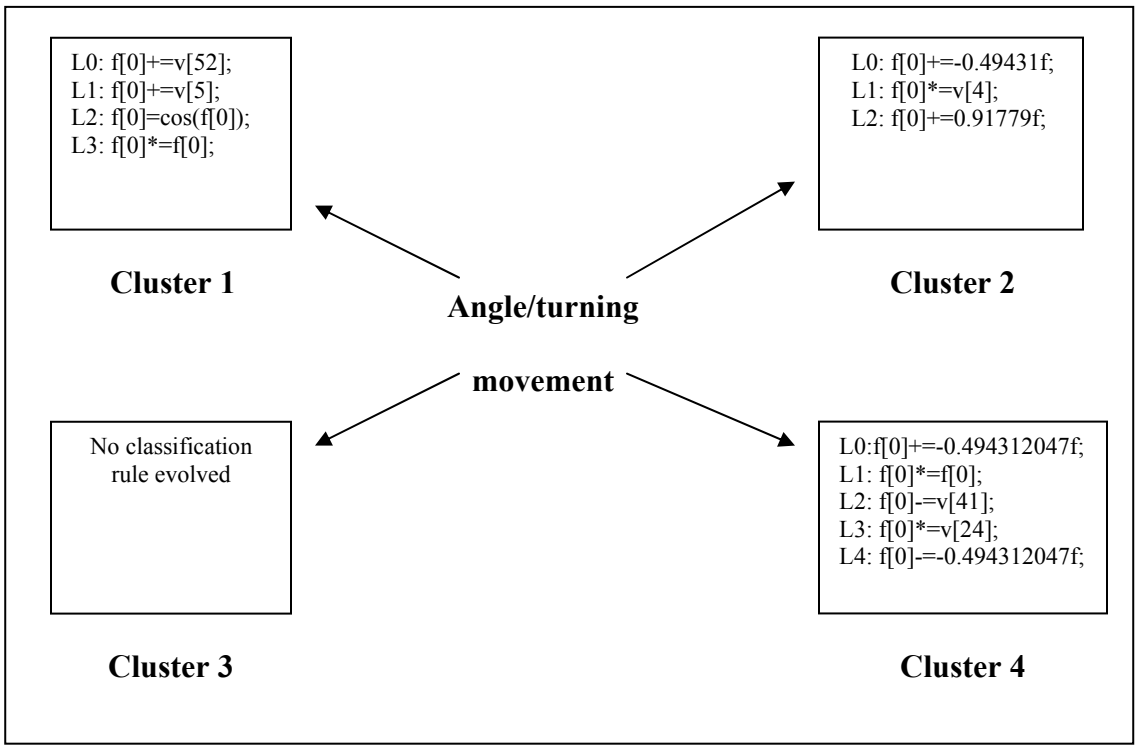


Figure 7-8 Non-severe / severe classification rules for angle / turning movement crashes

Figure 7-10 illustrates the results of the classification between severe and non-severe crashes. For Cluster 1 (WHR = 83.2677) crashes V_{52} (variable indicating the presence of a median with curb <= 6" and lawn) and V_5 (variable indicating dry surface condition) enter the classification rule developed. A careful observation at the entire rule for Cluster 1 indicate that the presence of

median with lawn and curb and also dry surface condition decrease the severity of the crash. The cosine function applied on f(0) reduces the value of f(0) when f(0) is higher. Das et al. (2008) also found dry surface conditions to favor less severe crashes probably because of resultant better friction the car is more in control. Hence, even if the crash occurs, the drivers could still be in control. The presence of lawn in the median could help in preventing multi-vehicle which more often results in severe crashes. In Cluster 2 (WHR = 81.944), as the variable V_4 (shoulder plus side walk width) increases the resulting crash tends to be less severe. Fatal crash rates are found to decrease with wider shoulder width (Kweon and Kockelman, 2005). Cluster 4 (WHR = 83.5804) results indicate that with good pavement condition (V_{41}) the crash severity will decrease. V_{24} (curve of roadway between 5^o and 8^o) also indicate low curvature. The entire rule indicates that with this curvature range the crashes occurring will be less severe. Souleyrette et al. (2001) found that the crash frequency had a direct association with the degree of curvature on horizontal surfaces.

7.3.1.3 Head-on Crashes

The results for the two types of analysis for the head-on crashes (one for injury and non-injury crashes; the other for severe and non-severe crashes) are illustrated in Figure 7-11 and Figure 7-12, respectively. In Cluster 1 (WHR = 70.1923) low skid values (V_{43}) result in increased likelihood of injury from a crash. Low skid values indicate poor traction control on roads which would increase the chances of loosing control of the vehicle during the event of a crash and thus leading to injury. Reduced friction could also lead to potentially dangerous head injuries on the

roadways (Finan et al., 2008). It is interesting to note that in Cluster 2 (WHR = 61.7116) the presence of non-restrictive median at sharper curves (V_{18}* V_{27}) lead to decreased probability of injuries. In Cluster 3 (WHR = 61.8879) paved median (V_{51}) is found to decrease the injuries. In Cluster 4 (WHR = 63.2472) crashes occurring during off-peak periods on roadways with surfaces other than blacktop (V_{38}/ V_{10}) decrease the injury probability. Results in Cluster 2 and 4 of head-on type crashes also indicate the capability of the GP methodology to discover interaction terms in the injury/no-injury classification.

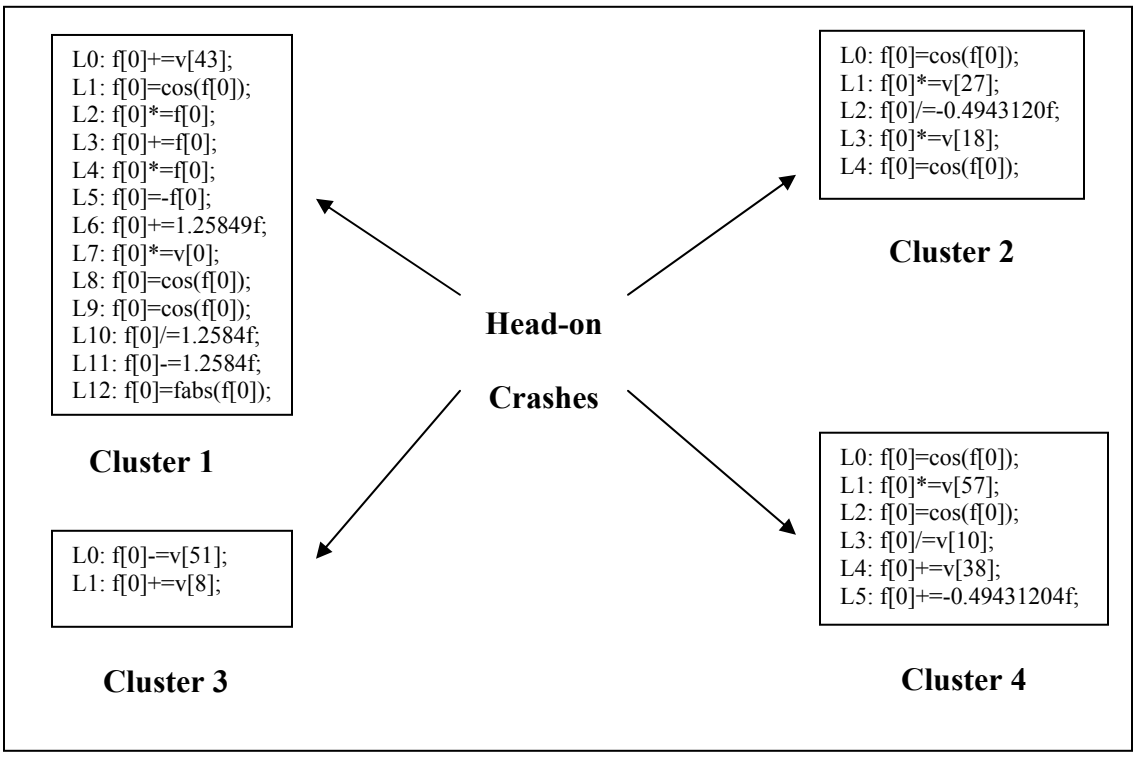


Figure 7-9 Non-injury / injury classification rules for head-on crashes

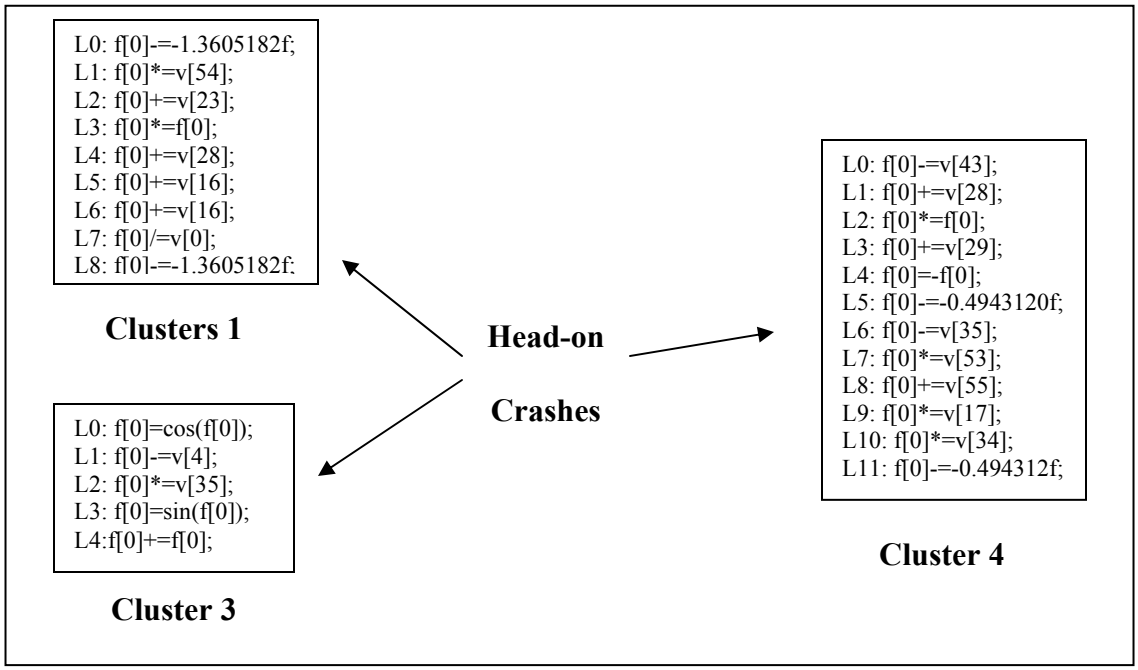


Figure 7-10 Non-severe / severe classification rules for head-on crashes

The results for the severity analysis are illustrated in Figure 7-12. In this analysis the Clusters 1 and 2 are combined to form one group (for the need of sufficient data). For Clusters 1 and 2 (WHR = 84.5273) variables like V_{23} (curvature between 4^o and 5^o) and V_{28} (ADT <= 31,000) increase the chances of severe crashes. Lower ADT means increased possible maneuvers during driving and hence the increased chances of potential conflicts. Lower ADT also indicates higher speeds, given a conflict occurs, and would potentially result in severe crashes. Restrictive openings in medians (V_{16}) also tend to increase the severity of crashes. However the crash severity would decrease with increase in surface width. This is in consistence with findings by Petritsch et al. (2007) who did an evaluation of geometric and operational characteristics for the safety of six-lane divided highways for the FDOT. Again in Cluster 3 (WHR = 82.1027), the presence of wide shoulder and side walk (V_4) decrease the severity of crashes. If the crash has

occurred during the afternoon peak period (V_{35}) then the resulting crash would be non-severe. The results are in line with a previous work by the author (Das et al., 2008). In Cluster 4 (WHR = 81.7513), again ADT less than 40,000 (V_{28} and V_{29}) leads to higher severity of injuries. As in Cluster 3, this cluster also has less severe injuries during afternoon peak traffic. Presence of curb and lawn median (V_{53}) helps avoid crossover head on crashes or reduce the intensity of it. Hence it would reduce the severity. If a head on crash occurs on the segment (V_{55}) then it would be more severe than if it would have occurred at any other roadway element. This could be attributed to higher vehicular speeds on segments than at intersections or access points. Restrictive opening and higher truck factor (V_{17} and V_{34}) results in higher severity of crashes. A study by Andreassen (2003) in Australia found that there are areas on corridors which should not have a higher truck percentage. Likewise the corridors with higher truck percentage should be flagged and more administrative measures should be taken to reduce the risk of crash occurrence and imminent severity due to crashes involving trucks.

7.3.1.4 Rear-end Crashes

The results for the two types of analysis for the rear-end crashes (one for injury and non-injury crashes; the other for severe and non-severe crashes) are illustrated in Figure 7-13 and Figure 7-14 respectively. In Cluster 1, the presence of paved and curbed median increase the likelihood of injury, while increase in maximum posted speed limit increase the probability of injury crashes on Cluster 2 corridors. In Cluster 3 rear-end crashes at intersections (V_{56}) are more injury prone even under good condition of the pavement (V_{41}). Surprisingly higher posted speed limits

tend to be safer in terms of injury occurrence for Cluster 4 corridors. One possible explanation could be that on longer stretches of roadway segments the driver gets used to the speed limit and after a while is more accustomed to the high speed traffic around it. Hence the injury probability might be reduced as the driver is more aware of the surrounding. The WHRs for the Clusters 1 through 4 are 56.3661, 53.0926, 52.495 and 54.1009 respectively.

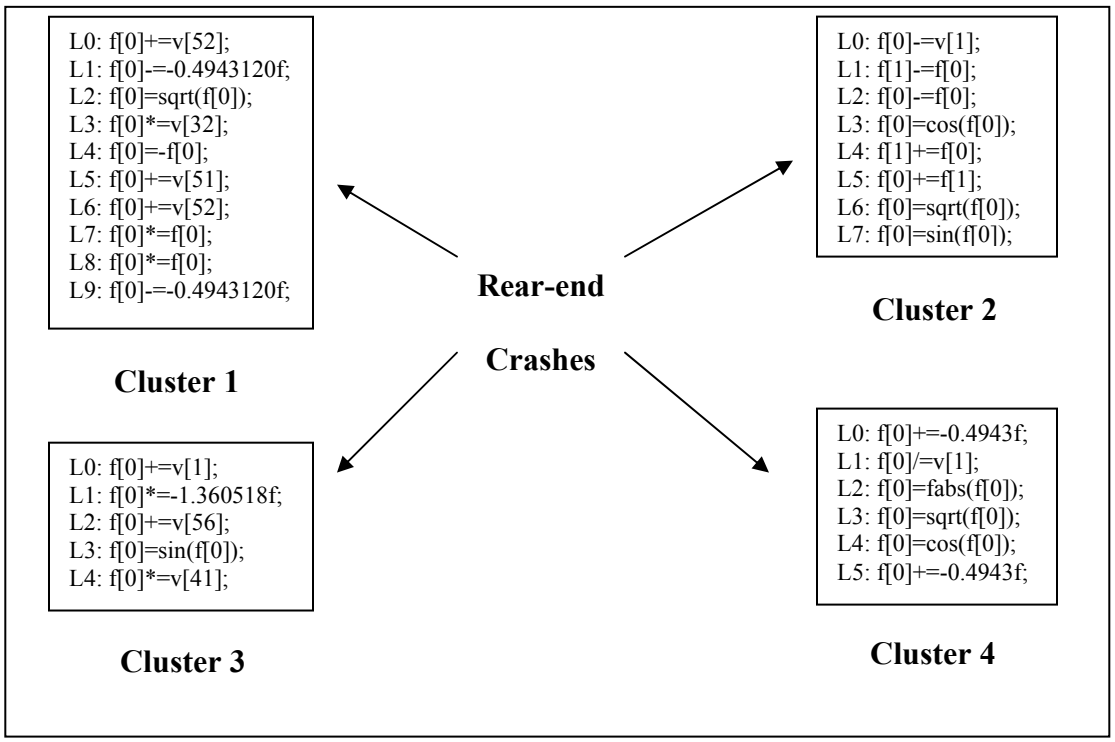


Figure 7-11 Non-injury / injury classification rules for rear-end crashes

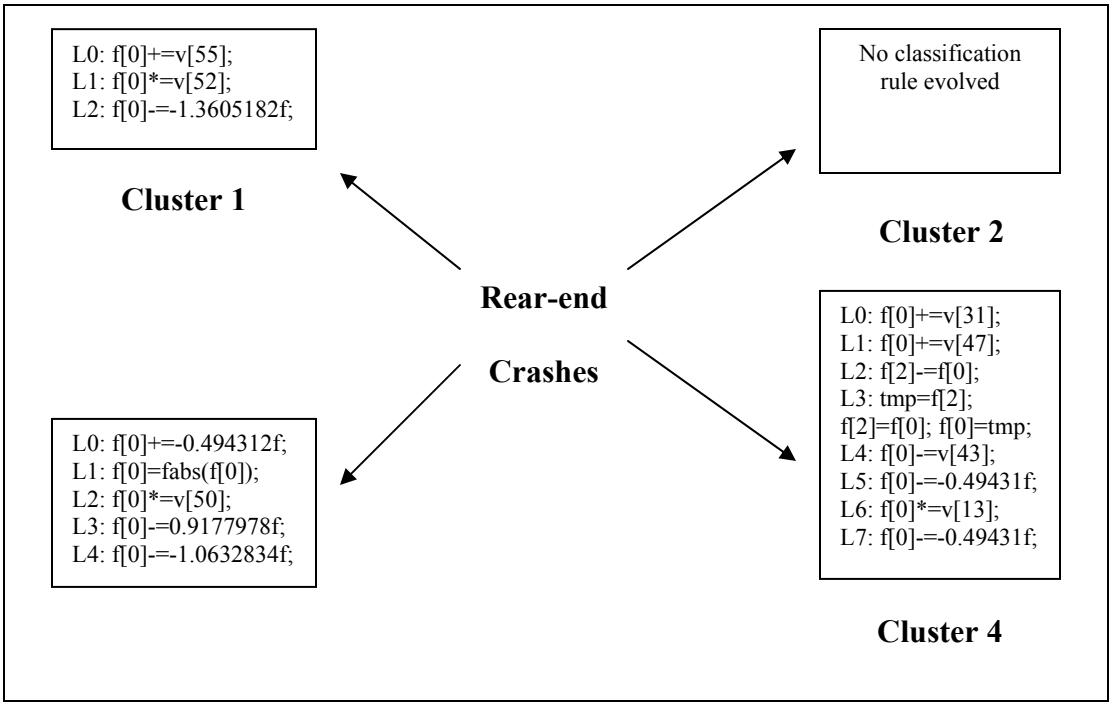


Figure 7-12 Non-severe / severe classification rules for rear-end crashes

In Figure 7-14 the results for the severity analysis are shown. In Cluster 1 (WHR = 91.8455), crashes related to segment (V_{55}) and on roadways with curb and lawn median (V_{52}) give rise to increased severity. In Cluster 3 (WHR = 91.4834) presence of lawn only median (V_{50}) leads to decreased severity of crashes. Lawn medians are generally wide medians. Wider medians lead to decreased crash rate (Gettis et al., 2005). Even though lawn medians are typically safer for head-on type of crash, yet the very presence of lawn medians can make the drivers make a move towards the lawn in case of imminent rear-end crash situation. This is different from the result obtained in Cluster 1, where curb and lawn median increase the severity. The presence of the curb makes it difficult for the driver to use the median space effectively for the drivers to avoid crashes. This could be a possible explanation as to why the crashes result in higher severity in

Cluster 1. In Cluster 4 (WHR = 89.4289) It was observed that V_{31} (ADT>= 52,000) causes increased severity of crashes. Thirty two percent of the crashes have speeds greater than 38 mph and thus indicating that a large number of vehicles were travelling at higher speeds (the number is large as the ADT is high). Thus crashes occurring at higher speeds would more likely lead to a severe crash. This indicates a higher speed variance. For the majority of slower vehicles (< 38 mph) severe crashes may occur due to the random aggressive behavior of drivers trying to make their way through a relatively low speed corridor. Nevarez et al. (2009) found ADT per lane to be significantly related to crash severity. A study by Pande and Abdel-Aty (2009) also finds severe rear-end crashes to be significantly related to ADT. A possible explanation to that could be the fact that the rear-end crashes, considered in this study, are occurring on high-speed arterials. In addition to that it must be observed that the severity of rear-end crashes is not entirely dependant on external factors. Rigid seat backs also contribute significantly to severity of injuries in rear-end crashes (Warner and Warner, 2008). Interestingly with the absence of street parking (V_{13}) the severity is found to diminish.

7.3.1.5 Concluding Remarks on Injury Severity Modeling

As stressed earlier in the study, classification is critical to our understanding of the variables of significance and their contribution to the safety problem at hand. In the present study the authors have set up a classification problem for the injury as well as severity of crashes. Typically in a classification problem the algorithm develops a set of rules which when followed leads to a particular category of the target variable. For example, in crash severity analysis when the binary

target variable represents severe/ non-severe crashes, the classification rule developed would lead to either severe crashes or non-severe crashes.

Classification using trees has been carried out since Breiman et al. (1984) came up with the Classification and Regression Tree (CART) algorithm. Different algorithms have been tried ever since to develop classification models or rules. The advantage or the feature that gives genetic programming the edge over any other existing classification algorithm is the fact that numerous models can be developed for the same dataset. The use of the concept of biological evolution helps the algorithm develop numerous models (by its capacity to perform multiple runs with randomized parameter settings), through the operators like crossover and mutation. A lower crossover frequency and a higher mutation frequency are implemented to prevent genetic drift from taking place. Genetic drift is the accumulation to a sub-optimal solution in the search space due to stochastic errors. The process of mutation always brings in novelty to the population of evolved generations. GP can also assemble teams of models than just individual models which makes it better than most classification algorithms which primarily work on just individual models. The individual models or teams model have been observed to have a lower error rate than other standard classification algorithms. Percent correct classification achieved on the validation data set for severe/non severe models were as high as 90% and more as indicated by the WHR values.

As mentioned earlier the two types of analyses carried out in the study includes: 1) injury and non-injury crashes; and 2) severe and non-severe crashes. Some of the results confirm to the

traditional well established patterns where as certain other results are not so common and do not confirm to convention. For angle/ turning movement crashes presence of parking and higher posted speed limits are responsible for more injury related crashes. Even low percentage of trucks can increase the chance of injury prone crashes. 'Curb and lawn' median and dry surface conditions decrease the severity of crashes where as poor pavement condition result in more severe crashes. Wider shoulders along with sidewalk also tend to make the roads safer from a severity point of view.

In case of head on crashes low ADT and median openings are the leading operational and geometric factors for severe crashes. Again wide shoulder and sidewalk result in less severe crashes. Crashes occurring on afternoon weekday peak periods also tend to be less severe. Lower skid resistance and the presence of 'curb and lawn' medians are again found to diminish the severity of the crash. Higher truck factor also results in increased severity of head-on crashes. Low skid values increase the injury probability of a crash while crashes occurring during off-peak periods are less injury prone.

Rear-end crashes at intersections are more likely to be injury prone as well as those at paved and curbed median segments of the roadways. Unlike the angle/ turning movement and the head-on crashes, the 'lawn and curb' median causes increased severity in rear-end crashes and similarly for higher ADT values. Absence of street parking also decreases the severity of rear-end crashes.

The results from the genetic programming classification are intuitive and their association with severity may be explained. Certain known results about severity of crashes have been confirmed while some new information is discovered about others. The 'lawn and curb' median are found to be safe for angle/ turning movement crashes and not so safe for rear-end crashes. Vision obstruction is a leading factor of severe crashes. Dry surface conditions, good pavements also reduce the severity of crashes. On-street parking, higher posted speed limits and lighting conditions do play a role in both injury related crashes and severe crashes.

It can be observed from the results that a lot of interaction terms are discovered in the classification approach for injury/no-injury and severe/non-severe crashes. The heuristic approach that GP applies has been observed to shed new light on the interaction between variables discussed in this study.

As it can be observed most of the variables of concern relate to geometric and operation factors. Event specific variables have not been included in this study for the sake of interpretability, generalization and the objectives of this study. However, it should be noted that the analysis could be carried with only those variables or by mixing them with geometric and traffic parameters. This could be a part of future investigation. On-street parking has been found to be a hazard for severe injury. Steps should be taken to either remove the facilities for parking or in the case where it is not possible, to restrict the parking hours. Pavement condition should be improved and wherever possible, 'curb and lawn' median should be designed. Higher truck percentage is found to increase severity; hence steps such as lane restriction for trucks or

rerouting them from flagged corridors should be taken. Betterment of lighting conditions on the roadways is always desired. Vision obstruction has traditionally been a problem; that however, is not only due to external factors. Nevertheless, transportation authorities should always take design initiatives for the drivers to have a clear view of the surroundings.

7.3.2 Crash Frequency Modeling

7.3.2.1 Data Preparation

Since occurrence of crash is a random event, all factors remaining constant, any given point has an equal probability of crash occurrence. Hence, the investigator divided the urban arterials into equal sections of length 0.5 miles, unlike the clusters of corridors prepared for the classification problem. In this part of the study only specific roadway geometric and design factors have been used, information for which, were completely available. Corridors with no crashes have also been included. The traffic and roadway geometric parameters for those corridor sections, the data for which was not available, have been imputed with data points generated randomly within the available range. Table 7-2 lists all the variables used in this analysis.

Table 7-2 Dependent / Independent variables used in crash frequency modeling

Variable Name	Variable Symbol	Description
Target or Dependent Variable		
Freq		Frequency of crashes
Continuous Independent Variables		
Mean_surface_width	V_0	Surface width (Continuous)
Mean_shld_width1	V_1	Shoulder width (Continuous)
maxSpeed	V_2	maximum posted speed limit (Continuous)
mean_sec_adt	V_3	ADT (Continuous)
mean_avg_t	V_4	Truck factor (Continuous)
mean_skid	V_5	Skid resistance (Continuous)
Categorical Independent Variables		
dry_surface_cond	V_6	Indicator for dry surface condition (Binary)
day_light_cond	V_7	Indicator for daylight crashes (Binary)
clear_weather	V_8	Indicator for clear weather crashes (Binary)
blacktop_surface_type	V_9	Indicator for blacktop surface (Binary)
no_defects_on_road	V_{10}	Indicator for no defects on roadway (Binary)
vision_no_obs	V_{11}	Indicator for vision obstruction (Binary)
MPW	V_{12}	Indicator for crashes in weekday morning peak (Binary)
APW	V_{13}	Indicator for crashes in weekday afternoon peak (Binary)
FSN	V_{14}	Indicator for crashes on Friday/ Saturday night (Binary)

Typically for regression analyses the R^2 value and error rate is reported. In this study the researcher chose the models with least error. Also, the same programs had the highest R^2 value. The crash analyses were carried out for three separate site locations, namely: 1) segments; 2) signalized intersections; and 3) access points. Separate models for mid-block segment related, signalized intersection related and access point related crashes for each crash type could help assess the safety situation better than by just having one model for all the roadway elements. For example, rear-end crashes on mid-block segments and signalized intersections, may have similar set of significant variables but the model form might be different. Angle crashes on mid-block segments might be affected by different roadway elements compared to the angle crashes on signalized or un-signalized intersections. Unilateral assumption of one model form explaining the crash occurrence phenomenon on all locations certainly makes the problem simpler but limits

the scope of understanding. The GP methodology used in this study has enabled the authors to establish different model forms for the different roadway elements. The overall model development structure is given in Figure 7-15. All the models were compared to the traditional Negative Binomial (NB) model. The mean square error (MSE) has been used as the metric to compare GP and NB models.

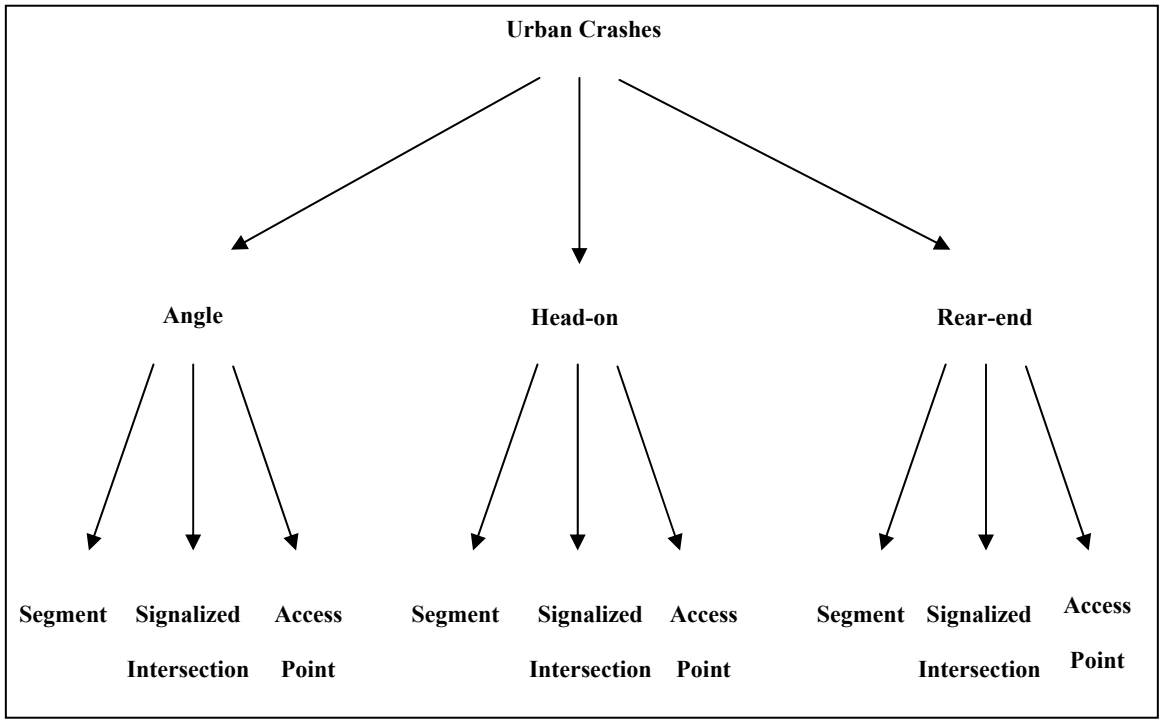


Figure 7-13 Overall model development structure

In the following sub-sections, which would discuss on the various models for different types of crashes, tables will be presented highlighting the percentage use of variables given in Table 7-2.

The percentage use is indicative of how frequently a particular variable has been used in the overall model development process.

7.3.2.2 Angle / Turning Movement Crashes

Most angle crashes occur near signalized intersections as compared to access points and mid-block segments. As previously argued the model form for the angle crashes should be different for the different roadway elements. In order to maintain uniformity the author first reports the mid-block segment related crash model followed by the signalized intersection related crashes and access point related crashes' models. The segment related crashes' model is given in Equation 7-1. The MSE for the model is 0.921 while that of corresponding NB model is 0.950.

Equation 7-1

$$f(0) = \sqrt{\left[\left(\frac{E}{D} - V_9 - V_6\right)^2 + V_{11}\right] \times 2^D \times V_{10}}$$

Where,

$$E = \left(\frac{2\left|5\left(a_1 + A^2\right) \times 2^{-A}\right|}{V_0} + A\right) \times 2^{-A} \times \frac{V_7}{a_2 \times V_0}$$

$$D = -A - \frac{\left|5\left(a_1 + A^2\right) \times 2^{-A}\right|}{V_0}$$

$$A = V_{12} + V_{13} + V_{14}$$

a_1=-1.550; a_2=0.176

Table 7-3 gives the percentage use of the variables in the overall development of the segment model for angle/ turning movement crashes.

Table 7-3 Variable use for segment model of angle/ turning movement crashes

Variable	Frequency
V_0	0.97
V_1	0.10
V_2	0.23
V_3	0.37
V_4	0.13
V_5	0.27
V_6	0.97
V_7	0.40
V_8	0.30
V_9	0.73
V_{10}	1.00
V_{11}	0.90
V_{12}	1.00
V_{13}	1.00
V_{14}	1.00

The signalized intersection model is as follows:

Equation 7-2

$$f(0) = \left| \left(2 \left| \frac{\sqrt{V_3 \times (1 + V_7 \times V_8)} - V_5}{V_2} - V_{12} - V_{13} \right| - 2V_{13} + a_1 - 2V_{12} \right) \times V_{11} \right| \times V_6 \times V_9$$

Where,

a_1=-0.559

The signalized intersection related crashes' GP model had a MSE of 29.794 while the traditional NB model for this case is 29.354. The frequency of use of the variables is given in Table 7-4.

Table 7-4 Variable use for signalized intersection model of angle/ turning movement crashes

Variable	Frequency
V_0	0.83
V_1	0.23
V_2	0.37
V_3	0.27
V_4	0.07
V_5	0.60
V_6	1.00
V_7	1.00
V_8	0.80
V_9	1.00
V_{10}	0.77
V_{11}	1.00
V_{12}	1.00
V_{13}	1.00
V_{14}	0.50

The access point related crashes' model is given in Equation 7-3. The MSE is observed to be 18.220 while the NB model outperforms with a MSE of 17.508.

Equation 7-3

$$f(0) = 2\left(\frac{V_3 \times V_9 \times a_1^5}{V_5} - V_{13}\right) V_7 \times V_{11} \times V_6 - V_{12} \times V_{11} \times V_6 + 2a_2 - V_{12} - 2V_{14}$$

Where,

a_1=0.288; a_2=1.501

Table 7-5 indicates the frequency of use of variables in the development of the above model.

Table 7-5 Variable use for access point model of angle/ turning movement crashes

Variable	Frequency
V_0	0.83
V_1	0.40
V_2	0.07
V_3	0.40
V_4	0.13
V_5	0.30
V_6	1.00
V_7	0.97
V_8	0.63
V_9	0.87
V_{10}	0.40
V_{11}	0.90
V_{12}	0.93
V_{13}	0.93
V_{14}	0.67

The models represented by Equation 7-1 through Equation 7-3 exhibits the complex relationship governing the angle crash occurrence at the different roadway element. The higher surface width (V_0) leads to less angle crashes at mid-block segments. Higher surface enables the driver to have more space for maneuvers that could lead to crash avoidance. Afternoon and morning peak periods in conjunction with no obstruction of vision and dry surface condition on blacktop surface roads leads to fewer angle crashes at signalized intersections. Higher skid values (V_5) leads to fewer angle crashes at both signalized intersections and access points. This result is intuitive and is supported by past research (Noyce et al., 2007). Higher traffic volume (V_3) intuitively reflects more number of angle crashes and the results confirm to it. Lower maximum posted speed limit (V_2) leads to higher number of angle crashes at signalized intersections which might indicate the presence of aggressive drivers on the road.

7.3.2.3 Head-on Crashes

Head-on crashes tend to occur when opposing traffic are close in lateral displacement. The mid-block segment crash model is presented in Equation 7-4. MSE value of 0.1 was observed for the GP model as compared to a MSE value of 0.236 of the equivalent NB model.

Equation 7-4

$$f(0) = \frac{\left(D \times 2^{C} - a_5\right)^2}{a_6}$$

Where,

$$D = \sqrt{\left[\cos\left(\sqrt{A+B}\right) - a_3\right]^2 \times V_9 + a_4 - V_{14}}$$

$$C = \sin\left(\frac{a_1 - A}{a_2} - V_5\right) + V_5 + V_7$$

$$B = A - 1 - a_1$$

$$A = \cos\left\{\left[2 + 2\cos(1) \times V_4 \times V_{10}\right]^2\right\}$$

a_1=-1.924; a_2=0.139; a_3= -0.184; a_4=0.723; a_5=0.292; a_6=1.253

The frequency of use of variables for the segment model of head-on crashes is given in the Table 7-6

Table 7-6 Variable use for segment model of head-on crashes

Variable	Frequency
V_0	0.13
V_1	0.17
V_2	0.30
V_3	0.17
V_4	0.50
V_5	0.47
V_6	0.17
V_7	0.43
V_8	0.17
V_9	0.97
V_{10}	1.00
V_{11}	0.53
V_{12}	0.67
V_{13}	0.57
V_{14}	1.00

The signalized intersection related crashes' model is as given in Equation 7-5.

Equation 7-5

$$f(0)=\left[\sin^2\left(\frac{C^2}{a_8^2}\right)-2\frac{A}{B}\sin\left(\frac{C^2}{a_8^2}\right)\right]\times V_{11}$$

Where,

$$C=\frac{A}{B}-V_6\times V_8\left(\frac{V_1-\frac{A}{B}}{a_7\times V_5}+V_9-V_{12}\right)+V_{13}+a_6$$

$$B=\left[\frac{\{A+\cos(2A)\}^2}{a_3}-a_4\right]\times V_0\times a_5-a_6$$

$$A=V_4\times V_{10}\left[a_1\times V_4\times\sin\left(V_4\times 2^{V_4}\right)-V_{14}-a_2\right]$$

a_1=0.288; a_2=-0.494; a_3=1.249; a_4= -0.730; a_5=0.106; a_6= -1.238; a_7=0.995; a_8= -1.924

The above GP model given by Equation 7-5 has a MSE of 0.06 while the NB model has a MSE of 0.154. Table 7-7 shows the values for frequency of use of variables in the above model.

Table 7-7 Variable use for signalized intersection model of head-on crashes

Variable	Frequency
V_0	0.63
V_1	0.33
V_2	0.17
V_3	0.07
V_4	0.53
V_5	0.27
V_6	1.00
V_7	0.33
V_8	1.00
V_9	1.00
V_{10}	1.00
V_{11}	0.97
V_{12}	0.93
V_{13}	0.80
V_{14}	0.33

The access point related crash model is given Equation 7-6.

Equation 7-6

$$f(0) = \left| H \times \left(1 - \frac{V_{14}}{a_5} \right) \right|$$

Where,

$$H = a_2 + G \times \left(D + F + G - V_7 \times V_{10} \right)$$

$$G = \left(V_6 \times F^2 - 3a_2 \right) \times V_9$$

$$F = \left[\left(\left| \cos\left(2^D \times E \right) \right| + a_4 \right) \times V_{10} \times V_{11} \right]^4 \times V_8^2$$

$$E = a_3 \times \frac{V_1}{V_2} \times \left[a_2 + V_{14} - 2^{A+B} \times V_{10} \times V_{11} \times (C - a_2)\right] + V_{13} - V_1$$

$$D = V_{14} + a_2 + V_{10} \times V_{11} \times 2^{A+B} \times (A + B - 1)$$

$$C = \left\{\left[\cos\left(-2^{A+B} \times (V_1 + B)\right)\right]^2 + V_9\right\}^2$$

$$B = (V_{12} + V_{14}) \times (V_{12} - V_0) + V_{13}$$

$$A = \frac{2\left[V_{14}(V_{12} - V_0) + V_1 - a_1\right]^2}{V_3}$$

a_1=1.505; a_2=-0.091; a_3=1.531; a_4=0.033; a_5=1.048

Table 7-8 presents the frequency of use for the variables used in the model development process for access point related head-on crashes.

Table 7-8 Variable use for access point model of head-on crashes

Variable	Frequency
V_0	0.33
V_1	0.23
V_2	0.53
V_3	0.17
V_4	0.70
V_5	0.10
V_6	0.40
V_7	0.33
V_8	0.43
V_9	1.00
V_{10}	1.00
V_{11}	1.00
V_{12}	0.60
V_{13}	0.57
V_{14}	1.00

The MSE of the GP model is 0.081 while the NB model has a MSE of 0.262. Models represented by Equation 7-4 through Equation 7-6 for the head-on crashes shows that traffic conditions during the Friday and Saturday night peak period (V_{14}) reduce the frequency for head-on crashes. This goes against intuition as this is the time of the week where drivers will be mostly speeding or could be under the influence. A possible explanation could be that a lower ADT on the corridors and the data reflects that 50% of the roadways have an ADT of less than 12,950. Higher average truck factor (V_4) increases the instances of head-on crashes on mid-block segments and signalized intersections. In a previous work it had been investigated that increased percentage of trucks leads to fatal head-on crashes (Abdelwahab and Abdel-Aty, 2004). Day light and good road condition decreases the instances of head on crashes near access points. At signalized intersections, the higher the surface width the higher the number of head-on crashes. The frequency of head-on crashes is observed to increase with higher skid resistances values on mid-block segments (V_5). The same observation is also reported during day light conditions (V_7).

7.3.2.4 Rear-end Crashes

The mid-block segment related crash model for rear-end crashes given in Equation 7-7 reveals the complex structure of the crash occurrence phenomenon. The MSE of the model is 11.234 where as the corresponding NB model has an MSE of 12.615.

Equation 7-7

$$f(0) = \left[\left|\left(B - V_6\right) \times V_6 - a_4\right| \times V_{10} + V_7 - V_{13}\right| + V_6 - V_{12}\Big] \times V_{11}$$

Where,

$$B = \left| \left[\left\{ \left(A + V_7 - V_{12} - V_{13} \right) \times a_3 - a_4 \right\} \times a_5 \right| + V_7 - V_{13} - a_4 - V_{12} \right] \times V_9 - V_{13} \Big|$$

$$A = \left| \left[\left\{ \left(\left| a_1 V_3 \right| \times a_1^{\,2} + a_2 \right) \times a_1 - V_{12} \right\} \times a_3 - a_4 \right] \times a_5 \right|$$

a_1=0.105; a_2=-0.828; a_3=1.048; a_4=1.779; a_5=1.253

Table 7-9 for frequency of use of variables used in the above model development is given below.

Table 7-9 Variable use for segment model of rear-end crashes

Variable	Frequency
V_0	0.50
V_1	0.03
V_2	0.10
V_3	1.00
V_4	0.23
V_5	0.40
V_6	0.90
V_7	0.87
V_8	0.07
V_9	0.30
V_{10}	0.63
V_{11}	1.00
V_{12}	0.90
V_{13}	1.00
V_{14}	0.23

The signalized intersection related crash model is as given by Equation 7-8. The MSE for the GP model is 57.632. The NB model in this particular case has a MSE of 81.808.

Equation 7-8

$$f(0) = \left| \frac{a_3 \left(B \times V_6 \times V_9 + V_3 \right)}{V_5} - 2V_{13} + V_{10} \right| - 2V_{12} + V_{10}$$

Where,

$$B = \left\{\left[\left(1 - a_1\right) \times A - V_3\right] \times 2^A \times a_2 + 2V_3\right\} \times V_7 \times V_8 + V_3$$

$$A = 2V_{10} - V_{12} - V_{13}$$

a_1=0.995; a_2=1.984; a_3=0.003

Table 7-10 representing the frequency of use of the signalized intersection model is given below.

Table 7-10 Variable use for signalized intersection model of rear-end crashes

Variable	Frequency
V_0	0.60
V_1	0.33
V_2	0.63
V_3	1.00
V_4	0.20
V_5	0.73
V_6	0.97
V_7	0.73
V_8	0.90
V_9	0.23
V_{10}	0.93
V_{11}	0.70
V_{12}	0.27
V_{13}	0.70
V_{14}	0.23

Finally, the access point related model for the rear-end crashes is:

Equation 7-9

$$f(0) = \left(a_8 \times B \times V_0 - V_{12} + V_7 + V_6 - V_{13}\right)^2 \times \frac{V_{11}}{a_9} \times V_9 - V_{12} + a_{10}$$

Where,

$$B = a_1 \times V_{10}\left[2a_6\left(a_4 \times A - v_0 + a_3\right) - V_8 - a_7\right]$$

$$A = \sqrt{\left(\frac{\left|a_1 \times V_5 - V_{11}\right| + a_2 + V_8}{a_3}\right) \times V_3} - V_2$$

Where

a_1=0.003; a_2=1.048; a_3=1.451; a_4=0.317; a_5=1.063; a_6=0.136; a_7=0.918; a_8=1.501; a_9=1.988; a_{10}=0.634. The frequency of use of variables is given in Table 7-11.

Table 7-11 Variable use for access point model of rear-end crashes

Variable	Frequency
V_0	0.70
V_1	0.40
V_2	0.23
V_3	0.43
V_4	0.17
V_5	0.33
V_6	1.00
V_7	0.97
V_8	0.70
V_9	0.97
V_{10}	0.77
V_{11}	0.97
V_{12}	1.00
V_{13}	1.00
V_{14}	0.27

For the above GP model given by Equation 7-9 the MSE is reported to be 10.264. In this case the NB model was found to perform marginally better with an MSE of 10.064. ADT (V_3) is significant in all the three model forms along with dry surface condition (V_6) and day light condition (V_7). The dry surface conditions probably indicate fine weather and more vehicles on the road. Hence improper maneuvers could result in higher number of collisions. Research has

shown that slippery road conditions lead to a higher probability of crash avoidance maneuvers as drivers would drive more cautiously during unfavorable conditions (Yan et al., 2008). It is interesting to note that when there is no vision obstruction (V_{11}) and during day time the rear-end crash frequency increases on segments and access points but not at signalized intersections. The same effect is also found for higher ADT. In any stretch of road where the ADT does not vary much, the result could be difficult to explain. However, speed variances will be more prominent in the mid-block segments and the access points than near the signalized intersections which increase crash risk (Pande et al., 2005). The ADT coupled with the speed variances could be a possible cause of the observation. Morning and afternoon peak periods (V_{12}, V_{13}) are observed to have fewer occurrences of rear-end crashes at all the roadway elements. A particular interaction term reflects that roadways with no defects (V_{10}) and higher surface width (V_0) would have less rear-end crashes near access points at higher posted speed limits (V_2). For convenience of the reader Table 7-12 presents all the variables entering the final models discussed in the study.

Table 7-12 Variables entering the various GP models

		Variables
Angle	**Segment Related**	V_0, V_6, V_7, V_9, V_{10}, V_{11}, V_{12}, V_{13}, V_{14}
	Signalized Intersection Related	V_2, V_3, V_5, V_6, V_7, V_8, V_9, V_{10}, V_{11}, V_{12}, V_{13}
	Access Point Related	V_3, V_5, V_6, V_7, V_9, V_{11}, V_{12}, V_{13}, V_{14}
Head-on	**Segment Related**	V_4, V_5, V_7, V_9, V_{14}
	Signalized Intersection Related	V_0, V_1, V_4, V_5, V_6, V_8, V_9, V_{10}, V_{11}, V_{12}, V_{13}, V_{14}
	Access Point Related	V_0, V_1, V_3, V_6, V_7, V_8, V_9, V_{10}, V_{11}, V_{12}, V_{13}, V_{14}
Rear-end	**Segment Related**	V_3, V_6, V_7, V_9, V_{10}, V_{11}, V_{12}, V_{13}
	Signalized Intersection Related	V_3, V_5,V_6, V_7, V_8, V_9, V_{10}, V_{12}, V_{13}
	Access Point Related	V_0, V_2, V_3, V_5, V_6, V_7, V_8, V_9, V_{10}, V_{11}, V_{12}, V_{13}

Table 7-13 depicts the MSE values of the GP models as well as the traditional NB models developed. The highlighted models are the ones with the lower MSE values.

Table 7-13 Observed validation dataset MSE for GP and NB models

		GP	NB
Rear-end	**Segment Related**	11.234	12.615
	Signalized Intersection Related	57.632	81.808
	Access Point Related	10.264	10.064
Angle	**Segment Related**	0.921	0.95
	Signalized Intersection Related	29.794	29.354
	Access Point Related	18.220	17.508
Head-on	**Segment Related**	0.1	0.236
	Signalized Intersection Related	0.06	0.154
	Access Point Related	0.081	0.262

7.3.2.5 Concluding Remarks on Crash Frequency Modeling

The methodology, which allows the researcher to choose among millions of programs, develops the models based on the simple concept of evolutionary biology. The developed models not only explain the data better but also show how the variation in parameter values affects the crash occurrence. The GP methodology uses the concepts of crossover and mutation to evolve models over time and available population. A lower crossover frequency and a higher mutation frequency are implemented to prevent genetic drift from taking place. Genetic drift is the accumulation to a sub-optimal solution in the search space due to stochastic errors. The process of mutation always brings in novelty to the population of evolved generations. GP can also assemble teams of models than just individual models which makes it better than most regression algorithms which primarily work on just individual models.

As can be observed, all the nine successfully developed GP models show the complex structure governing the phenomenon of the crash occurrence. The GP outperform the NB in terms of lower MSE values (for validation dataset only) in six of the nine models discussed in the study. A paired t-test for the MSE values for all the GP and NB models showed that there was a significant decrease in MSE values (83% confidence level). One of the reasons could be that the NB models work better with all continuous variables and in this present study we had a mix of continuous as well as categorical variables. The GP model builder which essentially evaluates through a host of models gives a new alternate non-linear structure even with categorical variables. As mentioned earlier the GP method optimizes the selection of the regression models over multiple runs and the selected model is then used for crash frequency prediction.

The parameter behavior discovered may not always be intuitive and the authors have tried their best to explain the plausible causes. For examples, the dry surface condition leads to increase frequency of crashes. Even under perfect conditions of driving, the driver may loose focus, thus resulting in improper maneuvers leading to crashes. Similarly it was regularly observed that crashes occur less during the dark hours. Though on roadway geometric, traffic and environmental factors are used in model development, the results also help in understanding driver characteristics. For example, Lower maximum posted speed limit leads to higher number of angle crashes at signalized intersections which indicates the presence of aggressive drivers on the road. The complex interaction terms appearing in the models are crucial to increased prediction accuracy, although they make the models more difficult to explain.

For any particular crash type, no single model based on one particular distribution could provide enough insight to understand the safety situation. The GP modeling approach gives the researcher independence for model development without restrictions of the distribution of data. The developmental structure (see, Figure 7-15) allows for separate models for mid-block segment related, signalized intersection related and access point related crashes for each crash type. This is important as the relationship of parameters explaining the frequency of any crash type on mid-block segments should be different from those occurring on signalized intersections or access points. However, the system of models developed allows cross referencing of the results among the roadway elements. With this twin approach the authors have been successful in building models which have proved to be a good alternative to the traditional methods of crash frequency estimation.

Readers could question the purpose of the comparison of the GP model with the basic NB model. The author would like to clarify that the purpose of the study is to introduce GP modeling into prediction of crash estimates and are not delving into advanced NB models (Lord et al., 2005; Miaou and Lord, 2003).

CHAPTER 8. GRAPHICAL PERCEPTION AND SENSITIVITY ANALYSES

8.1 Graphical Understanding and Introduction to Sensitivity Analysis

It is observed from the crash frequency modeling results in the previous chapter (see 7.3.2 Crash Frequency Modeling) that few of the continuous variables used in the model development process. As can be noted, the complexity of the models requires visual realization of the relationship of the crash frequency with varying input parameters. This chapter deals with the change in the frequency of crashes as the value of any particular continuous variable changes. Determining how important a parameter is to a function is different from understanding how the function changes with variation in the parameter. Sensitivity analysis is useful in determining the significance of a variable to a function. The absolute sensitivity S_i of a function $f(x_1,\ldots,x_n)$ towards a variable x_i is given by Equation 8-1 (Smith et al., 2008):

Equation 8-1

$$S_i = \frac{\partial f}{\partial x_i}$$

The absolute sensitivity is calculated at a normal operating point, which in the present study is the point when all other variables assume mean value. The absolute sensitivity value indicates which variable has the maximum effect on the result, for a constant change in the parameters. However, the relative sensitivity value (Equation 8-2) indicates which variables would have the

most impact on the variation of the output (Saltelli, et al., 2000). In this research the relativity sensitivity has been implemented.

Equation 8-2

$$S_i = \frac{\partial f}{\partial x_i} \times \frac{std(x_i)}{std(f)}$$

The mathematical formulation for Equation 8-2 requires derivatives to be evaluated. Hence, in the present case the relative sensitivity towards continuous input variables can only be assessed. The binary variables which do not satisfy the limit conditions cannot be assessed in the study.

8.1.2 Angle Crashes

The segment model for angle / turning movement crashes, given by Equation 7-1, has surface width (V_0) as the only continuous variable entering the model. The plots in Figure 8-1 show that as the surface width increases the frequency of crashes decrease for the morning and the afternoon peaks. The frequency increases during the Friday and Saturday night peak conditions. The results are also fixed for mean and/or modal values of the other input variables. In Figure 8-2 it can be observed that during off peak periods of the day the crash count decreases or increases with surface width depending on day light or no-day light condition respectively. The graphs were plotted using Mathcad^{TM} (version 14). In Figure 8-1 and Figure 8-2 frequency of crashes is on the y-axis while surface width (V_0) is plotted on the x-axis.

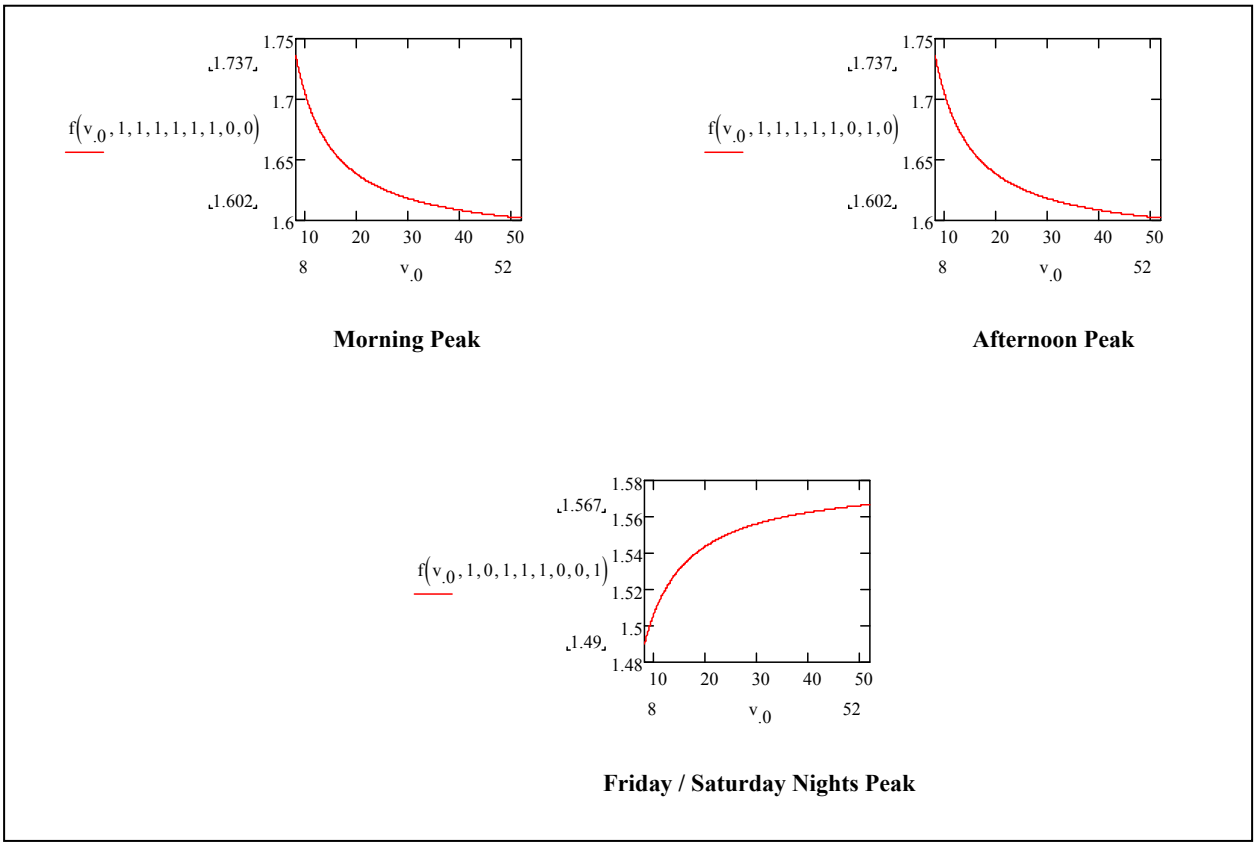


Figure 8-1 Crash Frequency versus Surface Width at different peak periods

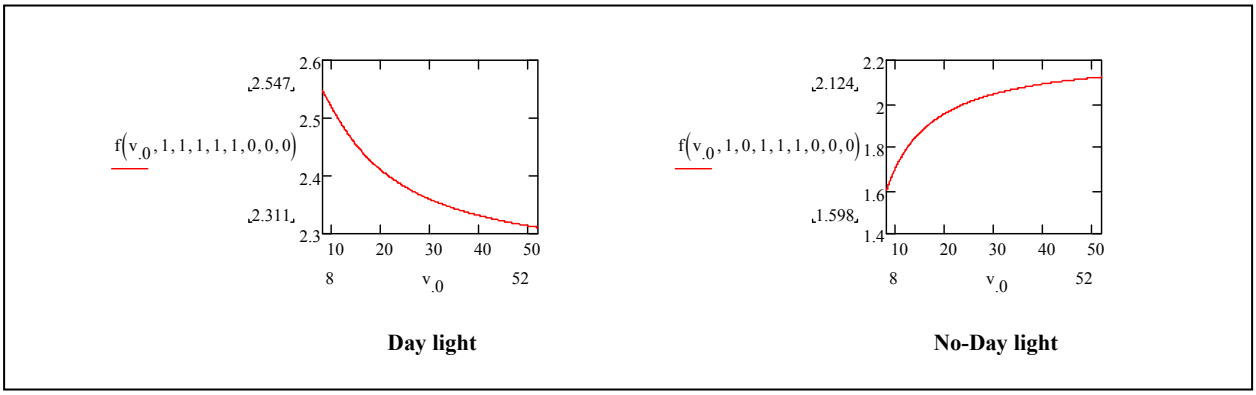


Figure 8-2 Crash Frequency versus Surface Width for Off Peak periods

As mentioned earlier the segment model has only surface width (V_0) as the continuous variable. The model's relative sensitivity towards the parameter evaluates to -0.001.

In the intersection model for angle / turning movement crashes given by Equation 7-2, it can be noted that three continuous variables enter the model namely: maximum posted speed limit (V_2), sectional ADT (V_3), and friction coefficient (V_5). Plot (a) in Figure 8-3 changing ADT (y-axis) and posted speed limit (x-axis) shows the crash count with changing ADT (y-axis) and posted speed limit (x-axis) while plot (b) illustrates the crash frequency with changing ADT (y-axis) and friction coefficient (x-axis). The results are for morning peak hours. The plots for afternoon peak hours and the off-peak hours are similar.

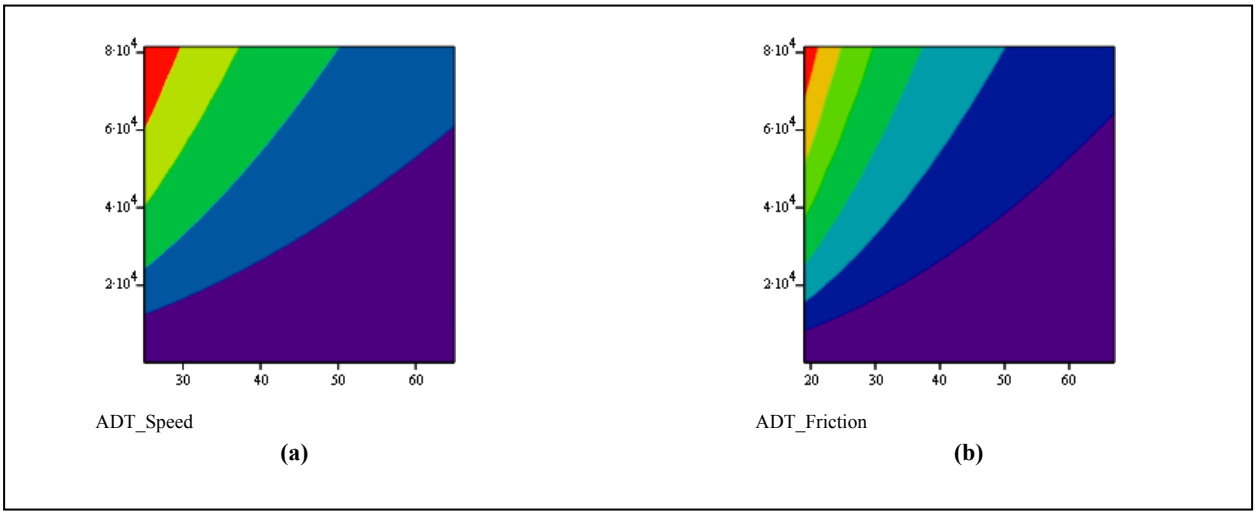


Figure 8-3 Crash Frequency contour plot with VIBGYOR increasing color patterns

The crash count is given by the contours which have been filed with VIBGYOR color patterns. This means that as we move from violet to red the crash frequency increases. For example in plot

(a) for any given posted speed limit, crash occurrences increase with increase of ADT. It can also be observed that the increase in crash frequency is greater at lower speed limit as the ADT goes up. This could be attributed to speed variation under the specific traffic conditions. For the intersection model ADT is the most important variable contributing to the crash count variance. The relative sensitivity is 0.29 as compared to 0.133 for maximum posted speed limit and 0.037 for friction coefficient. In case of the access model also, ADT has been found to be the most important continuous variable with a relative sensitivity value of 0.141 as compared to 0.048 for skid resistance.

8.1.3 Head-on Crashes

The segment model for head-on crashes is given by Equation 7-4 where the average truck factor (V_4) enters the model. The relative sensitivity for the average truck factor was also higher than the friction coefficient.

The intersection model for this type of crash, given by Equation 7-5 reveals interesting crash frequency patterns indicating differences between morning/ afternoon peaks and Friday/ Saturday night peaks. Figure 8-4 illustrates the crash count contour as surface width changes with shoulder width or skid resistance during the morning peak hours and the Friday/ Saturday night peak conditions. The afternoon peak patterns are similar to morning peak. Plots in Figure 8-5 reflect the migration of crash count prone conditions of average truck factor and friction coefficient. In Figure 8-4 the surface width is always on the x-axis.

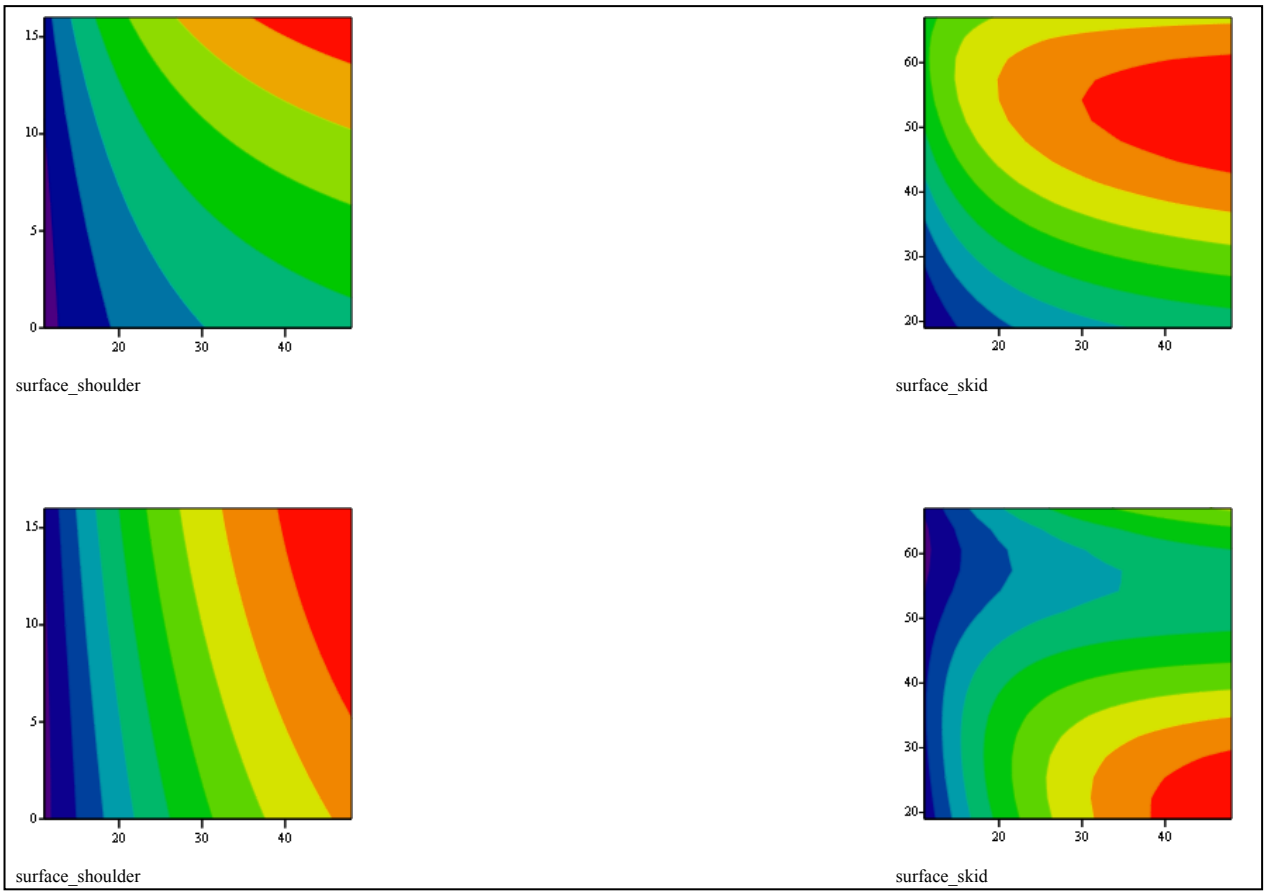


Figure 8-4 Crash Count patterns for Morning peak (top) and Friday/Saturday night peak (bottom)

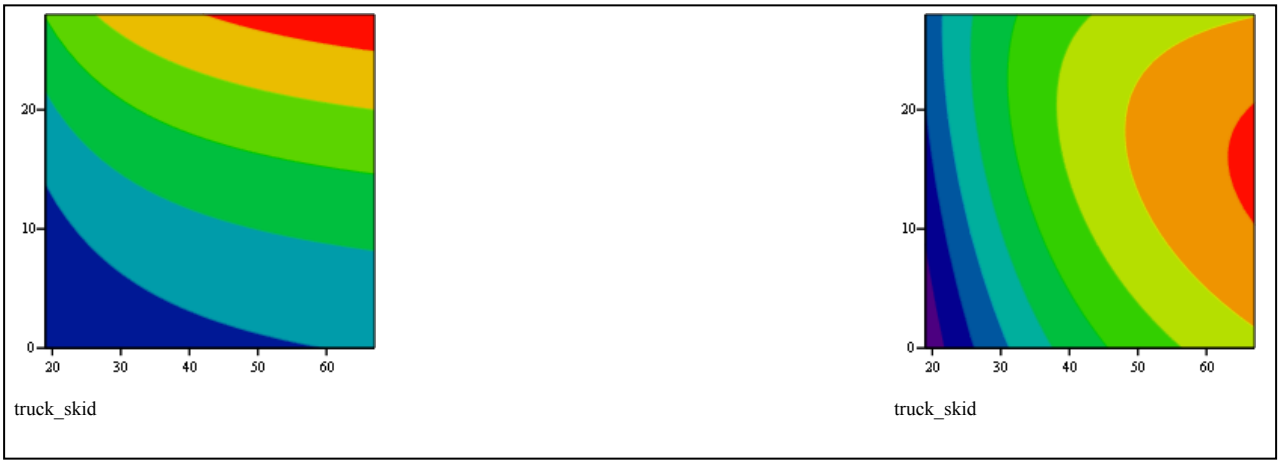


Figure 8-5 Crash Count patterns for Morning peak (left) and Friday/Saturday night peak (right)

In Figure 8-5 the average truck factor is on the y-axis while the skid resistance is on the x-axis. In Figure 8-4 and Figure 8-5 it is critical to observe how the ‘red; area changes. The shift is indicative of crash risk migration in time for different peak hours. Figure 8-6 and Figure 8-7 show surface plots indicating the crash frequency variation with change in ADT and shoulder width or maximum posted speed limit respectively. For both the figures the plot on the left is for morning peak while the plot on the right is for Friday/Saturday night peak conditions. In Figure 8-6 and Figure 8-7 the hidden axis is for the ADT. The vertical axis going away from the reader is the crash frequency.

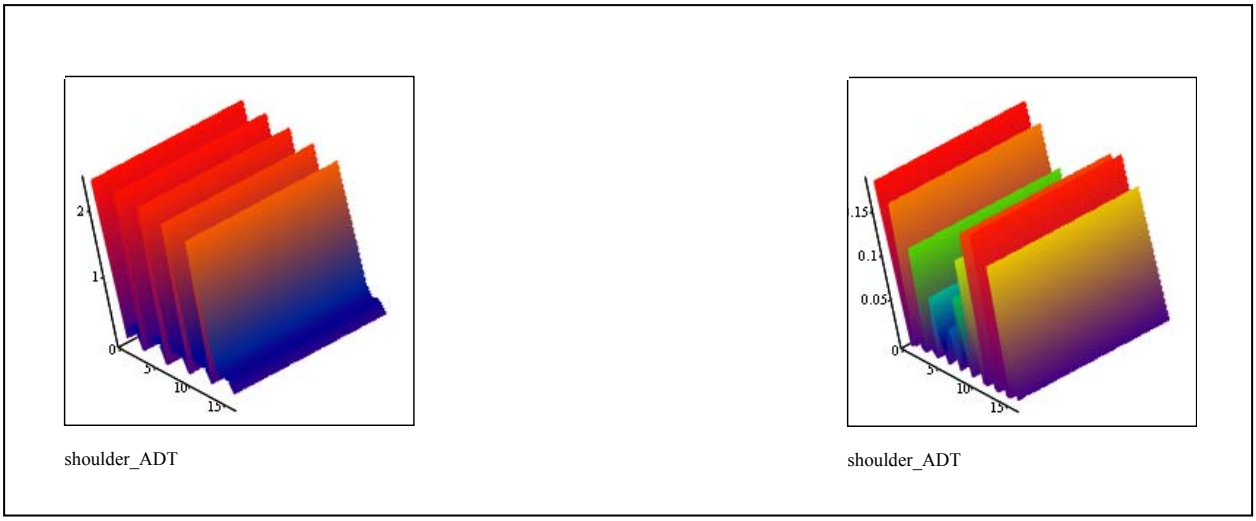


Figure 8-6 Crash Frequency variation with ADT and Shoulder width

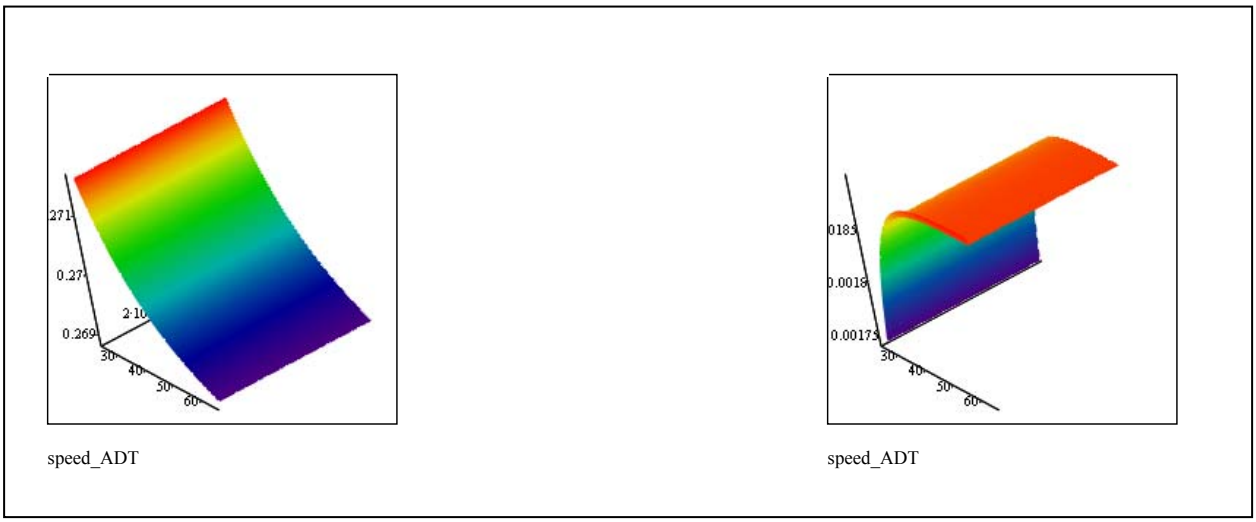


Figure 8-7 Crash Frequency variation with ADT and Maximum Posted Speed limit

The interpretation of the colors of the surface plots is the same as that of the contour plots.

8.1.4 Rear-end Crashes

The average sectional ADT (V_3) is common among all the three crash frequency models given by Equation 7-7 through Equation 7-9. Figure 8-8 and Figure 8-9 show the plots demonstrating how the frequency of crashes changes during the weekday morning peak hour and the weekday afternoon peak hour respectively with varying sectional ADT. In Figure 8-8 and Figure 8-9 the y-axis represents crash count and the x-axis represents the sectional ADT (V_3).

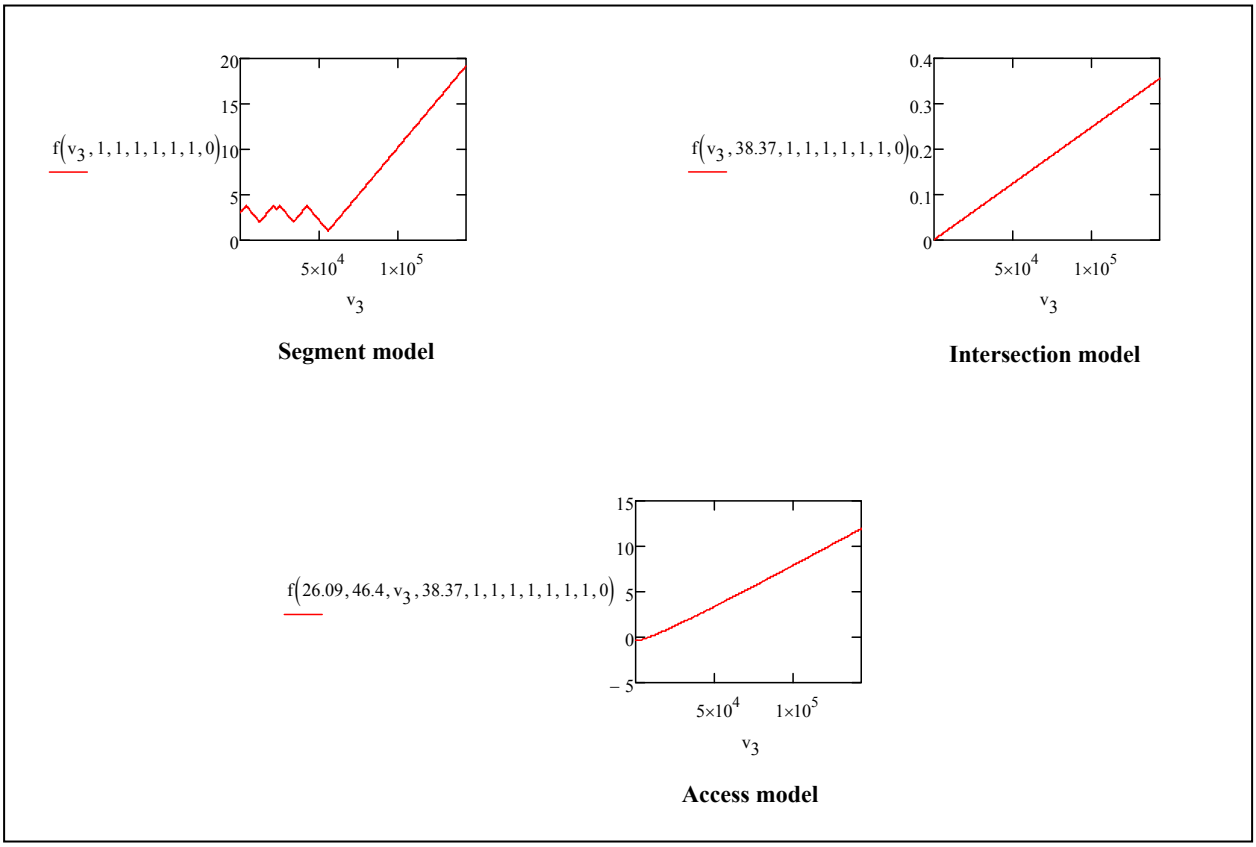


Figure 8-8 Crash Frequency versus ADT during morning peak hours

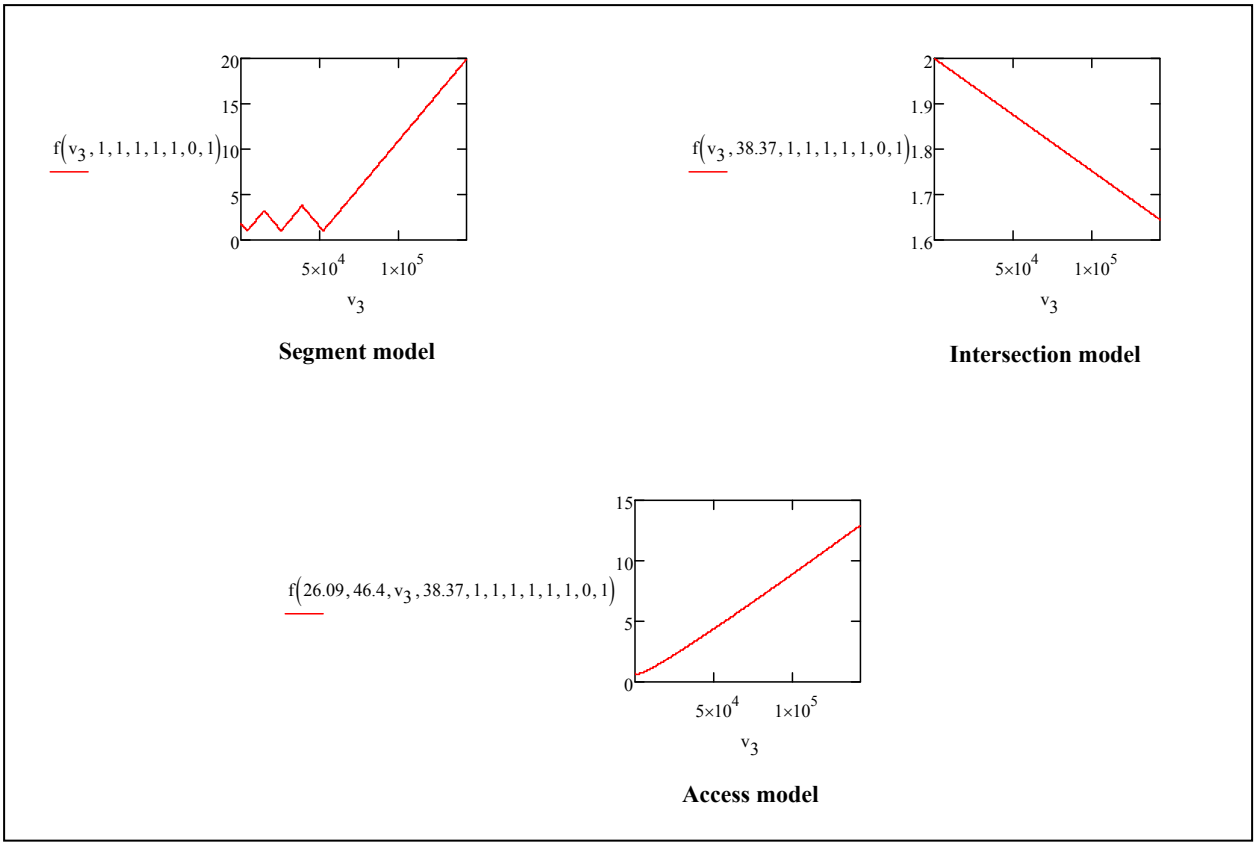


Figure 8-9 Crash Frequency versus ADT during afternoon peak hours

It is observed that at signalized intersections, the increase of ADT boosts the crash count during the morning peak hour and slows down the occurrence of crashes during the afternoon peak hour. The urgency to reach the work place during the morning peak hours could compel the drivers to be more aggressive on the roadways and hence, at the intersections, where cross traffic flow is also significantly high, there is an observed increase of crash counts. A more fascinating graph is observed for the segment models. During the morning peak hours there is an observed monotonic increase in crash frequency at ADT $\geq$ 55,299 and similar trend during the afternoon

peak hours at ADT ≥ 51,893. It can be observed that for lower ADT values, the crash count trend is fluctuating (see Figure 8-8 and Figure 8-9 – segment model plots).

For average conditions the frequency of crashes decreases with increase in the maximum posted speed limit for access related rear-end crashes. It was also interesting to observe that crash counts reach a maximum when the surface width is approximately 30 ft . Lower or higher values of surface width results in a decrease of crashes for access related rear-end crashes (Figure 8-10). As the skid resistance increases the crashes are found to be decreasing during the morning peak hours at signalized intersections while a reverse trend is observed during the afternoon peak hours during the weekdays (Figure 8-11). The y-axis represents crash frequency and the x-axis represents surface width (V_0) in Figure 8-10 and skid resistance (V_5) in Figure 8-11.

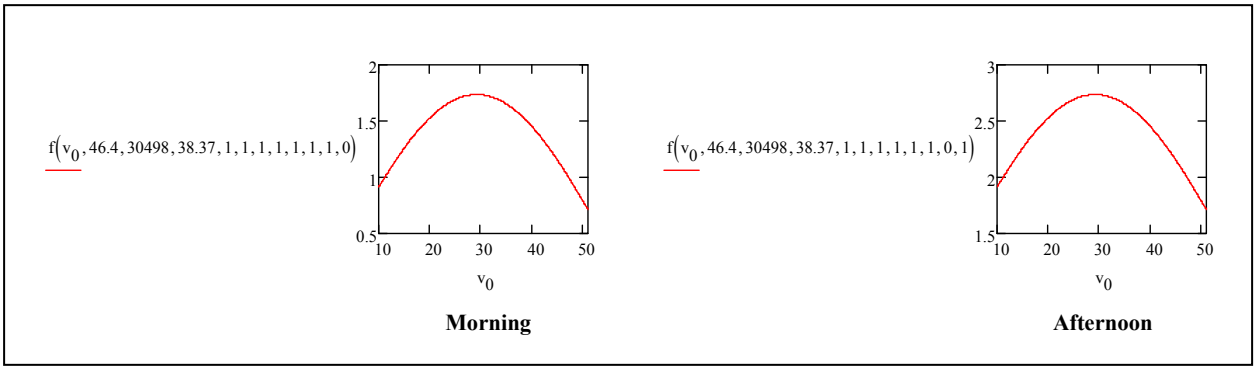


Figure 8-10 Crash Frequency versus Surface width

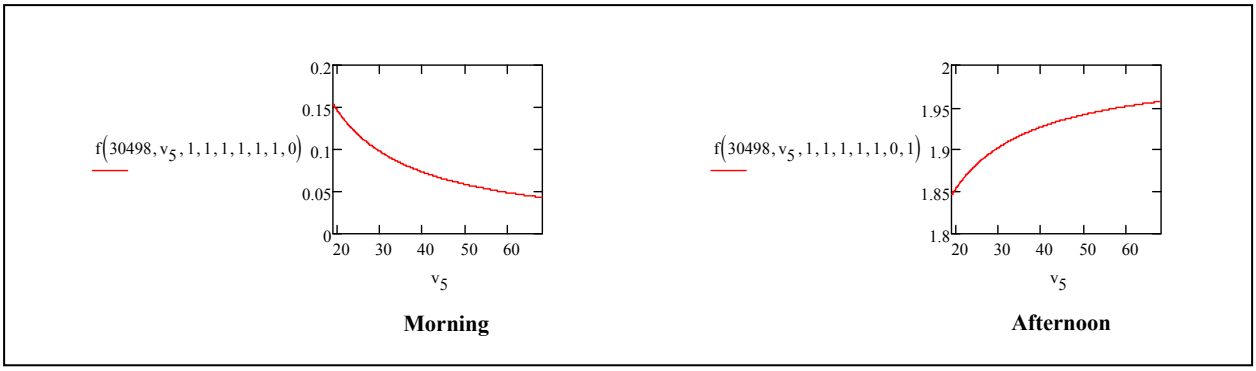


Figure 8-11 Crash Frequency versus Friction coefficient

Figure 8-12 shows the contour plots for varying crash occurrences with change in ADT and surface width or maximum posted speed limit during morning peak hours. Both the plots suggest that the crash count increase with increase of ADT. However, the crash occurrences are more with lower speeds. This is consistent with the results shown in Figure 8-7 and Figure 8-3 where the instances of head-on as well as angle/ turning movement crashes increase with lower speed limits during the morning peak hours. In Figure 8-12 the sectional ADT is the y-axis for both the plots.

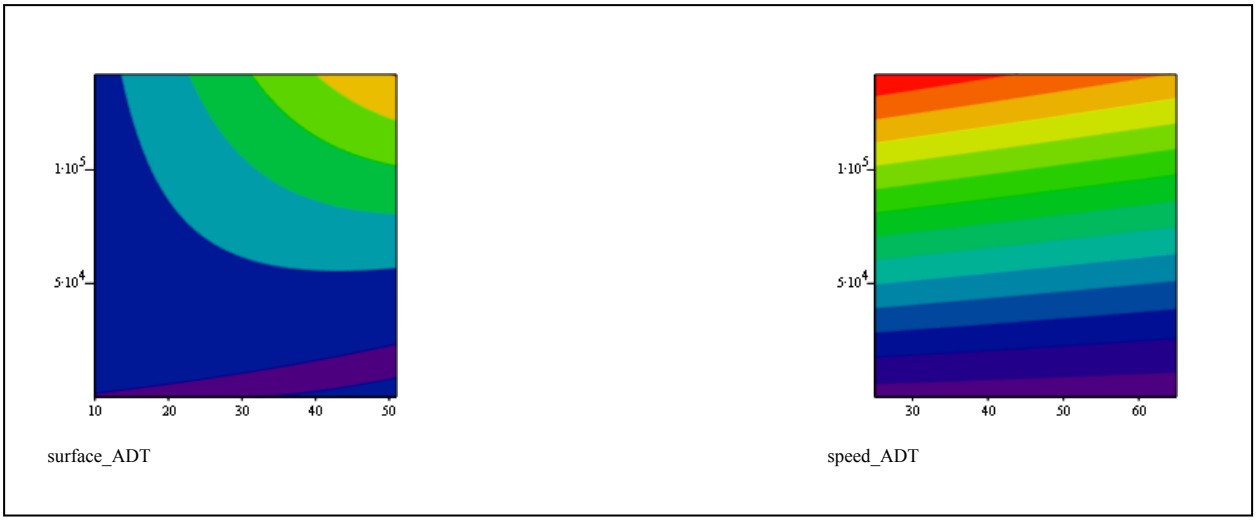


Figure 8-12 Crash Occurrence variation with ADT and Surface width (left) or Speed limit (right)

The relative sensitivity of the crash frequency towards ADT is -0.246 during the morning peak hours and is 0.246 during the afternoon peak hours, for the mid-block segment crashes. The difference in sign reflects the difference in the directionality of the variation in crash count. However, this value cannot be used further for comparative purpose due to the absence of other continuous variables in the model.

In the intersection related rear-end crash frequency model the relative sensitivity of the output towards ADT is higher (0.002) than that to the skid resistance (0.0009). This implies that ADT bring greater variation to the crash frequency than the friction coefficient of the pavement. In the access related rear-end crash count model four continuous variables enter the model. Among them, ADT was found to have the maximum effect on the variation of crashes (0.98). Maximum posted speed limit (0.009), surface width (0.007) and skid resistance (0.001) also effect the variation in crash count.

CHAPTER 9. CONCLUSIONS

9.1 Summary

Understanding the contributing factors for crash occurrence and injury severity resulting from a crash is essential for transportation safety analyses. Lack of proper understanding of the significant parameters can lead to inconclusive results thus rendering any research work impractical for implementation. Many researchers may argue that most of the contributing factors are already known; however, the dynamic nature of the transportation system has always posed a challenge to fully understand the behavior of the contributory factors. The existing body of knowledge consists of empirical, statistical, numerical and machine learning studies to augment our understanding of the safety situation. As stated in CHAPTER 1, the objective of the study is not only to enhance our perception of the safety of roadways but also to incorporate innovative applied methodologies in safety analysis. A key aspect of this research has been the application of machine learning algorithms in developing models for crash occurrence as well as injury severity classification. Missing data elements, under-reporting of crashes, faulty data entry, and non-uniform practice of law enforcement plague the transportation database. Hence, a thorough understanding of the transportation system is required, specifically from a safety point of view.

Understanding of the difference in the crash pattern of segments and intersections is elemental to the corridor safety approach, especially as it relates to injury severity. If one observes crashes only at the physical area of intersections; crashes would involve higher proportion of angle

and/or left turn crashes which tend to be more severe. However, as the definition of the intersection is changed to include some area around it (i.e., the influence area for an intersection is defined); rear-end and other groups of crashes would be included in the sample and the severity patterns may be altered. Research presented in CHAPTER 3 concludes that the set of significant factors change as the influence distance changes. Hence, the use of a fixed influence distance is ruled out. Another critical finding was the inter-dependency of the crash location with injury severity. In other words if injury severity patterns had to be studied it will be crucial that the corridor approach be used for the research objectives at hand.

It is critical to distinguish among segment-related, signalized intersection-related and un-signalized intersection-related crashes. State of Florida has a typical intersection influence radius of 250 ft. irrespective of the physical size, ADT of the intersecting roadways, number of lanes and demographics. In addition to this impractical influence radius, Florida has a 50 ft default intersection size. Since not all intersections are of the same size, no matter how good the officer is at guessing the location indicated in the crash report, it is a rough approximation. If the "site location" is used to determine the location of a crash, the only access related crashes that could be identified are those with site location value of 'driveway access'. This highlights the futility in using 'site location' to assign crashes. A closer study of crash reports revealed that traffic control in combination with the site location did a superior job in identifying the roadway element to be assigned to correctly. Hence the method of assigning a crash based on crash characteristics. CHAPTER 4 lays down the comprehensive rules used to assign crashes to appropriate roadway elements based on the 'site location', 'traffic control' and node information.

Though the researcher focused on a corridor approach for the study, the FDOT does not have an exact definition of a corridor. Hence it was critical that the analysis begins with a definition of a corridor. The roadway design of arterials which is essentially of three types: 1) Urban; 2) Sub-Urban and 3) Rural; was chosen as the defining parameter for creating homogenous sections. Corridors of similar lengths have their heterogeneity minimized. Hence, clustering was performed based on Partitioning around Medoids algorithm. Corridors were clustered in four groups and the range of corridors in each cluster is given in Table 5-1.

A data mining approach was adopted for injury severity analysis, especially tree algorithms which can generate classification rules. Classification is more appropriate for injury severity analysis as the response variable is binary or ordinal (for more than two categories of severity). The reason for choosing a data mining approach was its versatility to build a trained model and then validating it (supervised learning). However, recent research in theoretical statistics indicate that CART algorithm is biased towards selecting variables that are continuous in nature or have large number of categories. Hence, in the present work the researcher used the conditional inference tree as the classification algorithm (see CHAPTER 6). The input parameter's association with the target variable decides the importance of the parameter where as the split can be done by the regular split methods applied in CART. Failure to use seat belts, higher posted speed limits, alcohol / drug use, slow moving vehicles, higher skid resistance are some of the factors that contributed to increased severity of injuries sustained during crashes.

Essential difference between the crash occurrence phenomenon and the injury severity levels is the response type. Crash occurrence is a continuous integer response while the severity is an ordinal target. Since the crash occurrence and injury severity are fundamentally different phenomena it is not practical to have one model governing them. However, in CHAPTER 7 the researcher suggests independent approaches for building both the injury severity and crash frequency models under the broader umbrella of the heuristic GP, using the concepts of evolutionary biology like crossover and mutation. The process of model evolution takes places, through generations, with decreasing mean error as the objective function for crash count modeling and increasing hit rate as the objective function for injury severity classification. On-street parking at higher speed corridors increases the likelihood of injuries resulting from crashes. Higher shoulder width reduces injury severity where as restrictive median openings, lower ADT (indication of higher vehicle speeds), sharper curves and high truck percentage increase injury severity on the highways. The plots shown in CHAPTER 8 enhance our perception of the trends of crash counts, especially when the models are non-linear.

9.2 Recommendations

The pertinent question that a reader may ask is why is there a need of advanced machine learning algorithms, sophisticated data mining techniques, and statistical models to investigate the contributing factors. The need arises due to the fact that our roadway systems are dynamic. The environment around them changes, the driver behavior ranges from cautious to aggressive. The complexity of the equations which govern the crash occurrence phenomena and the classification

rules for injury severity is evidence enough to prove intuition wrong. Crash statistics mentioned in CHAPTER 1 reveal the grim safety situation. Even though the crash rates, fatality rates may have gone down significantly over the years the number of fatalities and incapacitating crashes are high. The only way to reduce the number of fatalities is through counter measures. Effective countermeasures can only be developed if our understanding of the causative factors is adequate. The 3 E's (engineering, education and enforcement) are the three basic strategies that have to be implemented effectively to observe any significant change in the safety situation of highways.

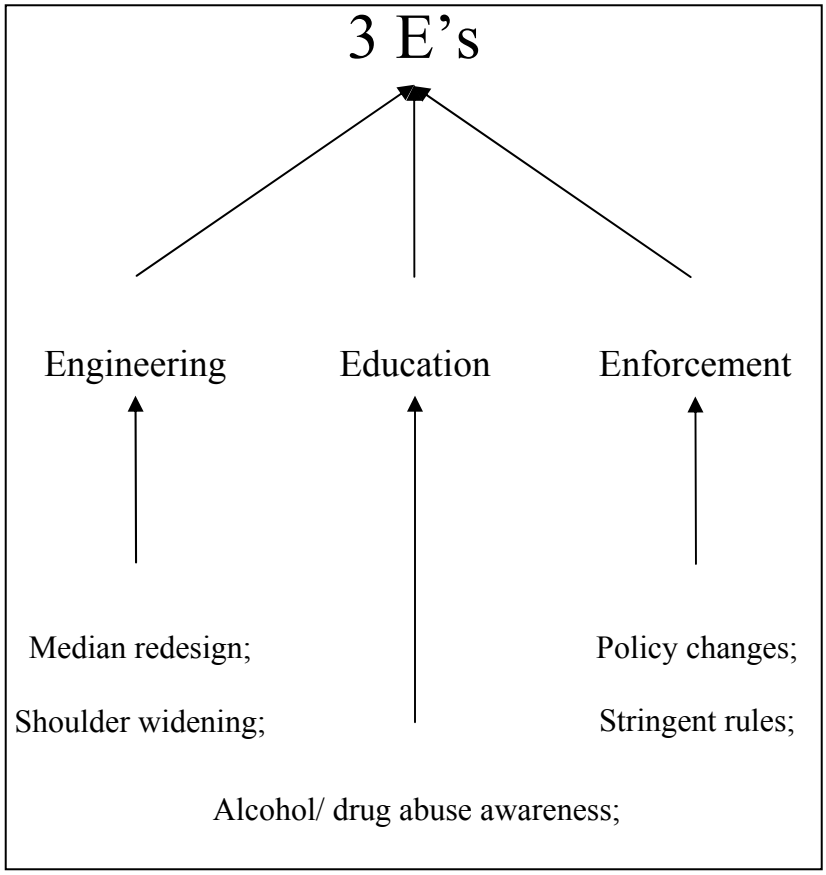


Figure 9-1 Bottom – Up approach for 3 E's implementation

The path from problem identification to solution implementation goes through understanding of contributing factors phase. One of the primary objectives of this extensive study was to identify significant factors and their trends in understanding injury severities. This is essential to recommend changes in design and policy for better planning of the roadway network. The following recommendations are based on some of the important findings of the present work. The evaluation of the recommendations could be the scope of future research work.

1. Crashes related to alcohol/ drug use by drivers are found to have higher injury severity when young children (up to 3 years old) and/or older people (more than 55 years) are present in the vehicle (see section 6.4.2.3). The physical vulnerability of people belonging to these age groups puts them at a higher risk of sustaining severe injuries in case of a crash. This is similar to reckless endangerment of another person. A law in California prohibits adults from smoking when minors are inside the vehicle. DUI is as much a public health issue as it is an unsafe driving problem. Hence, a harsher punishment could be imposed on drivers charged with DUI with passengers, especially those belonging to the above mentioned age groups, on board.

2. Presence of restrictive medians (gap between openings is 0.5 mile for posted speed limits greater than 45 mph and 0.25 mile otherwise) has been observed to be associated with higher injury severity in angle/ turning movement and sideswipe crashes (see sections 6.4.2.2, 6.4.2.6 and 7.3.1.3). To alleviate the situation operation strategies like increased time for left turn phase on nearby signalized intersections. A case by case analysis may

be carried out on these problematic roadway sections to observe the effects without adversely affecting the level of service for the intersections. Future roadway sections could use alternate design strategies in terms of using the more frequent directional medians. The directional medians have been observed to be better both in terms of safety and operations (Zhou et al., 2001).

3. Less than 45 mph posted speed limit together with a higher K-factor has been observed have higher proportion of severe angle/ turning movement and rear-end crashes on the urban arterials. (refer to sections 6.4.2.2 and 6.4.2.3) Design considerations may not permit the increase of speed limits. Authorities could make use of Advanced Traffic Information Systems (AITS) to educate drivers in planning the trips so as to avoid peak hour congestion. Proper dissemination of traffic information to the drivers could spread the congestion more uniformly through more number of hours leading to lower k-factor.

4. It was observed that higher friction coefficient is associated with severe injuries rear-end crashes (see section 6.4.2.3) and have also been associated with an increase in frequency of crashes (refer to sections 7.3.1.2, 7.3.1.3 and 7.3.1.4). Higher skid resistance means shorter braking distance. The drivers may over compensate for it by driving faster. Coming to a sudden stop due to better pavement friction and advanced braking system on the vehicles more often leads to internal movements inside the vehicle thus causing secondary severe injuries. Moreover, injuries sustained during crashes have been found to be more severe due to non-use of safety belts or other safety equipments. Mitigation of

impact could be a solution which may be from proper use of seatbelts and other equipment may lead to even safer braking. Even though authorities are presently doing their best to spread awareness on seatbelt use law, a more directed campaign to educate the drivers is required.

5. Slow and lighter vehicles like cycle, mopeds etc. always are riskier to drive on urban arterials as they compete for space on the roads without providing much physical protection to the driver (see sections 6.4.2.3 and 6.4.2.6). Wider cycling lanes could be tried as a potential countermeasure with proper before-after study to assess the safety improvement. Urban traffic control systems designed to recognize cyclists and give them priority. Diversion of traffic from roadways frequently used by higher number of cyclists could be implemented. Road signs like advanced stop signs for cyclists and the use of shared space concepts in future urban design could be tested. Strategies to encourage safe cycling on the roadways could be an impetus for green transportation.

6. The results reflected that with the increase of "shoulder + sidewalk" width, the severity of injury sustained decreased (refer to sections 7.3.1.2 and 7.3.1.3). This is particularly useful in the context of urban arterials as a lot of sections have very little inner and/or outer shoulder width. The presence of side walk gives the adequate cushion for recovery in case of a lane-departure crash. Wider sidewalks can not only provide the cushion but also allow more space for the urban pedestrian to avoid potential vehicle-pedestrian crash in which severe injuries are often sustained. Hence provision for adequate sidewalks in

the future roadway design should be incorporated and the right-of-way acquisition could be planned accordingly.

7. Higher truck factor on the roadways have been observed to be associated with increased severity of head-on injuries (see section 7.3.1.3). Strategies for restrictions on lane use could be tested specifically on the corridors where truck related crashes are higher.

LIST OF REFERENCES

Abdel-Aty, M. (2003). "Analysis of driver injury severity levels at multiple locations using probit models", Journal of Safety Research, Vol. 34, No. 5, pp. 597-603.

Abdel-Aty, M. and Abdelwahab, H. T. (2004). "Predicting injury severity levels in traffic crashes: a modeling comparison", Journal of Transportation Engineering, Vol. 130, No. 2, pp. 204-210.

Abdel-Aty, M. and A. H. As-Saidi, Using GIS to locate the high risk driver population. Swedish National Road and Transport Research Institute, 2000, pp. 111-126.

Abdel-Aty, M. and Keller, J. (2005). "Exploring the overall and specific crash severity levels at signalized intersections", Accident Analysis & Prevention, Volume 37, Issue 3, pp. 417-425.

Abdel-Aty, M. and Pande, A. (2006). "Comprehensive analysis of relationship between real-time traffic surveillance data and rear-end crashes on freeways", Transportation Research Record 1953, pp. 31-40.

Abdel-Aty, M., Pande, A., Das, A. and Knibbe, W. J. (2008). "Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems", Transportation Research Record 2083, pp. 153-161.

Abdel-Aty, M., Lee, C., Wang, X., Nawathe, P., Keller, J., Kowdla, S. and Prasad, H. (2006). "Identification of intersections' crash profiles/patterns", FDOT, Tallahassee, 2006, http://www.dot.state.fl.us/researchcenter/Completed_Proj/Summary_SF/FDOT_BC355_10_rpt.pdf Accessed June 10, 2007.

Abdel-Aty, M., Pemmanaboina, R. and Hsia, L. (2006). "Assessing crash occurrence on urban freeways by applying a system of interrelated equations", Transportation Research Record 1953, pp. 1-9.

Abdel-Aty, M. and Radwan, A. E. (2000). "Modeling traffic accident occurrence and involvement", Accident Analysis & Prevention, Vol. 32, No. 5, pp. 633-642.

Abdel-Aty, M. and Wang, X. (2006). "Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors", Transportation Research Record 1953, pp. 98-111.

Abdelwahab, H. and Abdel-Aty, M. (2004). "Investigating the effect of light truck vehicle percentages on head-on fatal traffic crashes", Journal of Transportation Engineering, Vol. 130, No. 4, pp. 429-437.

Anderson, M. L. (2008). "Safety for Whom? The Effects of Light Trucks on Traffic Fatalities", Journal of Health Economics, Vol. 27, No. 4, pp. 973-989.

Andreassen, D. (2003). "Aspects of Road Design and Trucks from the Analysis of Crashes", Institution of Professional Engineers New Zealand (IPENZ) Transportation Group Technical Conference Papers 2003.

Batra, S. and Kumar, S. (2008). "Airbag-Induced Fatal Subaxial Cervical Spinal Cord Injury in a Low-Velocity Collision", European Journal of Emergency Medicine, Vol. 15, No. 2, pp. 52-55.

Bedard, M., Guyatt, G. H., Stones, M. J. and Hirdes, J. P. (2002). "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities", Accident Analysis & Prevention, Vol. 34, No. 6, pp. 717-727.

Bichler-Robertson, G., G. Laycock, R. Clarke, R. Sampson, G. Cordner, G. Saville, R. Glensor, M. Scott and N. L. Vigne (2001). "Excellence in Problem-Oriented Policing: The 2001 Herman Goldstein Award Winners", US Department of Justice.

Bjornstig, U., Bjornstig, J. and Eriksson, A. (2008). "Passenger car collision fatalities - with special emphasis on collisions with heavy vehicles", Accident Analysis & Prevention, Vol. 40, No. 1, pp. 158-166.

Bonneson, J. A. and McCoy, P. T. (1997). "Effect of median treatment on urban arterial safety: an accident prediction model", Transportation Research Record, 1581, pp. 27-36.

Bowman, B. L., Vecellio, R. L., and Miao, J. (1995). "Vehicle and pedestrian accident models for median locations", Journal of Transportation Engineering, ASCE, Vol. 121, No. 6, pp. 531-537.

Brameier, M. and Banzhaf, W. (2007). "Linear genetic programming", Springer, New York.

Breiman, L. (2001). "Random forests", Machine Learning, Vol. 45, No. 1, pp. 5 – 32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). "Classification and regression trees", Wadsworth International Group, Belmont, California.

Breyer, J. P. and S. C. Joshua (1999). "Identifying and Implementing Corridor Safety Improvements: A Highway Safety Improvement Process and Safety Analysis Tools for Arizona", Publication FHWA-AZ 99-458. Arizona Department of Transportation and FHWA, U.S. Department of Transportation.

Brijs, T., Karlis, D., Van den Bossche, F. and Wets, G. (2003). "A Bayesian model for ranking hazardous sites", Flemish Research Center for Traffic Safety, Diepenbeek, Belgium.

Brown, H. C. and Tarko, A. P. (1999). "Effects of access control on safety on urban arterial streets", Transportation Research Record, 1665, pp. 68-74.

Carlin, B. P. and Louis, T. A. (2000). "Bayes and Empirical Bayes methods for data analysis", 2nd Edition, Chapman & Hall, Boca Raton.

Categorical Dependent Variable Models Using SAS, STATA, LIMDEP, and SPSS, Research Technologies at Indiana University. http://www.indiana.edu/~statmath/stat/all/cdvm/cdvm1.html#s12 Accessed June 4, 2007.

Ceylan, H. and Bell, M. G. H. (2004). "Traffic signal timing optimisation based on genetic algorithm approach, including drivers' routing", Transportation Research Part B: Methodological, Vol. 38, No. 4, pp. 329-342.

Chang, L. Y. and Mannering, F. (1999). "Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents", Accident Analysis & Prevention, Vol. 31, No. 5, pp. 579-592.

Chang, N. B. and Chen, W. C. (2000). "Prediction of PCDDs/PCDFs emissions from municipal incinerators by genetic programming and neural network modeling", Waste Management Research, Vol. 18, pp. 341-351.

Chang, L-Y and Wang, H-W. (2006). "Analysis of Traffic Injury Severity: An Application of Non-Parametric Classification Tree Techniques", Accident Analysis & Prevention, Vol. 38, No. 5, pp. 1019-1027.

Cheng, W. and Washington, S. P. (2005). "Experimental evaluation of hotspot identification methods", Accident Analysis & Prevention, Vol. 37, No. 5, pp. 870-881.

Dahir, S. and Wade L.G. (1990). "Wet-Pavement Safety Programs. NCHRP Synthesis of Highway Practice 158", Transportation Research Board, National Research Council, Washington, D.C.

Das, A., Abdel-Aty, M. and Pande, A. (2009). "Using conditional inference forests to identify the factors affecting crash severity on arterial corridors", Journal of Safety Research, Vol. 40, No. 4, pp. 317-327.

Das, A., Pande, A., Abdel-Aty, M. and Santos J. B. (2008). "Urban arterial crash characteristics related with proximity to intersections and injury severity", Transportation Research Record 2083, pp. 137-144.

Davis, G. A. and Yang, S. (2001). "Bayesian identification of high-risk intersections for older drivers via Gibbs sampling", Transportation Research Record 1746, pp. 84-89.

De Jong, K. A. (2006). "Evolutionary Computation: a unified approach", MIT Press, Cambridge, Massachusetts.

Delen, D., Sharda, R. and Bessonov, M. (2006). "Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks", Accident Analysis & Prevention, Vol. 38, No. 3, pp. 434-444.

Derrig, R. A., Segui-Gomez, M., Abtahi, A., and Liu, L. L. (2000). "The effect of population safety belt usage rates on motor vehicle-related fatalities", Accident Analysis & Prevention, Vol. 34, No. 1, pp. 101-110.

Deschaine, L. M. and Francone, F. D. (2004). "White paper: comparison of Discipulus™ Linear Genetic Programming software with Support Vector Machines, Classification Trees,

Neural Networks and Human Experts", http://www.rmltech.com/Comparison.White.Paper.pdf Accessed February 7, 2008.

Drummond, K. P., Hoel, L. A. and Miller, J. S. (2002). "A simulation-based approach to evaluate safety impacts of increased traffic signal density", retrieved January 30, 2007, from http://www.virginiadot.org/vtrc/main/online_reports/pdf/02-r17.pdf

Drummond, K. P., Hoel, L. A. and Miller, J. S. (2002). "Using simulation to predict safety and operational impacts of increasing traffic signal density", Transportation Research Record, 1784, pp. 100-107.

Dummeldinger, M., P. Henderson, B. G. Ward and V. Zambito (1994). "Corridor Safety Improvement Program: Impact Evaluation", University of South Florida, Center for Urban Transportation Research, Tampa, http://www.lib.usf.edu/cgi-bin/Ebind2h3.pl/cutr0253 . Accessed March 15, 2007.

Duncan, C., Khattak, A., & Council, F. (1999). "Applying the Ordered Probit Model to Injury Severity in Truck-Passenger Car Rear-End Collisions", Transportation Research Record 1635, pp. 63-71.

Elvik, R. (1997). "Evaluations of road accident black spot treatment: a case of the iron law of evaluation studies?", Accident Analysis & Prevention, Vol. 29, No. 2, pp.191-199.

Eluru, N. and Bhat, C. R. (2007). "A joint econometric analysis of seat belt use and crash – related injury severity", Accident Analysis & Prevention, Vol. 39, No. 5, pp. 1037-1049.

Evans, L. (1996). "Safety-belt effectiveness: the influence of crash severity and selective recruitment", Accident Analysis & Prevention, Vol. 28, No. 4, pp. 423-433.

Finan, J. D., Nightimgale, R. W. and Myers, B. S. (2008). "The Influence of Reduced Friction on Head Injury Metrics in Helmeted Head Impacts", Traffic Injury Prevention, Vol. 9, No. 5, pp. 483-488.

Flahaut, B., Mouchart, M., Martin, E. S. and Thomas, I. (2003). "The local spatial autocorrelation and the kernel method for identifying black zones: a comparative approach", Accident Analysis & Prevention, Vol. 35, No. 6, pp. 991-1004.

"Florida Corridor/Community Traffic Safety Program", NHTSA, summer 1996, http://www.nhtsa.dot.gov/people/outreach/safedige/Summer96/FHWA/Florida.html Accessed February 18, 2007.

Fontaine, M. D. and S. W. Read (2006). "Development and Evaluation of Virginia's Highway Safety Corridor Program", Publication FHWA/VTRC 06-R30. Virginia Department of Transportation and FHWA, U.S. Department of Transportation.

Francone, F. D. (1998). "Discipulus™ Software Owner's Manual", http://www.rmltech.com/Discipulus%20Owners%20Manual.pdf – Accessed March 23, 2008.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). "Bayesian Data Analysis", 2nd edition, Chapman & Hall, Boca Raton.

Gettis, J. L., Balakumar, R. and Duncan, L. K. (2005). "Effects of Rural Highway Median Treatments and Access", Transportation Research Record 1931, pp. 99-107.

Geurts, K., Wets, G., Brijs, T. and Vanhoof, K. (2004). "Identifications and ranking of black spots: sensitivity analysis", Transportation Research Record 1897, pp. 34-42.

Geurts, K. and Wets, G. (2003). "Black Spots Analysis Methods: A literature review", Flemish Research Center for Traffic Safety, Diepenbeek, Belgium.

Goldberg, D. E. (1989). "Genetic algorithms in search, optimization and machine learning", Addison-Wesley Publication, Reading, Massachusetts.

"Governor's Task Force on Highway Safety", State of Ohio Government Info and Services, August 4, 2005, http://corridorsafety.ohio.gov/ . Accessed March 2, 2007.

Greene, W. H. (2003). Econometric Analysis (5th Ed.), Pearson Education, USA.

Green, E. R. and K. R. Agent (2002). "Evaluation of High Traffic Crash Corridors", Publication KTC-02-8/SPR231-01-1F. Kentucky Department of Transportation.

Greibe, P. (2003). "Accident prediction models for urban roads", Accident Analysis & Prevention, Vol. 35, No. 2, pp. 273-285.

Hanley, K. E., Gibby, A. R. and Ferrara, T. C. (2000). "Analysis of accident-reduction factors on California State Highways", Transportation Research Record, 1717, pp. 37-45.

Harb, R., Yan, X., Radwan, E. and Su, X. (2009). "Exploring precrash maneuvers using classification trees and random forests", Accident Analysis & Prevention, Vol. 41, No. 1, pp. 98-107.

Harwood, D., Council, F., Hauer, E., Hughes, W. and Vogt, A. (2000). "Prediction of the expected safety performance of rural two-lane highways", Federal Highway Administration, Washington, D.C.

Hauer, E. (1986). "On the estimation of the expected number of accidents", Accident Analysis & Prevention, Vol. 18, No. 1, pp. 1-12.

Hauer, E., Harwood, D. W., Council, F. M. and Griffith, M. S. (2002). "Estimating safety by the empirical Bayes method: a tutorial", Transportation Research Record 1784, pp. 126-131.

"Highway 25 Safety Corridor". California Department of Transportation, November 13, 2003, http://www.dot.ca.gov/dist05/projects/hwy25/corridor.htm . Accessed March 6, 2007.

Highway Data Collection / Quality Control Section of Transportation Statistics Office, FDOT (2007). "The RCI Office handbook", http://www.dot.state.fl.us/planning/statistics/rci/officehandbook/fullrpt.pdf Accessed April 14, 2007.

Hiselius, L. W. (2004). "Estimating the relationship between accident frequency and homogeneous and inhomogeneous traffic flows", Accident Analysis & Prevention, Vol. 36, No. 6, pp. 985-992.

Holland, J. M. (1975). "Adaptation in Natural and Artificial Systems", University of Michigan Press, Ann Arbor.

Hothorn, T. Hornik, K., and Zeileis, A. (2006). "Unbiased recursive partitioning: a conditional inference framework", Journal of Computational and Graphical Statistics, Vol. 15, No. 3. pp. 651-674.

Hothorn, T. Hornik, K., and Zeileis, A. (2008). "A laboratory for recursive partitioning", http://cran.r-project.org/web/packages/party/party.pdf Accessed February 28, 2008.

Huang, H., Chor, C. H.and Haque, M. M. (2008). "Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis", Accident Analysis & Prevention, Vol. 40, No. 1, pp. 45-54.

Hughes, R. G. (1999). "Truck Safety in North Carolina: Effectiveness of NCDMV Efforts in FY99", University of North Carolina, University of North Carolina Highway Safety Research Center, Chapel Hill, http://www.hsrc.unc.edu/pdf/2000/cvspoo.pdf . Accessed March 9, 2007.

Hunter-Zaworski, K. M. and N. T. Price (1998). "Evaluation of the Corridor Safety Improvement Program: Phase 1 Final Report", Publication FHWA-OR-RD-98-20. Oregon Department of Transportation and FHWA, U.S. Department of Transportation.

Jernigan, J. D. (1999). "Comparative case studies of corridor safety improvement efforts", retrieved February 3, 2007, from http://www.virginiadot.org/vtrc/main/online_reports/pdf/00-r17.pdf

Jernigan, J. D. (1997). "Lessons learned from Virginia's pilot corridor safety improvement program", Publication VTRC 97-R11. Virginia Department of Transportation.

Jianming, M., Kockelman, K (2004). "Anticipating injury & death: controlling for new variables on Southern California highways", Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, DC.

Jones, B., A. Griffith and K. Haas (2002). "Effectiveness of Double Fines as a Speed Control Measure in Safety Corridors: Final Report", Publication FHWA-OR-DF-03-10. Oregon Department of Transportation and FHWA, U.S. Department of Transportation.

Jorgensen, R. E. (1966). "Evaluation of criteria for safety improvements on the highway", Westat Research Analysts Inc., Gaithersburg, Maryland, USA.

Kaufman, L. and Rousseeuw, P. J. (1990). "Finding groups in data: an introduction to cluster analysis", Wiley, New York.

Kim, K., Nitz, L., Richardson, J. and Li, L. (1995). "Personal and behavioral predictors of automobile crash and injury severity", Accident Analysis & Prevention, Vol. 27, No. 4, pp. 469-481.

Kockelman, K. M. and Kweon, Y. J. (2002). "Driver injury severity: an application ordered probit models", Accident Analysis & Prevention, Vol. 34, No. 3, pp. 313-321.

Koza, J. R. (1992). "Genetic programming: on the programming of computers by means of natural selection", MIT Press, Cambridge, Massachusetts.

Kweon, Y. J. and Kockelman, K. M. (2005). "Safety Effects of Speed Limit Changes: Use of Panel Models, Including Speed, Use, and Design Variables", Transportation Research Record 1908, pp. 148-158.

Levinson, H. S. (1999). "Access spacing and accidents – a conceptual analysis", retrieved February 3, 2007, from http://www.urbanstreet.org/2nd_sym_proceedings/Volume%201/Ec019_c1.pdf

Levinson, H. S. (2000). "Restrictive medians and two-way left turn lanes: some observations", Third National Access Management Conference, Federal Highway Administration, pp. 243-245.

Li, X., Lord, D. and Xie, Y. (2008). "Predicting motor vehicle crashes using support vehicle machine models", Accident Analysis & Prevention, Vol. 40, No. 4, pp. 1611-1618.

Liu, C. and Chen, C. L. (2009). "An Analysis of Speeding-Related Crashes: Definitions and the Effects of Road Environments", NHTSA Technical Report, National Center for Statistics and Analysis, NHTSA.

Long, S. J. (1997). “Regression Models for Categorical and Limited Dependent Variables”, Sage Publications, Thousand Oaks, CA.

Lord, D., Washington, S. and Ivan., J. (2005). “Poisson, Poisson-gamma and Zero Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory”, Accident Analysis & Prevention, Vol. 37, No. 1, pp. 35-46.

McGuigan, D. R. D. (1981). “The use of relationships between road accidents and traffic flow in black-spot identification”, Traffic Engineering and Control, Vol. 22, No. 8-9, pp. 448-453.

Ma, J., Kockelman, K. M. and Damien, P. (2008). “A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods”, Accident Analysis & Prevention, Vol. 40, No. 3, pp. 964-975.

Makkeasorn, A., Chang, N. B., Beaman, M., Wyatt, C., and Slater, C. (2006). “Soil moisture estimation in a semiarid watershed using RADARSAT-1 satellite imagery and genetic programming”, Water Resources Research, Vol. 42, pp. 1-15.

Malyshkina, N. and Mannering, F. L. (2008). “Effect of Increases in Speed Limits on Severities of Injuries in Accidents”, Journal of the Transportation Research Board, No. 2083, pp. 122-127.

Marshall, S. C. (2008). “The Role of Reduced Fitness to Drive Due to Medical Impairments in Explaining Crashes Involving Older Drivers”, Traffic Injury Prevention, Vol. 9, No. 4, pp. 291-298.

Martin, J. L. (2002). “Relationship between crash rate and hourly traffic flow on interurban motorways”, Accident Analysis & Prevention, Vol. 34, No. 5, pp. 619-629.

Melanie, M. (1996). "An introduction to genetic algorithms", MIT Press, Cambridge, Massachusetts http://www.netlibrary.com/urlapi.asp?action=summary&v=1&bookid=1337 . Accessed on February 5, 2008.

Miaou, S-P. (1996). "Measuring the goodness-of-it of accident prediction models", Federal Highway Administration, Washington, D.C.

Miaou, S.P. and Lord, D. (2003). "Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes", Transportation Research Record 1840, pp. 31-40.

Miaou, S. P. and Song, J. J. (2005). "Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence", Accident Analysis & Prevention, Vol. 37, No. 4, pp. 699-720.

Milton, J. and Mannering, F. (1998). "The relationship among highway geometrics, traffic related elements and motor-vehicle accident frequencies", Transportation, Vol. 25, No. 4, pp. 395-413.

Milton, J. C., Shankar, V. N. and Mannering, F. L. (2008). "Highway accident severities and the mixed logit model: An exploratory empirical analysis", Accident Analysis & Prevention, Vol. 40, No. 1, pp. 260-266.

Mountain, L., Fawaz, B. and Jarret, D. (1996). "Accident prediction models for roads with minor junctions", Accident Analysis & Prevention, Vol. 28, No. 6, pp. 695-707.

Mulinazzi, T. E. and Michael, H. L. (1967). "Correlation of design characteristics and operational controls and accident rates on urban arterials", Joint Highway Research Project, Purdue University, Lafayette, Indiana.

National Highway Traffic Safety Administration (2009). "Traffic safety facts 2008: A compilation of motor vehicle crash data from the Fatality Analysis Reporting System and the General Estimate System", Washington, D.C.

National Highway Traffic Safety Administration (2007). "Traffic safety facts 2006: A compilation of motor vehicle crash data from the Fatality Analysis Reporting System and the General Estimate System, 2006", Washington, D.C.

National Highway Traffic Safety Administration (2006). "Traffic safety facts 2005: A compilation of motor vehicle crash data from the Fatality Analysis Reporting System and the General Estimate System, 2006", Washington, D.C.

Nevarez, A., Abdel-Aty, M., Wang, X. and Santos, J. B. (2009). "Large-Scale Injury Severity Analysis for Arterial Roads: Modeling Scheme and Contributing Factors", Presented at the 88th Annual Meeting of the Transportation Research Board, Washington, DC.

Noland, R. B. and Oh, L. (2004). "The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data", Accident Analysis & Prevention, Vol. 36, No. 4, pp. 525-532.

Noyce, D. A., Bahia, H. U., Yambo, J., Chapman, J. and Bill, A. R. (2007). "Incorporating Road Safety into Pavement Management: Maximizing Surface Friction for Road Safety Improvements", Midwest Regional University Transportation Center, FHWA Report.

Obeng, K. (2008). "Injury Severity, Vehicle Safety Features, and Intersection Crashes", Traffic Injury Prevention, Vol. 9, No. 3, pp. 268-276.

O'Donnell, C. J. and Connor, D. H. (1996). "Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice", Accident Analysis & Prevention, Vol. 28, No. 6, pp. 739-753.

"Oregon Safety Corridor Program Guidelines". Oregon Department of Transportation, December 2006, http://www.oregon.gov/ODOT/TS/docs/Roadway/2006Safety_Corridor_Guidelines.pdf . Accessed March 10, 2007.

"Oregon Safety Corridors". Oregon Department of Transportation, January 19, 2007, http://www.oregon.gov/ODOT/TS/docs/Roadway/CorridorMasterList2007.pdf . Accessed March 10, 2007.

Pande, A. and Abdel-Aty, M. (2008). "Discovering indirect associations in crash data using probe attributes", Transportation Research Record 2083, pp. 170-179.

Pande, A. and Abdel-Aty, M. (2009). "A novel approach for analyzing severe crash patterns on multilane highways", Accident Analysis and Prevention, Vol. 40, No. 4, pp. 1320-1329.

Pande, A., Abdel-Aty, M. and Hsia, L. (2005). "Spatiotemporal variation of risk preceding crashes on freeways", Transportation Research Record, 1908, pp. 26-36.

Papayannoulis, V., Gluck, J. S., Feeney, K. and Levinson, H. S. (1999). "Access spacing and traffic safety", retrieved February 3, 2007, from http://www.onlinepubs.trb.org/onlinepubs/circulars/ec019/EC019_C2.PDF

Park, B., Messer, C. J. and Urbanik II, T. (2000). "Enhanced genetic algorithm for signal-timing optimization of oversaturated intersections". Transportation Research Record 1727, pp. 32-41.

Parker, M. R. (1983). "Design guidelines for raised and transversable medians in urban areas", Virginia Highway and Transportation Research Council, Charlottesville, VA.

Parker, M. R. (1990). "Simplified guidelines for selecting an urban median treatment – urban median information", Virginia Transportation Technology Transfer Center, Charlottesville, VA.

"Partitioning Around Medoids", UNESCO, http://www.unesco.org/webworld/idams/advguide/Chapt7_1_1.htm Accessed September 5, 2007.

Persaud, B. N., Lyon, C. and Nguygen, T. (1999). "Empirical bayes procedure for ranking sites for safety investigation by potential safety improvements", Transportation Research Record 1665, pp. 7-12.

Petritsch, T. A., S. Challa, H. F. Huang and R. Mussa (2007). "Evaluation of Geometric and Operational Characteristics Affecting the Safety of Six-Lane Divided Roadways", Florida Department of Transportation, http://www.dot.state.fl.us/research-center/Completed_Proj/Summary_SF/FDOT_BD543_05_rpt.pdf . Accessed March 7, 2007.

Quinlan, J., R. (1986). "Induction of decision trees", Machine Learning Vol. 1, pp.81-106.

Quinlan, J., R. (1993). "C4.5: Programs for Machine Learning", Morgan Kaufmann, San Mateo, California.

Rees, J. (2003). "Corridor management: identifying corridors with access problems and applying access management treatments, a U.S. 20 study", retrieved February 10, 2007, from http://www.ctre.iastate.edu/mtc/papers/2003/JREES.pdf

Saccomanno, F. F., Grossi, R., Greco, D. and Mehmood, A. (2001). "Identifying black spots along highway SS107 in Southern Italy using two models", Journal of Transportation Engineering, Vol. 127, No. 6, pp. 515-522.

Saltelli, A., Chan, K. and Scott, E. M. (2000). "Mathematical and statistical methods: sensitivity analysis", John Wiley & Sons Ltd., West Sussex, England.

Sawalha, Z., Sayed, T. and Johnson, M. (2001). "Evaluating safety of urban arterial roadways", Journal of Transportation Engineering, Vol. 127, No. 2, pp. 151-158.

Schlutler, P. J., Deely, J. J. and Nicholson, A. J. (1997). "Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model", The Statistician, Vol. 46, No. 3, pp. 293-316.

Smith, E. D., Szidarovszky, F., Karnavas, W. J. and Bahill, A. T. (2008). "Sensitivity analysis, a powerful system validation technique", The Open Cybernetics and Systemics Journal, Vol. 2, pp. 39-56.

Souleyrette, R., Kamyab, A., Hans, Z., Knapp, K. K., Khattak, A., Basavaraju, R. and Storm, B. (2001). "Systematic identification of high crash locations", Center for Transportation Research and Education, Iowa Department of Transportation.

Spainhour, L. K., D. Brill, J. O. Sobanjo, J. Wekezer and P. V. Mtenga (2005). "Evaluation of Traffic Crash Fatality Causes and Effects: A Study of Fatal Traffic Crashes in Florida from 1998-2000 Focusing on Heavy Truck Crashes", Florida Department of

Transportation, http://www.dot.state.fl.us/research-center/Completed_Proj/Summary_SF/FDOT_BD050_rpt.pdf . Accessed March 7, 2007.

Squires, C. A., and Parsonson, P. S. (1989). "Accident comparison of raised median and two-way left-turn lane median treatments", Transportation Research Record, 1239, Transportation Research Board, National Research Council, Washington D.C., pp. 30-40.

Strobl, C., Boulesteix, A. L., Zeileis, A. and Hothorn, T. (2007). "Bias in random forest variable importance measures: illustrations, sources and a solution", BMC Bioinformatics.

Sze, N. N. and Wong, S. C. (2007). "Diagnostic Analysis of the Logistic Model for Pedestrian Injury Severity in Traffic Crashes", Accident Analysis & Prevention, Vol. 39, No. 6, pp. 1267-1278.

Teklu, F., Sumalee, A. and Watling, D. (2007). "A Genetic Algorithm Approach for Optimizing Traffic Control Signals Considering Routing", Journal of Computer-Aided Civil and Infrastructure Engineering, Vol. 22, No. 1, pp. 31-43.

Tunaru, R. (2002). "Hierarchical Bayesian Models for Multiple Count Data", Austrian Journal of Statistics, Vol. 31, No. 2-3. pp. 221-229.

"The QLIM Procedure", SAS Institute Inc., Cary, North Carolina, USA, http://support.sas.com/rnd/app/papers/qlim. Accessed June 6, 2007.

Virginia's Surface Transportation Safety Executive Committee (2006). "Commonwealth of Virginia's Strategic Highway Safety Plan", Virginia Department of Transportation, http://www.virginiadot.org/info/resources/Strat_Hway_Safety_Plan_FREPT.pdf . Accessed March 10, 2007.

Walton, M. C., Horne, T. W. and Fung, W. K. (1978). "Design criteria for median turn lanes", Federal Highway Administration, Washington, D. C.

Wang, X. and Abdel-Aty, M. (2008). "Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models", Accident Analysis & Prevention, Vol. 40, No. 5, pp. 1674-1682.

Wang, X. and Abdel-Aty, M. (2006). "Temporal and spatial analyses of rear-end crashes at signalized intersections", Accident Analysis & Prevention, Vol. 38, No. 6, pp. 1137-1150.

Wang, X. and Abdel-Aty, M. and Brady, P. A. (2006). "Crash estimation at signalized intersections: significant factors and temporal effect", Transportation Research Record 1953, pp. 10-20.

Wang X., Abdel-Aty, M., Nevarez, A. and Santos J. B. (2008). "Investigation of safety influence area for four-legged signalized intersections: nationwide survey and empirical inquiry", Transportation Research Record 2083, pp. 86-95.

Warner, M. H. and Warner, C. Y. (2008). "Fatal and Severe Injuries in Rear Impact: Seat Stiffness in Recent Field Accident Data", SAE International.

"Washington/Corridor Safety Project". NHTSA, spring 1997, http://www.nhtsa.dot.gov/people/outreach/safedige/spring1997/n4-24.html Accessed March 6, 2007.

"Washington Cape Horn Corridor Traffic Safety Project". NHTSA, Traffic Safety Digest, Vol. 2, 2004, http://www.nhtsa.dot.gov/people/outreach/safedige/Volume-2-2004/Vol2_04_W04_WA.htm . Accessed March 16, 2007.

"Washington State Corridor Safety Program". Washington Traffic Safety Commission, 2006, http://www.corridorsafetyprogram.com/ . Accessed March 1, 2007.

Yamamoto, T., Hashiji, J. and Shankar, V. N. (2008). "Underreporting in traffic accident data, bias in parameters and the structure of injury severity models", Accident Analysis & Prevention, Vol. 40, No. 4, pp. 1320-1329.

Yan, X., Harb, R. and Radwan, E. (2008). "Analysis of factors of crash avoidance maneuvers using the general estimates system", Traffic Injury Prevention, Vol. 9, No. 2, pp. 173-180.

Zegeer, C. V., Huang, H. F., Stutts, J. C., Rodgman, E. and Hummer, J. E. (1994). "Commercial bus accidents characteristics and roadway treatments", Transportation Research Record 1467, pp. 14-22.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G. and Mao, Y. (2000). "Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario", Accident Analysis & Prevention, Vol. 32, No. 1, pp. 117-125.

Zhou, H., Lu, J. J., Castillo, N. and Yang, X. K. (2001). "Operational and safety effects of replacing a full median opening with directional median opening", ITE 2001 Annual Meeting and Exhibit, Chicago.

Zogby, J. J., T. E. Bryer and J. Tenaglia (1991). "Pennsylvania Corridor Highway Safety Improvement Program". TR News 154, May-June 1991, pp. 11-13.