ABSTRACT

Title: WEATHER IMPACT ON ROAD ACCIDENT SEVERITY IN MARYLAND

YUE LIU, M.S. in Civil and Environmental Engineering, 2013

Directed By: Professor and Chair Ali Haghani, Department of Civil and Environmental Engineering

This study was conducted to analyze and quantify the impact of weather factors on road accident severity, based on Maryland accident data during 2007-2010. In order to find a better model fitted related variables, three candidate models multinomial logit (MNL), ordered probit logit (OP), and neural networks were chosen to examine in SAS. The results showed that the Multilayer Perceptron Model in neural networks performed the best and is the accident severity model of choice.

During the model construction, eight factors related to weather condition were considered. They were: air temperature, average wind speed, total precipitation in the past 24 hours, visibility, slight, moderate, heavy precipitation and relative humidity. Based on the comparison criteria, we concluded that MNL regression is more interpretive than OP and Neural Networks models. All factors except visibility and heavy precipitation had significant impact on accident severity when considering the data from the entire Maryland highway system. Using MNL, a data subset with accident records only in a section of US route 50 was examined. After excluding the impact factors other than weather, a narrow significant variable set was obtained.

WEATHER IMPACT ON ROAD ACCIDENT SEVERITY IN MARYLAND

By

Yue Liu

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2013

Advisory Committee:
Professor Ali Haghani, Chair
Professor Paul Schonfeld
Assistant Professor Cinzia Cirillo

Acknowledgements

I am deeply grateful to my advisor Professor Ali Haghani for his dedication and support throughout my research. This study would not have been possible without his guidance and help. This journey of research allowed me to learn his ways of thinking and working and I am often impressed by his logical thoughts and wise solutions to difficult research questions.

I feel so honored to have Dr. Paul Schonfeld and Dr. Cinzia Cirillo to be my thesis committee members. Thank you all for your time and valuable guidance on my thesis.

Also, I would like to thank my colleagues, especially Dr. Masoud Hamedi and Ms. Azadeh Norouzi, for instructing and motivating me to complete this thesis. Without their efforts in acquiring and integrating data, I couldn’t finish this thesis so smoothly.

Finally, I want to thank my parents for their love, encouragement and strong support.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Research Motivation and Objectives

Recent statistics provided by Maryland State Highway Administration indicate that highway accident rates have decreased in recent years. However, due to high speeds and large traffic flow volumes, the fatality rate is still two times higher than that on local roads. According to an announcement by the Maryland State Highway Administration, despite a national decline in fatal crashes, Maryland fatalities increased between 2005 and 2006. In 2006, 101,889 motor vehicle crashes, or one every five minutes, occurred on Maryland's roadways, resulting in 53,615 injuries and 652 lives lost. From 2007 to 2011, these crashes had cost Maryland residents over $44 billion (Cumberland Area Long-Range Transportation, 2011).

The ultimate goal of accident analysis is to improve an agency’s ability to make future decisions in all components of a highway safety plan. These decisions can be aided by conducting formal effectiveness and administrative evaluations of ongoing and completed highway safety projects and programs. Analysis involves obtaining and processing quantitative information on the benefits and costs of implementing highway safety improvements. Estimations of benefits and costs reduce the dependence on engineering judgment and increase the ability of the agency to plan and implement future highway safety improvements which have the highest probability for success. Thus, scarce safety funds can be properly allocated to high pay-off improvements and diverted from those which are marginal or ineffective.

The accident prediction module in the previous studies (Randa Oqab Mujalli, 2011; Juan de O˜ na, 2011; Sunil Patil, 2011; Ali Tavakoli Kashani, 2012; and Ali S. Al-Ghamdi, 2002) estimated the expected crash severity on a highway using weather, road and traffic characteristics. This helped users to evaluate an existing highway, compare the relative safety performance of design alternatives, and assess the safety and cost-effectiveness of design decisions. Among all evaluating indicators, weather-related highway performance is a crucial aspect in measuring the safety of highway system. Severe weather conditions may have various impacts on the transportation system, involving the impacts on vehicle conditions, road conditions, and driver behavior (Osoro Mogaka Eric, 2011). These weather events can affect the transportation system both directly and indirectly (Federal Highway Administration Report, 1999). Especially in winter, heavy rains, snow, storms and freezing temperatures can result in a higher frequency of car crashes, and will also have higher opportunities to cause traffic congestion. On the other hand, people’s reactions to severe weather conditions may also lead to increased fuel consumption, delays and number of accidents (A. T. Kashani, 2009).

This research aims to achieve the overall goal of developing a better understanding of the impacts of weather on traffic accident severity and giving readers a perspective view of which weather elements impact accident severity significantly and how much the impact is.

The accident and log of messages data in the study period was acquired from the Center for Advanced Transportation Technology (CATT) Laboratory at the University of Maryland, College Park, and from the Coordinated Highway Action Response Team (CHART) reports.

The database was filtered and cleaned up. Weather condition databases were acquired from DOT archived data. The study area was divided into 5 regions and the nearest central weather tower station in each region was assigned to represent the weather condition in each region. The weather database was accumulated for a four-year study period and then joined to the main database based on closest weather tower station to the time and location of accident. In the case study part, subsets of data were filtered on certain sections with higher fatalities on US50. Running models on the selected section excluded the impact of traffic factors and made the result more accurate.

1.2 Organization of the Thesis

This thesis is organized in six chapters. The first two chapters focus on the macroscopic analysis and give the overview of the weather-related accident statistics and how the traffic elements are influenced by severe weather components. The third chapter introduces the three most widely used methodologies in constructing the relationship between weather and accident severity. It provides a comprehensive review of mathematical applications and uses goodness of fit methods to evaluate the different models. Chapter 4 briefly analyzes the content of accident and weather databases. Chapter 5 uses the three different models introduced in Chapter 3 to construct estimation models. The comparisons of their performances are given in Chapter 6, along with the case study on a selected road section. Conclusions and directions for future work follow in Chapter 7.

Chapter 2: Background and Literature Review

2.1 Weather-related Factors Impacting Accident Severity

2.1.1 Overall Impact Analysis

Weather-related crashes are defined by FHWA as those crashes that occur in adverse weather (i.e., rain, sleet, snow, and/or fog) or on slick pavement (i.e., wet pavement, snowy/slushy pavement, or icy pavement). Based on the fourteen-year's NHTSA data, 24% of crashes — approximately 1,511,000—on average are weather-related each year, and 7,130 people are killed and over 629,000 people are injured in weather-related crashes each year. Among all weather-related crashes, 75% happen on wet pavement and 47% happen during rainfall, which makes rain a major factor. Meanwhile, 15% of crashes happen during snow and only 3% happen in foggy weather. In terms of accident severity, among all weather-related accidents, 41.6% involve personal injury and 0.47% cause fatalities. Detailed statistics are shown in Table 2.1 (Road weather management program, FHWA, 2012).

Table 2. 1 Weather-Related Crash Statistics (Annual Averages)

Road Weather Conditions	Weather-Related Crash Statistics		
	Annual Rates (Approximately)	Percentages	
Wet Pavement	1,128,000 crashes	18% of vehicle crashes	75% of weather-related crashes
	507,900 persons injured	17% of crash injuries	81% of weather-related crash injuries
	5,500 persons killed	13% of crash fatalities	77% of weather-related crash fatalities
Rain	707,000 crashes	11% of vehicle crashes	47% of weather-related crashes
	330,200 persons injured	11% of crash injuries	52% of weather-related crash injuries
	3,300 persons killed	8% of crash fatalities	46% of weather-related crash fatalities
Snow/Sleet	225,000 crashes	4% of vehicle crashes	15% of weather-related crashes
	70,900 persons injured	2% of crash injuries	11% of weather-related crash injuries
	870 persons killed	2% of crash fatalities	12% of weather-related crash fatalities
Icy Pavement	190,100 crashes	3% of vehicle crashes	13% of weather-related crashes
	62,700 persons injured	2% of crash injuries	10% of weather-related crash injuries
	680 persons killed	2% of crash fatalities	10% of weather-related crash fatalities
Snow/Slushy Pavement	168,300 crashes	3% of vehicle crashes	11% of weather-related crashes
	47,700 persons injured	2% of crash injuries	8% of weather-related crash injuries
	620 persons killed	1% of crash fatalities	9% of weather-related crash fatalities
Fog	38,000 crashes	1% of vehicle crashes	3% of weather-related crashes
	15,600 persons injured	1% of crash injuries	2% of weather-related crash injuries
	600 persons killed	1% of crash fatalities	8% of weather-related crash fatalities

There are a number of factors that could cause road accidents during a bad weather condition. These include road condition, vehicle condition, and driver behavior, as shown in Figure2.1. Weather impacts traffic through several ways, among which visibility, precipitation, wind speed, and temperature are of most concern. Severe weather conditions affect drivers' capabilities, vehicles' stability and pavement's friction. (Kilpeläinen M., 2007). On the other hand, severe weather conditions also cause chaos in traffic flow and slow down the speed with which emergency response agencies can react. Hence, weather condition has significant impact in accident severity (Brodsky, 1988).

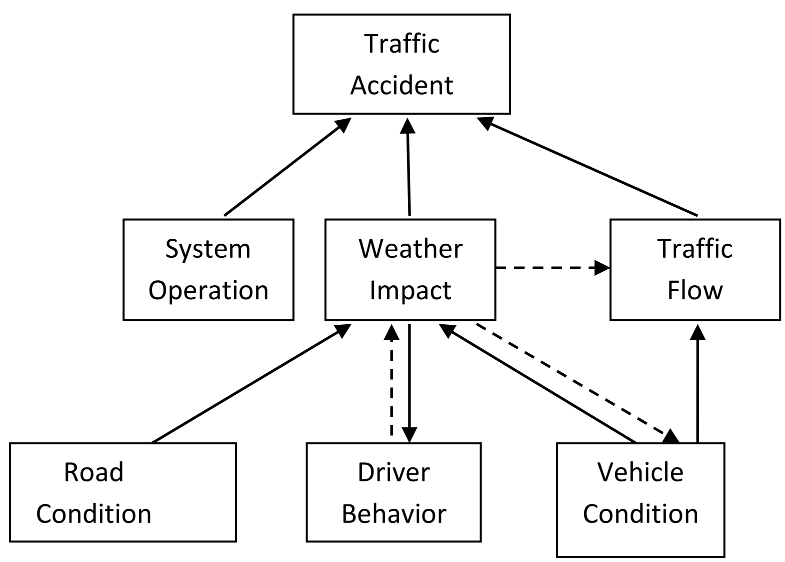

Figure 2.1 Factors in traffic accidents

In countries with severe winters and high annual precipitation, such as Canada and the UK, winter road safety is a source of concern for transportation officials. Driving conditions in winter can deteriorate and vary dramatically due to snowfall and ice formation, causing significant reduction in pavement friction and increasing the risk of accidents.

2.1.2 Impact on Vehicle Condition

Extreme weather has an effect on car equipment, especially electric cars, because it makes it difficult for cars to maintain an equivalent amount of electricity in winter to match their output in normal temperatures. The adverse impact of weather on cars' performance can be detected in various aspects, such as environment temperature, battery type, whether the car is designed to manage the battery's temperature and how well the condition of the car is maintained. In general, vehicles and car components suffer much more during extreme weather than on normal days, because difficult weather causes a lot of wear and tear on engine parts and transmission components. Salt that is often used as a de-icer during the winter months can also be very detrimental to vehicles. Moreover, wipers have higher probability to deteriorate very fast during freezing temperatures and snow and icy conditions, and the worst condition is that most normal wipers will simply stop working through layers of snow and ice (Jennifer Geiger, http://www.howstuffworks.com).

2.1.3 Impact on Road Condition

A thorough review of the previous studies indicates that both rain and snow, functioning as precipitation, can lead to a higher level of car accidents. For instance, Norrman et al. (2000) identified that the number of accidents is large on slippery roads. Winter road maintenance with massively direct and indirect costs, has stimulated significant interest in quantitative cost-benefit assessment of the system's security and mobility. In the past decade, a lot of research has been launched to determine the link between winter road safety, maintenance operations, and weather-related factors (Andreescu and Frost, 1998; Andrey et al., 2001; Handman, 2002; Knapp et al., 2002; Kumar and Wang, 2006). Severe weather can increase the cost of operation and maintenance of roads during winter for highway agencies, traffic management agencies, emergency management agencies, law enforcement agencies, and commercial vehicle operators (CVOs). Winter road maintenance accounts for about 20% of the state DOT budgets. Each year, more than 2.3 billion dollars are spent on snow and ice control operations by state and local agencies (Road weather management program, FHWA, 2012).

2.1.4 Impact on Driver Behavior

Older drivers are at relatively higher risk of causing collisions. Due to the fact that age is highly related with reductions in contrast sensitivity (Scialfa and Kline, 2007) and increases in response time even in well-practiced tasks (Voelcker-Rehage and Alberts, 2007), wet roads and fog can be particularly problematic for older drivers. Fog reduces contrast of the image. This affects the sense of distance, which can prompt rear end collisions (Broughton et al., 2007; Buchner et al., 2006). It also leads to an underestimation of how fast other vehicles are (Horswill and Plooy, 2008). Moreover, because objects have to be closer to become fully visible, fog reduces the amount of drivers’ reacting stimulation time. The increase of collision risk in fog puts even professional drivers under stress while driving in fog (Vivoli et al., 1993). Older drivers’ response to visibility challenges in the bad driving environment can be analyzed from both psychological and physiological aspects (Lana M. Trick, 2010, Scialfa, C. T., 1999).

2.1.5 Impact on Traffic Flow

Anyone who uses ground transportation has been affected by delays caused by various forms of weather. Whether it is rain or snow, ice or fog, the result in traffic flow is usually the same.

Travel delay increases as traffic flow slows down. Capacity can be reduced by lane submersion due to flooding and by lane obstruction due to accumulation of snow and debris. Due to road closures and access restrictions in hazardous conditions (e.g., large trucks in high winds), roadway capacity will also decrease.

On highways, slight rain or snow can reduce average speeds by 3 to 13%. Heavy rain can reduce average speeds by 3 to 16%. During heavy snow, average speeds can decline by 5 to 40%. Low visibility can cause speed reductions of 10 to 12%. Free-flow speed can be reduced by 2 to 13% in light rain and by 6 to 17% in heavy rain. Snow can cause free-flow speed to decrease by 5 to 64%. Speed variance can fall by 25% during rain. Light rain can decrease freeway capacity by 4 to 11% and heavy rain can cause capacity reductions of 10 to 30%. Capacity can be reduced by 12 to 27% in heavy snow and by 12% in low visibility. Light snow can decrease flow rates by 5 to 10%. Maximum flow rates can decline by 14% in heavy rain and by 30 to 44% in heavy snow (*Highway Capacity Manual 2000*). Details are shown in Table 2.2.

Table2.2: Freeway Traffic Flow Reductions due to Weather

Weather Conditions	**Freeway Traffic Flow Reductions**			
	Average Speed	**Free-Flow Speed**	**Volume**	**Capacity**
Light Rain/Snow	3% - 13%	2% - 13%	5% - 10%	4% - 11%
Heavy Rain	3% - 16%	6% - 17%	14%	10% - 30%
Heavy Snow	5% - 40%	5% - 64%	30% - 44%	12% - 27%
Low Visibility	10% - 12%			12%

To summarize, Table 2.3 from FHWA shows how weather factors impact on all four aspects (See http://www.ops.fhwa.dot.gov/weather/q1_roadimpact.htm).

Table2.3 : Weather Impacts on Roads, Traffic and Operational Decisions			
Road Weather Variables	**Driver Behavior Impacts**	**Traffic Flow Impacts**	**Road and Vehicle Condition Impacts**
Air temperature and humidity	N/A	N/A	• Road treatment strategy (e.g., snow and ice control) • Construction planning (e.g., paving and striping)
Wind speed	• Visibility distance (due to blowing snow, dust) • Lane obstruction (due to wind-blown snow, debris)	• Traffic speed • Travel time delay • Accident risk	• Vehicle performance (e.g., stability) • Access control (e.g., restrict vehicle type, close road) • Evacuation decision support
Precipitation (type, rate, start/end times)	• Visibility distance • Lane obstruction • Driver capabilities/ behavior	• Roadway capacity • Traffic speed • Travel time delay • Accident risk	• Vehicle performance (e.g., traction) • Road treatment strategy • Traffic signal timing • Speed limit control • Evacuation decision support • Institutional coordination • Pavement friction
Fog	• Visibility distance • Driver capabilities/ behavior	• Traffic speed • Speed variance • Travel time delay • Accident risk	• Road treatment strategy • Access control • Speed limit control
Pavement temperature	N/A	N/A	• Road treatment strategy • Infrastructure damage
Pavement condition	• Driver capabilities/behavior (e.g., route	• Roadway capacity • Traffic speed	• Infrastructure damage • Pavement friction

	choice)	• Travel time delay • Accident risk	• Vehicle performance • Road treatment strategy • Traffic signal timing • Speed limit control
Water level	N/A	• Traffic speed • Travel time delay • Accident risk	• Lane submersion • Access control • Evacuation decision support • Institutional coordination

2.2 Accident Severity Models

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). Palutikof (1991) stated that rain had significant effect in increasing the probability of traffic fatalities, while Sherretz & Farhar (1978) used a general linear regression model to indicate that accident rates increased when high quantity of precipitation occurred. During rainy days, the number of accidents can increase by 6% (Brotsky & Hakkert 1988). The correlated models connected with road geometric factors were usually created using multivariate analysis (Ogden, et al., 1994; Ogden and Newstead, 1994; Vogt, 1999). Corben and Foong, (1990) developed a multivariate linear regression model to predict accident rates at signalized intersections. 85% of the variance can be explained in this model, showing that the model performed well. In a FHWA study by Harwood, et al. (2000), experts combined quantitative data on accidents and other factors with the expert's judgment on design factors and expected impact of these design factors on the accident rate.

A multivariate logistic regression was 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. Evanco (1999) used a multivariate statistical analysis based on population to discuss the relationship between fatalities and accident response time.

The application of Artificial Intelligence techniques to analyze transportation problems is fairly recent. Abdelwahab et al. (2001) evaluated the performance of Neural Network (NN) conducted with the Levenberg-Marquardt algorithm and compared it with an ordered logit model. The results showed that the NN model (65.6% and 60.4% classification accuracy for the training and testing phases) performed better than ordered logit model (58.9% and 57.1% classification accuracy for the training and testing phases).

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 (Jianming Ma, 2006). However, there were only very few studies that included weather related characteristics in the analysis of accident severity. This is the focus of this research and will be described next.

Chapter 3: Methodology

3.1 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 utility function is basically the same as in a generalized linear regression model regarding the assumption that each error ε_{ki} for severity level k for observation i is independently-identically-distributed extreme value following a Gumbel distribution.

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,

$$S_{ik} = \boldsymbol{\beta}_k \mathbf{X}_{ik} + \varepsilon_{\mathrm{ik}} \tag{3.1}$$

The probability function for observation i ending in accident severity level k is:

$$P_i(k) = \frac{\exp(\alpha_k + \beta_k X_{ki})}{\sum_{k=1,2,3} \exp(\alpha_k + \beta_k X_{ki})} \tag{3.2}$$

The estimation of the model parameters can be carried out through the method of maximum likelihood. In addition to not accounting for the ordering of injury-severity outcomes, the multinomial logit model is particularly susceptible to correlation of unobserved effects from one injury severity level to the next. Such correlation causes a violation of the model's independence of irrelevant alternatives (IIA) property. On the plus side, traditional multinomial models do not impose the sometimes unrealistic parameter restrictions that traditional ordered probability models do. Further, if the IIA property holds, it can be shown that in the presences of underreporting of crashes all parameters will still be correctly estimated except for the constant term (see Washington et al., 2011).

3.3 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. The most common approach to the derivation of such models is to start by specifying a latent variable, Z, which is used as a basis for modeling the ordinal ranking of data. This unobserved variable is most often specified as a linear function for each crash observation, such that $Z = \beta X$, where X is a vector of variables determining the discrete ordering for each crash observation, β is a vector of estimable parameters, and ε is a disturbance term (Washington et al., 2011). With this, observed ordinal injury data, y, for each observed crash are defined as,

$$
\begin{aligned}
y &= 1 \text{ if } \mu_0 < z \le \mu_1, \\
y &= 2 \text{ if } \mu_1 < z \le \mu_2, \\
y &= 3 \text{ if } \mu_2 < z \le \mu_3, \qquad (3.3) \\
y &= \ldots. \\
y &= k \text{ if } \mu_{k-1} < z \le \mu_k,
\end{aligned}
$$

where the μ are estimable threshold parameters that define *y*, which corresponds to integer ordering and *k* is the highest integer ordered response. The μ are parameters that are estimated jointly with the model parameters β and, without loss of generality, μ_0 can be set to 0. The estimation problem then becomes one of determining the probability of *k* specific ordered responses for each crash observation, *i*. If the error term, ε, is assumed to be normally distributed across observations with a mean of zero and variance of one, an ordered probit model results. Setting the lower threshold, μ_0, equal to zero results in the outcome probabilities

$$P(y_i = k) = \Phi(u_k - \beta X) - \Phi(u_{k-1} - \beta X) \tag{3.4}$$

where μ_i and μ_{i-1} represent the upper and lower thresholds for injury severity *i*. Likewise, if the errors are instead assumed to be logistically distributed across observations, an ordered logit model results (Abdel-Aty, M., 2003).

3.4 Neural Network

Neural Networks, also known as Artificial Neural Networks, are usually discussed in terms of minimizing an error measure such as the least-squares criterion. The basic concept of NN is to build the data modeling process through an analogy to human brain behavioral characteristics, by applying parallel information processing algorithms. Such networks achieve the purpose by adjusting the large numbers of mutual connections between internal nodes, relying on the complexity of the brain system. ANN is a simulation of human thinking, which can be considered as a nonlinear dynamical system. The key point is to co-process information storage and perform parallel analysis. Although the structure of individual neurons is extremely simple and has limited functions, the behaviors the network system can achieve are colorful (Dursun Delen, 2006). Because the neural network can process massively parallel transmission and has the ability of self-organizing, adapting, and self-learning, it is particularly suitable to deal with the problems with many factors and conditions, especially problems with imprecise and vague information (Darçin, 2010).

Figure 3.1 is a schematic neuron flow, where a_1 –a_n are input vectors, w_1 –w_n are weights for each neural synapse, b is bias, and f is a transmission function, usually nonlinear. The output t can be formulated as

$$t = f(\overline{w}\overline{A'} + b) \tag{3.5}$$

Most NN training algorithms follow a similar scheme with:

- random initial estimates
- simple case-by-case updating formulas
- slow convergence or no convergence (Fouad N. Shoukry, 2005)

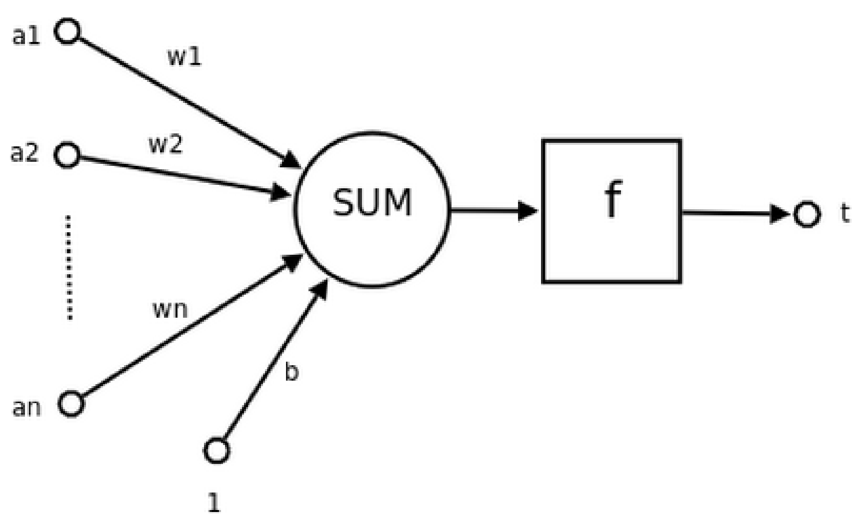

Figure 3.1 Neuron flow

Multiple neurons can compose the neural network in many ways, among which the most widely used expression is multilayer perceptron (MLP) shown in Figure 3.2.

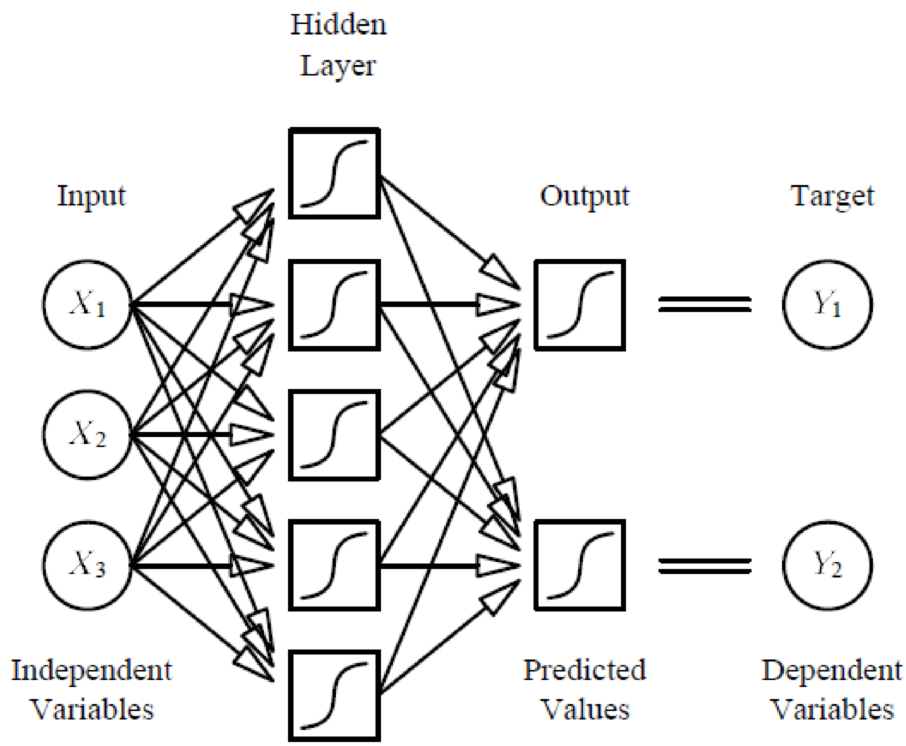

Figure 3.2. Multivariate Multiple Nonlinear Regression

Multilayer perceptrons (MLP) are general-purpose, flexible, nonlinear models that, given enough hidden neurons and enough data, can approximate virtually any function to any desired degree of accuracy. Multilayer perceptrons can be used when we have little knowledge about the form of the relationship between the independent variables and dependent variables (D. Chimba, 2009).

3.5 Model Evaluation and Selection

In general, based on the previous study, trade-offs of ordering the model have been considered, and new methods to take unobserved heterogeneity into account have been opened up in recent applications and models. The previous literatures (Kunt, 2012, Savolainen, 2011, Zhang, 2010, Sze, 2007, Nassiri, 2006, and Abdel-Aty, 2004) also 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 (Fan Ye, 2011). 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. But considering most recent works, an artificial neural network was applied more frequently and when compared with OP and MNL models, it performed better in Abdelwahab's research in 2001.

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, and the crucial weather-related factors have yet to be investigated under similar traffic and geometric environment. Detailed model comparison based on mean square of errors and log likelihood will be constructed in Chapter 5.

Chapter 4: Data

The initial data resource was provided by the Center for Advanced Transportation Technology (CATT) Laboratory in the Department of Civil and Environmental Engineering at the University of Maryland, College Park, and the Coordinated Highway Action Response Team (CHART) that reports for regions within the District of Columbia in Maryland, and Maryland Department of Transportation, State Highway Administration (SHA) and DOT archived data. The data were collected, selected and filtered by Norouzi (2012).

The study area is the roadway network in the State of Maryland. Figure 4.1 from Norouzi’s work shows the entire accident records in the study area.

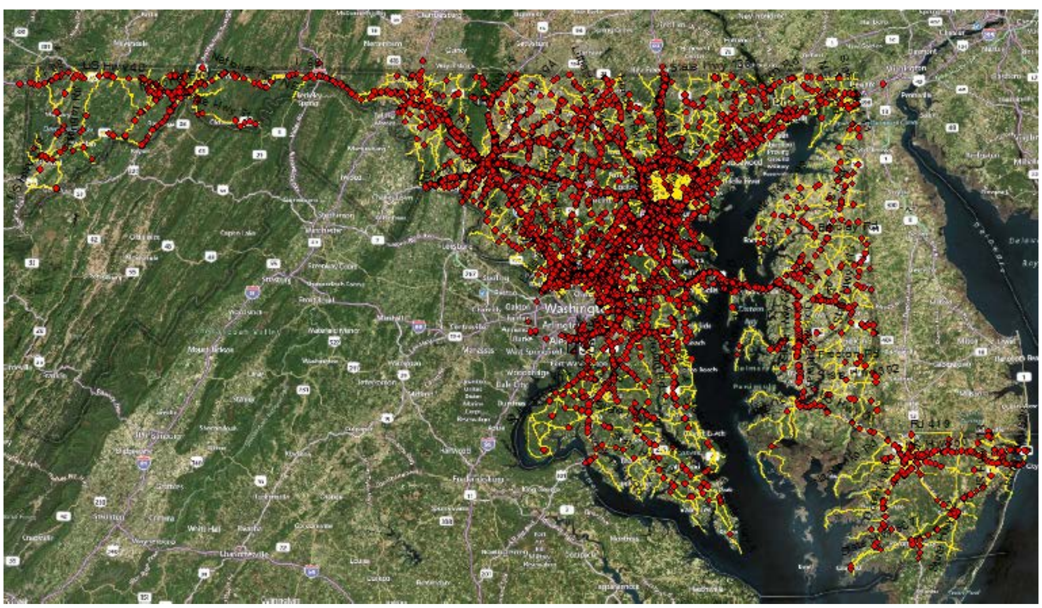

Figure 4.1 Accident records in study area

4.1 Accident Severity Data

After cleaning the initial data by removing data gaps and outliers, the number of accidents shrank from 38,718 to 20,469 for the four-year period of 2007 to 2010 in the entire State of Maryland.

The data set consisted of accident type (property damage, personal injury and fatality), road section, longitude and latitude, jurisdiction, time and other related information. Due to confidentiality concerns, access to police records and accident causes was not possible.

Locations of accidents were pinpointed on road network map for further analysis. The data set was restricted to freeways with higher accident rate, which allowed reducing the variability in the data set (Norouzi, 2012). Figure 4.2 shows the initial content of accident data.

	D	E	F	G	H	I	J	K	L	N	O	P	Q	R	S	T	U
1	agency_typ	location	time	date_	county_de	COUNTY_I	COUNTY_I	ID_PREFIX	ID_RTE_N	ID_MP	STATION_	FUNC_CLA	FUNC_CLA	BEGIN_SE	END_SECT	ROAD_SE	AADT_2
2	Collision, Property damage	MD 234 W	6:36:19	3/13/2010	Saint Mary	18	St Marys	MD	234	4.6	MD234-.30	6	RURAL MI	0.88	4.9	MD 236 TC	73
3	Collision, Personal injury	US 340 EA:	16:40:23	3/11/2010	Washingt	21	Washingt	US	340	0.44	US 340 at I	2	RURAL OT	0	2.27	VIRGINIA :	208
4	Collision, Personal injury	I-83 NORT	12:45.9	3/2/2010	Baltimore	3	Baltimore	RP	83	0.1	EXIT 16 RA	11	URBAN IN	0	0.18	HARRISBU	
5	Collision, Property damage	I-83 AT MT	18:24:09	3/19/2010	Baltimore	3	Baltimore	MD	137	3.08	MD137-.20	6	RURAL MI	1.87	7.94	MD 25 TO	52
6	Collision, Property damage	I-83 NORT	16:04:32	3/25/2010	Baltimore	3	Baltimore	MD	137	3.08	MD137-.20	6	RURAL MI	1.87	7.94	MD 25 TO	52
7	Collision, Property damage	I-695 OUTI	12:11:22	4/15/2010	Anne Arur	2	Anne Arur	RP	695	0.14	EXIT 3A RA	12	URBAN OF	0	0.17	BALTO BEI	
8	Collision, Personal injury	MD 2 NOR	17:32:06	4/20/2010	Anne Arur	2	Anne Arur	RP	695	0.14	EXIT 3A RA	12	URBAN OF	0	0.17	BALTO BEI	
9	Collision, Property damage	I-95 NORT	18:42:36	3/2/2010	Prince Ge	16	Prince Ge	MD	5	12.25	MD5-.30 N	12	URBAN OF	11.95	13.47	IS 95 TO M	656
10	Collision, Personal injury	I-95 INNEF	19:01:12	3/11/2010	Prince Ge	16	Prince Ge	MD	5	12.25	MD5-.30 N	12	URBAN OF	11.95	13.47	IS 95 TO M	656
11	Collision, Personal injury	I-95 INNEF	18:08:40	3/19/2010	Prince Ge	16	Prince Ge	MD	5	12.25	MD5-.30 N	12	URBAN OF	11.95	13.47	IS 95 TO M	656
12	Collision, Personal injury	I-95 OUTEI	18:25:58	3/19/2010	Prince Ge	16	Prince Ge	MD	5	12.25	MD5-.30 N	12	URBAN OF	11.95	13.47	IS 95 TO M	656
13	Collision, Property damage	I-95 INNEF	16:43:37	3/25/2010	Prince Ge	16	Prince Ge	MD	5	12.25	MD5-.30 N	12	URBAN OF	11.95	13.47	IS 95 TO M	656
14	Collision, Property damage	I-95 INNEF	16:00:50	4/2/2010	Prince Ge	16	Prince Ge	MD	5	12.25	MD5-.30 N	12	URBAN OF	11.95	13.47	IS 95 TO M	656
15	Collision, Property damage	MD 5 NOR	12:01:20	4/19/2010	Prince Ge	16	Prince Ge	MD	5	12.25	MD5-.30 N	12	URBAN OF	11.95	13.47	IS 95 TO M	656
16	Collision, Personal injury	I-95 OUTEI	18:45:28	4/26/2010	Prince Ge	16	Prince Ge	MD	5	12.25	MD5-.30 N	12	URBAN OF	11.95	13.47	IS 95 TO M	656
17	Collision, Personal injury	MD 924 N	04:16.7	4/13/2010	Harford	12	Harford	MD	924	2.18	MD924-.10	16	URBAN MI	0	4.17	TOLLGATE	185
18	Collision, Personal injury	MD 10 NO	12:18:27	4/20/2010	Anne Arur	2	Anne Arur	MD	10	3.2	MD10-.40	12	URBAN OF	1.71	3.6	MD 177 TC	466
19	Collision, Property damage	MD 10 NO	18:58:16	4/20/2010	Anne Arur	2	Anne Arur	MD	10	3.2	MD10-.40	12	URBAN OF	1.71	3.6	MD 177 TC	466
20	Collision, Personal injury	I-95 SOUTI	8:44:12	3/3/2010	Prince Ge	16	Prince Ge	MD	212	8.08	POWDER I	16	URBAN MI	8.08	8.365	IS 95 TO P	
21	Collision, Personal injury	I-95 SOUTI	17:54:53	3/18/2010	Prince Ge	16	Prince Ge	MD	212	8.08	POWDER I	16	URBAN MI	8.08	8.365	IS 95 TO P	
22	Collision, Personal injury	I-95 NORT	8:54:21	3/22/2010	Prince Ge	16	Prince Ge	MD	212	8.08	POWDER I	16	URBAN MI	8.08	8.365	IS 95 TO P	
23	Collision, Personal injury	I-95 NORT	8:51:20	3/23/2010	Prince Ge	16	Prince Ge	MD	212	8.08	POWDER I	16	URBAN MI	8.08	8.365	IS 95 TO P	

Figure4.2 Initial content of accident data

4.2 Weather Data

The weather data for this research were merged from the records of different regions published online by Maryland DOT. The initial format of the database was in the shape of month to month archived data collected from 49 weather tower stations and contained the following data fields: date and time, air temperature, humidity, average wind speed, wind gust, wind direction, precipitation type, precipitation intensity (light, medium, heavy), precipitation accumulation, rate (rate per hour in inches), visibility (miles) and surface temperature (see www.chart.state.md.us/). For simplicity, the area of research was divided into 5 regions of north, south, west, east and Washington, DC. The nearest central weather tower station in each region was assigned to represent the weather condition in that region. For instance, the weather station "I-68_Cumberland" was assigned to west region, "US 50 Kent Narrow Bridge" was assigned to east region, "I-895_Levering_Ave" was assigned to north region, "US-301_Potomac" was assigned to south region, and "I-270_I-370" was assigned to Washington, DC region. The weather data set was also accumulated for the four-year period of study (2007-2010). Figure 4.3 shows the format of weather database.

Date / Time repor	WeatherStation	AirTemperature	RelativeH	AverageW	WindGust	WindDire	Percipitat	Visibility	SurfaceTe	Precipitati
12/30/2011 23:59	DMS_WEATHER_US301	55	99	4	9	E	None	1.1	56 - 56	0.79
12/30/2011 23:54	DMS_WEATHER_I270	55	91	3	11	SW	None	1.1	58 - 59	0
12/30/2011 23:51	DMS_WEATHER_I895	43	75	1	6	N	None	1.1	50 - 53	0
12/30/2011 23:44	DMS_WEATHER_I895	50	96	1	3	W	None	1.1	56 - 57	0
12/30/2011 23:40	DMS_WEATHER_I895	68	50	19	37	S	None	1.1	68 - 68	0
12/30/2011 23:39	DMS_WEATHER_I895	77	40	2	6	S	None	1.1	92 - 96	0
12/30/2011 23:29	DMS_WEATHER_I895	52	98	1	2	W	None	1.1	60 - 62	0
12/30/2011 23:28	DMS_WEATHER_I270	34	99	6	10	SW	None	1.1	35 - 37	0
12/30/2011 23:19	DMS_WEATHER_I270	54	92	6	11	SW	Rain	1.1	56 - 56	0
12/30/2011 23:18	DMS_WEATHER_I270	67	26	1	4	SW	None	1.1	61 - 62	0
12/30/2011 23:09	DMS_WEATHER_I270	59	26	1	4	SW	None	1.1	60 - 62	0
12/30/2011 23:08	DMS_WEATHER_I270	63	37	4	14	SW	Rain	1.1	63 - 64	0
12/30/2011 23:04	DMS_WEATHER_I270	65	37	2	9	SW	None	1.1	70 - 71	0
12/30/2011 22:57	DMS_WEATHER_I270	66	25	4	14	SW	Rain	1.1	89 - 95	0
12/30/2011 22:54	DMS_WEATHER_I270	55	97	1	4	SW	None	1.1	55 - 57	0.51
12/30/2011 22:49	DMS_WEATHER_I895	48	99	2	4	NE	None	1.1	49 - 52	0.03
12/30/2011 22:43	DMS_WEATHER_I895	69	44	2	6	NW	None	1.1	82 - 87	0
12/30/2011 22:39	DMS_WEATHER_I895	49	99	1	2	W	None	1.1	57 - 60	0
12/30/2011 22:33	DMS_WEATHER_I270	40	74	4	13	SW	Unidentifi	1.1	42 - 45	0
12/30/2011 22:29	DMS_WEATHER_I270	59	33	0	4	SW	None	1.1	56 - 58	0
12/30/2011 22:24	DMS_WEATHER_I270	60	97	2	11	SW	None	1.1	62 - 127	0
12/30/2011 22:23	DMS_WEATHER_I270	50	54	4	21	SW	None	1.1	65 - 70	0
12/30/2011 22:19	DMS_WEATHER_I270	56	22	0	1	SW	None	1.1	61 - 61	0
12/30/2011 22:14	DMS_WEATHER_I270	61	29	0	0	SW	None	1.1	58 - 59	0

Figure 4.3 Weather Database Format

After data cleansing process on more than ten thousand records, the outliers were filtered and removed. Two databases, accident and weather, were joined over two dimensions through GIS tools. To be specific, the time of each accident was matched with the weather status, and each accident record was assigned to the closest weather tower station and matched with the weather condition at the time of accident. The matching process was performed using SQL queries coded in C++ (Norouzi, 2012).

In the final dataset, precipitation is measured using a rain gauge. Intensity is classified according to the rate of precipitation. Light rain describes rainfall which falls at a rate of less than 1 millimeter (0.039 in) per hour; Moderate rain describes rainfall with a precipitation rate of between 1 millimeter (0.039 in) and 4 millimeters (0.16 in) per hour. Heavy rain describes rainfall with a precipitation rate of greater than 4 millimeters (0.16 in). Three dummy variables are used in the regression to describe four stages of precipitation intensity. Visibility is defined as the distance (in miles) at which an object or light can be clearly discerned.

Dependent variable "cost" is used to measure accident severity. According to the "Average Economic Cost per Death, Injury, or Crash" (National Safety Council, 2010), the approximately calculable costs of motor-vehicle crashes are wage and productivity losses, medical expenses, administrative expenses, motor vehicle damage, and employers' uninsured costs. The costs of all these items for each death (not each fatal crash), injury (not each injury crash), and property damage crash are:

- Death $1,410,000
- Nonfatal Disabling Injury $70,200
- Property Damage Crash (including non-disabling injuries) $8,900

In the modeling part of this thesis, the probability of each crash category was considered in the calculation. Usually, we use 0, 1, and 2 to represent three categories. However, in order to explain the weather effect on different categories better, and to prevent the misinterpretation of numerical relation, here we use the cost in dollar as the value of dependent variable. As described in Chapter 3, using dollar numbers will not affect the SAS outputs, since only the percentages of three categories involve in it, not the accurate cost numbers are used.

This research attempted to apply different models with the dependent variable accident cost and weather-related variables (air temperature, average wind speed, precipitation total, intensity dummy (slight, moderate, heavy), visibility, and relative humidity). A list of all independent variables is provided below in Table 4.1.

Table 4.1 Independent variables for the accident severity model

Variable Notification	Definition
Air Temperature (F°)	The temperature indicated by a thermometer placed in an instrument shelter 1.5 to two meters above ground.
Average WindSpeed (MPH)	The mean wind speed over a specified period of time.
Precipitation Total (mm)	Total precipitation amount in past 24 hours, including rain and snow
Slight	1 when the precipitation intensity is slight, 0 otherwise
Moderate	1 when the precipitation intensity is moderate, 0 otherwise
Heavy	1 when the precipitation intensity is heavy, 0 otherwise
Relative Humidity (%)	The amount of moisture in the air compared to what the air can "hold" at that temperature. It doesn’t necessarily indicate precipitation intensity
Visibility (miles)	The distance at which an object or light can be clearly discerned
Cost (thousand dollars)	Approximately calculable measurement for accident severity

Part of the final data used in this research is shown in Figure 4.4.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	cost	location	time	date_	AirTemperature	RelativeHumidity	AverageWindSp	PrecipitationTotal	slight	moderate	heavy	Visibility
2	1	70.2	MD 234 W	6:36:19	3/13/2010	49	96	0	0	0	0	0	1.1
3	2	8.9	US 340 EA	16:40:23	3/11/2010	51	98	0	0	0	1	0	1.1
4	3	8.9	I-83 NORT	12:45.9	3/2/2010	48	97	1	0	1	0	0	0.3
5	4	70.2	I-83 AT MT	18:24:09	3/19/2010	27	95	0	0	0	0	0	1.1
6	5	8.9	I-83 NORT	16:04:32	3/25/2010	32	85	2	0	0	0	0	1.1
7	6	70.2	I-695 OUTI	12:11:22	4/15/2010	31	88	1	0	0	0	0	1.1
8	7	70.2	MD 2 NOR	17:32:06	4/20/2010	49	99	13	0	0	0	0	0
9	8	8.9	I-95 NORT	18:42:36	3/2/2010	31	89	1	0	0	0	0	1.1
10	9	8.9	I-95 INNEF	19:01:12	3/11/2010	48	39	12	0	0	0	0	1.1
11	10	8.9	I-95 INNEF	18:08:40	3/19/2010	49	38	9	0	0	0	0	1.1
12	11	8.9	I-95 OUTEI	18:25:58	3/19/2010	32	85	0	0	0	0	0	1.1
13	12	70.2	I-95 INNEF	16:43:37	3/25/2010	49	35	9	0	0	0	0	1.1
14	13	70.2	I-95 INNEF	16:00:50	4/2/2010	40	43	1	0	0	0	0	1.1
15	14	70.2	MD 5 NOR	12:01:20	4/19/2010	31	89	0	0	0	0	0	1.1
16	15	70.2	I-95 OUTEI	18:45:28	4/26/2010	30	87	0	0	0	0	0	1.1
17	16	8.9	MD 924 N	04:16.7	4/13/2010	45	35	1	0	0	0	0	1.1
18	17	70.2	MD 10 NO	12:18:27	4/20/2010	41	99	3	0	0	0	0	0
19	18	8.9	MD 10 NO	18:58:16	4/20/2010	39	99	4	0	0	0	0	0
20	19	8.9	I-95 SOUTI	8:44:12	3/3/2010	46	51	2	0	0	0	0	1.1
21	20	8.9	I-95 SOUTI	17:54:53	3/18/2010	48	43	5	0	0	0	0	1.1
22	21	8.9	I-95 NORT	8:54:21	3/22/2010	53	99	5	0	0	0	0	0
23	22	8.9	I-95 NORT	8:51:20	3/23/2010	43	93	0	0	0	0	0	1.1
24	23	8.9	I-95 NORT	16:05:53	3/24/2010	56	85	4	0	0	0	0	1.1
25	24	8.9	I-95 NORT	17:48:22	3/24/2010	57	91	4	0	0	0	0	1.1

Figure 4.4 Final data

4.3 Data description

The dependent variable is accident cost, which represents the accident severity level. The dataset after filtering contained 56.11% accidents leading to property damage, 41.18% accidents leading to personal injury, and 2.71% accidents leading to fatality, as shown below in Figure 4.5.

cost

cost	Frequency	Percent	Cumulative Frequency	Cumulative Percent
8.9	11485	56.11	11485	56.11
70.2	8429	41.18	19914	97.29
1410	555	2.71	20469	100.00

Figure 4.5 Analysis for dependent variable

Using mean and standard deviation analysis with SAS, the following output shown in Figure 4.6 can be obtained. From the figure, we can see that the standard deviation is especially large for the first two variables, air temperature and relative humidity, because these variables have significant seasonal variations in Maryland. From the means of each variable under three levels of accident severity, we can see that when severe accidents happen, the average values of air temperature, relative humidity, average wind speed, and visibility are higher than when accidents which lead to only property damage happen. Meanwhile, the means of precipitation total and intensity when severe accidents happen are lower compared to when accidents which lead to only property damage happen.

The MEANS Procedure

cost	N Obs	Variable	Label	N	Mean	Std Dev
8.9	11485	AirTemperature	AirTemperature	11485	55.7850239	19.3720521
		RelativeHumidity	RelativeHumidity	11485	68.6127993	25.3902202
		AverageWindSpeed	AverageWindSpeed	11485	3.6721811	3.2829097
		Visibility	Visibility	11485	1.3282978	2.5381814
		PrecipitationTotal	PrecipitationTotal	11485	0.2648820	1.3801226
		slight	slight	11485	0.0700044	0.2551655
		moderate	moderate	11485	0.0177623	0.1320921
		heavy	heavy	11485	0.0063561	0.0794749
70.2	8429	AirTemperature	AirTemperature	8429	58.0692846	19.1430045
		RelativeHumidity	RelativeHumidity	8429	67.4005220	25.4398652
		AverageWindSpeed	AverageWindSpeed	8429	3.5564124	3.2877436
		Visibility	Visibility	8429	1.4025982	2.8244922
		PrecipitationTotal	PrecipitationTotal	8429	0.1888884	1.1247094
		slight	slight	8429	0.0629968	0.2429716
		moderate	moderate	8429	0.0122197	0.1098719
		heavy	heavy	8429	0.0043896	0.0661125
1410	555	AirTemperature	AirTemperature	555	57.9171171	18.1571652
		RelativeHumidity	RelativeHumidity	555	76.3099099	24.2697199
		AverageWindSpeed	AverageWindSpeed	555	3.9639640	3.7511200
		Visibility	Visibility	555	1.8019820	3.9728972
		PrecipitationTotal	PrecipitationTotal	555	0.1427928	1.0186437
		slight	slight	555	0.0432432	0.2035877
		moderate	moderate	555	0.0072072	0.0846651
		heavy	heavy	555	0.0018018	0.0424476

Figure 4.6 Mean and standard deviation analysis with SAS

In order to provide a more reliable regression result, we should check first if multicollinearity exists among the variables. Using analysis of variance (ANOVA), we are able to test whether the means of variables are equal. This is accomplished by partitioning the total variance into the component that is due to true random error and the components that are due to differences between means. 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. The corresponding test shows whether it is significant enough to reject the null hypothesis that two variables have linear correlation. The output in Figure 4.7 shows that all the Pearson Correlation values are significant enough to reject the null hypothesis. For example, in terms of total precipitation and relative humidity, the Pearson Correlation value is 0.13526, close to 0, and the probability of having a larger absolute value than 0.13526 is less than 0.0001, so that there is no significant linear correlation between total precipitation and relative humidity. The same conclusion can be obtained from each pair of the eight variables.

The VIF quantifies the severity of multicollinearity. It measures how much the variance of an estimated regression coefficient is increased because of collinearity. The square root of the VIF tells how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model. In terms of our example, as shown in Figure 4.8, the VIF of air temperature is 1.08917 (the square root is 1.04363). This means that the standard error for the coefficient of that predictor variable is 1.04363 times as large as it would be if air temperature were uncorrelated with the other predictor variables. So this multiplier is small enough to show that the variance of air temperature won’t increase much because of collinearity. The same conclusion can be obtained from the VIF values of other variables.

The condition indices are the square roots of the ratio of the largest eigenvalue to each individual eigenvalue. The condition number indicates the potential sensitivity of the computed inverse to small changes in the original matrix. If the Condition Number is above 30, the regression is said to have significant multicollinearity. For example, in our case, the condition number for heavy precipitation is 11.578, which shows dependence might be starting to affect the regression estimate, but the effect is too weak and won't cause significant multicollinearity. Considering the condition indices of other variables, the multicollinearity can be ignored when we analyze the data.

The CORR Procedure

Pearson Correlation Coefficients, N = 20469
Prob > |r| under H0: Rho=0

	Air Temperature	Relative Humidity	Average Wind Speed	Visibility
PrecipitationTotal PrecipitationTotal	-0.16141 <.0001	0.13526 <.0001	0.15657 <.0001	-0.02691 0.0001
slight slight	-0.15533 <.0001	0.21910 <.0001	0.07492 <.0001	-0.02560 0.0002
moderate moderate	-0.07905 <.0001	0.12292 <.0001	0.04398 <.0001	-0.01497 0.0322
heavy heavy	-0.01856 0.0079	0.07228 <.0001	0.03444 <.0001	-0.01245 0.0750

Pearson Correlation Coefficients, N = 20469
Prob > |r| under H0: Rho=0

	Precipitation Total	slight	moderate	heavy
AirTemperature AirTemperature	-0.16141 <.0001	-0.15533 <.0001	-0.07905 <.0001	-0.01856 0.0079
RelativeHumidity RelativeHumidity	0.13526 <.0001	0.21910 <.0001	0.12292 <.0001	0.07228 <.0001
AverageWindSpeed AverageWindSpeed	0.15657 <.0001	0.07492 <.0001	0.04398 <.0001	0.03444 <.0001
Visibility Visibility	-0.02691 0.0001	-0.02560 0.0002	-0.01497 0.0322	-0.01245 0.0750
PrecipitationTotal PrecipitationTotal	1.00000	0.21108 <.0001	0.16674 <.0001	0.08411 <.0001
slight slight	0.21108 <.0001	1.00000	-0.03312 <.0001	-0.01969 0.0048
moderate moderate	0.16674 <.0001	-0.03312 <.0001	1.00000	-0.00917 0.1895
heavy heavy	0.08411 <.0001	-0.01969 0.0048	-0.00917 0.1895	1.00000

	Air Temperature	Relative Humidity	Average Wind Speed	Visibility
AirTemperature AirTemperature	1.00000	-0.13029 <.0001	-0.15361 <.0001	0.10805 <.0001
RelativeHumidity RelativeHumidity	-0.13029 <.0001	1.00000	-0.15192 <.0001	0.01573 0.0244
AverageWindSpeed AverageWindSpeed	-0.15361 <.0001	-0.15192 <.0001	1.00000	-0.02152 0.0021
Visibility Visibility	0.10805 <.0001	0.01573 0.0244	-0.02152 0.0021	1.00000

Figure 4.7 Pearson Correlation and hypothesis test

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	6241280	780160	15.45	<.0001
Error	20460	1033099880	50494		
Corrected Total	20468	1039341161			

Root MSE	224.70790	R-Square	0.0060
Dependent Mean	72.13261	Adj R-Sq	0.0056
Coeff Var	311.52057		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.31482	7.68791	1.21	0.2257
AirTemperature	AirTemperature	1	0.23574	0.08503	2.77	0.0056
RelativeHumidity	RelativeHumidity	1	0.60531	0.06593	9.18	<.0001
AverageWindSpeed	AverageWindSpeed	1	2.34200	0.49751	4.71	<.0001
Visibility	Visibility	1	1.99001	0.58387	3.41	0.0007
PrecipitationTotal	PrecipitationTotal	1	-2.70238	1.31204	-2.06	0.0394
slight	slight	1	-26.00311	6.68055	-3.89	<.0001
moderate	moderate	1	-37.25447	13.18672	-2.83	0.0047
heavy	heavy	1	-45.41450	21.56385	-2.11	0.0352

Parameter Estimates

Variable	Label	DF	Tolerance	Variance Inflation
Intercept	Intercept	1	.	0
AirTemperature	AirTemperature	1	0.91813	1.08917
RelativeHumidity	RelativeHumidity	1	0.87807	1.13886
AverageWindSpeed	AverageWindSpeed	1	0.91567	1.09209
Visibility	Visibility	1	0.98680	1.01337
PrecipitationTotal	PrecipitationTotal	1	0.88503	1.12991
slight	slight	1	0.89172	1.12143
moderate	moderate	1	0.94809	1.05475
heavy	heavy	1	0.98361	1.01666

Collinearity Diagnostics

			Proportion of Variation			
Number	Eigenvalue	Condition Index	Intercept	Air Temperature	Relative Humidity	Average WindSpeed
1	3.87209	1.00000	0.00259	0.00539	0.00625	0.01786
2	1.19225	1.80215	0.00047447	0.00287	0.00015274	0.00020277
3	1.00018	1.96759	1.379095E-7	0.00000404	2.997348E-7	0.00003080
4	0.99448	1.97321	0.00002182	0.00020344	0.00000160	0.00000267
5	0.72881	2.30497	0.00139	0.00240	0.00266	0.02387
6	0.67169	2.40098	0.00017791	0.00044924	0.00010518	0.02436
7	0.39965	3.11267	0.00323	0.01786	0.03056	0.76318
8	0.11197	5.88065	0.00031972	0.39911	0.48662	0.00953
9	0.02889	11.57771	0.99179	0.57171	0.47365	0.16096

Collinearity Diagnostics

	Proportion of Variation				
Number	Visibility	Precipitation Total	slight	moderate	heavy
1	0.01586	0.00483	0.00729	0.00208	0.00081286
2	0.04396	0.35724	0.10645	0.14939	0.03378
3	0.00031184	0.00000330	0.00987	0.10974	0.85101
4	0.00562	0.00557	0.36874	0.47756	0.03405
5	0.86042	0.11111	0.00004905	0.00507	0.00000456
6	0.03735	0.45239	0.43403	0.23195	0.07307
7	0.03242	0.04957	0.00375	0.00243	0.00036747
8	0.00369	0.01928	0.06861	0.02078	0.00488
9	0.00035698	0.00001774	0.00120	0.00099931	0.00203

Figure 4.8 ANOVA analysis with SAS

Moreover, we can use One-way Mutivariate Analysis of Variance (MANOVA) to test each variable's significance in a further step. SAS provides the p-value associated with the F statistic of a given independent variable. If the null hypothesis represents that the variable has no effect on the outcome, then the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. The smaller the p-value, the more strongly the test rejects the null hypothesis. For a given alpha level, say 0.05, if the p-value is less than 0.05, the null hypothesis is rejected. If not, then we fail to reject the null hypothesis. The multivariate tests of all eight variables show p-values less than 0.05 except "heavy", as in Figure 4.9. Here we set "heavy" as the source of the variability in the specified dependent variable "cost", Univariate output within MANOVA provides the p-value of 0.1572, which is greater than 0.05. Thus, we cannot reject the null hypothesis at the significance level of 5%, that is to say, "heavy" is not a strong predictor at 95% confidence level.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	101632	101632	2.00	0.1572
Error	20467	1039239529	50776		
Corrected Total	20468	1039341161			

R-Square	Coeff Var	Root MSE	cost Mean
0.000098	312.3914	225.3361	72.13261

Source	DF	Anova SS	Mean Square	F Value	Pr > F
heavy	1	101631.7518	101631.7518	2.00	0.1572

Figure 4.9 MANOVA on "heavy"

In multivariate output, MANOVA calculates four multivariate test statistics. The null hypothesis for each of these tests is the same: any independent variable has no effect on the dependent variable. As shown in Figure 4.10, Wilks' Lambda can be interpreted as the proportion of the variance in the outcomes that is not explained by an effect, so in this case, only 1-0.9897=0.0103 of the variance in the “cost” can be explained by the effect of weather factors. Pillai’s trace shows how much the effect from given variables contributes to the model, and Hotelling-Lawley’s trace is calculated to test the significance on the difference of the mean of two or more variables between the groups. Each of these two tests results in a small value (0.01029, and 0.01034) but less than 0.05 p-value, which means the weather impact on accident cost is small but still significant at 95% confidence level, and differences between the levels of the variables exist. The last line, Roy’s greatest root, should be ignored here because it only considers the first discriminant function while the independent variables have more than one dimension.

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall cost Effect
H = Type III SSCP Matrix for cost
E = Error SSCP Matrix

S=2 M=2.5 N=10228.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.98973846	13.22	16	40918	<.0001
Pillai's Trace	0.01028725	13.22	16	40920	<.0001
Hotelling-Lawley Trace	0.01034195	13.22	16	33475	<.0001
Roy's Greatest Root	0.00604134	15.45	8	20460	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

Figure 4.10 Overall MANOVA

To summarize this chapter, after merging and filtering, the final data include 20,469 observations in total. Each observation contains one accident and its corresponding severity level, time, location, and weather condition at the time of accident. Accident cost is selected to represent severity and 8 weather-related variables (air temperature, average wind speed, precipitation total, intensity dummy (slight, moderate, heavy), visibility, and relative humidity) are used in the next steps to examine the impact. The correlation test and analysis of variance show no correlation between each pair of independent variables. Based on the output of ANOVA, air temperature, average wind speed, visibility, relative humidity, precipitation accumulation and slight or moderate precipitation appear to have significant impact on accident severity; while heavy precipitation has less impact. Moreover, the overall multivariate test shows that differences exist between levels of the accident costs.

Chapter 5: Model Estimation and Performance Analysis

5.1 Multinomial Logit Model Estimation Result

Below we use “proc logistic” in SAS 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 cost=1410, i.e. severity level equals to fatality, is set to reference category. The output is shown in Figure 5.1.

```
                         The LOGISTIC Procedure

                           Model Information

Data Set                      WORK.COST
Response Variable             cost                      cost
Number of Response Levels     3
Model                         generalized logit
Optimization Technique        Newton-Raphson

        Number of Observations Read       20469
        Number of Observations Used       20469

                            Response Profile

                 Ordered                      Total
                   Value         cost     Frequency

                       1          8.9         11485
                       2         70.2          8429
                       3         1410           555

Logits modeled use cost=1410 as the reference category.

                       Model Convergence Status

     Convergence criterion (GCONV=1E-8) satisfied.

                         Model Fit Statistics

                                               Intercept
                               Intercept             and
                Criterion           Only      Covariates

                AIC            32239.194       32063.108
                SC             32255.047       32205.788
                -2 Log L       32235.194       32027.108

              Testing Global Null Hypothesis: BETA=0

Test                     Chi-Square       DF     Pr > ChiSq

Likelihood Ratio           208.0854       16         <.0001
Score                      210.5697       16         <.0001
Wald                       205.8602       16         <.0001
```

Figure 5.1 SAS output for MNL

The LOGISTIC Procedure

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AirTemperature	2	49.9295	<.0001
RelativeHumidity	2	86.6273	<.0001
AverageWindSpeed	2	23.7597	<.0001
Visibility	2	9.7479	0.0076
PrecipitationTotal	2	7.2996	0.0260
slight	2	10.4431	0.0054
moderate	2	7.6637	0.0217
heavy	2	3.9755	0.1370

Analysis of Maximum Likelihood Estimates

Parameter	cost	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	8.9	1	4.7678	0.2243	451.9130	<.0001
Intercept	70.2	1	4.2489	0.2261	353.2094	<.0001
AirTemperature	8.9	1	-0.00615	0.00244	6.3795	0.0115
AirTemperature	70.2	1	-0.00080	0.00246	0.1057	0.7451
RelativeHumidity	8.9	1	-0.0167	0.00190	77.4660	<.0001
RelativeHumidity	70.2	1	-0.0178	0.00191	86.6123	<.0001
AverageWindSpeed	8.9	1	-0.0546	0.0121	20.3992	<.0001
AverageWindSpeed	70.2	1	-0.0597	0.0123	23.7589	<.0001
Visibility	8.9	1	-0.0362	0.0117	9.5783	0.0020
Visibility	70.2	1	-0.0305	0.0118	6.7038	0.0096
PrecipitationTotal	8.9	1	0.1462	0.0758	3.7190	0.0538
PrecipitationTotal	70.2	1	0.1200	0.0762	2.4801	0.1153
slight	8.9	1	0.7026	0.2194	10.2520	0.0014
slight	70.2	1	0.7070	0.2211	10.2284	0.0014
moderate	8.9	1	1.1178	0.5132	4.7437	0.0294
moderate	70.2	1	0.8819	0.5179	2.8999	0.0886
heavy	8.9	1	1.5722	1.0106	2.4201	0.1198
heavy	70.2	1	1.2959	1.0171	1.6233	0.2026

Odds Ratio Estimates

Effect	cost	Point Estimate	95% Wald Confidence Limits	
AirTemperature	8.9	0.994	0.989	0.999
AirTemperature	70.2	0.999	0.994	1.004
RelativeHumidity	8.9	0.983	0.980	0.987
RelativeHumidity	70.2	0.982	0.979	0.986
AverageWindSpeed	8.9	0.947	0.925	0.970
AverageWindSpeed	70.2	0.942	0.920	0.965

Effect	cost	Point Estimate	95% Wald Confidence Limits	
Visibility	8.9	0.964	0.943	0.987
Visibility	70.2	0.970	0.948	0.993
PrecipitationTotal	8.9	1.157	0.998	1.343
PrecipitationTotal	70.2	1.127	0.971	1.309
slight	8.9	2.019	1.313	3.104
slight	70.2	2.028	1.315	3.127
moderate	8.9	3.058	1.118	8.362
moderate	70.2	2.416	0.875	6.666
heavy	8.9	4.817	0.665	34.919
heavy	70.2	3.654	0.498	26.825

Figure5.1 SAS output for MNL (Cont.)

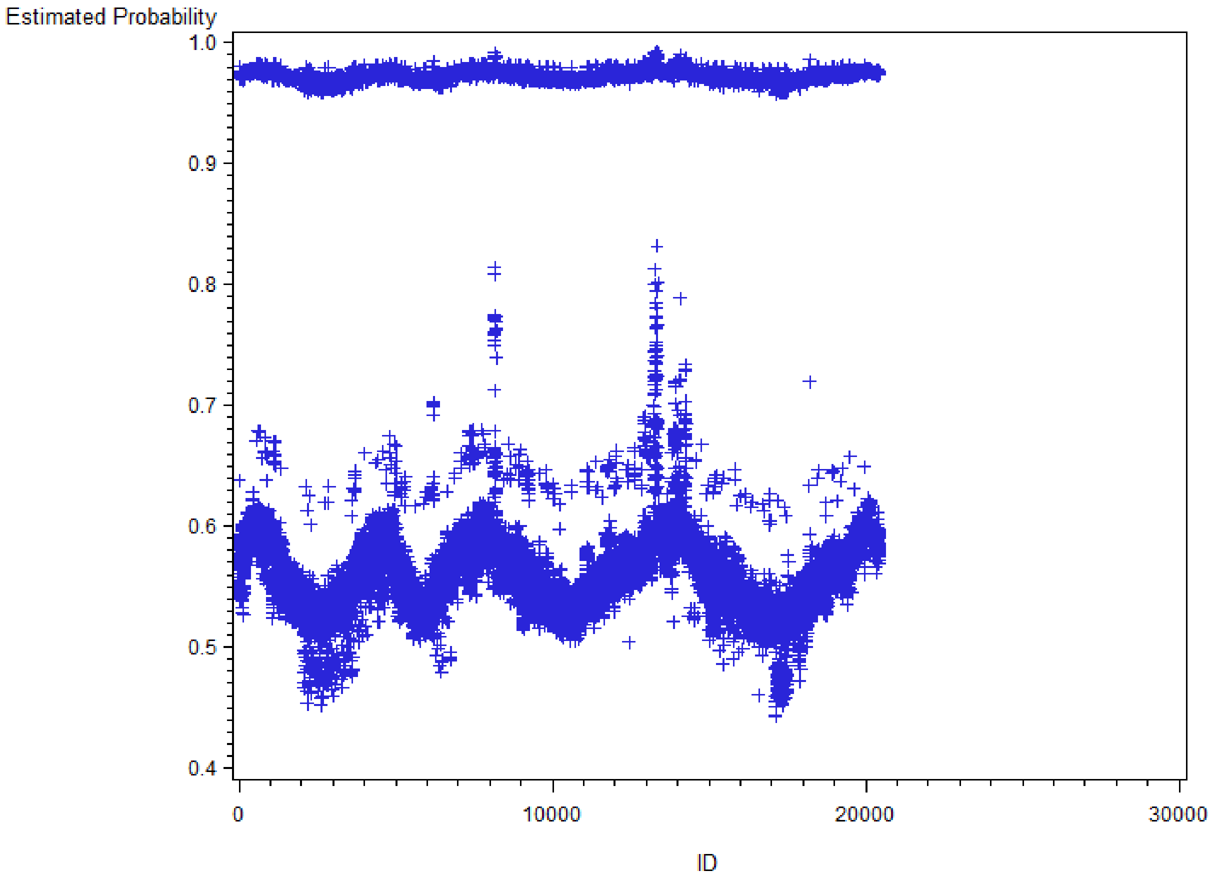

Figure5.1 SAS output for MNL (Cont.)

From the output we can see that the model has converged. Setting the reference category as observations with cost = 1410, the variables have different estimated coefficients when computing observations with cost=8.9 and cost= 70.2. In the test of global null hypothesis, SAS uses three tests: likelihood ratio, Score and Wald, to test the hypothesis that at least one factor has a significant impact on accident cost against the global null hypothesis that none of the factors has a significant impact. All p-values are less than 0.0001, so we can reject the null hypothesis at 99% confidence level, which tells us that our model as a whole fits significantly better than an empty model (i.e., a model with lack of regressors). Several model fit measures such as the Akaike Information Criterion (AIC) and the Schwarz Criterion (SC), are listed under Model Fit Statistics. Though criteria AIC and SC do not provide a test of a model in the sense of testing a null hypothesis, their values provide a means for model selection. The magnitude of the differences between the values of “intercept only” and “intercept and covariates” indicate how much better the model fits with covariates. Promising models give small values for these criteria. The chi-square test of each parameter shows significance at 99% confidence level.

The information we can get from coefficient signs is almost opposite with the output obtained by ANOVA, because when running regression with ANOVA, we should assume that all variables are categorical, but in this case variables such as air temperature and precipitation total are not categorical, so the signs in ANOVA output do not have much meaning and should be ignored. In the multinomial logit model, air temperature, average wind speed, relative humidity and visibility appear to have significant negative impact on accident cost, except the impact of air temperature to the accident with cost=70.2. All the precipitation related variables appear to have negative but less significant impact, except the slight precipitation.

Here we take some examples to describe the statistics “Odds Ratio Estimates” and results achieved after simple calculation.

- A one-unit increase in temperature is associated with a 1-0.994=0.006 unit decrease in the relative log odds of having an accident with cost of 8.9 vs. accident with cost of 1410; or a 1-0.999=0.001 unit decrease in the relative log odds of having an accident with cost of 70.2 vs. accident with cost of 1410.
- A one-unit increase in visibility value is associated with a 1-0.964=0.036 unit decrease in the relative log odds of having an accident with cost of 8.9 vs. accident with cost of 1410; or

a 1-0.970=0.030 unit decrease in the relative log odds of having an accident with cost of 70.2 vs. accident with cost of 1410.

- A one-unit increase in the precipitation total is associated with a 1.157-1=0.157 unit increase in the relative log odds of having an accident with cost of 8.9 vs. accident with cost of 1410; or a 1.127-1=0.127 unit increase in the relative log odds of having an accident with cost of 70.2 vs. accident with cost of 1410.

The analysis of odds ratio provides an easy way to interpret the different outcomes that different groups have on a particular scenario. For example, the impact on the probability of accident with cost of 8.9 caused by increase in precipitation total is 0.157/0.127-1=23.6% greater than the impact on the probability of accident with cost of 70.2. The odds ratio is a versatile and robust statistic, and similarly to the Pearson correlation coefficient, it can measure effect size and therefore provides information on the strength of relationship between two levels of a variable. Therefore, the analysis of odds ratio gives us a visual understanding of how weather factors have different impact in different accident categories.

The overall effects of estimators on accident cost are listed under "Type 3 Analysis of Effects". The Wald test shows that only heavy precipitation is not significant at the 95% confidence level. The log likelihood is -16013.554. The estimated probability values shown in the plot provide a visualized figure of how well the MNL model works.

To calculate the mean squared errors (MSE), let P_i be the true value of the probability of having an accident in observation i, and $\hat{P}_i$ be the predict value of P_i. The predicted probability of having all three categories of accidents for each observation can be obtained from SAS. Here we take the average of all observations under each category, then the MSE can be calculated using equation 5.1. The result is shown in Table 5.1.

$$MSE = \frac{1}{3}\sum_{i=1}^{3}(\hat{P}_i - P_i)^2 \tag{5.1}$$

Table 5.1 MSE calculation for MNL

Severity Category	**Actual prob.** P_i	**Average predicted prob.** $\hat{P}_i$
i=1 Property damage	0.5611	0.52737
i=2 Person injury	0.4118	0.47263
i=3 Fatality	0.0271	0
MSE_{MNL}	0.0055717	

5.2 Ordered Probit Model Estimation Result

This part shows an ordered probit regression analysis. In the following application, accident severity is the ordered dependent variable. An assumption has been made that the indexing in the model is a latent but continuous estimator and the related error is random and follows a normal distribution. The observed and coded discrete dependent variable, severity level (cost), is set up as the same with ordered value.

The result obtained by SAS is shown in Figure 5.2. The signs of coefficients of estimators are the same with MNL model, that is, the higher value of air temperature, average wind speed, relative humidity or visibility will lower the probability of severe accident occurring; the higher value of the precipitation total or intensity will increase the probability of severe accident occurring. However, not all factors show significant impact under the null hypothesis test, which must reconsider the meaning of their signs. Even though the signs of estimated parameters show that average wind speed has negative effect and slight precipitation has positive effect on accident cost, we cannot reject the null hypothesis that there is no significant impact at 95% confidence level. Therefore, we should not include these two variables in the best set.

The value of log likelihood of OP model is -16066.942, smaller than MNL model. Also, the mean square errors can be calculated in a same way: MSE_{op}=0.014682. Under these two criteria, we prefer using the multinomial logit model to the order probit model.

The Probit Procedure

Model Information

Data Set	WORK.COST	
Dependent Variable	cost	cost
Number of Observations	20469	
Name of Distribution	Normal	
Log Likelihood	-16066.94208	

Number of Observations Read	20469
Number of Observations Used	20469

Class Level Information

Name	Levels	Values
cost	3	8.9 70.2 1410

Response Profile

Ordered Value	cost	Total Frequency
1	8.9	11485
2	70.2	8429
3	1410	555

PROC PROBIT is modeling the probabilities of levels of cost having LOWER Ordered Values in the response profile table.

Algorithm converged.

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AirTemperature	1	48.9824	<.0001
RelativeHumidity	1	7.7616	0.0053
AverageWindSpeed	1	1.6807	0.1948
Visibility	1	5.7578	0.0164
PrecipitationTotal	1	7.4614	0.0063
slight	1	3.1630	0.0753
moderate	1	8.6559	0.0033
heavy	1	4.6017	0.0319

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.4122	0.0413	0.3311	0.4932	99.41	<.0001
AirTemperature	1	-0.0032	0.0005	-0.0041	-0.0023	48.98	<.0001
RelativeHumidity	1	-0.0010	0.0004	-0.0017	-0.0003	7.76	0.0053
AverageWindSpeed	1	-0.0035	0.0027	-0.0087	0.0018	1.68	0.1948
Visibility	1	-0.0073	0.0031	-0.0133	-0.0013	5.76	0.0164
PrecipitationTotal	1	0.0206	0.0075	0.0058	0.0353	7.46	0.0063
slight	1	0.0644	0.0362	-0.0066	0.1354	3.16	0.0753
moderate	1	0.2169	0.0737	0.0724	0.3613	8.66	0.0033
heavy	1	0.2599	0.1211	0.0224	0.4973	4.60	0.0319

Figure 5.2 SAS output with Ordered Probit Model

5.3 Neural Network Estimation Result

5.3.1 Model selection using Neuro Solution

First, we use the Neuro Solution software to check which model was the most suitable for regression. Three kinds of regression methods were built to compute the performance: linear regression, probabilistic neural network, and multilayer perceptron.

After running the learning and training process, the software compared the three networks in terms of mean square errors, mean absolute errors and residuals, and produced the performance metrics. Table 5.2 shows the results along with the best performing network.

MLP is the best performance model because of its low validation mean squared residual (error) compared with other regression models in neural networks.

Table 5.2 Summary of all three networks

	Training	Validation	Testing
Model Name	MSE	MSE	MSE
MLP-1-O-M (Multilayer Perceptron)	0.215153	0.18455	0.19083
LR-0-B-M (Linear Regression)	0.210452	0.19270	0.19744
PNN-0-N-N (Probabilistic Neural Network)	0.211488	0.19002	0.19367

5.3.2 Training of Accident Severity Data with Matlab

Multilayer perceptron is a class of artificial neural networks in which the layers are usually interconnected in a feed-forward way, that is to say, each neuron in one layer has a one-way direction to transport information to the neurons of the subsequent layer. Back-propagation is the most popular learning technique.

To set up the data, we use 70% of the total 20469 samples in training, 15% in testing and 15% in validation. Using 8 weather related variables as input and cost as target data, the networks contain two layers, sigmoid hidden neurons and linear output neurons. The structure is shown in Figure 5.3.

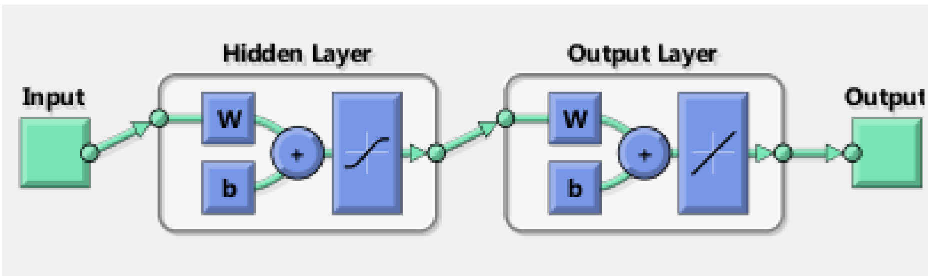

Figure5.3 Two layers Neural Networks structure

In the learning process, the weights of each connection are compared with the true value to minimize the error, and then this process is repeated until convergence is reached. To adjust weights properly, nonlinear optimization is applied that is called gradient descent. For this, the derivative of the error function with respect to the network weights is calculated, and the weights are then changed such that the error decreases (Warren S. Sarle, 1994). The procedure of MPL neural networks can be trained and tested with Matlab based on the following formulas:

$$g_j = a_j + \sum_{i=1}^{n_x} b_{ij} x_i \tag{5.2}$$

$$h_j = \frac{1}{(1+\exp(-g_j))} \tag{5.3}$$

$$q_k = c_k + \sum_{i=1}^{n_k} d_{jk} h_k \tag{5.4}$$

$$p_k = \frac{1}{(1+\exp(-q_k))} \tag{5.5}$$

$$r_k = y_k - p_k \tag{5.6}$$

where n_x = number of independent variables (inputs)

n_h = number of hidden neurons

x_i = independent variable

a_j = bias for hidden layer

b_{ij} = weight from input to hidden layer

g_j = net input to hidden layer

h_j = hidden layer values

c_k = bias for output (intercept)

d_{jk} = weight from hidden layer to output

q_k = net input to output layer

p_k = predicted value (output values)

y_k = depend variable (training values)

r_k = error.

During iterative training of a neural network, an epoch is a step through the entire training process, followed by testing of the verification set. It contains several iterations. Figure 5.4 shows that the mean square error meets convergence at epoch 119 with the minimal validation MSE of 0.171753.

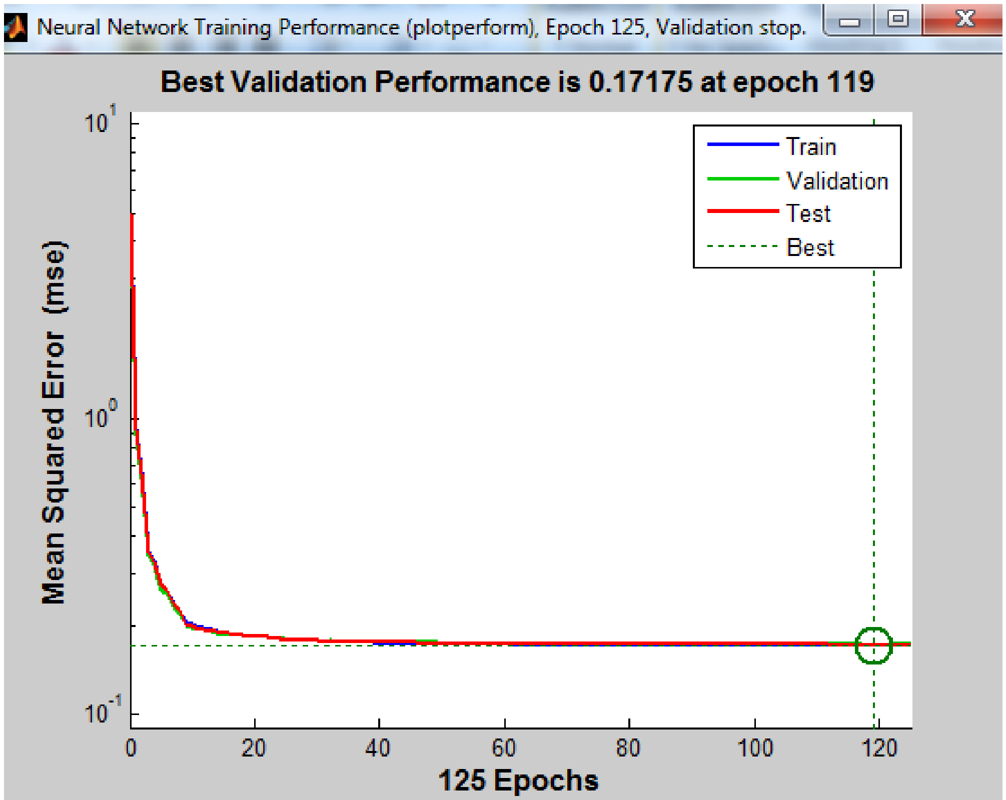

Figure 5.4 Trends of MSE

Table 5.3 lists the mean square of errors in each process. This result provides minimal MSE with various numbers of hidden layers, but comparing with traditional regression methods, the values are still large. No matter how we adjust the number of hidden layers or change the fitting problem to classification problem, the MSE values are always around 0.17.

Table 5.3 MSE of Neural Networks

Data set	Sample	MSE
Training	14329	0.170285
Validation	3070	0.171753
Testing	3070	0.171695

In order to find the degree of significance of each parameter, we repeat the same training and testing process 8 times. Each time, we delete one weather-related variable and calculate the MSE in validation set, as shown in Table 5.4.

Table5.4 Variable selection with neural networks

Subset of variables	Validation MSE	% improved
Initial set	0.171753	/
Without air temperature	0.170351	0.816
Without relative humidity	0.169099	0. 735
Without average wind speed	0.171829	-1.6144
Without visibility	0.173752	- 1.1191
Without precipitation total	0.170839	1.677
Without slight precipitation	0.167783	1.789
Without moderate precipitation	0.171533	-2.235
Without heavy precipitation	0.170726	0.47

Deleting air temperature, relative humidity, precipitation total, slight precipitation and heavy precipitation in the model decrease the validation MSE and increase the model accuracy, so these variables are less significant. On the other hand, deleting average wind speed, visibility and moderate precipitation in the model increase the validation MSE, which means these variables are important to be considered when generating the networks.

The Receiver Operating Characteristic (ROC) analysis shown in Figure 5.5 again proves the bad performance of neural networks. The diagonal divides the ROC space into two parts: Points above the diagonal represent good classification results (better than random), and points below the line poor results represent bad classification results (worse than random). All four curves are above but very close to the diagonal line, which means after training and testing process, the result doesn't have much improvement compared with randomly guessing performance.

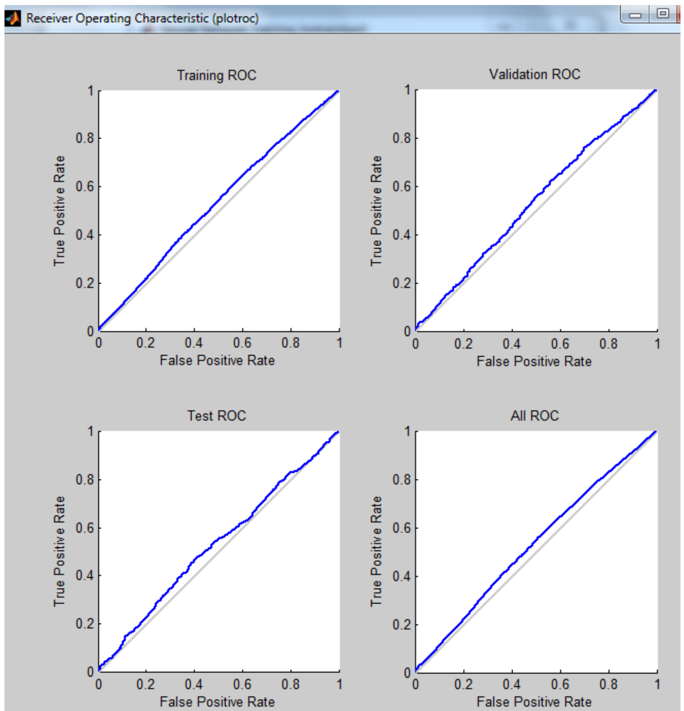

Figure 5.5 ROC analysis

The results demonstrate that Neural Network is not an effective tool to classify severity levels in crashes if appropriate input data is available. Comparing MSE of other methodologies mentioned before, the MLP in neural network has a lower accuracy.

5.4 Model Comparison and Selection

The comparison of all three methods is summarized by the values of log likelihood at convergence and MSE in Table 5.5.

Table 5.5 Model Comparison

	Log likelihood at convergence	MSE
Multinomial logit	-16013.554	0.0055717
Ordered Probit	-16066.942	0.014682
Neural Network (MLP)	N/A	0.171753

Based on the comparison of the criteria above, we found that the multinomial logit regression is more interpretive than the other two methods, since the former has higher value in log likelihood at convergence and lower value in MSE.

Moreover, though taking the ordinal information of dependent variable into account, the Order Probit model still does not have as good performance as expected compared with MNL model. This can be explained by the restriction of using an identical coefficient for an estimator across different accident severities. The OP model only allows one weather-related factor either to increase the probability of fatality and decrease the probability of property damage, or to decrease the probability of fatality and increase the probability of property damage, but it can't explain when both severity levels increase or decrease in probability. This may not be suitable when it comes to reality. Also, the OP model failed in explaining clearly what effect a factor has on estimating the probability of middle severity level, people injury.

The signs of coefficients in different models are summarized in Table 5.6 along with their significance.

Table5.6 Summary of coefficients in different models

Model / Variable	MNL		OP		NN
	Sign	Significance	Sign	Significance	Significance
Air temperature	-	N	-	Y	N
Relative humidity	-	Y	-	Y	N
Average wind speed	-	Y	-	N	Y
Visibility	-	Y	-	Y	Y
Precipitation total	+	N	+	Y	N
Slight	+	Y	+	N	N
Moderate	+	Y	+	Y	Y
Heavy	+	N	+	Y	N

Chapter 6: Case Study

In the previous chapters, the impact of weather-related variables was analyzed based on the accident record on the entire highway network in Maryland, and a relatively good result was generated with multinomial logit model. However, when we want to examine the effects of weather on accident severity, we must exclude the effect of other factors such as highway geometric characteristics and traffic elements. In the following analysis, a case study on US 50 is carried out to test the impact of weather condition on certain road sections, so that when we run the regression models, the costs of accidents are not affected by the characteristics of different roads, such as lane width, curve degree, pavement material, AADT and so on.

6.1 US Route 50

U.S. Route 50 is a major east–west route of the U.S. Highway system, stretching across Maryland. As shown in Figure 6.1, it passes the south end of the Baltimore-Washington Parkway and becomes the John Hanson Highway, a freeway to Annapolis. The freeway continues beyond Annapolis as the Blue Star Memorial Highway which crosses Chesapeake Bay on the Chesapeake Bay Bridge and continues to Queenstown where US 50 turns south, passing through Easton to Cambridge, and then east through Salisbury to Ocean City on the four-lane divided Ocean Gateway. US 50 ends near the Atlantic Ocean shore.

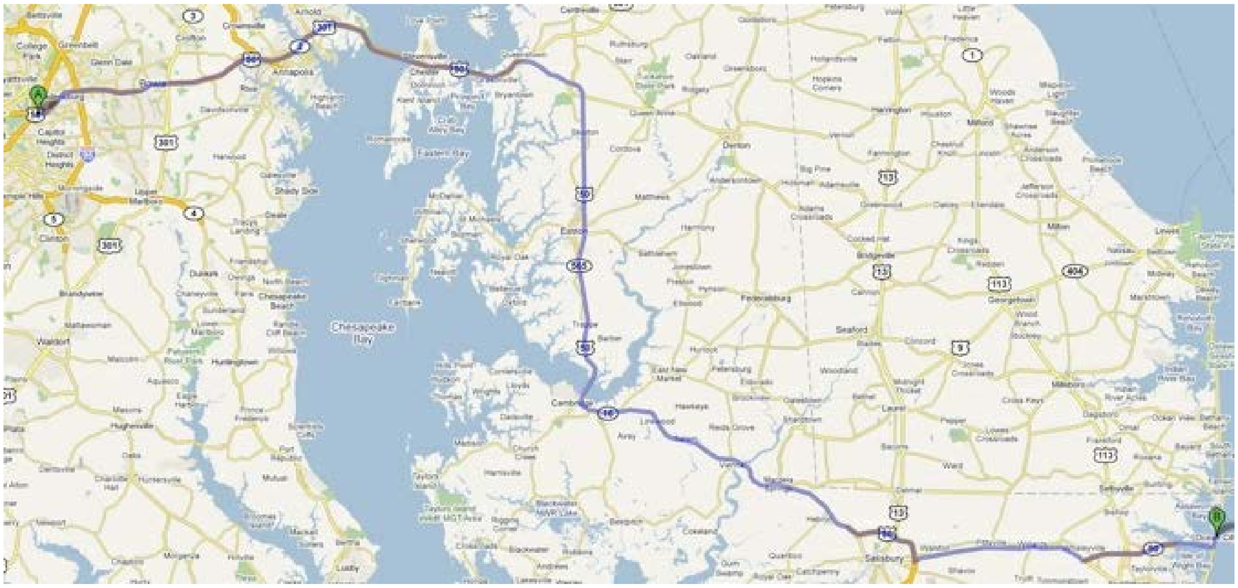

Figure 6.1 US route 50

In order to minimize the impacts caused by the wideness of weather change, we select a road section between latitude (38.9, 39.1) and longitude (-77, -76.2), and set the 970 accident records in total from 2007 to 2010 as targets. All accidents locations within this section are shown in Figure 6.2. The reason to choose this freeway is that compared with other highways, US 50 has more accident records in the four years period and a high fatality rate. Also, the accidents that happened in the part to the left of Chesapeake have more parallel weather condition and other traffic characteristics. In the following analysis, we take those 970 accident records to build a multinomial logit regression model, and compare the significance of each weather factor with the overall analysis. After eliminating the impact of different location, we have reasons to believe the result is more valuable and closer to reality.

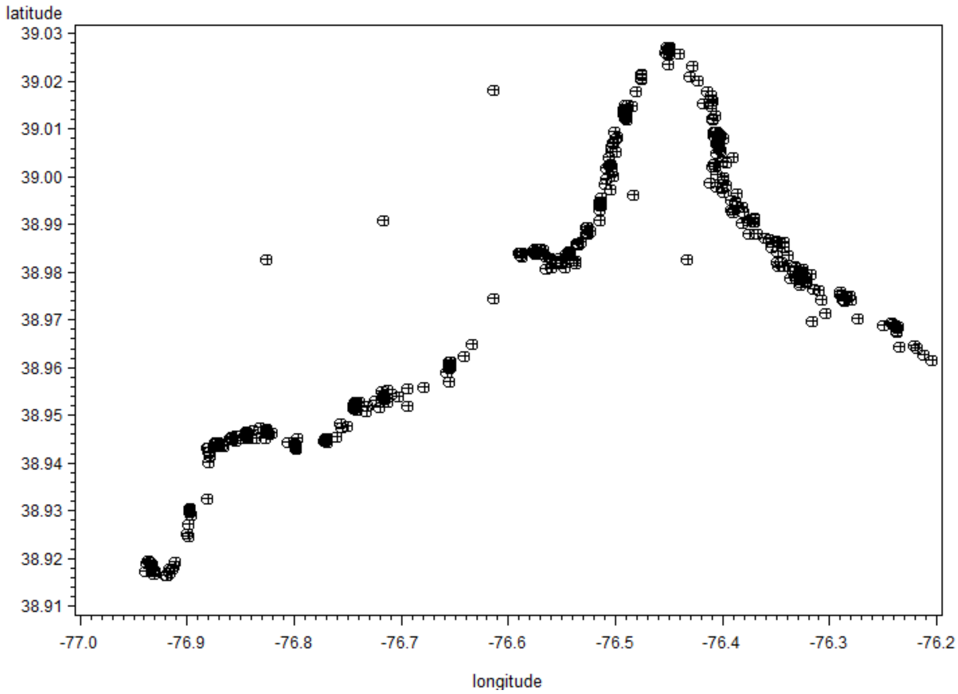

Figure 6.2 Accidents records on selected road section

6.2 Multinomial Logit Model Result and Analysis

In this part, we can also use SAS to estimate a multinomial logit regression model. Before that, the same procedure is performed to test the correlation. Using ANOVA, we are able to partition the total variance into the component that is due to true random error and the components that are due to differences between means. Again, three measures, Pearson Correlation, Variance Inflation Factor (VIF) and Condition index values, are used to test multicollinearity. The output in Figure 6.3 shows that all the Pearson Correlation values are significant enough to reject the hypothesis that two variables have collinearity. For example, in terms of total precipitation and relative humidity, the Pearson Correlation value is 0.11476, close to 0, and the probability of having a larger absolute value than 0.11476 is 0.0003, less than 0.05, so that there is no significant linear correlation between total precipitation and relative humidity at 95% confidence level. Similar conclusions can be obtained for each pair of the eight variables.

Pearson Correlation Coefficients, N = 970 Prob > |r| under H0: Rho=0								
	AirTemperature	RelativeHumidity	AverageWindSpeed	Visibility	PrecipitationTotal	slight	moderate	heavy
AirTemperature	1.00000	0.04643	-0.10582	0.12365	-0.17218	-0.14964	-0.03745	0.01184
		0.1484	0.0010	0.0001	<.0001	<.0001	0.0439	0.0126
RelativeHumidity	0.04643	1.00000	0.16727	0.02026	0.11476	0.12345	0.07239	0.03676
	0.1484		<.0001	0.5285	0.0003	0.0001	0.0242	0.0527
AverageWindSpeed	-0.10582	0.16727	1.00000	0.00826	0.18554	0.09282	0.06278	0.00029
	0.0010	<.0001		0.0972	<.0001	0.0038	0.0506	0.0927
Visibility	0.12365	0.02026	0.00826	1.00000	-0.04946	-0.04577	-0.02890	-0.01088
	0.0001	0.5285	0.0972		0.1237	0.1543	0.0686	0.0351
PrecipitationTotal	-0.17218	0.11476	0.18554	-0.04946	1.00000	0.26642	0.15250	0.05928
	<.0001	0.0003	<.0001	0.1237		<.0001	<.0001	0.0650
slight	-0.14964	0.12345	0.09282	-0.04577	0.26642	1.00000	-0.02823	-0.01696
	<.0001	0.0001	0.0038	0.1543	<.0001		0.3799	0.0978
moderate	-0.03745	0.07239	0.06278	-0.02890	0.15250	-0.02823	1.00000	-0.00689
	0.0439	0.0242	0.0506	0.0686	<.0001	0.3799		0.0303
heavy	0.01184	0.03676	0.00029	-0.01088	0.05928	-0.01696	-0.00689	1.00000
	0.0126	0.0527	0.0927	0.0351	0.0650	0.0978	0.0303	

Parameter Estimates									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation	Eigenvalue	Condition Index
Intercept	1	57.22689	24.47115	2.34	0.0196	.	0	3.90385	1.00000
AirTemperature	1	0.30369	0.26139	1.16	0.2456	0.93311	1.07168	1.22594	1.78448
RelativeHumidity	1	-0.21369	0.22937	-0.93	0.3517	0.94392	1.05941	1.00046	1.97536
AverageWindSpeed	1	-1.77920	1.25599	-1.42	0.1569	0.93479	1.06976	0.99837	1.97743
Visibility	1	-0.79076	1.26990	-0.62	0.5336	0.98169	1.01865	0.81213	2.19247
PrecipitationTotal	1	-0.36581	3.79245	-0.10	0.9232	0.85971	1.16319	0.64322	2.46358
slight	1	-8.91575	21.11405	-0.42	0.6729	0.89954	1.11168	0.31953	3.49536
moderate	1	-17.88337	47.43430	-0.38	0.7062	0.96542	1.03582	0.06930	7.50563
heavy	1	359.40706	77.27057	4.65	<.0001	0.99322	1.00683	0.02720	11.97987

Figure 6.3 Correlation analysis

The VIF measures how much the variance of an estimated regression coefficient is increased because of collinearity. For example, the VIF of air temperature is 1.07168 (the square root is 1.03522), this means that the standard error for the coefficient of that predictor variable is 1.03522 times as large as it would be if air temperature was not correlated with the other predictor variables. So this multiplier is small enough to show that the variance of air temperature would not increase much because of collinearity. The same conclusion can be obtained from the VIF values of other variables.

Then we can check the collinearity diagnostics. For example, the condition number for heavy precipitation is 11.97987, which shows dependence might be starting to affect the regression estimate, but the effect is way below 30, which is too weak and won't cause significant multicollinearity. Considering the condition indices of other variables, the multicollinearity can be ignored when we analyze the data.

The output is shown in Figure 6.4. From the global hypothesis test we can tell that at least one of the weather related factors has significant impact on accident cost. Regardless of the difference between each level of accident cost, from the last column of type 3 analysis of effect, we can see that air temperature, visibility, precipitation total, slight, moderate, and heavy precipitation all have p-values greater than 0.05, so they don't show any significant impact on accident cost at 95% confidence level. However, considering that weather factors can have different influence on different accident severity, we take a deeper look at analysis of maximum likelihood estimates, and find out that the set of variables with significant impact changes.

Here, cost=70.2, i.e. severity level equal to people injury, is set as the reference category. Two models are defined in this multinomial regression: one relating cost=8.9 to the reference category, cost=70.2 and another model relating cost=1410 to cost=70.2. The fourth column is the estimated multinomial logistic regression coefficient for the models. For example, 0.00860 is the multinomial logit estimate for a one unit increase in relative humidity for occurring an accident with cost=8.9 relative to occurring an accident with cost=70.2, given the other variables in the model are held constant. If relative humidity increases by one unit, the multinomial log-odds for preferring an accident with cost=8.9 to an accident with cost=70.2 would be expected to increase by 0.00860 unit while holding all other variables in the model constant. Moreover, under this assumption, the p-value is 0.0054, less than 0.05, so this impact is significant at 95% confidence level.

Further analysis on the comparison of different categories shows that while holding all other variables in the model constant, relative humidity and average wind speed have significant impact on the possibility of having an accident with property damage, and heavy precipitation has a significant impact on the possibility of having an accident with a fatality.

The LOGISTIC Procedure

Model Information

Data Set	WORK.US50
Response Variable	cost
Number of Response Levels	3
Model	cumulative logit
Optimization Technique	Fisher's scoring

Number of Observations Read	970
Number of Observations Used	970

Response Profile

Ordered Value	cost	Total Frequency
1	8.9	613
2	1410	12
3	70.2	345

Probabilities modeled are cumulated over the lower Ordered Values.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
143.2868	8	<.0001

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	44.2189	16	0.0002
Score	52.5156	16	<.0001
Wald	31.9354	16	0.0102

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AirTemperature	2	1.6364	0.4412
RelativeHumidity	2	7.8032	0.0202
AverageWindSpeed	2	13.6375	0.0011
Visibility	2	0.5502	0.7595
PrecipitationTotal	2	0.8663	0.6484
slight	2	0.3425	0.8426
moderate	2	0.5620	0.7550
heavy	2	4.3324	0.1146

Figure 6.4 MNL regression on US route 50 section

Analysis of Maximum Likelihood Estimates

Parameter	cost	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	8.9	1	-0.4005	0.3327	1.4485	0.2288
Intercept	1410	1	-4.3637	1.6347	7.1259	0.0076
AirTemperature	8.9	1	-0.00063	0.00362	0.0303	0.8618
AirTemperature	1410	1	0.0206	0.0166	1.5261	0.2167
RelativeHumidity	8.9	1	0.00860	0.00309	7.7241	0.0054
RelativeHumidity	1410	1	0.00114	0.0134	0.0072	0.9323
AverageWindSpeed	8.9	1	0.0672	0.0188	12.8226	0.0003
AverageWindSpeed	1410	1	-0.0487	0.1038	0.2198	0.6392
Visibility	8.9	1	-0.00847	0.0175	0.2327	0.6295
Visibility	1410	1	-0.1496	0.2565	0.3404	0.5596
PrecipitationTotal	8.9	1	0.00706	0.0603	0.0137	0.9069
PrecipitationTotal	1410	1	-0.8665	0.9403	0.8493	0.3568
slight	8.9	1	-0.1698	0.2908	0.3410	0.5593
slight	1410	1	-10.4046	264.1	0.0016	0.9686
moderate	8.9	1	0.5979	0.7976	0.5618	0.4535
moderate	1410	1	-9.9626	718.2	0.0002	0.9889
heavy	8.9	1	-14.5864	583.1	0.0006	0.9800
heavy	1410	1	3.4567	1.6609	4.3317	0.0374

Figure 6.4 MNL regression on US route 50 section (Cont.)

Comparing the result with what we obtained in the previous chapter, the regression on US 50 shows less significant weather related factors and different signs for these factors. One possible reason is that the analysis on overall Maryland highways overestimates the effect of weather factors because of the impacts of traffic and geometric elements. The other reason may be related to the error of small sample, especially when considering the signs of each parameter. However, regardless of the differences in these results, we can conclude that weather factors do have an impact on highway accident severity.

Chapter 7: Conclusion and Future Research

7.1 Identification of Factors Contributing to Accident Severity

The following statements are based on the MNL result using the accident data over the entire state of Maryland. In terms of the signs of each parameter, we cannot simply classify them as positive or negative when applying to other cases, because they may differ from region to region and be affected by other factors and it's hard to quantify their impact.

- Air temperature

  Air temperature has a significant effect on accident severity on highways. When air temperature increases during summer, strong sunlight can weaken drivers' sight, and the light will be refracted near the surface. High temperature can damage the engine and increase the probability of accident. Lack of proper vehicle maintenance can also have an impact on the effect of air temperature on engine performance.

- Average wind speed

  Average wind speed also has a significant effect on accident severity on highways. When wind speed is high, vehicles can be hard to control especially for the new drivers, and the visibility usually decreases as well. High wind speed weather, like hurricane and other hazard, will blow off branches and create debris on the road, which is a huge safety risk to drivers. Thus, higher wind speed has higher probability of causing a severe accident.

- Visibility

  Lower visibility can cause difficulty on driving, but it does not necessarily have impact on accident severity, because drivers will be more cautious and careful. They usually will slow down their speed and pay more attention to the road ahead, thus the possibility of severe accident is decreased. So visibility has no significant impact on accident severity.

- Total precipitation in the past 24 hours and precipitation intensity

  The accumulation of precipitation, no matter rain or snow, has a significant effect in accident severity, which is intuitive and can be easily understood. However, heavy precipitation has less power impacting accident severity comparing with other intensity. Heavy rain weakens the visibility and wets the road surface, which causes drivers to pay more attention while driving, thus with the decrease of traffic flow the probability of severe accident decreases.
- Relative humidity

  Based on the result, relative humidity also has a significant impact on accident severity. Similar with the impact of precipitation, the high relative humidity can affect vehicle's normal function and driver's feeling, but it may be hard to quantify the impact.

Moreover, based on small sample analysis reported in Chapter 6, we can also come to the following conclusions:

Comparing with other weather factors, relative humidity, average wind speed and heavy precipitation have more significant impact on accident severity. Also, between different levels of severity, a particular weather factor can have different impact. It is possible that while holding all other variables in the model constant, the relative humidity and average wind speed have significant impact on the possibility of having accident with property damage, but have insignificant impact or no impact on the possibility of having accident with people injury or fatality. And heavy precipitation may have a significant impact only on the possibility of having an accident with fatalities, but not on the other two. Therefore, the impact of weather factors on road accident severity should be analyzed case by case before specific conclusions are made.

7.2 Future Research Directions

In this research, data collected from 2007 to 2010 were used to build regression models with weather-related factors and highway accident severity. The conclusion was reached that a multinomial logit model has the best performance in explaining the data, and all factors except visibility and heavy precipitation have significant impact in accident severity. The same procedure was used on the US 50, and the results indicated that fewer factors were significant in impacting accident severity. However, this research has several drawbacks which are stated as follow.

Data is crucial for accident analysis. The quality of the research is highly dependent on the availability and quality of data. In this research, weather data was collected from five weather stations statewide, and each station represents the weather condition in an entire zone, which lacks precision. Also, if we can connect the non-accident data, such as total highway traffic flow in a certain area under the same weather condition and combine it with the same date and location, we would be able to compare the difference between severe weather condition and normal weather condition, which is more appropriate and accurate. To obtain more valuable data will be very helpful in improving future research.

Although it has been determined that most weather-related factors contribute in causing more severe accidents, more attention still needs to be given to the joint impact of geometric characteristic and environmental factors on accident severity. A lot of research has been conducted during the past decades to investigate the relationship between the geometric characteristics and car accidents. The next step of study should combine all significant factors together to build a more complicated and well-explained model.

This study is aimed at analyzing the impact of weather conditions on accident severity, but the study of weather-related factors and accident frequency is also highly recommended. The accident rate and severity regression models are similar with respect to some significant weather-related factors, but are also different with respect to other factors. Even for those common factors, the significance of influence may not be the same. By conducting an accident frequency regression model, we can estimate and predict both accident occurrence rate and severity. Thus, a model with two dependent variables can help governments to better analyze the impact of weather condition on traffic safety and make effective mitigation decisions.

References

1. Abdel-Aty, M., (2003), "Analysis of Driver Injury Severity Levels at Multiple Locations Using Ordered Probit Models," *Journal of Safety Research*, Volume 34, No. 5, pp. 597-603.
2. Abdel-Aty, M., H. T. Abdelwahab, (2004). Predicting Injury Severity Levels in Traffic Crashes: A Modeling Comparison, *Journal of Transportation Engineering*, Volume 130 pp. 204-210.
3. Abdelwahab, H. T., and M. A. Abdel-Aty, (2001), "Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections," *Transportation Research Record 1746*, pp. 6-13.
4. Al-Ghamdi, A. S, (2002), "Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity," *Accident Analysis and Prevention*, Volume 34, pp. 729–741
5. Bedard, M., G.H. Guyatt, M.J. Stones, and J. P. Hireds, (2002), "The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities," *Accident Analysis and Prevention*, Volume 34, pp. 717-727.
6. Brodsky, H., and A. S. Hakkert, (1988), "Risk of a Road Accident in Rainy Weather," *Accident Analysis and Prevention*, Volume 20, pp. 161–176.
7. Broughton, K., F. Switzer, and D. Scott, (2007), "Car Following Decisions under Three Visibility Conditions and Two Speeds Tested with a Driving Simulator," *Accident Analysis and Prevention*, Volume 39, pp. 106–116.

8. Buchner, A., M. Brandt, R. Bell, and J. Weise, (2006), "Car Backlight Position and Fog Density Bias Observer-Car Distance Estimates," *Human Factors*, Volume 48, No. 2, pp. 300–317.

9. Chimba, D., and T. Sando, (2009), "The prediction of highway traffic accident injury severity with neuromorphic techniques," *Advances in Transportation Studies* Volume 19, pp. 17-26.

10. Chong, M.M., A. Abraham, and M. Paprzycki, (2004), "Traffic Accident Analysis Using Decision Trees and Neural Networks," *International Conference on Applied Computing*, Volume 2, pp. 39-42.

11. Corben, B. F. and F. C. Wai, (1990), "Pro-Active Traffic Engineering Safety Study: Final Report Part 2 - Right-Turn-Against Crashes at Traffic Signals," Sponsored by Vic Roads, Accident Research Center, Monash University, Australia.

12. Darçin A., and B. Akba, (2010), "A neural network (NN) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics," *Scientific Research and Essays*, Volume 5, No. 19, pp. 2837-2847.

13. Delen, D., R. Sharda, and M. Bessonov, (2006), "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accident Analysis and Prevention*, Volume 38, pp. 434–444.

14. De O˜na, J., R. O. Mujalli, and F. J. Calvo, (2011), "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks," *Accident Analysis and Prevention*, Vol. 43, pp. 402–411.

15. Federal Highway Administration, Research, Development, and Technology Turner-Fairbank Highway Research Center, (1999), "Crash Models for Rural Intersections: Four-

Lane by Two-Lane Stop-Controlled and Two-Lane by Two-Lane Signalized," Publication No. FHWA-RD-99-128.

16. Federal Highway Administration, Road Weather Management Program, (2012), "How Do Weather Events Impact Roads," http://www.ops.fhwa.dot.gov/weather/q1_roadimpact.htm.

17. Geiger, J., "How to prepare your car for winter," http://www.howstuffworks.com/.

18. Haleem, K., M. Abdel-Aty, (2010), "Examining Traffic Crash Injury Severity at Unsignalized Intersections," *Journal of Safety Research*, Volume 41, No. 4, pp. 347-357.

19. Harwood, D.W., F.M. Council, E. Hauer, W.E. Hughes, and A. Vogt, (2000), "Prediction of the Expected Safety Performance of Rural Two-Lane Highways Report," US Department of Transportation, Federal Highway Administration, Publication Number FHWA-RD-99-207.

20. Horswill, M., A. Plooy, (2008), "Reducing contrast makes speeds in a video-based driving simulator harder to discriminate as well as making them appear slower," *Perception*, Vol.37, pp. 1269–1275.

21. Kashani, A. T., A. Shariat-Mohaymany, and A. Ranjbari, (2012), "Analysis of factors associated with traffic injury severity on rural roads in Iran," *Journal of Injury and Violence Research*, Volume 4, No. 1, pp. 36-41.

22. Kashani, A. T., A. Shariat-Mohaymany, and A. Ranjbari, (2011), "A Data Mining Approach to Identify Key Factors of Traffic Injury Severity," *Scientific Journal on Traffic and Transportation Research*, Volume 23, No. 1, pp. 11-17.

23. Khattak, A. J., A. J. Khattak, and F.M. Council, (2002), “Effects of work zone presence on injury and non-injury crashes,” *Accident Analysis and Prevention*, Volume 34, No. 1, pp. 19-29.

24. Kilpeläinen M., H. Summala, (2007), “Effects of weather and weather forecasts on driver behavior,” *Transportation Research Part F: Traffic Psychology and Behavior*, Volume 10, Issue 4, pp. 288-299.

25. Kockelman, K.M., Y.J. Kweon, (2002), “Driver Injury Severity: An Application of Ordered Probit Models,” *Accident Analysis and Prevention*, Vol. 34, No. 3, pp. 313-321.

26. Kunt, M.M., I. Aghayan, and N. Noii, (2012), “Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods,” *Transport*, Volume 26, No. 4, pp. 353-366.

27. Ma, J., and K. Kockelman, (2006), “Crash Frequency and Severity Modeling Using Clustered Data from Washington State,” IEEE Intelligent Transportation Systems Conference Proceedings, pp. 1621-1626.

28. Mogaka EO, Z. Ng´ang´a, J. Oundo, J. Omolo, and E. Luman, (2011), “Factors associated with severity of road traffic injuries,” Thika, Kenya, *The Pan African Medical Journal*, Volume 8, No. 20.

29. Mujalli, R.O., and J. de Oña, (2011), “A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks,” *Journal of Safety Research*, Volume 42, pp. 317–326.

30. Nassiri, H., and A. Edrissi1, (2006), “Modeling Truck Accident Severity on Two-Lane Rural Highways,” *Scientia Iranica*, Volume 13, No. 2, pp 193-200.

31. National Safety Council, (2010), “Average Economic Cost per Death, Injury, or Crash,”

http://www.nsc.org/news_resources/injury_and_death_statistics/Pages/EstimatingtheCostsofUnintentionalInjuries.aspx.

32. Norrman, J., M. Eriksson, and S. Lindqvist, (2000), "Relations between traffic accidents on slippery roads and winter road maintenance," *Climate Research*, Volume 15, pp. 185–193.

33. Norouzi, A., (2012), "Impact of Dynamic Message Signs on Occurrence of Road Accidents," Master's Thesis, University of Maryland, College Park.

34. O'Donnell, C.J., D.H.Connor, (1996), "Predicting the Severity of Motor Vehicle Accident Injuries Using Models of Ordered Multiple Choices," *Accident Analysis and Prevention*, Volume 28, No. 6, pp.739– 753.

35. Ogden, K. W., S. V. Newstead, P. K. Ryan, and S. Gantzer, (1994), "Factors Affecting Crashes at Signalized Intersections," Research Report No. 62, Monash University Accident Research Center, ISBN 0732600618, Sponsored by Department of Justice, Roads Corporation (VicRoads), Transport Accident Commission, and Royal Automobile Club of Victoria Ltd.

36. Palutikof, J. P., (1991), "Road accidents and the weather," Chapter 8 in *Highway Meteorology*, A. H. Perry and L. J. Symons, editors, pp. 163–189.

37. Patil, S., S. Geedipally, and D. Lord, (2011), "Analysis of Crash Severities using Nested Logit Model- Accounting for the Underreporting of Crash Data," *Accident Analysis & Prevention*, Volume 45, pp. 646-653.

38. Sarle, W.S., (1994), "Neural Network Implementation in SAS Software," Proceedings of the Nineteenth Annual SAS Users Group International Conference.

39. Sarle, W.S. (1994), "Neural Networks and Statistical Models," Proceedings of the Nineteenth Annual SAS Users Group International Conference.

40. Savolainen, P.T., F. L. Mannering, D. Lord, and M.A. Quddus, (2011), "The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives," *Accident Analysis and Prevention*, Volume 43, No. 5, pp. 1666–1676.

41. Scialfa, C. T., G. Ho, J. K. Caird, and T. Graw, (1999), "Traffic Sign Conspicuity: The Effects of Clutter, Luminance, and Age," Proceedings of 43rd Annual Meeting of the Human Factors and Ergonomics Society, pp. 108-112.

42. Sherretz, L. A., and B. C. Farhar, (1978), "An analysis of the relationship between rainfall and the occurrence of traffic accidents," *Journal of Applied Meteorology*, Volume 17, pp. 711–715.

43. Shoukry, F. N., (2005), "Artificial Neural Network in Classification of Severity Levels in Crashes with Guardrail," Master's thesis, West Virginia University.

44. Sze, N. N., and S. C. Wong, (2007), "Contributory Factors to Pedestrian Injury Severity in Traffic Crashes," Presentation at UC Berkeley.

45. Vivoli, G., M. Bergomi, S. Rovesti, G. Carrozzi, and A.Vezzosi, (1993), "Biochemical and hemodynamic indicators of stress in truck drivers," *Ergonomics*. Volume 36, No. 9, pp. 1089–1097.

46. Voelcker-Rehage, C., and J. L. Alberts, (2007), "Effect of motor practice on dual-task performance in older adults, "*The Journals of Gerontology Series B, Psychological sciences and social sciences,* Volume 62, No. 3, pp. 141–148.

47. Washington, S., M. Karlaftis, and F. Mannering, (2011), "*Statistical and econometric methods for transportation data analysis*," Second edition, Chapman and Hall/CRC, Boca Raton, FL.

48. Ye, F., (2011), “Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models,” Presented at the 90th Annual Meeting of the Transportation Research Board. See http://amonline.trb.org/12kbp6/12kbp6/1

49. Zhang, H., (2010), “Identifying and Quantifying Factors Affecting Traffic Crasg Severity in Louisiana,” PhD. Dissertation, Louisiana State University.