ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF CIVIL AND ENVIRONMENTAL ENGINEERING

Assessing the Predictive Abilities of Statistical Injury-Severity Prediction Modelling considering non-behavioral factors of accident

A Thesis in Road and Transportation Engineering

By Mulaw Amdu Belay

May 19, 2018

Addis Ababa

A Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science

The undersigned have examined the thesis entitled '**Assessing the Predictive Abilities of Statistical Injury-Severity Prediction Modelling considering non-behavioral factors of accident**' presented by **MULAW AMDU BELAY**, a candidate for the degree of **Master of Science** and hereby certify that it is worthy of acceptance.

Dr. Md Mahabubul Bari		12/05/2018
Advisor	Signature	Date
Dr. Getu Segni Tulu		
Internal Examiner	Signature	Date
Mr. Anteneh Afwork		
External Examiner	Signature	Date
Chairperson	Signature	Date

UNDERTAKING

I certify that research work titled “Assessing the Predictive Abilities of Statistical Injury-Severity Prediction Modelling considering non-behavioral factors of accident” is my own work. The work has not been presented elsewhere for assessment. Where material has been used from other sources it has been properly acknowledged/referred.

Mulaw Amdu Belay

ABSTRACT

In the developing country, Road Traffic Accidents are among the leading cause of death and injury; Ethiopia in particular experiences the highest rate of such accidents. Road Traffic Accidents cannot be absolutely eradicated, however, it is possible to prevent them to some extent as long as the contributing factors are identified and tackled appropriately. The driver behavior plays a crucial role in the occurrence of a crash; but, it is usually complex and unpredictable, and also focusing too much on the driver as the cause of a crash often masked the ability to see other causes that could reduce crash rates and crash severity. So it is important to figure out the role of the non-behavioral factors in traffic accidents, based on which cost-effective countermeasures can be recommended to reduce the chance of accidents. Previously, significant studies were undertaken to predict the magnitude of road traffic accidents and black spot areas in Addis Ababa City considering the driver as the main causing factor. However significant studies were not undertaken to predict the magnitude of accident severity considering the non-behavioral factors as the main causing factor. As a result, this study developed accident severity prediction models using Multinomial logistic regression that link accident severity to non-behavioral contributing factors. For this study, a total of 5251 traffic accident data from June 30/2011 to June 30/2016GC were collected from Yeka sub-city. Multinomial logistic regression was used to estimate the model parameters. The data set obtained for this study were applied to examine the goodness-of-fit regression models. The dependent variable used in this study was crash severity. As part of the study, the models have been tested to see how well they predict the accidents observed during a one year accident period. From the model developed, road type, road surface condition, crash type, maneuvering condition and lighting conditions were found as significant explanatory variables that influence the prediction of crashes in the yeka sub-city. This indicates that non-behavioral factors have an effect on the occurrence of accident.

Key Words: ***Road traffic accident, Non-behavioral factors, Accident severity Prediction Models, Multinoial Logistic regression, Crash Severity level.***

ACKNOWLEDGMENTS

It might be honest to state that a thesis work cannot be carried out by oneself without the help of the others. First of all, I would like to Thank Almighty God for his unlimited mercifulness and blessing of whom all success is accomplished. My deepest gratitude goes to my advisor, Dr. Mahabubul Bari, for his continuous valuable support, encouragement, and interest in my thesis work. I would like to thank my family and all the people whose patience, advice, encouragement, and support that helped me to complete this thesis. I would like also to acknowledge Addis Ababa Police Commission and Yeka sub-city Statistical technical Staffs for their willingness and support during collecting the required data. Thanks to all other governmental officers who helped me by giving different documents relevant to this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AADT	**Average Annual Daily Traffic**
AE	Around Entertainment
AH	Around Hospital
AH	Around Religious
AI	Around Industry
AIC	Akaike's Information Criteria
AM	Around Market
AN	Afternoon
AO	Around Office
AR	Around Resident
AS	Around School
ASL	Accident Severity Level
BM	Backward Movement driving condition
C	Cloudy Weather Condition
CH	Chilly Weather Condition
CPM	Crash Prediction Model
CRO	Crash hour
CRO	Crash Occurrence
CRT	Type of crash
CW	Cold Weather Condition
D	Drizzle Weather Condition
DA	Dark crash time
DAY	Days of the week
DEF	Daily expansion factors
DL	Day Light
DS	During Stopping
DW	Downward road alignment
E	Evening crash time
EDR	Entrance to Diverging road driving condition
EJR	Entrance to Junction road driving condition
ER	Earth Road
ER	Error Rates
ERA	Ethiopian Road Authority
ESQR	Entrance to Square road driving condition
F	Friday
FLJ	Five-leg Junction intersection type
FT	Fatal
GA	Good Asphalt
GLM	Generalized Linear Model
GOF	Goodness-of-fit
GR	Gravel Road
GW	Good Weather condition
HA	Hazy Weather Condition
HEF	Hourly expansion factors
HR	Heavy Rain
HZ	Highly Zigzag road alignment
INO	Intersection occurrence
IS	Island Separated road type

LAU	Land Use
LE	Local Exit driving condition
LIC	Lighting Condition
LM	Late Morning crash time
LR	Likelihood ratio
LT	Left turning driving condition
M	Maneuvering driving condition
M	Monday
MAC	Defendant Vehicle maneuvering condition
MEF	monthly expansion factor
MLE	Maximum Likelihood Estimation
MN	Mid-night crash time
MO	Morning crash time
N	Night crash time
NGRL	Night with good road light
NO	Noon crash time
NPRL	Night with Poor road light
NWRL	Night without Road Light
O	Others
OT	Overturning Crash type
OTP	Overturning to Passenger Crash type
OW	One-way road type
PA	Poor Asphalt
PAC	Type of Road pavement
PDO	Property Damage Only
R	Roundabout intersection type
RC	Rail Crossing intersection type
ROC	Road Condition
ROG	Road Geometry
ROT	Road Type
RT	Right turning driving condition
RTA	Road Traffic Accident
S	Saturday
SA	Straight ahead driving condition
SA	Straight ahead road alignment
SBS	Two-way Separated with broken line road type
SHS	Straight and Highly Slopping road alignment
SI	Serious Injury
SLI	Slight Injury
SLS	Two-way Separated with Solid line road type
SPSS	Statistical Package for the Social Sciences
SQ	Square intersection type
SR	Sunrise
SS	Sunset
SSS	Straight and Slightly Slopping road alignment
SU	Sunday
SUD	Straight with up and down road alignment
SZ	Slightly Zigzag road alignment
T	Tuesday
TH	Thursday

TS	T-shape intersection type
TW	Undivided Two-way road type
UD	Undefined
UT	U-turning driving condition
UW	Upward road alignment
VTI	Vehicle to Inert Crash type
VTP	Vehicle to pedestrian Crash type
VTSV	Vehicle to Stopped Vehicle Crash type
VTV	Vehicle to Vehicle Crash type
VTVI	Vehicle to Vehicle to Inert Crash type
VTVP	Vehicle to Vehicle to Pedestrian Crash type
VTVPV	Vehicle to Vehicle to Parked Vehicle Crash type
W	Wednesday
WEC	Weather Condition
WOI	Without Intersection
XS	X-shape intersection type
YS	Y-shape intersection type

CHAPTER 1 INTRODUCTION

1.1 Background

Road accidents are very common all over the world and annual global road crash statistics (Association for Safe International Road Travel, 2013) states that nearly 1.3 million people die in road crashes each year, on average 3,287 deaths a day with an additional 20-50 million are injured or disabled. More than half of all road traffic deaths occur among young adult ages between 15 to 44 years. Road traffic crashes rank as the 9th leading cause of death and accounts for 2.2% of all deaths globally. As different researchers mentioned unless an action is taken, road traffic injuries are predicted to become the fifth leading cause of death by 2030 (CHENGYE, RANJITKAR, & Prakash, 2010)

In the developing Country, Road Traffic Accidents are also among the leading cause of death and injury; Ethiopia in particular experiences the highest rate of such accidents. According to the latest WHO (Organization, 2014) data published in May 2014, Road Traffic Accidents Deaths in Ethiopia reached 15,015 or 2.50% of total deaths. Along with the continuous increase in automobiles and many other reasons, the number of traffic accidents in Addis Ababa city (Capital City of Ethiopia) has been increasing from year to year. (Pande) (Organization, 2014)

In general speaking, factors that contribute to road crashes may include inappropriate driver behavior, congested traffic, unanticipated roadway geometrical change and adverse weather condition etc. (Nicholas J.Garber, 2009). The driver behavior plays a crucial role in the occurrence of a crash; however, it is usually complex and unpredictable. But, it is important to figure out the role of the non-behavioral factors in traffic accidents, based on which cost-effective countermeasures can be recommended to reduce the chance of accidents and also to estimate current or future safety performance. In the past, when current or future safety performance estimates for a roadway were needed in the USA, they have been developed by one of four approaches: averages from historical accident data, predictions from statistical models based on regression analysis, results of before-after studies, and expert judgments made by experienced engineers. (Administration U. D., 2000) (Pande)

Among the above four approaches used in the USA, predictions from statistical models based on regression analysis were presented in this study. As a result, the main aim of this study was to develop a model that can predict the probability of the occurrence of traffic accident severity to the future by assessing the predictive abilities of different statistical models so that it is possible to understand the severity of the existing transportation condition and also it can make the decision-maker more conscious to take some steps to improve the condition of transportation system.

Many projects have been conducted on modeling crashes. The existing models include single variable and multivariate deterministic models, stochastic multivariate models, and artificial neural network. After a long period of application of deterministic models, researchers began to realize that the characteristics of crashes are discrete, sporadic, and non-negative (J.Garber, 2001). Researchers, therefore, started applying stochastic models to describe the occurrence of crashes. Apart from the modeling techniques, the causal variables considered in those crash models have shifted from single variable to multiple variables too. (Jonsson, 2005) (Marko Renčelj, 2009)

Various relationships were studied, such as the relationships between the number of lanes and crash rates, traffic volume and crash rates, shoulder and lane widths and crash rates. Traffic volume was believed to have a significant influence on the occurrence of crashes. A U-shaped curve for the relationship between the crash rate and traffic volume was shown. However, no consistent relationships between the occurrence of crashes and geometric parameters have been indicated yet as reported by Williams. (WILLIAMS, 2009)

This study assessed different statistical models to get the best fit model that shows the correlation between traffic accident severity and non-behavioral contributing factors including geometric characteristics of the road, environmental factors, and traffic conditions. Thus identification of non-behavioral related factors which contribute towards crash severity is very important in improving safety or reducing the severity of the accident. This study aims at examining the significant variables that affect the occurrence of an accident such as the occurrence of the intersection, road type, weather condition, road surface condition, road alignment, land use pattern and maneuvering condition.

Multinomial logistic regression, Binary logit and Binary probit models were used in this study to estimate the effect of the statistically significant non-behavioral factors on the probability of the occurrence crash severity.

The Accident severity Prediction Models developed in this study should help in estimating the probability of occurrence of the accident given the types of significant variable. In addition, it can be used during the decision-making process in the management of road safety.

1.2 Statement of the problem

The costs of fatalities and injuries due to road traffic accidents (RTAs) have a tremendous impact on societal well-being and socioeconomic development. According to the latest WHO data published in May 2014 Road Traffic Accidents Deaths in Ethiopia reached 15,015 or 2.50% of total deaths. With more vehicles and traffic, the capital city of Addis Ababa takes the lion's share of the risk, with an average of 20 accidents being recorded every day and even more going unreported. [(Authority, 2015)] As such, one cannot improve safety without successfully relating accident frequency and severity to the causative variables.

Therefore one of the most critical gaps in the management of road safety in our country is focusing too much on the driver as the cause of a crash which often masked the ability to see other causes that could reduce crash rates and crash severity. In other words, something about the road may lead the driver to make a mistake, or the driver may make the mistake and the road may not allow for recovery from the mistake.

By developing accident severity prediction models it can be easy to assess the safety performance and identify hazardous locations needing treatment for both existing and proposed roads by considering non-behavioral factors only. This study developed Accident Severity Prediction Model by assessing the predictive abilities of different statistical models, and the developed model was tested using a one year collected traffic crash data.

The other issue is that, even though there are several decision-making styles in practice, almost all of them does not use modeling as a basic building block. The acceptability of modeling or a particular modeling approach within a decision style is very important. Models which end up being ignored by decision-makers not only represent wasted resources & effort but resulted in frustrated analyst and planners. This study recommended that the contribution

of modeling (which is a latest best fit model) can make to improved decision-making and planning if adopted as an effective to decision making process since it possible to estimate crash probabilities from the developed best fit model.

1.3 Research Question

In this study three research questions have been formulated and specific answers need to be obtained. The table below shows objectives with the specific questions to address them.

- Does non-behavioral factors influence the occurrence of the crash?
- Does the developed statistical Accident Prediction Model estimate nearly the same as the collected traffic accident data?
- Is Accident prediction model important during decision making process

If "Yes" is the answer, what shall be the importance of a particular modeling approach to have an effective decision making process?

- Null – Hypothesis: Statistical Accident Prediction Model is very important during decision making;
- The alternative – Hypothesis: Statistical Accident Prediction Model is not very important during the decision-making process.

Table 1-1 Research objective with its questions

Research objectives	Research questions
To develop and assess statistical Accident Prediction Model	• What statistical distribution best describes the data? • What are the main independent variables/parameters for developing the model? • Which methods have been used to develop APM? • What are the selection criteria for different statistical regression analysis for APM?

1.4 The objective of the study

1.4.1 General Objective

The general objective of the study was to develop statistical Accident severity Prediction Model and to assess its predictive ability considering non-behavioral factors of the accident.

1.4.2 Specific Objectives

The specific objective of this study was to correlate traffic accident severity with non-behavioral factors which include road environment condition, traffic condition and the road characteristic features.

To do the above objective, the activities conducted in this study are:

- To select the main causing non-behavioral factors of road traffic accidents based on tests of proportion.
- To determine which independent variables have a statistically significant effect on the dependent variable.
- To develop a model showing the relation between main causing non-behavioral factors with crash severity level.
- To test and investigate whether the developed model can be used as an input during decision-making process or not.

1.5 Scope and Limitation of the study

1.5.1 Scope of the Study

This paper focuses on the issue of Road traffic accidents and statistical accident prediction model development in Yeka Sub-city of Addis Ababa City administration.

1.5.2 Limitation of the study

In this sub-city, a six-year traffic accident data was collected for Model development (a four year data used) and Model result testing purpose (a two year data used), however, the study has the following limitations;

- For the sake of making the research manageable, this study has been limited in scope (only considers the effect of non-behavioral factors on road traffic accident), time (only a four year traffic accident data has been collected for model development and added a two year data for model result testing), and coverage areas (only considers one sub-city in Addis Ababa).
- Road traffic accident data from Yeka sub-city police commission Bureau is in the form of Hardcopy, and it makes very difficult to collect a six-year accident data considering all parameters in the booklet. Therefore for this study, only non-behavioral data was collected excluding vehicle-related factors.
- During data collection in MS-excel sheet, translation of Amharic words from daily traffic accident recording sheets to standard words was tried but not for all (in some cases direct translation was used).

- Because of the impossibilities of getting the true underlying number of accidents, this study assumes the police reported accident data as a representative sample of all accidents occurring.
- Expansion factors were used for determination of AADT.
- A 3-hour classification system was used for analyzing accident based on the crash hour for data collection purpose.
- Besides, possible efforts were exerted to overcome the above constraints and to accomplish the desired work successfully.

1.6 Significance of the study

Due to the problem of road safety, the costs of traffic fatalities and injuries can be huge not only in terms of money, but also physical and mental sufferings. Road crashes add a burden to the whole society, such as medical emergency services, hospitalization, traffic congestion, social welfare, and insurance systems in the long term, which may negatively influence the productivity and competitiveness of a society as stated by (Tsui, 2006). These costs can significantly be reduced if the probability of occurrence of crash can be better estimated and countermeasures can be properly carried out and strictly enforced.

Knowing the most significant variables/ factors plays a key role in addressing road safety problems. In this study, significant variables are correlated with the dependent variable for estimating the probability of occurrence of crash in the future. As a result the developed best fit model can be used as an input during decision making process in road safety management and based on the model and other criteria, cost effective countermeasure can be applied.

Generally the methodology followed in this study is worthy in reducing the costs brought by the traffic road crashes.

1.7 Thesis Report Organization

The Study is comprised of five Chapters. The first chapter consists of a background of the study, statement of the problem, objectives of the Research, the significance of the Study, limitation of the Study and organization of the Study. The second chapter comprises a review of related literature to support this study which provides an introduction to accident prediction models, elements of model development, assessing the predictive abilities of statistical accident models and the like. Chapter three is on research methodology which describes data

source and method of data collection, materials, and software used, selected models and model estimation technique; assessment of model performance. Chapter four (result and discussion) consists of the construction of modeling parameters, the correlation between variables, model results and analysis, model result testing and assessing the predictive ability of models. Chapter five consists of conclusions and recommendations which proposes future research areas.

CHAPTER 2 LITERATURE REVIEW

Transport is the movement of people and goods from one place to another. But according to (Belachew M., 1997), transport also comprises movement of information. Similarly, Transportation is the conveyance of people, properties and information from one place to another or it is the repositioning of people, properties, and information over space. The availability of highway transportation has provided several advantages that contribute to a high standard of living. However, several problems related to the highway mode of transportation exist. As (Nicholas J.Garber, 2009) reported, these problems include highway-related crashes, parking difficulties, congestion, and delay.

The type of transport which exhibits accident that drastically affects the well-being of the people and economy of the nations is the one which involves the movement of people and or goods from one place to the other. Several Road Traffic Accident (RTA) incidences occur throughout the world at every fraction of times in a day. Whatever the reason, where ever the scene and whoever the victim is, RTAs remain as a headache for everyone.

The most shocking and emerging reality of RTA is that it will continue affecting the survival of several lives across the planet. Consequently, (UN, 2009), remains pessimistic in road traffic accident cases where it projected that road traffic injury will be the fifth – leading cause of death globally by 2030. However, WHO (2004) projected that RTA crashes which were ranked at the 9th leading cause of burden of disease by 2002 could rank at the 3rd cause of burden of disease by 2020. If the current trend in motorization continues increasing in the same or similar manner for the coming decade.

Road Traffic Accident is any vehicle accident occurring on a public highway which includes collision between vehicles and animals, vehicles and pedestrians or vehicles and stuck obstacles. Single vehicle accidents that involve a single vehicle, which means without another road user, are also enclosed (Safecarguide., 2016,).

In a similar manner (Ajit G. and S. Ripunjoy, 2004) have mentioned that Accident is an occasion, occurring abruptly, unpredictably and inadvertently under unforeseen circumstances. As (Berhanu G., 2000) reported that, the manifestations of RTA are sporadic and random in space and time. Seemingly, (Segni G., 2007) have also outlined that an

accident is a rare, random, multi-factor event always preceded by a situation in which one or more road users have failed to cope with the road environment.

In this regard, Road Traffic Accident (RTA) can be defined as an accident that occurred on a way or street open to public traffic; resulted in one or more persons being killed or wounded, and at least one stirring vehicle was intricate. Therefore, RTA is a smash between vehicles; between vehicles and pedestrians; between vehicles and animals; or between vehicles and geographical or architectural obstacles.

In general speaking, factors that contribute to road crashes may include inappropriate driver behavior, congested traffic, unanticipated roadway geometrical change and adverse weather condition etc. The driver behavior plays a crucial role in the occurrence of a crash; however, it is usually complex and unpredictable. But, it is important to figure out the role of the non-behavioral factors in traffic accidents, based on which cost-effective countermeasures can be recommended to reduce the chance of accidents and also to estimate current or future safety performance.

In the past, when current or future safety performance estimates for a roadway were needed in the USA, they have been developed by one of four approaches: averages from historical accident data, predictions from statistical models based on regression analysis, results of before-after studies, and expert judgments made by experienced engineers. Their general concept is described below. (Administration U. D., 2000) (Pande)

2.1 Estimates from Historical Accident Data

As (Marko Renčelj, 2009)reported, Historical accident data are an important indicator of the safety performance of a roadway, but they suffer from the weakness of being highly variable. Given this high variability, it is difficult to estimate the long-term expected accident rate using a relatively short-duration sample of 1 to 3 years of accident data. As (Marko Renčelj, 2009) reported that, if a location has experienced no accidents in the past several years, it is certainly not correct to think that it will never experience an accident, yet the available data for that site alone provide an insufficient basis for estimating its long-term expected safety performance. (Marko Renčelj, 2009)

A high-accident location is a roadway section or intersection identified because it experienced more than a specified threshold number of accidents during a recent period (typically 1 to 3 years). Each high-accident location is investigated by the engineering staff of the responsible highway agency and, at locations where a particular accident pattern is clearly evident and an appropriate countermeasure is feasible, an improvement project may be programmed and constructed. As reported by (Marko Renčelj, 2009) the decision making concerning such projects often involves a benefit-cost or cost-effectiveness calculation based on the expected percentage reduction in accidents from the level of recent accident experience. However, both statistical theory and actual experience show that, because of the random nature of accidents, locations with high short-term accident experience are likely to experience fewer accidents in the future even if no improvement is made. This phenomenon, known as *regression to the mean*, makes it difficult both to identify potential problem locations and estimate the potential (or actual) effectiveness of improvements made at such locations. (Administration U. D., 2016) (Marko Renčelj, 2009)

2.2 Estimates from Expert Judgment

As mentioned by (D.W. Harwood, 2000) Expert judgment, developed from many years of experience in the highway safety field, can have an important role in making reliable safety estimates. Experts may have difficulty in making quantitative estimates with no point of reference, but experts are usually very good at making comparative judgments (e.g., A is likely to be less than B, or C is likely to be about 10 percent larger than D). Thus, experts need a frame of reference based on historical accident data, statistical models, or before-and-after study results to make useful judgments. But according to the limitation of experts' experience, there might be some defects in this method. (Administration U. D., 2016) (D.W. Harwood, 2000)

2.3 Estimates from Before-and-After Studies

'Before-and-after studies have been used for many years to evaluate the effectiveness of highway improvements in reducing accidents. However, as Hardwood (2000) reported, most before-and-after studies have design flaws such that the study design cannot account for the effects of regression to the mean. Therefore, the potential user of the before-and-after study results cannot be certain whether they represent the true effectiveness of the potential

improvement in reducing accidents or an overoptimistic forecast that is biased by regression to the mean'. (D.W. Harwood, 2000)

'Safety experts are generally of the opinion that, if the potential bias caused by regression to the mean can be overcome, a before-and-after study may provide the best method to quantify the safety effect of roadway geometric and traffic control features'. (Administration U. D., 2016) (D.W. Harwood, 2000)

2.4 Estimates from Statistical Models

Safety analysts have, for many years, applied statistical techniques to develop models to predict the accident experience of roadways and intersections. Such models are developed by obtaining a database of accident and roadway characteristics (e.g., traffic volumes, geometric design features, and traffic control features) data from highway agency records, selecting an appropriate functional form for the model, and using regression analysis to estimate the values of the coefficients or parameters in that model.

Historically, most such models were developed with multiple regression analysis. Recently, researchers have begun to use Poisson and negative binomial regression analyses which are theoretically better suited to accident data based on small counts (i.e., zero or nearly zero accidents at many sites). Regression models are very accurate tools for predicting the expected total accident experience for a location or a class of locations, but they have not proved satisfactory in isolating the effects of individual geometric or traffic control features. (Administration U. D., 2016) (Marko Renčelj, 2009)

Among the above four approaches used in USA, predictions from statistical models based on regression analysis was presented in this study in case of Addis Ababa City, in particular Yeka sub-city. As a result, the main aim of this study was to develop a statistical model that can predict the probability of occurrence of traffic accident to the future, and to assess the predictive abilities of different statistical models so that it is possible to understand the severity of the existing transportation condition and also it can make the decision-maker more conscious to take some steps to improve the condition of our transportation system.

2.4.1 Statistical Accident Severity Model

There are two types of crash analysis methods that used with bicycle safety studies: 1) predictive model types and 2) discrete choice model types. Predictive models and discrete choice models have distinct differences in purpose. Predictive models typically have very few variables and literally predict the likelihood of a crash outcome based upon the future existence of a specific variable. Discrete models analyze past crashes in order to define the degree of influence a certain variable could have on the injury-severity if that variable is present. Typically, discrete models have a much higher number of variables compared to predictive models. Popularly used predictive models in crash analysis include Conventional Linear Models, Generalized Linear Model- Poisson Models, Negative Binomial Regression Models and Empirical Bayes Method. The second type of bicycle crash analysis used frequently is a discrete choice model. Popular discrete choice models used throughout various academic disciplines and industries in crash analysis include Multinomial Logit/Probit model, Nested Logit/Probit model, mixed multinomial logit/probit model and random utility models. The most common discrete choice model is the multinomial logit model.

A number of accident Severity models have been developed to estimate the expected accident severities or frequencies on roads as well as to identify various factors associated with the occurrence of accidents. Most previous researchers have focused on both behavioral and non-behavioral factors (such as traffic flow characteristics, road characteristics, environmental conditions and similar parameters), and the statistical methods used by researchers are mainly dependent on the nature of the response variable and various methodological issues associated with the data. The response variable of existing accident severity models is generally either a binary outcome (e.g., injury or non-injury) or a multiple outcome (e.g., fatality, disabling injury, evident injury, possible injury, no injury or property damage).

As reported by many researchers, an effective road safety management requires a good insight into the factors that are believed to be related to road traffic accidents. Based on this framework, several research studies have been conducted over the years aiming at identifying factors that may influence both the frequency and the severity of road traffic accidents. However, as pointed out by (Savolainen, 2011), one has to be aware that the factors influencing accident frequency may vary from the ones affecting the severity; hence, it is suggested that their analysis should be performed individually. In the area of accident severity research, continuous efforts have been conducted in order to investigate the relationship

between the level of severity (dependent variable) and a set of explanatory variables, which usually include: driver attributes (e.g., age and gender), vehicle features (e.g., body type, vehicle age and number of vehicles involved in the accident), road characteristics (e.g., number of lanes, road surface conditions, intersection control and types of road), and accident characteristics (e.g., accident's main cause). Occasionally, the influence of other variables on accident severity like speed limit, day of the week, time of the day, average traffic characteristics (AADT), weather and traffic conditions have also been scrutinized (Manner, 2013) (Rui Garridoa, 2014)

Various techniques have been applied to the analysis of accident severity data. The earlier traffic accident studies used ordinary or normal linear regression models, which follow the assumption of a normal distribution for the dependent variable, a constant variance for the residuals, and the linear relationship existing between dependent and independent variables. This assumption is misleading because the occurrence of crashes is random, discrete, and rare. As (Turner, 1996) reported, there are at least three reasons why accidents are not normally distributed. The first reason is that the frequency of accidents is a discrete variable and hence should have a discrete distribution, while the normal distribution is a continuous distribution. The second reason why the normal distribution is inappropriate is that the variance in the number of accidents is not constant for all values of the explanatory variables, as is assumed with the normal distribution. The final reason the normal distribution is inappropriate is that it does not take account of the non-negativity of the accident counts. In situations where the response variable does not have a normal distribution or when its distribution cannot be approximated using a normal distribution, discrete outcome models can often be used to fit a regression line to the data. (Turner, 1996), (J.Garber, 2001), (Tulu, 2015)

Multinomial logistic regression (MLR) models are widely used in transportation to study the relationship between the categorical dependent variable and a set of continuous and categorical independent predictor variables collected through surveys. Washington et al. (2011) developed a MLR model consisting of 18 independent variables covering driver factors, traffic flow, distance, and number of signals etc. to study factors that influence driver's selection of route on their morning commute to work. Yan et al. (2009) utilized MLR to study the impact of potential factors such as driver factors, road layout, and environmental conditions on rear-end truck to car, car to truck, and car to car crashes. A MLR model was

developed by Morfoulaki et al. (2007) to identify the factors contributing to service quality and customer satisfaction (*very satisfied, satisfied, somewhat dissatisfied, and very dissatisfied*) with a public transit service in Greece. Gkritza et al. (2006) conducted an empirical study using multinomial logit models to investigate the socio-economic and demographic factors that significantly affect passenger satisfaction with airport security screening process. (Washington, 2011) (Yan, 2009) (Morfoulaki, 2007) (Gkritza, 2006)

Bédard et al. (2002) applied a multivariate logistic regression to assess the independent contribution of driver, accident, and vehicle characteristics to fatal injuries sustained by drivers. They found that factors such as increased driver age, female gender, blood alcohol concentration greater than 0.30, non-belted drivers, driver side impacts, travelling speeds greater than 70 mph, and older vehicles were associated to higher odds of fatal outcomes; on the opposite, a 25 cm increase in vehicle size (wheelbase) was translated into a reasonable reduction of the fatality risk. (Bedard, 2002)

A multivariate logistic regression was also applied by Bedard et al. (2002) to determine the relation between driver characteristics and vehicle conditions to accident fatality rate. O'Donnell and Connor (1996) evaluated the probabilities of four levels of accident severity as a function of driver properties and they compared the Ordered Logit and Ordered Probit criterions. Kockelman and Kweon (2002) applied ordered probit model to study the risk of different injury levels sustained under all crash types. Khattak et al. (2002) applied an ordered probit modeling approach in their study to investigate dependent variables including vehicle property, pavement, driver, and environmental characteristics that can potentially cause more severe accident with older drivers. (Bedard, 2002) (Khattak, 2002) (Kockelman, 2002) (O'Donnell, 1996)

The ordered probit modelling methodology was also used by Abdel-Aty (2003) in an attempt to analyze the driver injury severity levels at several roadway entities. Three separated models were developed for signalized intersections, roadway sections, and toll plazas. Older and male drivers, those not wearing a seat belt, drivers of passenger cars (i.e., vehicle type), vehicles struck at driver's side (i.e., point of impact/accident characteristics), and those who speed revealed a higher probability of a severe injury in all the models. Other variables were found significant only in specific models: alcohol, dark lighting conditions, and the existence of horizontal curvatures affected the likelihood of injuries in the roadway sections' model; a driver's error was significant in the model for signalized intersections; vehicles equipped with

an electronic toll collection affected the likelihood of higher injury severity in the toll plazas' model; lastly, both signalized intersections and roadway sections models revealed higher probability of injuries in rural areas. Furthermore, the author tested other modelling approaches, namely a multinomial logit and a nested logit with different nesting structures, and compared the results produced by those models with the results provided by the ordered probit model. This comparison showed that the ordered probit model, besides being simpler, has also produced better results than the multinomial logit; by comparing the ordered probit model with the best nesting structure of the nested logit model, the author recommended the former to model driver injury severity because, in spite of the similarity found in the results provided by both models, the latter has introduced considerable complexity in the modelling process. (Abdel-Aty, 2003)

To summarize, most of the previous research was focused on the number of accidents and the factors that could increase the accident frequency. The research that considered accident severity generally took into account all factors that had relations with accidents. However, there were only very few studies that included environmental related characteristics in the analysis of accident severity. This is the focus of this research and will be described next. Further to the above; many more researches have been conducted by different researchers and professionals to develop Accident Severity Prediction Models by considering necessary elements for model development.

2.4.2 Elements of Model Development

As (WILLIAMS, 2009) stated that, the following elements have been identified to be necessary for any model development: appropriate model form and error structure, procedure for selecting the model explanatory variables, procedure for outlier analysis, and methods for assessing model goodness of fit. (WILLIAMS, 2009), (Miaou & Lum, 1993)

2.4.2.1 Model Form and Error Structure

(WILLIAMS, 2009), reported that, Miauo and Lum (1993) investigated the statistical properties of two conventional linear regression models and identified potential limitations of these models in developing vehicle crashes and highway geometric design relationships. It was demonstrated that the conventional linear regression models lack the distributional property to describe adequately random, discrete, non-negative, and typically sporadic, vehicle crash events on the road.

As a result, these models were not appropriate to make probabilistic statements about the occurrences of vehicle crashes on the road, and the test statistics derived from these models were questionable. Several other kinds of literature have supported the unsatisfactory property of ordinary linear regression models in developing vehicle crashes, traffic flow, and highway geometric relationships. (J.Garber, 2001), (Miaou, Shankar et al., & Lord et al., 1994, 1997, 2005), (Heinz Hautzinger, 2007)

Currently, Generalized Linear Regression Model (GLM) is used almost exclusively for the development of Crash Prediction Models; and logistic regression or logit regression models are the most common models used in analyzing influential factors for traffic crashes. (Sawalha & Sayed, 2006) said the mathematical form used for any CPM (Crash Prediction Model) should satisfy two conditions:

1. It must yield logical results (i.e. it must not lead to the prediction of the non-negative number of crashes and it ensure a prediction of zero crash frequency for zero values of the exposure variables,
2. To use generalized linear regression in modeling procedure, there must exist a known link function that can linearize this form for the purpose of coefficient estimation.

The above conditions according to Sawalha and Sayed (2006) are satisfied by a model form that consists of the product of the power of the exposure measures multiplied by an exponential function incorporating the remaining explanatory variables. Such a model form can be linearized by the logarithm link function; expressed mathematically in the following model form:

$$E(Y) = a_o L^{a1} Q^{a2} exp \sum_j b_j x_j \quad (1)$$

Where E(Y) is the expected prediction crash frequency, L is the section length, Q is some function of flow, xj is variables describing road geometry or environment of the road or any variable additional to L and Q, Expo is the exponential function e=2.718282 and a0, a1, a2, bj are the model parameters. (WILLIAMS, 2009) (Marko Renčelj, 2009) (Jonsson, 2005)

As (Fred Mannering, 2010) described that, in addition to Model form, important data and methodological issues should be identified in the crash severity literature. These issues have been shown to be a potential source of error in terms of incorrectly specifying statistical

models which may lead to erroneous crash-severity predictions and incorrect inferences relating to the factors that determine the frequency of crashes.

The accident severity investigations attempted to examine the influence of driver attributes, vehicle features, accident characteristics, and road, weather, and traffic conditions on the severity outcome. A wide variety of methodological approaches have been used to fulfil this propose. Recently, Savolainen et al. (2011) conducted a research, which intended to assess the characteristics of road accidents severity data and the methodological approaches most commonly used for the analysis of such data. The authors highlighted that the "appropriate methodological approach can often depend heavily on the available dataset, including the number of observations, quantity and quality of explanatory variables, and other data-specific characteristics". Still, they found that the majority of the modelling approaches were framed in the discrete response models which include: binary response models (e.g., binary probit and binary logit), ordered discrete response models (e.g. bivariate ordered probit and generalized ordered logit), unordered multinomial discrete response models (e.g. multinomial logit, Markov switching multinomial logit, nested logit, and mixed logit) (Savolainen, 2011)

From the above methodological issues, some of it which support this study was selected for further explanation, and described herein below.

2.4.2.1.1 Logistic Regression

In statistics, logistic regression or logit regression is a type of probabilistic statistical classification model. It is also used to predict a binary response from a binary predictor, used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. That is, it is used in estimating empirical values of the parameters in a qualitative response model. The probabilities describing the possible outcomes of a single trial are modeled, as a function of the predictor variables, using a logistic function. Frequently logistic regression is used to refer specifically to the problem in which the dependent variable is binary—that is, the number of available categories is two—and problems with more than two categories are referred to as multinomial logistic regression or, if the multiple categories are ordered, as ordered logistic regression. (Geedipally, 2005)

2.4.2.1.2 *Multinomial Logit Model*

Multinomial logit regression is suitable for modeling nominal outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the regression variables.

According to the literature, since the dependent variable, accident severity, has a discrete nature, discrete choice models are identified as the most suitable approach. Among all the discrete choice models, the multinomial logit model (MNL) is the easiest and most widely used in predicting accident severity. One primary feature of MNL models is that they do not recognize any order in injury levels. This means that the probabilities of property damage, people injuries, or fatalities occurring as a result of each weather factor do not follow the same order as the accident severity level. For example, if the regression result shows that a higher air temperature may increase the possibility of accidents with injuries compared to accidents with property damage, we cannot conclude that a higher air temperature may also increase the possibility of accidents with fatality. Because of this feature, MNL models do avoid certain restrictions posed by standard ordered models, because they allow variables to have opposing effects regardless of injury order. An MNL model assumes that the unobserved factors are uncorrelated over the alternatives, also known as the independence of irrelevant alternatives assumption.

The multinomial logit is a widely used discrete choice model in various fields. The formula of the logit model was first introduced by Luce in 1959 from the assumption of probability of choice (Train, 2009). Following this, there have been a number of contributions to the development of this model (McFadden, 1974) (Train, 2009) (Tulu, 2015)

The probability expression of the multinomial logit regression is:

$$P_i(k) = \frac{\exp(\alpha_k + \beta_k X_{ik})}{\sum_{k=1}^{n} \exp(\alpha_k + \beta_k X_{ik})} \quad (2)$$

The general framework used to model the degree of injury severity sustained by a crash that involves individual begins by defining a linear function S that determines the injury outcome k for observation i as

$$S_{ik} = \beta_k X_{ik} + \varepsilon_{ik} \quad (3)$$

Where:

ε_{ik} →An error term for severity level k and observation i

S_{ik}→ Linear function for severity level k and observation i

$P_i(k)$ → Probability function for severity level k and observation I

β_k →Parameters to be estimated by the model

X_{ik} is the independent variables

The estimation of the model parameters can be carried out through the method of maximum likelihood.

A logit model for multiple injury severity was first used last century (Shankar, Mannering, & Barfield, 1996). Even though other methods have been introduced more recently, multinomial logit model continues to be used as an analytical methodology of injury severity (Lee & Abdel-Aty, 2005; Obeng & Rokonuzzaman, 2013; Wang, et al., 2013). Nowadays scholars promote the use of multinomial logit model injury severity analysis due to the problem of underreporting, and misclassification of crash data (Islama & Mannering, 2006). Flexibility is considered the main advantage of this method. The details will be discussed in the following section of ordered response model. (Wang, 2013) (Shankar V. M., 1996) (Obeng, 2013) (Lee, 2005) (Islama, 2006)

2.4.2.1.3 *Nested Multinomial Logit Model*

This model has been used by researchers in a variety of circumstances, including transportation data analysis. It is appropriate to apply this model if the set of alternatives can be fragmented into subset (Train, 2009). The nested logit can overcome the problem of independence of irrelevant alternatives, and resolves independence of irrelevant alternatives problem by grouping alternatives that form share the nests. (Train, 2009) (Lee, 2005)

2.4.2.1.4 Ordinal Response Models

The proportional odds model assumes an ordering to the categories. Researchers tend to model this approach (Washington, et al., 2011). The ordered crash data like property damage, slight injury, serious injury and fatal injury are naturally in order. However, the natural orders of crashes are often disturbed. Ordinal data, such as degree of severity in pedestrian crashes, are preferentially modelled using ordered probability structure methods; however, there are circumstances that affect the outcome of the models (Washington, et al., 2011). The major

problem of the characteristics of crash injury severity data that should be considered in selecting the appropriate methodology are the underreporting of crashes, the ordinal nature of the data, omitted variable bias, accounting for fixed parameter effects, small sample size, unobserved heterogeneity, within-crash correlation, and spatial and temporal correlations (Savolainen, et al., 2011; Ye & Lord, 2011). In such situations, even though the data are ordinal, the ordered probability model may not be suitable for the data to be modelled. These authors suggest that caution should be taken in selecting ordered and unordered models since a trade-off is being made between forming ordered responses and giving up the flexibility provided by unordered outcome models. Other authors express concern in selecting appropriately between ordered or unordered structure models (Islama & Mannering, 2006). All agree that ordinal injury severity data could be affected by the availability of air bags in cars, which reduce the degree of injury severity (fatality) and increase the number of slight injuries, thus introducing bias in the natural order. Moreover, the underreporting of injuries is also a potential problem in modelling using the ordered probability model, especially in developing countries where many cases of serious and minor injury are not completely recorded. More complex models are needed like the mixed logit model to address unobserved heterogeneity. (Washington, 2011) (Islama, 2006)

2.4.2.1.4.1 Ordered Probit Model

As mentioned in the analysis of MNL model, one of the significant drawbacks is that an MNL model doesn't consider the ordering information for accident severity (ranked as fatality, personal injury, property damage). The ordered probit (OP) model, however, addresses the problem of independence of irrelevant alternatives and includes the ordered discrete data (Kockelman, 2001). In order to apply an OP model here, we assume that the sample is large enough so that all unobserved components of utility have normal distributions. Accounting for the ordinal nature of injury data (for example, ranging from no-injury, to possible injury, to evident injury, to disabling injury, to fatal injury) is an important consideration in crash injury-severity modeling (O'Donnell, 1996). To account for the ordinal nature of the data, traditional ordered probability models have been widely applied. (Kockelman, 2002) (O'Donnell, 1996)

2.4.2.1.5 Model Evaluation and Selection

The previous literatures stated that the appropriate means often depended mainly on the available dataset, including sample size, quantity and quality of explanatory variables, as well

as specific characteristics of other data. So far, there is no consensus on which model is the best, because the model selection criteria are often determined by the achievability and nature of the data. In some research papers, ordinal models were more popular than nominal models because nominal models use the same coefficient for estimators among different accident severity and restrict how variables affect outcome probabilities. The advantage of nominal models is their simplicity and overall performance when the sample is small and lacks detail. Some researchers directly compared accident severity models, such as Abdel-Aty (2003), who preferred the OP model to the MNL and ML models, while another study by Haleem and Abdel-Aty (2010), led to a conclusion that the binary probit model performed better compared to the OP and NL models. (Abdel-Aty, 2003) (Haleem, 2010)

Overall, although continuous progress has been made in accident severity modeling over the years, the best performance methodology has yet to be found. Different method should be applied under different conditions and restrictions.

2.4.2.2 Procedure for Outlier Analysis

Data collected may contain odd or extreme observations called outliers. Outliers occur in a set of data either because they are really different from the rest of the data or because errors took place during data collection and recording. These extreme observations may have an effort on the model equation. Outlier analysis could be carried out to identify these influential observations.

2.4.3 Methods for Assessing Model Goodness of Fit

Researchers in this area have mentioned two major statistical measures used in assessing the goodness of fit a model. The first statistical approach is called Pearson Chi-squared statistic, which is defined as:

$$Pearson\ \chi^2 = \sum [y_i - E(Y_i)/Var(Y_i)] \quad (4)$$

Where y_i and E (Yi) are the observed and estimated crash frequencies; and Var(Yi) is the estimated variance of y_i

The other goodness of fit statistics is called Deviance. The deviance is the likelihood ratio test statistic measuring twice the difference between the maximized log-likelihoods of the studied model and full or saturated model. The full model has as many parameters as there

are observations so that the model fits the data perfectly. Therefore, the full model, which possesses the maximum log-likelihood achievable under the given data, provides a baseline for assessing the goodness of fit of an intermediate model.

Both the Pearson Chi-squared and scaled deviance have exact Chi-squared distributions, and the estimated value divided by its degrees of freedoms should be close to one. (i.e. D/(N-p) or χ^2/(N-p)=l. (Bauer & D.W. Harwood, 1996), have mentioned that if the values of these ratios are between 0.8 and 1.2 for a given model, then the model can be considered to appropriate for representing the data.

2.4.3.1 Criteria for assessing the quality of Accident Prediction Model

Accidents are a very complex phenomenon - hence models also need to be complex in order to faithfully reproduce the main features of reality. As (Marko Renčelj, 2009) mentioned, yet, the *"art"* of model building is, and will always be, the art of making the right simplifications. A good model is not necessarily an immensely complex model that perfectly fits the data in every detail. A good model is rather the simplest possible model that adequately fits the data. The following criteria are proposed by Marko Rencelj for assessing the quality of APM: (Marko Renčelj, 2009) (Reurings, 2005)

- As a basis for developing a model, the "probability distribution" of accidents in the original data set should be tested, this test should include several of the most commonly used probability distributions for accidents;
- The "residual terms" of the model should be specified according to the same probability distribution as the original data set, the structure of residuals should always be tested;
- Separate models should be developed for accidents at different levels of severity. As a minimum, separate models are required for fatal accidents, injury accidents (sometimes including fatal accidents) and property-damage-only accidents;
- Separate models should be developed for different types of roadway elements. Roadway elements include: road sections, intersections, bridges, tunnels, curves, railroad-highway grade crossings;
- Data on exposure should be decomposed to the maximum extent possible;
- The functional form used to describe the relationship between each independent variable and accidents should be explicitly chosen and reasons given for the choice. Alternative functional forms should be tested as a basis for the choice made;
- Explanatory variables should be entered stepwise into the model;

- The correlations between explanatory variables should be examined to avoid including very highly correlated variables in the model;
- The possible presence of omitted variable bias should always be discussed. It is understood that no APM can be "complete" by including absolutely every conceivable variable that may influence accident occurrence.
- The predictive performance of an APM should be tested. This is done by applying the model to a data set that was not used in developing the model. (Marko Renčelj, 2009)

2.4.4 Assessing potential source of errors in Predictive Models

There are many sources of error in APM. The most frequently discussed sources of error include omitted variable bias, bias due to co-linearity among explanatory variables, wrong functional form for relationships between variables. Possibly the most common form of omitted variable bias in current APM is the incompleteness of exposure data. Explanatory variables in APM tend to be correlated may lead to unstable estimates of the coefficients. (WILLIAMS, 2009)

2.4.5 Types of variables

When working with statistics, it's important to recognize the different types of data: numerical (discrete and continuous), categorical, and ordinal. *Data* are the actual pieces of information that can be collected, retrieved and analyzed for different purposes.

In statistics, a **variable** has two defining characteristics: A variable is an attribute that describes a person, place, thing or idea; and the value of the variable can "vary" from one entity to another. Variables can be divided into different schemes like dependent and independent variable, qualitative and quantitative variables, Categorical and Continuous Variables, Qualitative vs. Quantitative Variables, Discrete vs. Continuous Variables, Univariate vs. Bivariate Data, and etc.

2.4.5.1 Dependent and Independent Variables

An independent variable sometimes called an experimental or predictor variable, is a variable that is being manipulated in an experiment in order to observe the effect on a dependent variable, sometimes called an outcome variable. The dependent variable is simply that, a variable that is dependent on an independent variable(s). (Wikipedia t. f., 2016)

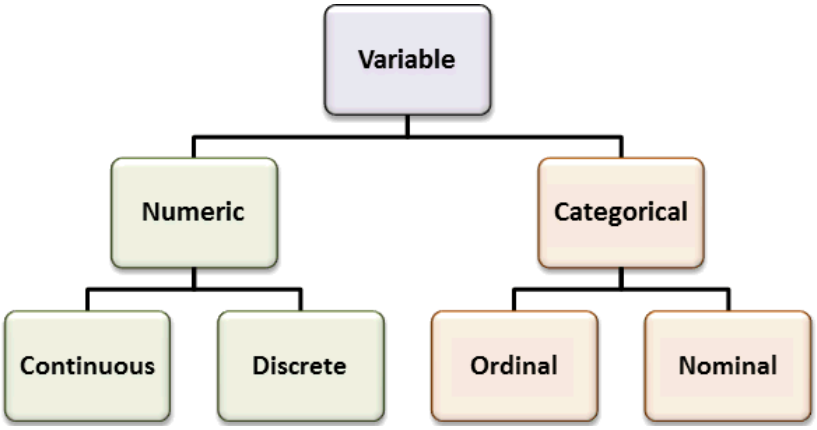


Figure 2-1Types of variables flowchart

2.4.5.2 Categorical and Continuous Variables

Categorical variables have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'. Categorical variables are qualitative variables and tend to be represented by a non-numeric value. Categorical variables are also known as discrete or qualitative variables. Categorical variables can be further categorized as either nominal, ordinal or dichotomous. (Wikipedia t. f., 2016)

- A nominal variable is a categorical variable in which observations can take a value that is not able to be organized in a logical sequence.
- An ordinal variable is a categorical variable in which observations can take a value that can be logically ordered or ranked.
- Dichotomous variables are nominal variables which have only two categories or levels.

Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. It may be further described as either continuous or discrete:

- A discrete variable is a numeric variable in which 0bservations can take a value based on a count from a set of distinct whole values.
- A continuous variable is a numeric variable in which observations can take any value between a certain set of real numbers.

2.4.5.3 Qualitative vs. Quantitative Variables

Variables can be classified as **qualitative** (categorical) or **quantitative** (numeric).

- Qualitative variables take on values that are names or labels.

- Quantitative variables are numeric. They represent a measurable quantity.

2.4.5.4 Univariate vs. Bivariate Data

Statistical data are often classified according to the number of variables being studied.

- Univariate data. Univariate analysis is the simplest form of quantitative (statistical) analysis. (Wikipedia, n.d.)
- Bivariate data. Bivariate data involves the analysis of two variables for determining the empirical relationship between them.
- Multivariate data: Describes a case in terms of Analysis of two or more variables simultaneously.

2.4.6 Selection of Relevant Variables for Analysis and Modeling

As (José M. Pardillo Mayora & Rubio, 2003) stated that, the first step in the development of the model was to define a set of variables based on roadway characteristics and to analyze its correlation with accident rates or frequency. Research efforts aimed to characterize and quantify the relationships between crash rates/frequency and traffic and roadway infrastructure characteristics that have been on-going for over 40 years.

The Federal Highway Administration (Administration U. D., 2000) show that over 50 different roadway variables have been identified as having some influence on crash rates. These variables pertain to different roadway features: horizontal and vertical alignment (i.e.: degree of curvature, grade, sight distance, existence of spiral transition, etc.); cross section (i.e.: roadway width, lane width, shoulder width, etc.); roadside features (i.e.: roadside hazard rating, existence of guardrails, roadside slope, obstacle free zone, etc.); intersections and interchanges (i.e.: intersection layout, intersection angle sight distances, channelization, etc.); and access control (i.e.: driveway density, access channelization, etc.).

In the development of crash prediction models, it is crucial to determine which among these variables capture to a greater extent the effects of highway design and operation on safety. (José M. Pardillo Mayora & Rubio, 2003), (J.Garber, 2001)

In consonance with earlier studies, the decision on which variables should be retained in the model was based on either of the following methods: The chi-squared test of independence, principal component analysis, variable identification factor, Backward Elimination (Top down approach), Forward Selection (Bottom up approach) and Stepwise Regression

(Combines Forward/Backward) (Miaou & Lum, 1993) (José M. Pardillo Mayora & Rubio, 2003), (J.Garber, 2001) (Ronald, Raymond, & Sharon, 2007)

There seems to be a belief among many safety researchers that the more variables in an APM the better the model. Some researchers have even reported models containing variables with highly insignificant parameters based on the belief that such variables would still improve model prediction. Explanatory variables that have statistically significant model parameters, on the other hand, contribute to the explanation of the variability of the accident data, and their inclusion in the model, therefore, improves its fit to this data.

If an APM is to be used for studying the safety of the particular set of locations used to develop it, then a more accurate study would result by using a model that fits the accident data as closely as possible. This best-fit model is achieved by including all the available statistically significant explanatory variables. (Sawalha Z. , 2003)

2.5 Crash as a Bernoulli trial

A crash is, in theory, the result of a Bernoulli trial (Lisa, K.S., & B.David et al., 2005). Whenever vehicles enter highway segment, an intersection or any other type of entity (a trial) of a given transportation network, each value may result in an outcome that may be classified as a success (crash) or failure (no crash). (WILLIAMS, 2009)

If there are N independent trials (vehicles passing through a road segment, an intersection, etc.) that give rise a Bernoulli distribution with the probability of success (crashes) = p and q= (1-p) is the probability of failure (no crash). The appropriate probability model that accounts for a series of Bernoulli trials is known as the binomial distribution (Miaou, Shankar et al., & Lord et al., 1994, 1997, 2005), and is given as:

$$P(Y = n) = \binom{N}{n} P^n (1 - P)^{N-n} \quad (5)$$

Where Y is the random variable that records the number of success, n=0, 1, 2…N. the mean and variance of the binomial distribution are E(Y) =Np and Var(Y) =Npq respectively.

For typical motor vehicle crashes where the event has a very low probability of the occurrence and a large number of trials exists, the binomial distribution is approximated by a Poison

distribution. Under the binomial distribution with parameters N and p, let $p = \frac{\lambda}{N}$, so that a large sample size N will be offset by the diminution of p to produce a constant mean number of events λ for all values of p. then as N→∞

$$P(Y = n) = \binom{N}{n} (\frac{\lambda}{N})^n \left(1 - \frac{\lambda}{N}\right)^{N-p} \cong \frac{\lambda^n}{n!} e^{-\lambda} \quad (6)$$

Where, n=10, 1, 2…N and λ is the mean of Poisson distribution (Miaou, Shankar et al., & Lord et al., 1994, 1997, 2005)

The mean or the expected value of the Poisson distribution Y is assumed to be equal to its variance. That is, E (Yi) = Var (Yi) =λ

Where E (Yi) is the expected number of crashes and Var (Yi) is the variance of observed number of crashes. For a given set of explanatory variables, the expected number of crashes (E (Yi) =λ) can be estimated using the formulation, $\ln \lambda = \beta x_i$

Where x is a vector of explanatory variables and β is a vector of parameters to be estimated.

However, as (WILLIAMS, 2009) mentioned, in some cases, this method has limitations in applying to real-world data due to the assumption of equal mean and variance. When the data is over-dispersed or under-dispersed (i.e. variance is greater or less than the mean), the use of Poisson regression models to make probabilistic statements about the occurrences of vehicle crashes will overstate or understate the likelihood of the crashes of the road (Miaou, Shankar et al., & Lord et al., 1994, 1997, 2005).

Many previous studies have found that crash data tend to over-dispersed in many situations with the variance being significantly higher than the mean. In such cases, any inferences made based on Poisson model estimation may lead to wrong conclusions. As a result of this, many researchers recommend using alternative methods in analyzing crash data, when the data is over-dispersed (WILLIAMS, 2009).

2.6 Road Traffic Accident Classification

There is no definite and consistent classification method of road traffic accidents worldwide. Some countries keep only simple records classifying accident into total serious (heavy)

injuries, and minor (light) injuries or as total injury and property damage only. Also as indicated by (Hobbs, 1979), the comparison of accident statistics between countries is made difficult because common definitions are not used.

For example, death is defined differently in different countries. Death within 30 days in Britain, at the scene in Portugal, within 24 hours in Spain, within 6 days in France, 7 days in Italy and within a year in the USA. Among these, the definition within 30 days is mainly accepted and the case is true in Ethiopia. Therefore, accidents are classified according to the severity of the accident emphasizing whether a person is killed or injured into fatal, serious, and slight and damage only accidents. (Bitew Mebratu, 2002) (Tulu, 2015).

(Tulu, 2015) and (Bitew Mebratu, 2002), described Classification of crashes based on severity in Ethiopia as:

- Fatal crash or Killed: 'A human casualty who dies within 30 days after the collision due to injuries received in the crash. (Tulu, 2015) (Bitew Mebratu, 2002).
- Serious injury: seriously injured is a person hospitalized, other than for observation, for more than 24 hours (Tulu, 2015) (Bitew Mebratu, 2002).
- Slight injury is a person hospitalized for less than 24 hours. (Tulu, 2015)
- Property damage: is non-injury crashes (property damage e.g. roadside objects, vehicle etc.)

2.7 Road traffic in Ethiopia

Most of the road deaths in developing countries involve vulnerable road users such as pedestrians and cyclists. In Ethiopia, (Mekonnen, 2007) reported that, pedestrian injuries account for 84% of all road traffic fatalities compared with 32% in Britain and 15% in the United States of America. In contrary, in the heavily motorized countries, drivers and passengers account for the majority of road deaths involving children (Bunn, F.T., & al., 2003). Similarly, (Mekonnen, 2007) quoted that, RTA in Ethiopia is a serious problem. The RTA death rate is estimated to be 130 per 10,000 vehicles. Of the total victims of RTA who lost their lives, over half are pedestrians, out of whom 30% are children.

Based on a five-year average record, (Mekonnen, 2007) found as, 81% are caused due to drivers error, 5% due to a vehicle defect, 4% due to pedestrian error, 1% due to road defects and 9% due to other problems in Ethiopia. (Mekonnen, 2007)

In addition to this (Segni G., 2007) added another responsible reason of RTA occurrences in Ethiopia like driving without respecting right-hand rule, failure to give way for vehicles and pedestrians, overtaking in snaky horizontal curves, following too close to the vehicle in front, improper turning, and speeding. These causes contribute to 73% of the total accident in the year 2004/05 in Ethiopia but the other possible reasons accounted for less than 27%.

2.7.1 Road traffic accident reporting system in Ethiopia

As stated by (UN, 2009), similar to most countries of the world, police is responsible for traffic accident investigation and reporting in Ethiopia. According to the Ethiopian transport regulation, a driver of a vehicle involved in a road accident shall notify the nearest police station immediately if the accident involves personal injury and within twenty-four hours if it involves property damage only. According to the regulation, all accidents are reportable. However it is not practical, Because of this, the reporting of nonfatal accidents is uncertain.

The accident is recorded in a daily report book at a local police station or traffic office. Periodic summaries of aggregate road accident records are made and sent to the immediate higher police department.

(Segni G., 2007), reported considering the quality of daily accident reporting format. The content of the road accident reporting, as it exists now, misses relevant details of an accident report required for any road safety improvement works. The reporting form, in the daily report book, is not designed to include details of each vehicle and road user involved in an accident. The report, further, does not contain details of the road section and precise location of an accident. The location of an accident is usually reported broadly by “Kebelle and Wereda” or the name of the surroundings.

Besides, The terminology of accident details does not have a uniform definition even among the staff members at a police station. In addition to the indicated limitations of accident reporting, there is no established system of computerized accident data bank to store detailed information on individual road traffic accidents occurring in the country.

Moreover, there is no system of periodic road traffic accident analysis and dissemination system to give information on road traffic accident trends, specific accident problems so that stakeholders are aware and aim to improve the situation.

2.7.2 Factors contributing to accident

Road traffic crash results from a combination of factors related to the components of the system including roads, the setting, vehicles and road users, and the way they interact. Some factors contribute to the occurrence of a collision and are therefore part of crash causation. Other factors aggravate the effects of the collision and thus contribute to trauma severity. Some factors may not appear to be directly related to road traffic injuries. Some causes are immediate, but they may be underpinned by medium-term and long-term structural causes. However, the contributing factors in each set of circumstances generally fall into four components of the road traffic system: Lisa, K.S., & B.David et al., (2005)

- Road users' errors or human-related factors,
- Vehicle defects,
- Deficiencies in the road and its environment.

In this study, among different contributing crash factors, roadway related and environmental related factors were selected for modeling and further data reduction techniques were applied based on their significance as described in the foregoing section.

2.7.2.1 Roadway related factors

Since the entire process of road transport is conducted on roads, the quality, size and engineering characteristics of the roads will have a considerable contribution to the increase or decrease of RTA risks. WHO (2004) supports this idea by saying that, the road network has an effect on crash risk because it determines how road users perceive their environment and deliver instructions for road users, through signs and traffic panels, on what they should be doing. Many traffic management and road safety engineering measures work through their influence on human behavior. Some variables regarding the road related causes of RTA are discussed as to below.

2.7.2.2 Road Environment

Road environments have impacts on occurrences of road traffic accidents. In developed countries, there are continuous efforts to meet the safety standards of roads through safety audit during the planning, designing, and operation stage. (Terje A., 1998) Indicates that in Africa road network is mounting fast, preservation standards have started improving lately.

Berhanu (2000) reports that in Ethiopia, the police have limited road and traffic engineering skill in general and thus they underestimate the contribution of roads and environments to traffic accidents and especially they lack training on the subject area.

2.7.2.2.1 Roadway Characteristics.

The roadway's conditions like the quality of pavements, shoulders, traffic control devices and intersections, can be a factor in a crash. Fewer traffic control devices and complex intersections with excessive signage lead to confusion. Highways must be designed for adequate sight distance for designed speed for the drivers to have sufficient perception-reaction time. The Traffic signs and signals should provide enough time for decision sight distance when the signal changes from green to red.

Another important factor is the frictional force between the pavement and tires. If the tires lose contact with the pavement then the vehicle starts fishtailing. (Lisa, K.S., & B.David et al., 2005)

2.7.2.2.2 Environmental related factors

The climatic and environmental conditions can also be a factor in transportation crashes. Supporting this idea (Lisa, K.S., & B.David et al., 2005); (Alister C. OBE and B. Simon, 2011) argued that, Weather on roads can contribute to crashes: for example wet pavement reduces friction and flowing or standing water greater than 1/8"deep can cause the vehicle to hydroplane. Many several crashes have occurred during conditions of smoke or fog, which can reduce visibility. Vehicles traveling at high rate of speed are unable to see the slowing and or stopped vehicles in front of them which can lead into multi-vehicle pileup. Glare can reduce driver visibility, especially on the east-west roadway during the hours of sunrise and sunset. During foggy conditions glare off of street lights and stop lights can also affect visibility. Wind gusts can affect vehicle stability. Slippery road (due to weather), deposit on the road, animal or object in the carriageway, poor or defective road surface, Inadequate or masked signs or road markings are also responsible for the disaster caused by environmental characteristics to RTA.

2.8 Key outcomes from the literature review

The key outcomes from literature review are:-

- Deterministic models did not show consistent results regarding the relationships between crashes and the geometric elements.

- Discrete choice models showed great potential in obtaining the true models of crashes

- A lot of improvements has been made in the previous research of modeling crashes. The modeling techniques shifted from conventional regression to stochastic regression models.

- The initial research emphasized the relationships between highway geometric variables and crashes, while current research focuses more on exploring the relationships between traffic variables and crashes under a certain geometric characteristics

- Various traffic and geometric variables were shown to have significant influences on the occurrences of crashes.

- The relationship between the crash rate and traffic volume presented a "U-shaped.

- Different measures of evaluation were used including coefficient of determination, log likelihood, and AIC.

- Various traffic and geometric variables were shown to have significant influences on the occurrences of crashes

- Many researchers performed tests on the independent variables before bringing them into modeling.

CHAPTER 3 METHODOLOGY OF THE STUDY

This study analyzed a four year vehicle crash data collected in Yeka sub-city from 2011 to 2014 G.C. In the model building part of the methodology, factors such as geometric and environmental factors, and traffic factors were used in SPSS 23 to come up with the factors that influence most to the crash severity. A multinomial logistic regression model was built by using SPSS 23 in this study. Logistic regression measures the relationship between a categorical dependent variable and one or more independent variables by using probability scores as the predicted values of the dependent variable. Developing and assessing a model requires some data and information, so the required information (reviewing the previous literature) and the required data was collected and described in detail below.

In this study the methodology involves the following steps:

a) Reviewing previous literature
b) Identification of Study Area
c) Data source and data collection
d) Reduction of Dummy/design variables
e) Selection of Variables for Analysis and Modelling
f) Choice of model form and Traffic Crash Model Development
g) Assessment of Model Performance
h) Model Result testing

3.1 Reviewing previous literature

In chapter two (Literature review), the previous studies were summarized; and from this literature review modeling techniques, selection of variables, model performance measures and model result testing procedures were identified. Among different contributing crash factors roadway related and environmental related factors were selected for modeling and further data reduction techniques were applied based on their significance. Discrete choice models were used in this study (as described in the literature review, when researchers began to realize that the characteristics of crashes are discrete, sporadic, and random) to describe the occurrence of crash severity.

3.2 Identification of Study Area

As study area Yeka sub-city (one of the sub-city of Addis Ababa out of ten sub-cities which located in North-East part) was selected. The selection of this sub-city was based on convenience of transportation. So the selection technique applied to this study was non-probability sampling (also known as non-random sampling) for the purpose of cost-effectiveness and time-effectiveness. From different non-probability sampling approach, convenience sampling was chosen in order to collect a sample convenient to the researcher.

3.3 Data Source & Data Collection

Primary and secondary data sources were the best sources for this study. Primary data sources such as basic field survey was considered to collect the actual traffic flow; whereas the secondary data were collected from Addis Ababa police commission which was Traffic accident data; from Yeka sub-city police commission which was Traffic accident data, and also from the city government of Addis Ababa road and transport bureau/Ministry of transport which was a document on Accident prediction system in the country (but no any document regarding on it), Ethiopia. The collection and preparation of accident and traffic volume data was the most time-consuming task in this study.

Care was taken in collecting the required data since the quality of the data have an effect on accident model development. As (Tulu, 2015) stated, the quality of the data collected in a particular study has a large effect on the accident prediction model development. The major factor concerned in accident prediction studies is the under-reporting of accidents. Taking account of this factor is difficult, and as (Turner, 1996) reported, not many researchers have attempted to do on this.

Good quality data has the potential to improve problem identification, analysis, and prioritization of a specific safety problem. (Tulu, 2015) mentioned the six data quality characteristics in which accident data should exactly reflect the characteristics of the statistical object (accuracy), avoid missing values (completeness), be suitable and adequate to the parameters (validity), be up-to-date (timeliness), comprehend all relevant parameters (coverage) and be easily retrieved and processed (availability).

In practice, however, these objectives are sometimes far from their ideal condition.

Timeliness: - this indicates that Information should be available within a specific timeframe to allow for meaningful analysis of the current status of the issue under investigation.

In case of this study, there was no problem of timeliness due to up-to-date data was used.

Accuracy: Information within the database should be correct and reliable in describing the data element it purports to describe. Accuracy is typically enhanced through the practice of conducting consistency checks and validations on the data being entered into the database. This needs an in-depth investigation to conduct it.

Completeness: Information within the database should be complete in terms of all reportable instances of the event/ characteristic being reported and available within the database, and all required data elements within the record should be completed with appropriate responses.

For the case of this study, the required data was completed without missing since the collection of data was on non-behavioral parameters

Consistency/Uniformity: Information collected should be consistent among all reporting jurisdictions with all reporting jurisdictions using the same reporting threshold and reporting the same information on a standard data collection form(s). Ideally, information was reported using nationally accepted and published guidelines and standards.

This was the main problem in this study. There is nationally accepted data collection format but the main problem is on filling the format. For the same environment, different traffic police give a different name for that environment in which the accident occurred. This was solved due to merging different levels based on test of proportion.

Integration: By using common data elements, information in one database should be capable of being linked with information from other databases. An example of integration is the linkage of crash data with roadway inventory data by having a common location element in each database.

Accessibility: Information within the database should be readily available to all eligible users of the information.

In practice, however, these objectives are sometimes far from their ideal condition. This study tried to consider some of the above data quality assessment parameters.

3.3.1 Traffic accident data

Road crash data can be obtained from various sources, including police, hospital, and insurance reports. Currently, the main source of crash data in Ethiopia is the police report database; and in Addis Ababa City Accidents are recorded by the traffic police on daily basis. This study was based on a secondary data obtained from Yeka sub-city Police Commission booklet compiled by traffic police officers. The total observation collected for this study were 5251 total traffic accidents occurred within four years (2011/12-2016/17)

As (Tulu, 2015) mentioned, to enable systematic analysis of the road crash problem, the following information about each crash is required. (Tulu, 2015)

- Where crashes occur: location by map coordinates, road name, road classification, and road Layout and type of traffic control. In Yeka sub-city traffic accident booklet, type of the area that is nearest to the incident are specified as around college, factory, religious, recreation, office complex, hospital, residential, and others; this was also one of the explanatory variable considered in this thesis.
- When crashes occur: - by year, month, and date of the month, day of week and time of day; this is also the case in Yeka sub-city
- Who was involved: - people, vehicles, animals, and roadside objects;
- What was the result of the crash: fatal, personal injury (seriously and slightly), or property damage only. In Yeka sub-city Accidents are classified as fatal, serious injury, light injury and property damage only.
- What were the environmental conditions: light condition, weather, and pavement surface condition; Light conditions includes daylight, dark hour with good, street, dark hour with poor street light, dark hour with no street light; and Weather condition includes fine, mist/fog, cloudy, light rain, heavy rain, hot, cold wind, other

As explained above and shown on figure 3-1, detailed collection regarding on-road crashes includes the following attributes: date, day, and time of crash; vehicle type, vehicle service year, vehicle type, vehicle ownership, vehicles deficiencies; driver sex, age and education; weather, road, and illumination conditions; pedestrian crash type, collision type; degree of crash severity; pedestrian sex and age; and location of crash etc.

It is difficult to achieve comprehensive traffic accident data collection in this sub-city due to high collection costs and limited research time. To manage time and cost constraints this

research only considers 13 road crash attributes/parameters out of more than 25 attributes/parameters in the booklet or accident record sheet as shown in table 3-1.

Table 3-1 Collected Traffic Accident Parameters for the study

All Parameters in the booklet (accident record sheet)	Collected Parameters for this study
crash in the days of the week, Crash hour, driver's age, drivers' sex, driver's educational background, based on Driver's relation with vehicles, driver's experience, driver's license rank, vehicle service year, vehicle type, vehicle ownership, vehicles deficiencies, land use, road type, road geometry, road condition, intersection type, pavement type, lighting condition, weather condition, defendant vehicle maneuvering condition, collision type, Type of crash, Type of severity	crash in the days of the week, Crash hour, and land use, road type, road geometry, road condition, intersection type, pavement type, lighting condition, weather condition, defendant vehicle maneuvering condition, collision type, Type of crash, Type of severity

To collect the required data and stored in Excel sheets, a sufficient number of enumerators were deployed. The data were collected from a hard copy booklet forms and transferred into Excel spreadsheets with the help of the prepared drop down template for each attribute. The prepared drop down template was a supporter for a data collectors to have the same similar choices for each attribute/parameters and also to have the same spelling on the Excel sheet which was further important during data coding and filtering.

For example, figure 3-1shows prepared drop-down template for attribute of illumination conditions that indicates there are only seven choices (daylight, sunset, sunrise, a night with good road light, a night with poor road light, a night without road light and others) to be selected by the data collector. The cell only accepts among the prepared choices in the drop-down template, and all the data were collected in this way.

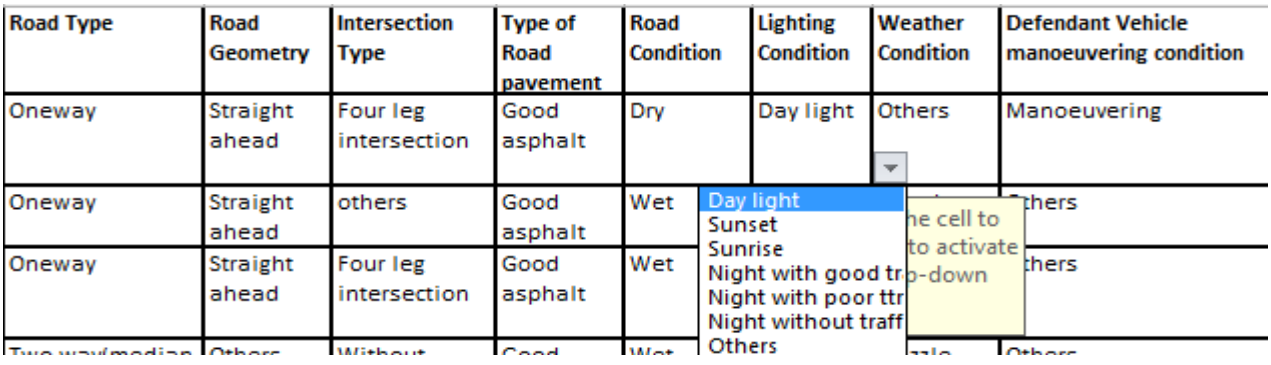


Figure 3-1 Sample drop-down format for lighting condition (illumination conditions)

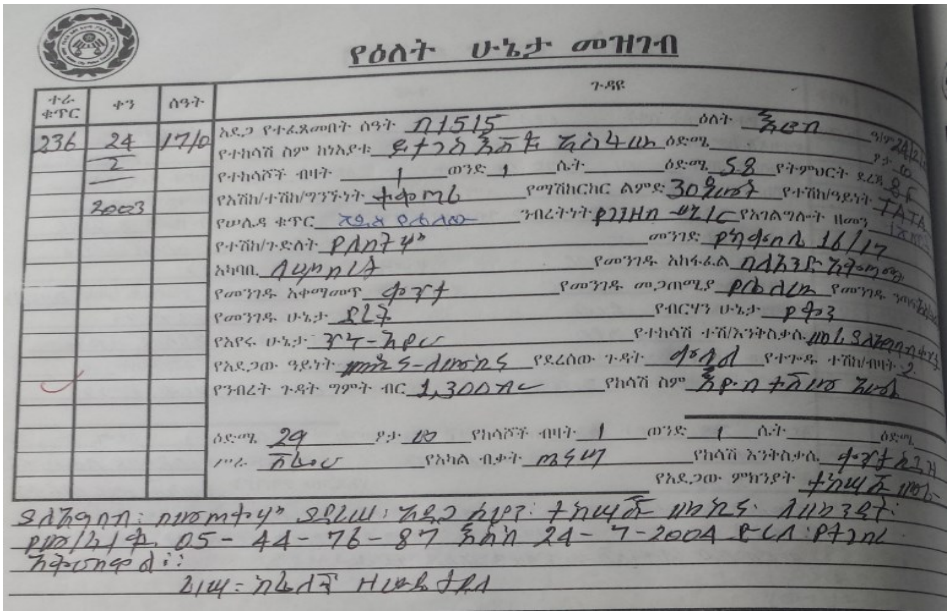


Figure 3-2 Sample of daily recording traffic accident booklet

3.3.2 Traffic volume data

Traffic Volume Data Collection and projections of traffic volumes are basic requirements for developing and assessing accident prediction models. The most common method of collecting traffic volume data (manual method) was used in this study by assigning a person to record traffic as it passes. Among the three methods of the Manual count, recording data onto tally sheets is the simplest means of conducting manual counts which were used in this study. The data were recorded with a tick mark on a pre-prepared field form. A watch or stopwatch was used to measure the desired count interval (15 minutes interval was used).

ERA vehicle classification and a 3-hour traffic volume data were collected on the selected section as shown in Appendix B; and hourly, daily, and monthly expansion factors were used to determine AADT. Hourly expansion factors (HEF) are used to expand counts of durations shorter than 24 hours to 24-hour volumes by multiplying the hourly volume for each hour during the count period by the HEF for that hour and finding the mean of these products. Daily expansion factors (DEF) are used to determine weekly volumes from counts of 24-hour duration by multiplying the 24-hour volume by the DEF. The AADT for a given year may be obtained from the ADT for a given month by multiplying this volume by the monthly expansion factor (MEF). Collected traffic volumes data for selected sections are shown in Appendix E.

AADT using adjustment factor were mathematically calculated using the following formula:

$$AADT = V \times f24 \times fd \times fs \quad (7)$$

Where

- AADT→Average Annual Daily Traffic
- V →Average Hourly Traffic
- f_{24} →Expansion Factor for 24 Hours
- f_d →Expansion Factor for Day Of The Week
- f_s →Expansion Factor for Month of a Year

3.4 Identification of Response and Explanatory Variables in the study

As described in the literature review, Highway geometric characteristics, traffic variables, and other contributing variables have been selected in stochastic regression models for study. Section 2.4.6 describes the types of variables that may be used in statistical analysis like dependent Vs Independent, Discrete Vs Continuous, Qualitative Vs Quantitative, Univariate Vs Multivariate etc. The following sections describe the data sets utilized in this study.

The dependent variable used in this study was crash severity which can be stratified as Property Damage only, slight injury and serious injury, and the coding or replacing system for analyzing in SPSS software is shown in table 3-2 below. A multinomial variable was created for crash severity and used as the dependent variable. The percentage of serious injury crash (SI) is 8.68%, the percentage of slight injury crash (SLI) is 8.68%, and the percentage of property-damage-only (PDO) crash is 82.63% (Table 3-2).

Table 3-2 Coded Variables related to Types of Severity for analyzing in SPSS

Variables	Serious Injury	Slightly Injury	Property Damage only
Coding System	SI	SLI	PDO
Observation	456	456	4339
Percentage	8.68%	8.68%	82.63%

An independent variable sometimes called an experimental or predictor variable, is a variable that is being manipulated in an experiment in order to observe the effect on a dependent variable sometimes called an outcome variable. Independent variables in this study include three sets of variable (variables related to Environmental conditions, Variables related to road features and others).

3.4.1 Variables related to Environmental conditions

Variables related to environmental conditions include weather condition related variables, road surface condition, land use pattern and lighting condition.

Weather Condition related variables: - The variables related to Weather Condition involved in the accident-severity obtained from the crash data collecting booklet are: Cloudy, Chilly, Drizzle, Hazy, Good Weather, Cold Weather, and Heavy rain; and the coding or replacing system for analyzing in SPSS software is shown in the table 3-3 below. After applying tests of proportion (which is described below), only significant variables were selected for further analysis; and insignificant variables were merged into significant variables. As a result weather condition was treated as a binary variable by classifying as Good weather and Bad weather. Others which includes heavy rain, cold weather, hazy, drizzle, chilly and cloudy were insignificant and merged with bad weather (new level for insignificant variables).

Based on the four year data, 94.78% happen on good weather condition and the remaining happen during bad weather condition. In terms of accident severity, 17.56% involve personal injury during good weather condition and 13.87% cause personal injury during bad weather condition. Detailed statistics are shown in Table 3-3

Table 3-3Coded Variables related to Weather Condition

Variables	Cloudy	Chilly	Drizzle	Hazy	Good weather	Cold Weather	Heavy Rain
Coding System	C	CH	D	HA	GW	CW	HR
Observation	113	7	67	19	4977	5	63
Percentage	2.15%	0.13%	1.28%	0.36%	94.78	0.1%	1.2%

Road surface condition related variables: - The variables related to Road surface condition involved in accident-severity include dry road surface condition and wet road surface condition. Ninety percent (90%) of the total crash occurred in dry road surface as against 10% which occurred in wet road surface condition. Most personal injury crashes occurred on dry road surface condition.

Land use pattern:- In the daily traffic crash data recording sheet, land use pattern has so many levels but during data coding it was categorized in to eight attributes or levels which

include around school, around office, around market, around church, around industry, around factory, around entertainment, around residential area and others; but It is more convenient to have as few levels of variables as possible in order to simplify the model interpretation and to avoid difficulties during running the model processing in SPSS. As a result based on test of proportion, The following categories or levels of land use have been used: Around Factory or Industry (AFI), Around Institutional Areas or Populated non-business District (ANBD), Around Commercial Zones or Business District (ABD), Around Residential Areas (AR) and others or none which include places other than the above lists like around roundabout and the like. The number of accidents is higher in populated districts like populated-business districts (entertainment, market), and populated non-business districts (school, hospital, church, organizational office etc.) compared to residential area. It was observed that over two-third (70%) of all road traffic accident in yeka sub-city occurred in the central business districts.

Illumination Condition related variables: - The variables related to Illumination/Lighting Condition involved in the accident obtained from the crash data collecting booklet are: Daytime with sufficient daylight, Twilight, Sun rising, Night with sufficient light, Night with insufficient light, Night without light; and the coding or replacing system for analyzing in SPSS software is shown in the table 3-4 below. The four year data highlighted that 45% of accident severity took place during daylight.

Furthermore, illumination/lighting condition was treated as a binary variable by classifying as Day Light and Darkness which include sunset, a night with poor road light, a night without road light and night with a good light. Based on the year data, 72.0% happen on daylight condition and the remaining happen during other lighting conditions. In terms of accident severity, 16.0% involve personal injury during daylight condition and 20.6% cause personal injury during other lighting conditions. Detailed statistics are shown in Table 3-4

Table 3-4 Coded Variables related to Lighting Condition

Variables	Day Light	Sunrise	Sunset	A night with good road light	A night with Poor road light	Night without Road Light
Coding System	DL	SR	SS	NGRL	NPRL	NWRL
Observation	3770	256	298	607	233	87
Percentage	72.0%	4.8%	5.7%	11.56%	4.43%	1.67%

3.4.2 Variables related to road features

Road Junction (Intersection Type) related variables: - The variables related to road junction involved in the accident obtained from the crash data collecting booklet are: Midblock/without Intersection (WOI), Y-junction(YS), T-junction(TS), Roundabout(R), Four-leg Junction(SQ), Five-leg Junction(FLJ), and Rail Crossing(RC) and others (O); and the coding or replacing system for analyzing in SPSS software is indicated in the abbreviated way. Furthermore, road junction or intersection type related variables were merged and treated as a binary variable by classifying as without intersection and with intersection (which include Roundabout, square, T-shape, X-shape, Y-shape, Five-leg junction, Rail crossing and others)

Lanes/Medians (Road Type) related variables: - The variables related to lanes/medians involved in the accident obtained from the crash data collecting booklet are: One-way, Undivided two-way, divided two-way (Median), Two-way divided with solid lines road marking, Two-way divided with broken lines road marking; and the coding or replacing system for analyzing in SPSS software is shown in table 3-5 below.

Table 3-5 Coded Variables related to Road Type

Variables	One-way	Undivided Two-way	Island Separated	Two-way Separated with broken line	Two-way Separated with Solid line
Coding System	OW	TW	IS	SBS	SLS
Observation	1582	1930	1562	101	76
percentage	30.12%	36.75%	29.75%	1.92%	1.45%

Road Alignment (Road Geometry) related variables: - The variables related to road alignment involved in the accident obtained from the crash data collecting booklet are: Tangent road with flat terrain, Tangent road with mild grade and flat terrain, Tangent road with mountainous terrain and escarpment, Tangent road with rolling terrain, Gentle horizontal curve, and Steep grade upward with mountainous terrain; and the coding or replacing system for analyzing in SPSS software is shown in the table 3-6 below. Furthermore, road alignment was treated as a binary variable by classifying as straight ahead (Flat) and non-straight ahead(Rolling) which include straight with slightly sloping, straight with up and down, straight with highly slopping, slightly zigzag, highly zigzag, downward and upward.

Table 3-6 Coded Variables related to Road Alignment/Geometry

Variables	Straight ahead	Straight with up and down	Straight and Slightly Slopping	Straight and Highly Slopping	Slightly Zigzag	Highly Zigzag	Down ward	Upward
Coding System	SA	SUD	SSS	SHS	SZ	HZ	DW	UW

Road Pavement Condition related variables: - The variables related to Road pavement Condition involved in the accident obtained from the crash data collecting booklet are: Good Asphalt Condition, Distressed/Poor Asphalt Condition, Gravel Condition, and Earth Condition; and the coding or replacing system for analyzing in SPSS software is shown in the table 3-7 below.

Table 3-7 Coded Variables related to Road Pavement Condition

Variables	Good Asphalt	Poor Asphalt	Gravel Road	Earth Road
Coding System	GA	PA	GR	ER

3.4.3 Other Variables

Crash related variables: - The variables related to Types of Crash involved in the accident obtained from the crash data collecting booklet are: Overturning (OT), Vehicle to Inert (VTI), Vehicle to Pedestrian (VTP), Vehicle to Vehicle (VTV), Vehicle to Vehicle to Inert (VTVI), Vehicle to Vehicle to Inert to Pedestrian (VTVIP), Vehicle with parked Vehicle (VTS), Vehicle to Vehicle to Parked Vehicle (VTVSV) and Overturning to passenger (OTP); and the coding or replacing system for analyzing in SPSS software is indicated in the abbreviated way.

Based on tests of proportion, these variables were also categorized as vehicle-to-vehicle, vehicle-to-inert, vehicle-to-pedestrian and vehicle-overturning. Insignificant variables were also merged with significant variables(vehicle to vehicle to inert, vehicle to vehicle to passenger and vehicle to stopped vehicle was merged with vehicle-to-vehicle; vehicle to inert to pedestrian was merged with vehicle-to-inert; vehicle to pedestrian to vehicle and vehicle to pedestrian to inert was merged with vehicle-to-pedestrian.

Defendant Vehicle Maneuvering Condition related variables: - The variables related to Defendant Vehicle Maneuvering Condition involved in the accident obtained from the crash

data collecting booklet are: Backward movement (BM), During Stopping(DS), Entrance to Diverging road (EDR), Local Exit (LE), Entrance to Junction /merging road (EJR), U-turning, Entrance to square road (ESQR), Left turning (LT), right turning (RT), Straight through (ST), Maneuvering (MA) and Undefined (UD); and the coding or replacing system for analyzing in SPSS software is indicated in the abbreviated way.

Almost all maneuvering related variables were significant based on test of proportion which was not used here for further reduction of this variable. Researchers classify maneuvering conditions as entering or leaving intersection, making turns other than at an intersection, straight through and parking. In this study, defendant vehicle maneuvering condition was classified as straight through, maneuvering (which include maneuvering, backward movement, and parking), entering or leaving intersection (which include entrance to diverging road, entrance to junction road, entrance to square road); making turns other than intersection (which include left turn, right turn, U-turn, local exit); this classification was done considering researchers classification and data collecting booklet sheet.

Crash Hour related variables: - The variables related to Crash Hour involved in the accident obtained from the crash data collecting booklet are the actual time in which the crash happened but during collected data processing a 3-hour classification was used to use the prepared drop-down template and for simplification. It includes (Mid-night, Dark, Late morning, Morning, Noon, Afternoon, Evening and Night); and the coding or replacing system for analyzing in SPSS software is shown in table 3-8 below

Table 3-8 Coded Variables related to Crash Hour

Variables	Midnight	Dark	Late Morning	Morning	Noon	Afternoon	Evening	Night	Undefined
Coding System	MN	DA	LM	MO	NO	AN	E	N	UD

Days of the week related variables: - The variables related to the day of the week involved in the accident obtained from the crash data collecting booklet are: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday; and the coding or replacing system for analyzing in SPSS software is shown in table 3-9 below.

Table 3-9 Coded Variables related to days of the week

Variables	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday

Coding System	M	T	W	TH	F	S	SU

3.5 Significant Variables for Regression Modeling

In literature review section different methods were described for selection of significant variables, But in this study, using all of these methods is believed unnecessary since the relationship of each variable to the existence of an injury in an accident can be easily identified by the chi-squared test of independence. Thus, a chi-squared test was used to identify and select explanatory variables which have a statistically significant association with the dependent variable. There are five steps to conduct this test or Statistical software (SPSS) makes this determination much easier:-

- **Step 1:** Formulate the hypotheses.
- **Step 2:** Specify the expected values for each cell of the table (when the null hypothesis is true).
- **Step 3:** To see if the data give convincing evidence against the null hypothesis, compare the observed counts from the sample with the expected counts, assuming H_0 is true.
- **Step 4:** Compute the test statistic.
- **Step 5:** Decide if chi-square is statistically significant

Sample Cross tabulation:

Step 1: Formulate the hypotheses

- ✓ **Null Hypothesis: H_0:** There is no significant association between types of road geometry and occurrence of an accident
- ✓ **Alternative Hypothesis: H_a:** There is a significant association between types of road geometry and occurrence of an accident

Step 2: Specify the expected values for each cell of the table (when the null hypothesis is true). The expected values specify what the values of each cell of the table would be if there was no association between the two variables. The formula for computing the expected values requires the sample size, the row totals, and the column totals.

$$expected\ coun = \frac{row\ total * column\ total}{table\ total} \quad (8)$$

Step 3: To see if the data give convincing evidence against the null hypothesis, compare the observed counts from the sample with the expected counts, assuming H_0 is true. The observed values are the actual counts computed from the sample. Statistical software will compute both the expected and observed counts for each cell when conducting a chi-square test. The table below shows the table that SPSS creates for the two variables. In each cell, the expected and observed value is shown. (Table 3-10).

Table 3-10 Sample observed counts of road type Vs types of severity

Road Type	**Accident severity type**				
	Fatal	Serious Injury	Slight Injury	Property Damage Only	Total Accident
OW	4	9	155	330	1581
TW	10	240	183	1490	1930
IS	4	93	103	1361	1562
SS	0	8	5	63	76
BLS	0	6	7	87	100
Total	18	436	454	4332	5251

Crosstabs

[DataSet1] C:\Users\HP\Desktop\SPSS data\Coded data\Untitled2.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
RT * Type_of_severity	5251	100.0%	0	0.0%	5251	100.0%

RT * Type_of_severity Crosstabulation

			Type_of_severity					
			1	FT	PDO	SI	SLI	Total
RT	1	Count	0	0	1	0	1	2
		Expected Count	.0	.0	1.6	.2	.2	2.0
	BLS	Count	0	0	87	6	7	100
		Expected Count	.2	.3	82.5	8.3	8.6	100.0
	IS	Count	1	4	1361	93	103	1562
		Expected Count	3.3	5.4	1288.6	129.7	135.1	1562.0
	OW	Count	3	4	1330	89	155	1581
		Expected Count	3.3	5.4	1304.3	131.3	136.7	1581.0
	SLS	Count	0	0	63	8	5	76
		Expected Count	.2	.3	62.7	6.3	6.6	76.0
	TW	Count	7	10	1490	240	183	1930
		Expected Count	4.0	6.6	1592.2	160.3	166.9	1930.0
Total		Count	11	18	4332	436	454	5251
		Expected Count	11.0	18.0	4332.0	436.0	454.0	5251.0

Figure 3-3 Sample crosstab output from SPSS statistical software

Step 4: Compute the test statistic

The chi-square statistic compares the observed values to the expected values. This test statistic is used to determine whether the difference between the observed and expected values is statistically significant. The formula for the statistics is:-

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} \quad (9)$$

Table 3-11 Sample Crosstab statistics

Road Type		Accident severity type				
		Fatal	Serious Injury	Slight Injury	Property Damage Only	Total Accident
OW	observed	4	89	155	1330	1581

	expected	5.4	131.3	136.7	1304.3	1581
TW	observed	10	240	183	1490	1930
	expected	6	160.3	166.9	1592.2	1930
IS	observed	4	93	103	1361	1562
	expected	5.4	12.7	135.1	1288.6	1562
SLS	observed	0	8	5	63	76
	expected	0.3	6.3	6.6	62.7	76
BLS	observed	0	6	7	87	100
	expected	0.3	8.3	8.6	82.5	100
Total		18	436	454	4332	5251

$$\chi^2 = \frac{(4-5.4)^2}{5.4} + \frac{(89-131.3)^2}{5.4131.3} + \frac{(155-136.7)^2}{136.7} + \frac{(1330-1304.3)^2}{1304.3} + \frac{(10-6.6)^2}{6.6} + \frac{(240-160.3)^2}{160.3} +$$
$$\frac{(183-166.9)^2}{166.9} + \frac{(1490-1592.2)^2}{1592.2} + \frac{(4-5.4)^2}{5.4} + \frac{(93-129.7)^2}{129.7} + \frac{(103-135.1)^2}{135.1} + \frac{(1361-1288.6)^2}{1288.6} +$$
$$\frac{(0-0.3)^2}{0.3} + \frac{(8-6.3)^2}{6.3} + \frac{(5-6.6)^2}{6.6} + \frac{(63-62.7)^2}{62.7} + \frac{(0-0.3)^2}{0.3} + \frac{(6-8.3)^2}{8.3} + \frac{(7-8.6)^2}{8.6} + \frac{(87-82.5)^2}{82.5} =$$
100.037

The above example shows the observed and expected values for the above sample. If these values are entered into the formula for the chi-square tests statistic, the value obtained is 100.037.

Step 5: Decide if chi-square is statistically significant: - The final step of the chi-square test of significance is to determine if the value of the chi-square test statistic is large enough to reject the null hypothesis. Statistical software makes this determination much easier.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	100.037[a]	20	.000
Likelihood Ratio	97.442	20	.000
Linear-by-Linear Association	27.562	1	.000
N of Valid Cases	5251		

a. 12 cells (40.0%) have expected count less than 5. The minimum expected count is .00.

Figure 3-4 Sample Chi-square test

For the purpose of this analysis, only the Pearson Chi-Square statistic is needed. The p-value for the chi-square statistic is .000, which is smaller than the alpha level of .05. Therefore, there is enough evidence to reject the null hypothesis.

Conclusion: Evidence from the sample shows that there is a significant association between types of road geometry and occurrence of accident.

Significant variables were selected in a similar way as described above using statistical software (SPSS) and for all study variables, the cross-tabulation was listed in Appendix I.

3.5.1 Variables Used and Its Correlation

The collected data were grouped into three sets of variables in which variable set one includes environmental-condition related variables, variable set two includes road feature related variables and others. Also, data on annual average daily traffic were collected and used in the injury-severity model development. The explanatory variables in each of these categories given in Table 3-12 below were incorporated into the model after its correlation test.

Table 3-12 Explanatory variables in each category

Category of explanatory variables	Variable levels	Variable type
Road Characteristic		
Road type	5	Dummy
Road geometry	2	Binary
Pavement type	2	Binary
Intersection type	2	Binary
Environmental characteristic		
Weather condition	2	Binary
Lighting condition	2	Binary
Land use pattern	5	Dummy
Road surface condition	2	Binary
Other Variables		
Days of the week	7	Dummy
Crash type	4	Dummy
Vehicle maneuvering condition	4	Dummy
Annual Average Daily Traffic	Continuous	Continuous

3.6 Traffic Crash-Severity Model Development

Numerous efforts have been devoted to investigate crash severity as related to roadway design features, environmental and traffic conditions. Numbers of accident severities were estimated by using discrete output models. These models relate accident severities to some available explanatory variables. The multinomial logit model is used as the model to estimate significant influences on injury-severities of vehicle crashes at Yeka sub-city.

Various techniques have been applied to the analysis of accident severity data. The statistical methods used by researchers are mainly dependent on the nature of the response variable and various methodological issues associated with the data. The response variable of existing accident severity models is generally either a binary outcome (e.g., injury or non-injury) or a multiple outcome (e.g., fatality, disabling injury, evident injury, possible injury, no injury or property damage). The response variable in this study was the crash severity, which consisted of three levels of crash and it was modeled as a nominal variable. The main objective of this study was to investigate the complex relationships between the crash severity and the predictor variables by using the logistic regression modeling.

Statistical software SPSS 23 was used to estimate a multinomial logistic regression model. In practice, when estimating the model, the model coefficients of the reference group are set to zero. Since 3 levels of severity exist, only (3-1) distinct sets of parameters can be identified and estimated, so severity level equals to property damage only, is set to reference category. A MNL model is estimated as the base model to investigate vehicle crash severities in Yeka sub-city. A significance value of $\alpha = 0.05$ is used to include or reject variables within the model.

In the next sub-section, the logistic regression introduction, the logistic regression equation, and the logistic regression model fit are discussed.

3.6.1 Multinomial Logistic Regression

The multinomial logit is a widely used discrete choice model in various fields. It is suitable for modeling nominal outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the regression variables. According to the literature, since the dependent variable, accident severity, has a discrete nature, discrete choice models are identified as the most suitable approach. Among all the discrete choice models, the multinomial logit model (MNL) is the easiest and most widely used in predicting accident severity. One primary feature of MNL models is that they do not recognize any order in injury levels. This means that the probabilities of property damage, slight injuries, or serious injuries occurring as a result of each weather factor do not follow the same order as the accident severity level.

3.6.1.1 Basic Model Structure

The equation representing the relationship between the response variable and the explanatory variable is called the structural model. The formula of the logit model was first introduced by Luce in 1959 from the assumption of probability of choice (Train, 2009). Following this, there have been a number of contributions to the development of this model (McFadden, 1974, 2001).

The specific form of the logistic regression model is:

$$\pi(x) = E(Y/x) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \quad (10)$$

The transformation of the $\pi(x)$logistic function is known as the logit transformation:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_i x_i \quad (11)$$

where: E(Y)is the dependent variable, in this case, the expected number of accidents occurred for a given time period

β_i are parameters to be estimated by the model

$x_i' s$ are the independent variables

The estimation of the model parameters can be carried out through the method of maximum likelihood.

3.7 Model Estimation Technique

The Maximum Likelihood Estimation (MLE) were used to estimate the discrete output models.

3.8 Assessment of Model Performance

It is important to note that an objective assessment of the predictive performance of a particular model can be made through the evaluation of several Goodness-of-fit (GOF) criteria. In this study the following performance assessing criteria were used for the model selection and measurement of model performance, and described below

- Log-likelihood ratio.
- Akaike Information Criteria
- Deviance,
- Pearson's chi-square,
- Deviance to the degree of freedom ratio
- Pearson's chi-square to degree of freedom ratio
- Error Rates

1. Likelihood ratio (LR) statistic: - it describes the difference between log-likelihood statistics for the full and reduced model. Its value is given by the following formula:

$$Ratio\ of\ likelihood = -2ln\hat{L}_1 - \left(-2ln\hat{L}_2\right) = -2\ln\left(\frac{\hat{L}_1}{\hat{L}_2}\right) \quad (12)$$

Where:

The degrees of freedom for LR statistics are equal to the difference between the numbers of parameters in the two models; $-2ln\hat{L}_1$ and $-2ln\hat{L}_2$ are log likelihood statistics of the two compared models; and $\hat{L}_1$and $\hat{L}_2$ are the maximized likelihood values of the two models.

2. Akaike Information Criteria

Log-likelihood was used in computing Akaike Information Criteria (AIC) and corrected AIC (AICC) for reduced models.

$$AIC = -2\ Log\ L + 2k \quad (13)$$

$$AICC = -2\ Log\ L+2(k+1)n/(n-k-2) \quad (14)$$

Where Log L is the log likelihood;

K is the number of estimated parameters; n is the number of observations.

The smaller the AIC or AICC value, the better the model. As the sample size increases, there is an increasing tendency to accept the more complex model when selecting a model based on AIC.

3. Pearson chi-square:- The Pearson chi-square statistic divided by its degrees of freedom, n – p provides another measure of fit of the model. Asymptotically, this value tends toward

one. Generally, if the Pearson chi-square ratio is between 0.8 and 1.2, this is an indication that the model can be assumed to be appropriate in modeling the data.

4. Mean Deviance: - the deviance of the model containing all the parameters (including the intercept) divided by its degrees of freedom, n - p provides a test for over-dispersion and a measure of fit of the model. Asymptotically, this value tends toward one.

5. Error Rates: - To test how well the models perform, the Error Rates (ER) were computed for selected models. As (Miaou & Lum, 1993) mentioned, ER was defined as the sum of absolute values of the difference between the observed and estimated relative frequencies over the observed relative frequency, and given as shown below.

$$r_k = \frac{|f_k - \hat{f}_k|}{f_k} \tag{15}$$

Where:

r_k is the error rate for k occurrence of events; f_k is the percentile of observations with k occurrence of events among the total data set;

$\hat{f}_k$ is the estimated percentile of observations with k occurrence of events, i.e. relative frequency of k occurrence of events;

$$\hat{f}_k = \sum \hat{p}(y_i = k)/n \tag{16}$$

Where $\hat{p}(y_i = k)$ is the estimated probability of k occurrence of events?

3.9 Model Validation

In statistics, regression validation is the process of deciding whether the numerical results quantifying hypothesized relationships between variables, obtained from regression analysis, are acceptable as descriptions of the data.

In this study model result testing and Goodness-of-fit (GOF) measures were used to conduct external model validation.

3.10 Software Used

To simplify the task of working with regression, there are plenty software's on the market nowadays. SPSS which stands for Statistical Package for the Social Sciences is one of the

famous in computing regression. SPSS Statistics will generate quite a few tables of output for a Generalizer linear regression analysis. Some of them are described below.

1. The first table in the output is the **Model Information** table. This confirms the name of the dependent variable, the probability distribution and the link function.
2. The second table, **Case Processing Summary**, shows how many cases were included in the analysis (the "Included" row) and how many were not included (the "Excluded" row), as well as the percentage of both.
3. The **Categorical Variable Information** table highlights the number and percentage of cases in each group of each independent categorical variable in the analysis.
4. The **Goodness of Fit** table provides many measures that can be used to assess how well the model fits.
5. The **Omnibus Test** table. It is a likelihood ratio test of whether all the independent variables collectively improve the model over the intercept-only model (i.e., with no independent variables added).
6. The Tests of Model Effects table displays the statistical significance of each of the independent variables in the "Sig." column:

The general steps followed in this study are summarized on the flow chart shown below (Figure 3-5).

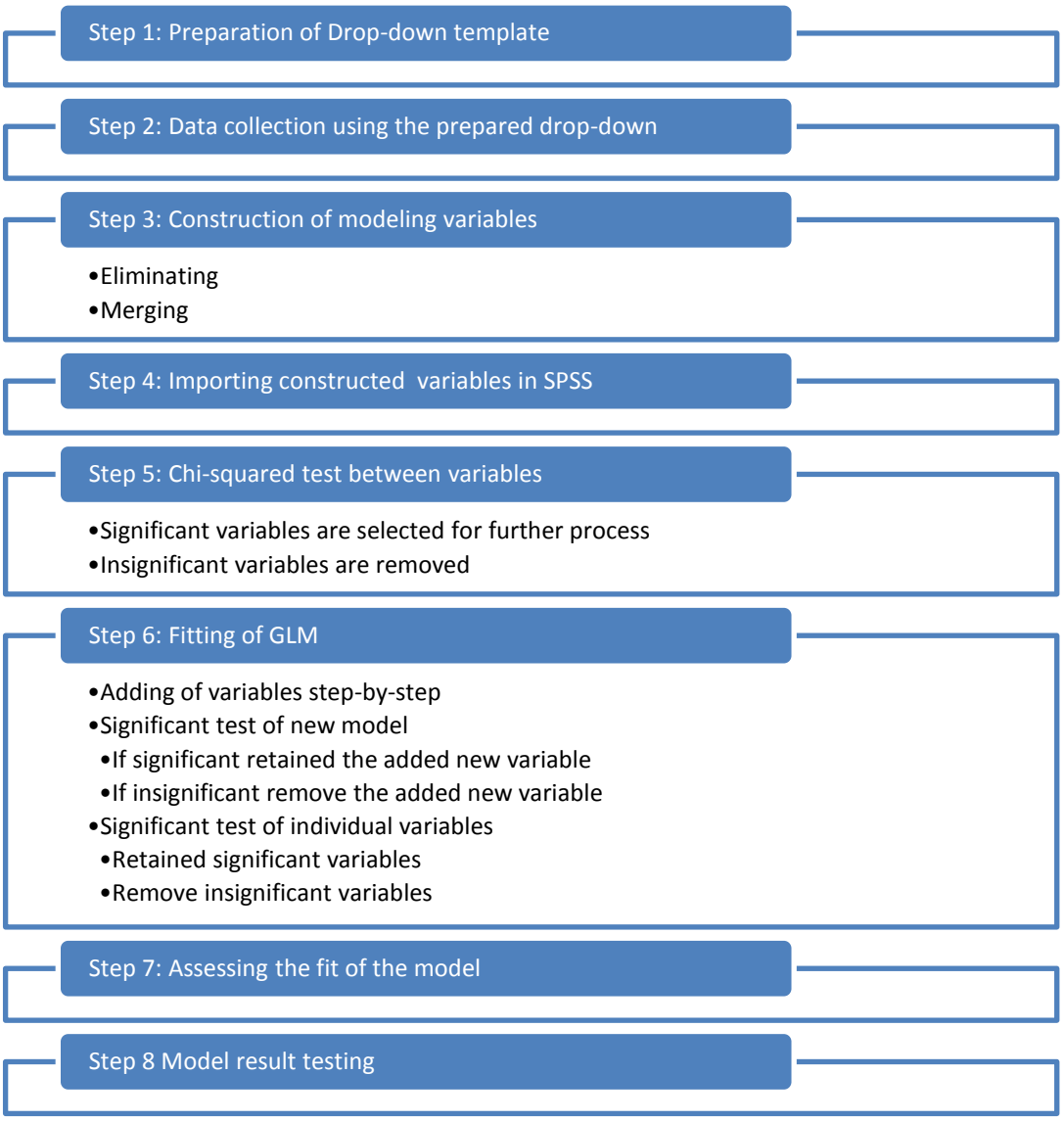


Figure 3-5 Flow chart for methodology of the study

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Results and Interpretation

4.1.1 General Crash Analysis

During the period of 2011 up to 2016G.C., a total of 8109 total accident occurred in Yeka sub-city as shown in table 4-1 for each year. The yeka sub-city alone accounted for more than 12% of all road traffic accidents in Addis Ababa. Table 4-1 presents the distribution or trend of total accident per year for both yeka sub-city and Addis Ababa city. The trend seems linear incremental with 0.943 coefficient of determination as shown in the figure. The four-year data (from 2011- 2014) was used for model development and the remaining two years data was used for model result testing.

Table 4-1 a six-year accident data

Year	2011	2012	2013	2014	2015	2016
Total accident in Yeka	781	1007	1626	1838	2082	2175
Total accident in Addis	11529	15815	17904	17732	20432	

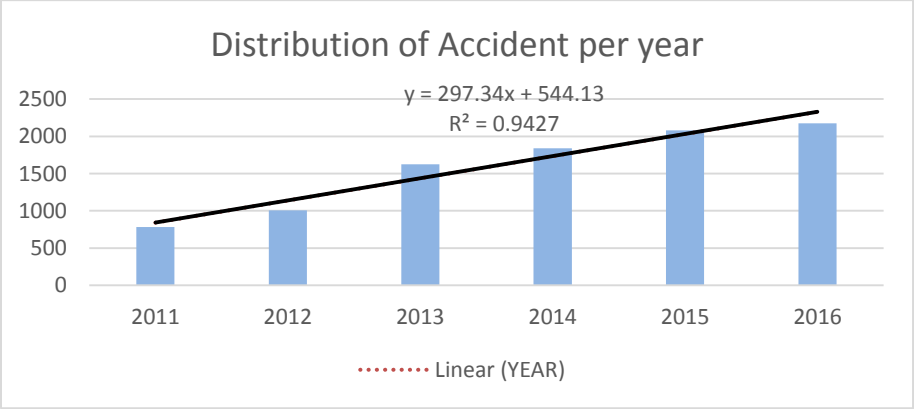


Figure 4-1 Distribution of accident per year

Several road features, environmental and traffic related non-behavioral factors which are considered to have an effect on the occurrence of vehicle related traffic crashes were analyzed by using logistic regression in order to determine the most significant ones. A total of fourteen variables were selected for exploratory analysis to investigate characteristics of predictor

variables and screen out the most influential ones. The multinomial logistic regression model which uses maximum likelihood estimation method was applied to estimate statistically the effects of these variables in contributing to the occurrence of vehicle crash severity levels. Predictor variables were tested at a 95% significance level.

The modelling procedure started with the assessment of the presence of multicollinearity in the dataset. After the multicollinearity checking, all the variables were included in the model for each set of variables, and the backward selection process was used in such manner that only the variables significant at the 0.05 significance level were retained. It is not appropriate for the variable representing factors with strong correlation to exist simultaneously in a model. In consonance with earlier studies, the decision on which variables should be retained in the model was based on two criteria. The first criterion was whether the estimated parameter was significant at the 95% confidence level (p-value less than 5%). The second criterion was whether the addition of the variable to the model cause a significant drop in the deviance ratio at the 95% confidence level. Generally, the basic process was as follows and explained below:

1. Studies of the individual variables (dependent as well as independent)
2. Construction of modeling variables from gathered data
3. Studies of the correlation between pairs of variables
4. Model estimation
5. Assessing the fit of the model.

4.1.2 Studies of dependent as well as independent variables

Road traffic crash data were supplied by the Yeka sub-city Police Commission for the period July 2011 to June 2016 (six years), which was the latest data available at the time. As explained in the methodology part daily crash recording booklet in Yeka sub-city includes the following crash database variables/attributes: time of day, day of week, education, age and gender of drivers, driving experience, driver's relationship with vehicle (employee/owner/other), vehicle service years, vehicle type, vehicle ownership, road type, land use, median and junction types, terrain, pavement type, pavement conditions, illumination, weather conditions, casualty type, and reason for the crash.

It is difficult to achieve comprehensive traffic accident data collection considering all attributes in this sub-city due to collecting costs and limited research time. To manage time and cost constraints this research only considers 14 road crash attributes/parameters out of more than 20 attributes/parameters in the booklet or accident record sheet. These attributes

were selected based on the reason that almost all of them seems non-behavioral factors to the occurrence of traffic crash which was the interest of this study; and as (Turner, 1996) stated that since predicting the occurrence of road accidents is difficult task as accidents are affected by so many factors, one has to resort to drastic oversimplifications.

Traffic crash factors can be classified as behavioral and non-behavioral factors. Behavioral factors include driver and vehicle-related factors like drivers age, education, sex, experience; vehicles type, service year, deficiencies and vehicle ownership; whereas non-behavioral factors include road and environmental characteristics like the location of crash, intersection type, road surface condition, weather condition, lighting condition and the like. So this study was conducted by considering the non-behavioral factors of a traffic accident.

As (J.Garber, 2001) stated that Highway geometric characteristics, traffic variables, and other contributing variables have been selected in stochastic regression models for study. Likewise in this study road and environmental characteristics have been selected for the modelling process. Due to the complexity of the occurrence of crashes which is dependent/target variable, multiple factors were considered and applied in the modeling of crashes. Those factors which were treated as explanatory variables includes road geometry, intersection type (occurrence of intersection), road type (occurrence of median), road surface condition, land use characteristic, and environmental conditions such as lighting and weather conditions; and the statistics of some attributes were described below.

The composition of property damage only and Injury with respect to road alignment occurred on straight ahead section was 82.67% and 17.33% respectively; and the non-straight section contributed to 17.61% of injury and 82.39% of property damage only crashes. As table 4-2 and Figure 4-1 shows that 86.92% occurred on the straight section. These findings are in line with other studies conducted in Ethiopia (Tulu, 2015).

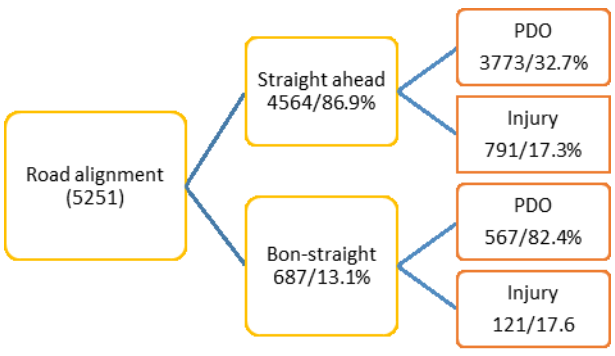


Figure 4-2Sample statistics for Explanatory variables

There is also an association between different land use patterns/types and the occurrence of crashes. The analysis indicates that most of injury and property damage only crashes occurred in and around central business districts (32.39%) and residential areas (27.19%). Populated non-business districts like school, hospital, office and praying rooms are also crash prone areas, second to central business districts (31.08%). Table 4-2 shows that 17.52% injury and 82.48% of property damage only occurred in central business districts; and 18% injury and 82% property damage only in residential areas under consideration of the four year period. This is consistent with the findings of another study which was carried out in Addis Ababa (Schneider, et al., 2008a) as explained by (Segni G., 2007). As (Tulu, 2015).stated that the high occurrence of crashes in these areas may be explained by the complexity of the road environment, mixed traffic and built-up property along these roads that attracts mixed road users with variation across time and location.

The absence of median strips or barriers also has a significant effect in increasing crashes. According to the data, 54.48% of injury crashes occurred on undivided roadways; and median separated roads accounted for 20.34% and 79.66% of injury and property damage only crashes respectively.

The variation in road traffic crashes by time of day may not reflect variations in traffic volumes, and most crashes occur during daylight. As shown in table 4.2 68.89% of crash occurred in the daytime of which 83.05% of property damage only and 16.95% injury crashes which includes both serious and slight injury. So, based on this statistics both the occurrence of accident and the level of severity follow the same pattern in which more than 80% of it occurred during the daytime.

Pavement surface condition is also important in terms of stability of vehicles when a vehicle goes off the road. As (WILLIAMS, 2009) mentioned that Roads with good asphalt are found to have lower crash and severity rate than similar road having poor asphalt. But based on the data 82.59% of property damage only and 17.41% injury accidents occurred on good asphalt (97.35% compared to poor asphalt) as shown in table 4-2 which was not consistent with the findings of another. This may need further an in-depth investigation.

Approximately 31.69% of crashes occurred around within intersections which was not consistent with the previous findings as Williams mentioned that the occurrence of crash increases with an increase in access density.

Table 4-2 observed frequency of explanatory variables

Explanatory variable		**Observed frequency**		**Percentage (%)**		
	Reduced level	Injury	PDO	Injury	PDO	Total
Crash hour	Daytime (DT)	622	3048	16.95	83.05	69.89% DT
	Night time(NT)	290	1291	18.34	81.66	
Crash type	Vehicle-to-inert(VTI)	67	592	10.17	89.83	71.32% VTV
	Vehicle-to-vehicle(VTV)	498	3247	13.3	86.7	
	Vehicle-to-pedestrian(VTP)	328	422	43.73	56.27	14.28% VTP
	Overturning (OT)	19	78	19.59	80.41	
Land use pattern	Business districts (ABD)	286	1346	17.52	82.48	31.08% ABD
	Industry/factory(AFI)	44	162	21.36	78.64	
	Populated non-business district (ANBD)	285	1416	16.75	83.25	32.39
	Residential area (AR)	257	1171	18	82	27.19
	Others (O)	40	244	14.08	85.92	
Lighting condition	Daylight (DL)	702	3615	16.26	83.74	82.21
	Darkness (DN)	210	724	22.48	77.52	
Maneuvering condition	Straight through(ST)	411	2980	12.12	87.88	64.58
	Entering/leaving intersection	67	138	32.68	67.32	460.74
	During turning (DT)	317	451	41.28	58.72	14.63
	Maneuvering (MA)	99	550	15.25	84.75	12.36
	Others/undefined(UD)	18	220	7.56	92.44	
Pavement condition	Good asphalt (GA)	890	4222	17.41	82.59	97.35
	Poor asphalt(PA)	22	117	15.83	84.17	
Occurrence of intersection	With intersection(WI)	222	1442	13.34	86.66	31.69
	Without inter…(WOI)	690	2897	19.24	80.76	
	Median separated(MS)	497	1946	20.34	79.66	46.52

Occurrence of median	Without median(WOM)	415	2393	14.78	85.22	
Road alignment	Straight ahead(SA)	791	3773	17.33	82.67	86.92
	Non-straight(NSA)	121	566	17.61	82.39	
Road surface condition	Dry(DR)	796	3695	17.72	82.28	85.53
	Wet (WE)	116	644	15.26	84.74	
Weather condition	Good weather(GW)	880	4139	17.53	82.47	95.58
	Bad weather (BW)	32	200	13.79	86.21	

4.1.3 Construction of modeling variables from gathered data

The dataset used in this study was derived from a sample of 5251 accidents reported in traffic police records in Yeka sub-city, one of the sub-city of Addis Ababa. The data was done manually because of the lack of computerization. All the data collected from the police records were classified as property damage only (PDO) accident, injury accident (serious or slight injury classification is available) and fatal accident which was not recorded at sub cities instead at federal level, but in some occasions serious injury become fatal if the injured person died within 30 days.

The explanatory variables (independent variables) are categorical. Since some of the categorical variables have several levels, as described in the methodology part, a collection of design variables (or dummy variables) was needed to represent the data. SPSS 23 software package was used for coding design variables/dummy variables. It is more convenient to have as few levels of design variables as possible in order to simplify the model interpretation and to avoid difficulties during running the model processing in SPSS. In other words, the more levels of design variables the model include, the more difficult the interpretation becomes and the occurrence of extremely small events/observations may lead to difficulties during running the model. Thus, an attempt was made in the early stages of this study to reduce the number of level of design variables.

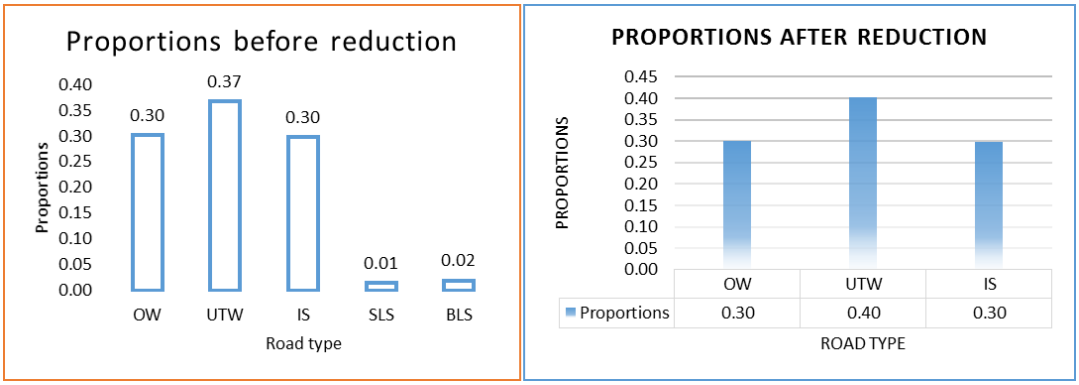


Figure 4-3 Sample Study variable before and after reduction

Table 4-3 Levels of variables before and after reduction

Before reduction		After reduction		
Categorical variable	Levels (Design variables)	Categorical variable	Levels (Design variables)	Variable type
Crash hour	8	Crash hour	2	Binary (nominal)
Crash type	11	Crash type	4	Dummy (nominal)
Days of the week	7	Days of the week	7	Dummy (nominal)
Intersection type	8	Occurrence of intersection	2	Binary (nominal)
Land use	11	Land use	4	Dummy (nominal)
Lighting condition	7	Lighting condition	2	Binary (nominal)
Maneuvering condition	13	Maneuvering condition	4	Dummy (nominal)
Pavement type	4	Pavement type	2	Binary (nominal)
Road Condition	2	Surface condition	2	Binary (nominal)
Road geometry	9	Road alignment	2	Binary (nominal)
Road type	5	Occurrence of median	3	Dummy (nominal)
Weather condition	2	Weather condition	2	Binary (nominal)
Severity type	4	Severity type	3	Ordinal

Looking at the proportion of the levels of the study variables (see the sample Fig.4-2), some of it was merged with other levels. Thus, the hypothesis testing technique for proportions was used in this study to decide whether the number of levels for a design variable could be reduced or not. The level of road type before reduction were five in number which includes one-way, two-way solid line separated, two-way broken line separated, two-way median separated and undivided two-way. But the proportions of two-way solid line separated and two-way broken line separated were insignificant as shown graphically above.

As a result, these levels of road type was merged within undivided two-way. So the level of road type after reduction becomes three which include one-way, undivided two-way and median separated. Table 4-3 shows the number/levels of design variables before and after reduction, and the reduction process for all study variables was described graphically in Appendix I.

4.1.4 Correlation between variables

The most common measures of correlation are Pearson Correlation, Variance Inflation Factor (VIF) and Condition index values. The value of Pearson Correlation, which can range from -1 to 1, is a measure of the strength of the linear relationship between two variables. A value of -1 indicates a perfect negative linear relationship between variables, a value of 0 indicates no linear relationship between variables, and a value of 1 indicates a perfect positive relationship between variables

In this study, thirteen variables have been tested for inclusion in the models. Twelve of them are categorical variables. Prior to the development of accident prediction models, Pearson's correlation analysis for aggregate data was processed to find out the linear relationship within every two independent factors. It is not appropriate for the variables representing factors with strong correlation to exist simultaneously in a model.

To be able to take correlations into account when selecting variables for the models, the correlations between variables have been studied; but it is difficult to give a brief overview over the correlations between categorical variables. Appendix F shows a correlation matrix for the variables. For example days of the week correlates with crash hour; lighting condition correlates slightly with most other variables, but no strong correlations are found. It negatively correlated with weather condition. The same conclusion can be obtained from each pair of variables.

4.1.5 Variable Selection

In this study, three sets of independent variables were incorporated into the model. The first sets of independent variable included all the candidate independent variable; the second sets of independent variable included variables related to environmental characteristics(land use pattern, lighting condition, road surface condition and weather condition); and the last sets of

independent variables included variables related to road characteristics (occurrence of median, occurrence of intersection, pavement condition and road alignment).

In consonance with earlier studies, the decision on which variables should be retained in the model was based on two criteria. The first criterion was whether the estimated parameter was significant at the 95% confidence level (p-value less than 5%). The second criterion was whether the addition of the variable to the model cause a significant drop in the scaled deviance at the 95% confidence level. Although a large number of variables were collected and considered for inclusion in the full model development, only variables with significant estimated coefficients (p-value less than 5%) were maintained in the model.

Chi-squared test of independence was used to identify and select explanatory variables which have a statistically significant association with the dependent variable. As explained in the literature review, the chi-square test can be used to estimate how closely the distribution of a categorical variable matches an expected distribution (the goodness-of-fit test), or to estimate whether two categorical variables are independent of one another (the test of independence). Based on the chi-squared test of independency executed in SPSS 23 statistical software, six explanatory variables become significant (p-value less than 5%). These include occurrence of intersection, lighting condition, days of the week, road type, maneuvering condition and occurrence of median. For all study variables the cross tabulation is listed on Appendix G.

4.1.6 Model Estimation Results

By specifying the dependent variable, the explanatory variables, the error structure and the link function, the modeling process was processed. Model parameters (coefficients) were estimated using maximum likelihood approach. SPSS 23 is used to estimate a multinomial logistic regression model. The main objective of this study was to investigate the relationships between the crash severity and the predictor variables by using the logistic regression modeling. The first part of this analysis was checking ordinal or nominal of dependent variable crash severity by using test of parallel lines in SPSS. From the test of parallel lines results shown in Table 4-4 above, the significant level is less than 0.05, so it is extremely significant. We can conclude that the dependent variable crash severity is nominal, so the nominal multinomial logistic regression model was built. In practice, when estimating the model, the model coefficients of the reference group are set to zero. Since 3 levels of severity exist, only (3-1) distinct sets of parameters can be identified and estimated, so severity level

equals to property damage only, is set to reference category. The output for each dataset is described herein below.

Table 4-4 Test of Parallel Lines in SPSS Output

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	71.819			
General	40.141	31.678	2	.000

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

4.1.6.1 Multinomial logistic Model results

Initial effort to develop the multinomial logistic regression was made by using the available variables in the dataset. A nominal multinomial logistic regression model was built in SPSS 23 to predict the likelihood of vehicle crash being property damage only, slight injury or serious injuries. The P-value is the important criterion in maximum likelihood estimates for identifying the influential factors for vehicle crashes. If the P-value is less than 0.05, it means this factor is significant to the dependent variable crash severity. If the P-value is less than 0.001, it means this factor is extremely significant to the dependent variable crash severity.

The following Tables 4-5 through 4-10 present the results of the Multinomial Logit model. The Likelihood ratio test (Table 4-5, 4-7 & 4-9) shows the contribution of each variable to the model; and the remaining tables (Table 4-6, 4-8 &4-10) indicates parameter estimates which include the logistic regression coefficient, Wald test, and odd ratio for each of the predictors (environmental conditions, road features and traffic condition variables).

4.1.6.1.1 Results for Road features related Variables

Initial effort to develop the multinomial logistic model was made by using the available road feature related variables in the dataset. From road feature related factors, road type and occurrence of intersection had significant effect for the crash severity levels with P-value less than 0.05(Table 4-5). As shown on Table 4-5 road type significantly reduced the likelihood of an injurious and property damage only vehicle crash. However others (road alignment and pavement type) did not influence the likelihood of crash severity within 5% error. For the road type, all levels (Undivided two-way, divided two-way & one-way) influences the

occurrence of slight crash severity level, but only divided two-way influences the likelihood of serious crash severity level (Table 4-6). So it means the crash which is taken place in divided two-way will result more serious injury than other factors. On the other hand intersection type is statistically significant for both slight and serious injury levels (Table 4-6).

Table 4-5Likelihood Ratio Tests for Road features

Likelihood Ratio Tests

Effect	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	209.690	288.484	185.690[a]	.000	0	.
Road Type	275.445	327.975	259.445	73.755	4	.000
Road Geo	208.844	274.505	188.844	3.154	2	.207
Intersection occurrence	228.476	294.138	208.476	22.786	2	.000
Pavement condition	207.652	273.314	187.652	1.962	2	.375

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Table 4-6 indicates parameter estimates which include the logistic regression coefficient, Wald test, and odd ratio for each road feature related predictors. The odds ratio presents levels comparison in each influential factor to indicate how each factor affects the analysis results. For the road type, the odds ratio of divided two-way versus undivided two-way for slight and serious crash severity level are 1.159 and 1.680 respectively as presented in Table 4-6. This means divided two-way is more dangerous than undivided two-way for both crash severity levels. The odds ratio of one-way versus undivided two-way for slight and serious crash severity level are 1.194 and 0.709 as shown in Table 4-6. Therefore, it means one-way is more dangerous than undivided two-way for slight crash severity level, but the case is vice versa for the serious crash severity level.

For intersection type, the odds ratio of with intersection versus without intersection for slight crash severity level is 1.357 as shown on Table 4-6. This means with intersection is more dangerous than without intersection for slight crash severity levels. The odds ratio of with

intersection versus without intersection for serious crash severity level is 1.610 (Table 4-6). Therefore, it means once again with intersection is more dangerous than without intersection for serious crash severity level.

Table 4-6Parameter Estimates for road features

Parameter Estimates

Severity_level[a]		B	Std. Error	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
						Lower Bound	Upper Bound
Slight injury	Intercept	-2.589	.336	.000			
	[divided two-way=1]	.428	.128	.001	1.534	1.194	1.970
	[one-way=2]	.409	.133	.002	1.505	1.159	1.954
	[undivided two-way=3]	0[b]	.	.	.	.	.
	[non-straight=1]	-.104	.140	.456	.901	.685	1.185
	[straight ahead=2]	0[b]	.	.	.	.	.
	[with intersection=1]	.305	.114	.007	1.357	1.085	1.695
	[without intersection=2]	0[b]	.	.	.	.	.
	[good asphalt=1]	-.095	.288	.741	.909	.517	1.599
	[poor asphalt=2]	0[b]	.	.	.	.	.
Serious injury	Intercept	-3.613	.419	.000			
	[divided two-way=1]	.764	.125	.000	2.146	1.680	2.743
	[one-way=2]	-.049	.151	.745	.952	.709	1.280
	[undivided two-way=3]	0[b]	.	.	.	.	.
	[non-straight=1]	.236	.158	.135	1.266	.929	1.725
	[straight ahead=2]	0[b]	.	.	.	.	.
	[with intersection=1]	.476	.118	.000	1.610	1.277	2.031
	[without intersection=2]	0[b]	.	.	.	.	.
	[good asphalt=1]	.461	.371	.214	1.586	.766	3.285
	[poor asphalt=2]	0[b]	.	.	.	.	.

a. The reference category is: 3.00.

b. This parameter is set to zero because it is redundant.

4.1.6.1.2 Results for Environmental condition related Variables

The explanatory variables related to environmental condition includes weather condition, land use pattern, lighting condition and road surface condition. Multinomial logistic model

was also conducted using environmental related variables. From these variables, road surface condition and lighting condition had significant effect for the crash severity levels with P-value less than 0.05(Table 4-7). However others (land use pattern and weather condition) did not influence the likelihood of crash severity within 5% error. For the road surface condition, dry road surface level and daylight influences the occurrence of serious crash severity level, but not slight crash severity level, instead wet surface condition and darkness influences the likelihood of slight crash severity level (Table 4-8). So it means the crash which is taken place in dry road surface condition and day light will result more serious injury than other factors (Table 4-8).

Table 4-7Likelihood Ratio Tests for environmental conditions

Likelihood Ratio Tests

Effect	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	293.664	398.722	261.664[a]	.000	0	.
Land Use	287.456	339.986	271.456	9.792	8	.280
Road surface condition	317.006	408.932	289.006	27.342	2	.000
Lighting condition	341.605	433.531	313.605	51.941	2	.000
Weather condition	291.586	383.513	263.586	1.923	2	.382

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Table 4-8 indicates parameter estimates which include the logistic regression coefficient, Wald test, and odd ratio for each environmental condition related predictors. For the road surface condition, the odds ratio of dry versus wet for slight and serious crash severity level are 1.145 and 2.368 respectively as presented in Table 4-8. This means dry road surface condition is more dangerous than wet road surface condition for both crash severity levels. The odds ratio of one-way versus undivided two-way for slight and serious crash severity level are 1.194 and 0.709 as shown in Table 4-6. Therefore, it means one-way is more

dangerous than undivided two-way for slight crash severity level, but the case is vice versa for the serious crash severity level.

For intersection type, the odds ratio of with intersection versus without intersection for slight crash severity level is 1.357 as shown on Table 4-6. This means with intersection is more dangerous than without intersection for slight crash severity levels. The odds ratio of with intersection versus without intersection for serious crash severity level is 1.610 (Table 4-6). Therefore, it means once again with intersection is more dangerous than without intersection for serious crash severity level.

Table 4-8Parameter Estimates for environmental conditions

Parameter Estimates

Severity_level[a]		B	Std. Error	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
						Lower Bound	Upper Bound
Slight injury	Intercept	-2.203	.278	.000			
	[Around business district=1]	-.158	.129	.222	.854	.663	1.100
	[Around Factory=2]	.086	.254	.735	1.090	.662	1.794
	[Around non-business district=3]	-.066	.125	.601	.937	.733	1.197
	[Around Residence=4]	-.239	.241	.320	.787	.491	1.261
	[Others=5]	0[b]	.	.	.	.	.
	[Dry=1]	.136	.161	.399	1.145	.835	1.571
	[Wet=2]	0[b]	.	.	.	.	.
	[Daylight=1]	-.205	.144	.156	.815	.614	1.081
	[Darkness=2]	0[b]	.	.	.	.	.
	[Bad weather=1]	.089	.248	.720	1.093	.673	1.776
	[Good weather=2]	0[b]	.	.	.	.	.
Serious injury	Intercept	-2.646	.324	.000			
	[Around business district=1]	.120	.128	.351	1.127	.877	1.449
	[Around Factory=2]	.394	.241	.102	1.483	.925	2.379
	[Around non-business district=3]	-.074	.132	.575	.929	.717	1.203
	[Around Residence=4]	-.330	.263	.209	.719	.429	1.203
	[Others=5]	0[b]	.	.	.	.	.
	[Dry=1]	.862	.175	.000	2.368	1.681	3.337
	[Wet=2]	0[b]	.	.	.	.	.
	[Daylight=1]	-.953	.127	.000	.386	.301	.494
	[Darkness=2]	0[b]	.	.	.	.	.
	[Bad weather=1]	.381	.294	.194	1.464	.823	2.603

[Good weather=2]	0[b]	.	.	.	.	.

a. The reference category is: 3.00.

b. This parameter is set to zero because it is redundant.

4.1.6.1.1 Results for Traffic condition related Variables

To see the effects of traffic condition related variables, multinomial logistic model was also conducted. The Collected traffic condition related variables include traffic volume, defendant vehicle maneuvering condition and crash type. All traffic condition related variable are extremely significant with P-value less than 0.001 (Table 4-9). As shown in table 4-10, traffic volume had significant effect on the occurrence of both slight and serious injuries with P-value less than 0.001. For the crash type, all levels (overturning, vehicle-inert, vehicle-pedestrian & vehicle-vehicle) influences the occurrence of both slight and serious crash severity level (Table 4-6). For vehicle maneuvering type, turning and maneuvering levels had significant effect on the likelihood of slight crash severity level. So it means the crash which is taken place in during turning and maneuvering will result more slight injury than other factors. On the other hand serious injury is statistically influenced by during turning, during entering or leaving intersection and during manuevering (Table 4-6).

Table 4-9Likelihood Ratio Tests for traffic condition variables

Likelihood Ratio Tests

Effect	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	Df	Sig.
Intercept	404.715	499.866	368.715[a]	.000	0	.
AADT	473.479	558.058	441.479	72.764	2	.000
Cash type	761.904	825.338	737.904	369.189	6	.000
Maneuvering condition	584.283	637.145	564.283	195.568	8	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Table 4-10 indicates parameter estimates which include the logistic regression coefficient, Wald test, and odd ratio for each traffic condition related predictors. The odds ratio presents

levels comparison in each influential factor to indicate how each factor affects the analysis results. For the crash type, the odds ratio of vehicle-pedestrian is higher than odd ratio of overturning for both slight and serious crash severity level. For the maneuvering condition, straight through had higher odd ratio indicating that dangerous effect for the occurrence of crash.

Table 4-10Parameter Estimates for traffic condition variables

Parameter Estimates

Severity_level[a]		B	Std. Error	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
						Lower Bound	Upper Bound
Slight injury	Intercept	-4.500	2.248	.045			
	AADT	.618	.234	.008	1.856	1.172	2.937
	[Overturning=1]	-2.847	.304	.000	.058	.032	.105
	[Vehicle-inert=2]	-4.964	1.128	.000	.007	.001	.064
	[Vehicle-pedestrian=3]	-2.737	1.079	.011	.065	.008	.537
	[Vehicle-vehicle=4]	0[b]	.	.	.	.	.
	[During turning=1]	-2.364	.291	.000	.094	.053	.166
	[Around inter=2]	-1.035	1.129	.359	.355	.039	3.247
	[Maneuvering=3]	-2.154	.528	.000	.116	.041	.327
	[Straight through=4]	.018	1.069	.987	1.018	.125	8.268
	[Others=5]	0[b]	.	.	.	.	.
Serious injury	Intercept	-16.958	2.403	.000			
	AADT	1.894	.247	.000	6.645	4.094	10.784
	[Overturning=1]	-3.408	.246	.000	.033	.020	.054
	[Vehicle-inert=2]	-7.376	.855	.000	.001	.000	.003
	[Vehicle-pedestrian=3]	-1.933	.517	.000	.145	.052	.399
	[Vehicle-vehicle=4]	0[b]	.	.	.	.	.
	[During turning=1]	-1.214	.282	.000	.297	.171	.516
	[Around inter=2]	3.714	.488	.000	41.026	15.768	106.742
	[Maneuvering=3]	-.851	.391	.029	.427	.198	.919
	[Straight through=4]	-.761	1.350	.573	.467	.033	6.588
	[Others=5]	0[b]	.	.	.	.	.

a. The reference category is: Property damage only.

b. This parameter is set to zero because it is redundant.

4.1.7 Model Result Testing

The results from fitting all the explanatory variables simultaneously is shown in Appendix J; and the result indicates that CRT (Crash type), PAC(Pavement type) and MAC (maneuvering conditions) are significant variables influencing the occurrence of slight crash severity. On the other hand LIC (Lighting condition), MAC (Maneuvering condition) and CRT (crash type) are found as significant variables.

The predicted logit model was established using the above three statically significant variables out of ten explanatory variables for both injury severity levels. However, others were dropped because they did not contribute significantly to the model.

The fitted model for serious injury prediction is

$$\begin{aligned}\ln(odds) = {} & 0.667LIC1 - 3.367CRT1 - 6.722CRT2 - 1.727CRT3 \\ & - 1.283MAC1 + 3.592MAC2 - 0.757MAC3 \\ & - .348MAC4\end{aligned} \quad (17)$$

OR

$$P(\,Serious\ Injury) = \frac{e^{0.667LIC1-3.367CRT1-6.722CRT2-1.727CRT3-1.283MAC1+3.592MAC2-0.757MAC3-.384MAC4}}{1+e^{(0.667LIC1-3.367CRT1-6.722CRT2-1.727CRT3-1.283MAC1+3.592MAC2-0.757MAC3-.384MAC4)}} \quad (18)$$

Where:

- ✓ P(Serious Injury)→the probability of severity level to be serious injury
- ✓ LIC1→1 if the lighting condition is daylight; 0 otherwise
- ✓ CRT1→1 if the crash type is overturning ; 0 otherwise
- ✓ CRT2→1 if the crash type is vehicle-inert; 0 otherwise
- ✓ CRT3→1 if the crash type is vehicle-pedestrian, 0 otherwise
- ✓ MAC1→1 if the vehicle movement is during turning, 0 otherwise
- ✓ MAC2→1 it the vehicle mov't is leaving/entering intersection, 0 otherwise
- ✓ MAC3→1if the vehicle movement is during maneuvering, 0 otherwise
- ✓ MAC4→1 if the vehicle movement is straight through, 0 otherwise

To assess the predictive abilities of the developed models, a one-year data was used for model testing. The relative frequencies of the raw data and the estimated models and error rates between them were calculated and compared as shown in table 4-11 and 4-12.

Table 4-11 Frequencies of the Raw Data and Estimated Models (2016)

	Significant predictors								Probabilities
Cases	LIC1	CRT1	CRT2	CRT3	MAC1	MAC2	MAC3	MAC4	
Case1	0	0	0	1	0	1	0	0	0.866
Case2	1	0	0	0	0	0	0	0	0.661
Case3	1	0	0	1	0	0	0	0	0.257
Case4	1	0	0	1	0	0	1	0	0.142
Case5	0	0	0	1	0	0	1	0	0.078

By comparing the estimated probabilities of the above cases, case one which indicate that the occurrence of serious injury during daylight condition, interring/leaving an intersection and vehicle-to-pedestrian crash type accounted higher probability than the other cases.

Based on the model test result, the null hypothesis of this study which stated that "statistical Accident Prediction Model are very important during decision-making process" was tested using Chi-square test as shown below.

Table 4-12 Chi-square test for model result testing

Cases	Significant predictors				
	Observed	Estimated	Error rate	Test statistics	P-value
Case1	117	133	0.137	1.923	0.34
Case2	106	101	0.05	0.248	0.12
Case3	64	39	0.64	16.03	0.001
Case4	8	22	1.75	8.91	0.003
Case5	10	12	0.2	0.333	0.26

The chi-square statistic compares the observed values to the estimated values. This test statistic is used to determine whether the difference between the observed and estimated values is statistically significant. The chi-square tests statistic and the P-value is shown in Table 4-12.

The p-value for the three cases (case 1, 2 &5) is higher than the alpha level of .05; as a result the value of the chi-square test statistic is large enough to reject the null hypothesis. Therefore, there is enough evidence to reject the null hypothesis.

As a result, based on the above evidence, the developed statistical Accident Prediction Model doesn't estimate nearly the same as the collected traffic accident data within 5% level of significance.

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

In this study road traffic crashes in the Yeka sub-city have been analyzed, and statistical models developed and assessed to identify contributory factors that are likely to affect these crashes. Before developing the accident prediction models an extensive review of both local and overseas accident prediction studies, and associated statistics papers and books was carried out to determine what model forms and statistical techniques were being used by other researchers.

During this review, a diverse range of model forms and statistical techniques were encountered. A lot of improvements has been made in the previous research of modeling crashes. The modeling techniques shifted from conventional regression to stochastic regression. The initial research emphasized the relationships between highway geometric variables and crashes as Single-variate and multivariate deterministic models.

Although the previous research is helpful in identifying the attribution of different causal variables to the occurrence of crashes, further studies are still needed in order to obtain consistent and concise conclusion about the relationships between the occurrence of crashes and its causal factors. Further research is still needed to obtain more reliable and consistent crash models.

After considering previous studies and reviewing several statistical techniques it was decided that normal regression was not appropriate for developing accident severity prediction models and instead multinomial logistic regression was necessary.

Multinomial logistic regression was used to estimate the model parameters. The data set obtained for this study were applied to examine the goodness-of-fit regression models. The dependent variable used in this study was crash severity. As part of the study, the models have been tested to see how well they predict the accidents observed during a one year accident period.

From the model developed, road type, road surface condition, crash type, maneuvering condition and lighting conditions were found as significant explanatory variables that

influence the prediction of crashes in the yeka sub-city. This indicates that non-behavioral factors have an effect on the occurrence of accident.

Based on the findings of this study, the following conclusions are made:

- The results have showed that multinomial logistic regression models fit the data and the variables road type, road surface condition, crash type, maneuvering condition and lighting conditions have clear influences in terms of influencing on traffic accidents.
- Significant influential factors that increase the likelihood of vehicle crashes injury severity have been identified.
- Multinomial logit regression model showed a great potential in predicting the probability of occurrence of serious injury severity type.
- The closeness of the estimation and raw data indicates that multinomial logit regression modeling methods can be used to describe the probabilities of crash events.
- Since multinomial logistic regression doesn't estimates nearly the observed data with minimum error rates, there is an evidence to reject the null hypothesis which states that "Statistical Accident Prediction Model are very important during decision making process".

5.2 Recommendation

Road traffic may be considered as a system in which the human, the vehicle and the road and its environment interact each other. In order to improve the efficiency and safety of road traffic, it is important to relate accident frequency and severity to the causative variables. Even though the contributing factors of an accident is a multi-factor, it is more advisable to analyze each factors individually, and develop a simplified accident predictive model.

In this study accident prediction model relating the occurrence of accident to non-behavioral factors was developed (which was qualitative). But if the traffic data collecting format is changed (if include quantitative parameters like lane width, section length), quantitative researches should be performed by relating road and environmental factors with the occurrence of accidents.

In addition to the above, Automotive and Dynamic Engineers should conduct a research and develop a predictive model considering only vehicle related factors, similarly, to change the road user behavior, attitude or knowledge in order to increase road safety, Behaviorist and Doctor/Epidemiologist should conduct a research and develop a predictive model considering human related factors.

The contribution of the above mentioned research topics will minimize the occurrence of accident if best fit prediction models are used as a tool for decision making process and cost effective countermeasure is implemented. To use accident prediction models as a tool for decision making process, reliable and accurate relationship between the occurrences of crash and contributing factors should be developed.

Traffic accident database is a prerequisite for any traffic accident reduction and prevention measures. It is a vital source of factual information for politicians and administrators, researcher, traffic and road engineers, organizations engaged in driver training, the police who makes the accident reports. So traffic police officers should be supported with professionals for collecting full information regarding on road characteristic features (should include Class of road/road number, Carriage type/ no. of lanes, Road width, Road shoulder width, section length, Speed limit, Junction type and the like) and Environmental features (which include Light conditions, Road lighting, Road surface condition (dry, wet, etc.), Road surface quality, Weather, Type of traffic control, Road geometry, traffic control device and the like). These parameters should be included in the daily accident recording sheet.

The last recommendation goes to Addis Ababa police commission statistical officers to use or prepare a standard daily accident recording booklet just not by writhing a statement but by putting a mark or a thick sign.

Traffic police should be trained at national and local levels or supported with professionals to collect accurate and reliable traffic accident data or there should be three experts (traffic/transport engineer/ expert, automotive/dynamic experts & behaviorist or epidemiologist) during traffic accident data collection. Traffic and transport engineer should prepare standard accident data collecting format and record daily accident data by themselves. Likewise automotive or dynamic engineers/experts and behaviorist or epidemiologists experts should prepared their own standard accident data collecting format. Lastly there should be another expert who compile the data collected by the three experts.

REFERENCES

Abdel-Aty, M. (2003). *Analysis of Driver Injury Severity Levels at Multiple Locations Using Ordered Probit Models,.*

Administration, U. D. (2000). *Prediction of the expected safety performance of rural Two-lane highways. United State: Development and Technology Turner-Fairbank highway research center.*

Administration, U. D. (2016, 03 08). *Federal Highway Administration Research and Technology. Retrieved from www.fhwa.dot.gov/publications/research/safety/99.207*

Ajit G. and S. Ripunjoy. (2004). *A Statistical Analysis of Road Traffic Accidents in Dibrugarh city. INDIA.: Assam.*

Al-Ghamdi, A. S. (2001). *Using logistic regression to estimate the influence of accident factors on accident severity. Riyadh 11421, Saudi Arabia: College of Engineering King Saud Uniersity.*

Alister C. OBE and B. Simon. (2*011). Licensed to Skill. England and Wales: Institute of Advanced Motorists Limited.*

Authority, E. R. (2015, DECEMBER). *Accident Monitoring and Analysis. Retrieved from Road safety Information Page: http://www.rta.gov.et/planningresearch.htm*

Bauer, K., & D.W. Harwood. (1996). *"Statistical Models of At-Grade Intersection Accident. Federal Highway Administration.*

Bedard, M. G. (2002). *The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities.*

Belachew M. (1997). *"Some Thoughts on Intra-Urban Transport Problems in Ethiopia, The Case of the Anbessa City Bus Transport." . Journal of Development Research 19(1).*

Berhanu G. (2000). *Effects of Road and Traffic Factors on Road Safety in Ethiopia . Trodhium Norway: .*

Bhadra, D. (2005). *Choice of Aircraft Fleets in the U. S. Domestic Scheduled Air Transportation System: Findings from a Multinomial Logit Analysis.*

Bitew Mebratu. (2002). *Taxi Traffic Accidents in Addis Ababa: Causes, Temporal and Spatial Variations, And Consequences . Addis Ababa: Addis Ababab University.*

Bunn, F.T., C., & al., e. (2003). *Traffic calming for the prevention of road traffic injuries: Systematic review and meta-analysis.*

CHENGYE, P., RANJITKAR, & Prakash. (2010). *Modeling Motorway Accidents using Negative Binomial REgression;. New Zealand: Department of Civil & Environmental Engineering, University of Auckland, Auckland 1142.*

D.W. Harwood, F. C. (2000). *Prediction of the Expected Safety Performance of Rural Two-Lane Highways. Georgetown: U.S. Department of Transportation, Federal Highway Administration .*

Fred Mannering, D. L. (2010). *The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. Texas A&M University: Forthcoming in Transportation Research.*

Geedipally, S. R. (2005). *Analysis of Traffic Accidents before and after resurfacing - A statistical approach . Sweden: Swedish National Road and Transport Research Institute.*

Gkritza, K. D. (2006). *Airport Security Screening and Changing Passenger Satisfaction:.*

Haleem, K. M.-A. (2010). *Examining Traffic Crash Injury Severity at Unsignalized Intersections.*

Hauer, E. (1986). *On the Estimation of the Expected Number of Accidents Accident Analysis and Prevention.*

Heinz Hautzinger, C. P. (2007). *Analysis Methods for Accident and Injury Risk Studies.*

Hobbs, F. (1979). *Traffic Planning and Engineering,second edition, . New York: Pergamon Press.*

Islama, S. &. (2006). *Driver aging and its effect on male and female single-vehicle accident injuries:.*

J.Garber, N. (2001). *Stochastic Models Relating Crash Probabilities With Geometric And Corresponding Traffic Characteristics Data. Viriginia: A U.S. DOT University Transportation Center.*

Jonsson, T. (2005). *Predictive Models for Accidents on Urban Links, A focus on vulnerable road users. Lund Institute of Technology, Department of Technology and Society, Traffic Engineering.*

José M. Pardillo Mayora, C. P., & Rubio, R. L. (2003). *Relevant Variables for Crash Rate Prediction in Spain´s Two Lane Rural Roads. Madrid Polytechnic University: presentation and publication review to the Transportation Research.*

Khattak, A. J. (2002). *Effects of work zone presence on injury and non-injury crashes.*

Kockelman, K. Y. (2002). *Driver Injury Severity: An Application of Ordered Probit Models.*

Lee, C. &.-A. (2005). *Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida.*

Lisa, K.S., & B.David et al. (2005). *Evaluation of Traffic Crash Fatality Causes and Effect. Florida State,: Florida State University.*

Maher M.J., S. I. (1996). *A comprehensive methodology for the fitting predictive accident models.*

Manner, H. &.-Z. (2013). *Analyzing the severity of accidents on the German Autobahn. Accident Analysis and Prevention, .*

Marko Renčelj. (2009). *The methodology for predicting the expected level of traffic safety in the different types of level intersections. Slovenia: Universita' Degli Studi Di Trieste.*

McFadden, D. (1974). *Frontiers in Econometrics. New York: Academic Press.*

Mekonnen, T. (2007). *Empirical analysis on traffic accidents involving human injuries, the case of addis ababa. addis ababa: addis ababa university office of graduate program; faculty of science, department of statistics.*

Miaou, 1., Shankar et al., 1., & Lord et al., 2. (1994, 1997, 2005).

Miaou, S. (1994). *The relationship between truck accidents and geometric design of road section: Poisson versus negative binomial regressions.*

Miaou, S.-p. L. (1993). *Modeling Vehicel Accidents and Highway Geometric Design Relationships. Accident Analysis and Prevention.*

Mohammed, M. (2011). *Costing Road Traffic Accidents in Ethiopia. MSc, . Addis Ababa: Addis Ababa University.*

Morfoulaki, M. Y. (2007). *Estimation of Satisfied Customers in Public Transport Systems.*

Nicholas J.Garber. (2009). *Traffic and Highway Engineering, fourth edition. Virginia: University of Virginia.*

O'Donnell, C. D. (1996). *Predicting the Severity of Motor Vehicle Accident Injuries Using Models of Ordered Multiple Choices.*

Obeng, K. &. (2013). *Pedestrian injury severity in automobile crashes.*

Organization, W. H. (2014).

Pande, M. A.-A. (n.d.). *The viability of real-time prediction and prevention of traffic accidents.*

Renčelj, M. (2009). *The methodology for predicting the expected level of traffic safety in the different types of level intersections.* Univesità degli studi di Trieste.

Reurings, M. J. (2005). *Accident Prediction Models and Road safety Impact Assessment: a state-of-the-art. ripcord -iserest, .*

Ronald, W. E., Raymond, M. H., & Sharon, M. L. (2007). *Probability & Statistics for Engineers & Scientists. Pearson Education International.*

Rui Garridoa, A. B. (2014). *Prediction of road accident severity using the ordered probit model. Coimbra, Portugal: Department of Civil Engineering, University of Coimbra.*

Safecarguide. (2016,, April 25). *Safecarguide. Retrieved from http//www.safecarguide.com/exp/intro/idx.htm.*

Sascia Canale, S. L. (n.d.). *The reliability of urban road Nework Accident forecat models. Catania: Universita degli Studi di Catania.*

Savolainen, P. M. (2011). *The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. . Accident Analysis and Prevention,.*

Sawalha, Z. (2003). *Statistical Issues in Traffic Accident Modeling. Research Associate Department of Civil Engineering University of British Columbia.*

Sawalha, Z., & Sayed, T. (2006). *Traffic Accident Modelling: Some Statistical Issues.*

Segni G. (2007). *Causes of Road Traffic Accidents and Possible Counter Measures on Addis Ababa-Shashemene Roads. . Addis Ababa : Addis Ababa University.*

Shankar, V. M. (1995). *Effect of roadway geometrics and environmental factors on rural accident frequencies.*

Shankar, V. M. (1996). *Statistical analysis of accident severity on rural freeways. .*

Terje A. (1998). *Road Safety in Africa Appraisal Of Road Safety.*

Train, K. (2009). *Discrete Choice Methods with Simulation. University of California, Berkeley: Cambridge University Press.*

Transportation Reasearch Board Executive Commitee . (2005). S*implified Shear Design of Structural Concrete Members. NCHRP, NCHRP REPORT 549.*

Tsui, M. K. (2006). *Pedestrian Crashes in Commercial and Business Areas: A Case Study of Hong Kong.*

Tulu, G. S. (2015). *Pedestrian crashes in ethiopia: Identification of Contributing Factors through Modelling of Exposure and Road Environment Variables. Queensland University of Technology.*

Turner, S. (1996). *Estimating Accident in A Road Network.* Univercity of Canterbury.

UN. (2009). *United Nations Economic Commission for Africa. Road Safety in Ethiopia.*

Wang, Y. Y. (2013). *Injury severity of pedestrian crashes in Singapore.*

Washington, S. P. (2011). *Logistic Regression, Discrete Outcome Models and Ordered Probability Models.* Chapman & Hall/CRC.

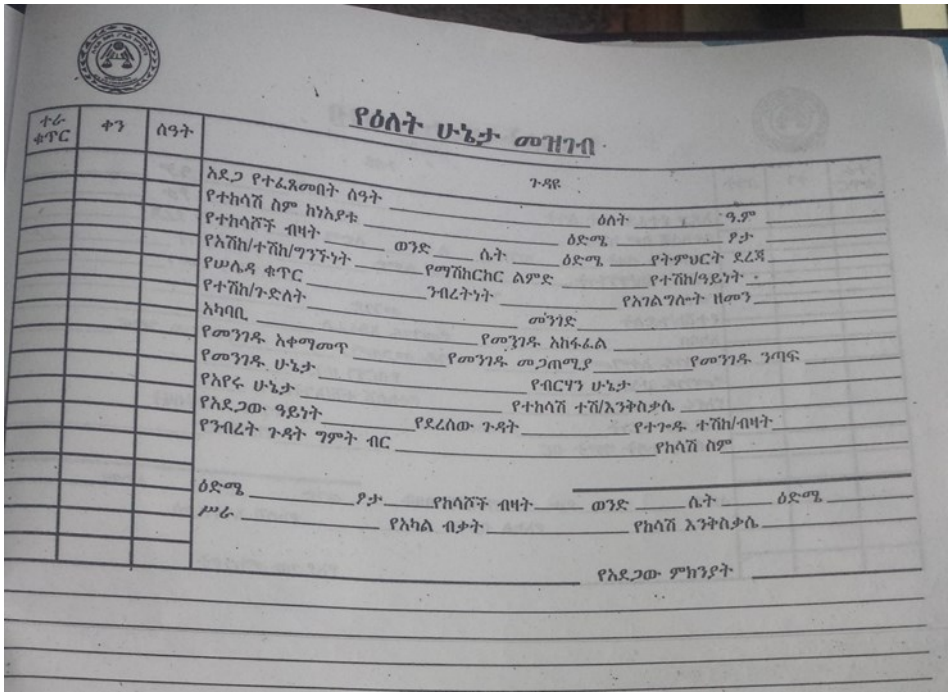
WILLIAMS, A. (2009). *Crash Prediction Model for two-lane Rural Highways in the Ashanti Region of Ghana. Kumasl: Kwame Nkrumam University Of Science And Technology.*

Yan, X. E. (2009). *Analysis of Truck-Involved Rear-End Crashes Using Multinomial Logistic Regression.*

Zakaria, M., Ueda, T., Wu, Z., & Meng, L. (2009). E*xperimental Investigation on Shear Cracking Behavior in Reinforced Concrete Beams with Shear Reinforcement.*

APPENDIX

Appendix A: Daily Accident Recording sheet



የዕለት ሁኔታ መዝገብ

ተራ ቁጥር	ቀን	ሰዓት	ጉዳዩ

አደጋ የተፈጸመበት ሰዓት _____ ዕለት _____ ዓ.ም _____
የተከሳሽ ስም ከነአያቱ _____ ዕድሜ _____ ያታ _____
የተከሳሾች ብዛት _____ ወንድ ____ ሴት ____ ዕድሜ _____ ያትምህርት ደረጃ _____
የአሽከ/ተሽከ/ግንኙነት _____ የማሽከርከር ልምድ _____ የተሽከ/ዓይነት _____
የሠሌዳ ቁጥር _____ ንብረትነት _____ የአገልግሎት ዘመን _____
የተሽከ/ጉድለት _____ መንገድ _____
አካባቢ _____ የመንገዱ አከፋፈል _____
የመንገዱ አቀማመጥ _____ የመንገዱ መጋጠሚያ _____ የመንገዱ ንጣፍ _____
የመንገዱ ሁኔታ _____ የብርሃን ሁኔታ _____
የአየሩ ሁኔታ _____ የተከሳሽ ተሽ/እንቅስቃሴ _____
የአደጋው ዓይነት _____ የደረሰው ጉዳት _____ የተጐዱ ተሽከ/ብዛት _____
የንብረት ጉዳት ግምት ብር _____ የከሳሽ ስም _____

ዕድሜ _____ ያታ _____ የከሳሾች ብዛት ____ ወንድ ____ ሴት ____ ዕድሜ _____
ሥራ _____ የአካል ብቃት _____ የከሳሽ እንቅስቃሴ _____

የአደጋው ምክንያት _____

Appendix B: Filled sample accident data recording sheet

የዕለት ሁኔታ መዝገብ

ተራ ቁጥር	ቀን	ሰዓት	ጉዳዩ
236	24 / 2 / 2003	17/0	

አደጋ የተፈጸመበት ሰዓት ስ1515 ዕለት እሁድ ዓ.ም 2003
የተከሳሽ ስም ከነአያቱ ይታገስ እሸቱ ሽብሩ ዕድሜ _____ ያታ _____
የተከሳሾች ብዛት 1 ወንድ 1 ሴት _____ ዕድሜ 58 የትምህርት ደረጃ 8ኛ
የአሽከ/ተሽከ/ግንኙነት ተቀጣሪ የማሽከርከር ልምድ 30 ዓመት የተሽከ/ዓይነት TATA
የሠሌዳ ቁጥር ጽ3ሐ 04ለ10 ንብረትነት የንግዘመ ሚ/ር የአገልግሎት ዘመን 1ዓ.ም
የተሽከ/ጉድለት የለበትም መንገድ የካቲት 16/17
አካባቢ ለሚኮሪት የመንገዱ አከፋፈል በለአንድ አቅጣጫ
የመንገዱ አቀማመጥ ቀጥታ የመንገዱ መጋጠሚያ የቤቲዘሀ የመንገዱ ንጣፍ
የመንገዱ ሁኔታ ደረቅ የብርሃን ሁኔታ የቀን
የአየሩ ሁኔታ ጥሩ-እየሮ የተከሳሽ ተሽ/እንቅስቃሴ ሠሜት ዓለቀጥታ አቅጣጫ
የአደጋው ዓይነት የመጀመሪያ-ለመጋጨ የደረሰው ጉዳት ሞተል የተጐዱ ተሽከ/ብዛት 2
የንብረት ጉዳት ግምት ብር 1,300 ሽር የከሳሽ ስም እያሱ ተሸገር ሹሩ

ዕድሜ 29 ያታ ነው የከሳሾች ብዛት 1 ወንድ 1 ሴት _____ ዕድሜ _____
ሥራ ሹፌር የአካል ብቃት መደሀ የከሳሽ እንቅስቃሴ ቀጥታ አቅጣ
የአደጋው ምክንያት ተሽከ ሰ ሠሜ

ዓለአቅጣጫ: ለመጠምሃ" ዓደረሠ: ሹፌሯ ሰሃየ: ተሽከሸ ሠሜት: ለመንገዳ: የመ/ከ/ቁ 05-44-76-87 እስከ 24-7-2004 ዓ.ም የታገስ አቅመጽቀ ል::
ሪፖ: አሬሰች ዘውዴ ተፈ

Appendix C: Collected Traffic accident data (Sample)

Crash hour	Day of the week	Street Nam	Land Use	Road Type	Road Geometry	Intersectio n Type	Type of Road pavement	Road Condit ion	Lighting Condition	Weather Condition	Defendant Vehicle manoeuvering condition	Type of crash	Type of severity
08:00-09:	Tuesday	K/08/15	Around Chur	Island Sep	Straight ah	T Sha	Good As	Wet	Sunset	Good We	Left turning	Vehicle to	Propery Dama
10:00-11:	Wednesd	K/11/12	Around Chur	Island Sep	Straight ah	Without	Good As	Wet	Sunrise	Good We	Straight ahead	Vehicle to	Propery Dama
12:00-13:	Wednesd	K/19	Around Ente	Island Sep	Straight ah	Without	Good As	Dry	Day Ligh	Good We	Right turning	Vehicle to	Slight Injury
10:00-11:	Sunday	K/13/14	Around Offic	Island Sep	Straight ah	Square	Good As	Dry	Day Ligh	Good We	Straight ahead	Vehicle to	Propery Dama
09:00-10:	Sunday	K/13/14	Around Scho	One-way	Straight ah	Without	Good As	Dry	Day Ligh	Good We	Backward mov	Vehicle to	Propery Dama
07:00-08:	Tuesday	K/13/14	Around Offic	Island Sep	Straight ah	Roundabo	Good As	Dry	Day Ligh	Good We	Straight ahead	Vehicle to	Propery Dama

Appendix D: Coded Traffic accident data (sample)

Crash_ hour	Day_of_t he_week	Land_ Use	Road_ Type	Road_Geo metry	Intersectio n_Type	Type_of_Ro ad_pavemen t	Road_C ondition	Lighting_ Condition	Weather_Con dition	manoeu vering_c ondition	Type_of _crash	Type_of_ severity
AN	TH	AI	TW	SA	SQ	GA	DR	NPRL	GW	SA	VTV	PDO
LM	W	AO	IS	SA	R	GA	DR	NPRL	GW	SA	VTV	PDO
D	T	AM	IS	SA	R	GA	DR	NPRL	C	ESQR	VTI	PDO
NO	F	AO	IS	SA	TS	GA	DR	NPRL	GW	RT	VTV	SI
MN	M	AE	IS	SA	XS	G	DR	NPRL	GW	SA	VTV	PDO

Appendix E: Collected Traffic Volume data (sample)

	Collected Traffic Volume on April, MEF = 0.97																
		Day 1 (Monday)	DEF=5.42								**Day 4(Monfday)**						
Section Name: British Embacy Google Earth:- Comoros Street				**Direction: 4killo to megenagna**			Lane type 3 lane		***Section Name: British Embacy***				**Direction: Megenagna to 4killo**			Lane type 3 lane	
15 Minutes Interval	**Vehicle type**							EF	15 Minutes Interval	**Vehicle type**							
	Car	Minibus	Mid Bus	Bus	Truck	T/Trailer	Total			Car	Minibus	Mid Bus	Bus	Truck	T/Traile	Total	
01:00-01:15	745	227	121	0	2	0	1095	8	01:00-01:15	625	118	91	5	2	0	841	
01:15-01:30	810	248	119	0	0	0	1177		01:15-01:30	715	114	116	3	0	1	949	
01:30-01:45	835	267	104	9	4	0	1219		01:30-01:45	830	154	118	8	1	0	1111	
01:45-02:00	900	259	123	10	5	0	1297		01:45-02:00	855	180	103	4	0	0	1142	
Hourly Volume	**3290**	**1001**	**467**	**19**	**11**	**0**	**4788**		Hourly Volum	**3025**	**566**	**428**	**20**	**3**	**1**	**4043**	
02:00-02:15	915	195	70	4	7	0	1191	7.5	02:00-02:15	870	205	65	4	0	1	1145	
02:15-02:30	940	239	63	7	3	0	1252		02:15-02:30	874	241	60	3	3	0	1181	
02:30-02:45	916	244	59	3	0	2	1224		02:30-02:45	825	241	62	3	0	2	1133	
02:45-03:00	847	265	43	4	4	2	1165		02:45-03:00	795	255	54	2	0	1	1107	
Hourly Volume	**3618**	**943**	**235**	**18**	**14**	**4**	**4832**		Hourly Volum	**3364**	**942**	**241**	**12**	**3**	**4**	4566	
03:00-03:15	602	262	28	7	3	0	902	10	03:00-03:15	640	258	33	5	0	2	938	
03:15-03:30	525	267	30	8	2	0	832		03:15-03:30	465	254	31	2	0	1	753	
03:30-03:45	530	234	25	2	5	2	798		03:30-03:45	320	223	25	0	1	0	569	
03:45-04:00	515	215	21	2	1	0	754		03:45-04:00	321	206	28	3	1	1	560	
Hourly Volume	**2172**	**978**	**104**	**19**	**11**	**0**	**3286**		Hourly Volum	**1746**	**941**	**117**	**10**	**2**	**4**	2820	
24-Hr Volume(A	**25058**	**8286.83**	**2179.5**	**159**	**101**	**10**	**35801.3**		**24-Hr Volum**	**22297**	**7001**	**2133.83**	**116.7**	**22.17**	**26**	**31596**	
Weekly Volume	**135816**	**44914.6**	**11812.9**	**862**	**547.4**	**54.2**	**194043**		**Weekly Volu**	**1E+05**	**37945.4**	**11565.4**	**632.3**	**120.1**	**140.9**	**171252**	
AADT	**18820**	**6223.89**	**1636.93**	**119**	**75.86**	**7.51057**	**26888.8**		**AADT**	**16746**	**5258.15**	**1602.63**	**87.62**	**16.65**	**19.53**	**23731**	
		Day 2 (Tuesday), DEF=6.36									**Day 5(Wednsday)**						
Section Name: School of tommorow Fikremariam Aba Techan Street				**Direction: 4killo to megenagna**			Lane type 3 lane		***Section Name: School of tommorow***				**Direction: Megenagna to 4killo**			Lane type 3 lane	

Appendix F: Correlation matrix

Correlations

	Crash hour	Days	Land use	Road type	Roadgeometry	Intersectiontype	Pavementcondition	Roadcondition	Lighting condition	Weather condition	Maneuvering condition
Crashhour	**1**	.075	-.017	-.012	-.027	-.008	-.015	.009	.010	-.005	.008
Days	.075	**1**	.012	.025	-.001	-.013	-.012	.016	-.009	-.004	-.015
Landuse	-.017	.012	**1**	.008	-.007	.010	-.007	-.023	.014	-.018	.005
Roadtype	-.012	.025	.008	**1**	.025	.119	.001	-.070	.007	-.001	.017
Roadgeometry	-.027	-.001	-.007	.025	**1**	.060	.036	.115	.033	-.008	.001
Intersectiontype	-.008	-.013	.010	.119	.060	**1**	-.038	.068	.110	.001	-.025
Pavement condition	-.015	-.012	-.007	.001	.036	-.038	**1**	-.019	.002	-.074	.047
Roadcondition	.009	.016	-.023	-.070	.115	.068	-.019	**1**	.189	-.091	-.010
Lightingcondition	.010	-.009	.014	.007	.033	.110	.002	.189	**1**	-.021	.019
Weathercondition	-.005	-.004	-.018	-.001	-.008	.001	-.074	-.091	-.021	**1**	.094
Maneuveringcondition	.008	-.015	.005	.017	.001	-.025	.047	-.010	.019	.094	**1**

Appendix G: Chi-Square test (Sample)

ROT * ASL

Crosstab

			ASL		Total
			INJ	PDO	
ROT	TW	Count	200	1362	1562
		Expected Count	271.3	1290.7	1562.0
	OW	Count	250	1332	1582
		Expected Count	274.8	1307.2	1582.0
	IS	Count	462	1645	2107
		Expected Count	365.9	1741.1	2107.0
Total		Count	912	4339	5251
		Expected Count	912.0	4339.0	5251.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	55.884[a]	2	.000
Likelihood Ratio	55.923	2	.000
Linear-by-Linear Association	54.005	1	.000
N of Valid Cases	5251		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 271.29.

ROC * ASL

Crosstab

			ASL		Total
			INJ	PDO	
ROC	WE	Count	116	644	760
		Expected Count	132.0	628.0	760.0
	DR	Count	796	3695	4491
		Expected Count	780.0	3711.0	4491.0
Total		Count	912	4339	5251
		Expected Count	912.0	4339.0	5251.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.743[a]	1	.098		
Continuity Correction[b]	2.575	1	.109		
Likelihood Ratio	2.821	1	.093		
Fisher's Exact Test				.108	.053
Linear-by-Linear Association	2.743	1	.098		
N of Valid Cases	5251				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 132.00.

b. Computed only for a 2x2 table

Appendix H: Tests of Proportions (Sample)

Hypothesis testing for proportions on Road type

Null Hypothesis: *the two proportions are equal* ($P_1 = P_2$)

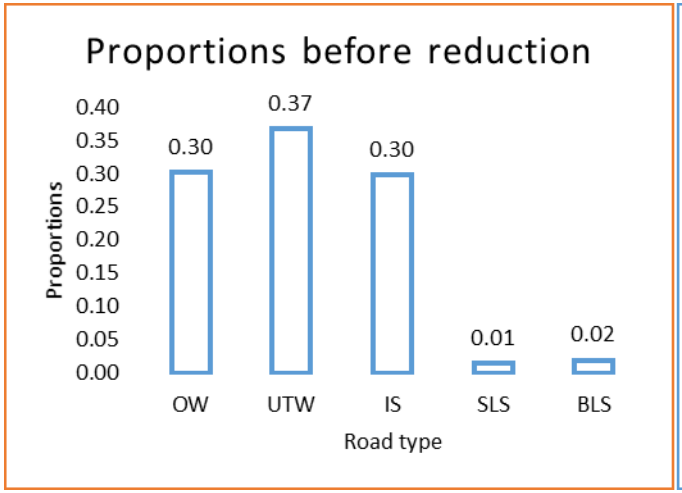
Alternative Hypothesis: $P_1 \neq P_2$ *or* $P_1 > P_2$ *or* $P_1 < P_2$

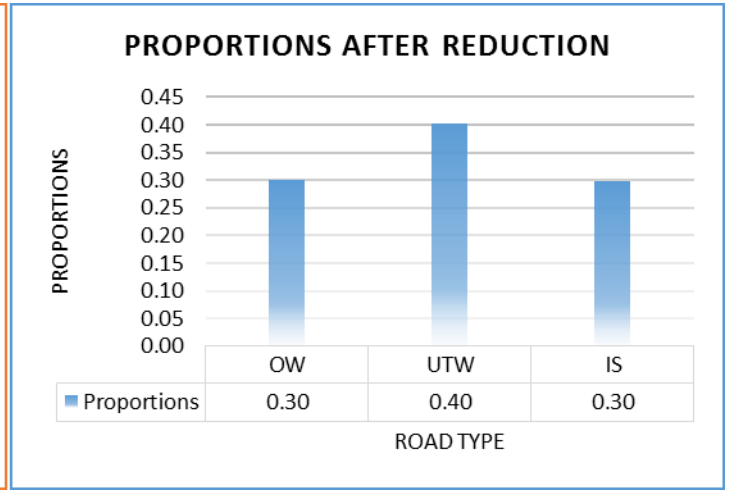
$$\sqrt{\frac{\hat{P}(1-\hat{P})}{n_1} + \frac{\hat{P}(1-\hat{P})}{n_2}} = S_{p1-p2} \qquad Z = \frac{p_1 - p_2}{S_{p_1-p_2}}$$

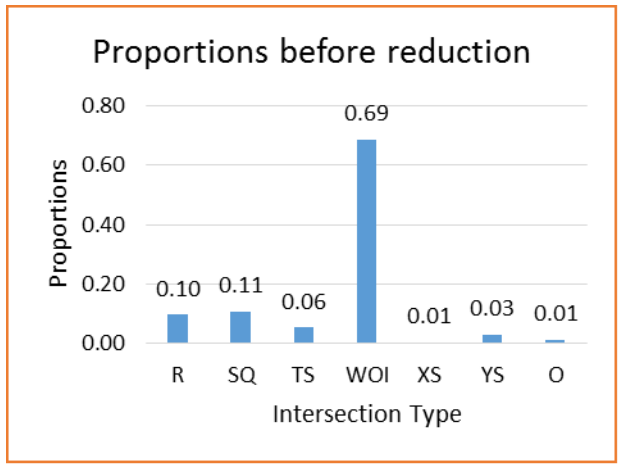
$$P_1 = \frac{x_1}{n_1} \;\&\; P_2 = x_2/n_2 \qquad \hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \qquad P - value = (1 - \Phi(Z))$$

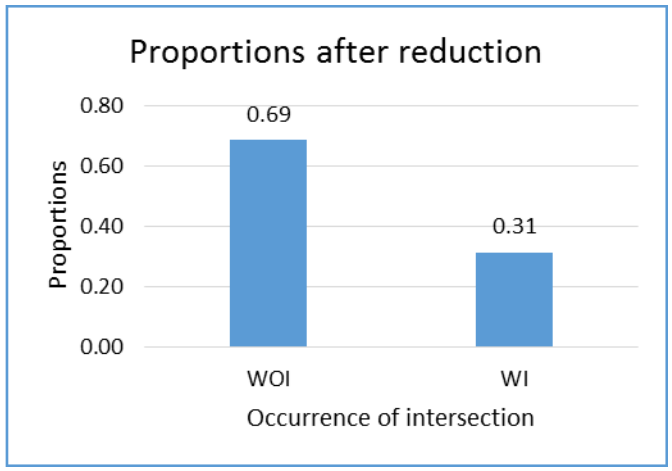
			Injuries	PDO										
Total Crash Observat	5,251.00		919.00	4,332.00										
	Injuries		Property damage											
Road type	Frequenc y	%age	Frequenc y	%age	Number of observatio ns in PDO n1	Number of observatio ns in INJ n2	Proportio n1 p1	Proportio n2 p2	difference between proportions p1-p2	Weighted value of P1 and P2 P^	Estimated Standared Error Sp1-p2	Z-stat Value Z	One tailed P-value P-value	Remark
OW	251.00	0.27	1,330.00	0.31	4,332.00	919.00	0.27	0.31	0.03	0.28	0.02	2.08	**0.02**	Selected
TW	440.00	0.48	1,490.00	0.34	4,332.00	919.00	0.48	0.34	0.13	0.46	0.02	7.45	**0.00**	selected
IS	201.00	0.22	1,361.00	0.31	4,332.00	919.00	0.22	0.31	0.10	0.24	0.02	6.20	**0.00**	selected
SLS	14.00	0.02	64.00	0.01	4,332.00	919.00	0.02	0.01	0.00	0.02	0.00	0.10	**0.46**	merged
BLS	14.00	0.02	87.00	0.02	4,332.00	919.00	0.02	0.02	0.00	0.02	0.00	1.06	**0.14**	merged

Appendix I: Study variables before and after reduction (Sample)









Appendix J: Multinomial logit model results

Parameter Estimates

Severity_level[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Slight injury	Intercept	-4.554	3.653	1.554	1	.213			
	Weather condition	.417	.701	.353	1	.553	1.517	.384	5.996
	Lighting condition	-.275	.337	.665	1	.415	.760	.392	1.471
	Road surface condition	.124	.369	.114	1	.736	1.132	.550	2.333
	Pavement type	1.205	.567	4.513	1	.034	3.336	1.098	10.141
	Intersection occurrence	-.282	.269	1.102	1	.294	.754	.446	1.277
	Road geometry	-.068	.500	.018	1	.892	.934	.351	2.489
	LogAADT	.555	.393	1.990	1	.158	1.742	.806	3.764
	[Overturning=1]	-2.857	.309	85.474	1	.000	.057	.031	.105
	[Vehicle-inert=2]	-5.228	1.202	18.927	1	.000	.005	.001	.057
	[Vehicle-pedestrian=3]	-2.989	1.088	7.539	1	.006	.050	.006	.425
	[Vehicle-vehicle=4]	0[b]	.	.	0	.	.	.	.
	[During turning=1]	-2.404	.294	66.934	1	.000	.090	.051	.161
	[Around inter=2]	-1.167	1.192	.959	1	.328	.311	.030	3.219
	[Maneuvering=3]	-2.484	.582	18.205	1	.000	.083	.027	.261
	[Straight through=4]	-.454	1.151	.156	1	.693	.635	.067	6.056
	[Others=5]	0[b]	.	.	0	.	.	.	.

2.00	[divided two-way=1]	-.067	.420	.025	1	.873	.935	.411	2.130
	[one-way=2]	-.576	.588	.961	1	.327	.562	.178	1.778
	[undivided two-way=3]	0[b]	.	.	0	.	.	.	.
	[Around business district=1]	-.449	.377	1.417	1	.234	.638	.305	1.337
	[Around Factory=2]	.038	.453	.007	1	.933	1.039	.428	2.523
	[Around non-business district=3]	-.361	.363	.992	1	.319	.697	.342	1.419
	[Around Residence=4]	0[b]	.	.	0	.	.	.	.
	Intercept	-9.193	3.791	5.881	1	.015			
	Weather condition	-.676	.840	.649	1	.421	.508	.098	2.637
	Lighting condition	.667	.254	6.890	1	.009	1.949	1.184	3.207
	Road surface condition	-.537	.391	1.881	1	.170	.585	.271	1.259
	Pavement type	-.977	.681	2.058	1	.151	.376	.099	1.430
	Intersection occurrence	-.257	.260	.979	1	.322	.773	.465	1.287
	Road geometry	-.680	.485	1.971	1	.160	.506	.196	1.309
	LogAADT	1.309	.391	11.190	1	.001	3.701	1.719	7.967
	[CRT=1]	-3.367	.252	178.879	1	.000	.035	.021	.057
	[CRT=2]	-6.722	1.035	42.202	1	.000	.001	.000	.009
	[CRT=3]	-1.727	.537	10.356	1	.001	.178	.062	.509
	[CRT=4]	0[b]	.	.	0	.	.	.	.
	[MAC=1]	-1.283	.288	19.887	1	.000	.277	.158	.487
	[MAC=2]	3.592	.484	55.013	1	.000	36.295	14.049	93.764

[MAC=3]	-.757	.402	3.554	1	.059	.469	.213	1.030
[MAC=4]	-.348	1.458	.057	1	.811	.706	.041	12.286
[MAC=5]	0[b]	.	.	0	.	.	.	.
[ROT=1]	.526	.410	1.646	1	.199	1.693	.758	3.783
[ROT=2]	.603	.509	1.404	1	.236	1.827	.674	4.952
[ROT=3]	0[b]	.	.	0	.	.	.	.
[LAU=1]	.118	.315	.140	1	.708	1.125	.607	2.086
[LAU=2]	-.093	.419	.049	1	.825	.911	.401	2.074
[LAU=3]	-.003	.304	.000	1	.993	.997	.550	1.810
[LAU=5]	0[b]	.	.	0	.	.	.	.

a. The reference category is: 3.00.

b. This parameter is set to zero because it is redundant.