A COMPREHENSIVE SEVERITY ANALYSIS OF LARGE VEHICLE CRASHES

by

Haluk Laman
B.Sc. Department of Civil Engineering, Cukurova University, 2009

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

Fall Term
2012

ABSTRACT

The goal of this thesis is to determine the contributing factors affecting severe traffic crashes (severe: incapacitating and fatal - non-severe: no injury, possible injury, and non-incapacitating), and in particular those factors influencing crashes involving large vehicles (heavy trucks, truck tractors, RVs, and buses). Florida Department of Highway Safety and Motor Vehicles (DHSMV) crash reports of 2008 have been used. The data included 352 fatalities and 9,838 injuries due to large vehicle crashes.

Using the crashes involving large vehicles, a model comparison between binary logit model and a Chi-squared Automatic Interaction Detection (CHAID) decision tree model is provided. There were 13 significant factors (i.e. crash type with respect to vehicle types, residency of driver, DUI, rural-urban, etc.) found significant in the logistic procedure while 7 factors found (i.e. posted speed limit, intersection, etc.) in the CHAID model. The model comparison results indicate that the logit analysis procedure is better in terms of prediction power.

The following analysis is a modeling structure involving three binary logit models. The first model was conducted to estimate the crash severity of crashes that involved only personal vehicles (PV). Second model uses the crashes that involved large vehicles (LV) and passenger vehicles (PV). The final model estimated the severity level of crashes involving only large vehicles (LV). Significant differences with respect to various risk factors including driver,

vehicle, environmental, road geometry and traffic characteristics were found to exist between those crash types and models. For example, driving under the influence of Alcohol (DUI) has positive effect on the severity of PV vs. PV and LV vs. PV while it has no effect on LV vs. LV. As a result, 4 of the variables found to be significant were similar in all three models (although often with quite different impact) and there were 11 variables that significantly influenced crash injury severity in PV vs. PV crashes, and 9 variables that significantly influenced crash injury severity in LV vs. PV crashes.

Based on the significant variables, maximum posted speed, number of vehicles involved, and intersections are among the factors that have major impact on injury severity. These results could be used to identify potential countermeasures to reduce crash severity in general, and for LVs in particular. For example, restricting the speed limits and enforcing it for large vehicles could be a suggested countermeasure based on this study.

ACKNOWLEDGMENTS

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Mohamed Abdel-Aty, who has supported me throughout my thesis with his patience and knowledge. I would also like to thank Dr. Omer Tatari and Dr. Mohamed Ahmed for their invaluable suggestions and kind acceptance to be in the committee.

Special thanks to Jaeyoung Lee, Rongjie Yu, and Muamer Abuzwidah for their guidance and help in the use of statistical software. I would like to thank my colleagues and my research group, in particular for their warm companionship and encouragements.

Thanks are also due to all my friends here in Orlando and back home: Akif Sahin, Taha Bal, Mustafa Ozkul, Tolga Ercan, Suphi Civelek, Bugrahan Aslan, Tolga Yardimci, Selcuk Bildik, Mustafa Akdag, and Yunus Emre Acikgonul.

Last but not least, I am eternally grateful for my parents; Mustafa and Sirin, my sisters Mihriban Tugba and Elif Betul and my intended wife Arzu Arslan. Without their love and support I would never have completed this task.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CHAID	Chi-squared Automatic Interaction Detection
DF	Degrees of Freedom
DHSMV	Department of Highway Safety and Motor Vehicles
DUI	Driving Under Influence
FDOT	Florida Department of Transportation
FHSMV	Florida Highway Safety and Motor Vehicles
LV	Large Vehicle
PV	Passenger Vehicle
VMT	Vehicle Miles Travelled

CHAPTER ONE: INTRODUCTION

Deaths, injuries and traffic congestions keep traffic safety as a prominent research topic in the field of transportation engineering. The nature and extent of roadway crashes vary by a wide range depending on roadway types and facility, driver characteristics and land-use patterns among other factors. Since crashes associate with complex interactions of numerous factors, micro level crash analysis (e.g., road specific crash analysis, crash specific safety analysis, event specific analysis) allows more insight for causes of a crash.

According to the Florida Highway Safety and Motor Vehicles the death rate in Florida is 1.5 per 100 million Vehicle Mile Travelled (VMT) (Florida Highway Safety and Motor Vehicles, 2008). From the 363,206 crashes reported, 693,832 vehicles were involved. These caused 2,983 casualties and 199,658 injuries (FHSMV, 2008). Data maintained by the DHSMV (Department of Highway Safety and Motor Vehicles) in 2008 indicated that 282 persons were killed and 9,159 were injured out of 22,277 in crashes involving large vehicles in Florida.

1.1 Research Motivation and Objectives

Studies analyzing crash injury severity often focus on crash frequencies. Multiple factors effect crash frequency and severity.

- roadway geometrics
- traffic conditions
- roadway and environmental conditions
- driver and vehicle characteristics

In this study, these factors are considered in order to make a statement regarding large truck and bus safety in Florida.

As shown in this thesis the factors provided above, that affect the crash frequency and severity, will be analyzed based on crash injury severity through several logistic regression models and a Chi-squared Automatic Interaction Detection (CHAID) decision tree model.

The objective of this study is to focus on the injury severity caused by large vehicles. The LV's are grouped as heavy trucks, truck tractors, RVs, and buses.

1.2 Organization of the Thesis

The thesis has been organized in the following format. Following the introductory chapter a detailed literature review is provided in chapter two; previous studies conducted in large vehicle crash severity analysis have been

critically reviewed along with the different groups of considered regressors in the corresponding studies. This chapter also summarizes the crash severity analyses from different groups of crashes.

The next chapter (chapter three) describes data preparation steps. Datasets and variables used in the analysis are explained in this chapter. The statistical modeling approach of this study is described in chapter four.

The following chapter (chapter five) provides some preliminary analysis, which includes descriptive statistics from different datasets and distributions of crash factors. This is examined through large vehicle involvement based on severity.

Models and results from the analyses are presented in chapter six. This chapter provides a comprehensive discussion regarding the association (direction and magnitude) of different significant parameter estimates from different models developed in this study. The final chapter consists of the summary, conclusions and recommended future work.

CHAPTER TWO: REVIEW OF LITERATURE

There are two sections in this chapter. A synthesis of literature on the analysis of injury severity of such crashes with particular focus on large vehicle crashes is presented in the first section. The injury severity analysis in traffic safety is a widely researched area. The second section provides also crash injury severity analysis literature for different types of vehicles.

2.1 Large Vehicle Related Crash Injury Severity Analysis

Chang and Mannering (1999) analyzed the injury severity and vehicle occupancy for truck-involved crashes and non-truck-involved crashes of nationwide US data by estimating nested logit models. Variables which significantly increase the severity only for truck-involved crashes are higher speed limits and type of collision. Injury severity is noticeably worsened if the crash has a truck involved. The effects of trucks are more significant for multi-vehicle crashes than single-vehicle crashes. Abdel-Aty and Abdelwahab (2004) also developed the same type of models to show the association between large vehicle type crashes (light truck vehicle, vans, and sport utility vehicles (SUV)) on drivers' visibility and rear-end collisions. According to the results, drivers' visibility, speed and inattention have the largest effect on being involved in a rear-end collision.

Khattak et al. (2003) developed binary probit models to examine the injury severity on large truck rollovers for only single-vehicle crashes. The results stated that driver behaviors as speeding, use of alcohol or drug, traffic violations have higher risk factors in single-vehicle truck crashes.

Lyman and Braver in 2003 made exploratory data analysis for 25 years of US nationwide large truck crashes by exposure measures such as; occupant fatalities per 100,000 population, per 10,000 licensed drivers, per 10,000 registered trucks, and per 100 million vehicle miles traveled. Trends in occupant deaths in large truck involved crashes are shown in the results. USDOT (2006) provided an exploratory analysis conducted to a sample of large truck involved crashes which all include a fatality or an injury from crash reports for 33 months at 24 sites in 17 states. As a result, it is shown that 87 % of the crashes have occurred due to driver actions and poor driving decisions. 13% of the coded reasons were the weather conditions, or roadway problems.

Cantor et al. (2010) focused on truck prediction modeling using poisson regression models. Driver age, weight, gender, and employment stability etc. are significantly related to the likelihood of crash occurrence. Poorly maintained vehicles have also poor safety performance according to the results.

Numerous researches have been used logistic regression in the crash severity analysis. Khorashadi et al. (2005) used the 4 years of California crash data and analyzed by multinomial logit models to determine the differences in

rural and urban driver injury severities (both passenger-vehicle and large-truck driver injuries) in crashes that involve large trucks. Intersection related crashes at rural areas result in a significantly increase in a likelihood of severe/fatal injury. In both area type DUI is the most influential variable to be involved in a severe/fatal crash. It is also shown that geometrics, environmental conditions, and driver actions have also significant effects on severe/fatal crash occurrences.

Nassiri and Edrissi in 2006 made a comparison between neural networks and logit modeling using vehicle crash data on two-lane rural highways for truck crashes. The results of both models have significant factors such as roadways, vehicles, environment, and drivers. The research by Chen and Chen (2011) shows truck driver injury severities' differences between single-vehicle and multi-vehicle crash types by estimating mixed logit models. In this paper the analysis revealed that several risk factors may lead to more severe injuries of truck drivers such as; age, asleep or fatigued driver, carrying hazardous material, wide median, truck overturn, etc.

A different approach to injury severity analysis was used in Islam and Hernandez' (2011) study which is random parameters tobit regression modeling with crash rates instead of crash frequency. US nationwide crash database is used. The exposures were truck miles traveled and ton-miles of freight. Road surface condition, road geometry, time, day, and month of the crash were all found significant.

Lemp et al. (2011) said that size and weight regulations of large trucks triggered by safety concerns. They used Heteroskedastic ordered probit models to study the impact of vehicle, occupant, driver, and environmental characteristics on injury severity outcomes for those involved in crashes with large trucks. In the results it is mentioned that non-bright lighting conditions or road surface conditions are increasing the fatality risk of the crashes while the number of truck-trailers are also increasing the likelihood of fatality. The same approach was developed to analyze the injury severities of all persons involved in a large truck crash by Zhu and Srinivasan in 2011. Driver behaviors such as; DUI, illegal drug use, inattention were found to be significant predictors on severity. Drivers' familiarity with the vehicle is also a significant factor which is also related to the owner is driver variable in this research.

In a different research, Zhu and Srinivasan (2011) analyzed the factors affecting the overall injury severity of large truck crashes of a national recent data sample with empirical models. Results provides numerous significant variables such as; driver distraction (truck driver), alcohol use (car drivers), and emotional factors (car drivers).

Finally, Chang and Chien (2013) used non-parametric Classification and regression tree (CART) method to establish the empirical relationship between injury severity outcomes and driver/vehicle characteristics under 2005-2006 truck involved crash data from national freeways in Taiwan. Results are showing that

drunk driving is the most important determinant for the injury severity of truck crashes on freeways. Vehicle types, number of vehicles involved in the crashes are also significant factors on severity of the truck involved crashes.

2.2 Crash Injury Severity Analysis

Shankar et al. (1996) used a nested logit model to estimate the crash severity on rural freeways with a 5-year data from a 61 km section of a rural interstate in Washington State. The estimation results show that environmental conditions, highway design, crash type, driver characteristics, and vehicle attributes have valuable effect on the crash severity.

Chen (1997) developed a series of discrete categorical analyses to determine the association of crash location, type, and driver variables and the severity of the resulting crash using the HSIS data for the years 1994-1997. Car-semitrailer crashes and rural areas found the most likely types to be involved in a severe crash. Desapriya et al. (2006) compared severity of alcohol related vs. non-alcohol related motor vehicle crashes with odds ratios and CI's. Also looks at severely damaged vehicles besides of injury severity.

Kuhnert et al. (2000) presented the advantages of non-parametric models such as CART and MARS (multivariate adaptive regression splines) which can provide more informative and attractive models than logistic regression models. Chang and Wang in 2006 also used the CART model from 2001 crash data for

Taipei, Taiwan. The results indicate that the vehicle type is the most significant variable associated with the crash severity. Pedestrians, motorcycle and bicyclists have higher risks of being involved in a severe crash. Das and Abdel-Aty (2009) developed conditional inference forests, which are ensembles of individual CART algorithms, are applied for identifying traffic/highway design/driver-vehicle information significantly related to fatal/severe crashes on urban arterials for different crash types. Alcohol/drugs and higher posted limits contribute to severe crashes.

Artificial Neural Networks are also widely used in severity analysis. Abdelwahab and Abdel-Aty (2002) developed MAP (fuzzy ARTMAP) neural networks to analyze the injury severity for drivers involved in traffic crashes at highways, signalized intersections, and toll plazas. Models for each crash location type show vehicle speed at the time of crash increases the likelihood of high injury severity. Drivers in passenger cars are also more likely to have a severe crash than those who drive vans or pickup trucks. Rural area, nighttime, and drunk driver crashes have also higher risk to be involved in severe crashes according to the results. Abdel-Aty and Abdelwahab (2004) developed another ANN model; MLP (multilayer perceptron), ART (fuzzy adaptive resonance theory) and a calibrated ordered probit model in order to compare based on injury severity level. According to the results; gender, vehicle speed, seat belt use, type of vehicle, point of impact, and area type (rural vs. urban) affect the likelihood of

injury severity levels. Female and/or drunk drivers have higher chances of experiencing a severe injury. Nighttime and rural areas are riskier in terms of driver injury severity. Speeding have positive effect on the severity of the crash (not the speed limit, speed ratio). Finally, Delen et al. in 2006 used eight binary MLP neural networks model to estimate the potentially non-linear relationships between the severity and crash related factors. Seat-belt use, driving under the influence of alcohol or drugs, age and gender of the driver, and vehicle role in the accident found to be influential on the outcome of the crash. The weather conditions did not seem to affect the severity level of injury.

Logistic regression models are the most popular methodology in severity analysis. Al-ghamdi (2002) made the binary dependent variable as fatal or non-fatal in the logistic regression model in order to examine the contribution of several variables to crash severity. Location and cause of crash found the most significant variables. For the cause of crash, speed is the highest level while the road section is the highest level influencing the severity. Binary logit models were also performed by Kieliszewski in 2006 to further scope predictor variables to identify traffic event characteristics with respect to severity level, maneuver type, and conflict type. Another binary logistic regression modeling procedure was used by Sze and Wong (2007) to determine the association between the probability of fatality and severe injury and all contributory factors. Das et al. (2008) used simultaneous estimations as probit and logit models to identify

factors contributing injury severity on intersections on an urban arterial corridor. As a result, more severe crashes occur on blacktop surfaces, and segments with higher speed limit, wider pavement surface, and lower and median AADT. In some cases dry pavement conditions is also significant contributing the severity.

Nevarez et al. (2009) used the logistic regression models in two phases. First phase included all drivers and roadway locations. The second phase involved an extension of these models, controlling by crash types. The crash types models showed important contributing factors such as speeding, use of alcohol or drug, type of vehicle. Huang et al. in 2010 developed multilevel ordered logit model methodology to identify the contribution of influential factors and injury severity level under fog or smoke related traffic crashes. According to the results, higher speed, undivided, no sidewalk, two lane rural roads, and at night without street light crashes are riskier in terms of injury severity level. Theofilatos et al. in 2012 used two binary logistic regression models to estimate the probability of fatality/severe injury versus slight injury inside and outside the urban areas. As a result, involvement of motorcycles, bicycles, and buses were significantly riskier based on severity for outside urban areas, while weather conditions and involvements of buses or motorcycles were significantly riskier inside urban areas.

Ordered probit models are also common in analyzing crash severity. For example; Kockelman and Kweon in 2002 examined the risk of different injury

severity levels with this method under a model structure; all crash types, two-vehicle crashes, and single-vehicle crashes. According to the results, pickups and SUV's are less safe than passenger vehicles under single-vehicle crash conditions. Light trucks protect their drivers better than any other vehicles. Abdel-Aty (2003) analyzed driver injury severity levels using the ordered probit and nested logit modeling methodology. Roadway sections, signalized intersections, and toll plazas in Central Florida are considered. Alcohol, lighting conditions affected the severity level on roadway sections' model. Passenger cars and those who speed have higher risk to experience a severe crash. Abdel-Aty and Keller (2005) used the same model and tree-based regression methodology and adopted in the research to understand the factors that contributes the injury severity at intersections. Ordered probit model results show that higher speed limit decreases the severity level while crashes involving a pedestrian/bicyclist had the highest probability to be involved in a severe crash. Tree-based regression model also indicates the higher posted limit on the minor roads significantly affected lower injury severity levels. Haleem and Abdel-Aty (2010) estimated three approach to analyze the crash injury severity level at three- and four-legged unsignalized intersections: First, ordered probit model with five levels of injury severity; second approach is a binary probit model with severe vs. non-severe injury; and last approach dealt fitting a nested logit model. Results are showing important effects of traffic volume and driver factors on injury severity.

Last but not least, linear genetic programming (LGP) method is used to distinguish the relationship of geometric and environmental factors with injury related crashes and severe crashes by Das and Abdel-Aty (2010). As a result, dry surface conditions, good pavement conditions, wider shoulders, and sidewalk widths decrease the severity of crashes. Higher posted limit is found to make the injuries more possible according to the results of LGP.

Table 1: Summary of Reviewed Related Literature

Author, year	Design-Respondent	Methodology	Major finding and significant factors
Chang and Mannering, 1999	truck/non truck-severity	nested logit	higher speed limits, truck involvement
Khattak et al., 2003	large truck-single veh.-severity	binary probit	speeding, DUI
Lemp et al., 2011	large truck-severity	heteroskedastic ordered probit	non-bright lighting and road surface conditions
Khorashadi et al., 2005	large truck-severity	multinomial logit	intersection related, rural areas, DUI
Theofilatos et al., 2012	area type-severity	binary logit	involvement of motorcycles, bikes, buses (urban); weather conditions (rural)
Chang and Chien, 2013	truck-severity	CART-tree	DUI, type of vehicle, number of vehicles involved
Das and Abdel-Aty, 2009	arterial corridors-severity analysis	CART-tree	alcohol/drugs, higher speed limits
Abdel-Aty and Abdelwahab, 2004	light-truck rear-end	nested logit	visibility, speed, inattention of drivers
Cantor et al., 2010	truck-occurrence	poisson regression	driver age, weight, gender
Islam and Hernandez, 2011	large truck-fatality	tobit regression	road surface condition, road geometry
Zhu and Srinivasan, 2011	large truck-severity analysis	heteroskedastic ordered probit	DUI, inattention of drivers
Chang and Wang, 2006	vehicle type-severity	CART-tree	pedestrians, bicyclists, motorcycle involvements

Table 1 provides a summary of some of the reviewed literature. The severity analysis in this thesis follows a similar pattern to the literature that has been presented in this chapter. Logistic regression models and a CHAID decision tree model were developed and analyzed. The prediction power of logistic procedure was compared with CHAID model. In this study, new factors that were not discussed in previous literature were introduced, such as the bus or truck involvement, blacktop/concrete road surface type comparison, shoulder existence of the roadway, and residence of Florida. The preparation of the data used in the models will be presented in the next chapter.

CHAPTER THREE: DATA PREPARATION

The source of data for this study is the Florida Traffic Crash Reports, maintained by the Department of Highway Safety and Motor Vehicles. These crash reports are used by law enforcement officers in Florida to report traffic crashes to the Department of Highway Safety and Motor Vehicles. In this chapter; the DHSMV data, the datasets and the variables used in the analysis will be elaborated.

The crash data have been obtained from the Department of Highway Safety and Motor Vehicles (DHSMV), for year 2008. The DHSMV traffic crash database is a relational database consisting of nine files. Each file deals with a specific aspect of a traffic crash. The files are as follows:

1. Events file; contains general information about the crash event characteristics and circumstances. This is the "parent file" of the database.
2. Vehicles file; contains information about each vehicle and their actions in the traffic crash.
3. Drivers file; contains information about each driver involved and condition or action of the driver that contributed to the crash.
4. Property file; contains information about property (other than vehicles) damaged in the crash

5. Pedestrians file; deals with information on any pedestrians involved in the traffic crash (demographic and casual).

6. Violations file; lists the citations (if any) issued in connection with the traffic crash, by statute number. (limited to the first eight citations issues per party)

7. Passengers file; provides information about any passengers involved in the traffic crash.

8. ComVeh file; contains information about commercial vehicles and carriers involved in crashes.

9. D.O.T. Site Location file; contains additional information about Department of Transportation crash locations occurring on state roads only.

3.1 Preparation of Datasets Used in the Analysis

There were four different datasets used in the modeling procedures. The first dataset (Dataset-A) consisted of the large vehicle (LV) crashes. It was prepared by choosing the crashes which contained at least one LV out of all types in crashes. Second dataset (Dataset-B) only involved the passenger vehicles vs. passenger vehicle crashes. The passenger vehicles (PV) were grouped as automobile, van, light truck, and medium truck. They can be defined as smaller vehicles compared to the LV's. Third dataset (Dataset-C) was prepared by choosing only the LV vs. PV crashes out of other type of vehicle crashes. Dataset-A is different than dataset-C in which the first may also involve

different type of vehicles which were not defined as LV or PV (i.e. motorcycle, bike, etc.). Dataset-C, only contained LV vs. PV crashes. Last dataset (Dataset-D) was defined as the crashes occurred between LV's which means that only two or more LV's were involved in those crashes. The variables taken from the drivers' file of DHSMV were all chosen for the LV drivers for these crash datasets accept of dataset-B which is LV vs. LV crashes dataset. In dataset-B, driver characteristic variables were not LV drivers' but one of the involved PV drivers' characteristics. Exploratory analyses and five different modeling procedures are estimated using the above mentioned four datasets.

Missing values were found for many of the variables. The value of certain variables could be more likely to be missing for severe crashes while the value of other variables could be more likely to be non-severe crashes. Therefore, removing the missing values would skew the sample. So, it is chosen to retain all cases by imputing with the most frequent level of each variable.

3.2 Predictor and Response Variables Considered in the Analysis

In this study, the variables used in the models are crash injury severity, type of vehicle, lighting condition, weather, rural/urban, owner is driver, residence code, road surface type, road surface conditions, type of shoulder, alcohol/drug use, site location, on-off roadway, divided/undivided highway, total number of vehicles, posted speed. These variables were defined as follows.

- Crash injury severity: This variable is from the events file. So, it contains every person involved in the crash. The levels are as seen in Table 1. The dummy codes are also given as it is used in the models. Incapacitating and fatal levels are grouped as severe crashes while the rest of the levels are defined as non-severe.

Table 2: DHSMV Crash Severity Levels

Severity Level	Description	Dummy code
1	No injury	0
2	Possible Injury	0
3	Non-incapacitating evident injury	0
4	Incapacitating injury (Severe)	1
5	Fatal (within 30 days)	1

- Type of vehicle: This variable is from the vehicles file. The classification is as seen in Table 2. The large trucks group contains; heavy truck (05), truck-tractor (06), motor home (07) and the buses group contains; bus (driver + seats for 9-15) (08), bus (driver + seats for over 15) (09) in the models.

Table 3: DHSMV Type of Vehicle Classification

Code	Description
01	Automobile
02	Van
03	Light Truck/Pick Up (2 or 4 rear tires)
04	Medium Truck (4 rear tires)
05	Heavy Truck (2 or more rear axles)
06	Truck-Tractor (Cab - Bobtail)
07	Motor Home (RV)
08	Bus (driver + seats for 9-15)
09	Bus (driver + seats for over 15)
10	Bicycle
11	Motorcycle
12	Moped

13	All Terrain Vehicle
14	Train
15	Low Speed Vehicle
77	Other
0	Unknown and/or Dummy Record

- Lighting condition: This variable is from the events file. The classification is as seen in Table 3. Non-bright lighting conditions are defined as, (05) dark (no light), dusk, and dawn in the models.

Table 4: DHSMV Lighting Condition Classification

Code	Description	Dummy Code
01	Daylight	0
02	Dusk	1
03	Dawn	1
04	Dark (Street Light)	0
05	Dark (No Light)	1
88	Unknown	0

- Weather: This variable is from the events file. The classification is as seen in Table 4. Only events occurred in rainy weathers are considered in the modeling.

Table 5: DHSMV Weather Classification

Code	Description	Dummy Code
01	Clear	0
02	Cloudy	0
03	Rain	1
04	Fog	0
77	All Other	0
88	Unknown	0

- Rural/urban: This variable is from the events file. The dummy codes were defined as, rural – 0, urban – 1 in the analysis.

- Owner is driver: This variable is from the vehicles file. The dummy codes were defined as, the driver is not the owner – 0, owner is driver – 1 in the analysis.

- Residence code: This variable is from the drivers file. The classification is as seen in Table 6. Drivers whom are residents of the state of Florida were coded as (0), and the non-resident drivers were coded as (1) in the modeling procedure.

Table 6: DHSMV Residence Code Classification

Code	Description	Dummy Code
1	1 County Of Crash	0
2	2 Elsewhere In State	0
3	3 Non-Resident	1
4	4 Foreign	1
5	5 Unknown	1

- Road surface type: This variable is from the events file. The classification is as seen in Table 7. In this research the road surface type variable was used to compare the injury severity level between blacktop surface type and concrete surface type. The blacktop surface type was coded as (0) while the concrete surface type was coded as (1).

Table 7: DHSMV Road Surface Type Classification

Code	Description
01	Slag/Gravel/Stone
02	Blacktop
03	Brick/Block
04	Concrete
05	Dirt
77	All Other
88	Unknown

- Road surface conditions: This variable is from the events file. The classification is as seen in Table 8. The road surface conditions variable was coded in the models as, dry (0) and others (1) which defined as bad road conditions.

Table 8: DHSMV Road Surface Conditions Classification

Code	Description
01	Dry
02	Wet
03	Slippery
04	Icy
77	All Other
88	Unknown

- Type of shoulder: This variable is from the events file. The classification is as seen in Table 9. The type of shoulder variable was coded in the models as, unpaved (1) and others (0).

Table 9: DHSMV Type of Shoulder Classification

Code	Description
01	Paved
02	Unpaved
03	Curb
88	Unknown
00	N/A

- Alcohol/drug use: This variable is from the drivers file. The classification is as seen in Table 10. The alcohol/drug use variable was coded in the models as; not drinking or using drugs, pending BAC test results, unknown (0) means non-alcohol/drug use, alcohol - under Influence, drugs - under influence, alcohol & drugs - under influence, had been drinking (1) means DUI (Driving under Influence).

Table 10: DHSMV Driver Alcohol/Drug Use Classification

Code	Description
1	1 Not Drinking or Using Drugs
2	2 Alcohol - Under Influence
3	3 Drugs - Under Influence
4	4 Alcohol & Drugs - Under Influence
5	5 Had Been Drinking
6	6 Pending BAC Test Results
0	0 Unknown and/or Dummy Record

- Site location: This variable is from the events file. The classification is as seen in Table 11. The site location variable was considered only for intersection related crashes or not. It was coded in the models as; At

Intersection, Influenced by Intersection (1) means intersection related crash while the rest of the classes were (0) means not intersection related crash.

Table 11: DHSMV Site Location Classification

Code	Description	Dummy Codes
01	Not at Intersection/RR X-ing/Bridge	0
02	At Intersection	1
03	Influenced by Intersection	1
04	Driveway Access	0
05	Railroad	0
06	Bridge	0
07	Entrance Ramp	0
08	Exit Ramp	0
09	Parking Lot - Public	0
10	Parking Lot – Private	0
11	Private Property	0
12	Toll Booth	0
13	Public Bus Stop Zone	0
77	All Other (Explain in Narrative)	0

- On-off roadway: This variable is taken from the events file. The dummy codes for the modeling is as follows; on roadway (0), off roadway (1).

- Divided/undivided highway: This variable is taken from the events file. The dummy codes for the modeling is as follows; divided highway (0), undivided highway (1).

- Total number of vehicles: This variable is taken from the events file. It is the sum of all vehicles involved in the crash. In this research it was used

as; more than two vehicle involved crashes (code: 1) and two or less vehicles involved in the crash (code: 0).

- Posted speed: This variable is taken from the vehicles file in order to code the speed variable as crashes occurred on less than 45 mph posted speed limit roadway (code: 0) or more than 44 mph posted speed limit roadway (code: 1). The classification is based on the median (46 mph) of the speed limits.

- Involved vehicle type: This variable is prepared for the dataset of the severity analysis of large vehicle involved crashes. The description of the levels for this variable is as seen in Table 12.

Table 12: Involved Vehicle Type

Level	Description
0	Large Vehicle – Large Vehicle Crashes
1	Large Vehicle – Passenger Vehicle Crashes
2	Single Large Vehicle Crashes
3	Large Vehicle – Bicyclist/Pedestrian/Moped Crashes
4	Large Vehicle – Motorcycle Crashes

To sum up, brief descriptions of all the variables used in the series of binary logistic regression models and decision tree model are as seen in Table 13.

Table 13: Variable Description

Variable Name	Definition
Injury Severity	Target variable: 1, severe/fatal; 0, non-severe
Lighting	1, Bright lighting; 0, non-bright lighting
Rain	1, Rainy; 0, not rainy
Rural/urban	1, Urban area; 0, rural area
Owner is Driver	1, Owner is driver; 0, owner and driver are not same
Florida Resident	1, Florida resident; 0, non-resident
Blacktop/Concrete	1, Blacktop; 0, concrete
Shoulder	1, No shoulder; 0, with shoulder
Road Surface Condition	1, Bad road condition; 0, dry
DUI	1, DUI; 0, non-alcohol/drug use
Intersection Related	1, Intersection related; 0, not intersection related
On/Off Roadway	1, Off roadway; 0, on roadway
Divided/Undivided	1, Undivided Highway; 0, divided highway
Bus/Truck	1, Large truck; 0, bus
More Than 2 Vehicles	1, More than two vehicles; 0, two or less vehicles
Speed	1, More than 44 mph posted limit; 0, less than 45 mph posted limit
Involved Vehicle Type	1, LV-PV; 2, single LV; 3, LV-bicyclist/pedestrian/moped; 4, LV-motorcycle; 0, LV-LV

The preparation of datasets used and variables conducted in crash injury severity analysis have been elaborated in this chapter. The methodology used for the modeling procedures will be explained in the next chapter.

CHAPTER FOUR: METHODOLOGY

The statistical models used in this thesis are binary logistic regression and CHAID decision tree procedure. In this chapter these two model methodologies will be explained.

4.1 Binary Logistic Regression Model

In order to analyze the crash injury severity in large vehicle involved crashes, binary logistic regression models were estimated with the consideration of statistically significant factors.

The formula of the logistic model is as follows (Greene, 2003):

$$\mathrm{Prob}(\mathrm{Y}=1|\mathrm{x}) = \frac{\exp^{\mathrm{x}\beta}}{1+\exp^{\mathrm{x}\beta}}$$

where β is a vector of the coefficient estimates of the parameters and X is a vector of independent variables. Odds ratio is a measure of association which approximates relative risk or in other words, how much more likely it is for the outcome to be present among those with x = 1 than among those with x = 0. (Hosmer and Lemeshow, 2000)

4.2 Chi-squared Automatic Interaction Detection (CHAID) Decision Tree Model

CHAID uses a Chi-square splitting criterion as indicated by its name. More specifically, it uses the p-value of the Chi-square:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O: the frequencies observed.

E: the frequencies expected.

The main characteristics of CHAID are:

(1) CHAID determines for each potential predictor the optimal *n*-ary split it would produce at each node, and selects the predictor on the basis of these optimal splits. (Ritschard, 2010).

(2) The search for a split on an input peaks gradually. Initially a branch for each value of the input signal is assigned. Branches merged alternately split and again seems justified by the p-values. The CHAID algorithm by Kass ends when no merge or split again provides a corresponding p-value. The last split is adopted. An alternative, sometimes called the exhaustive process still divides merge to form a binary split, and then takes the split with the lowest p-value among all the algorithms considered. Once a split is assumed for an input, its p-value is adjusted, and the input with the best matched p-value is selected as the split variable. When the p-value is set to be smaller than a threshold the user

specified, then the node is split. When all the adjusted p-values of the splitting variables in the unsplit nodes are above the user-specified threshold, the tree construction ends.

The two methodologies of models have been described in this chapter. The next chapter will be providing descriptive statistics.

CHAPTER FIVE: PRELIMINARY ANALYSIS

This chapter presents the preliminary exploration of the nature and characteristics of the variables in the final prepared datasets which were described in Chapter 3. The preliminary analysis included descriptive statistics and exploratory analysis for the crashes involving large vehicle, only large vehicle crashes, large vehicle vs. personal vehicle crashes, and only personal vehicle crashes.

There are 22,632 crashes involving large vehicles (LV) and 265,848 crashes not involving LV's. So, the LV involved crashes are 8% of the whole population of crashes. The incapacitating and fatal crash proportions in crash frequencies of LV and non-LV crashes are provided in Figure 1. The percentage of incapacitating crashes in the LV is slightly less than the non-LV crashes while the LV crashes have higher proportion of fatal crashes.

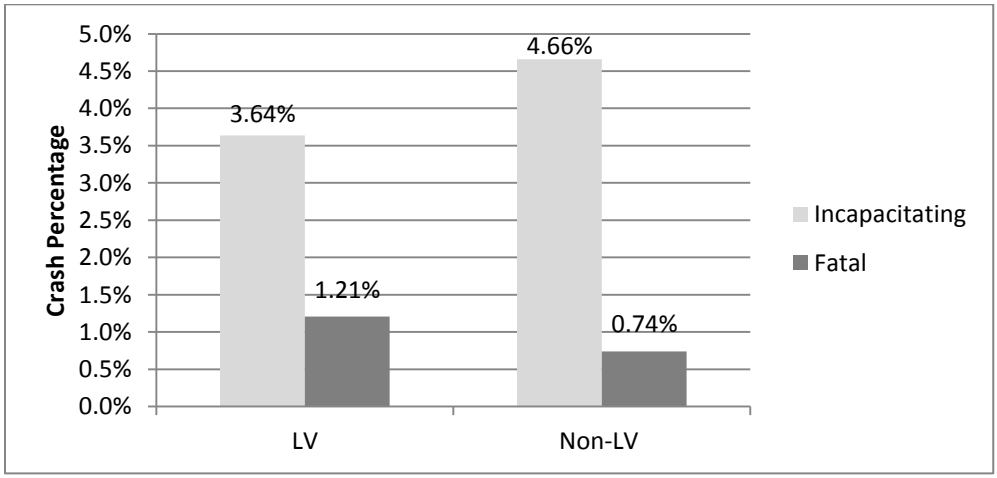

Figure 1: Distribution of LV Involvement by Incapacitating and Fatal Crash Percentages

A distribution of crash groups, such as LV (large vehicle) vs. LV, single LV, LV vs. PV, LV vs. motorcycle, and LV vs. bike/ped/moped, by severe crash (incapacitating and fatal) proportions out of the number of crashes for each group is provided in Figure 2. The number of crashes occurred for each group is as follows;

- LV-LV: 2,662 (severe: 92),
- Single LV: 3,368 (severe: 125),
- PV-LV: 16,356 (severe: 844),
- LV-bike/ped: 92 (severe: 35),
- LV-motorcycle: 153 (severe: 47).

The LV vs. motorcycles have the largest percentage of being severe crashes while LV vs. LV crash groups have the smallest percentage of being a severe crash.

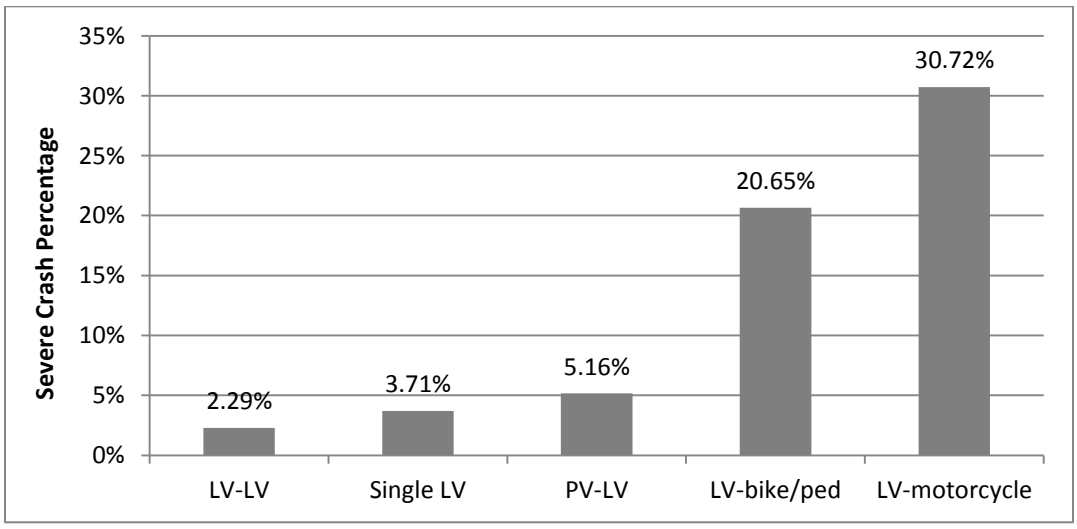

Figure 2: Distribution of Crash Groups by Severe Crash Percentages

Table 14 provides the Chi-square and p-values of variables by crash groups such as; large vehicle (LV) vs. LV, LV vs. passenger vehicle (PV), and PV vs. LV. The non-severe and severe crash frequencies are used in these descriptive statistics.

Table 14: Chi-square and p-values of Variables by Crash Groups

	LV vs. LV		LV vs. PV		PV vs. PV	
Variables	**Chi-sq**	**p-value**	**Chi-sq**	**p-value**	**Chi-sq**	**p-value**
Lighting Condition	62.98986	<0.001	134.7827	<0.001	1211.654	<0.001
Rural-Urban	32.61907	<0.001	119.9021	<0.001	1498.326	<0.001
Owner is Driver	0.166368	0.6834	2.290597	0.1302	29.57266	<0.001
Blacktop/Concrete	2.835911	0.0922	5.617744	0.0178	95.53463	<0.001
Shoulder	0.018321	0.8923	23.2749	<0.001	636.5581	<0.001
Road Surface Conditions	2.146991	0.1428	1.44567	0.2292	30.98036	<0.001
DUI	0.139421	0.7089	116.9739	<0.001	1896.291	<0.001
Intersection	6.796141	0.0091	5.135433	0.0234	118.9802	<0.001
On/off Roadway	19.1524	<0.001	22.63856	<0.001	17.28701	<0.001
Number of Vehs.	3.207499	0.0733	236.6069	<0.001	127.7066	<0.001
Speed Limit	79.84101	<0.001	242.2691	<0.001	2103.031	<0.001
Bus/Truck	2.027187	0.1545	37.34104	<0.001	-	-

The p-values indicate that the lighting conditions, rural-urban, blacktop-concrete, intersection, on/off roadway, number of vehicles, and speed limit variables are associated with the severity of vehicle involvement types at the 90% confidence (α=0.10). It is also shown that DUI and shoulder existence are variables significant (α=0.10) in both cases (LV vs. PV, PV vs. PV). Bus/truck variable is only significant (α=0.10) in LV vs. PV crashes. And finally, road surface conditions and 'owner is driver' variables are only significant for PV vs.

PV crashes. The distributions of these variables by vehicle involvement type severe crash percentages are illustrated in the following figures.

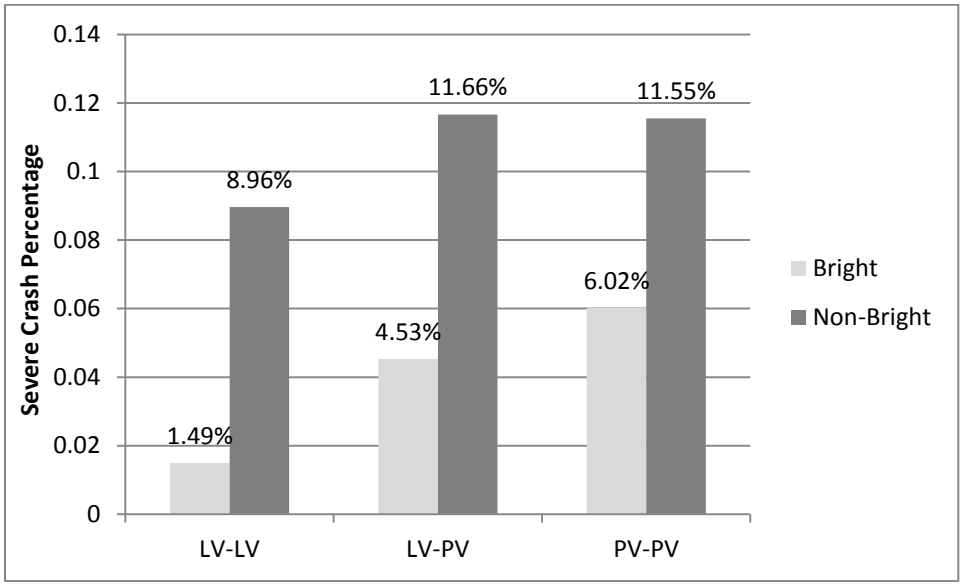

Figure 3: Distribution of Lighting Conditions by Severe Crash Percentages of Crash Groups

According to Table 14 lighting conditions are associated with the severity of vehicle involvement types at the 99% confidence (α=0.01). The non-bright conditions, severe crash percentages are higher than bright conditions as shown in Figure 3. The LV vs. PV crashes have the highest proportion of severe crashes at non-bright lighting conditions.

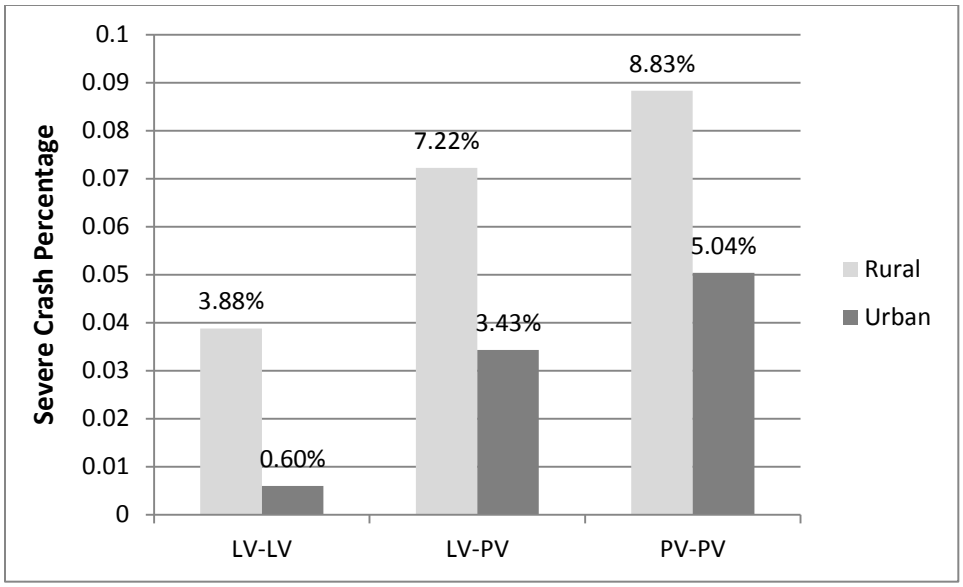

Figure 4: Distribution of Area Type by Severe Crash Percentages of Crash Groups

Table 14 shows that the area types are associated with the severity of vehicle involvement types at the 99% confidence (α=0.01). Severe crash percentages for rural areas are higher than the percentages for urban areas in three of the distributions as shown in Figure 4. The PV vs. PV crashes have the highest proportion of severe crashes at rural areas.

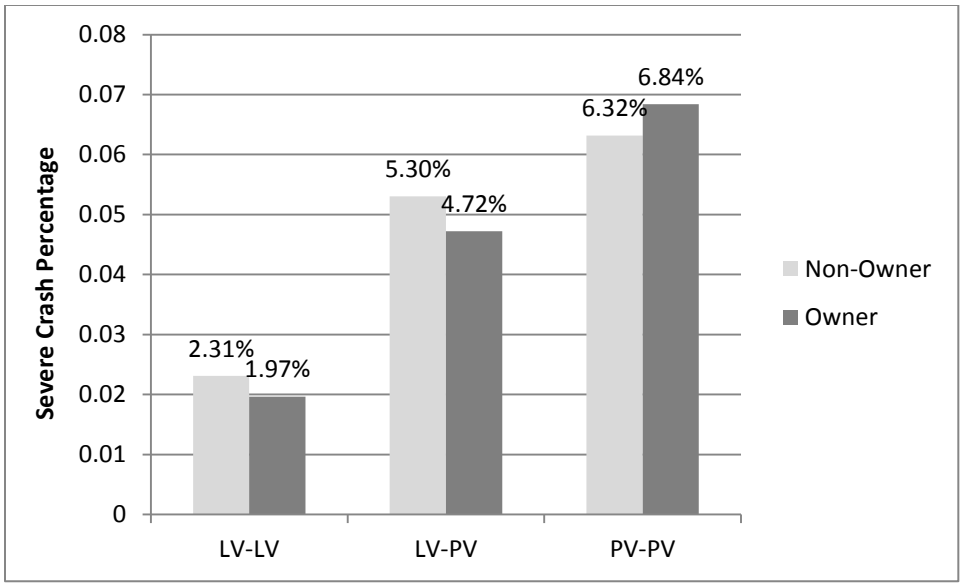

Figure 5: Distribution of 'Owner is Driver' by Severe Crash Percentages of Crash Groups

According to Table 14, the 'owner is driver' variable is associated with the severity of PV vs. PV crash group at the 99% confidence (α=0.01). The severe crash percentages of non-owner drivers in LV vs. LV, and LV vs. PV crashes are higher than the owners. The percentages of severe crashes of owners of the vehicles are higher than the non-owners of the vehicles in PV vs. PV crashes as shown in Figure 5. It is also shown that PV vs. PV crashes have the highest proportion of severe crashes in terms of the variable 'owner is driver'.

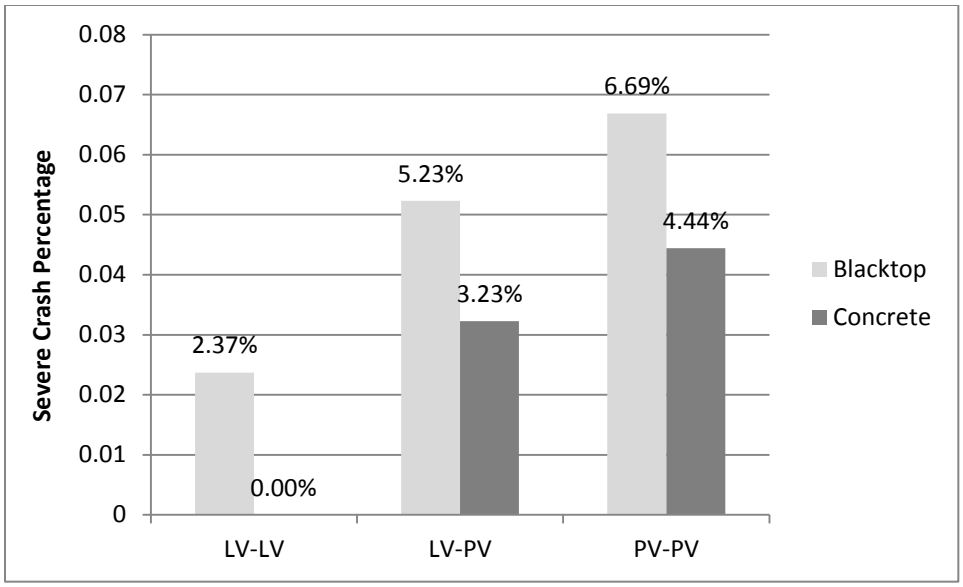

Figure 6: Distribution of Blacktop/Concrete by Severe Crash Percentages of Crash Groups

Table 14 shows that the blacktop/concrete road surface types are associated with the severity of vehicle involvement types at the 90% confidence (α=0.10). The severe crash percentage of blacktop (asphalt) surface type is higher than it is in concrete surface types as shown in Figure 6. The PV vs. PV crashes have the highest proportion of severe crashes at blacktop surface type.

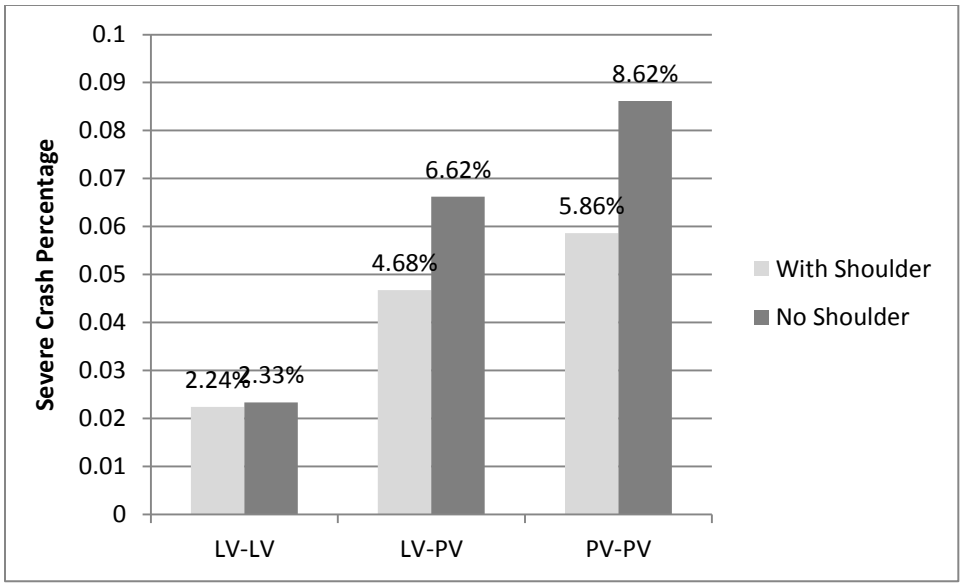

Figure 7: Distribution of Roadway Shoulder by Severe Crash Percentages of Crash Groups

According to Table 14, it is seen that the shoulder existence of the roadway is associated with the severity of LV vs. PV and PV vs. PV vehicle involvement types at the 99% confidence (α=0.01). Experiencing severe crash percentages of roadways without shoulders are higher than roadways with shoulders as shown in Figure 7. PV vs. PV crashes have the highest severe crash proportion at roadways without shoulders.

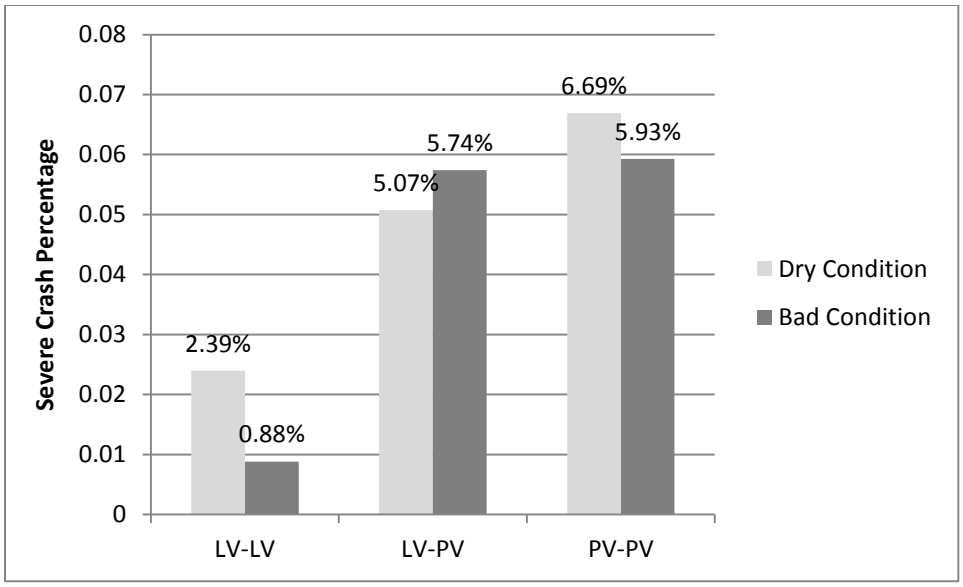

Figure 8: Distribution of Road Condition by Severe Crash Percentages of Crash Groups

Table 14 indicates that the condition of the roadway surface is associated only with the severity of PV vs. PV vehicle involvement type at the 99% confidence (α=0.01). In PV vs. PV and LV vs. LV crashes' severe percentages of dry road conditions are higher than bad road conditions while in LV vs. PV, bad condition severe crash percentages are higher as shown in Figure 8.

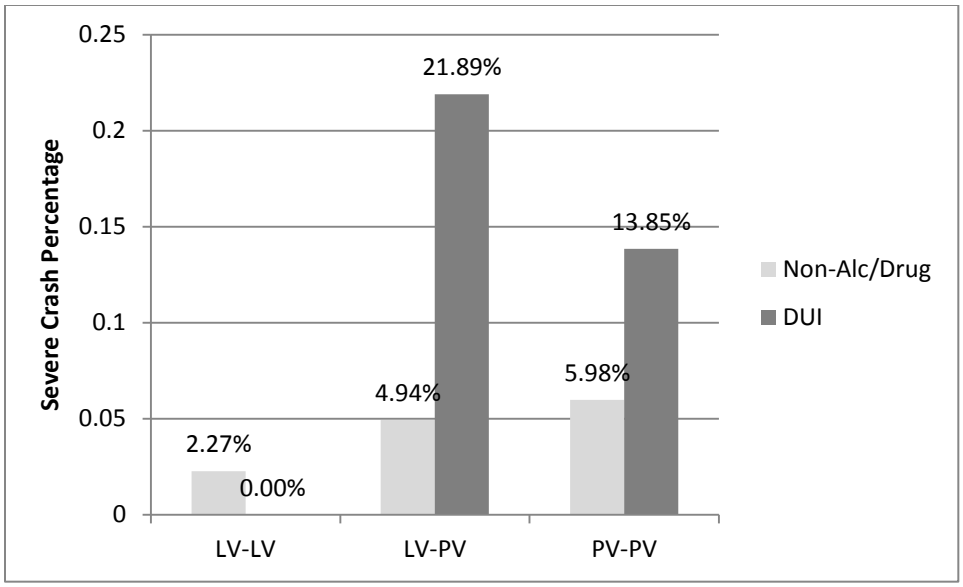

Figure 9: Distribution of Alcohol/Drug Use by Severe Crash Percentages of Crash Groups

According to Table 14, the alcohol/drug use of drivers is associated with the severity of LV vs. PV and PV vs. PV vehicle involvement types at the 99% confidence (α=0.01). Drivers which are driving under the influence of alcohol or drugs (DUI) have higher proportions of severe crashes than the ones not using alcohol or drugs while driving as shown in Figure 9. The LV vs. PV crashes have the highest proportion of severe crashes of DUI drivers.

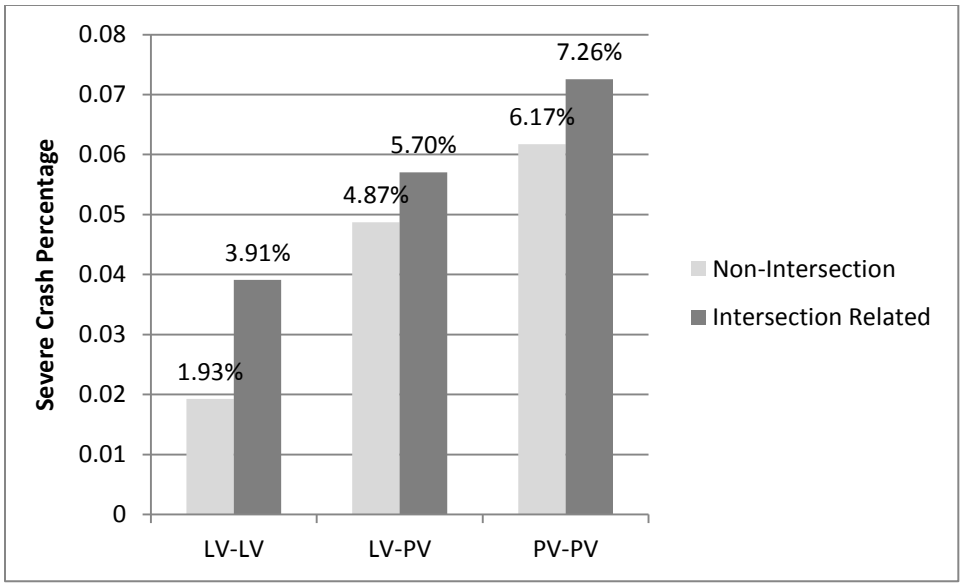

Figure 10: Distribution of Intersection Type by Severe Crash Percentages of Crash Groups

Table 14 shows that the intersections are associated only with the severity of vehicle involvement types at the 95% confidence (α=0.05). The severe crash percentages at intersection related locations are higher than non-intersection locations as shown in Figure 10. The PV vs. PV crashes have the highest proportion of severe crashes at intersections.

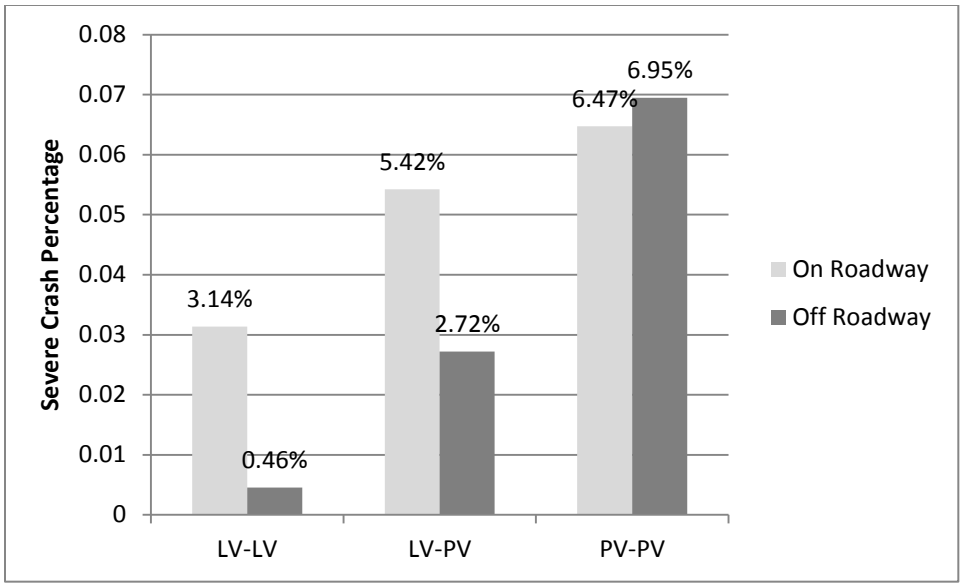

Figure 11: Distribution of On/Off Roadway by Severe Crash Percentages of Crash Groups

According to Table 14, the on/off roadway variable is associated with the severity of vehicle involvement types at the 99% confidence (α=0.01). The severe crash percentages on roadway have higher proportion than off roadways in LV vs. LV and LV vs. PV crashes while off roadways have higher severe crash percentages in PV vs. PV crashes as shown in Figure 11. In addition, the PV vs. PV crashes have the highest proportions of severe crashes at on/off roadway crashes.

.

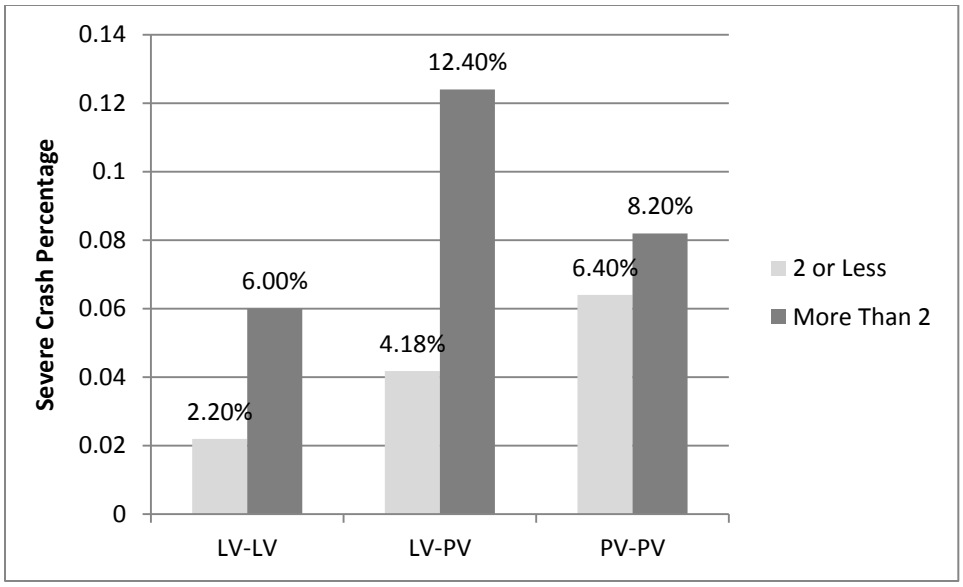

Figure 12: Distribution of Number of Vehicle Involved by Severe Crash Percentages of Crash Groups

Table 14 indicates that the number of vehicles involved is associated only with the severity of vehicle involvement types at the 90% confidence (α=0.10). Severe crash percentages of more than 2 vehicles involved crashes are higher in all crash types as shown in Figure 12. The LV vs. PV crashes have the highest proportion of severe crashes with more than 2 vehicles involved.

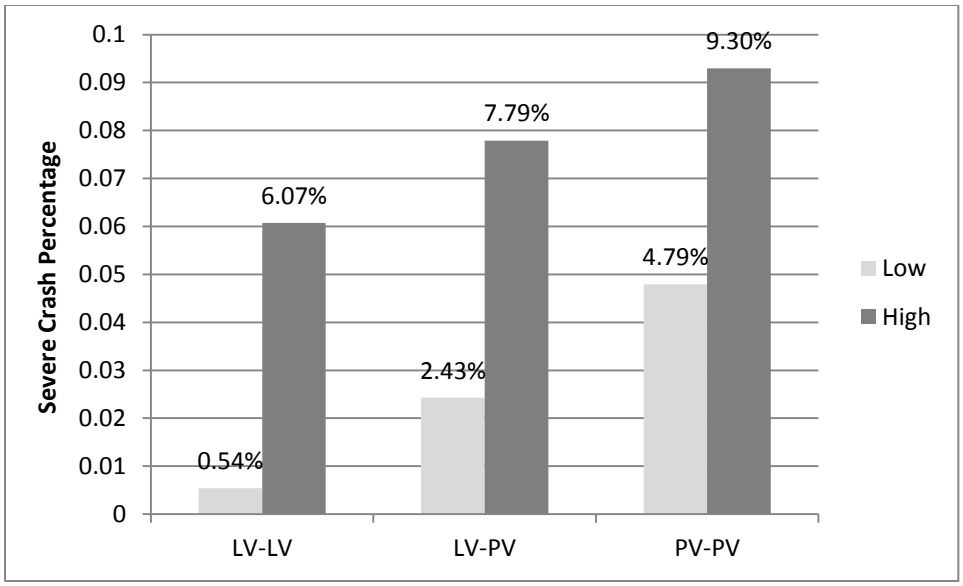

Figure 13: Distribution of Speed Limit by Severe Crash Percentages of Crash Groups

According to Table 14 the maximum speed limit is associated with the severity of vehicle involvement types at the 99% confidence (α=0.01). The severe crash percentages of roadways with higher (>=45mph) speed limits are higher than roadways with lower (<44mph) speed limits as shown in Figure 13. The PV vs. PV crashes have the highest proportion of severe crashes at high speed limits.

Descriptive analysis as well as distributions for each variable by vehicle involvement types were provided and described in this chapter. The following chapter will be dealing with the models and their results which will involve several statistical models with similar datasets used in this chapter.

CHAPTER SIX: MODELS AND RESULTS

After the exploratory analysis of the crashes involving large vehicles' injury severity provided in the preliminary analysis chapter, the modeling procedures are presented in this chapter. This chapter has been divided into two major sections. First section deals with two different types of models under dataset-A which is large vehicle involved crashes considering numerous predictor variables based on injury severity as a response variable. The second section discusses the modeling of PV vs. PV crashes (dataset-B, model-A), LV vs. PV crashes (dataset-C, model-B), and LV vs. LV crashes (dataset-D, model-C) again based on the injury severity as a binary outcome. A modeling structure has been developed with these three crash datasets in order to compare the contributing factors. SAS® and SAS Enterprise Miner® software programs have been used for the analysis conducted in this chapter. Both sections provide separate discussion for the modeling results.

6.1 Severity Analysis of Large Vehicle Involved Crashes

In this section a binary logistic regression model and a CHAID decision tree model were fitted to establish relationships between large vehicle involved crash events characteristics and injury severity. The severe crashes are defined as incapacitating and fatal crashes as it is mentioned in Chapter 3. The dataset

has 1,096 severe crashes out of 22,631 observations. Due to large difference between non-severe and severe crash frequencies the dataset is normalized by sampling. The sampling procedure uses all the observations with the rare occurrence (severe crashes), and then takes a random sample of the remaining data. A 30 percent to 70 percent proportional split was used which means that the final data have 30% and 70% severe and non-severe crashes respectively. There were 3653 observations and 1,096 severe crashes after sampling the raw data. No noteworthy differences detected in the significant variables between the models before and after the sampling procedure. First, the binary logit modeling procedure is explained and the results are discussed. Secondly, the CHAID decision tree modeling procedure is elaborated with the results. Finally, a model comparison is presented at the end in order to evaluate the two modeling procedures in terms of prediction power.

6.1.1 Binary Logistic Regression Model

In this model, severe vs. non-severe crashes were used as a binary outcome. Table 15 summarizes the model results. The p-values are shown to identify significant variables in the model. Three measures of goodness-of-fit, e.i. likelihood ratio, score and Wald Chi-square, of the model were used to show the statistical significance of the model at significance level less than 0.001. The alpha levels for each variable are also defined in Table 15 in order to understand

the confidence intervals of the probabilities for severity. Receiver Operating Characteristic (ROC) curve of the model is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. Regarding predictive power, c (the area under ROC curve) has a value of 0.754. The methodology of binary logit modeling procedure is presented in chapter 4.

The significant variables are shown in Table 15; crash groups based on vehicle types involved, residence code of the driver, roadway surface type, shoulder existence of the roadway, maximum speed limit, area type, driving under influence of alcohol or drugs, lighting conditions, owner is driver, on/off roadway crashes, intersection related crashes, number of vehicles involved, and bus or truck. The group of crashes based on the vehicles involved is a nominal variable with 5 levels which are PV vs. LV crashes (1), single LV crashes (2), bike/pedestrian/moped vs. LV crashes (3), and motorcycle vs. LV crashes (4) and the base level, LV vs. LV crashes (0).

Table 15: Binary logit model for injury severity under LV involved crashes

Goodness-of-fit tests

Test	Chi-square	Pr>ChiSq
Likelihood ratio	659.0482	<.0001
Score	635.1141	<.0001
Wald	507.1716	<.0001

Prediction power

Measure	Statistic
c (area under ROC curve)	0.754

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0.9291	0.2291	16.4471	<.0001***
LV-LV (0)	-	-	-	-
PV-LV (1)	-0.8220	0.1226	44.9852	<.0001***
Single LV (2)	-0.6922	0.1486	21.6876	<.0001***
Bike/Ped.- LV (3)	1.4609	0.3488	17.5401	<.0001***
Motorcycle- LV (4)	1.4461	0.2845	25.8443	<.0001***
Non-Resident (1)	-0.1451	0.0531	7.4749	0.0063***
Blacktop (0) -Concrete (1)	-0.2512	0.1195	4.4175	0.0356**
No Shoulder (1)	0.0787	0.0452	3.0334	0.0816*
PostedSpeed (>=45mph (1))	0.5203	0.0464	125.696	<.0001***
Rural (0)-Urban(1)	0.1714	0.0430	15.8911	<.0001***
DUI (1)	0.9578	0.1352	50.1844	<.0001***
Lighting - bright (0), non-bright (1)	0.3409	0.0573	35.4186	<.0001***
Owner is Driver (1)	-0.0888	0.0479	3.4370	0.0638*
On Roadway (0), Off Roadway (1)	-0.1442	0.0666	4.6909	0.0303**
Intersection (1)	0.1596	0.0440	13.1570	0.0003***
More Than 2 Vehicles (1)	0.5391	0.0560	92.6109	<.0001***
Bus (0), Truck (1)	0.0924	0.0531	3.0205	0.0822*

*** Significant at α=0.01, ** Significant at α=0.05, * Significant at α=0.10

"In all cases, base conditions are defined as zero"

6.1.1.1 Discussion of Results

With respect to the significant factors found in this model, LV vs. LV crashes are more likely to be severe compare to the PV vs. LV crashes. Singe large vehicle crashes have less probability, involving in a severe crash in contrast to LV vs. LV crashes while bike/pedestrian/moped vs. LV crashes, as well as the motorcycle vs. LV crashes have more probability involving in a severe crash than the base type of crash. The single LV crashes vs. LV-LV crashes result is also consistent with the study of Chang and Mannering (1999). It is also found that the residency of the driver has a negative effect on the severity, means if the driver is a resident he/she is a riskier driver in terms of severity. Moreover, road surface type variable was defined as blacktop or concrete and the model estimates that the blacktop surface type is riskier compare to concrete surface types. Roads without shoulders have a positive effect on severity. Posted speed is another significant factor which is showing the roadways with speed limits of 45 mph or higher are more risky in terms of crash severity. Severe crashes are more likely to occur in urban areas comparing to rural areas. Driving under influence of alcohol or drugs increases the risk of being involved in a severe crash. Non-bright lighting conditions such as nighttime without a streetlight, and dusk/dawn times have positive effect on the severe crashes. It is more likely to be involved in a severe crash for the driver who is the owner as well of the vehicle. On roadway crashes, intersection related crashes, and crashes involved more than

two vehicles are also more likely to be severe crashes. Large trucks are riskier in contrast to buses according to the LV involved crashes severity model results.

6.1.2 CHAID Decision Tree Model

CHAID decision tree modeling procedure has been conducted to dataset-A (LV involved crashes). The decision trees give the importance of variables, in addition to help the analyst to better interpret the results. The advantage of using trees in severity analysis is that it helps to determine the values of parameters contributing more to the severity. A series of predictor variables found significant affecting the qualitative target variable of injury severity level in an attempt to identify the important patterns of the LV involved crashes. Predictor variables were presented in Table 16. Figure 14 provides the results of the CHAID decision tree map, which has 14 terminal nodes. It shows that the variables used in this model are the primary splitters in the decision tree, implying that these variables were critical in classifying the injury severity for LV involved crashes.

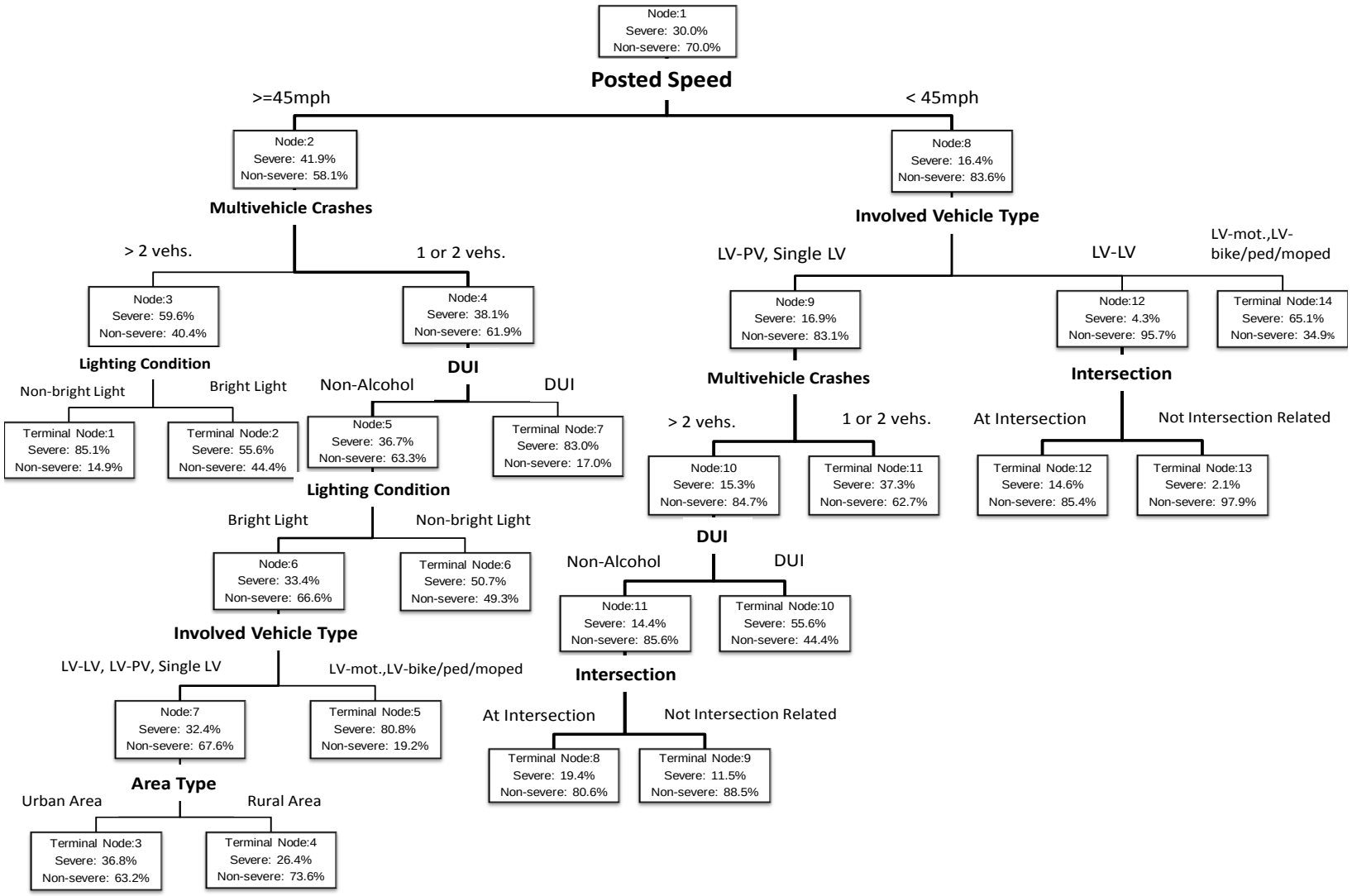

Figure 14: CHAID Decision Tree Map

6.1.2.1 Discussion of Results

The interpretation of CHAID results is straightforward. The initial split at node 1 is based on the variable of posted speed limit. This indicates that the single best variable to classify the injury severity of LV involved crashes is whether or not occurred at roadways with 45mph or more posted speed limit. CHAID directs the crashes occurred at 45mph or more speed limited to the left, forming node 2 and those crashes occurred at speed limit below 45mph to the right, forming node 8. CHAID further splits node 2 based on the multivehicle crashes variable and directs the crashes involved more than two vehicles to the left, forming node 3; one or two vehicle involved crashes to the right, forming node 4. CHAID further splits node 3 based on lighting condition variable and directs the crashes occurred in non-bright light to left, forming terminal node 1; crashes in bright light conditions to the left, forming terminal node 2. As indicated by terminal node 1, if the crash occurred at 45mph or more speed limited roadway with more than 2 vehicles with non-bright light conditions, the tree predicts the severity of injury to this crash is most likely to be severe (85.1%). At terminal node 2, the tree predicts that more than two vehicles involved crashes at high speed limits with bright light conditions the crashes are more likely to be severe (55.6%). The tree further splits node 4 to who was involved in a crash while driving under influence to right, forming terminal node 7; who is not DUI to

the left, forming node 5. Terminal node 7 is showing that crashes occurred at high speed limits with one or two vehicles are 83% more likely to be severe if the driver is DUI. CHAID splits node 5 based on lighting condition again and directs the non-alcohol or drug used drivers to bright light conditions, forming node 6; non-bright conditions to terminal node 6. CHAID predicts that terminal node 6 has 50.7% probability to be a severe crash. At node 6 the data is split based on the involved vehicle type to the crashes. Terminal node 5 is likely to be severe (80.8%). CHAID further split node 7 based on area type. Terminal node 3 which is the crash occurred at 45mph or more speed limited roadway with more than two vehicles and been used under influence of alcohol or drugs with bright light conditions at urban areas have 36.8% probability to be severe crashes while rural areas have 26.4% at terminal node 4. The prediction of injury severity likelihood can be obtained by continuing down the tree branches, with this splitting rule, until a terminal node is reached.

According to the right side of the tree (i.e., nodes 8–12 and terminal nodes 8–14) for the crashes occurred at low speed limits, 5 of the 7 terminal nodes (except for terminal nodes 10 and14) show that the injury severity is most likely to be no-injury regardless of what the contributing factors are. For example, terminal node 9 which is LV vs. PV or single LV (1 or 2 vehicle involved) crashes occurred at low speed limits have 37.3% probability of being a severe crash. It can be clearly seen that the injury severity likelihoods predicted by the crashes

occurred at higher speed limits are substantially more severe than those by the lower speed limits. This indicates that speed limit of the roadway the crash occurred is the most influential factor to severity. Table 16 is providing the predicted importance of the variables by CHAID. With respect to the importance order, the type of vehicle involved to the LV crash is following the speed limit variable. More than two vehicle involvement, DUI, lighting condition, area type, and intersection relation of the crashes are following variables respectively in the CHAID importance order.

Table 16: Variable importance predicted by CHAID

Variable Name	Importance
Posted speed limit	1.00000
Involved vehicle type	0.57590
Number of vehicles involved	0.55533
DUI	0.49121
Lighting conditions	0.42328
Rural or urban	0.23599
Intersection relation	0.20324

6.1.3 Model Comparison of Logistic Regression and CHAID Decision Tree

In this section, a comparison between the binary logistic regression model and CHAID decision tree model will be presented. Both models were conducted to the LV involved crashes dataset. The prediction powers of two models were determined by the area under the ROC curve. The sum of squared errors was also provided for each model. Figure 15 is providing the ROC curves in one

sensitivity-specificity diagram. The areas under the ROC curves (c-value) are provided in Table 17.

Table 17: Statistical Models by Area under the ROC curve (c-value)

Logistic Regression Model	0.754
CHAID Decision Tree Model	0.744

The sum of squared error for the regression model is 1266.618, while the tree models' is 1279.527.

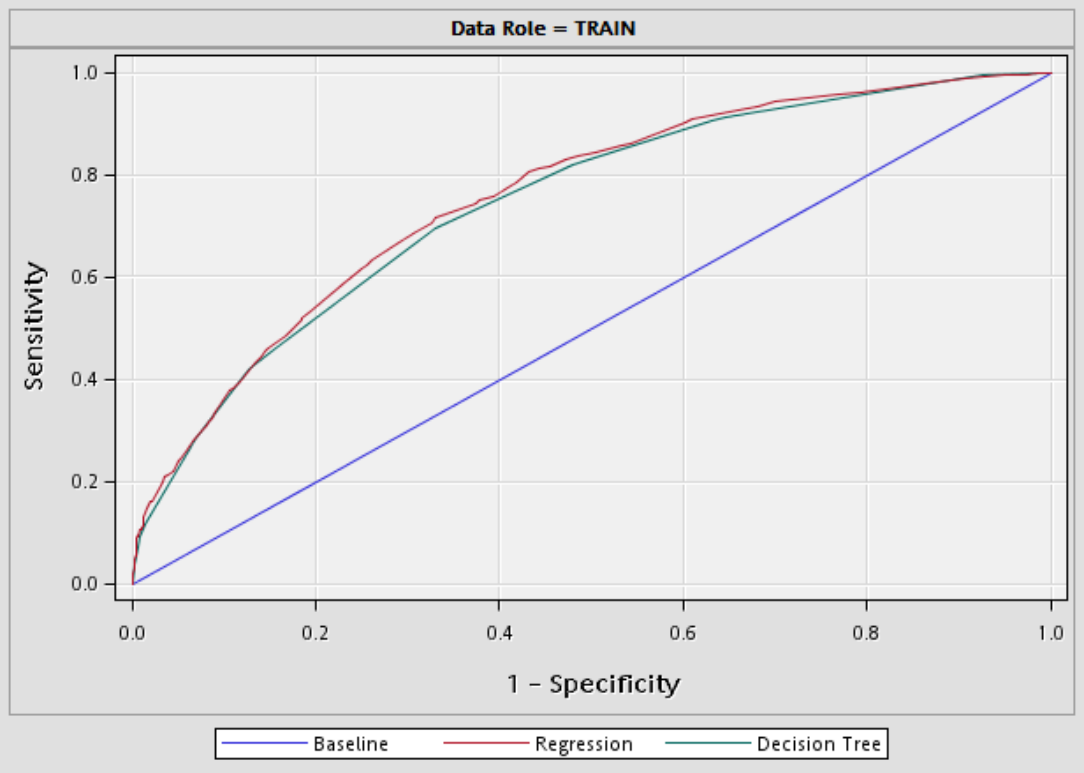

Figure 15: ROC curves of regression and tree models

As a result, both the areas under the ROC curves and squared errors of the regression model seem better in terms of prediction power compare to the

CHAID decision tree model. However, there is a difference in the number of significant factors. CHAID could have a higher prediction power with greater number of covariates. Hence, these two methodologies are comparable.

6.2. Severity Analysis of a Modeling Structure

This section has three binary logistic models based on a modeling structure. Dataset-B (PV vs. PV crashes), dataset-C (PV vs. LV crashes), and dataset-D (LV vs. LV crashes) were used to estimate three models respectively, PV vs. PV model-A), PV vs. LV (model-B), and LV vs. LV (model-C) binary logistic regression models. Data preparation of each type of crash dataset was explained in chapter 3. The PV's and LV's were grouped as followed by the "type of vehicle" variable in vehicle dataset of DHSMV crash reports.

Passenger Vehicle: Automobile, Van, Light Truck (Pick-up, 2 or 4 rear tires), Medium Truck (4 rear tires).

Large Vehicle: Large Truck (2 or more rear axles), Truck Tractor (Cab-Bobtail), Motor Home (RV), Bus (driver + seats for 9-15), Bus (driver + seats for over 15).

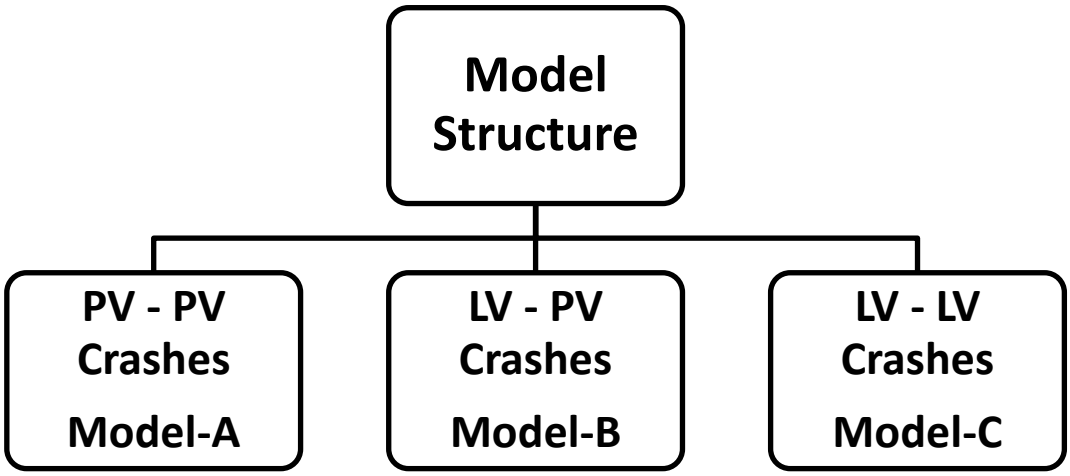

Figure 16: The Structure of Crash Types-Severity Models

The modeling structure was built in order to compare and contrast the three different crash group datasets. Regarding the results of three models, the significant variables will be elaborated to compare the differences among these crash groups and find the uniqueness for each of them in terms of injury severity at the end of this section.

6.2.1 Personal Vehicle vs. Personal Vehicle Crashes Model

In this section a binary logistic regression model (model-A) fitted to dataset-B (PV vs. PV crashes) which is the crashes only between/among passenger vehicles based on injury severity. The dataset has 17,502 severe crashes out of 265,848 observations. Due to the large difference between non-severe and severe (severe: incapacitating and/or fatal) crash frequencies the

dataset is normalized by sampling. The sampling procedure uses all the observations with the rare occurrence (severe crashes), and then takes a random sample of the remaining data. A 30 percent to 70 percent proportional split is used which means 30% of the data is severe and 70% is non-severe crashes. There were 58,340 observations and still 17,502 severe crashes after sampling the raw data. No noteworthy differences detected in significant variables between the models before and after the sampling procedure.

In the model, severe crashes vs. non-severe crashes were used as a binary outcome. Table 18 summarizes the model results. The p-values are shown to identify the significant variables in the model. Three measures of goodness-of-fit of the model, likelihood ratio, score and Wald Chi-square, show the statistical significance of the model at significance level less than 0.001. The alpha levels for each variable are also defined in Table 18 in order to understand the confidence intervals of the probabilities for severity. Regarding predictive power, c (the area under ROC curve) has a value of 0.660.

Table 18: Binary logit model for injury severity under PV vs. PV crashes

Goodness-of-fit tests

Test	Chi-square	Pr>ChiSq
Likelihood ratio	3340.2979	<.0001
Score	3362.7339	<.0001
Wald	3150.7427	<.0001

Prediction power

Measure	Statistic
c (area under ROC curve)	0.648

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	-0.3721	0.0353	111.293	<.0001***
No Shoulder (1)	0.1153	0.0105	120.688	<.0001***
Speed (>=45mph (1))	0.2984	0.00977	932.410	<.0001***
Rural (0), Urban (1)	-0.1631	0.0101	262.535	<.0001***
DUI (1)	0.3279	0.0186	311.949	<.0001***
Lighting – Bright (0), Non-bright (1)	0.2338	0.0139	283.384	<.0001***
Blacktop (0), Concrete (1)	-0.1215	0.0254	22.9241	<.0001***
Road Condition – Dry (0), Bad (0)	-0.1177	0.0137	73.5740	<.0001***
On Roadway (0), Off Roadway (1)	0.0364	0.0121	9.0577	0.0026***
Owner is Driver (1)	0.0579	0.00943	37.7314	<.0001***
Intersection (1)	0.1436	0.0101	204.129	<.0001***
More Than 2 Vehicles (1)	0.1276	0.0146	76.4227	<.0001***

*** Significant at α=0.01, ** Significant at α=0.05, * Significant at α=0.10
"In all cases, base conditions are defined as zero"

This model has eleven significant factors contributing the injury severity outcome. The variables used in this model were elaborated in Chapter 3. These variables are respectively; shoulder existence, maximum speed limit, area type,

driving under influence of alcohol or drugs, lighting conditions, roadway surface type, roadway surface condition, on/off roadway crashes, owner is driver, intersection related crashes, and number of vehicles involved. The results will be explained in the discussion of results.

6.2.2 Large Vehicle vs. Personal Vehicle Crashes Model

There is a binary logistic regression model (model-B) fitted to the dataset-C which is the crashes only between/among Large Vehicles and Passenger Vehicles based on severity. The dataset has 846 severe crashes out of 16,448 observations. The large difference between non-severe and severe (severe: incapacitating and/or fatal) crash frequencies leads to normalize the dataset by sampling. The sampling procedure uses all the observations with the rare occurrence (severe crashes), and then takes a random sample of the remaining data. A 30 percent to 70 percent proportional split is used which means 30% of the data is severe and 70% is non-severe crashes. There were 1,974 observations and still 846 severe crashes after sampling the raw data. No noteworthy differences detected in significant variables between the models before and after the sampling procedure.

In the model, severe crashes vs. non-severe crashes were used as a binary outcome. Table 19 summarizes the model results. The p-values are shown to identify the significant variables in the model. Three measures of

goodness-of-fit of the model, likelihood ratio, score and Wald Chi-square, show the statistical significance of the model at significance level less than 0.001. The alpha levels for each variable are also defined in Table 19 in order to understand the confidence intervals of the probabilities for severity. Regarding predictive power, c (the area under ROC curve) has a value of 0.733.

Table 19: Binary logit model for injury severity under LV vs. PV crashes

Goodness-of-fit tests

Test	Chi-square	Pr>ChiSq
Likelihood ratio	423.3376	<.0001
Score	410.1762	<.0001
Wald	347.5966	<.0001

Prediction power

Measure	Statistic
c (area under ROC curve)	0.733

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	0.2907	0.1708	2.8946	0.0889*
No Shoulder (1)	0.0896	0.0504	3.1554	0.0757*
Speed (>=45mph (1))	0.4887	0.0515	89.9709	<.0001***
Rural (0), Urban (1)	-0.2465	0.0477	26.7138	<.0001***
DUI (1)	0.8094	0.1373	34.7550	<.0001***
Lighting – Bright (0), Non-bright (1)	0.3278	0.0647	25.6644	<.0001***
Owner is Driver (1)	-0.1055	0.0509	4.3003	0.0381**
On Roadway (0), Off Roadway (1)	-0.1643	0.0903	3.3065	0.0690*
Intersection (1)	0.2328	0.0486	22.9724	<.0001***
More Than 2 Vehicles (1)	0.5778	0.0563	105.207	<.0001***

*** Significant at α=0.01, ** Significant at α=0.05, * Significant at α=0.10
"In all cases, base conditions are defined as zero"

This model has nine significant factors contributing to the injury severity outcome. The variables used in this model were elaborated in Chapter 3. These variables are respectively; shoulder existence, maximum speed limit, area type, driving under influence of alcohol or drugs, lighting conditions, owner is driver, on/off roadway crashes, intersection related crashes, and number of vehicles involved. A detailed explanation of the results will be provided in the discussion of results.

6.2.3 Large Vehicle vs. Large Vehicle Crashes Model

A binary logistic regression model (model-C) fitted to the dataset-D which is the crashes only between/among Large Vehicles based on injury severity. The dataset has 61 severe crashes out of 2,692 observations. Due to the large difference between non-severe and severe (severe: incapacitating and/or fatal) crash frequencies the dataset is normalized by sampling. The sampling procedure uses all the observations with the rare occurrence (severe crashes), and then takes a random sample of the remaining data. A 30 percent to 70 percent proportional split is used which means 30% of the data is severe and 70% is non-severe crashes. There are 203 observations and still 61 severe crashes after sampling the raw data. No noteworthy differences detected in significant variables between the models before and after the sampling procedure.

In the model, severe crashes vs. non-severe crashes were used as a binary outcome. Table 20 summarizes the model results. The p-values are shown to identify the significant variables in the model. Three measures of goodness-of-fit of the model, likelihood ratio, score and Wald Chi-square, show the statistical significance of the model at significance level less than 0.001. Regarding predictive power, c (the area under ROC curve) has a value of 0.866.

Table 20: Binary logit model for injury severity under LV vs. LV crashes

Goodness-of-fit tests

Test	Chi-square	Pr>ChiSq
Likelihood ratio	85.0683	<.0001
Score	74.3417	<.0001
Wald	50.0661	<.0001

Prediction power

Measure	Statistic
c (area under ROC curve)	0.866

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	-0.6567	0.3021	4.7243	0.0297**
Speed (>=45mph (1))	1.1241	0.2297	23.9543	<.0001***
Rural (0), Urban (1)	-0.8335	0.2474	11.3548	0.0008***
Lighting – Bright (0), Non-bright (1)	0.5084	0.2369	4.6047	0.0319**
Intersection (1)	0.8241	0.2508	10.7945	0.0010***

*** Significant at α=0.01, ** Significant at α=0.05, * Significant at α=0.10
"In all cases, base conditions are defined as zero"

There are four significant factors contributing to the injury severity outcome. The variables used in this model were elaborated in Chapter 3. These

variables are respectively; maximum speed limit, area type, lighting conditions, and intersection related crashes. The results will be explained in detail in the discussion of results.

6.2.4 Discussion of Results

The results of three models in the modeling structure will be discussed in this section.

With respect to the significant factors in model-A; roadways without shoulders, blacktop road surface type compare to concrete, and dry road surface conditions are more likely to have severe instead of non-severe (No injury, Possible Injury, Non-incapacitating evident injury) crashes. Posted speed limit (1, >=45mph; 0, <44mph) is the most significant factor in model-A in terms of coefficients, which has a positive effect on the crash injury severity. Rural areas are more likely to experience more severe crashes than urban areas. Driving under influence of alcohol or drugs is also found to increase the injury severity of PV vs. PV crashes. The crashes occurred in non-bright lighting conditions (dark without street light, dusk, and dawn) have positive effect on injury severity. Moreover, off roadway crashes, intersection related crashes and more than two vehicles involved crashes were found to have positive affect the injury severity of PV vs. PV crashes. Last but not least, the ‘owner is driver’ is a significant factor which can be concluded as owner of the vehicle is more likely to be involved in a

severe crash. This could be explained as the large vehicles are mostly commercial vehicles. So, the drivers are most likely not to be the owners of the vehicles.

Regarding to the results of model-B; shoulder existence of the roadway, posted speed limit, rural vs. urban, driving under influence of alcohol or drugs, lighting conditions, intersection relation, and more than two vehicles involvement variables can be concluded in the same way with the model-A results, mentioned above. Nevertheless, there are two factors with opposite signs which mean they don't have the same affect. First, on roadway crashes instead of off roadway crashes are more likely to be severe for LV vs. PV crashes. And second is the owner is driver variable which is concluded as the non-owner of vehicles has a higher probability to be involved in a severe crash in model-B.

Model-C is the LV vs. LV crash type model and the results of this model indicates four significant factors contributing to the crash injury severity binary outcome. These factors were; posted speed limit, rural vs. urban areas, lighting conditions, and the intersection relation of the crash. The effects of these variables can be explained in the exact same way with the model-A and model-B results.

Although, all the variables were used in all three of the models, the significant factors for each model have dissimilarities. The differences among three models in terms of significant factors and their effect on the models are summarized in Table 21. As it is seen in the table, posted speed limit, lighting

condition, and intersection variables are affecting the injury severity positively and rural areas are more likely to have severe crashes in all three models. DUI, more than two vehicles, and shoulder have positive effect on crash injury severity outcome in model-A and model-B. Off roadway crashes are more likely to be severe in model-A while on roadway crashes are riskier in model-B. Owner is driver factor has significant positive effect on model-A, and a significant negative effect on model-B. Blacktop-concrete and road surface condition variables are only significant with a negative effect on the injury severity binary outcome of model-A. To sum up, it is distinguished that LV vs. LV crashes have the smallest number of contributing factors to the crash injury severity while PV vs. PV crashes have the largest number of predictor variables.

Table 21: Variable descriptions and their effects on the models

Variable Description	Model-A (PV-PV)	Model-B (LV-PV)	Model-C (LV-LV)
Posted Speed Limit: 1, >=45mph; 0, <44mph	+	+	+
Lighting Condition: 1, Bright lighting; 0, non-bright lighting	+	+	+
Intersection: 1, Intersection related; 0, not intersection related	+	+	+
Rural-Urban: 1, Urban area; 0, rural area	-	-	-
DUI: 1, DUI; 0, non-alcohol/drug use	+	+	
Number of Vehicles: 1, more than 2 vehicles; 0, 2 or less vehicles	+	+	
Shoulder: 1, No shoulder; 0, with shoulder	+	+	
On/Off Roadway: 1, Off roadway; 0, on roadway	+	-	
Owner is Driver: 1, Owner is driver; 0, non-owner	+	-	
Blacktop-Concrete: 1, Blacktop; 0, concrete	-		
Road Surface Condition: 1, Bad road condition; 0, dry	-		

In this chapter five different models and their results were discussed and presented as well as comparisons between/among some of them. The overall summary and conclusion of the thesis will be given in the next chapter.

CHAPTER SEVEN: CONCLUSIONS

The main objectives of this study were to investigate the characteristics of large vehicle crashes in order to identify the contributing factors to injury severity levels. Severe is considered as incapacitating and fatal. Large vehicles are consider as: heavy trucks, truck-tractors, RVs, buses with 9-15 seats, and buses with over 15 seats.

To achieve this purpose, three different statistical approaches were proposed. First the descriptive statistics, second is the binary logistic regression modeling, and third is the CHAID decision tree modeling.

Descriptive statistics were examined to get the distribution of severe crashes / fatal crashes for LV-PV (LV vs. PV crashes) and PV-PV groups through various factors which were addressed by researchers. In this part, crash severity level, environmental conditions, large vehicle involvement, passenger vehicle involvement, motorcycle involvement, bike / pedestrian involvement, and driver characteristics (i.e. DUI, residence etc.) were discussed for both crash groups.

The main results are:

(1) Non-LV involved crashes are more likely to have incapacitating injuries than LV involved ones; however, the fatality rate is significantly high in LV involved crashes.

(2) There are several factors (i.e. lighting conditions, DUI) influencing the injury severity of PV vs. PV, LV vs. PV, and LV vs. LV crashes. The bad lighting

conditions, high speed limits, no-shoulder roadways, driving under the influence of alcohol or drugs, intersections, blacktop road surface, rural areas, and multiple vehicle pile-ups prove to have positive affect on the injury severity in all three crash groups.

Analyzing crash severity by type of vehicle is considered crucial criteria not only because it reflects the importance and danger of large vehicle crashes but also because it reveals differences between a large vehicle crash and smaller vehicle crash. Crash severity is affected by various factors including driver characteristics, vehicle characteristics, environmental factors, and roadway features.

Fully understanding the impacts that these factors worsen the crash severity is beneficial for selecting proper countermeasures to reduce the crash severity of large vehicle crashes. Furthermore, this insight can help identify solutions for decreasing the severity and fatality rates of crashes.

A logistic regression binary output was used to estimate the crash severity models for large vehicle involved crashes. According to the results of crash injury severity modeling and the analysis of LV involved crashes, some conclusions can be given:

(1) Residence of the driver, owner is driver (Zhu and Srinivasan in 2011 supports this result), number of vehicles involved, lighting condition, alcohol/drug use of drivers, roadway section with/without shoulder, rural or urban area,

blacktop/concrete road surface type, on/off roadway, intersection related/not related site location, posted speed limit, whether a bus or truck was involved, and different vehicle types appear as the main influence to large vehicle crash severity. Findings of Lemp et al. (2011) strengthens the results in this model.

(2) The factors of resident drivers, non-owner drivers, more than two vehicle crashes, non-bright light condition, DUI drivers, roads without shoulder, urban, blacktop surface type, on roadway crashes, intersection related locations, higher speed limit, and truck involved crashes are more likely to reduce the severity of LV involved crashes. The crash type variable findings indicate that LV vs. LV crashes were more likely to be severe when compared to LV vs. PV and single LV crashes. Furthermore, the LV vs. motorcycle and LV vs. bike/ped/moped crashes have more probability to be severe crashes.

(3) Non-owner drivers could induce LV crash severity. The reason may be that most drivers of LV's are not owners of the truck or buses, because those vehicles are more likely to be commercial vehicles.

(4) Drivers who are Florida residents are more likely to be involved in severe crashes. This finding could be explained as the unfamiliar drivers with the roadways drive more careful.

(5) Based on the magnitudes of the variable coefficients, the variables of maximum speed limit, number of vehicles involved, and the type of crash all have a major impact on the crash severity level. Thus proving the restriction to driving

speed as a principle factor for the safety of LV's and vehicles involved in a crash with LV's.

Furthermore, a CHAID decision tree model is also conducted to the LV involved crashes dataset. According to the results of CHAID:

(1) There are seven variables which came out to be significant. The importance of the variables for severity is respectively: posted speed limit, Involved vehicle type, Number of vehicles involved, DUI, Lighting conditions, Rural or urban, and Intersection relation of the crashes. Chang and Chien (2013) also found similar factors affecting the large truck crash severity with non-parametric models.

(2) The decision tree indicates 14 terminal nodes of different crash scenarios based on the contributing factors, with their probabilities to be severe crashes.

A comparison of the two models mentioned above has also been provided. The comparison results indicated that the regression and CHAID decision tree models are comparable.

A modeling structure is also built in order to analyze the PV (personal vehicle) vs. LV (large vehicle) crashes, LV vs. LV crashes, and PV vs. PV crashes. The main benefit of this modeling structure is its ability to show three different small and large vehicle crash combinations at the same time, and compare the results of them. Binary logit modeling procedure has been used for those three models. The main results of the modeling structure are:

(1) Higher speed limits, non-bright lighting conditions, rural areas, and intersection related factors are reducing the likelihood of severity in LV vs. LV crashes. Findings in this model are also consistent with Khorashadi et al. (2005)'s study.

(2) In addition to the contributing factors in model-A, LV vs. PV crashes severity is positively affected by DUI drivers, more than two vehicles involvement, and no shoulder factors. Owner is driver and, on/off roadway variables have negative effect in this model.

(3) The PV vs. PV crashes crash severity is influenced by two more factors compared to the model-B. These factors are blacktop surface of roadway and bad road surface conditions. The owner is driver and on/off roadway variables have opposite effects on severity in contrast model-B.

Based on these statistical analyses for large vehicle involved crashes, several countermeasures can be suggested:

(1) The maximum speed limits for large vehicles should be reduced in order to control the severe crashes occurring due to high speed limits. Speed limit signs could also be adjusted. Some dynamic signs such as changeable message signs with radar and speed feedback signs could be effective to reduce driver speed.

(2) Lighting conditions should be improved. Streetlights at all types of roadways should also be revised and be opened even in sunrise and sunset times.

(3) Intersections are also important site locations in terms of crash severity. So, intersection safety improvements are also needed to reduce the LV involved severe crashes in particular.

The limitation in this study was the use of one year data from the state of Florida. However, the crash data from the state of Florida may not represent the entire nations' crash characteristics. Thus, it is recommended that in the future studies, several years of crash data from different regions be used.

This study analyzed the crash injury severity considering the dimensions of vehicles. The importance of vehicle sizes should be further studied to include different crash scenarios such as; different type of vehicles involvement, crash types, and more site locations. Furthermore, interactions among various variables such as gender and ages of the drivers could be used.

APPENDIX: MODELS BEFORE SAMPLING THE DATASETS

Table 22: Binary logit model for injury severity under LV involved crashes (raw data)

Goodness-of-fit tests

Test	Chi-square	Pr>ChiSq
Likelihood ratio	956.0431	<.0001
Score	1222.0246	<.0001
Wald	906.8366	<.0001

Prediction power

Measure	Statistic
c (area under ROC curve)	0.756

Total Frequency

Non-severe	21,535
Severe	1,096

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	-0.3355	0.1606	69.1126	<.0001
PV-LV	-0.7839	0.0812	93.2043	<.0001
Single LV	-0.6834	0.1076	40.3577	<.0001
Bike/Ped.- LV	1.3531	0.2229	36.8374	<.0001
Motorcycle- LV	1.3337	0.1653	65.0635	<.0001
Non-Resident	-0.1513	0.0434	12.1562	0.0005
Blacktop-Concrete	-0.2577	0.1014	6.4525	0.0111
No Shoulder	0.0679	0.0360	3.5560	0.0593
Speed (>=45mph)	0.5119	0.0392	170.6453	<.0001
Rural-Urban	0.1811	0.0346	27.3392	<.0001
DUI	0.8878	0.0796	124.3886	<.0001
Lighting	0.3157	0.0424	55.4748	<.0001
Owner is Driver	-0.1204	0.0386	9.7378	0.0018
On/Off Roadway	-0.1129	0.0558	4.0928	0.0431
Intersection	0.1815	0.0351	26.8196	<.0001
More Than 2 Vehicles	0.5192	0.0409	161.3840	<.0001
Bus/Truck	0.0902	0.0443	4.1464	0.0417

Table 23: Binary logit model for injury severity under PV vs. PV crashes (raw data)

Goodness-of-fit tests

Test	Chi-square	Pr>ChiSq
Likelihood ratio	4535.5801	<.0001
Score	4931.2757	<.0001
Wald	4652.0633	<.0001

Prediction power

Measure	Statistic
c (area under ROC curve)	0.646

Total Frequency

Non-severe	248,346
Severe	17,502

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	-2.1838	0.0297	5395.1861	<.0001
No Shoulder	0.1045	0.00877	142.1448	<.0001
Speed (>44mph)	0.2948	0.00836	1244.5427	<.0001
Rural-Urban	-0.1689	0.00857	388.3335	<.0001
DUI	0.3252	0.0144	511.7329	<.0001
Lighting	0.2329	0.0111	443.1320	<.0001
Blacktop-Concrete	-0.1254	0.0226	30.7665	<.0001
Road Condition	-0.1103	0.0118	87.3431	<.0001
On/Off Roadway	0.0430	0.0102	17.7558	<.0001
Owner is Driver	0.0455	0.00799	32.4539	<.0001
Intersection	0.1212	0.00844	206.3274	<.0001
More Than 2 Vehicles	0.1278	0.0122	110.2879	<.0001

Table 24: Binary logit model for injury severity under LV vs. PV crashes (raw data)

Goodness-of-fit tests

Test	Chi-square	Pr>ChiSq
Likelihood ratio	597.0322	<.0001
Score	693.4086	<.0001
Wald	572.9103	<.0001

Prediction power

Measure	Statistic
c (area under ROC curve)	0.732

Total Frequency

Non-severe	15,602
Severe	846

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	-1.7478	0.1253	194.5339	<.0001
No Shoulder	0.1194	0.0406	8.6625	0.0032
Speed (>44mph)	0.4774	0.0445	115.1814	<.0001
Rural-Urban	-0.2129	0.0396	28.9300	<.0001
DUI	0.8473	0.0925	83.9173	<.0001
Lighting	0.3496	0.0485	52.0027	<.0001
Owner is Driver	-0.0960	0.0417	5.3040	0.0213
On/Off Roadway	-0.1774	0.0793	5.0044	0.0253
Intersection	0.1983	0.0389	25.9769	<.0001
More Than 2 Vehicles	0.5322	0.0418	162.3106	<.0001

Table 25: Binary logit model for injury severity under LV vs. LV crashes (raw data)

Goodness-of-fit tests

Test	Chi-square	Pr>ChiSq
Likelihood ratio	584.641	<.0001
Score	136.1933	<.0001
Wald	85.5409	<.0001

Prediction power

Measure	Statistic
c (area under ROC curve)	0.860

Total Frequency

Non-severe	2,631
Severe	61

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	-3.6855	0.2363	243.2405	<.0001
Speed (>44mph)	0.9896	0.1796	30.3636	<.0001
Rural-Urban	-0.6835	0.1969	12.0552	0.0005
Lighting	0.5846	0.1427	16.7840	<.0001
Intersection	0.3979	0.1505	6.9890	0.0082

LIST OF REFERENCES

Abdel-Aty, MA, & Abdelwahab, H. (2004). Predicting injury severity levels in traffic crashes: a modeling comparison. Journal of transportation …, (April), 204–210. Retrieved from http://ascelibrary.org/doi/abs/10.1061/(ASCE)0733-947X(2004)130:2(204)

Abdel-Aty, Mohamed. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research, 34(5), 597–603. doi:10.1016/j.jsr.2003.05.009

Abdel-Aty, Mohamed, & Abdelwahab, H. (2004). Modeling rear-end collisions including the role of driver's visibility and light truck vehicles using a nested logit structure. Accident; analysis and prevention, 36(3), 447–56. doi:10.1016/S0001-4575(03)00040-X

Abdel-Aty, Mohamed, & Keller, J. (2005). Exploring the overall and specific crash severity levels at signalized intersections. Accident; analysis and prevention, 37(3), 417–25. doi:10.1016/j.aap.2004.11.002

Abdel-Aty, M a, & Radwan, a E. (2000). Modeling traffic accident occurrence and involvement. Accident; analysis and prevention, 32(5), 633–42. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10908135

Abdelwahab, H. T., & Abdel-Aty, M. a. (2002). Investigating Driver Injury Severity in Traffic Accidents Using Fuzzy ARTMAP. Computer-Aided Civil and Infrastructure Engineering, 17(6), 396–408. doi:10.1111/1467-8667.00286

Al-ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. Accident analysis and prevention, 34, 729–741.

Cantor, D. E., Corsi, T. M., Grimm, C. M., & Özpolat, K. (2010). A driver focused truck crash prediction model. Transportation Research Part E: Logistics and Transportation Review, 46(5), 683–692. doi:10.1016/j.tre.2009.08.011

Chang, L. Y., & Mannering, F. (1999). Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. Accident; analysis and prevention, 31(5), 579–92. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10440555

Chang, L.-Y., & Chien, J.-T. (2013). Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Safety Science, 51(1), 17–22. doi:10.1016/j.ssci.2012.06.017

Chang, L.-Y., & Wang, H.-W. (2006). Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident; analysis and prevention, 38(5), 1019–27. doi:10.1016/j.aap.2006.04.009

Chen, F., & Chen, S. (2011). Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accident analysis and prevention, 43(5), 1677–1688

Chen, L. wan. (1997). Applying Categorical Data Analysis to Multi-way Contingency Table-Location , Accident Type , and Related Factors With Severity. University of North Carolina Forrest M. Council, Highway Safety Research Center

Das, A., & Abdel-Aty, M. (2010). A genetic programming approach to explore the crash severity on multi-lane roads. Accident; analysis and prevention, 42(2), 548–57. doi:10.1016/j.aap.2009.09.021

Das, A., Abdel-Aty, M., & Pande, A. (2009). Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. Journal of safety research, 40(4), 317–27. doi:10.1016/j.jsr.2009.05.003

Das, A., & Abdel-Aty, M. a. (2011). A combined frequency–severity approach for the analysis of rear-end crashes on urban arterials. Safety Science, 49(8-9), 1156–1163. doi:10.1016/j.ssci.2011.03.007

Das, A., Pande, A., Abdel-Aty, M., & Santos, J. (2008). Urban arterial crash characteristics related with proximity to intersections and injury severity. Transportation Research Record, (November 2007), 1–14. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Urban+arterial+crash+characteristics+related+with+proximity+to+intersections+and+injury+severity#0

Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident; analysis and prevention, 38(3), 434–44. doi:10.1016/j.aap.2005.06.024

Desapriya, E., Pike, I., & Raina, P. (2006). Severity of alcohol-related motor vehicle crashes in British Columbia: case - control study. International journal of injury control and safety promotion, 13(2), 89–94. doi:10.1080/17457300500172685

Greene, WH. (2003) Econometric analysis, Prentice Hall, New Jersey, 5th Edition

Haleem, K., & Abdel-Aty, M. (2010). Examining traffic crash injury severity at unsignalized intersections. Journal of safety research, 41(4), 347–57. doi:10.1016/j.jsr.2010.04.006

Hosmer, DG., Lemeshow, S. (2005) Applied logistic regression. Wiley, New York, 2nd Edition

Islam, M., & Hernandez, S. (2011). An empirical analysis of fatality rates for large truck involved crashes on interstate highways. onlinepubs.trb.org, 1–19. Retrieved from http://onlinepubs.trb.org/onlinepubs/conferences/2011/RSS/2/Islam,M.pdf

Khattak, A., Schneider, R., & Targa, F. (2003). Risk factors in large truck rollovers and injury severity: analysis of single-vehicle collisions. Transportation Research Board …, (January). Retrieved from http://www.ltrc.lsu.edu/TRB_82/TRB2003-000331.pdf

Khorashadi, A., Niemeier, D., Shankar, V., & Mannering, F. (2005). Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accident; analysis and prevention, 37(5), 910–21. doi:10.1016/j.aap.2005.04.009

Kieliszewski, C. (2006). Twisted metal| An investigation into observable factors that lead to critical traffic events. Retrieved from http://gradworks.umi.com/31/97/3197975.html

Kockelman, K. M., & Kweon, Y.-J. (2002). Driver injury severity: an application of ordered probit models. Accident; analysis and prevention, 34(3), 313–21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11939360

Kuhnert, P. M., Do, K.-A., & McClure, R. (2000). Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis, 34(3), 371–386. doi:10.1016/S0167-9473(99)00099-7

Lee, C., & Abdel-Aty, M. (2005). Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida. Accident; analysis and prevention, 37(4), 775–86. doi:10.1016/j.aap.2005.03.019

Lee, C., & Abdel-Aty, M. (2008). Presence of passengers: does it increase or reduce driver's crash potential? Accident; analysis and prevention, 40(5), 1703–12. doi:10.1016/j.aap.2008.06.006

Lemp, J. D., Kockelman, K. M., & Unnikrishnan, A. (2011). Analysis of large truck crash severity using heteroskedastic ordered probit models. Accident; analysis and prevention, 43(1), 370–80. doi:10.1016/j.aap.2010.09.006

Lyman, S., & Braver, E. R. (2003). Occupant deaths in large truck crashes in the United States: 25 years of experience. Accident; analysis and prevention, 35(5), 731–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12850074

Morrow, R. (n.d.). Fog and Smoke Related Crashes in Florida : Identifying Crash Characteristics , Spatial Distribution and Injury Severity.

Nassiri, H., & Edrissi, A. (2006). Modeling Truck Accident Severity on Two-Lane Rural Highways. Scientia Iranica, 13(2), 193–200. Retrieved from http://64.130.220.45/En/VEWSSID/J_pdf/95520060208.pdf

National Technical Information Service, Springfield, V. 22161. (2006). Report to Congress on the Large Truck Crash Causation Study. U.S. Department of Transportation, Federal Motor Carrier Safety Admisintration, (March).

Nevarez, P., Abdel-Aty, M., & Wang, X. (2009). Large-Scale Injury Severity Analysis for Arterial Roads: Modeling Scheme and Contributing Factors. … Research Board 88th …, (July 2008). Retrieved from http://trid.trb.org/view.aspx?id=880629

Pande, A., & Abdel-Aty, M. (2009). A novel approach for analyzing severe crash patterns on multilane highways. Accident; analysis and prevention, 41(5), 985–94. doi:10.1016/j.aap.2009.06.003

Ritschard, G. (2010). CHAID and earlier supervised tree methods. Retrieved from http://www.unige.ch/ses/metri/cahiers/2010_02.pdf

Sawalha, Z., & Sayed, T. (2006). Traffic accident modeling: some statistical issues. Canadian Journal of Civil Engineering, 1124(January), 1115–1124. doi:10.1139/L06-056

Shankar, V, Mannering, F., & Barfield, W. (1996). Statistical analysis of accident severity on rural freeways. Accident Analysis & Prevention, 28(3), 391–401. Retrieved from http://www.sciencedirect.com/science/article/pii/0001457596000097

Shankar, Venkataraman, & Mannering, F. (1996). An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. Journal of Safety Research, 27(3), 183–194. Retrieved from http://www.sciencedirect.com/science/article/pii/0022437596000102

Stein, H. S., & Jones, I. S. (1988). Crash involvement of large trucks by configuration: a case-control study. American journal of public health, 78(5), 491–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1349325&tool=pmcentrez&rendertype=abstract

Sze, N. N., & Wong, S. C. (2007). Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accident; analysis and prevention, 39(6), 1267–78. doi:10.1016/j.aap.2007.03.017

Theofilatos, A., Graham, D., & Yannis, G. (2012). Factors affecting accident severity inside and outside urban areas in Greece. Traffic injury prevention, 13(5), 458–67. doi:10.1080/15389588.2012.661110

Wang, X., & Abdel-Aty, M. (2008). Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. Accident; analysis and prevention, 40(5), 1674–82. doi:10.1016/j.aap.2008.06.001

Wang, Z. (2008). Modeling crash severity and speed profile at roadway work zones. University of South Florida Scholar Commons. Retrieved from http://scholarcommons.usf.edu/etd/555/

Yamamoto, T., Hashiji, J., & Shankar, V. N. (2008). Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident; analysis and prevention, 40(4), 1320–9. doi:10.1016/j.aap.2007.10.016

Yamamoto, T., & Shankar, V. N. (2004). Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. Accident; analysis and prevention, 36(5), 869–76. doi:10.1016/j.aap.2003.09.002

Zhu, X., & Srinivasan, S. (2011a). Modeling occupant-level injury severity: An application to large-truck crashes. Accident; analysis and prevention, 43(4), 1427–37. doi:10.1016/j.aap.2011.02.021

Zhu, X., & Srinivasan, S. (2011b). A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accident; analysis and prevention, 43(1), 49–57. doi:10.1016/j.aap.2010.07.007