Research Report No. UVACTS-5-15-74
December 2001

Stochastic Models Relating Crash Probabilities With Geometric And Corresponding Traffic Characteristics Data

by

Dr. Nicholas J. Garber

Lei Wu

A Research Project Report
For the National ITS Implementation Research Center
A U.S. DOT University Transportation Center

Dr. Nicholas J. Garber
Department of Civil Engineering
Email: njg@virginia.edu

Lei Wu
Department of Civil Engineering
Email: lw3h@virginia.edu

Center for Transportation Studies at the University of Virginia. During the late 1940s, the University of Virginia Department of Engineering began an ongoing partnership with the research branch of the Virginia Department of Transportation's Virginia Transportation Research Council (VTRC). Today, the Center for Transportation Studies (CTS) organizes the transportation academic and research activities within the CE Department. CTS receives substantial support from the USDOT, particularly from two federal University Transportation Center Grants- the Mid-Atlantic Universities Transportation Center (MAUTC) funded through Pennsylvania State University, and the ITS Implementation Research Center (ITS), funded through George Mason University. It is also supported by contracts with the Virginia DOT, VTRC and other private and public agencies.

1

CTS Website
http://cts.virginia.edu

Center for Transportation Studies
University of Virginia
351 McCormick Road, P.O. Box 400742
Charlottesville, VA 22904-4742
804.924.6362

Abstract

Several kinds of models have been developed in modeling the occurrence of crashes, but most of these models have deficiencies and lack good results. Single and multivariate deterministic models have illustrated some influences of causal factors on crashes, but the inherent deterministic characteristic of these models makes explaining crash events a difficult task for these kinds of models. Stochastic regression models, such as Poisson, Negative Binomial, and Zero Inflated Poisson (ZIP) have been explored to account for the discrete and stochastic characteristics of crashes. However, no consistent results have been illustrated yet. The possible reasons for the deficiencies of existing models could be attributed to the modeling methodologies or the data set used. We have opportunities to obtain corresponding traffic data related to the time when crashes take place from the Smart Travel Lab at University of Virginia. Those data include volume, speed, and occupancy.

Based on the review of the existing models, data obtained from the Smart Travel Lab were used in the application of several stochastic regression models including Poisson, negative binomial, zero inflated Poisson, and zero inflated negative binomial regression models. The selected variables include crash counts, volume, speed, occupancy, curvature, exposure, and standard deviation of speed. Negative binomial and ZIP were shown to be preferred modeling methods for this study. Significant positive relationships were also identified between the occurrence of crashes and volume, standard deviation of speed, and exposure. These relationships could be applied to provide powerful support to the decision making of incident management in Intelligent Transportation Systems.

Table of Contents

List of Figures

List of Tables

Glossary of Terms and Abbreviations

Alltime

Alltime reflects the amount of exposure over which the dependent variable was observed. It is 365 days*3 years=1095 days·years in this study.

Coef.

Coef. represents coefficients.

Count

The number of crashes within a specific hour throughout the three years from July 1st, 1997 to July 1st, 2001.

Curve

Curve is the surrogate measure measuring the average curvature for a specific segment of roadway. It is estimated from the length of circular curve, length of transition curve, radius of circular curve, and total length of section.

Expose

Expose is the product of the length of a specific segment of roadway and the number of days of a specific weekday group.

Freq.

Freq. represents frequency.

Inflate

Inflated indicates the inflated model part.

No. of Obs.

No. of Obs. Represents the number of observations.

Occu

Occu is the average occupancy of a specific hour during one of the four weekday groups throughout the three years from July 1st, 1997 to July 1st, 2001.

P>|Z|

P value of the z test.

Speed

The average speed of a specific hour during one of the four weekday groups throughout the three years from July 1st, 1997 to July 1st, 2001.

Std.Err.

Std. Err. represents standard error.

Stdsp

Stdsp is the standard deviation of speeds within a specific one hour during one of the four weekday groups throughout the three years from July 1st, 1997 to July 1st, 2001.

Volume

The average hourly traffic volume per lane of a specific hour during one of the four weekday groups throughout the three years from July 1st, 1997 to July 1st, 2001.

Z

Z represents the z test statistic.

_cons

_cons is the constants in the inflated and base models.

95% Conf. Interval

95% Conf. Interval represents 95% confidence interval.

CHAPTER 1: INTRODUCTION

Many projects have been conducted on modeling crashes. The existing models include single variable and multivariate deterministic models, stochastic multivariate models, and artificial neural network. After a long period of application of deterministic models, researchers began to realize that the characteristics of crashes are discrete, sporadic, and non-negative. Researchers, therefore, started applying stochastic models to describe the occurrence of crashes. Jovanis et al.[1] proposed the idea of applying the Poisson distribution in modeling the occurrence of crashes to overcome the shortcoming of conventional regression models. In 1990, Garber and Joshua [2] performed stochastic crash prediction models using Poisson regression models. Later on, more stochastic models were explored, such as Zero Inflated Poisson[3], negative binomial regression models[3]. Meanwhile, outside the highway vehicle crash modeling area, many studies about zero inflated models were performed. For example, Lambert [4] published a paper on ZIP regression models and their application to defects in manufacturing. Greene [5] further discussed the estimation of zero inflated Poisson and zero inflated negative binomial regression models. Gan [6] proved the uniqueness and consistence of maximum likelihood estimators for zero inflated models under appropriate regularity conditions. Overall, researchers think zero inflated regression models are sound theoretically and practically in describing the count data with excess zero occurrences which is very common in nature. This kind of models have not yet been fully developed in the study of highway vehicle crashes. In another study, Awad et al.[7] compared linear regression and artificial neural networks approaches.

Apart from the modeling techniques, the causal variables considered in those crash models have shifted from single variable to multiple variables. Various relationships were studied, such as the relationships between the number of lanes and crash rates,[8] traffic volume and crash rates, [8,9,10] shoulder and lane widths and crash rates.[11,12,13] Traffic volume was believed to have significant influence upon the occurrence of crashes. A U-shaped curve for the relationship between the crash rate and traffic volume was shown. However, no consistent relationships between the occurrence of crashes and geometric parameters have been indicated yet. Due to the complexity of the occurrence of crashes, multiple factors were considered and applied in the modeling of crashes. Those factors included road geometry, traffic variables, and environmental conditions such as lighting and weather. In Garber and Ehrhart's study[14], speed variance was shown to have significant influence upon the occurrence of crashes.

Overall, major deficiencies related to current studies are the following:

- Too many independent variables did not help the clear indication of the effects of those variables on the occurrence of crashes.
- Too much information reduces the reliability of models.
- The application of advanced modeling techniques such as zero inflated Poisson models placed emphasis on geometric parameters while traffic volume was usually the only traffic variable included.
- Daily traffic volume account was adopted in the previous research on the modeling of crashes, which could not reflect the important hourly traffic variation pattern.
- Inappropriate measures were used to evaluate the goodness-of-fit of the estimated models, for example R^2 and Akaike Information Criteria (AIC). R^2 has been shown to

be good only for deterministic models, whereas Poisson and negative binomial regression models are stochastic. Although AIC is applicable to nested models, it has been used to compare Poisson and ZIP. According to Greene [5], Poisson is not nested to ZIP regression models. Similarly, negative binomial is not nested to zero inflated negative binomial regression models.

The existing studies indicate that zero inflated models could be used to examine the influence of traffic variables on the occurrence of crashes. These zero inflated models include various forms of zero inflated Poisson and zero inflated negative binomial regression models. They are extensions of Poisson and negative binomial regression models, which account for greater frequency of zero events than those predicted by the conventional generalized linear regression models. Since preliminary investigation of crash data indicated many zero crashes, zero inflated models were therefore included in this study. Traffic parameters and curvature were selected as independent variables in this study. The influences of other factors were standardized by selecting observations under certain consistent conditions. For example, crashes in bad weather (rain, fog, snow, and so on) were filtered to eliminate the effects of the bad weather. Also, only crashes that occurred in the daytime were selected for analysis to eliminate the influence of lighting conditions. Due to the relative short length of each road section selected (one to three miles) and general flat terrain, geometric parameters other than curvature were not included in this study.

1.1 Purpose and Scope

The purpose of this study is to describe the relationships between the crash probability and traffic and geometric characteristics.

This study is limited to a selected section of Interstate 64 within Norfolk, Virginia Beach, and Chesapeake in Virginia. This road section was selected because both the crash and traffic data are available since July 1st, 1998. Both the crash data and corresponding traffic data come from the database of Integrated Transportation Systems Management (ITSM) of the Smart Travel Lab at the University of Virginia. The time period of the data set is from July 1st, 1998 through July 1st, 2001.

1.2 Study Objectives

The specific objectives of this study include:

- To review the existing crash modeling methods and identify the feasible methodologies;
- To examine the identified relationships between the occurrence of crashes and related causal factors including traffic and geometric variables;
- To apply the selected methods of modeling using real data from the Smart Travel Lab (STL) at the University of Virginia; and
- To identify how the independent variables contribute to the occurrence of crashes based on the obtained models.

CHAPTER 2: LITERATURE REVIEW

The literature on crash modeling has been thoroughly examined. The review identified the causal factors of crashes, relationships between crashes and causal factors, variables selected in the models, and different modeling methodologies. The literature review was summarized according to the following different modeling methods:

- Single and multivariate deterministic models
- Stochastic models
- Multiple-logistic models
- Artificial Intelligence approaches
- Fault tree analysis
- Classification and Regression Tree (CART) analysis

2.1 Model Description

2.1.1 Single and Multivariate Deterministic Models

The main task of reviewing the deterministic models is to identify established relationships between the crash rate and traffic variables, highway geometric configurations, and environmental factors. Deterministic models have a major flaw: assuming that the number of crashes is continuous and the error of the dependent variable is normally distributed with a constant variance. This assumption is misleading because the occurrence of crashes is random, discrete, and rare. Despite this error, the results of previous studies indicate the influences of various factors on the occurrence of crashes.

At first, researchers only paid attention to relationships between crashes and different single variables. Various relationships were studied, such as the relationships between the number of lanes and crash rates,[8] traffic volume and crash rates,[8,9,10] shoulder and lane width and crash rates.[11,12,13] Traffic volume was believed to have a significant influence upon the occurrence of crashes. It is generally believed that the single-vehicle crash rate decreases while traffic volume increases, whereas the multiple-vehicle crash rate increases with increasing traffic volume. Also, it is commonly accepted that the relationship between crash rates and traffic volumes presents a "U" shape.[9,10]

In one of the studies, Zegeer et al.[12] reviewed 30 highway safety studies and selected four, and from which they extracted data to develop relationships between vehicle safety and lane width, shoulder width, and shoulder type. These three factors were indicated to have significant effects on highway vehicular safety.

In two other studies, Glennon [15,16] studied the effects of alignment and sight distance on highway safety by reviewing previous literature. He[16] noted in his review that there was no clear effect of improved intersection sight distance on highway safety. Garber et al.[14] studied the influence of causal factors on the occurrence of crashes. Speed variance was shown to have a positive relationship with the crash rate in this study. However, relationships regarding the influence of other variables, such as highway geometric parameters, upon crash rates were not consistent. This may be due to the omission of influencing variables, the lack of ample consistent data, or the inherent disadvantage of deterministic models.

Recognizing the complexity of crashes, researchers realized the necessity of using multivariate models in modeling crashes. However, although these models account for

the influences of multiple factors, they are more difficult to analyze. Graphs have been used to describe the influence of some variables while other contributing factors were kept fixed. In these cases, they could be regarded as single-variate models. At most, only three-dimensional graphs could be drawn, making it difficult to pictorially express the interactions among more than three variables. Both qualitative and quantitative models were used to describe the relationships between causal factors and safety in single-variate models and multivariate models, while researchers used more quantitative multivariate models

Mohamedshah et al.[17] used data from the Highway Safety Information System (HSIS) to formulate several general deterministic models and selected the multivariate linear regression model with the highest R^2. Comprehensive data were used in this study. Although modeling methods in that study remained unimproved, the achieved models were better. Garber et al.[14] developed multivariate deterministic models of highway crashes. In their study, several significant independent variables were identified, which included mean speed, standard deviation of speed, flow per lane, lane width, and shoulder width. Three kinds of deterministic models were performed, which included multiple linear regression, robust regression, and multivariate ratio of polynomials models. The results of this study showed that the relationships between crash rate and speed, flow, and geometric characteristics were not linear. The final models showed that speed variance had a significant influence on the crash rate.

Persaud et al. [18] noted several problems of current modeling research, which included:

- ADT was used instead of corresponding traffic volumes, which is more appropriate than the former in describing traffic volumes when crashes occur.
- The crash rate (crashes per unit of traffic) was widely used. In doing so, researchers made an assumption that crashes are proportional to traffic intensity, which is not necessary correct.
- Conventional deterministic regression models assumed that the error of the dependent variable follows the normal distribution.

To compensate for these disadvantages, Persaud et al. used the hourly traffic volume and hourly crash count in this study and applied the Generalized Linear Modeling Computer Package (GLIM). GLIM allows for the specification of a negative binomial error structure for the dependent variable.

Single and multivariate deterministic models show the influences of some causal factors over the occurrence of crashes, but the resulting quantitative relationships were not consistent. Deterministic models lack the ability to explain the stochastic occurrence of crashes. This deficiency could be one of the major reasons contributing to the performance of deterministic models.

2.1.2 Stochastic Models

What completely differentiates stochastic models from deterministic models is the assumption rooted under the former models: the occurrence of vehicular crashes is random. Early in 1989, Okamoto et al. [19] suggested that the occurrence of traffic crashes is stochastic. In 1990, Garber et al.[2] first developed several Poisson regression models to describe the occurrence of crashes. Various studies [3,20,21,22,23,24,25]further examined the goodness-of-fit of Poisson regression models. More stochastic models were proposed

other than Poisson regression models, which included ZIP[3], Negative Binomial [3,21,26,27], and Extended Negative Binomial[24] regression models. While dependent variables in these models are stochastic, the link functions are deterministic. The link functions are used to connect the mean of crash counts with independent variables.

2.1.2.1 Model Forms

2.1.2.1.1 Poisson Regression Models

Poisson models assume that vehicle crashes are independent and follow Poisson distribution. Miaou et al.[23] proposed two kinds of multiplicative models and one kind of revised Poisson model.

- Multiplicative Poisson Regression Model 1

$Y_i \sim$ ind Poisson (μ_i)

Or $p(Y_i = y_i) = p(y_i) = \dfrac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \qquad i = 1,2,...,n.$

Where:

$p(y_i)$— the probability of the occurrence of y_i crashes for a given time period on roadway segment i.

Y_i— the number of crashes for a given time period for roadway segment i;

μ_i— mean value of crashes occurred for a given time period;

$$\mu_i = E(Y_i) = \nu_i \left[e^{\sum_{j=1}^{k} x_{ij}\beta_j} \right] \qquad i = 1,2,3,...,n.$$

x_{ij}— the j^{th} independent variable for roadway segment i;

β_j— the coefficient for the j^{th} independent variable;

ν_i — traffic exposure for roadway segment i.

- Multiplicative Poisson Regression Model 2

A different function was used as the link function. The symbols remain the same as above.

$$\mu_i = E(Y_i) = \nu_i \left[\beta_1 \left(\prod_{j=2}^{k} (1 + x_{ij})^{\beta_j} \right) \right] \qquad i = 1,2,3,...,n$$

- Revised Multiplicatve Poisson Regression Model

$Y_i \sim$ ind Poisson (μ_i)

Or $p(Y_i = y_i) = p(y_i) = \dfrac{\mu_i^{y_i} e^{-\mu_i}}{y_i} \qquad i = 1,2,...,n.$

Where

$$\mu_i = E(Y_i) = \nu_i^{\beta_0} \left[e^{x'_i \beta} \right] = \nu_i^{\beta_0} \left[e^{\sum_{j=1}^{k} x_{ij} \beta_j} \right] \qquad i = 1,2,3,...,n$$

Where β_0 is the coefficient for v_i.

The Poisson regression model assumes that the variance of the dependent variable is equal to the mean. However, in many applications, count data were found to display extra variation or over-dispersion relative to a Poisson model. The over-dispersion means the real variance is greater than that computed from the Poisson models. Miaou et al.[23] indicated that the violation of the assumption that the mean equals the variance does not change parameter estimates but causes the underestimation of the variances of the estimated coefficients. Several methods were proposed to remedy this deficiency:

- Multiply μ_i with the over-dispersion parameter τ and use the resulting $\tau\mu_i$ as the variance of Y_i instead of μ_i.[3]

$$\tau = \frac{\chi^2}{N - p}$$

Where

χ^2 is the Pearson's chi-squared statistic for the model;

$$\chi^2 = \sum_i (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i$$

Where:

y_i — the number of crashes or the crash rate for i roadway segment;

μ_i — mean value of crashes occurred;

N is the number of observations; and

p is the number of parameters considered in the model.

- Ivan et al.[25] suggested using quasi-likelihood rather than maximum likelihood estimation. He indicated that "Both techniques use the same log-likelihood function, but quasi-likelihood estimation does not make any assumption about the distribution because it allows for separate mean and variance structures by computing the dispersion parameter and assuming that the variance is equal to the product of the mean and the dispersion. "
- Miaou[3] suggested dividing the t-statistics obtained from the Poisson regression model by $\tau^{1/2}$ to get better estimates of the t-statistics. The significance level for acceptance was reduced somewhat after that adjustment, but, as the author indicated, the relationship between truck crashes and influencing factors remained unchanged.

2.1.2.1.2 Zero Inflated Poisson (ZIP) Regression Models

Miaou[3] applied the following ZIP model in the prediction of highway vehicular crashes:

$$p(Y_i = y_i) = e^{-\theta r_i} \qquad if \quad y_i = 0$$

$$= \left(\frac{1 - e^{\theta r_i}}{1 - e^{-r_i}}\right)\frac{r_i^{y_i} e^{-r_i}}{y_i!} \qquad if \quad y_i = 1,2,3,...,n.$$

Where:

$0 < \theta \leq 1$, when $\theta = 1$, the ZIP regression model is essentially Poisson regression model;

$p(y_i)$ — the probability of the occurrence of y_i crashes on roadway segment i.

Y_i — the number of crashes for a given time period for roadway segment i;

r_i — the mean value of crashes for a given time period occurred;

$$r_i = \nu_i \left[e^{\sum_{j=1}^{k} x_{ij}\beta_j} \right] \qquad i = 1,2,3,...,n$$

x_{ij} — the j^{th} independent variable for roadway segment i;

β_j — the coefficient for the j^{th} independent variable;

ν_i — traffic exposure for roadway segment i;

The mean and variance of Y_i are

$$\mu_i = E(Y_i) = \left(\frac{1 - e^{-\theta r_i}}{1 - e^{-r_i}}\right) r_i$$

And

$$Var(Y_i) = \mu_i + \left(\frac{1 - e^{r_i(\theta - 1)}}{e^{\theta r_i} - 1}\right)\mu_i^2 \quad = \mu_i + \Phi_i \mu_i^2$$

Where Φ_i is a function of r_i and θ. When 0< θ <1, the variance is greater than the expectation, which allows over-dispersion in the data.

2.1.2.1.3 Negative Binomial Regression Models

In Miaou 's study[3], the negative binomial regression model was of the form:

$$p(Y=y_i)=\frac{\Gamma(\frac{1}{\alpha}+y_i)}{\Gamma(\frac{1}{\alpha})\Gamma(y_i+1)}(\frac{1}{1+\alpha\mu_i})^{\frac{1}{a}}(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y_i} \qquad y_i = \ 0,\ 1,\ 2\cdots$$

Where

$$\mu_i = E(Y_i) = \nu_i \left[e^{\sum_{j=1}^{k} x_{ij}\beta_j} \right] \qquad i = 1,2,3,...,n$$

And

$$Var(Y_i) = \mu_i + \alpha\mu_i^2$$

Where: Γ(.)= Gamma function; α= rate of over-dispersion.

2.1.2.1.4 Extended Negative Binomial Regression Models

Miaou[1] proposed extended negative binomial model to account for the heterogeneity within one segment of a roadway. The variability could result from significant changes in geometric characteristics within a single roadway segment. While the model form remained the same, one single segment was further divided into subsections. Then the mean of crashes was estimated for each subsection within one segment using constant coefficients for all the subsections. The expected mean for the entire section was obtained as a weighted sum of the means of all subsections with a single segment. By examining each subsection separately, the extended negative binomial regression model accounted for the variability within one segment. Vogt et al.[24] applied

[1] Source: Vogt, A., Bared, J., 1998.

extended negative binomial regression models to describe the occurrence of crashes for two-lane rural segments and intersections.

2.1.2.2 Comparison of Stochastic Models

The Poisson regression model has a more concise model form than other stochastic regression models. It has only one parameter to estimate. However, as mentioned above, over-dispersion is a major limitation with the Poisson regression model. The overwhelming occurrence of zero crashes contributes to the inequality of the variance and expected values in the estimated Poisson regression models. Zero inflated Poisson regression models have been adapted to account for the large percentile occurrence of zero crashes in a road segment during a specific time period. The interpretation of ZIP is difficult although it is more flexible than the Poisson regression model. The negative binomial regression models correct the over-dispersion caused by Poisson regression model. However, the computation to estimate model parameters is more comprehensive. Extended negative binomial models provide one method to account for the heterogeneity within single roadway segments.

2.1.2.3 Model Estimation and Selection

The Maximum Likelihood Estimation (MLE) and the quasi-likelihood estimation were used to estimate the stochastic models. These two methods do not assume that the error of the dependent variable follows normal distribution.

Many studies [2,3,14,20,21,22,23,24,26] used Akaike's Information Criterion (AIC) to measure the model performance and t-statistic to evaluate the significance of selected variables. AIC is defined as follows:

$$AIC = -2ML + 2K$$

Where ML is the maximum log-likelihood and K is the number of free parameters in the model. The best model has the minimum AIC value. Pearson's chi-square statistic and likelihood ratio statistic were also used to assess the overall goodness-of-fit of models.[20] Apart from mathematical criteria, Miaou[3] suggested the following aspects could be examined for model acceptance:

- Regression parameters should be consistent and have expected algebraic signs;
- These estimated models should make good engineering sense; and
- Estimated crash involvement should be consistent with the observed value.

2.1.2.4 Variable Selection

Highway geometric characteristics, traffic variables and other contributing variables have been selected in stochastic regression models for study. The number of variables selected is more than that in conventional regression models.

Garber et al.[2] studied the relationship between the probability of large-truck crash involvement and the following independent variables: number of lanes, lane width, shoulder width, and curvature change ratio. They also included absolute mean slope, slope change rate, segment length, AADT, mean speeds for trucks, non-trucks, and all vehicles, speed variance for all vehicles and trucks. Other variables selected were difference in mean speeds between trucks and non-trucks. Slope change rate, curvature change rate and AADT. They also used segment length, percent of large trucks, speed difference in the final models because of their significant influence on the large-truck crashes. Miaou et al.[22] selected yearly dummy variables, AADT per lane, horizontal curvature, vertical grade, and deviation of paved inside (or left) shoulder width as the

covariates. Other than focusing on highway geometric and traffic factors, Fridstr∅m et al.[27] applied the following variables in their models: weather, daylight, road investment and maintenance expenditure, crash reporting routines, vehicle inspection, law enforcement, seat belt usage, proportion of inexperienced drivers, and alcohol sales additional to exposure and traffic density. Karlaftis et al.[26] studied the influence of VMT, population, total road mileage, the proportion of city mileage, time variable, and the proportion of urban roads in total VMT on the crashes. Hadi et al.[21] indicated that transformations of the variables should also be investigated for possible inclusion in the final models.

2.1.2.5 Deficiency of Stochastic Models

Stochastic models have been proven effective to describe the occurrence of crashes. However, the complex nature of crashes makes the generalization of stochastic models a difficult task. There are differences among various models with respect to the variables selected and the significance of variables in the estimated models. Also, some dubious variables were adopted in the models, such as the AADT and crash rate. AADT does not represent the true traffic condition when the crash occurs. By using the crash rate as the dependent variable, a linear relationship has been assumed between the number of crashes and associated traffic exposure. The other problem with the crash rate is that the dependent variables in stochastic models could only be event counts while the crash rate is not an integer. Therefore, the stochastic models are not good for the modeling of the crash rate. Also, the probability itself is not as direct as the crash number or crash rate in quantifying the crashes. For people unfamiliar with the probability of certain number of crashes, this might be a slight problem.

2.1.3 Multiple-Logistic Models

2.1.3.1 Multivariate Logistic Regression Models

A logistic model is designed to describe the probabilities of count variables. The figure of logistic model is S-shaped. The value of dependent variable improves significantly when the independent variable or the function of independent variables of the model reaches certain thresholds. Garber et al.[2] used multiple logistic regression models to analyze the relationship between the probability of truck crash involvement and highway geometric and traffic variables. Lin et al.[28] applied time dependent logistic regression model to analyze the relationship between safety and truck driver service hours. Lin et al.[28] studied the relationship between safety and truck driver service hours. Several other independent variables were also selected, which included age, experience, multi-day driving pattern, and off-duty time before the trip of interest. He developed logistic regression models and found that the driving time had a strong influence on safety performance. However, the first four hours of driving time had the least effect on safety. After that the accident risk increased significantly until the ninth hour. Below is the form of a multiple logistic regression model.

$$f(z) = \frac{e^{-z}}{1+e^{-z}}$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Where:

k is the number of variables

X_i is the i^{th} variable

β_i is the i^{th} coefficient

Z is an index that combines the x's.

2.1.3.2 Assess the Significance of the Variables and Goodness of Models

Apart from likelihood ratio test, stepwise procedure was used to select the independent variables and AIC was used to measure the goodness-of-fit of models[2]. The wald test statistic is also a good option to measure the significance of the variables. Wald test statistic follows the Z distribution computed by dividing the estimated coefficient ($\hat{\beta}$)by its standard errors ($s\hat{\beta}$).

$Z = \frac{\hat{\beta}}{s\hat{\beta}}$ is approximately N(0,1),

LR $\approx Z^2_{Wald}$ in large samples.

Likelihood ratio (LR) statistic describes the difference between log likelihood statistics for the nested and full model. It has an approximate chi-square distribution in large samples. Its value is given in the following formula:

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

Where:

The degrees of freedom for LR statistics are equal to the difference between the number of parameters in the two models;

$-2\ln\hat{L}_1$ and $-2\ln\hat{L}_2$ are log likelihood statistics of the two compared models; and

$\hat{L}_1$ and $\hat{L}_2$ are the maximized likelihood values of the two models.

When the test sample is large, the likelihood ratio statistic and the corresponding squared wald statistic give approximately the same value. However, when dealing with

small to moderate samples, the two statistics may give different results. The likelihood ratio statistic has been shown to be better than the wald statistic. The wald statistic is convenient to use because only the full model needs to be fitted.

Another good measure for logistic models is the percentage of subjects in the data set that are classified correctly. Correct Classification Rate (CCR) is such a measure. Sometimes, models with similar CCR could have considerably different R^2 because the predicted values are rounded when the CCR is computed. R square could be used as a supplementary measure to CCR when various models are compared.

$$CCR = \frac{C}{T}$$

Where:

C: number of correct classifications;

T: total number of classifications.

2.1.4 Artificial Intelligence Approaches

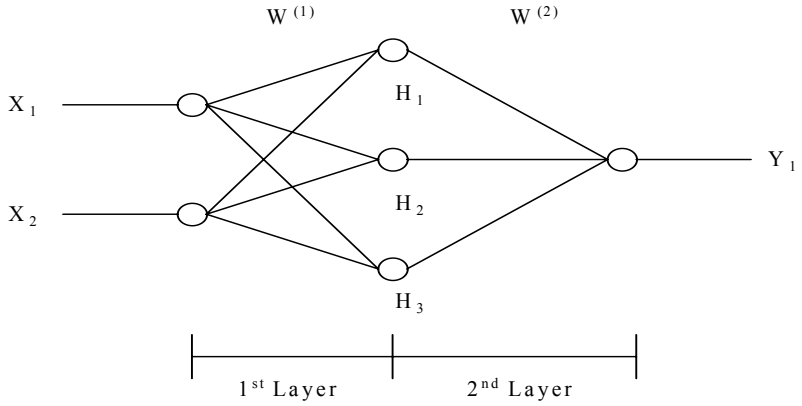
Artificial Neural Networks (ANN) and fuzzy methods belong to the paradigm of artificial intelligence. ANN and fuzzy methods have been applied in highway safety research.

2.1.4.1 Artificial Neural Networks

In Award et al.'s[7] study, apart from the linear regression, ANN and a hybrid system combining fuzzy logic and neural networks were applied to the data. The hybrid system took advantage of the properties and strengths of both fuzzy logic and neural networks. The final models had four variables: gore-to-taper distance, ramp traffic volume, main road traffic volume, and truck percentages on the main road. The

dependent variable was the truck crash frequency at each ramp location. The study indicated that the two ANN approaches showed good performance in identifying different patterns of crashes in the training data while their performance with test data was unsatisfactory. The authors thought that ANN techniques were good choices in analyzing highway vehicular crashes because simple models could not represent the complex relationships between crashes and causal factors. Apparently, ANN techniques require more training data to obtain satisfactory results. Vogt et al.[29] stated in their literature review that ANN could be a good alternative modeling method to the stochastic regression model. A typical neural network was presented in that report.

Figure 2.1 A Typical Neural Network



Where:

$W^{(1)}$, $W^{(2)}$ — weight matrices, not limited only to two;

X_1, X_2 — independent variables, not limited only to two variables;

H_1, H_2 — hidden units, not limited only to two variables, which can be

$H_j = f(\sum_k W_{jk}^{(1)} X_k)$;

Y_1 — dependent variable, which is

$Y_i = f(\sum_k W_{ij}^{(2)} H_j)$

$f(*)$ is the activation function, which could be one of the following forms : step function, linear function, ramp function, and sigmoid function. The sigmoid function is of "S" shape. The simplest form of the sigmoid function is:

$$f(y) = \frac{1}{1+e^{-y}}$$

While the activation function tends to be the same for all the units and variables, it is not required to be universal for the whole network.

2.1.4.2 Fuzzy Methods

Vaija[30]discussed fuzzy methods and applied them in the study of safety. Three different fuzzy methods were discussed in that study: the simple fuzzy expert system, fuzzy linear regression and fuzzified linear programming. The simple fuzzy expert system is based on fuzzy simulation. It can accommodate heterogeneous and partially inconsistent data of different ranges of accuracy. For the fuzzy linear regression, the error term is assumed to be proportional to the indefiniteness of the whole system instead of being the deviation between the observed and estimated values of the dependent variable in conventional regression analysis. Thus, the error term can be described by the fuzziness of the parameters. The fuzzified linear programming is a modification of traditional linear programming. The uncertainty is considered to enter the system when the expert knowledge is used to specify values of and the relationships among variables. Although that paper was about the process control and accident analysis, it was very

helpful in the modeling of highway crashes since vagueness is common among all kinds of accident process. Fuzzy methods discussed in that paper presented good modeling alternatives.

2.1.4.3 Hybrid Methods

To combine merits of different methods, it is useful to fuse them. Award et al.[7] applied hybrid system using fuzzy logic and neural networks to predict crash frequency. The following fusion were taken in applications of many areas: (1) Neural networks for designing fuzzy systems; (2) Fuzzy systems for designing neural networks; (3) Evolutionary computing for the design of fuzzy systems; and (4) Evolutionary computing in training and generating neural networks.

2.1.5 Fault Tree (FT) Analysis

Garber et al.[2,31] performed fault tree analysis to examine the major factors associated with crashes and the interactions among those factors. The occurrence of a crash was regarded as a process ending with an undesired outcome. The outcome, i.e. a crash, was defined as the top event as it was located on the top of the fault tree. In their study[31], a fault tree consisted of several paths. The authors stated that "these paths were defined such that all possible events or actions leading to the occurrence of crashes are sufficiently described." The possible events in different paths were defined as basic events. The probabilities of basic events in the fault tree were assessed according to the crash data in Virginia from 1984 through 1986. The authors first determined the basic events as major factors which might be one of the driver, vehicle, and environmental factors. Then, interactions between the major factors were accounted as secondary

factors. The probability of a top event was determined through the probabilities of major and secondary factors.

The advantages of fault tree analysis include:

- It can be used to identify the causal factors of crashes clearly and clarify the whole possible processes;
- The probability of a crash can also be obtained; and
- Effective strategies can be provided in accordance with the major and secondary factors. However, a FT analysis needs the incorporation of expert knowledge to decide the major and secondary factors, which might introduce subjective errors. Also, due to the complexity of crashes and interactions among different causal factors, it is hard to separate the influence of a single factor from other factors. Therefore, determining the pure probabilities of basic events is very challenging.

2.1.6 Classification and Regression Tree (CART) Analysis

CARTs are non-parametric procedures for explaining and/or predicting either a categorical or continuous response. Hakkert et al.[32] used the classification and regression tree analysis as a preliminary tool to explain the relationships between independent variables and road crashes. Also, CART was used to identify significant variables for further analysis. It is adaptable in dealing with high dimensional and non-homogeneous data set. The tree structure is very helpful to clarify the relationships between independent variables and crash event and interactions among independent variables.

2.2 Summary of Literature Review and Findings

2.2.1 Variable Selection

- Various traffic and geometric variables were shown to have significant influences on the occurrences of crashes. Important traffic variables include volume, speed variance, and speed. The relationship between the crash rate and traffic volume presented a "U" shape. It was indicated that the larger the speed variance, the larger the crash rate. While speed remained an important factor in the occurrence of crashes, no consistent pattern was found between crashes and the speed. The following geometric variables were identified to be important: curvature, lane width, shoulder width, number of lanes, shoulder type, and grade. Deterministic models did not show consistent results regarding the relationships between crashes and the geometric elements. Recent stochastic regression models showed that adverse geometric conditions showed negative impacts upon crashes. Including as many as possible independent variables was suggested to account for over-dispersion. However, Elvik[33] thought that the accuracy of prediction models depended on whether the safety effect of each variable included is significant or not. Many researchers performed tests on the independent variables before bringing them into modeling.

2.2.2 Modeling Methods

- Single-variate and multivariate deterministic models explored relationships between crashes and the influencing factors. Many of those relationships were qualitative, which incorporated the expert knowledge and past experience. The modeling techniques were relatively primitive and data were not very good.
- Stochastic regression models showed great potential in obtaining the true models of crashes. Several stochastic regression models were applied and their theoretical

disadvantages were compensated. However, the following difficulties make the final success still futuristic:

- Past research did not pay enough attention to traffic flow parameters. The most often-studied traffic variable was ADT. While the ADT does not represent the true traffic volume, hourly traffic volume can reflect the true traffic volume when crashes occur better. Flow data within shorter period can be used which would depend on the availability of data.
- The used data were subject to both sampling and non-sampling errors.
- The resulting models were of limited context without widespread applications to support their credibility.

- Fault tree analysis can clearly identify the causal factors and the whole process of crashes, but it is not feasible in large- scale data modeling. Fault tree analysis needs the probabilities of the occurrence of a certain number of crashes caused by various single factors to calculate the probability of the occurrence of a certain number of crashes under certain circumstances. This probability is hard to be determined because of the difficulty to separate the influences of different factors.
- CART can be used preliminarily to analyze the relationships between independent variables and crashes and identify critical independent variables to be included in the models.
- Fuzzy methods are consistent with the characteristics of crashes.
- ANN was applied in transportation since the early 1990s. There is only one study performed applying ANN to the modeling of truck crashes. In this study, the performance of ANN techniques was not satisfactory. However, neural networks showed promise in

analyzing the complex relationships between the occurrence of crashes and its causal factors.

- Both stochastic regression models and artificial intelligence are worth further research. The specific artificial intelligence systems could be the separate techniques, such as artificial neural networks, or hybrid methods combining the different methods of artificial intelligence. However, equal attention should be paid to details, such as the traffic volume variable, lengths of the selected section of roadway, and cluster analysis of the selected section of roadway. After all these improvements, the model performance should be able to arrive at a higher level.

2.3 Conclusion

A great deal of improvements have been made in the previous research of modeling crashes. The modeling techniques shifted from conventional regression to stochastic regression and artificial intelligence network models. Different measures of evaluation were used including coefficient of determination, log likelihood, and AIC. Also, the availability of good data was improved. Researchers can expect better data to be applied in the modeling of crashes. The objective of modeling crashes was transferred from providing criteria and assessment for highway construction and maintenance to supporting advanced traffic management, incident management, and emergency management. The initial research emphasized the relationships between highway geometric variables and crashes, while current research focus more on exploring the relationships between traffic variables and crashes under a certain geometric characteristics. Although the previous research is helpful in identifying the attribution of

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

In this study, two major efforts were made to contribute to the improvement of existing crash modeling research:

- The traffic data at the time when the crashes occurred were used. These traffic data consist of hourly traffic volume per lane, speed, standard deviation of speed, and occupancy. They were referred to in this study as corresponding traffic characteristics.
- Both conventional stochastic (Poisson and negative binomial) regression models and a specific form of zero inflated Poisson regression models have been shown to be effective in describing the occurrence of crashes. The data set obtained for this study were applied to examine the goodness-of-fit of both conventional stochastic and more zero inflated stochastic regression models. These zero inflated stochastic regression models include zero inflated Poisson and zero inflated negative binomial regression models, which were first applied in the crashing modeling.

CHAPTER 3: METHODOLOGY

The following steps are included in this chapter:

- literature review
- data collection
- data screening, reduction, and aggregation
- Variables in the models

3.1 Literature Review

The previous studies were summarized in chapter 2. Independent variables, modeling techniques, and feasible model performance measures were identified for this study based on the literature review. Traffic volume, speed, occupancy, speed deviation, curvature, and exposure of crashes were selected in the models. Stochastic regression models were used in this study. The selected stochastic regression models consist of the Poisson, negative binomial, zero inflated Poisson, and zero inflated negative binomial regression models. Log likelihood value and vuong were used to measure the goodness of fit of models. Vuong[5] was developed specifically to measure the model performance of zero inflated models. More details are given in chapter four.

3.2 Data Collection

Traffic, crash, and geometric data were applied in this study. The data were obtained from two sources: the Smart Travel Lab (STL) at the University of Virginia (UVA) and measurement of the digital Virginia Department of Transportation (VDOT)

county map. The traffic and crash data were extracted from the Integrated Transportation Systems Management (ITSM) database in the STL at UVA[A] while curvature was measured from the VDOT county map using AutoCAD 14.

The total length of the selected road section is about 15.5 miles. This selected section of road was further defined in the oracle database in STL as 11 road segments, whose length vary from 0.95 to 2.68 miles. The traffic data are updated in ITSM in approximately every two minutes, which means very time-specific traffic data can be obtained from this database. The following figure and table show the basic information of the selected road segments.

Table 3.1 Basic Information of Selected Roadway Segments

Road Segment	City	Location(from-to)	Length(mile)
E64-03	Chesapeake	Greenbrier -Indian river	2.68
E64-02	Virginia beach	Indian river -Twin bridges	1.59
E64-01	Norfolk	Twin bridges -64/44 interchange	1.08
W64-01	Norfolk	64/44 interchange -Northhampton	2.01
W64-02	Norfolk	Northampton -Military	1.07
W64-03	Norfolk	Military -Norview	1.24
W64-04	Norfolk	Norview - Chesapeake	0.95
W64-05	Norfolk	Chesapeake - Tidewater	1.00
W64-06	Norfolk	Tidewater - 64 HOV ramp	1.05
W64-07	Norfolk	64 HOV ramp - Bay Ave	1.78
W64-08	Norfolk	Bay Ave - 4th View	1.08

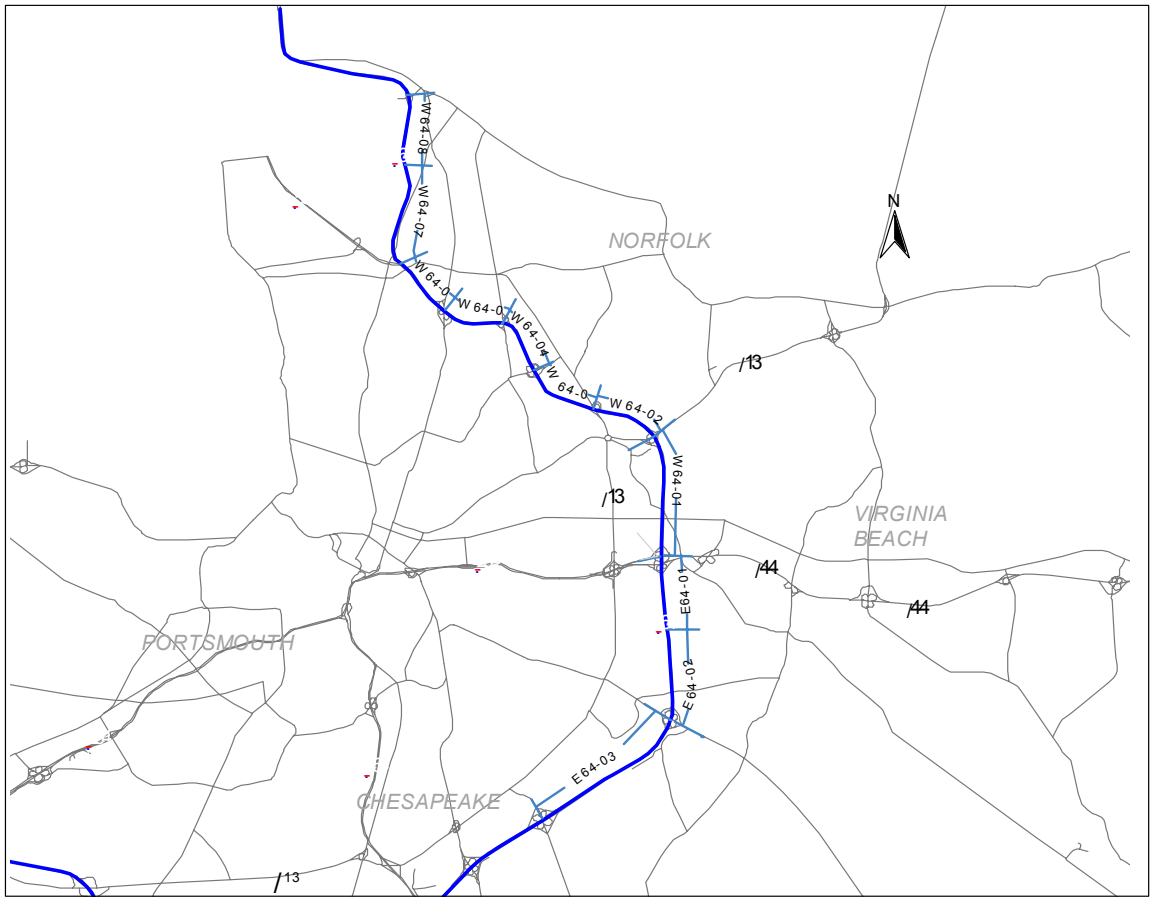
The data collection consisted of the following four tasks:

- Traffic data collection
- Crash data collection

[A] The author would like to thank Ling Qin for providing the SQL query for the extraction of traffic data.

- Curvature Measurement
- Exposure Determination

Figure 3.1 Roadway Segment Location



3.2.1 Traffic Data Collection

The traffic data include hourly traffic volume per lane (vehicle/hour/lane), occupancy (%), and speed (mile/hour) for each segment of roadway in both directions. The occupancy is the percentage of time during which vehicles are over detectors for a specific time, which is defined as:

$$occupancy = C_k * \frac{q}{\overline{\mu_s}}$$

Where:

C_k is a constant;

q is the traffic volume; and

$\overline{\mu_s}$ is the space mean speed.

The traffic data were collected over the whole three years from July 1st, 1998 to July 1st, 2001. In order to reflect the temporal traffic variation and its true influence over the occurrence of crashes, the data collection period was first divided into four weekday groups: (1) Monday, (2) Tuesday, Wednesday, and Thursday, (3) Friday, and (4) Saturday and Sunday. Then, each weekday group was divided into 24 hours and the traffic data were collected accordingly. The oracle database provides 6 sets of data every ten minutes within one hour. For example, from 7:00 am to 8:00 am, each set of traffic volume, speed, and occupancy could be obtained at 7:00 am, 7:10 am, 7:20 am, 7:30 am, 7:40 am, and 7:50 am separately. The hourly traffic volume and occupancy were obtained as the average of the six sets of data within each hour. The hourly speed was obtained as the weighted average of the ten minutes speed data while the traffic volumes were used as the weights. The standard deviation of speed was obtained as:

$$stdsp = \sqrt{\sum_{i=1}^{6}(vol_i * (speed_i - \overline{speed})^2)/(\sum_{i=1}^{6} vol_i - 1)}$$

Where:

stdsp: hourly standard speed deviation of a particular hour,

vol_i : the ith ten minutes traffic volume within the corresponding hour,

$speed_i$: the i[th] ten minutes speed within the corresponding hour,

$\overline{speed}$: the weighted average speed of the corresponding hour,

Where $\overline{speed} = (\sum_i vol_i * speed_i) / \sum_i vol_i$

3.2.2 Crash Data Collection

The time period for the crash data is also three years from July 1[st], 1998 to July 1[st], 2001. The following information was collected for crashes. The crashes were obtained from the database recording all incidents by selecting only those incidents whose type is "accident". Preliminary examination of the crash data indicated that the occurrence of secondary crashes is rare in this study. Therefore, it is reasonable to assume that crashes were mainly affected by geometric configurations, traffic variables, and environmental factors. The information provided the basis for data screening, data reduction, and data aggregation.

- Tms_Call_Number: a specific nine digit number with a dash between the fourth and fifth number for each crash recorded in the database of ITSM.
- Time: day of the year, day of the week, and time of the day were extracted for each crash.
- Weather: the weather conditions when crashes occurred. Crashes happened on rain, snow, fog, and other inclement weather were screened out to eliminate the influence of weather upon the occurrence of crashes.
- Lane: the cross sectional location where specifically the crashes happened were extracted for each crash. Only those crashes happened on mainline of freeway were selected.
- Roadway segment: the specific segment of roadway where the crash happened

- Direction: for each section of roadway, the direction might be east or west. Each direction of the same section of road has a different set of data. They were coded as different segments of roadway in the Oracle database.

3.2.3 Curvature Measurement

The actual curvature of the selected roadway sections was not available. Digital Virginia State county map was used instead. Those roadway sections were selected and imported into AutoCAD. Dimensions of curves, transition lines, and straight lines were measured using AutoCAD. The reason for using AutoCAD was that it can give the dimension of lines and curves automatically. After obtaining the lengths of straight lines, transition lines, and length and radius of each curve within one segment of roadway, the average Curvature Change Rate (CCR) of each roadway section was calculated easily. The CCR was used to describe the curvature of each segment of roadway.

CCR is defined as the absolute sum of the angular changes in horizontal alignment divided by the length of the highway segment.

$$CCR = \left[\sum_{i=1}^{n} \left| \frac{L_i}{R_i} \right| + \sum_{j=1}^{n} \frac{L_s}{2R_i} \right] \frac{(57.3)(1609)}{L} \text{deg/ } mile$$

Where:

L_i = length of circular curve i (m)

L_s = length of transition curves (m)

R_i = radius of circular curve i (m)

L = total length of section (m)

3.2.4 Exposure Determination

3.2.4.1 Exposure

The exposure of crashes is defined as:

$$\exp ose = L * N$$

Where:

Expose is the exposure of each crash (mile*day)

L is the length of the segment of roadway where the crash occurred (mile)

N is the number of days of the weekday groups when the crash occurred (day)

3.2.4.2 Temporal Exposure

The temporal exposure reflected the length of time for which the variable of crash counts was observed, which is given as alltime.

$$alltime = 3 years * 365 days$$

Where alltime is the temporal exposure of the crash counts.

3.3 Data Screening, Reduction, and Aggregation

Only crashes that happened on mainline of roadway were extracted from the database for study. Those related with ramps and interchanges were removed to eliminate the possible influence of facilities other than mainline roadway. The crash data were reduced by weather and lighting conditions. The crashes happened during inclement weather were sorted out in order to eliminate the influence of weather. As a result, 330 crashes were removed from the data set. Those inclement weather conditions included cold and ice, fog, rain, snow, national disaster, sleet, and missing records. Table 3.2 gives the detailed numbers of crashes in each class and the corresponding percentiles. The crashes happened during the time period of 8:00:00 pm through 7:00:00 am in the

morning of the next day were excluded to eliminate the influence of unfavorable lighting condition. Thus, 329 more crashes were removed from the data. After these two screenings, 1018 out of the total of 1677 crashes were kept for analysis. (please refer to table 3.3 for the crash distribution by hour of the day) Finally, the remaining crashes were aggregated to hourly counts by day of week over the three studied years. The hourly counts means the number of crashes occurred during each hour.

Table 3.2 Crash Distribution By Weather

Weather	Number	Percentage
Clear	1030	61.42%
Cool	79	4.71%
Cloudy	213	12.70%
Warm	6	0.36%
Hot/Humid	19	1.13%
Cold/Ice	29	1.73%
Fog	13	0.78%
Rain	244	14.55%
Snow	11	0.66%
National Disaster	1	0.06%
Sleet	2	0.12%
Missing	30	1.79%
Summary	1677	100.00%
Selected	1347	80.32%
Rejected	330	19.68%

Table 3.3 Crash Distribution By Hour of the Day

Time Period	Number	Percentage
0:00:00-6:59:00	219	16.26%
7:00:00-19:59:59	1018	75.58%
20:00:00-24:00:00	110	8.17%
Total	1347	100.00%

As mentioned above, raw traffic data including volume, occupancy, and speed were collected and sorted by hour of the day, day of the week over the studied three years. In order to make the counts of crashes statistical significant, the crash data were aggregated in three years instead of one year. So were the traffic data. Several road segments were found to have invalid traffic data because of the malfunction or breakdown of detectors. As a result, four sets of data were eliminated in the further analysis: E64-02 westbound, W64-01 westbound, W64-04 eastbound and W64-08 westbound. As a result of the data reduction, 936 records were obtained for analysis. The data consisted of the following information: location, weekday, time, hourly crash count, traffic volume, occupancy, speed, and standard deviation of speed.

3.4 Variables in the Models

3.4.1 Variable Selection

The dependent variable used in the link function in this study was the number of crashes. The crash rate was not considered because stochastic regression models only deal with count numbers.

Table 3.4 Frequencies of Different Number of Crashes

C rash Count	0	1	2	3	4	5	6	7	8	9	10	13	15	16	17
No. of Obs.	564	214	64	26	27	11	10	5	2	4	2	1	2	1	2
Freq. (%)	60.3	22.9	6.8	2.8	2.9	1.2	1.1	0.5	0.2	0.4	0.2	0.1	0.2	0.1	0.2

Independent variables include traffic volume, speed, standard deviation of speed, exposure, and occupancy. Also, temporal exposure was used as a constant in the link function. While as many as four traffic variables were included in this study, the only

selected geometric variable was curvature. The curvatures changed drastically within the selected roadway (please refer to table 3.5). Therefore, it is necessary to include the curvature in the models to examine its influence over the occurrence of crashes. Other geometric variables were also showed to have significant influences over the occurrence of crashes, such as lane width, shoulder width, number of lanes, and grade. However, in this particular case, those geometric characteristics are relatively consistent. It is reasonable to assume that section of interstate 64 has uniform lane width and shoulder width. Also, that roadway segment selected is located in relatively level terrain. The influence of those factors is consistent over the whole roadway studied. Those factors can be considered as fixed environmental conditions to which the models are constrained. Therefore, exclusion of lane width, shoulder width, and grade would not hurt the model performance.

Table 3.5 Descriptive Statistics For Independent Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
Volume (veh./hr./lane)	1292.6	433.8	250.6	2245.2
Speed (mile/hr.)	58.1	5.6	31.2	65.0
Standard deviation of speed ((mile/hr)^2)	0.901	1.243	0.006	7.792
Occupancy	7.2%	3.9%	0.5%	23.4%
Exposure (mile*day)	2.47	1.58	0.95	8.05
Curvature (degree/mile)	43.8	21.7	6.9	78.5

3.4.2 Examination of Variables

The occupancy is directly related with the traffic volume and space mean speed as indicated above. The speed obtained from STL is the time mean speed. However, Subramanyan [34] showed that the time mean speed and space mean speed were related by a linear relationship with very little difference between the two for the basic freeway segments in his study. Thus, the occupancy is also related closely with the traffic volume and speed data from STL. Two sets of independent variables would be included in the models. One set of models include occupancy, standard deviation of speed, and curvature while the other set include traffic volume, speed, standard deviation of speed, and curvature.

CHAPTER 4: STOCHASTIC REGRESSION MODELS

When researchers began to realize that the characteristics of crashes are discrete, sporadic, and random, stochastic regression models were applied to describe the occurrence of crashes instead of deterministic models. The most widely applied stochastic regression models consist of Poisson and negative binomial regression models. Zero inflated Poisson and negative binomial regression models are modifications of Poisson and negative binomial regression models. They were proposed to handle excess zeros in the data set. Examples of zero inflated data include:

- Number of people infected by a certain disease per household
- Number of delinquency of sixty days or more on a credit account
- Number of crashes occurred on a road within a specific time period

The zero inflated models are two-regime models. In the first stage, a binomial phenomenon is assumed. The probability of no event occurring is p, while the probability of non-zero events is 1-p. Then, if any event does occur, conventional stochastic regression models such as Poisson or negative binomial are used to describe the probability of any particular number of events. The previous studies proved the effectiveness of zero inflated regression models.

4.1 Model Forms

Poisson and negative binomial regression models belong to the generalized linear model, while zero inflate Poisson and zero inflated negative binomial regression models are their extensions. The counts of vehicle crashes are independently distributed with mean, while the mean is linearly related with a set of covariates. In the case of this study,

the covariates include volume, speed, occupancy, speed variance, exposure, and curvature.

4.1.1 Poisson Regression Model

The probability distribution for a Poisson random variable is given by

$$prob[Y = y_i \mid t_i] = p(y_i \mid t_i) = \frac{e^{-(t_i\lambda_i)}(t_i\lambda_i)^{y_i}}{y_i!}, y_i = 0,1,...$$

Where y_i is the number of crashes of the i[th] observation

t_i is the time interval

λ_i is the mean number of crashes

Let t_i be the unit time interval,

$$E[y_i \mid x_i, t_i = 1] = Var[y_i \mid x_i, t_i = 1] = \lambda_i = e^{\beta' x_i}$$

Where β' is the coefficients matrix

x_i is the covariate matrix for the i[th] observation.

When an exposure is applied,

$$E[y_i \mid x_i, t_i = 1] = Var[y_i \mid x_i, t_i = 1] = \lambda_i = v_i e^{\beta' x_i}$$

Where v_i is the exposure for the i[th] observation.

Poisson regression model assumes that the variance of the data is equal to the mean, while in many applications, count data were found to display extra variation or over-dispersion relative to a Poisson model. Over-dispersion (the ratio of variance over mean is greater than 1) and under-dispersion (the ratio of variance over mean is less than 1) exist when the variance of the data was greater or less than that the Poisson models

indicate. Apart from the restriction of the Poisson distribution itself, heterogeneity of sub-population and large percentile zero-occurrence of crashes may also attributed to the over- and under-dispersion.

4.1.2 Negative Binomial Regression Model

Some density functions Π(λ) have been used to cope with the restrictive equality of the mean and variance for the Poisson distribution, such as

$$p(y,\lambda) \rightarrow p(y,\lambda)\Pi(\lambda)$$

When Π(λ) is Gamma distribution, the mixture density becomes negative binomial distribution.

$$p(Y=y_i)=\frac{\Gamma(\frac{1}{a}+y_i)}{\Gamma(\frac{1}{a})y_i!}(\frac{1}{1+a\ \mu_i})^{\frac{1}{a}}(1-\frac{1}{1+a\mu_i})^{y_i}$$

Where:

α is the parameter of negative binomial model;

$$\mu_i = E(Y_i) = \lambda_i = v_i e^{\beta' x_i}$$

$$Var(Y_i) = \mu_i(1+\alpha\mu_i)$$

Other symbols have the same meaning as those in the Poisson regression model.

The ratio of variance over expectation of the negative binomial model is greater than 1.

$$\frac{Var(y_i)}{E(y_i)} = 1+\alpha E(y_i)$$

4.1.3 Zero Inflated Poisson and Negative Binomial Regression Models

Zero inflated regression models are two regime models. In the first stage, according to Greene, "The idea underlying the models is that binomial probability model governs the binary outcome of whether a count number is zero or positive number"[5]. In this study, the model of this stage is defined as the inflated model. Then, the positive part of the distribution is described by Poisson, negative binomial or other stochastic distributions. Similarly, the model of this stage is defined as the base model. Essentially, zero inflated models are mixture models. Mullahy[2] proposed the hurdle model. Miao[3] applied this form of model in modeling the occurrence of crashes.

$$\Pr ob[y_i = 0] = \rho$$

$$\Pr ob[y_i = k] = \left[\frac{1-\rho}{1-e^{-\lambda_i}}\right]\frac{e^{-\lambda_i}\lambda_i^k}{k!}, k = 1,2,\cdots$$

Where ρ is a parameter between 0 and 1 to indicate the probability of zero events in the binary process.

Another model form proposed[5] is

$$\Pr ob[y_i = 0] = \psi + (1-\psi) f(0)$$

$$\Pr ob[y_i = j] = (1-\psi) f(j), j = 1,2,\cdots$$

Where ψ is a parameter between 0 and 1. Greene[5] indicated that a single parameter does not reflect the effects of covariates in ψ. Lambert[4] proposed the following zero inflated Poisson model, which was applied in this study:

---

[2] Sited from Greene

$$Y_i \sim 0 \qquad with \quad probability \quad p_0$$
$$\sim Poission \quad (\lambda_i) \qquad with \quad probability \quad 1-p_0,$$
$$So \quad that$$
$$Y_i = 0 \qquad with \quad probability \quad p_0 + (1-p_0)e^{-\lambda_i}$$
$$= k \qquad with \quad probability (1-p_0)e^{-\lambda_i}\lambda_i^k / k!,$$
$$k = 1,2,\cdots$$

Where

p_0 could be represented by probability model incorporating the effects of covariates, such as logit or probit model. When using the logit model,

$$p_0 = \frac{e^{r^{'}w_i}}{1+e^{r^{'}w_i}}$$

$r^{'}$ is the coefficients matrix and w_i is the ith covariate. It is possible that $r^{'}w_i = \tau\beta^{'}x_i$. Thus,

$$p_0 = \frac{\lambda_i^{\tau}}{1+\lambda_i^{\tau}}.$$

In this study, $r^{'}w_i \neq \tau\beta^{'}x_i$, which means that different variables are included in the inflated and base models.

When the probit model is applied, $p_0 = \Phi(r^{'}w_i)$. Φ is the cumulative normal distribution function.

The ratio of variance over expectation is also greater than 1.

$$E(y_i) = (1-q_0) * \lambda_i$$

$$Var(y_i) = \lambda_i(1-q_0)(1+\lambda_i * q_0)$$

$$\frac{Var(y_i)}{E(y_i)} = 1 + \lambda_i * q_0$$

Zero inflated negative binomial regression model is also a dual regime model. The only difference is to use negative binomial distribution to count for the positive counts of the occurrence of events.

$$Y_i \sim 0 \qquad with \quad probability \quad p_0$$

$$\sim Negative \quad Binomial \quad (\lambda_i) \qquad with \quad probability \quad 1 - p_0,$$

So that

$$Y_i = 0 \qquad with \quad probability \quad p_0 + (\frac{1}{1+\alpha * \lambda_i})^{\frac{1}{\alpha}}$$

$$= k \qquad with \quad probability (1-p_0) * (\frac{\Gamma(\frac{1}{a} + y_i)}{\Gamma(\frac{1}{a}) y_i!} (\frac{1}{1+a\ \lambda_i})^{\frac{1}{a}} (1 - \frac{1}{1+a\ \lambda_i})^{y_i}),$$

$$k = 1,2,\cdots$$

$$E(y_i) = (1-q_0)\lambda_i$$

$$Var(y_i) = (1-q_0)\lambda_i(1+(q_0+\alpha)\lambda_i)$$

$$\frac{Var(y_i)}{E(y_i)} = 1 + \left[\frac{q_0+\alpha}{1-q_0}\right] * E\ (y_i)$$

Obviously, the ratio of variance over expectation is greater than 1 for the ZINB model.

4.2 Model Estimation Technique

Maximum likelihood estimation method has been widely used in estimating Poisson, negative binomial, and zero inflated regression models.[4,5,6] Gan [6] proved the uniqueness and consistence of maximum likelihood estimators for zero inflated models.

According to Hayter[35],"If a data set consists of observations $x_1,\ldots,x_n$ from a probability distribution $f(x,\theta)$ depending on one unknown parameter θ, the maximum likelihood estimate $\hat{\theta}$ of the parameter is found by maximizing the likelihood function.

$$L(x_1,...,x_n,\theta) = f(x_1,\theta)\times...\times f(x_n,\theta).\text{"}$$

Similarly, when the parameters need to be estimated are more than one, the likelihood function is

$$L(x_1,...,x_n,\theta_1,...\theta_k) = f(x_1,\theta_1,...\theta_k)\times...\times f(x_n,\theta_1,...\theta_k).$$

k is the total number of parameters. k is 1 for Poisson distribution and 2 for negative binomial distribution. The detailed likelihood functions have been skipped. Interested readers can refer to Greene for further details. A commercial statistics software named STATA was used in this study.

4.3 Model Selection Criteria

4.3.1 Log Likelihood Value

According to the definition of maximum likelihood estimation method, the estimated parameters are the best when the maximum likelihood is obtained. It is also sustained for the maximum log likelihood. Since the data remain the same, it is comparable among different models.

4.3.2 Akaike Information Criteria

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

AIC = -2 Log L+2k

AICC = -2 Log L+2(k+1)n/(n-k-2).

Where Log L is the log likelihood;

K is the number of estimated parameters;

n is the number of observations.

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

AICC is better than AIC by incorporating the numbers of both estimated parameters and observations. In the case of this study, the sample of observations would keep the same. Thus, AIC and AICC should be consistent in this context. Nested models are constructed such that a simpler model can be obtained from a more complex model by eliminating one or more parameters from the more complex model. If not related in this way, models are not nested. In this study, both nested and non-nested would be modeled, selected, and tested.

4.3.3 Vuong Test Statistic

Voung test statistic (V) was proposed for non-nested models by Vuong[3] to compare the fitness of zero inflated Poisson model (or zero inflated negative binomial)

[3] referred from Greene

versus Poisson model (or negative binomial model).

$$V = \frac{\sqrt{N}\ \overline{m}}{S_m}, \qquad where\ \ m_i = \log\left[\frac{f_1(y_i)}{f_2(y_i)}\right]$$

N : *number of observations*

$\overline{m}$: *mean of* m_i

S_m : *standard deviation of* m_i

f_1 *and* f_2 *are two competing probability models*

V has a standard normal distribution. There are three possible outcomes:

- The absolute value of V is less than a threshold value such as 1.96 for 0.95 confidence level, then neither model is preferred by the test result.
- V is a large positive value, then model 1 is preferred.
- V is a large negative value, then model 2 is preferred.

4.4 Model Testing Technique

To test how well the models perform, the Error Rates (ER) were computed for selected models. ER was defined as the sum of absolute values of difference between the observed and estimated relative frequencies over the observed relative frequency. As Miaou [3] has shown:

$$r_k = \frac{\left| f_k - \hat{f_k} \right|}{f_k} \qquad k = 0,\ \ 1,\ \ \ldots\ \ n$$

Where:

r_k is the error rate for k occurrence of events;

f_k is the percentile of observations with k occurrence of events among the total data set;

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

$\hat{f}_k = \sum_i \hat{p}(y_i = k)/n$,

Where $\hat{p}(y_i = k)$ is the estimated probability of k occurrence of events under the prevailed covariate values and estimated parameters.

CHAPTER 5: RESULTS

Four kinds of stochastic regression models were estimated using the obtained data. They included Poisson, negative binomial, zero inflated Poisson, and zero inflated negative binomial regression models. The statistical software, STATA, contains the estimation function for stochastic models including the zero inflated models. Currently, it is the only commercial statistical program which deals with zero inflated regression models. It was used in this study

5.1 Model Estimation

5.1.1 Model Selection Criteria

The following criteria were used for the model selection and measurement of model performance:

- Reasonable algebraic signs of independent variables;
- Good p-value of each independent variable;
- Relatively low log likelihood value of the estimated models;
- Vuong values used to compare the zero inflated and the corresponding stochastic regression models were greater than 1.96 for the selection of zero inflated models;
- Relatively low AIC value when the nested and full models were compared; and
- Concise model form.

5.1.2 Two Sets of Zero Inflated Models

Both the logit and probit probability models were used to estimate the probability of zero events in the binary process. These models act as inflated models in the zero inflated regression models. Their performances were consistent. Independent variables such as volume, standard deviation, occupancy, and exposure were included in the logit

or probit models to incorporate the effects of those variables. It depended on the model performance to decide which variables to include or exclude.

5.1.3 Variable Selection

5.1.3.1 Two Sets of Independent Variables

As mentioned in chapter 3, two sets of models were tested. The first set of models included volume, speed, standard deviation of speed, curvature, and exposure as the independent variables. The second set of models included occupancy, standard deviation of speed, curvature, and exposure. The occupancy was separated from the volume and speed in the models. The reason for this separation was that the occupancy is directly related with the volume and speed. Therefore, the redundancy was avoided by considering the occupancy and the volume and speed in two sets of models.

5.1.3.2 Curvature

The sign of coefficients of curvature were shown to be negative for both sets of estimated models, which was not reasonable. Therefore, the curvature was excluded and new sets of models were developed. The new sets of models had the same structure as the models with the curvature. The models including the curvature were regarded as full models and the models without the curvature were nested models relative to the full models. it was found out that the two kinds of models generated close means of crash counts. For example, the means of the full and nested Poisson regression models are $mean_i^{full} = \lambda_i^{full} = 1095 * e^{(0.0009099*volume - 0.075077*speed + 0.0777812*stdsp - 0.006*curve + 0.19749*\exp ose - 4.592716)}$ and $mean_i^{nested} = \lambda_i^{nested} = 1095 * e^{(0.0009003*volume - 0.0802691*speed + 0.0755886*stdsp + 0.2026498*\exp ose - 4.536242)}$. Applying these two models to the whole data set, it was found that the average absolute

difference between these two means is 0.089227, which generates little difference in the probabilities of various crashes. Thus, the difference can be ignored.

The corresponding AIC and AICC values were compared among the nested and full models to see which set of models had better performance. The following tables (Table 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, and 5.8) show the performance of the two kinds of models. Each one of the four stochastic regression models was presented for illustration. The inflated models of these presented are logit models. During the modeling process, all models that can be selected showed that the curvature had the negative algebraic sign.

Table 5.1 Full Poisson Regression Model (P)

```
Poisson regression     Log likelihood = -1111.8009

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .0009099   .0000887    10.252   0.000     .0007359    .0010838
       speed |  -.0750077   .0051985   -14.429   0.000    -.0851965   -.0648189
       stdsp |   .0777812   .0245785     3.165   0.002     .0296082    .1259541
       curve |  -.0006000   .0027163    -3.558   0.000    -.0149892   -.0043414
      expose |   .1974899   .0113324    10.831   0.000     .1005298    .1449519
       _cons |  -4.592716    .384622   -11.941   0.000    -5.346562   -3.838871
     alltime |  (exposure)
------------------------------------------------------------------------------
```

Table 5.2 Nested Poisson Regression Model (P')

```
Poisson regression     Log likelihood = -1118.216

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .0009003   .0000898    10.028   0.000     .0007244    .0010763
       speed |  -.0802691   .0050662   -15.844   0.000    -.0901986   -.0703395
       stdsp |   .0755886   .0245365     3.081   0.002      .027498    .1236792
      expose |   .2026498    .011408    11.040   0.000     .1035884    .1483071
       _cons |  -4.536242   .3912108   -11.595   0.000    -5.303001   -3.769483
     alltime |  (exposure)
------------------------------------------------------------------------------
```

Table 5.3 Full Negative Binomial Regression Model (NB)

```
Negative binomial regression      Log likelihood = -1035.5609
```

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
volume	.0010988	.0001271	8.642	0.000	.0008496	.001348
speed	-.0838335	.0101925	-8.225	0.000	-.1038104	-.0638566
stdsp	.0527957	.0446464	1.183	0.237	-.0347096	.140301
curve	-.0063011	.0038063	-2.664	0.008	-.0175988	-.0026782
expose	.2099129	.0186026	7.013	0.000	.0940013	.1669221
_cons	-4.367009	.6634074	-6.583	0.000	-5.667263	-3.066754
alltime	(exposure)					
/lnalpha	-.3735875	.1522498			-.6719918	-.0751833
alpha	.6882607	.1047876	6.568	0.000	.5106904	.9275734

Table 5.4 Nested Negative Binomial Regression Model (NB')

Negative binomial regression Log likelihood = -1039.1043

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
volume	.0010614	.0001267	8.378	0.000	.0008131	.0013097
speed	-.0932521	.0096841	-9.629	0.000	-.1122326	-.0742715
stdsp	.0414902	.0446943	0.928	0.353	-.0461091	.1290895
expose	.2136536	.0186723	7.111	0.000	.0961896	.1693837
_cons	-4.037511	.6569174	-6.146	0.000	-5.325046	-2.749977
alltime	(exposure)					
/lnalpha	-.3476521	.1507794			-.6431742	-.0521299
alpha	.7063446	.1065022	6.632	0.000	.5256214	.9492055

Table 5.5 Full Zero Inflated Poisson Regression Model (ZIP)

Zero-inflated poisson regression Log likelihood = -1089.367

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
volume	.0007873	.0001027	7.662	0.000	.0005859	.0009886
speed	-.0670401	.0055424	-12.096	0.000	-.077903	-.0561771
stdsp	.0635775	.0250987	2.533	0.011	.0143849	.1127701
curve	-.0065580	.0029037	-3.634	0.000	-.016243	-.0048607
expose	.1837378	.0126456	9.030	0.000	.0894089	.1389786

```
    _cons |   -4.518031   .4047575    -11.162   0.000      -5.311342   -3.724721
 alltime | (exposure)
---------+--------------------------------------------------------------------
inflate  |
    _cons |   -1.053938    .206294     -5.109   0.000      -1.458267   -.6496093
------------------------------------------------------------------------------
Vuong Test of Zip vs. Poisson:         2.337   Prob > Z    0.990
```

Table 5.6 Nested Zero Inflated Poisson Regression Model (ZIP')

```
Zero-inflated poisson regression  Log likelihood  = -1096.042

------------------------------------------------------------------------------
   count |       Coef.   Std. Err.       z     P>|z|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
  volume |     .000759   .0001041      7.288   0.000       .0005549    .0009631
   speed |   -.0728084    .005395    -13.496   0.000      -.0833823   -.0622345
   stdsp |    .0594375   .0251423      2.364   0.018       .0101594    .1087155
  expose |    .1857808   .0127957      9.024   0.000       .0903845    .1405426
   _cons |   -4.404276   .4130305    -10.663   0.000      -5.213801   -3.594751
 alltime | (exposure)
---------+--------------------------------------------------------------------
inflate  |
   _cons |   -.9906476   .2002664     -4.947   0.000      -1.383163   -.5981325
------------------------------------------------------------------------------
Vuong Test of Zip vs. Poisson:         2.381   Prob > Z    0.991
```

Table 5.7 Full Zero Inflated Negative Binomial Regression Model (ZINB)

```
Zero-inflated negative binomial regression        Log likelihood  = -1038.618
------------------------------------------------------------------------------
   count |       Coef.   Std. Err.       z     P>|z|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
count    |
  volume |     .001396   .0001358     10.282   0.000       .0011299    .0016621
   speed |   -.0831402   .0106012     -7.843   0.000      -.1039182   -.0623623
   stdsp |    .0557153   .0467788      1.191   0.234      -.0359695    .1474001
   curve |   -.0053439   .0038893     -2.211   0.027      -.0162211   -.0009754
  expose |    .2429839   .0193619      7.800   0.000       .1130669    .1889642
   _cons |   -4.959753   .6885356     -7.203   0.000      -6.309257   -3.610248
 alltime | (exposure)
---------+--------------------------------------------------------------------
inflate  |
   _cons |   -98.38801          0          .       .      -98.38801   -98.38801
---------+--------------------------------------------------------------------
/lnalpha |   -.3205518   .1496446     -2.142   0.032      -.6138498   -.0272538
---------+--------------------------------------------------------------------
   alpha |    .7257485   .1086043                           .5412631    .9731143
------------------------------------------------------------------------------
Vuong Test of Zinb vs. Neg. Bin: Std. Normal -1.241       Prob > Z     0.1072
```

Table 5.8 Nested Zero Inflated Negative Binomial Regression Model (ZINB')

```
Zero-inflated negative binomial regression        Log likelihood  = -1041.561

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .0013248   .0001342     9.875   0.000     .0010619    .0015877
       speed |  -.0958184   .0100809    -9.505   0.000    -.1155765   -.0760602
       stdsp |   .0310062   .0465437     0.666   0.505    -.0602177    .1222302
      expose |   .2390355    .019308     7.694   0.000     .1107186    .1864045
       _cons |  -4.341601   .6785342    -6.398   0.000    -5.671503   -3.011698
     alltime |  (exposure)
-------------+----------------------------------------------------------------
inflate      |
       _cons |  -118.2495          0        .       .      -118.2495   -118.2495
-------------+----------------------------------------------------------------
    /lnalpha |  -.3035684   .1481558    -2.049   0.040    -.5939485   -.0131883
-------------+----------------------------------------------------------------
       alpha |   .7381794   .1093656                       .5521428    .9868983
------------------------------------------------------------------------------
Vuong Test of Zinb vs. Neg. Bin: Std. Normal  -1.153       Prob > Z      0.1244
```

Table 5.9 Comparison of Full and Nested Models

Models	P	P'	NB	NB'	ZIP	ZIP'	ZINB	ZINB'
Log likelihood	-1112	-1118	-1036	-1039	-1089	-1096	-1039	-1042
K	6	5	7	6	7	6	8	7
N	936	936	936	936	936	936	936	936
AIC	2236	2246	2085	2090	2193	2204	2093	2097
AICC	2238	2249	2087	2092	2195	2206	2095	2099
ΔAIC	10		5		11		4	
ΔAICC	11		5		11		4	

Note:

1. K is the number of parameters estimated
2. N is the total number of observations
3. ΔAIC is the difference of AIC values between the full and nested models
4. ΔAICC is the difference of AICC values between the full and nested models

Table 5.9 shows that the nested and full models have very close AIC and AICC values. The differences of AIC and AICC between the nested and full models are negligible considering the fact that the AIC and AICC values are at the magnitude of 2000. The full models are therefore not any better than the nested models according to the AIC or AICC values. Since the truck traffic on the selected section of I-64 is light, also curvature has much more influence on crashes involving trucks than those involving

passenger cars, it is very reasonable to discard the full models and select the nested models for further examination.

5.1.3.3 Speed and Standard Deviation of Speed

By examining the models including volume, speed, standard deviation of speed, and exposure, some general trends were found for all the estimated models. Poisson and ZIP have good performance. They have small p values for independent variables, which are less than 0.05. Vuong values for selected ZIP are greater than 1.96. However, negative binomial and ZINB regression models do not have such good performance. Not all the p values of independent variables are less than 0.05. The above nested negative binomial model (Table 5.4) indicates that the p value of standard deviation of speed is 0.353. The above zero inflated negative binomial model (Table 5.8) also shows that the p value of standard deviation of speed is 0.505. The large p values indicate the insignificance of the variables in the estimated models.

Then, the models including occupancy, standard deviation of speed, and expose were also examined. It was found that these models present good results. Table 5.10 and 5.11 show the examples of negative binomial and zero inflated negative binomial regression models. All the p values in table 5.10 are 0.000, which indicates that all the independent variables in the model are significant including the constant.

Table 5.10 An Example of Negative Binomial Regression Model

Negative binomial regression Log likelihood = -1094.3011

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
occu	12.67423	1.335323	9.492	0.000	10.05705	15.29142
stdsp	.2872677	.0407773	7.045	0.000	.2073457	.3671896
expose	.172161	.0203826	5.249	0.000	.0670493	.1469478

_cons	-9.000146	.1456089	-61.810	0.000	-9.285535	-8.714758
lalltime	(exposure)					
/lnalpha	.1400335	.1229362			-.100917	.3809841
alpha	1.150312	.141415	8.134	0.000	.9040081	1.463724

Table 5.11 shows that all the p values for the independent variables except that of alpha are very good. Also, the vuong value is 2.675, which is larger than 1.96. This indicates that this ZINB model is acceptable.

Table 5.11 An Example of ZINB Regression Model

Zero-inflated negative binomial regression Log likelihood = -1094.3011

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
occu	15.45177	1.374715	11.240	0.000	12.75738	18.14616
stdsp	.1963106	.0397685	4.936	0.000	.1183657	.2742555
expose	.103988	.0199116	3.246	0.001	.0256032	.1036553
_cons	-8.792633	.1478045	-59.488	0.000	-9.082325	-8.502942
lalltime	(exposure)					
inflate						
occu	26.18221	6.120585	4.278	0.000	14.18608	38.17834
stdsp	-3.941304	1.162919	-3.389	0.001	-6.220583	-1.662024
expose	-1.186127	.2363584	-3.119	0.002	-1.200437	-.2739288
/lnalpha	-.1653139	.1475507	-1.120	0.263	-.4545079	.12388
alpha	.8476276	.125068			.6347602	1.13188

Vuong Test of Zinb vs. Neg. Bin: Std. Normal 2.675 Prob > Z 0.9963

While standard deviation of speed does not have good p values in the models including volume and speed, it is shown to be significant in the models including occupancy. It might be other variables instead of standard deviation of speed that caused the poor performance of certain models. More trials indicated that including speed and standard deviation of speed in the same model did not present good result for negative binomial and zero inflated negative binomial regression models. Manual stepwise

variable selection was performed to include only significant variables in the models. Previous studies indicate that the independent variables selected for study all have significant influence over the occurrence of crashes. Therefore, those variables were kept as many as possible in order to fully describe the crash. Among the four independent variables, each one was removed from the model to test the result. It was found that volume and exposure are absolutely necessary in order to have acceptable p values and reasonable log likelihood values. Also, they are needed for large vuong values in the case of zero inflated models. For the inclusion of the speed or standard deviation of speed, they did not generate much difference with respect to Poisson and negative binomial regression models with similar model structure. The p values of independent variables are all good and the log likelihood values are very close. However, by comparing the competing zero inflated regression models including ZIP and ZINB, it was found that both of the two kinds of models could have good results if different structures of inflated models were used. Thus, it was decided that the speed and standard deviation of speed would be included in different sets of models. Their performance would be evaluated and compared in the model testing section.

5.1.4 Transformations

Various transformations were tried for independent variables, such as log, square, and sin terms of the independent variables. Transformations were tried throughout the modeling process. Several general conclusions could be drawn about the effect of transformations in this study. First, transformations did not change the signs of independent variables. Second, transformations could change p value of the independent variable. Third, transformations did not have significant influence over the log likelihood

value. Transformations did not improve the model performance significantly. Therefore, no transformation was used in order to keep the model forms concise.

5.1.5 Results

Finally, three sets of models were developed with respect to the independent variables included in the models. The inflated part for zero inflated models includes logit and probit models.

Table 5.12 Different Sets of Models

Model	Variables In the Base Model	Variables In the Inflated Model
1	Occu, Stdsp, Expose	Selected from Occu, Stdsp, Expose
2	Volume, Speed, Expose	Selected from Volume, Speed, Expose
3	Volume, Stdsp, Expose	Selected from Volume, Stdsp, Expose

Note: Poisson and negative binomial regression models do not have inflated models.

5.1.5.1 The First Set of Models

Table 5.13 The Final Result(I): Poisson Regression Model

```
Poisson regression                      Log likelihood = -1231.0627

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        occu |    13.28704   .7075363    18.779   0.000     11.90029    14.67378
       stdsp |    .2804112   .0187751    14.935   0.000     .2436126    .3172098
      expose |    .18732895   .0110761    10.511   0.000      .094717    .1381344
       _cons |   -9.085271   .0969487   -93.712   0.000    -9.275287   -8.895256
  alltime    |  (exposure)
------------------------------------------------------------------------------
```

Table 5.13 gives the following actual model form[4]

[4] λ_i is the mean occurrence rate of crashes per 1095 days•years.

$$p(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, y_i = 0,1,...$$

$$\lambda_i = 1095 * e^{(-9.085271 + 13.28704 * occu + 0.2804112 * stdsp + 0.18732895 * \exp ose)}$$

Table 5.14 The Final Result(I): Negative Binomial Regression Model

```
Negative binomial regression                    Log likelihood = -1094.3011
------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        occu |   12.67423   1.335323     9.492   0.000     10.05705    15.29142
       stdsp |   .2872677   .0407773     7.045   0.000     .2073457    .3671896
      expose |   .1721607   .0203826     5.249   0.000     .0670493    .1469478
       _cons |  -9.000146   .1456089    -61.810   0.000    -9.285535   -8.714758
     alltime |  (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |   .1400335   .1229362                      -.100917    .3809841
-------------+----------------------------------------------------------------
       alpha |   1.150312    .141415     8.134   0.000     .9040081    1.463724
------------------------------------------------------------------------------
```

Table 5.14 gives the following model form:

$$p(Y = y_i) = \frac{\Gamma(\frac{1}{1.150312} + y_i)}{\Gamma(\frac{1}{1.150312})y_i!}(\frac{1}{1+1.150312\lambda_i})^{\frac{1}{1.150312}}(1-\frac{1}{1+1.150312\lambda_i})^{y_i}$$

$$\lambda_i = 1095 * e^{(-9.000146 + 12.67423 * occu + 0.2872677 * stdsp + 0.1721607 * \exp ose)}$$

Table 5.15 The Final Result(I): ZIP Regression Model (Logit)

```
Zero-inflated poisson regression                Log likelihood = -1163.077

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        occu |   11.40821   .8177328    13.951   0.000     9.805485    13.01094
       stdsp |   .2051468    .022978     8.928   0.000     .1601108    .2501827
      expose |   .1380964   .0135287     6.344   0.000     .0593117    .1123432
       _cons |  -8.223972   .1136425    -72.367   0.000    -8.446707   -8.001237
     alltime |  (exposure)
-------------+----------------------------------------------------------------
inflate      |
       stdsp |  -.165369   .0843004    -1.962   0.050    -.3305949   -.0001432
      expose |  -.1199754   .0278855    -2.674   0.007    -.1292198   -.0199107
------------------------------------------------------------------------------
Vuong Test of Zip vs. Poisson:          4.373   Prob > Z    1.000
```

Table 5.15 gives the following model form:

$$P(0) = \; p_0 + (1-p_0)e^{-\lambda_i}$$
$$P(y_i) = (1-p_0)e^{-\lambda_i}\lambda_i^{y_i} / y_i!, \qquad y_i = 1,2,\cdots$$

$$p_0 = \frac{e^{(-0.165369*stdsp - 0.1199754*\exp ose)}}{1+e^{(-0.165369*stdsp - 0.1199754*\exp ose)}}$$

$$\lambda_i = 1095 * e^{(-8.223972 + 11.40821*occu + 0.2051468*stdsp + 0.1380964*\exp ose)}$$

Table 5.16 The Final Result(I): ZIP Regression Model (Probit)

```
Zero-inflated poisson regression                Log likelihood  = -1163.095

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       occu |     11.4021   .816986    13.956   0.000      9.80084     13.00337
      stdsp |    .2049714   .0228834     8.957   0.000      .1601206    .2498221
     expose |    .1378998   .0135159     6.341   0.000      .0592146    .1121959
      _cons |   -8.22176    .1132222   -72.616   0.000     -8.443671   -7.999849
 all_time   |  (exposure)
-------------+----------------------------------------------------------------
inflate      |
      stdsp |   -.1016934   .0492777    -2.064   0.039     -.1982759   -.0051109
     expose |   -.0740611   .0166252    -2.769   0.006     -.0786141   -.0134445
------------------------------------------------------------------------------
Vuong Test of Zip vs. Poisson:          4.364   Prob > Z    1.000
```

Table 5.16 gives the following model form:

$$P(0) = \; p_0 + (1-p_0)e^{-\lambda_i}$$
$$P(y_i) = (1-p_0)e^{-\lambda_i}\lambda^{y_i} / y_i!, \qquad y_i = 1,2,\cdots$$

$$p_0 = \Phi(-0.1016934 * stdsp - 0.0740611 * \exp ose)$$

Φ is the cumulative normal distribution function.

$$\lambda_i = 1095 * e^{(-8.22176+11.4021*occu+0.2049714*stdsp+0.1378998*\exp ose)}$$

Table 5.17 The Final Result(I): ZINB Regression Model (Logit)

Zero-inflated negative binomial regression Log likelihood = -1076.224

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
occu	15.45177	1.374715	11.240	0.000	12.75738	18.14616
stdsp	.1963106	.0397685	4.936	0.000	.1183657	.2742555
expose	.1039884	.0199116	3.246	0.001	.0256032	.1036553
_cons	-8.792633	.1478045	-59.488	0.000	-9.082325	-8.502942
alltime	(exposure)					
inflate						
occu	26.18221	6.120585	4.278	0.000	14.18608	38.17834
stdsp	-3.941304	1.162919	-3.389	0.001	-6.220583	-1.662024
expose	-1.186127	.2363584	-3.119	0.002	-1.200437	-.2739288
/lnalpha	-.1653139	.1475507	-1.120	0.263	-.4545079	.12388
alpha	.8476276	.125068			.6347602	1.13188

Vuong Test of Zinb vs. Neg. Bin: Std. Normal 2.675 Prob > Z 0.9963

Table 5.17 gives the following model form:

$$P(0) = p_0 + (1-p_0) * (\frac{1}{1+0.8476276*\lambda_i})^{\frac{1}{0.8476276}}$$

$$P(y_i) = (1-p_0) * (\frac{\Gamma(\frac{1}{0.8476276}+y_i)}{\Gamma(\frac{1}{0.8476276})y_i!}(\frac{1}{1+0.8476276*\lambda_i})^{\frac{1}{0.8476276}}(1-\frac{1}{1+0.8476276*\lambda_i})^{y_i}),$$

$y_i = 1,2,..$

$$p_0 = \frac{e^{(26.18221*occu-3.941304*stdsp-1.186127*\exp ose)}}{1+e^{(26.18221*occu-3.941304*stdsp-1.186127*\exp ose)}}$$

$$\lambda_i = 1095 * e^{(-8.792633+15.45177*occu+0.1963106*stdsp+0.1039884*\exp ose)}$$

Table 5.18 The Final Result(I): ZINB Regression Model (Probit)

```
Zero-inflated negative binomial regression      Log likelihood  = -1076.097

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        occu |   15.39846   1.376479    11.187   0.000     12.70061    18.09631
       stdsp |   .1978606   .0398325     4.967   0.000     .1197902    .2759309
      expose |   .1045362   .0199828     3.251   0.001     .0258042    .1041353
       _cons |  -8.795223   .1476709    -59.560   0.000    -9.084653   -8.505794
  alltime |  (exposure)
-------------+----------------------------------------------------------------
inflate      |
        occu |   14.91926    3.20143     4.660   0.000     8.644575    21.19395
       stdsp |  -2.302216    .668659    -3.443   0.001    -3.612763   -.9916683
      expose |  -.6840073   .1261036    -3.371   0.001    -.6722718   -.1779548
-------------+----------------------------------------------------------------
/lnalpha |  -.1558964   .1465879    -1.064   0.288    -.4432035    .1314106
-------------+----------------------------------------------------------------
       alpha |   .8556478   .1254276                      .6419766    1.140436
------------------------------------------------------------------------------
Vuong Test of Zinb vs. Neg. Bin: Std. Normal  2.756       Prob > Z     0.9971
```

Table 5.18 gives the following model form:

$$P(0) = p_0 + (1-p_0) * (\frac{1}{1+0.8556478*\lambda_i})^{\frac{1}{0.8556478}}$$

$$P(y_i) = (1-p_0) * (\frac{\Gamma(\frac{1}{0.8556478}+y_i)}{\Gamma(\frac{1}{0.8556478})y_i!}(\frac{1}{1+0.8556478*\lambda_i})^{\frac{1}{0.8556478}}(1-\frac{1}{1+0.8556478*\lambda_i})^{y_i}),$$

$y_i = 1,2,..$

$$p_0 = \Phi(14.91926 * occu - 2.302216 * stdsp - 0.6840073 * \exp ose))$$

Φ is the cumulative normal probability function

$$\lambda_i = 1095 * e^{(-8.795223+15.39846*occu+0.1978606*stdsp+0.1045362*\exp ose)}$$

5.1.5.2 The Second Set of Models

Table 5.19 The Final Result(II): Poisson Regression Model

```
Poisson regression                              Log likelihood = -1122.7057

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .0008843   .0000913     9.687   0.000     .0007054    .0010632
       speed |  -.0880675   .0043339   -20.320   0.000    -.0965619   -.0795732
      expose |   .1935204   .0113083    10.636   0.000     .0981097    .1424376
       _cons |  -3.960136   .3447918   -11.486   0.000    -4.635915   -3.284356
    alltime  |  (exposure)
```

Table 5.19 gives the following model form:

$$p(y_i) = \frac{e^{-\lambda_i}\lambda_i^{\,y_i}}{y_i!}, y_i = 0,1,...$$

$$\lambda_i = 1095 * e^{(-3.960136+0.0008843*volume-0.0880675*speed+0.1935204*\exp ose)}$$

Table 5.20 The Final Result(II): Negative Binomial Regression Model

```
Negative binomial regression                    Log likelihood =  -1039.535
------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .0010584   .0001272     8.322   0.000     .0008092    .0013077
       speed |   -.098854   .0076574   -12.910   0.000    -.1138622   -.0838457
      expose |   .2097919   .0185103     7.044   0.000      .094107    .1666659
       _cons |  -3.660617   .5195995    -7.045   0.000    -4.679014   -2.642221
    alltime  |  (exposure)
-------------+----------------------------------------------------------------
/lnalpha |  -.3398058   .1497743                          -.633358   -.0462536
-------------+----------------------------------------------------------------
       alpha |   .7119085   .1066256     6.677   0.000     .5308064    .9547998
------------------------------------------------------------------------------
```

Table 5.20 gives the following model form:

$$p(Y=y_i)=\frac{\Gamma(\frac{1}{0.7119085}+y_i)}{\Gamma(\frac{1}{0.7119085})y_i!}(\frac{1}{1+0.7119085*\lambda_i})^{\frac{1}{0.7119085}}(1-\frac{1}{1+0.7119085*\lambda_i})^{y_i}$$

$$\lambda_i=1095*e^{(-3.660617+0.0010584*volume-0.098854*speed+0.2097919*\exp ose)}$$

Table 5.21 The Final Result(II): ZIP Regression Model (Logit)

```
Zero-inflated poisson regression              Log likelihood  = -1078.957

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .0003361   .0001257     2.674   0.007     .0000898    .0005825
       speed |  -.0829075   .0052875   -15.680   0.000    -.0932708   -.0725442
      expose |   .1538476   .0150062     6.372   0.000     .0662053    .1250284
       _cons |  -2.943303   .4183755    -7.035   0.000    -3.763304   -2.123302
    altime   |  (exposure)
-------------+----------------------------------------------------------------
inflate      |
      volume |  -.0017374    .000311    -5.587   0.000    -.0023469   -.0011278
       speed |   .0376874   .0072403     5.205   0.000     .0234966    .0518781
      expose |  -.1810950   .0506755    -2.221   0.026    -.2118735    -.013229
------------------------------------------------------------------------------
Vuong Test of Zip vs. Poisson:          3.398    Prob > Z    1.000
```

Table 5.21 gives the following model form:

$$P(0)= \quad p_0+(1-p_0)e^{-\lambda_i}$$

$$P(y_i)=(1-p_0)e^{-\lambda_i}\lambda_i^{y_i}/y_i!, \qquad y_i=1,2,\cdots$$

$$p_0=\frac{e^{(-0.0017374*volume+0.0376874*speed-0.1810950*\exp ose)}}{1+e^{(-0.0017374*volume+0.0376874*speed-0.1810950*\exp ose))}}$$

$$\lambda_i=1095*e^{(-2.943303+0.0003361*volume-0.0829075*speed+0.1538476*\exp oose)}$$

Table 5.22 The Final Result(II): ZIP Regression Model (Probit)

```
Zero-inflated poisson regression              Log likelihood  =  -1078.55

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
count        |
      volume |    .000339   .0001249     2.713   0.007     .0000941    .0005838
       speed |  -.0827737   .0052447   -15.782   0.000    -.0930531   -.0724942
      expose |   .1528379   .0148719     6.387   0.000      .065841    .1241378
       _cons |  -2.950776   .4162002    -7.090   0.000    -3.766514   -2.135039
     alltime |  (exposure)
-------------+----------------------------------------------------------------
inflate      |
      volume |  -.0010551   .0001768    -5.968   0.000    -.0014016   -.0007085
       speed |   .0231252   .0043035     5.374   0.000     .0146906    .0315599
      expose |  -.115175     .029487    -2.428   0.015    -.1293749   -.0137881
------------------------------------------------------------------------------
Vuong Test of Zip vs. Poisson:         3.444   Prob > Z    1.000
```

Tabel 5.22 gives the following model form:

$$P(0) = \quad p_0 + (1 - p_0)e^{-\lambda_i}$$

$$P(y_i) = (1 - p_0)e^{-\lambda_i}\lambda_i^{y_i} / y_i!, \qquad y_i = 1,2,\cdots$$

$$p_0 = \Phi(-0.0010551 * volume + 0.0231252 * speed - 0.115175 * \exp ose)$$

Φ is the cumulative normal distribution function.

$$\lambda_i = 1095 * e^{(-2.950776 + 0.000339 * volume - 0.0827737 * speed + 0.1528379 * \exp oose)}$$

Table 5.23 The Final Result(II): ZINB Regression Model (Logit)

```
Zero-inflated negative binomial regression    Log likelihood  =  -1029.368

------------------------------------------------------------------------------
```

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
volume	.0011134	.0001274	8.741	0.000	.0008638	.0013631
speed	-.0956421	.007489	-12.771	0.000	-.1103204	-.0809639
expose	.1616817	.0189952	5.290	0.000	.0632559	.1377157
_cons	-3.696897	.5146413	-7.183	0.000	-4.705575	-2.688218
alltime	(exposure)					
inflate						
expose	-16.61638	3.828047	-2.698	0.007	-17.82999	-2.824319
_cons	16.4897	6.306136	2.615	0.009	4.129901	28.8495
/lnalpha	-.5411487	.163512	-3.310	0.001	-.8616264	-.2206709
alpha	.5820793	.095177			.4224744	.8019805
Vuong Test of Zinb vs. Neg. Bin: Std. Normal			2.262		Prob > Z	0.9882

Table 5.23 gives the following model form:

$$P(0) = p_0 + (1-p_0) * (\frac{1}{1+0.5820793 * \lambda_i})^{\frac{1}{0.5820793}}$$

$$P(y_i) = (1-p_0) * (\frac{\Gamma(\frac{1}{0.5820793} + y_i)}{\Gamma(\frac{1}{0.5820793}) y_i!} (\frac{1}{1+0.5820793 * \lambda_i})^{\frac{1}{0.5820793}} (1 - \frac{1}{1+0.5820793 * \lambda_i})^{y_i}),$$

$y_i = 1,2,..$

$$p_0 = \frac{e^{(16.4897 - 16.61638 * \exp ose)}}{1 + e^{(16.4897 - 16.61638 * \exp ose)}}$$

$$\lambda_i = 1095 * e^{(-3.696897 + 0.0011134 * volume - 0.0956421 * speed + 0.1616817 * \exp ose)}$$

Table 5.24 The Final Result(II): ZINB Regression Model (Probit)

Zero-inflated negative binomial regression Log likelihood = -1029.297

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
volume	.0011136	.0001274	8.742	0.000	.0008639	.0013633
speed	-.0956922	.0074892	-12.777	0.000	-.1103708	-.0810136
expose	.1620084	.0189685	5.308	0.000	.0635114	.1378664
_cons	-3.695898	.51492	-7.178	0.000	-4.705123	-2.686673
alltime	(exposure)					

inflate					
expose	-10.26771	2.308006	-2.765	0.006	-10.90504 -1.857816
_cons	10.19261	3.812776	2.673	0.008	2.719708 17.66552
/lnalpha	-.5399951	.1634059	-3.305	0.001	-.8602647 -.2197255
alpha	.5827511	.0952249			.4230501 .8027391

Vuong Test of Zinb vs. Neg. Bin: Std. Normal 2.278 Prob > Z 0.9886

Table 5.24 gives the following model form:

$$P(0) = p_0 + (1 - p_0) * (\frac{1}{1 + 0.5827511 * \lambda_i})^{\frac{1}{0.5827511}}$$

$$P(y_i) = (1 - p_0) * (\frac{\Gamma(\frac{1}{0.5827511} + y_i)}{\Gamma(\frac{1}{0.5827511}) y_i!} (\frac{1}{1 + 0.5827511 * \lambda_i})^{\frac{1}{0.5827511}} (1 - \frac{1}{1 + 0.5827511 * \lambda_i})^{y_i}),$$

$y_i = 1,2,..$

$p_0 = \Phi(10.19261 - 10.26771 * \exp ose)$

Φ is the cumulative normal distribution function.

$$\lambda_i = 1095 * e^{(-3.695898 + 0.0011136 * volume - 0.0956922 * speed + 0.1620084 * \exp ose)}$$

5.1.5.3 The Third Set of Models

Table 5.25 The Final Result(III): Poisson Regression Model

Poisson regression Log likelihood = -1236.6436

count	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
volume	.001373	.0000847	16.214	0.000	.0012071 .001539
stdsp	.2625704	.0187723	13.987	0.000	.2257774 .2993634
expose	.2786986	.011105	15.598	0.000	.1514469 .1949778
_cons	-10.10727	.1517684	-66.597	0.000	-10.40473 -9.809805
alltime	(exposure)				

Table 5.25 gives the following model form:

$$p(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, y_i = 0,1,...$$

$$\lambda_i = 1095 * e^{(-10.10727+0.001373*volume+0.2625704*stdsp+0.2786986*\exp ose)}$$

Table 5.26 The Final Result(III): Negative Binomial Regression Model

```
Negative binomial regression                    Log likelihood =  -1084.3266
------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .0014075   .0001332    10.566   0.000     .0011464    .0016686
       stdsp |   .3325017   .0412491     8.061   0.000     .2516549    .4133485
      expose |   .2770803   .0206218     8.351   0.000     .1317885    .2126246
       _cons |  -10.24142   .2315827    -44.224   0.000    -10.69532   -9.787528
     alltime |  (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |   .1026405   .1213168                      -.135136     .340417
-------------+----------------------------------------------------------------
       alpha |   1.108093   .1344302     8.243   0.000     .8735971    1.405534
------------------------------------------------------------------------------
Likelihood ratio test of alpha=0:     chi2(1) =   304.63   Prob > chi2 = 0.0000
```

Table 5.26 gives the following model form:

$$p(Y = y_i) = \frac{\Gamma(\frac{1}{1.108093} + y_i)}{\Gamma(\frac{1}{1.108093})y_i!}(\frac{1}{1+1.108093*\lambda_i})^{\frac{1}{1.108093}}(1-\frac{1}{1+1.108093*\lambda_i})^{y_i}$$

$$\lambda_i = 1095 * e^{(-10.24142+0.0014075*volume+0.3325017*stdsp+0.2770803*\exp ose)}$$

Table 5.27 The Final Result(III): ZIP Regression Model (Logit)

```
Zero-inflated poisson regression                Log likelihood  = -1170.054

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .0015089   .0000932    16.192   0.000     .0013263    .0016916
       stdsp |   .1644969   .0211808     7.766   0.000     .1229833    .2060105
      expose |   .1791929    .012802     8.699   0.000     .0862777    .1364605
       _cons |  -9.53743   .1730749    -55.106   0.000    -9.876651    -9.19821
     alltime |  (exposure)
-------------+----------------------------------------------------------------
inflate      |
```

volume	.0008557	.0002494	3.431	0.001	.0003669	.0013445
stdsp	-.7507441	.2212027	-3.394	0.001	-1.184293	-.3171948
expose	-.621399	.1169855	-3.301	0.001	-.6154894	-.1569148

Vuong Test of Zip vs. Poisson: 4.129 Prob > Z 1.000

Table 5.27 gives the following model form:

$$P(0) = p_0 + (1-p_0)e^{-\lambda_i}$$

$$P(y_i) = (1-p_0)e^{-\lambda_i}\lambda_i^{y_i} / y_i!, \qquad y_i = 1,2,\cdots$$

$$p_0 = \frac{e^{(0.0008557*volume - 0.7507441*stdsp - 0.621399*\exp ose)}}{1+e^{(0.0008557*volume - 0.7507441*stdsp - 0.621399*\exp ose)}}$$

$$\lambda_i = 1095 * e^{(-9.53743 + 0.0015089*volume + 0.1644969*stdsp + 0.1791929*\exp ose)}$$

Table 5.28 The Final Result(III): ZIP Regression Model (Probit)

Zero-inflated poisson regression Log likelihood = -1170.536

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
volume	.0015212	.0000909	16.728	0.000	.0013429	.0016994
stdsp	.1651558	.0211502	7.809	0.000	.1237021	.2066094
expose	.1770297	.01258	8.746	0.000	.0853684	.134681
_cons	-9.556837	.1707638	-55.965	0.000	-9.891528	-9.222146
alltime	(exposure)					
inflate						
volume	.0005421	.0001428	3.796	0.000	.0002622	.000822
stdsp	-.4487386	.1236118	-3.630	0.000	-.6910133	-.2064638
expose	-.401727	.0668604	-3.734	0.000	-.3807189	-.1186311

Vuong Test of Zip vs. Poisson: 4.102 Prob > Z 1.000

Table 5.28 gives the following model form:

$$P(0) = p_0 + (1-p_0)e^{-\lambda_i}$$

$$P(y_i) = (1-p_0)e^{-\lambda_i}\lambda_i^{y_i} / y_i!, \qquad y_i = 1,2,\cdots$$

$$p_0 = \Phi(0.0005421 * volume - 0.4487386 * stdsp - 0.401727 * \exp ose)$$

Φ is the cumulative normal distribution function.

$$\lambda_i = 1095 * e^{(-9.556837 + 0.0015212 * volume + 0.1651558 * stdsp + 0.1770297 * \exp ose)}$$

Table 5.29 The Final Result (III): ZINB Regression Model (Logit)

```
Zero-inflated negative binomial regression      Log likelihood  = -1072.463

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      volume |   .0014793   .0001328    11.138   0.000      .001219    .0017396
       stdsp |   .3262934   .0407186     8.013   0.000     .2464863    .4061004
      expose |   .2185748   .0209906     6.472   0.000     .0947042    .1769859
       _cons |  -10.07697   .2314543    -43.538   0.000    -10.53061   -9.623324
    alltime  | (exposure)
-------------+----------------------------------------------------------------
inflate      |
      expose |   -16.9573   4.067587    -2.591   0.010    -18.51136   -2.566709
       _cons |   16.93089   6.671905     2.538   0.011     3.854196    30.00758
-------------+----------------------------------------------------------------
/lnalpha     |  -.0912809   .1327536    -0.688   0.492    -.3514731    .1689114
-------------+----------------------------------------------------------------
       alpha |   .9127613   .1211723                      .7036508    1.184015
------------------------------------------------------------------------------
Vuong Test of Zinb vs. Neg. Bin: Std. Normal  2.475       Prob > Z     0.9933
```

Table 5.29 gives the following model form:

$$P(0) = p_0 + (1 - p_0) * (\frac{1}{1 + 0.9127613 * \lambda_i})^{\frac{1}{0.91276133}}$$

$$P(y_i) = (1 - p_0) * (\frac{\Gamma(\frac{1}{0.9127613} + y_i)}{\Gamma(\frac{1}{0.9127613}) y_i!} (\frac{1}{1 + 0.9127613 * \lambda_i})^{\frac{1}{0.9127613}} (1 - \frac{1}{1 + 0.9127613 * \lambda_i})^{y_i}),$$

$k = 1,2,..$

$$p_0 = \frac{e^{(16.93089 \ -16.9573 \ * \exp \ ose \ )}}{1 + e^{(16.93089 \ -16.9573 \ * \exp \ ose \ )}}$$

$$\lambda_i = 1095 * e^{(-10.07697 + 0.0014793 * volume + 0.3262934 * stdsp + 0.2185748 * \exp ose)}$$

Table 5.30 The Final Result(III): ZINB Regression Model (Probit)

Zero-inflated negative binomial regression Log likelihood = -1072.458

count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
count						
volume	.0014798	.0001328	11.142	0.000	.0012195	.0017401
stdsp	.3262291	.040751	8.005	0.000	.2463585	.4060997
expose	.2195514	.0209532	6.512	0.000	.0953845	.1775197
_cons	-10.08195	.2311406	-43.618	0.000	-10.53498	-9.628926
alltime	(exposure)					
inflate						
expose	-10.72279	2.445224	-2.725	0.006	-11.45681	-1.871706
_cons	10.70862	4.016218	2.666	0.008	2.836972	18.58026
/lnalpha	-.0881219	.1325587	-0.665	0.506	-.3479322	.1716883
alpha	.9156492	.1213772			.7061468	1.187308

Vuong Test of Zinb vs. Neg. Bin: Std. Normal 2.477 Prob > Z 0.9934

Table 5.30 gives the following model form:

$$P(0) = p_0 + (1 - p_0) * (\frac{1}{1 + 0.9156492 * \lambda_i})^{\frac{1}{0.9156492}}$$

$$P(y_i) = (1 - p_0) * (\frac{\Gamma(\frac{1}{0.9156492} + y_i)}{\Gamma(\frac{1}{0.9156492}) y_i!} (\frac{1}{1 + 0.9156492 * \lambda_i})^{\frac{1}{0.9156492}} (1 - \frac{1}{1 + 0.9156492 * \lambda_i})^{y_i}),$$

$k = 1,2,..$

$p_0 = \Phi(10.70862 - 10.72279 * \exp ose)$

Φ is the cumulative normal distribution function.

$$\lambda_i = 1095 * e^{(-10.08195 + 0.0014798 * volume + 0.3262291 * stdsp + 0.2195514 * \exp ose)}$$

5.2 Model Result Examination

In order to examine the performance of the estimated models, the following tasks were performed:

1. Examine the coefficients of variables in different models;

2. Examine the diagnostic statistics of models; and

3. Examine the sign of speed.

5.2.1 Examine the Coefficients of Variables In Different Models

For all the three sets of models, the four stochastic regression models have consistent algebraic sign for each variable in the base models (please refer to table 5.32, 5.33, and 5.34). For the inflated models of zero inflated models, the algebraic signs of variables are also consistent with the exception of volume. In table 5.21 and 5.22, the signs of volume are negative. While in table 5.27 and 5.28, the signs of volume are positive. The different model structure might have attributed to the different signs of the same variable. The inflated models in table 5.21 and 5.22 include volume, speed, and

exposure, while the inflated models in table 5.27 and 5.28 include volume, standard deviation of speed, and exposure. The signs of the exposure and standard deviation of speed remain negative consistently in the inflated models. The coefficients for each variable in different models have close values. The coefficients of corresponding variables in Poisson and NB models are very close to each other.

The logit and probit models used in the zero inflated models did not generate much difference. The corresponding coefficients in the base models are nearly identical. The two inflated models do not have the same structure, but probit and logit inflated models generate nearly identical results (table 5.31). The log likelihood and vuong statistics only changed slightly from using logit to probit as the inflated model. For example, for the ZIP models in table 5.32, the log likelihood changed from -1163.08 to -1163.1 and vuong changed from 4.373 to 4.364. These changes are so slight that they can be disregarded. Therefore, the logit and probit models can be applied interchangeably in this study.

Table 5.31 Comparison of the Logit and Probit Inflated Models

Δp_0	1st set of models		2nd set of models		3rd set of models	
	ZIP	ZINB	ZIP	ZINB	ZIP	ZINB
Maximum	0.007301	0.03	0.0147301	0.010324	0.023784	0.01129613
Mean	0.001065	0.00537	0.00352911	0.001224	0.0082594	0.002078
Minimum	2.40E-06	1.464E-13	9.032E-06	1.0985E-51	4.81E-05	1.09639E-06

Note: Δp_0 is the difference of p_0 s between the logit and probit inflated models

5.2.2 *Examine the Diagnostic Statistics*

By comparing each set of models, it is found out that NB models have better log likelihood values than those of Poisson models. In table 5.31, the difference of log

likelihood values between the Poisson and NB models is 136.7616. In table 5.31 and 5.33, the corresponding differences are 83.1707 and 152.371 separately. This consistently significant difference indicates that the NB is superior to the Poisson models.

The values of vuong statistics are uniformly greater than 1.96, which favors the selected zero inflated models over the corresponding base stochastic models. With the same model structure, ZIP is better than the Poisson model and ZINB is better than the NB model. The p value of α in ZINB models is higher than 0.05, which indicates the insignificance of α in the models. When the α in the NB model is close to 0, the NB turns into the Poisson model.

Table 5.32 The Summary of Model Results (1st Set of Models)

Model		Poisson(I)	NB(I)	ZIP(I)		ZINB(I)	
				Logit	Probit	Logit	Probit
Base Model	Occu	13.28704	12.67423	11.40821	11.4021	15.45177	15.3985
	Stdsp	0.2804112	0.287268	0.205147	0.204971	0.196311	0.197861
	Expose	0.18732895	0.17216075	0.138097	0.137900	0.103988	0.1045363
	Cons	-9.085271	-9.000146	-8.22397	-8.22176	-8.79263	-8.79522
Inflated Model	Occu					26.18221	14.91926
	Stdsp			-0.16537	-0.10169	-3.94130	-2.30222
	Expose			-0.11998	-0.07406	-1.18613	-0.68401
α			1.150312			0.847628	0.855648
Log Likelihood		-1231.0627	-1094.3011	-1163.08	-1163.1	-1076.22	-1076.1
Vuong				4.373	4.364	2.675	2.756

Table 5.33 The Summary of Model Results (2nd Set of Models)

Model		Poisson(II)	NB(II)	ZIP(II)		ZINB(II)	
				Logit	Probit	Logit	Probit
Base Model	Volume	0.000884	0.001058	0.000336	0.000339	0.001113	0.001114
	Speed	-0.0880675	-0.098854	-0.08291	-0.08277	-0.09564	-0.09569
	Expose	0.193520	0.209792	0.153848	0.152838	0.161682	0.162008
	_Cons	-3.960136	-3.660617	-2.9433	-2.95078	-3.6969	-3.6959
Inflated Model	Volume			-0.00174	-0.00106		
	Speed			0.037687	0.023125		
	Expose			-0.18110	-0.11517	-16.6164	-10.26771
	_Cons					16.4897	10.19261
α			0.711909			0.58208	0.582751
Log Likelihood		-1122.7057	-1039.535	-1078.96	-1078.55	-1029.37	-1029.30
Vuong				3.398	3.444	2.262	2.278

Table 5.34 The Summary of Model Results (3rd Set of Models)

Model		Poisson(111)	NB(111)	ZIP(111)		ZINB(111)	
				Logit	Probit	Logit	Probit
Base Model	Volume	0.001373	0.0014075	0.001509	0.001521	0.001479	0.00148
	Stdsp	0.26257	0.3325017	0.164497	0.165156	0.32629	0.326229
	Expose	0.27870	0.27708	0.17919	0.17703	0.21857	0.21955
	_Cons	-10.1073	-10.24142	-9.53743	-9.55684	-10.077	-10.082
Inflated Model	Volume			0.000856	0.000542		
	Stdsp			-0.75074	-0.44874		
	Expose			-0.6214	-0.4017	-16.9573	-10.7228
	_Cons					16.9309	10.70862
α			1.108093			0.91276	0.915649
Log Likelihood		-1236.644	-1084.327	-1170.05	-1170.54	-1072.46	-1072.46
Vuong				4.129	4.102	2.475	2.477

Similarly, when the α in the ZINB model is close to 0, it turns into the ZIP model. Therefore, in this study, the selected ZINB models can be replaced by the ZIP models with the same model structures without loss in their performance.

5.2.3 The Sign of Speed

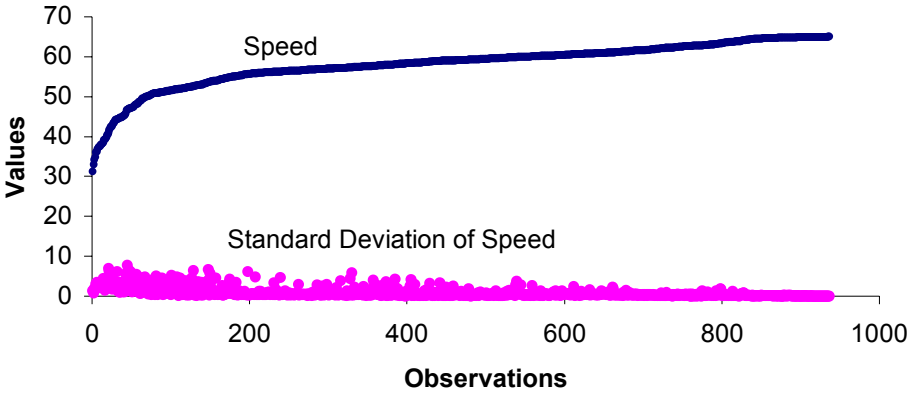
In the base models, the sign of speed is negative, while the sign of standard deviation of speed is positive. It was found that the correlation coefficient between speed and standard deviation of speed is -0.61. This negative correlation coefficient suggests that a negative relationship exists between the speed and standard deviation of speed. Table 5.35 shows the mean speed and standard deviation of speed for each speed interval. When the speed is higher than a certain value, 40 mph in this case, the higher the speed, the lower the standard deviation of speed. For example, In the highest speed interval of 60-65 mph, the standard deviation of speed is 0.34, which is much lower than those of lower speed intervals. Figure 5.1 shows the speed and standard deviation of speed for the 936 observations. A general trend was found that when the speed is low, the maximum standard deviation of speed is high.

Table 5.35 Speed and Standard Deviation of Speed

Speed Interval (mph)	Number of Observations	Frequency	Speed(mean)	Stdsp(mean)
<35	4	0.4%	33.33	1.51
35-39.9	14	1.5%	37.98	2.97
40-44.9	21	2.2%	43.35	3.39
45-49.9	30	3.2%	47.87	3.37
50-54.9	110	11.8%	52.57	1.87
55-59.9	372	39.7%	57.83	0.76
60-65	385	41.1%	62.54	0.34

Note: 1. The unit of speed is mph. The unit of standard deviation of speed is mph^2.
2. The mean standard deviation of speed was calculated as the average value of derived standard deviation of speed within the corresponding speed interval.

Figure 5.1 Speed and Standard Deviation of Speed



5.3 Model Result Testing

The same data was used for model testing as that of model estimation. The relative frequencies of the raw data and the estimated models and error rates between them were calculated and compared. Please refer to chapter 4 for the details of relative frequency.

5.3.1 Compare the Relative Frequency and Error Rate

The relative frequency and error rate were calculated for each estimated model and the raw data (please see Table 5.36 for the details). In this study, the crash counts were classified as 0, 1,…,6, and more. By examining table 5.36 and figure 5.2, 5.3, 5.4, it was found that Poisson regression models underestimated the probability of zero crashes,

while they gave higher estimates for all other crash counts. The Poisson regression models also have the largest error rates among the four kinds of regression models. This is consistent with all of the three sets of models. The other three kinds of models gave close estimation of probabilities for each crash count although zero inflated Poisson models have slight higher error rates. The error rates are consistent with log likelihood values of these models.

Table 5.36 Relative Frequencies of the Raw Data and Estimated Models

Count #/ freq(%)		0	1	2	3	4	5	6	>6	Error Rate
Raw		564/ 60. 3	214/ 22. 9	64/ 6. 8	26/ 2. 8	27/ 2. 9	11/ 1. 2	10/ 1. 1	19/ 1. 9	N/ A
I	P	475/ 50. 7	268/ 28. 6	108/ 11. 5	43/ 4. 6	19/ 2	9/ 1	5/ 0. 5	10/ 1	3. 2
	NB	566/ 60. 5	195/ 20. 9	81/ 8. 6	38/ 4. 1	20/ 2. 1	11/ 1. 2	7/ 0. 7	17/ 1. 9	1. 4
	ZI P	555/ 59. 3	176/ 18. 8	102/ 10. 9	50/ 5. 4	24/ 2. 6	12/ 1. 3	7/ 0. 7	10/ 1. 1	2. 7
	ZI NB	571/ 61	187/ 20	82/ 8. 7	39/ 4. 2	21/ 2. 2	12/ 1. 3	7/ 0. 8	18/ 1. 9	1. 5
I I	P	490/ 52. 4	270/ 28. 9	100/ 10. 7	36/ 3. 8	15/ 1. 6	8/ 0. 8	5/ 0. 5	13/ 1. 4	2. 9
	NB	562/ 60	209/ 22. 3	80/ 8. 5	34/ 3. 7	17/ 1. 8	10. 0/ 1	6/ 0. 6	20/ 2. 1	1. 7
	ZI P	559/ 60	179/ 19. 1	99/ 10. 6	46/ 4. 9	21/ 2. 3	11/ 1. 2	6/ 0. 7	14/ 1. 5	2. 3
	ZI NB	566/ 60. 5	204/ 21. 8	81/ 8. 6	35/ 3. 7	17/ 1. 8	1- Sep	6/ 0. 6	18/ 1. 9	1. 7
I I I	P	471/ 50. 3	273/ 29. 1	111/ 11. 8	43/ 4. 6	18/ 18. 7	8/ 0. 9	4/ 0. 4	10/ 1. 0	3. 5
	NB	569/ 60. 7	194/ 20. 7	80/ 8. 5	37/ 4	20/ 2	11/ 1. 2	7/ 0. 7	19/ 2	1. 5
	ZI P	539/ 57. 6	190/ 20. 3	102/ 10. 9	50/ 5. 3	25/ 2. 6	13/ 1. 3	7/ 0. 7	11/ 1. 2	2. 6
	ZI NB	573/ 61. 2	188/ 20	80/ 8. 6	38/ 4. 1	20/ 2. 1	11/ 1. 2	7/ 0. 7	19/ 2	1. 5

Note: # means the number of observations.

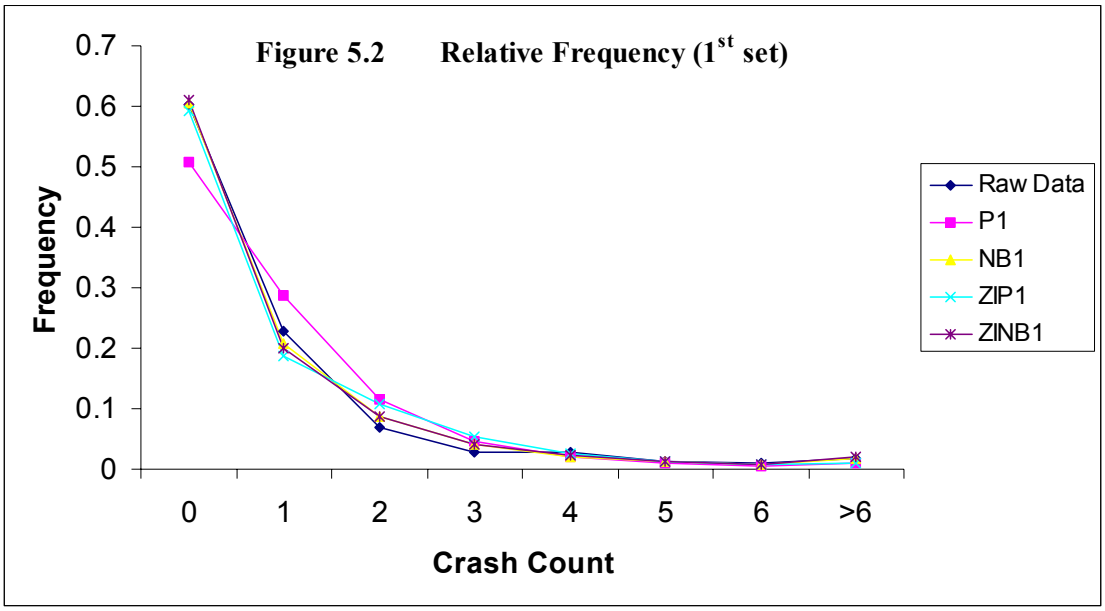


Figure 5.2 Relative Frequency (1st set)
Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 >6
Crash Count
Raw Data
P1
NB1
ZIP1
ZINB1

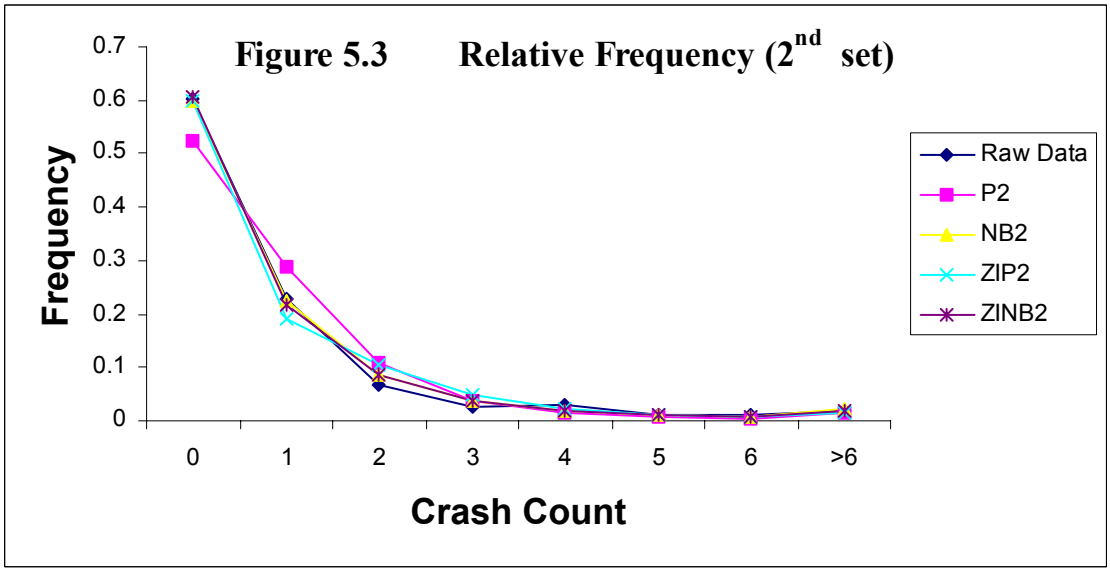
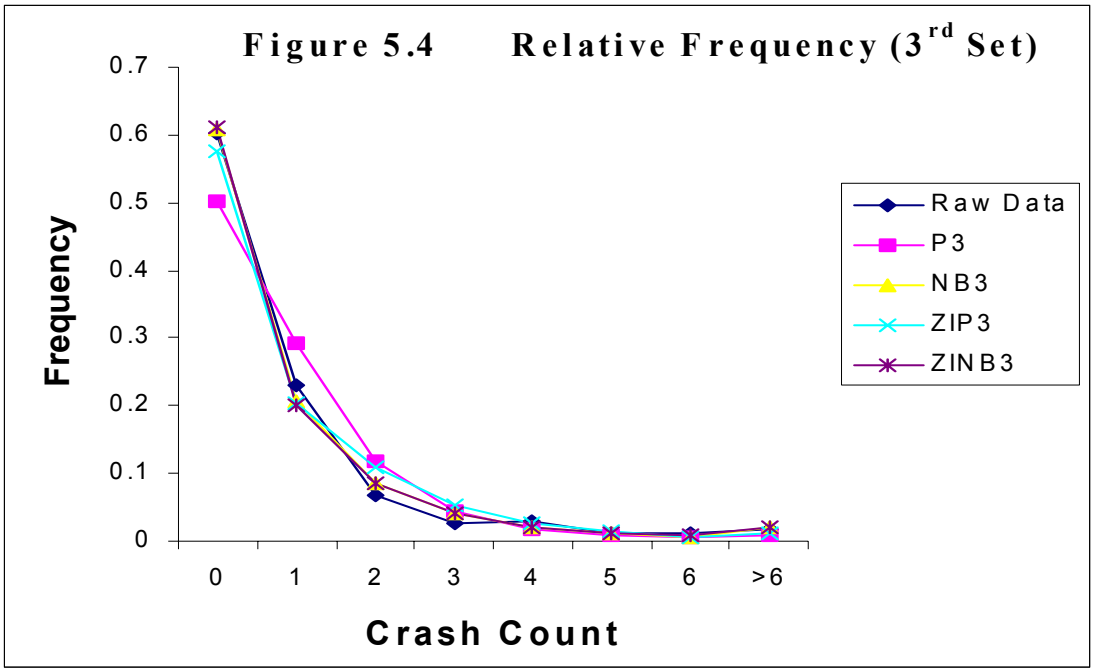


Figure 5.3 Relative Frequency (2nd set)
Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 >6
Crash Count
Raw Data
P2
NB2
ZIP2
ZINB2



5.3.2 p_0 of *Zero Inflated Regression Models*

Zero inflated models consist of two steps. In the first step, the probability of whether any event occurs at all is identified. In this study, p_0 is used to indicate the probability of zero crash counts and $1-p_0$ is used to indicate the probability of one or more crashes. During the model testing process, it was found that large percentages of small p_0s were obtained for ZINB models when compared with those for ZIP models. Taking the example of ZINB(II), 624 among 936 observations have p_0 smaller than 0.01. This suggests that the two-step ZINB regression models are not necessary for this study because of the insignificance of the first step. On the contrary, when a large proportion of the observed data have significantly high p_0, this suggests the importance

of using the two step ZIP models. Even though the performance of ZINB regression models is not inferior to those of NB and ZIP, ZINB is not a good alternative for this study.

5.4 Graphical Representation of Zero Inflated Models[36]

Recall that the expectation of zero inflated Poisson model is $E(X) = (1 - p_0) * \lambda$, a rectangle area is helpful in comparing the safety of various road sections. For example, road section A can be regarded better than road section B if and only if $(1 - p_0)^A \leq (1 - p_0)^B$ and $\lambda^A \leq \lambda^B$ (with at least one inequality being strict). Figure 5.5, 5.6 and 5.7 indicate the three conditions under which road section A is safer than road section B.

Figure 5.5 Comparison of Road Section A and B

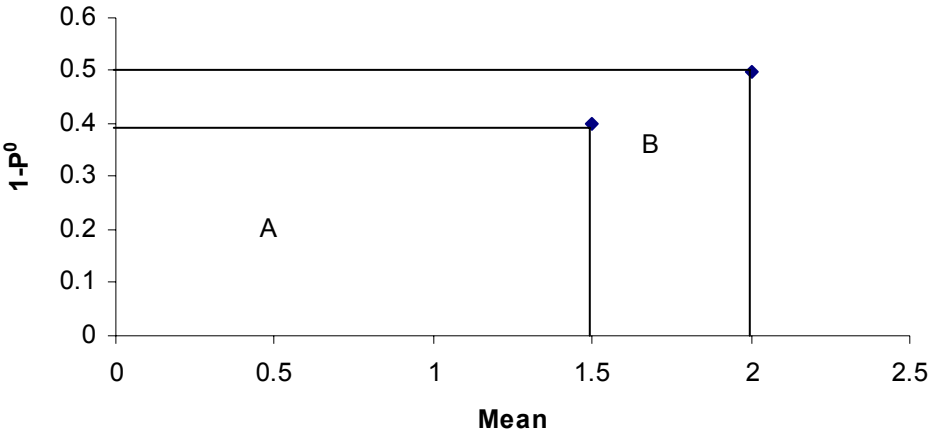


Figure 5.6 Comparison of Road Section A and B

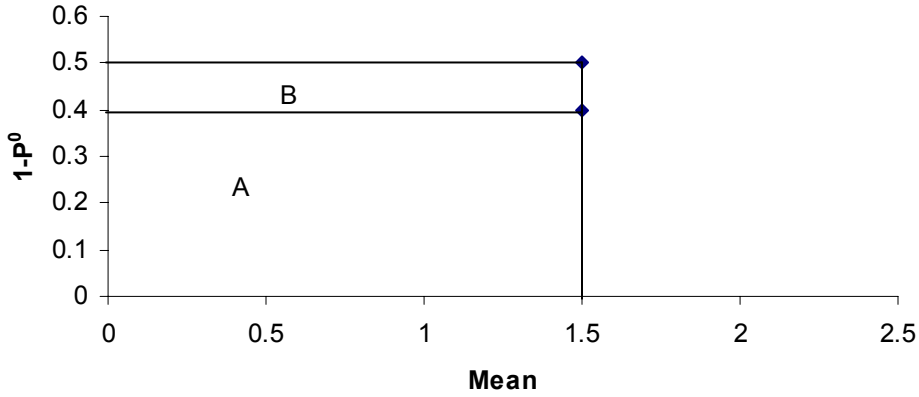
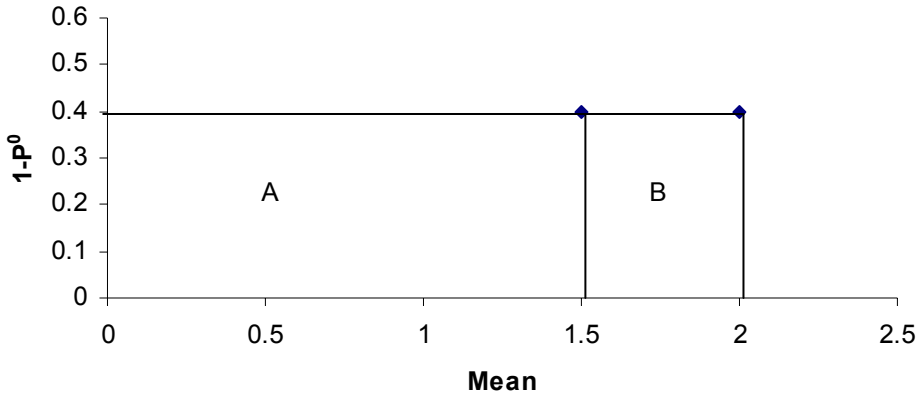


Figure 5.7 Comparison of Road Section A and B



5.5 Conclusions

1. Base on the above discussion, NB and ZIP are preferred over Poisson and ZINB regression models. Poisson regression models underestimate the probability of zero crashes , while overestimating probabilities of crashes more than zero. ZINB regression models present good results. However, the large percentage of very small P_0 s in ZINB models suggests that the two-regime ZINB regression models are unnecessary in this study. Both NB and ZIP regression models gave good results. NB regression models had slightly better log likelihood and error rate values than those of ZIP. However, the study showed that there is no significant difference in the performance of negative binomial and zero inflated Poisson regression models.

2. Three sets of independent variables were applied. No significant difference was found among them. They include: (1) occupancy, standard deviation of speed, and exposure; (2) volume, speed, and exposure; and (3) volume, standard deviation of speed, and exposure. No significant difference were found among these three sets of models.

3. Logit and probit models are good forms for use in zero infalted models. They can be used interchangeably.

4. In the base models, all the independent variables have positive algebraic signs except speed. Traffic volume, occupancy, standard deviation of speed, and exposure were shown to have significant positive relationships with the mean of the dependent variable of the link function, i.e. the number of crashes. The larger those variables, the larger the mean of number of crashes. On the contrary, speed was shown to have significant negative relationship with the mean of number of crashes. In this study, speed and standard deviation of speed showed fairly strong negative correlation with each other, which indicates that the negative sign of speed is reasonable in this study.

Chapter 6 CONCLUSIONS

6.1 Conclusions

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

- The closeness of the estimation and raw data indicates that stochastic regression modeling methods can be used to describe the probabilities of crash events.
- The use of corresponding traffic data (those occurring at the time of the crashes) reflected the true influences of the independent variables on the occurrence of crashes.
- The models that have been developed can be used in the field to predict the probability of a certain number of crashes under different geometric and traffic conditions, for which data can be obtained directly in the field.
- These models will facilitate the use of real time data in the field to develop congestion relieving strategies that do not have negative impacts on safety.

6.2 Recommended Further Research Efforts

6.2.1 Other Modeling Methods

Apart from the application of stochastic regression models, other promising methodologies include artificial neural networks, fuzzy methods, and genetic algorithms. Through the genetic evolution method, an optimal solution can be found and represented by the final winner of the genetic game.

6.2.2 Development of Additional Estimation and Evaluation Procedures

Additional programs are needed to accommodate more flexibility in the estimation and evaluation of stochastic models. For example, STATA only uses the

maximum likelihood estimation method for stochastic regression models, but it could also be feasible to use the quasi-likelihood estimation method.

6.2.3 Wide Area Examination of Stochastic Regression Models

This study has shown the effective modeling methods and significant independent variables in the models. However, more roadway sections need to be taken under study to verify and support the findings of this study before the wide area application of this study.

6.2.4 Application Of Models In Advanced Transportation Management Systems

One of the major objectives of crash modeling research is to support the traffic management including regular and real time management. Accurate and reliable relationships between the occurrence of crashes and highway geometric and traffic conditions under a certain environment could present useful insight to the potential corresponding safety and traffic operation performance. Therefore, more research should be performed to incorporate the models into traffic management systems.

REFERENCES

1 Jovanis, P., Chang, H., 1986. Modeling the relationship of accidents to miles traveled. Transportation Research Record 1068, 42-51.

2 Garber, N., Joshua, S., 1990. Traffic and geometric characteristics affecting the involvement of large trucks in accidents. VDOT Project No.: 9242-062-940, Virginia Transportation Research Council, University Station, Charlottesville, Virginia.

3 Miaou, S., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention, 26, 471-482.

4 Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1-14.

5 Greene, W., 1994. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working paper. Stern School of Business, Economics Department, New York University, New York, New York.

6 Gan, N., 2000. General zero-inflated models and their applications. Dissertation. Department of Statistics, North Carolina State University, Raleigh, North Carolina.

7 Awad, W., Jason, B., 1998. Prediction models for truck accidents at freeway ramps in Washington state using regression and artificial intelligence techniques. Transportation Research Record 1635, 30-36.

8 Lundy, R., 1965. Effect of traffic volumes and number of lanes on freeway accident rates. Highway Research Board Number 99, Highway research Board of the National Academy of Sciences, National Research Council, Washington, DC.

9 Hall, J.W., Pendleton, O.J., 1990. Rural accident rate variations with traffic volume. Transportation Research Record 1281, 62-70.

10 Gwynn, D., 1967. Relationship of accident rates and accident involvements with hourly volumes. Traffic Quarterly, Vol. XXI, No. 3, 407-418.

11 Perkins, E, 1957. Relationship of accident rate to highway shoulder width. Highway Research Board Bulletin 151, 13-14.

12 Zegeer, C., Stewart, R., Council, F., Neuman, T., 1994. Accident relationships of roadway width on low-volume roads. Transportation Research Record 1445, 160-168.

13 Zegeer C., Deacon, J., 1987. Effect of lane width, shoulder width, and shoulder type on highway safety. Relationship between safety and key highway features: state of the art report 6. Transportation Research Board, Washington, DC.

14 Garber, N., Ehrhart, A., 2000. Effect of speed, flow, and geometric characteristics on crash frequency for two lane highways. Transportation Research Record 1717, .

15 Glennon, J., 1987. Effect of alignment on highway safety. Relationship between safety and key highway features: state of the art report 6. Transportatin Research Board, Washington, DC.

16 Glennon, J., 1987. Effect of sight distance on highway safety. Relationship between safety and key highway features: state of the art report 6. Transportatin Research Board, Washington, DC.

17 Mohamedshah, Y., Paniati, J., Hobeika, A., 1993. Truck accident models for interstates and two-lane rural roads. Transportation Research Record 1407, 35-41.

18 Persaud, B., Dzbik, L., 1992. Relating freeway accidents to traffic, geometric and operational factors. Compendium of Technical Papers. Institute of Transportation Engineers Annual Meeting. Washington, DC, 298-301.

19 Okamoto, H., Koshi, M., 1989. A method to cope with the random errors of observed accident rates in regression analysis. Accident Analysis and Prevention. Vol.21, 317-332.

20 Miaou, S., Hu, P., Wright, T., Rathi, A., Davis, S., 1992. Relationship between truck accidents and highway geometric design: a Poisson regression approach. Transportation Research Record 1376, 10-18.

21 Hadi, M., Aruldhas, J., Chow, L., Wattleworth, J., 1995. Estimating safety effects of cross-section design for various highway types using negative binomial regression. Transportation Research Record 1500, 169-177.

22 Miaou, S., Lum, H., 1993. Statistical evaluation of the effects of highway geometric design on truck accident involvements. Transportation Research Record 1407, 11-23.

23 Miaou, S., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis and Prevention, Vol.25, 689-709.

24 Vogt, A., Bared, J., 1998. Accident Models for Two-lame Rural Segments and Intersections. Transportation Research Record 1635, TRB, National Research Council, Washington, DC, pp18-29.

25 Ivan, J., Pasupathy, R., Ossenbruggen, P., 1999. Differences in causality factors for single and multi-vehicle crashes on two-lane roads. Accident Analysis and Prevention. Vol.31, 695-704.

26 Karlaftis, M., Tarko, A., 1998. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention. Vol.30, 425-433.

27 Fridstr∅m, L., Ingebrigtsen, S., 1991. An aggregate accident model based on pooled, regional time-series data. Accident Analysis & Prevention. Vol.23, 363-378.

28 Lin, T., Jovanis, P., Yang, C., 1993. Modeling the safety of truck driver service hours using time-dependent logistic regression. Transportation Research Record 1407, 1-10.

29 Vogt, A., Bared, J., 1998. Accident Models for two-lane rural roads: segments and intersections. FHWA-RD-98-133. Pragmatics, Inc., McLean, Virginia.

30 Vaija, P., 1987. Application of fuzzy methods to process safety control and to accident analysis. Acta Polytechnical Scandinavica, Chemical Technology and Metallurgy Series. No. 180, Helsinki 1987, 26pp.

31 Joshua, S., Garber, N., 1992. A causal analysis of large vehicle accidents through fault-tree analysis. Risk Analysis, Vol 12, No. 2, 173- 187.

32 Hakkert, A., Hocherman, I., Mensah, A., 1996. Levels of Safety on Interurban Roads. Transportation Research Record 1553, 95-102.

33 Elvik, R., 1996. Does prior knowledge of safety effect help to predict how effective a measure will be. Accident Analysis and Prevention, Vol.28, 339-347.

34 Subramanyan, S., 2000. The feasibility of developing congestion mitigation measures that incorporate crash risk. A case study: Hampton Roads area. Master Thesis, University of Virginia, Charlottesville, Virginia.

35 Hayter, A., 1996. Probability and statistics for engineers and scientists. PWS Publishing Company.

36 Böhning, D., Dietz, E., Schlattmann, P., 1997. Applications of latent trait and latent class models in the social science, chapter 32 zero-inflated count models and their applications in public health and social science. Waxmann Publishing Co., ISBN 3-89325-464-1. 333-344.

Appendix A: Sample SQL Codes Used To Extract Crash Data from the Smart Travel Lab At University Of Virginia

```
spool C:\e-64-01-w.txt

COLUM CRASH  FORMAT A15
COLUM HOUR  FORMAT A6
COLUM DATEX FORMAT A11
COLUM LANE  FORMAT A25
COLUM WEATHER FORMAT A8
BREAK ON noo SKIP 4
REM ** INSERT SELECT STATEMENT
SELECT
        A.TMS_CALL_NUMBER as CRASH,
        TO_CHAR(B.BEGIN,'hh24:mi') AS HOUR,
        TO_CHAR(B.BEGIN, 'mm-dd-yyyy') AS DATEX,
        A.LANE,
        B.WEATHER
FROM
        HR.INC_ROADWAY A,
        HR.INCIDENT B
WHERE
        A.LOCATION ='E64-01'
        AND upper(A.DIRECTION) in ('WEST','WEST BOUND')
        AND upper(B.TYPE) = 'ACCIDENT'
        AND A.TMS_CALL_NUMBER = B.TMS_CALL_NUMBER
        AND B.BEGIN BETWEEN  to_date('07-01-1998','mm-dd-yyyy') and
to_date('07-01-2001','mm-dd-yyyy')

/
spool off
```

Appendix B: Sample Crash Data

TMS-CALL Number	DATE	WEEKDAY	TIME	WEATHER	LANE	LOCATION	DIRECTION
1998-16021	11/15/98	1	19:32:00	Clear	Right Shoulder	W64-06	East Bound
1999-10836	6/27/99	1	17:30:00	Cloudy	Right Shoulder, Gore	W64-06	East Bound
1999-22126	11/14/99	1	16:01:00	Clear	Right Shoulder, Reversible	W64-06	East Bound
2001-04344	2/18/01	1	22:29:00	Cool	5, Right Shoulder	W64-06	East Bound
2001-12766	5/6/01	1	23:00:00	Clear	1	W64-06	East Bound
1998-10231	8/10/98	2	14:18:00	Clear	Right Shoulder	W64-06	East Bound
1999-17888	9/20/99	2	17:33:00	Cloudy	Left Shoulder, 1	W64-06	East Bound
2000-18730	7/17/00	2	11:23:00	Clear	Right Shoulder	W64-06	East Bound
2001-01260	1/15/01	2	8:24:00	Clear	Left Shoulder, 3, Right Shoulder	W64-06	East Bound
1999-12278	7/13/99	3	19:20:00	Cloudy	Gore	W64-06	East Bound
2000-10564	5/9/00	3	15:46:00	Clear	Right Shoulder	W64-06	East Bound
2000-30046	10/24/00	3	16:36:00	Clear	5	W64-06	East Bound
2001-22750	7/24/01	3	12:55:00	Cloudy	Right Shoulder	W64-06	East Bound
1999-22916	11/24/99	4	11:32:00	Cloudy	5, Right Shoulder	W64-06	East Bound
2000-25539	9/13/00	4	9:13:00	Clear	Left Shoulder	W64-06	East Bound
2001-22946	7/25/01	4	17:10:00	Hot / Humid	4	W64-06	East Bound
2001-22947	7/25/01	4	17:13:00	Hot / Humid	4	W64-06	East Bound
1998-08536	7/9/98	5	17:40:00	Clear	Left Shoulder, 1	W64-06	East Bound
1998-15420	11/5/98	5	15:52:00	Clear	Left Shoulder	W64-06	East Bound
1998-17480	12/10/98	5	17:49:00	Clear	3, Shoulder Lane	W64-06	East Bound
1999-03817	3/11/99	5	16:27:00	Clear	5, Right Shoulder	W64-06	East Bound
1999-06361	4/22/99	5	10:03:00	Warm	Right Shoulder	W64-06	East Bound
2000-14774	6/15/00	5	12:51:00	Clear	5	W64-06	East Bound
2001-07506	3/22/01	5	17:47:00	Cloudy	5, Right Shoulder	W64-06	East Bound
2001-18413	6/21/01	5	8:26:00	Clear	Right Shoulder	W64-06	East Bound
1999-02352	2/12/99	6	8:25:00	Clear	Left Shoulder, 1, 2	W64-06	East Bound
2000-04386	3/3/00	6	10:19:00	Clear	1, 2, 3	W64-06	East Bound

Appendix C: Sample Traffic Data

STATIONID	WEEKDAY	TIME	VOLUME	OCCUPANCY	SPEED
104	1	0:00	719.1115	1.272917	49.25625
104	1	0:10	611.3884	1.211579	48.9379
104	1	0:20	583.725	1.207292	47.87604
104	1	0:30	528.1469	0.884694	45.59286
104	1	0:40	482.2212	0.817172	44.59192
104	1	0:50	422.209	0.797	41.984
104	1	1:00	381.9515	0.820202	43.11111
104	1	1:10	287.9313	0.819192	43.53636
104	1	1:20	251.3354	0.767677	42.85253
104	1	1:30	243.7238	0.859406	42.88812
104	1	1:40	215.4404	1.016162	43.40707
104	1	1:50	248.087	1.04	42.249
104	1	2:00	246.7726	1.030392	41.44314
104	1	2:10	286.5153	1.380612	42.39286
104	1	2:20	311.3485	1.140206	43.20825
104	1	2:30	323.6697	1.128283	44.13939
104	1	2:40	287.938	1.105	43.112
104	1	2:50	193.499	1.305051	42.91111
104	1	3:00	174.7612	1.285437	42.89223
104	1	3:10	160.5901	1.316832	42.00297
104	1	3:20	141.5529	1.275	39.5125
104	1	3:30	115.3629	1.255238	38.9
104	1	3:40	103.681	1.259048	38.01238
104	1	3:50	92.68058	1.275728	39.09806
104	1	4:00	86.62941	1.288235	37.12157
104	1	4:10	75.91942	0.972816	37.16311
104	1	4:20	65.32981	0.953846	38.21346
104	1	4:30	66.61165	0.951456	38.89418
104	1	4:40	62.69615	0.946154	38.18462
104	1	4:50	64.8781	0.937143	35.90286
104	1	5:00	76.17076	0.956604	38.39623
104	1	5:10	71.98824	0.962745	45.55392
104	1	5:20	78.35631	0.957282	44.78252
104	1	5:30	104.5447	0.959223	44.72621
104	1	5:40	115.3515	1.009901	47.01386
104	1	5:50	130.1124	0.979048	43.85429
104	1	6:00	137.4702	0.961538	41.35962
104	1	6:10	140.9943	0.96381	43.60095
104	1	6:20	158.5231	0.965385	43.17692
104	1	6:30	174.4	1.036893	45.00874
104	1	6:40	199.9874	0.912621	45.57476
104	1	6:50	228.8495	0.821905	44.45048
104	1	7:00	245.1914	0.879048	44.76857
104	1	7:10	232.6238	0.87619	45.46571