OVER- AND UNDER-DISPERSED CRASH DATA: COMPARING THE CONWAY-MAXWELL-POISSON AND DOUBLE-POISSON DISTRIBUTIONS

A Thesis

by

YAOTIAN ZOU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2012

Major Subject: Civil Engineering

Over- and Under-dispersed Crash Data: Comparing the Conway-Maxwell-Poisson and Double-Poisson Distributions

OVER- AND UNDER-DISPERSED CRASH DATA: COMPARING THE CONWAY-MAXWELL-POISSON AND DOUBLE-POISSON DISTRIBUTIONS

A Thesis

by

YAOTIAN ZOU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Dominique Lord
Committee Members,	Yunlong Zhang
	Thomas E. Wehrly
Head of Department,	John Niedzwecki

August 2012

Major Subject: Civil Engineering

ABSTRACT

Over- and Under-dispersed Crash Data: Comparing the Conway-Maxwell-Poisson and Double-Poisson Distributions. (August 2012)

Yaotian Zou, B.E., Southeast University

Chair of Advisory Committee: Dr. Dominique Lord

In traffic safety analysis, a large number of distributions have been proposed to analyze motor vehicle crashes. Among those distributions, the traditional Poisson and Negative Binomial (NB) distributions have been the most commonly used. Although the Poisson and NB models possess desirable statistical properties, their application on modeling motor vehicle crashes are associated with limitations. In practice, traffic crash data are often over-dispersed. On rare occasions, they have shown to be under-dispersed. The over-dispersed and under-dispersed data can lead to the inconsistent standard errors of parameter estimates using the traditional Poisson distribution. Although the NB has been found to be able to model over-dispersed data, it cannot handle under-dispersed data. Among those distributions proposed to handle over-dispersed and under-dispersed datasets, the Conway-Maxwell-Poisson (COM-Poisson) and double Poisson (DP) distributions are particularly noteworthy. The DP distribution and its generalized linear model (GLM) framework has seldom been investigated and applied since its first introduction 25 years ago.

The objectives of this study are to: 1) examine the applicability of the DP distribution and its regression model for analyzing crash data characterized by over- and under-dispersion, and 2) compare the performances of the DP distribution and DP GLM with those of the COM-Poisson distribution and COM-Poisson GLM in terms of goodness-of-fit (GOF) and theoretical soundness. All the DP GLMs in this study were developed based on the approximate probability mass function (PMF) of the DP distribution.

Based on the simulated data, it was found that the COM-Poisson distribution performed better than the DP distribution for all nine mean-dispersion scenarios and that the DP distribution worked better for high mean scenarios independent of the type of dispersion. Using two over-dispersed empirical datasets, the results demonstrated that the DP GLM fitted the over-dispersed data almost the same as the NB model and COM-Poisson GLM. With the use of the under-dispersed empirical crash data, it was found that the overall performance of the DP GLM was much better than that of the COM-Poisson GLM in handling the under-dispersed crash data. Furthermore, it was found that the mathematics to manipulate the DP GLM was much easier than for the COM-Poisson GLM and that the DP GLM always gave smaller standard errors for the estimated coefficients.

ACKNOWLEDGEMENTS

First and foremost, I would like to give my sincere gratitude to my advisor, Dr. Dominique Lord for his tremendous and constant help on completing the thesis. His guidance, comments and suggestions trained me in a professional manner and ensured the research on the right track. The concern and encouragement he gave me inspired me to find a way to get though all the difficulties in preparing the thesis.

I would also like to appreciate the help from the committee members, Dr. Yunlong Zhang and Dr. Thomas Wehrly for their advice and reviews on this thesis. Specials thanks are given to Dr. Thomas Wehrly for his thoughtful answers on my statistics-related questions.

Particularly, I would like to thank Dr. Srinivas Geedipally for his detailed review on the research and his help on accessing the data and simulation codes.

I am also grateful to my colleagues and friends, including Pei-fen Kuo, Yajie Zou, Fan Ye, and Lingzi Cheng who have been willing to offer their help, support and comments.

Last but not the least, I am specially thankful to my parents who financially supported me for my graduate study. They are always standing by me and encouraging me through ups and downs.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

1. INTRODUCTION

Traffic crashes have been huge negative impacts on the human health and economic development. Much time and effort have been devoted by researchers to pinpoint factors that influence traffic crashes and propose countermeasures to reduce the crash occurrences. However, due to the limited access of individual driver's information, it is difficult to identify factors influencing the number and severity of crashes and evaluate their effects on traffic safety. Instead of focusing on the individual information, most researchers approach the crash cause study from a long-term statistical view. They have been trying to associate the factors of interest with the frequency of crashes that occurs in a given space (roadway or intersection) and time period (Lord and Mannering, 2010). Therefore, statistical models have been widely used to analyze the relationship between traffic crashes and factors such as road section geometric design, traffic flow, weather, etc. The most important application of those statistical models established on the historical data lies in its capability of predicting the number of crashes on the newly built or upgraded roads (Lord, 2000).

The Poisson distribution is commonly used to model count data. In traffic safety analysis, it has been frequently used to model the number of crashes for various entities such as roadway segments and intersections over a given time period. However, the Poisson distribution has only one parameter which requires the variance equals the mean

---

This thesis follows the style of Accident Analysis and Prevention.

and it does not allow for the flexibility of variance varying independently of the mean. In practice, traffic crash data are often over-dispersed (i.e., the sample variance is larger than the sample mean) (Lord et al. 2005). On rare occasions they have been shown to be under-dispersed (i.e., the sample variance is smaller than the sample mean) and this often happens when the sample mean value is low (Lord and Mannering, 2010). The over-dispersed and under-dispersed data would lead to the inconsistent standard errors of parameter estimates using the traditional Poisson distribution (Cameron and Trivedi, 1998).

In light of the limitations of the traditional Poisson models and the wide presence of under- and over-dispersion in traffic crash data, it is important for researchers to examine the application of innovative statistical methods for analyzing crash data. In order to handle the over-dispersion, a large number of statistical methods have been proposed ranging from the most commonly used model mixed-Poisson (such as the negative binomial or NB) to those most recent models such as the neural and Bayesian neural networks, latent class or mixture model, gamma count model and support vector machine model (Abdelwahab and Abdel-Aty, 2002; Xie et al., 2007; Depaire et al., 2008; Park and Lord, 2008; Oh et al., 2006; Li et al., 2008). The NB is the most widely used model because it has closed form equation and the mathematical relationship between the mean and the variance is very easy to manipulate (Hauer, 1997). It should be noted that traditional distributions such as the Poisson or NB cannot handle under-dispersion.

To handle the data characterized by under-dispersion, researchers proposed alternative models such as the weighed Poisson (Castillo and Perezcasany, 2005), the

generalized Poisson models (Consul, 1989) and the gamma count distribution (Winkelmann, 1995). However, these models suffered from their theoretical or logical soundness. In the generalized Poisson model, the bounded dispersion parameter when under-dispersion occurs greatly diminishes its applicability to count data (Famoye, 1993). As for the gamma distribution, two parameterizations have been proposed by researchers. One parameterization is based on the continuous gamma density function (Daniels et al., 2010), which does not allow the count to be equal to zero. Based on gamma waiting time distribution, another parameterization assumes that observations are not independent where the observation for time t-1 would affect the observation for time t (Winkelmann, 1995; Cameron, 1998). This would become unrealistic if the time gap between the two observations is large.

Among the distributions that have been examined in the literature, two distributions that can handle both under- and over-dispersion are particularly noteworthy. One is the Conway-Maxwell-Poisson (COM-Poisson) (Conway and Maxwell, 1962; Shmueli, 2005; Kadane et al., 2006) and the other is the Double-Poisson (DP) (Efron, 1986). Albeit first introduced in 1962, the statistical properties of the COM-Poisson have not been extensively investigated until recently the COM-Poisson distribution and its generalized regression model (GLM) have been found to be very flexible to handle count data (Guikema and Coffelt, 2008; Geedipally, 2008; Sellers et al., 2011; Francis et al. 2012). As for the DP, its distribution has seldom been investigated and applied since its first introduction 25 years ago.

1.1 Problem Statement

In traffic safety analysis, a large number of distributions have been proposed to analyze the number of crashes on various entities, such as roadway segments and intersections, for a given time period. In practice, traffic crash data are often over-dispersed. On rare occasions, they have shown to be under-dispersed. The over-dispersed and under-dispersed data can lead to the inconsistent standard errors of parameter estimates using the traditional Poisson distribution. Although the NB distribution has been found to be able to model over-dispersed data, it cannot handle under-dispersed data.

Among the distributions that can handle under-dispersed data, two distributions are particularly noteworthy. They are the COM-Poisson and DP, both of which can handle data characterized by under-, equi- and over-dispersion. The COM-Poisson distribution and COM-Poisson GLM have been found to be very flexible to handle count data. While for the DP, its distribution has seldom been investigated and applied since its first introduction 25 years ago.

Therefore, it is of interest to examine the applicability of the DP distribution and its regression model for analyzing crash data characterized by over- and under-dispersion. For a new distribution like the DP, it is important to first evaluate the distribution before dealing with the regression model. So there is a need to compare the performances of the DP distribution and DP GLM with those of the COM-Poisson distribution and COM-Poisson GLM in terms of goodness-of-fit (GOF) and theoretical soundness.

1.2 Study Objectives

This study focuses on the applicability of different distributions and their GLMs for analyzing the crash data characterized by under- and over-dispersion. Specifically, the DP and COM-Poisson models will be further explored and compared in terms of their potential capability of handling both under- and over-dispersed data.

- Evaluating of the Performance of the DP Distribution

  The performance of the DP distribution will be assessed and compared to other distributions with no covariates considered. Nine scenarios of simulated data with three means (high, medium and low) and three levels of dispersion (under-, equi-, and over- dispersion) will be examined in this study. The simulated data will be generated by different distributions. Comparisons on GOF statistics of simulated data fitted by the DP and COM-Poisson will be conducted. The GOF statistics of simulated data fitted by other distributions such as the Poisson, NB, and gamma count model will also be given as a reference.

- Comparing the GLM Performance for Over-dispersed Data

  The performance of the DP GLM in handling over-dispersed crash data will be compared with that of the NB model and COM-Poisson GLM. Two observed over-dispersed datasets along with two different and commonly used link functions will be used to establish the GLMs in order to eliminate the potential bias of using only one dataset or one link function.

- Comparing the GLM Performance for Under-dispersed Data

  The performance of the DP GLM in handling under-dispersed crash data will be compared with that of the NB model and COM-Poisson GLM. Pairwise comparisons will be first conducted between the DP GLM with other two models. Then an overall comparison among the three models will be provided.

1.3 Outline of the Thesis

The outline of this thesis is as follows:

Section 2 provides an overview on the statistical models proposed to handle the over-and under-dispersion of traffic crash data. The limitation of each model will also be discussed. The COM-Poisson and DP models will be mainly introduced at the end of this section.

Section 3 evaluates the performance of the DP distribution using nine mean-dispersion scenarios of simulated data. The performance of the DP distribution is compared to that of the COM-Poisson distribution. The GOF statistics of simulated data fitted by other distributions such as the Poisson, NB, and gamma count are also given as a reference.

Section 4 summarizes the performance of the DP GLM in analyzing the traffic crash data characterized by over-dispersion. The results on the NB model and COM-Poisson GLM are also presented. This section further investigates the effects of the key covariates and conducts the residual checking and the variance analysis. At the end of

this section, the use of the normalizing constant in the probability mass function of the DP GLM will be discussed.

Section 5 investigates the performance of the DP GLM in analyzing the under-dispersed traffic crash data. The comparison results with the COM-Poisson GLM and gamma count model are also summarized. Further interpretation on the effects of key covariates is also given.

Section 6 summarizes the main findings of this research. It also documents future work directions at the end.

2. BACKGROUND

This section provides an overview on the statistical models proposed to handle the over- and under-dispersion of traffic crash data. The characterization of each model and their corresponding GLM framework will be described. The limitation of each model will also be discussed. The COM-Poisson and DP models will be mainly introduced at the end of this section.

2.1 Poisson Model

The Poisson distribution is a discrete probability distribution to describe the number of occurrences in a given interval of time or space. The average rate of the occurrences is known and the occurrence of one event is independent of the occurrence of others. Crashes are mostly characterized by rareness, discreteness and randomness. Lord et al. (2005) indicated that crashes can be best characterized as Bernoulli trails with low probability and large number, which makes the number of crashes can be characterized as Poisson trials. The Poisson distribution is frequently used to model the crash data characterized by the variance increasing with the increase of the mean.

The probability mass function (PMF) of the Poisson distribution is:

$$P(y_i|\lambda_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \tag{2.1}$$

where y_i is the number of crashes per year for site i, and λ_i is the mean crashes per year.

The mean and variance of the Poisson distribution is given by:

$$E(Y) = Var(Y) = \lambda_i \tag{2.2}$$

For the Poisson regression model, the expected number of crashes per year λ_i is linked to the explanatory variables x_i such as the traffic flows and geometric design factors by the following link function:

$$\lambda_i = \exp(x_i \beta) \tag{2.3}$$

where the vector β is the coefficients to be estimated.

The limitation of the Poisson model lies in that it requires the variance is equal to the mean. In practice, traffic crash data are often over-dispersed which means the variance is larger than the mean. The over-dispersion arises from the unobserved differences across sites (Washington et al., 2003) and unmeasured uncertainties associated with the observed or unobservable variables (Lord and Park, 2008). On rare occasions the crash data have been shown to be under-dispersed and this often happens when the sample mean value is low (Lord and Mannering, 2010). The over-dispersed and under-dispersed data would lead to the inconsistent standard errors for the parameter estimates using the traditional Poisson distribution (Cameron and Trivedi, 1998).

2.2 Negative Binomial Model

The NB (or Poisson-gamma) is the most widely used model in analyzing crash data. It has been found to serve as a good alternative to handle over-dispersion and the mathematics to manipulate the relationship between the mean and variance is relatively

simple (Hauer, 1997). Furthermore, its regression model has been well incorporated in many statistical software such as SAS (SAS Institute Inc., 2002) and R (R Development Core Team, 2006).

The NB distribution was first used to model the random number of successes Y until a predefined number of r of failures based on a sequence of Bernoulli trials. The PMF of the NB distribution is:

$$P(Y = y; r, p) = \binom{y + r - 1}{y} (1-p)^r (p)^y; r = 0,1,2,...,0 < p < 1 \qquad (2.4)$$

The parameter p is the probability of success in each trial and it is calculated as:

$$p = \frac{r}{\mu + r} \qquad (2.5)$$

where,

μ = E(Y) = mean of the observations;

r = inverse of the dispersion parameter alpha (i.e. $r = 1/\alpha$).

When the parameter r is extended to a real, positive number, its PMF can be rewritten using the gamma function:

$$P(Y = y; r, p) = \frac{\Gamma(r + y)}{\Gamma(r) \times y!} (1-p)^r (p)^y; r > 0, 0 < p < 1 \qquad (2.6)$$

And it can be shown (Casella and Berger, 1990):

$$Var(Y) = r \frac{p}{(1-p)^2} = \frac{1}{r} \mu^2 + \mu \qquad (2.7)$$

Based on the Equations (2.4) and (2.5), the PMF of the NB distribution can be re-parameterized as:

$$P(Y=y;r,\mu)=\frac{\Gamma(r+y)}{\Gamma(r)\times\Gamma(y+1)}(\frac{r}{\mu+r})^{r}(\frac{\mu}{\mu+r})^{y};r>0,0<p<1 \qquad (2.8)$$

This PMF shown in Equation (2.8) has been frequently used to model vehicle crash count data.

In the NB regression model, μ is linked to the covariates:

$$\mu_i=\exp(x_i\beta) \qquad (2.9)$$

The NB distribution is also known as the Poisson-gamma distribution. The Poisson-gamma distribution is based on another parameterization in which the number of crashes Y_i is Poisson distributed with its conditioned mean μ_i:

$$Y_i\big|\mu_i\sim Po(\mu_i),i=1,2,...,n \qquad (2.10)$$

The mean of the crashes is given by:

$$\mu_i=\lambda_i\exp(\varepsilon_i) \qquad (2.11)$$

The $\exp(\varepsilon_i)$ is assumed to follow a gamma distribution for all site i:

$$\exp(\varepsilon_i)\big|\alpha\sim gamma(r,r) \qquad (2.12)$$

Despite of its popularity in traffic crash data analysis, the NB models suffers limitation in fitting data characterized by under-dispersion. The NB could theoretically handle under-dispersion by setting its shape parameter as negative ($Var(Y)=\mu+(-\alpha)\mu^2$). However, doing that would make the conditioned mean of the Poisson no longer gamma distributed and lead to a misspecification of its PDF (Clark and Perry, 1989; Saha and Paul, 2005) and unreliable parameter estimates (Lord et al, 2010).

2.3 Gamma Count Model

The gamma count model was proposed by Winkelmann (1995) to model over- and under-dispersed count data. Oh et al. (2006) applied the gamma count model to analyze rail-highway crossing crashes and the data were found to be under-dispersed. The gamma count model for count data is given as:

$$\Pr(y_i = j) = Gamma(\alpha j, \lambda_i) - Gamma(\alpha j + \alpha, \lambda_i) \tag{2.13}$$

where $\lambda_i = \exp(\beta X_i)$ and λ_i is the mean of the crashes.

$$Gamma(\alpha j, \lambda_i) = 1, \quad if \quad j = 0, \tag{2.14}$$

$$Gamma(\alpha j, \lambda_i) = \frac{1}{\Gamma(\alpha j)} \int_0^{\lambda_i} u^{\alpha j - 1} e^{-u} du, \; if \quad j > 0, \tag{2.15}$$

where α is the dispersion parameter. If $\alpha < 1$, there is over-dispersion, if $\alpha > 1$ there is under-dispersion, and if $\alpha = 1$, there is equi-dispersion and the gamma count model collapses to the Poisson model.

The conditional mean function is given by:

$$E[y_i | X_i] = \sum_{i=1}^{\infty} j Gamma(\alpha j, \lambda_i) \tag{2.16}$$

The cumulative distribution function is given by:

$$\begin{aligned} F(T|\alpha, \lambda_i) &= \int_0^T \frac{\lambda_i^{\alpha j}}{\Gamma(\alpha j)} u^{\alpha j - 1} e^{-\lambda_i u} du, \alpha > 0, \lambda_i > 0 \\ &= \frac{1}{\Gamma(\alpha j)} \int_0^{\lambda_i T} u^{\alpha j - 1} e^{-u} du, \; j = 0, 1, \ldots \\ &= Gamma(\alpha j, \lambda_i T) \end{aligned} \tag{2.17}$$

Even though the gamma count model can provide a good fit for the crash data, its assumption has limited its applicability. The gamma count model assumes that observations are not independent where the observation for time t-1 would affect the observation for time t (Winkelmann, 1995; Cameron, 1998). This would become unrealistic if the time gap between the two observations is large. For instance, a crash that occurred at time t cannot directly influence another one that will occur six months after the first event.

2.4 The Conway-Maxwell-Poisson Model

In order to model queues and service rates, Conway and Maxwell (1962) first introduced the COM-Poisson distribution as a generation of the Poisson distribution. However, this distribution was not widely used until Shmueli et al. (2005) further examined its statistical and probabilistic properties. Kadane et al. (2006) developed the conjugate distributions for the parameters of the COM-Poisson distribution. The PMF of the COM-Poisson for the discrete count can be given by Equations (2.18) and (2.19):

$$P(Y=y)=\frac{1}{Z(\lambda,\nu)}\frac{\lambda^{y}}{(y!)^{\nu}} \tag{2.18}$$

$$Z(\lambda,\nu)=\sum_{n=0}^{\infty}\frac{\lambda^{n}}{(n!)^{\nu}} \tag{2.19}$$

For $\lambda>0$ and $\nu\geq 0$. Where y is a discrete count; λ is a centering parameter which is often approximately equal to the mean; ν is the shape parameter of the COM-Poisson distribution. The COM-Poisson distribution allows for both under-dispersed ($\nu>1$) and over-dispersed ($\nu<1$) data, and it is a generalization of some well-known distributions.

In the formulation, setting $\nu = 0$, $\lambda < 1$ yields the geometric distribution; $\nu \to \infty$ yields the Bernoulli distribution in the limit; and $\nu = 1$ yields the Poisson distribution. The flexibility of the COM-Poisson distribution greatly expands its use for count data.

The first two central moments of the COM-Poisson distribution are given by Equations (2.20) and (2.21):

$$E[Y] = \frac{\partial \log Z}{\partial \log \lambda} \tag{2.20}$$

$$Var[Y] = \frac{\partial^2 \log Z}{\partial \log^2 \lambda} \tag{2.21}$$

The COM-Poisson distribution does not have closed-form expressions for its moments in terms of the parameters λ and ν. The approximation of the mean can be achieved by different approaches including (i) using the mode, (ii) including only the first few terms of Z when ν is large, (iii) bounding E[Y] when ν is small, and (iv) using an asymptotic expression for Z in Equation (2.18). Using the last approach, Shmueli et al. (2005) derived the approximation in Equations (2.22) and (2.23).

$$E[Y] \approx \lambda^{1/\nu} + \frac{1}{2\nu} - \frac{1}{2} \tag{2.22}$$

$$Var[Y] \approx \frac{1}{\nu} \lambda^{1/\nu} \tag{2.23}$$

When ν is close to one, the centering parameter λ is approximately equal to the mean. When ν gets small, λ differs substantially form the mean. For the over-dispersed data, ν would be expected to be small and thus a COM-Poisson GLM based on the

original COM-Poisson formulation would be very difficult to interpret and use for the over-dispersed data.

In order to circumvent the problem, Guikema and Coffelt (2008) proposed a re-parameterization of the COM-Poisson distribution to provide a clear centering parameter. They substituted $\mu = \lambda^{1/\nu}$ and then the new formulation of the COM-Poisson distribution is summarized in Equations (2.24) and (2.25):

$$P(Y = y) = \frac{1}{S(\mu,\nu)}(\frac{\mu^y}{y!})^\nu \tag{2.24}$$

$$S(\mu,\nu) = \sum_{n=0}^{\infty}(\frac{\mu^n}{n!})^\nu \tag{2.25}$$

Correspondingly, the mean and variance of Y are given by Equations (2.26) and (2.27) in terms of the new information and the asymptotic approximations of the mean and variance of Y are given by Equations (2.28) and (2.29):

$$E[Y] = \frac{1}{\nu}\frac{\partial \log S}{\partial \log \mu} \tag{2.26}$$

$$V[Y] = \frac{1}{\nu^2}\frac{\partial^2 \log S}{\partial \log^2 \mu} \tag{2.27}$$

$$E[Y] \approx \mu + 1/2\nu - 1/2 \tag{2.28}$$

$$Var[Y] \approx \mu/\nu \tag{2.29}$$

The approximations are especially accurate once $\mu > 10$. This new parameterization makes the integral part of μ the mode and μ as a reasonable centering parameter. The substitution allows ν to keep its role as a shape parameter. That is, $\nu < 1$ leads to over-dispersion and $\nu > 1$ to under-dispersion.

Based on the new parameterization, Guikema and Coffelt (2008) developed a COM-Poisson GLM framework to model discrete count data using Bayesian framework in WinBUGS (Spiegelhalter, 2003). The modeling framework is shown in Equations (2.30) and (2.31). It should be noted that the model framework is a dual-link GLM in which both the mean and variance depend on the covariates.

$$ln(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{2.30}$$

$$ln(\nu) = \gamma_0 + \sum_{j=1}^{q} \gamma_j z_j \tag{2.31}$$

The established GLM framework can handle under- and over-dispersed datasets, as well as datasets that contain intermingled under- and over-dispersed counts (only for dual-link models because the dispersion characteristic is captured using the covariate-dependent shape parameter). In the dual-link GLM, the variance can vary with the covariate values, which is especially useful when high values of some covariates tend to be variance-decreasing and low values of other covariates tend to be variance-increasing or vice versa. It should be noted that parameter estimation for the dual-link GLM is complex and difficult.

With the derivation of the likelihood function of the COM-Poisson GLM by Sellers and Shmueli (2010), the maximum likelihood estimation (MLE) of the parameters of a COM-Poisson GLM was greatly simplified compared with the Bayesian estimating method. The MLE formulation did not allow for a varying shape parameter. The MLE codes in R for the COM-Poisson GLM could be found here: http://cran.r-project.org/web/packages/compoisson/index.html (R Development Core Team, 2006).

Geedipally (2008) examined the performance of the COM-Poisson GLM in the context of single link.

2.5 The Double Poisson Model

Based on the double exponential family, Efron (1986) proposed the double Poisson distribution. The double Poisson model has two parameters μ and θ, with its approximate probability mass function given as:

$$P(Y=y)=f_{\mu,\theta}(y)=(\theta^{1/2}e^{-\theta\mu})(\frac{e^{-y}y^{y}}{y!})(\frac{e\mu}{y})^{\theta y}, y=0,1,2,..., \quad (2.32)$$

The exact double Poisson density is given as:

$$P(Y=y)=f_{\mu,\theta}(y)=c(\mu,\theta)\,f_{\mu,\theta}(y) \quad (2.33)$$

where the factor $c(\mu,\theta)$ can be calculated as:

$$\frac{1}{c(\mu,\theta)}=\sum_{y=0}^{\infty}f_{\mu,\theta}(y)\approx 1+\frac{1-\theta}{12\mu\theta}(1+\frac{1}{\mu\theta}) \quad (2.34)$$

With $c(\mu,\theta)$ which is a normalizing constant nearly equal to 1. The constant $c(\mu,\theta)$ ensures that the density sums to unity. The expected value and the standard deviation (SD) referring to the exact density $f_{\mu,\theta}(y)$ are:

$$E(Y)\approx\mu, \quad (2.35)$$

$$SD(Y)\approx(\frac{\mu}{\theta})^{1/2} \quad (2.36)$$

Thus, the double Poisson model allows for both over-dispersion ($\theta<1$) and under-dispersion ($\theta>1$). When $\theta=1$, the double Poisson distribution collapses to the Poisson distribution.

Based on the approximate probability mass function, i.e. Equation (2.32), the maximum likelihood estimation (MLE) for μ and θ is given as:

$$\mu=\bar{y}=\frac{\sum_{y=0}^{\infty} n_y \times y}{\sum_{y=0}^{\infty} n_y} \tag{2.37}$$

$$\theta=\frac{1}{2(\frac{\sum_{y=0}^{\infty} n_y \times y \times \ln(y)}{\sum_{y=0}^{\infty} n_y}-\bar{y}\times\ln[\bar{y}])} \tag{2.38}$$

where n_y denotes the observed frequency of count equal to y.

It should be noted that the MLE for θ does not seem to be applicable when $y=0$ due to the presence of $\ln(y)$ in Equation (2.38). However, the limit of $y\times\ln(y)$ approaches 0 when y is getting close to 0, thus $n_y \times y \times \ln(y)=0$ approximately equals 0.

For the DP GLM, the expected number of crashes per year μ_i is linked to the explanatory variables x_i by the following link function (similar to the traditional Poisson):

$$\mu_i=\exp(x_i\beta) \tag{2.39}$$

where the vector β is the coefficients to be estimated.

A disadvantage of the DP distribution is that its results are not exact since the normalizing constant $c(\mu,\theta)$ has no closed form solution (Winkelmann, 2008; Hilbe,

2011). Considering the inclusion of the normalizing constant would substantially increase the non-linearity of the PMF which makes the MLE is difficult to achieve, all the DP GLMs in this thesis are developed based on the PMF without the NC. More discussions on the use of the normalizing constant could be found in Section 4.5.3.

2.6 Other Models

Apart from the aforementioned models, researchers have introduced other statistical count models for analyzing vehicle crash data. These models include: the zero-inflated model (Shankar et al, 1997; Carson and Mannering, 2001; Qin et al, 2004), Poisson-lognormal model (Miaou et al., 2003; Lord and Miranda-Moreno, 2008), Bayesian neural networks (Abdelwahab and Abdel-Aty, 2002; Xie et al., 2007), latent class or mixture model (Depaire et al., 2008; Park and Lord, 2008), support vector machine model (Li et al., 2008), multivariate models (Tunaru, 2002; Park and Lord, 2007), etc.

It should be noted that the zero-inflated model is a dual-state model and its zero state cannot appropriately reflect the actual crash-data generating process (Lord et al., 2005; Wedagama et al., 2006; Ma et al, 2008). Other aforementioned models are complex and most of them do not have a closed form, which causes difficulty in estimating parameters.

2.7 Summary

This section has provided a brief overview on a variety of statistical models that have been proposed to model traffic crash data. The NB has been the most popularly used model due to the wide presence of over-dispersed crash data. However, most models such as the NB have difficulty in handling the crash data characterized by under-dispersion. The models proposed to handle the under-dispersed data were mainly introduced in this section. The focus of this section was to present the statistical properties and GLM frameworks of two models, the DP model and COM-Poisson model, both of which can handle over-, equi- and under-dispersed count data. The limitations of the commonly used models were also discussed in this section.

Since the DP model has seldom been investigated and applied after its introduction 25 years ago, it is of great interest to examine the applicability of the DP distribution and its regression model for analyzing crash data. Meanwhile, there is also a need to compare its performances with those of the COM-Poisson model and other models that can handle either over- or under-dispersed count data. Thus, the following sections provide the results on the detailed comparisons between the DP and other models in handling simulated count data (Section 3) as well as observed crash data characterized by over-dispersion (Section 4) and under-dispersion (Section 5).

3. PERFORMANCE OF THE DOUBLE-POISSON DISTRIBUTION

Of all the available distributions that have been proposed in the literature, two distributions that can handle both over- and under-dispersion are of interest. They are the COM-Poisson (Conway and Maxwell, 1962; Shmueli et al., 2005; Kadane et al., 2006) and DP distributions (Efron, 1986) (note: the distribution proposed by Efron should not to be confused with the Double Poisson model documented in Lao et al. (2011)). The properties of the COM-Poisson have been investigated extensively and several researchers have found that both the distribution and regression model are very flexible to handle count data (Sellers et al., 2011; Francis et al., 2012). On the other hand, although the DP has been introduced over 25 years ago, this distribution has never been fully investigated. In fact, very few researchers have applied or used the DP distribution or model for analyzing count data since its introduction.

The primary objective of this section is to examine the potential applicability of the DP distribution for analyzing count data characterized by both over- and under-dispersion. The study objective was accomplished using simulated data for nine different mean-variance relationships (or scenarios). Before tackling the performance of the regression model, it is important to first evaluate the performance of the distribution, similar to how other new distributions have first been investigated in the past (Shmueli et al., 2005; Lord and Geedipally, 2011). This section focuses on the distribution only and covariates will not be considered. The DP distribution was compared with the COM-Poisson distribution using various GOF statistics. Although the gamma count model is

technically not adequate, the DP distribution was also compared with this distribution for the under-dispersed simulated datasets. For over-dispersion, the DP distribution was compared with the NB distribution.

3.1 Simulation Protocol

In order to compare the general performance of different distributions before the development of GLMs, simulated data were first generated due to its flexibility to control the mean and dispersion level.

Nine scenarios were examined for three sample mean levels (high, medium and low) and three levels of dispersion (under-, equi- and over-dispersion). The discrete count data were initially simulated using the COM-Poisson distribution, since this distribution has already been shown to handle under-, equi- and over-dispersion. To examine potential bias with using only one distribution to simulate data, counts were also simulated using the traditional Poisson and NB distributions for the equi-dispersion and over-dispersion respectively. A total of 2,000 observations were simulated for each scenario. The three mean values were obtained by setting μ = 0.5, 1, and 5 (recall that $\mu = \lambda^{1/\nu}$ in the COM-Poisson; μ is also defined as the mode). The levels of dispersion were: ν = 1.3, 1 and 0.5 representing under-, equi- and over-dispersion, respectively. Corresponding input values of the Poisson and NB parameters were set to get the similar simulated data characteristics (i.e., the mean and variance/mean ratio) with that of the COM-Poisson.

For each scenario, different distributions were fitted based on their characteristics of handling dispersion. All scenarios were fitted using the DP and COM-Poisson distributions. The gamma count, Poisson and NB distributions were only employed to fit the under-dispersed data, equi-dispersed data and over-dispersed data, respectively. Recall that the gamma count is technically a distribution that is not adequate for crash data analysis, since crash data rarely influence each other directly at different time periods. For each of the aforementioned scenarios, five simulation runs were conducted. The GOF measures for each run were computed and then the average GOF values for all five runs.

3.2 Parameter Estimation

In order to fit the double Poisson distribution, parameters were first estimated based on the observed frequency for each count using Equations (2.37) and (2.38). Then, the approximated predicted probabilities and frequencies were calculated for each count using Equation (2.32). After considering the normalizing constant documented in Equations (2.33) and (2.34), the exact predicted probability and frequency for each count were calculated.

For the COM-Poisson distribution, the estimated parameters can be calculated according to the mean and variance of the data with Equations (2.22) and (2.23). However, the mean and variance are just the approximations and will not provide the proper estimates. Thus, the MCMC implementation of the COM-Poisson GLM proposed

by Guikema and Coffelt (2008) in MATLAB (2011) was used for the parameter estimation and likelihood calculation.

Since there are no closed forms for the expected value and variance of gamma count distribution, the software LIMDEP 8.0 was used to obtain the predicted likelihood for each count (Greene, 2002). The 'gamma probabilities' under the 'Poisson' command in LIMDEP can be used to fit the given count data. The NB distribution was assessed using the well-known method documented in various textbooks (Cameron and Trivedi, 1998).

3.3 Goodness-of-fit

Different methods were used to assess the GOF of the distributions. They include: the Pearson's Chi-squared test, the likelihood ratio test and the log-likelihood value. Like the Pearson's Chi-squared statistic (Chi-Sq), the likelihood Ratio statistic (LR) has approximately a Chi-squared distribution and the null hypothesis is rejected for a reasonable fit for large values of likelihood ratio statistic. The log-likelihood statistic (LogL) was calculated by taking the logarithm of the estimated likelihood for each observation. The sum of those log-likelihoods was then obtained for comparing those different distributions.

Besides, given that the degree of freedom (DF) for different distributions might differ within the same scenario, the value of Chi-Sq divided by DF (Chi-Sq/DF) was also provided as an alternative for those three GOFs. The smaller the Chi-Sq/DF, the better the fit. Those GOF statistics are given as:

$$Chi-Sq=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i} \tag{3.1}$$

$$LR=2\sum_{i=1}^{n}O_i*Log(\frac{O_i}{E_i}) \tag{3.2}$$

$$LogL=\sum_{i=1}^{n}Log(P_i) \tag{3.3}$$

$$Chi-Sq/DF=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i*DF} \tag{3.4}$$

$$DF=n-(p+1) \tag{3.5}$$

where,

O_i is the observed frequency for the category of count equal to i;

E_i is the expected frequency for the category of count equal to i;

P_i is the expected likelihood for the category of count equal to i;

n is the number of total categories;

p is the number of parameters used in fitting the distribution.

3.4 Comparison of Results

Nine scenarios of simulated data with three means (high, medium and low) and three levels of dispersion (under-, equi-, and over- dispersion) were examined in this study. Comparisons on GOFs of simulated data fitted by the DP and COM-Poisson distributions were conducted. The GOFs of simulated data fitted by other distributions such as NB, gamma and Poisson were also be given as a reference. The results were

presented by the level of dispersion: under-, equi- and over-dispersion. GOFs for each run as well as the average on all five runs were included.

3.4.1 Under-dispersion

All the under-dispersed data were simulated under the COM-Poisson distribution. Tables A.1 to A.3 in Appendix show the results for under-dispersed simulated data for the high, medium and low sample means, respectively. In each table, all five runs show consistent comparison results. The three tables show that the COM-Poisson and gamma count distributions provide better fit than that for the DP distribution. Since the estimated parameter is the mode of the COM-Poisson, this may not always be equal to the sample mean. This characteristic nonetheless does not directly affect the GOF analyses. Additional information about this characteristic can be found in Lord et al. (2008a).

Table 3.1 summarizes the GOF statistics of the averaged five run values for all the under-dispersion scenarios using COM-Poisson simulated data. In terms of the ratio Chi-Sq/DF, the DP distribution seems to provide a good fit, but only when the mean is high. The difference in fit is larger for the Chi-Sq and LR than for the LogL. It is interesting to note that the gamma count distribution works better than the DP distribution for under-dispersion.

Table 3.1 Summary of GOFs for under-dispersion (COM-Poisson simulated data)

Mean Type	Distributions	GOF			
		Chi-Sq	LR	LogL	Chi-Sq/DF
High	DP	8.8	9.2	-4165.2	0.98
	COM-P	7.6	7.7	-4164.4	0.85
	Gamma	7.3	7.4	-4164.3	0.81
Medium	DP	23.0	24.7	-3186.0	4.99
	COM-P	3.7	3.7	-3175.4	0.92
	Gamma	3.7	3.7	-3175.4	0.93
Low	DP	3.6	3.4	-1585.9	3.58
	COM-P	1.1	1.1	-1584.7	1.12
	Gamma	1.0	1.0	-1584.6	1.03

3.4.2 Equi-dispersion

Two distributions, the COM-Poisson and traditional Poisson were used to generate the equi-dispersed data. Tables A.4 to A.6 in Appendix tabulate the results for the equi-dispersed COM-Poisson simulated data for the high, medium and low sample means based on each run, respectively. Table 3.2 summarizes the GOF statistics averaged on the five runs for all the equi-dispersion scenarios using the COM-Poisson simulated data. Likewise, Tables A.7 to A.9 in Appendix tabulate the results for the Poisson simulated data for each run and Table 3.3 summarizes the GOF statistics averaged on the five runs for all equi-dispersion scenarios.

As can be seen from Tables 3.2 and 3.3, the COM-Poisson simulated data and Poisson simulated data give similar comparison results. The COM-Poisson and Poisson provides a good fit, while the DP is not as good as the other two. Comparing the sample

mean values, the DP works better for the high sample mean. Although the values of Chi-Sq, LR and LogL for the COM-Poisson are smaller than those for the Poisson, we cannot arbitrarily conclude that the COM-Poisson is better than Poisson. Rather, when one needs to take into account the number of estimated parameters, which show the Poisson to be very close to the COM-Poisson. The reason the Poisson not the best distribution overall is explained by the fact that the mean and variance are not exactly equal for all three simulated datasets.

Table 3.2 Summary of GOFs for equi-dispersion (COM-Poisson simulated data)

Mean Type	Distributions	GOF			
		Chi-Sq	LR	LogL	Chi-Sq/DF
High	DP	11.9	12.4	-4415.7	1.14
	COM-P	9.6	9.6	-4414.4	0.96
	Poisson	10.3	10.3	-4414.7	0.93
Medium	DP	22.3	23.4	-3440.1	4.29
	COM-P	6.5	6.5	-3431.6	1.31
	Poisson	8.9	8.8	-3432.6	1.49
Low	DP	1.3	1.3	-1863.5	1.33
	COM-P	0.9	0.9	-1863.3	0.87
	Poisson	1.7	1.7	-1863.8	0.86

Table 3.3 Summary of GOFs for equi-dispersion (Poisson simulated data)

Mean Type	Distributions	Goodness-of-Fit			
		Chi-Sq	**LR**	**LogL**	**Chi-Sq/DF**
High	DP	10.8	11.4	-4416.6	1.05
	COM-P	9.1	9.2	-4415.4	0.89
	Poisson	10.3	10.3	-4416.0	0.94
Medium	DP	20.3	21.6	-3430.6	4.05
	COM-P	4.4	4.5	-3422.0	0.89
	Poisson	6.3	6.4	-3422.9	1.06
Low	DP	0.6	0.6	-1837.9	0.62
	COM-P	0.5	0.5	-1837.9	0.53
	Poisson	0.8	0.8	-1838.0	0.39

3.4.3 Over-dispersion

Two distributions, the COM-Poisson and NB, were used to generate the over-dispersed data. Tables A.10 to A.12 in Appendix tabulate the results for the over-dispersed COM-Poisson simulated data for the high, medium and low sample means based on each run, respectively. Table 3.4 summarizes the GOF statistics averaged on the five runs for all the over-dispersion scenarios using the COM-Poisson simulated data. Likewise, Tables A.13 to A.15 in Appendix tabulate the results for the NB simulated data for each run and Table 3.5 summarizes the GOF statistics averaged on the five runs for all over-dispersion scenarios.

As can be seen from Tables 3.4 and 3.5, the COM-Poisson simulated data and NB simulated data give similar comparison results. The COM-Poisson and NB provide a good fit for all mean values, while the DP is not as good for the medium mean and low

sample mean values, especially when fitting the NB simulated data. For the high sample mean, the DP provides a good fit.

Table 3.4 Summary of GOFs for over-dispersion (COM-Poisson simulated data)

Mean Type	Distributions	GOF			
		Chi-Sq	LR	LogL	Chi-Sq/DF
High	DP	25.1	26.9	-5041.9	1.77
	COM-P	14.7	14.9	-5021.1	1.05
	NB	27.4	27.2	-5027.0	1.88
Medium	DP	18.7	19.3	-4008.6	2.34
	COM-P	7.8	8.0	-4003.0	0.98
	NB	13.9	14.4	-4006.0	1.62
Low	DP	6.4	6.5	-2638.3	2.13
	COM-P	3.1	3.1	-2635.4	0.86
	NB	2.9	2.9	-2635.1	0.72

Table 3.5 Summary of GOFs for over-dispersion (NB simulated data)

Mean Type	Distributions	GOF			
		Chi-Sq	LR	LogL	Chi-Sq/DF
High	DP	31.5	33.6	-4807.3	2.46
	COM-P	18.7	17.7	-4799.4	1.48
	NB	9.9	9.7	-4795.8	0.76
Medium	DP	23.4	23.7	-3616.1	3.90
	COM-P	9.4	9.3	-3682.8	1.47
	NB	9.3	9.4	-3608.7	1.46
Low	DP	13.9	13.2	-1906.3	8.69
	COM-P	3.3	3.4	-1899.8	1.67
	NB	2.7	2.7	-1899.4	1.35

3.5 Discussion

For all nine scenarios, the COM-Poisson performs better than the DP. The DP has been shown to provide a better fit when the mean is high for all types of dispersion. It should be noted that the COM-Poisson may be expected to be better than the DP in fitting COM-Poisson simulated data. The primary reason why the DP works better for high sample mean values is related to the observations that are equal to zero. In calculating the values of Chi-Sq and LR, all the observations are grouped into several categories, and the final values of Chi-Sq and LR are aggregated based on the value of the Chi-Sq and LR for each of those categories. In this study, it was found that very often the category for observations equal to zero had exceptionally large Chi-Sq and LR values compared to other categories. This artificially increases the total or final Chi-Sq and LR values, indicating a poorer fit. When the mean increases, the total Chi-Sq and LR values get less affected since the proportion of zeros becomes smaller.

The hypothesis as to why DP cannot provide a good fit when the observations equal to zero might be related to the approach used for calculating the likelihood. In the approximate PMF of Efron's DP distribution (see Equation (2.32)), the denominator is zero for observations equal to zero, which is not solvable. To circumvent this problem, the author calculated the limits of the likelihood when observation value approached zero in writing the thesis. The validity and accuracy of this approach might need to be further examined.

Overall, the differences observed in statistical fit between the DP and COM-Poisson distributions were not enormous, especially when you compare the differences

in fit between the NB and the recently introduced Negative-Binomial-Lindley distribution used for analyzing crash data characterized by a large amount of zeros (Lord and Geedipally, 2011; Geedipally et al. 2011). The latter comparison shows a wider difference between the two distributions (NB and NB-L) and the gap increases as the data become more dispersed. The fact that the DP is not clearly superior to existing distributions, such as the NB distribution, probably explains why it has not been used extensively by researchers and practitioners.

Although the COM-Poisson fits all the data much better than the DP, the comparison on their performance of handling under-dispersed data is yet to be determined since all the under-dispersed data in this section were simulated by the COM-Poisson distribution and the COM-Poisson may be expected to generate better results than other distributions. Thus, it is of great interest to examine the GLMs, particularly in terms of their performance of handling under-dispersion. Besides, the DP GLM has already been developed by the original author who developed this distribution (Efron, 1986) and it is possible to examine its stability in the context of a regression model.

3.6 Summary

The primary objective of this section was to examine the potential applicability of the DP distribution for analyzing count data characterized by both over- and under-dispersion. The study objective was accomplished using simulated data for nine different mean-dispersion relationships (or scenarios). Five runs each with 2,000 observations

were simulated under each scenario by different distributions. The under-dispersed data were simulated by the COM-Poisson distribution; the equi-dispersed data were simulated by the COM-Poisson and Poisson distributions; and the over-dispersed data were simulated by the COM-Poisson and NB distributions.

For each scenario, different distributions were fitted based on their characteristics of handling dispersion. All scenarios of data were fitted using the DP and COM-Poisson. The gamma count model, Poisson and NB were only employed to fit under-dispersed data, equi-dispersed data and over-dispersed data, respectively. Four different GOF statistics were used to evaluate and compare the performances of different distributions.

The simulation results showed that the COM-Poisson performs better than the DP for all nine scenarios, and that the DP works better for high mean scenarios independent of the type of dispersion. The lack of fit for the DP in low mean scenarios is due to its inadequacy for fitting "zero" observations. It should be noted that the comparison on their performance of handling under-dispersed count data is yet to be determined since all the under-dispersed data in this section was simulated by the COM-Poisson distribution and the COM-Poisson may be expected to generate better results than other distributions. Thus, it is of great interest to examine the DP GLMs, particularly in terms of their performance of handling under-dispersion. Next section will investigate the applicability of the DP GLMs in handling over-dispersed data.

4. APPLICATION OF THE DOUBLE POISSON GLM TO CRASH DATA CHARACTERIZED BY OVER-DISPERSION

Over-dispersion is a commonly seen characteristic in traffic crash data. Its presence has made the traditional Poisson distribution unable to handle traffic crash data in most cases. Unlike the Poisson, the NB distribution allows its variance varying independent of the mean by including a dispersion parameter ($Var(Y) = \mu + \alpha\mu^2$), which has made the NB the most widely applied distribution in handling crash data. Recently, the COM-Poisson distribution has been examined in terms of its capability of handling over-dispersed and under-dispersed data (Shmueli et al., 2005; Kadane et al., 2006; Guikema and Coffelt, 2008; Geedipally, 2008). It has been found that the COM-Poisson GLMs can fit the over-dispersed data as well as the NB models (Guikema and Coffelt, 2008; Lord et al., 2008a).

The objectives of this section are to evaluate the application of the DP GLMs for analyzing motor vehicle crash data characterized by over-dispersion and conduct the comparison analysis between the DP GLMs, NB models and COM-Poisson GLMs. Although the results in Section 3 have showed that the DP distribution can handle over-, equi-, and under-dispersed count data, it is of more interest to see how good the DP GLM can link the crash data to the variables that affect traffic safety and how much influence those variables can affect the expected crash data. In this section, all the results of the DP GLMs are compared with those of the NB models. Two over-dispersed datasets along with two different and commonly used link functions are examined in

order to eliminate the potential bias of using only one dataset or one link function. The performance of the COM-Poisson GLMs is also given as a reference in the second dataset. All the DP GLMs in this thesis are developed based on the PMF without the normalizing constant. At the end of this section, a discussion is provided about whether or not the inclusion of the normalizing constant in the DP GLM improves the performance of the model.

4.1 Data Description

In this section, two over-dispersed datasets were used to examine the performance of the DP GLMs and compare those with other models. It should be noted that the two datasets were collected for different transportation elements: one for the intersection (crashes occurred within 250ft of the intersection center) and the other for the roadway segment (crashes occurred beyond the 250ft of the intersection center). Given that the variables that contribute to the occurrences in intersection related crashes and non-intersection related crashes are different (Highway Safety Manual – AASHTO, 2010), two link functions were employed to conduct the GLM analysis for the two datasets respectively.

The first dataset (Texas segment data or Texas data) recorded crashes that occurred on 4-lane rural undivided and divided roadway segments for five years (from 1997 to 2001) in Texas. The dataset has been used to develop the statistical models and accident modification factors in the project NCHRP 17-29 (Lord et al., 2008b). The dataset were provided by the Texas Department of Public Safety (DPS) and the Texas

Department of Transportation (TxDOT). Only undivided segments crash data were used in this study. Other than the crash number records classified by the year and severity level, information about variables such as annual average daily traffic (AADT), shoulder width, number of intersections was also reported along those segments. Table 4.1 shows the summary statistics for the key variables in this dataset.

Table 4.1 Summary statistics of variables for the Texas data

Variables		Min.	Max.	Mean	Std Err	Obs
Response	KABCO [a] Crashes for Five Years	0	97	2.84	5.69	1499
	KAB [b] Crashes for Five Years	0	19	0.63	1.60	1499
Offset	Segment Length (Miles)	0.1	6.275	0.55	0.67	1499
Variables	AADT (veh/day)	402	24800	6613.61	4010.01	1499
	Lane Width (ft)	9.75	16.5	12.57	1.59	1499
	Shoulder Width (Right + Left) (ft)	0	40	9.96	8.02	1499
	Right Shoulder Width (ft)	0	24	13.65	3.65	1499
	Median Width (ft)	1	240	47.71	28.87	1499
	Number of Intersections	0	47	2.33	2.62	1499
	Number of Horizontal Curves	0	16	0.70	1.32	1499

[a] KABCO crashes: crashes with severity level of fatality, injury type A, injury type B, injury type C and property damage only

[b] KAB crashes: crashes with severity level of fatality, injury type A, injury type B

The second over-dispersed dataset (Toronto intersection data or Toronto data) were collected on urban 4-legged signalized intersections in Ontario, Toronto from the year 1990 to 1995. A total number of 54, 869 crashes that occur on 868 intersections were reported. The dataset were found to be of good quality and then applied to many

studies (Lord, 2000; Miaou and Lord, 2003). The crash data for the year 1995 with a total number of 10,030 crashes were used in this study. Traffic flow data from both major approaches (Major AADT) and minor approaches (Minor AADT) were recorded in the crash report. Table 4.2 shows the summary statistics for the key variables in this dataset.

Table 4.2 Summary statistics of variables for the Toronto data

Variables		Min.	Max.	Mean	Std Err	Obs
Response	Crashes	0	54	11.56	10.02	868
Variables	AADT for Major Approach (veh/day)	5469	72178	28044.8	10660.4	868
	AADT for Minor Approach (veh/day)	53	42644	11010.2	8599.4	868

4.2 Link Function

For the Texas data, the GLM frameworks are developed for the DP and NB models. Although Lord et al. (2008a) established the COM-Poisson GLM with the same dataset, its link function is not the same as that in this study and thus the results of the COM-Poisson GLM are not included for this dataset. In this study, the response variable is the mean of KABCO crashes (crashes with severity level of fatality, injury type A, injury type B, injury type C and property damage only) or KAB crashes (crashes with severity level of fatality, injury type A, injury type B). The unit for both response variables is in crashes per year. It should be noted that the length of the segment and the year number 5 are handled as the offset term. The link function for the two models is the same as that in the project NCHRP 17-29 (Lord et al., 2008b):

$$\mu_i = \beta_0 L_i F_i^{\beta_1} e^{\{\beta_2 LW_i + \beta_3 SW_i + \beta_4 CURVE\_DEN_i\}} \quad (4.1)$$

where,

μ_i = the mean number of KABCO crashes or KAB crashes per year for segment i;

L_i = length of the segment in miles for segment i;

F_i = the traffic flow volume (or AADT) in veh/day for segment i;

LW_i = lane width in ft for segment i,

SW_i = total shoulder width (both sides) in ft for segment i;

$CURVE\_DEN_i$ = number of horizontal curves per mile located on the segment i,

β_0 β_1 β_2 β_3 β_4 = coefficients

For the other dataset, the Toronto data, the GLM frameworks are established for the DP, NB and COM-Poisson models. It should be noted that the response variable for the DP GLM and NB model is different than that for the COM-Poisson GLM. The response variable μ_i for the former two models is the mean of crashes per year (see Equations (2.35) and (2.39) for the DP model, and see Equations (2.10) and (2.12) for the NB model), whereas the response variable μ_i for the COM-Poisson model is the mode of crashes per year (see Equation (2.30)). The link function is:

$$\mu_i = \beta_0 F_{Maj\_i}^{\beta_1} F_{Min\_i}^{\beta_2} \quad (4.2)$$

where,

μ_i = the mean (for the NB model and DP GLM) or the mode (for the COM-Poisson GLM) of crashes per year for intersection i;

F_{Maj_i} = entering flow for the major approach (or AADT) for intersection i ;

F_{Min_i} = entering flow for the minor approach (or AADT) for intersection i ;

$\beta_0, \beta_1, \beta_2$ = coefficients.

Since the link function for the COM-Poisson GLM is established on the mode not on the mean as the case with the DP GLM and NB model, the coefficients of the COM-Poisson GLM cannot be directly compared to the other two models. However, the mean and variance for the COM-Poisson GLM can be obtained (see Equations (2.26) and (2.27)) or approximated (see Equations (2.28) and (2.29)) according to their relationship to the mode μ_i and shape parameter v, making the direct comparisons on GOF statistics of the three models comparable.

Instead of calculating the predicted mean and variance of the COM-Poisson models according those equations, the researcher simulated 100,000 samples with the estimated μ_i and v obtained from the COM-Poisson GLMs and then took the mean and variance of those samples as the predicted mean and variance in this study.

4.3 Goodness-of-fit

The GOF statistics of the DP GLM and COM-Poisson GLM will be compared for both under- and over-dispersed data as the following:

- Akaike Information Criterion (AIC)

  As a measure of GOF considering the influence of parameters for estimated models, AIC is defined as:

  $$AIC = -2\log L + 2p \tag{4.3}$$

  where L is the maximized value of the likelihood function for the estimated model, and p is the number of parameters in the statistical model. By penalizing models with a large number of parameters, the AIC attempts to select the model that best explains the data with a minimum of parameters. Lower the AIC, better the model.

- Mean Prediction Bias (MPB)

  MPB is used to measure the magnitude and direction of the average model bias. MPB is calculated using the following equation:

  $$MPB = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{4.4}$$

  where n is the sample size, $\hat{y}_i$ and y_i are the predicted and observed crashes at site i respectively. When the model over-predicts crashes, MPB is positive and when the model under-predicts crashes, MPB is negative.

- Mean Absolute Deviance (MAD)

  MAD is the average of the absolute deviations and it measures the average mis-prediction of the model. The model closest to 0 is considered to be the best. It is computed by the following equation:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{4.5}$$

- Mean Squared Predictive Error (MSPE)

  MSPE is often used to access the error associated with a validation or external data set. The model closest to 0 is considered to be the best. It can be calculated by the following equation:

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{4.6}$$

4.4 Parameter Estimation Method

The parameters of the DP GLMs were estimated using the NLMIXED procedure in SAS (SAS Institute Inc., 2002). The NLMIXED procedure was designed to fit nonlinear mixed models, that is, models with nonlinear random and fixed effects. The statement PROC NLMIXED offers an interface for specifying and coding a user-defined conditional distribution. PROC NLMIXED provides a variety of optimization techniques to maximize an approximation to the likelihood.

In this study, the approximate PMF of the DP distribution and its GLM link function were coded using SAS programing statements. The default quasi-Newton

algorithm was applied to obtain estimated parameters along with their approximate standard errors. Although NLMIXED procedure was often intended for mixed effects model, in this study it was used to fit models with only fixed effects. It should be noted that for the COM-Poisson GLM, its coefficients were estimated using the Bayesian framework in WinBUGS (Spiegelhalter et al., 2003). For the NB model, the GENMOD procedure in SAS was used to obtain the estimated parameters using MLE.

4.5 Comparison Results

The comparison results between the three models for the Texas data and Toronto data will be presented. The comparison conducted between the DP GLM with and without the normalizing constant will also be given at the end of this subsection.

4.5.1 Texas data

Although there are a variety of models that can handle data characterized by over-dispersion, the NB is the most widely applied model to establish the link between crashes and covariates in traffic crash data analysis. And the mathematics to manipulate the relationship between the mean and the variance structures for the NB model is very simple (Hauer, 1997). So in this subsection, the DP GLMs was first compared with the NB models in terms of GOF statistics. All those conclusions are based on the significance level as 0.1.

Table 4.3 summarizes the results of the DP GLMs and NB models in fitting the Texas data. As shown in the table, the NB models and DP GLMs established under the

two response scenarios (KABCO crashes as the response variable and KAB crashes as the response variable) generate consistent comparison results between the two models. The shape parameters of the DP GLMs in both scenarios are significantly less than 1, indicating the presence of the over-dispersion.

Table 4.3 Comparison of results between DP GLMs and NB models using the Texas data

Estimated Parameters and Standard Errors				
Variables	**KABCO as response variable**		**KAB as response variable**	
	NB[a]	**DP**	**NB**	**DP**
Intercept ($Ln(\beta_0)$)	-7.9488 (0.4060)[b]	-7.8341 (0.3713)	-6.8242 (0.5470)	-6.8743 (0.4793)
F (β_1)	0.9749 (0.0440)	0.9773 (0.0400)	0.7768 (0.0580)	0.8021 (0.0507)
LW (β_2)	-0.0533 (0.0170)	-0.06497 (0.0146)	-0.0844 (0.0230)	-0.0957 (0.0199)
SW (β_3)	-0.0100 (0.0030)	-0.01049 (0.0029)	-0.0114 (0.0050)	-0.0125 (0.0038)
CURVE_DEN (β_4)	0.0675 (0.0120)	0.09291 (0.0106)	0.0635 (0.0160)	0.0876 (0.0140)
Dispersion Parameter	0.3906 (0.0360)	0.5099 (0.0186)	0.3793 (0.0570)	0.8204 (0.0300)
Goodness-of-fit Statistics				
AIC	5134.772	5266.1	3198.728	3324.7
MAD	1.702	1.714	0.826	0.835
MSPE	11.236	10.631	2.727	2.559

[a] The results for NB models are directly taken from the project NCHRP 17-29 (Lord et al., 2008b).

[b] Values in parentheses are the standard errors for the estimated parameters.

As for the GOF statistics in both scenarios shown in Table 4.3, the values of MSPE for the DP GLMs are slightly less than those for the NB models whereas the values of AIC and MAD for the DP GLMs are bigger than those for the NB models. So

we can infer that the DP GLMs fits the data almost the same or slightly worse than the NB models especially considering the difference in GOF statistics between the two models is not pronounced. It should be noted that even though the values of coefficients are very similar for the two models, the standard errors of those coefficients for the DP GLMs are always smaller than those for the NB models in the two response scenarios.

Figure 4.1 presents the distributions of crash numbers for the observed and predicted crashes in two response scenarios. Most sites have KABCO and KAB crashes less than 5. The observed crashes are more scattered than the predicted crashes.

Figure 4.2 shows the scatter plots for the predicted vs. observed crashes in the two response scenarios. It can be seen from both scenarios that all the data points are quite evenly scattered along the reference line Y=X, and that both the DP GLMs and NB models can give an equally reasonably good fit to the data. It is interesting to note that compared with the NB models, the data points for the DP GLMs are more closer to the reference line when observed crashes is larger than 10 in KABCO crashes and larger than 5 in KAB crashes, which indicates that the DP GLMs provide more accuracy in predicting larger crash numbers than the NB models.

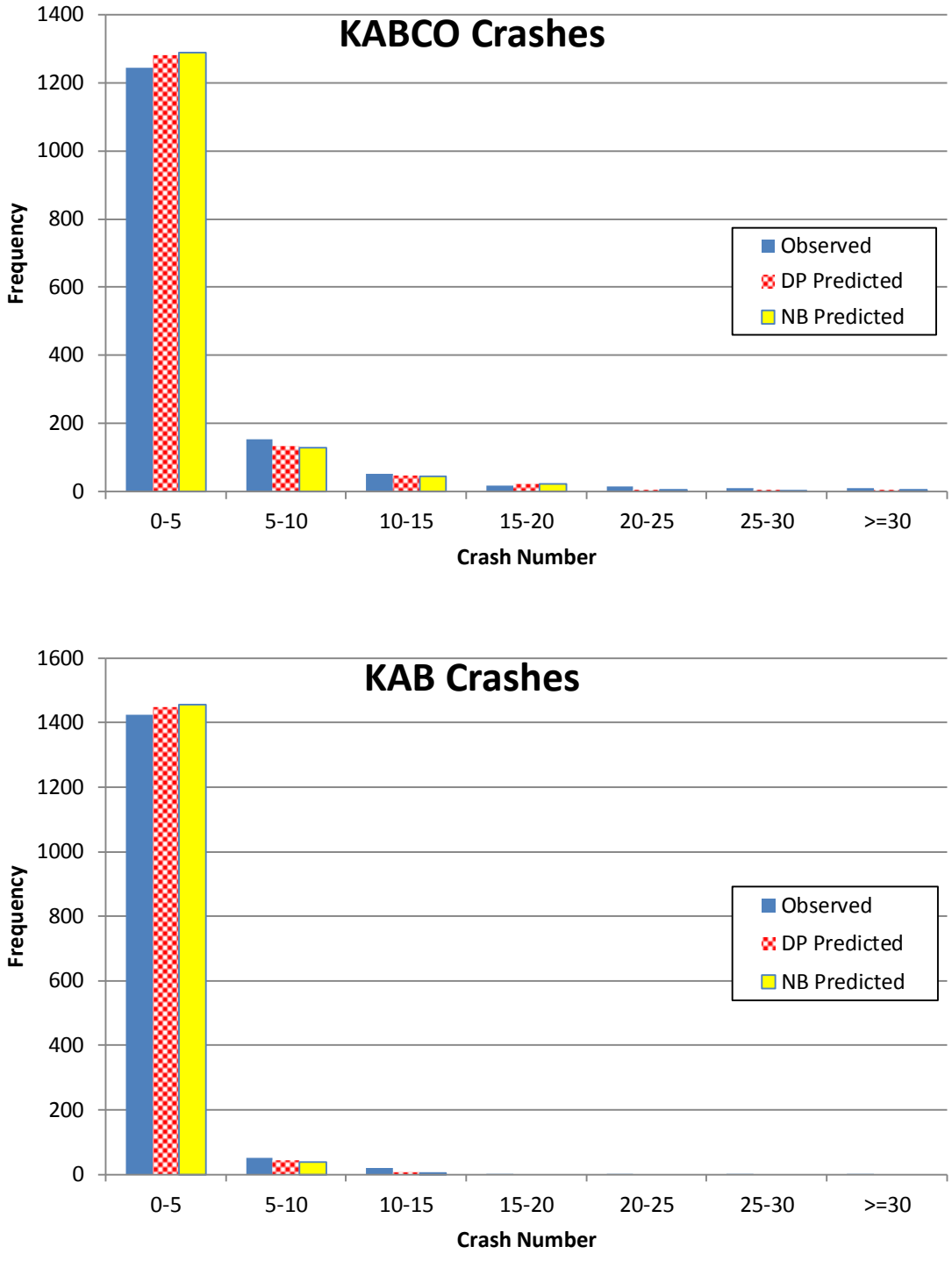


Figure 4.1 Frequencies of observed and predicted crashes for the Texas data (KABCO and KAB crashes)

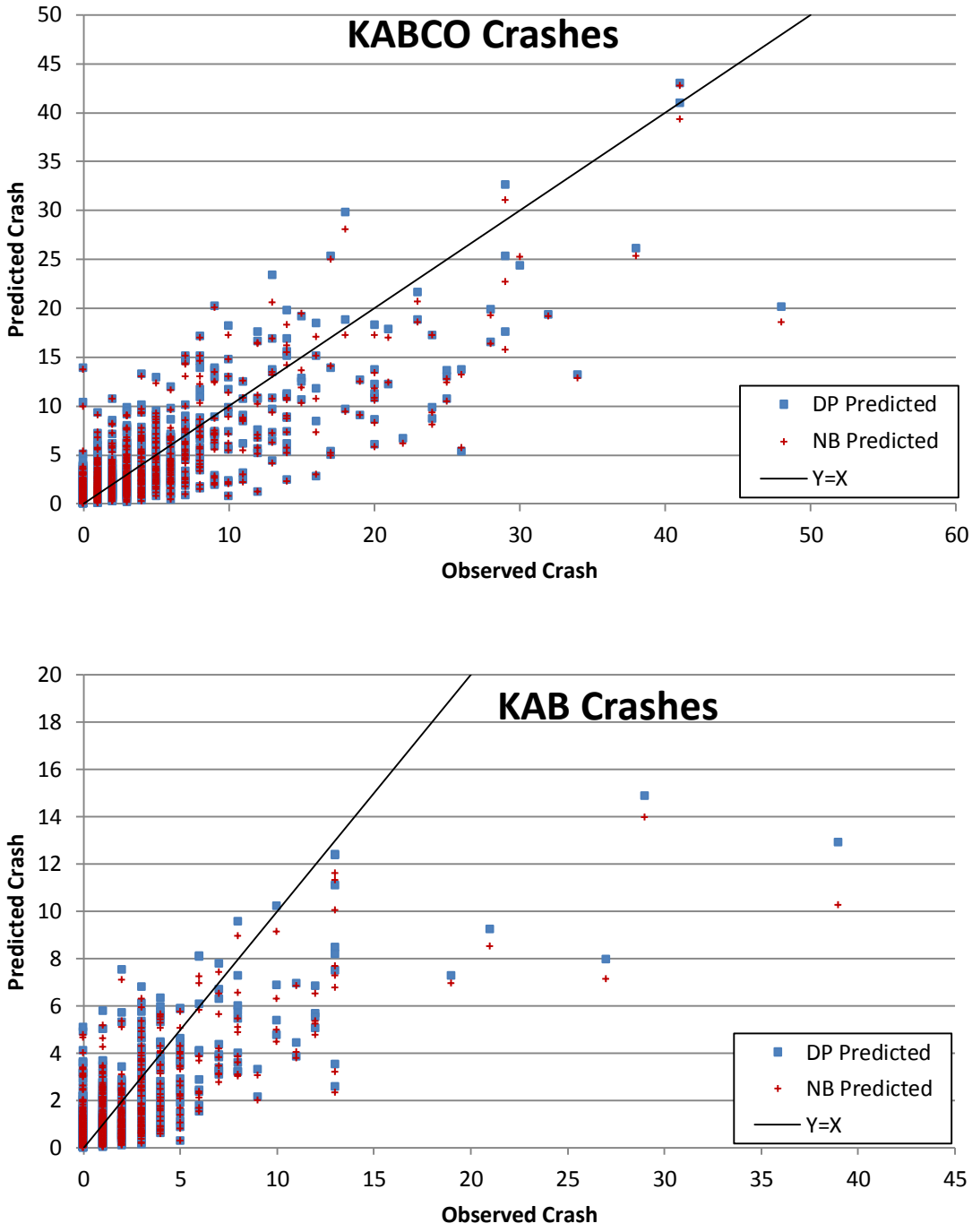


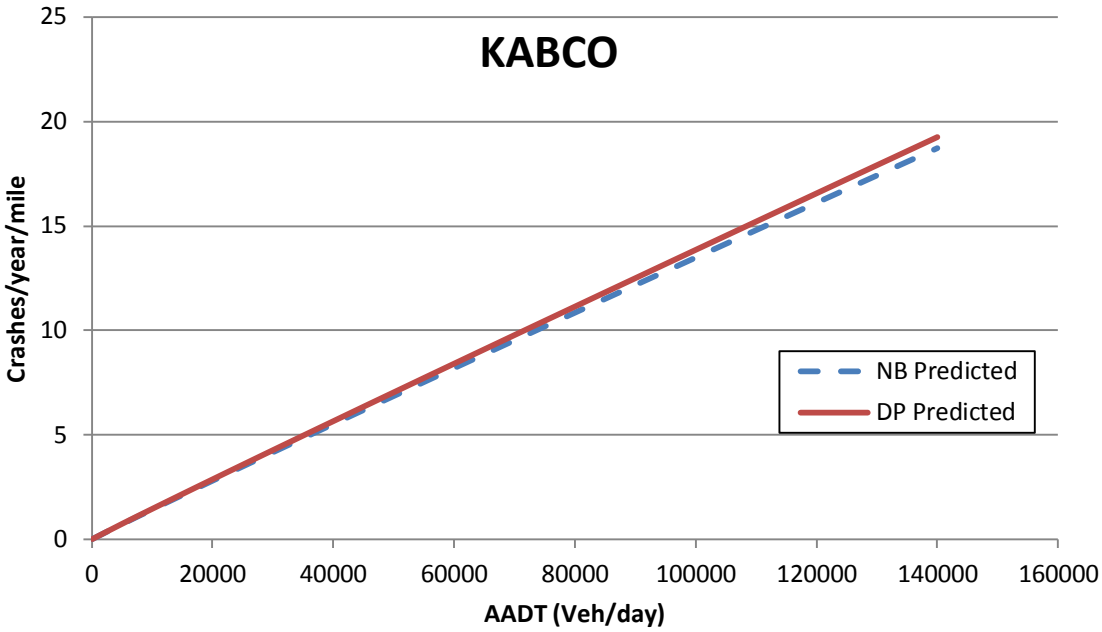
Figure 4.2 Predicted vs. observed crashes for the Texas data (KABCO crashes and KAB crashes)

Figure 4.3 illustrates the DP GLM and NB model predicted crashes against the variable AADT for the two response scenarios when controlling other variables at their average. It can be seen that the estimates of the DP GLMs are slightly higher than those of the NB models. The difference on the predictions between the two models increases with the increase of the AADT.

Figure 4.4 shows the cumulative residual (CURE) plots for the AADT variable under the two response scenarios of the Texas data. A CURE plot can be used to measure how the model fits the data with respect to each covariate by plotting the trend of the cumulative residuals with the increase of the interested variable (Hauer and Bamfo, 1997). Cumulative residuals oscillating closely around the value zero indicates a better fit to the data. In Figure 4.4, the plots were adjusted to make the final cumulative value equal to zero. It can be seen from both scenarios that the DP GLMs and NB models can give a similarly good fit to the data. The DP GLMs have smaller bias than the NB model when the AADT is larger than 15000 but perform slightly worse when the AADT is in the region from 8000 to 15000.

Figures 4.5 and 4.6 show the comparison on the crash variance predicted by the DP GLMs and NB models under the two response scenarios respectively. The predicted crash variance is calculated as the square of the residual for each model. The residual is equal to the difference between the predicted and the observed crashes. The reference line in each figure shows the theoretical relationship between the mean and variance for each model (for the DP, $Var(Y) \approx E(Y)/\theta$); for the NB, $Var(Y) = E(Y) + \alpha E(Y)^2$). As

we can see from the Figures 4.5 and 4.6, most data points fall evenly along their own theoretical line.



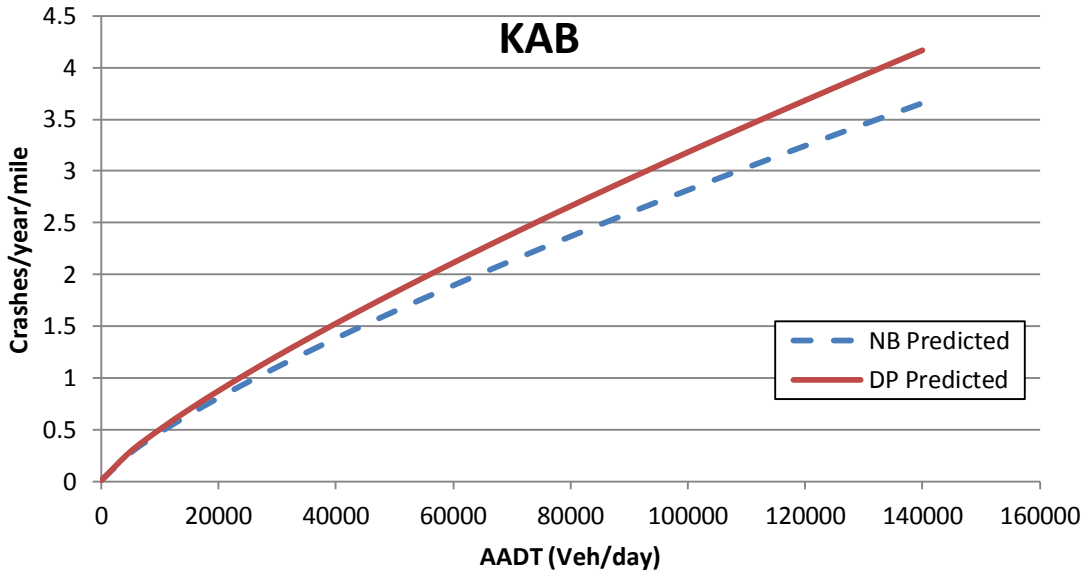


Figure 4.3 Estimated values (crashes/year) for the Texas data (KABCO crashes and KAB crashes)

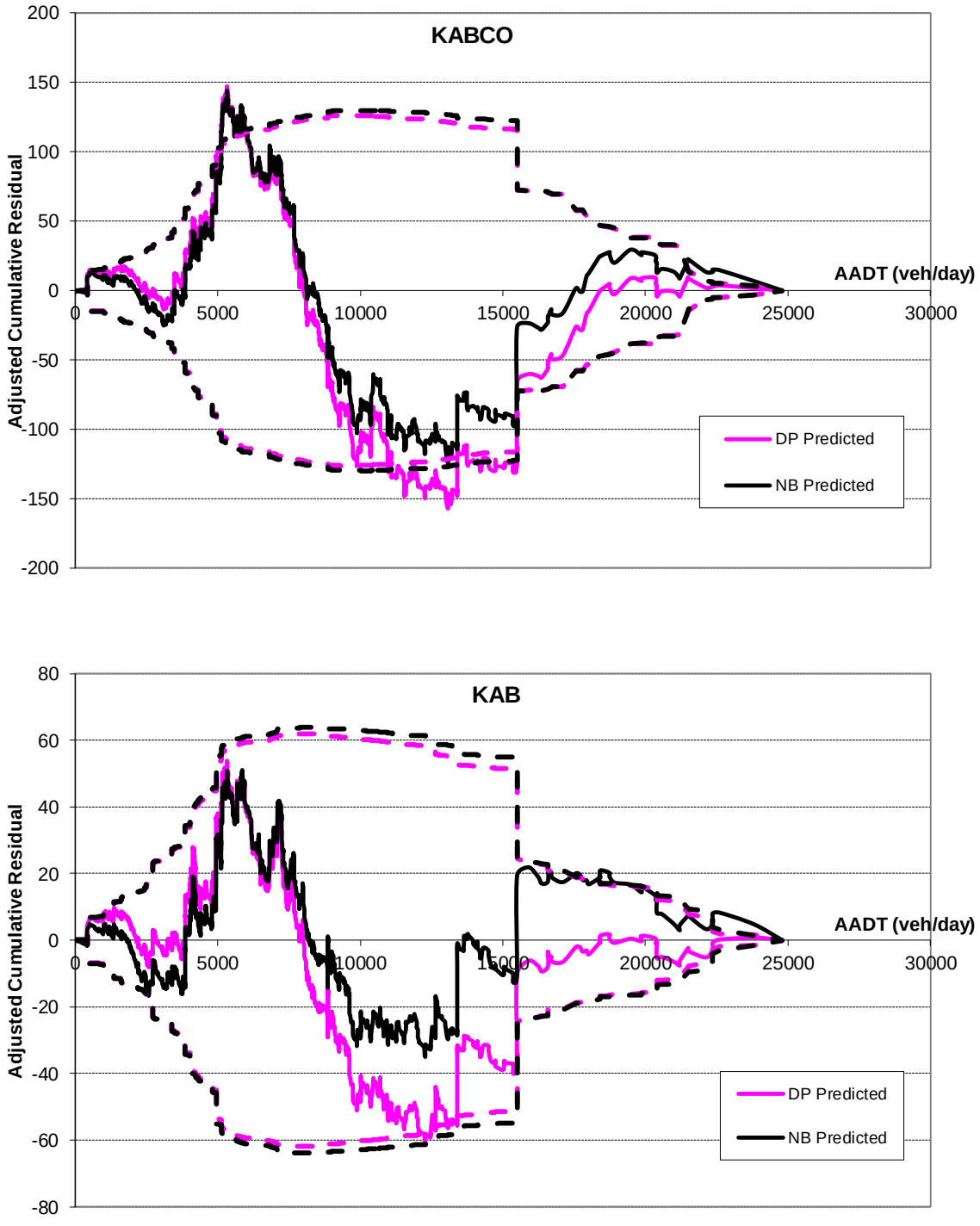
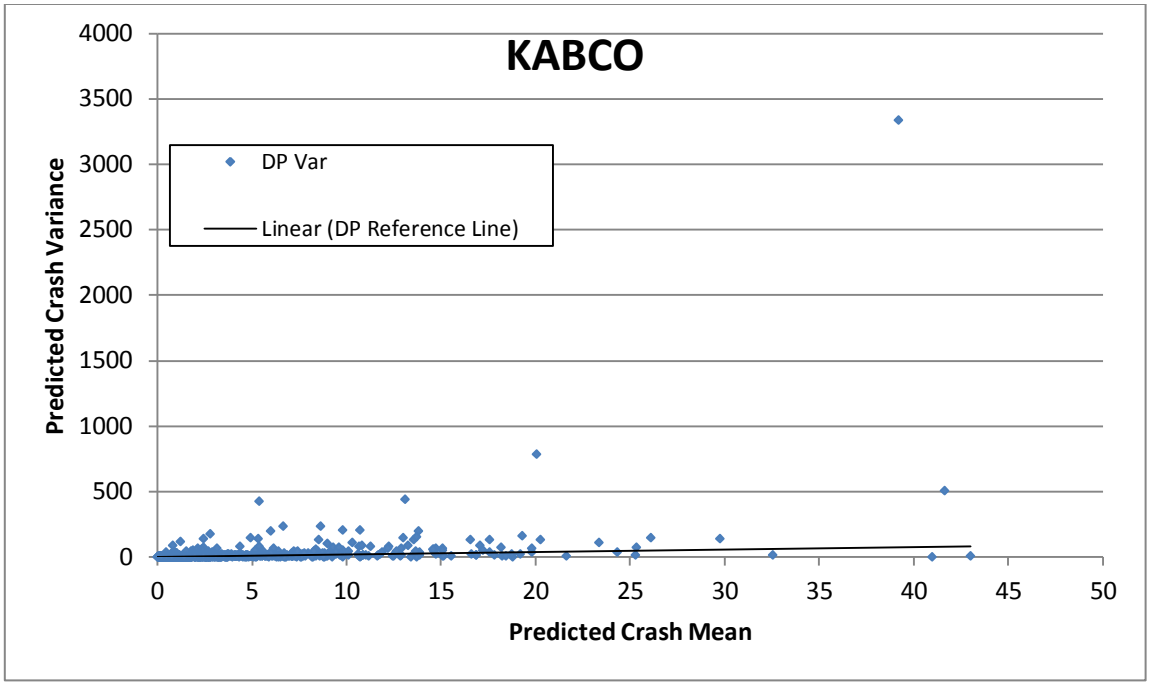
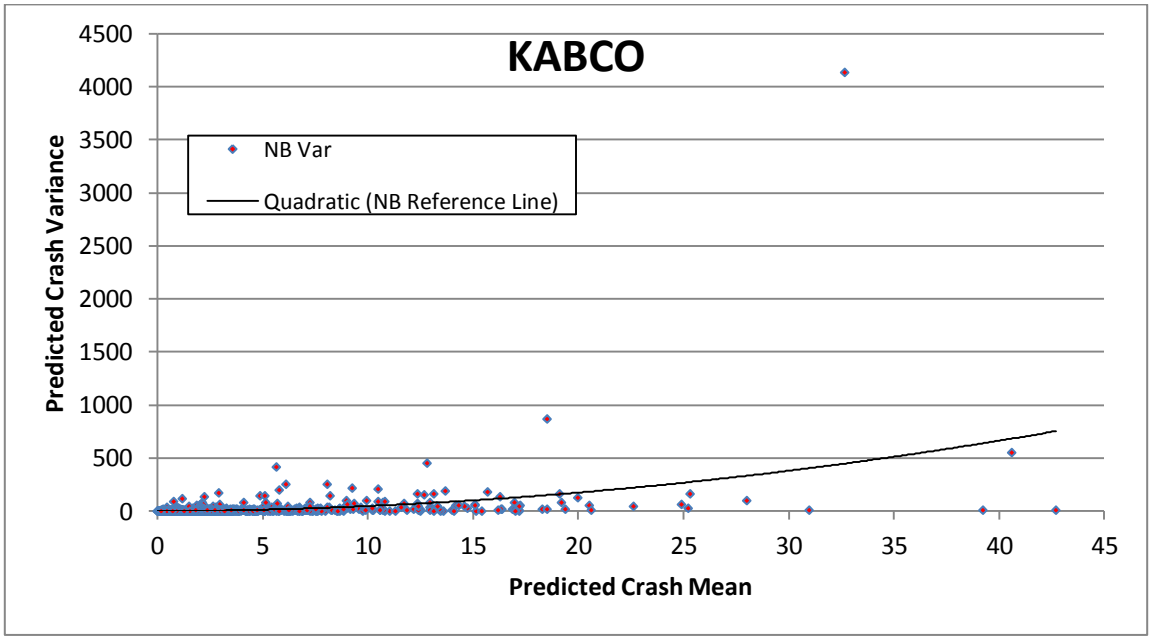


Figure 4.4 Cumulative residual plots for the Texas data against variable AADT (KABCO crashes and KAB crashes)
Note: Dotted lines represent ±2 standard deviance

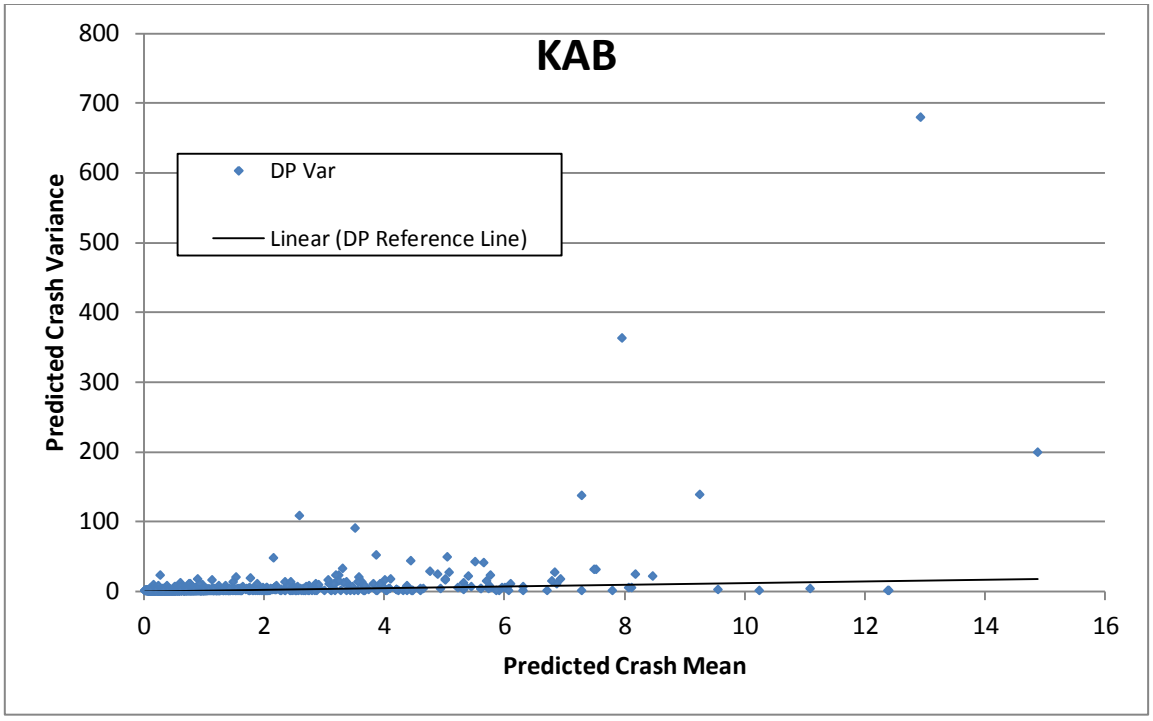


a) DP GLM

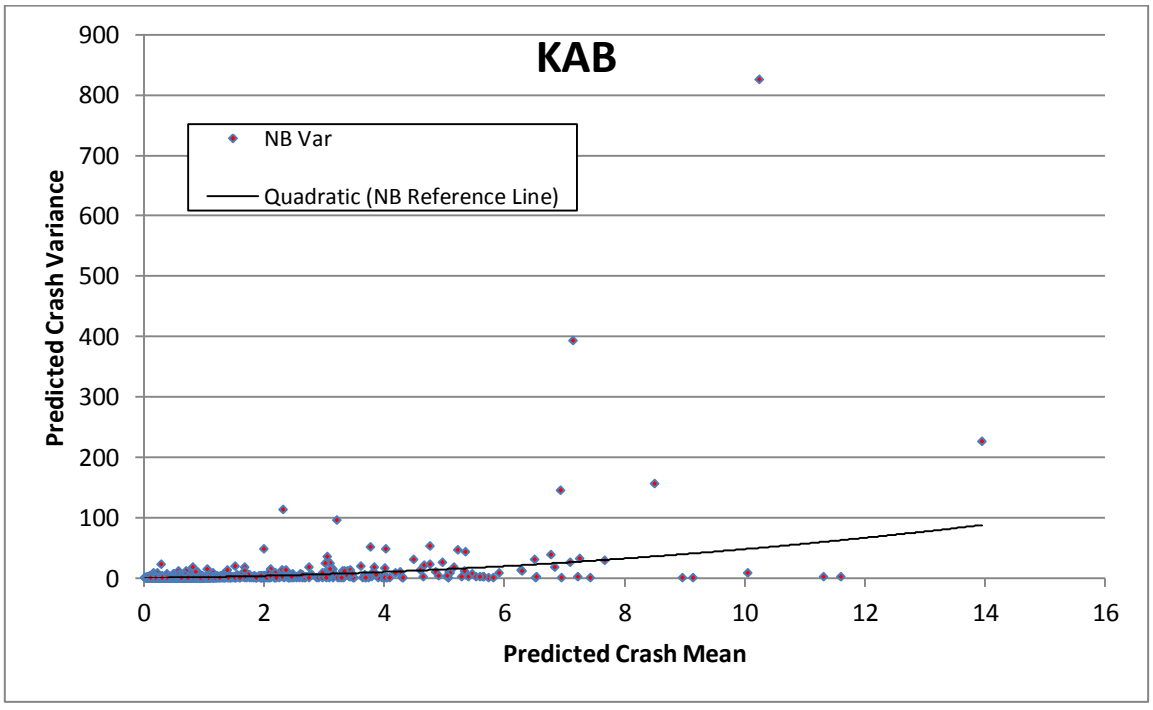


b) NB model

Figure 4.5 Predicted crash variance vs. predicted crash mean for the Texas data (KABCO crashes)



a) DP GLM



b) NB model

Figure 4.6 Predicted crash variance vs. predicted crash mean for the Texas data (KAB crashes)

4.5.2 Toronto data

Recently it has been found that the COM-Poisson GLM can fit the over-dispersed data as well as the NB model (Guikema and Coffelt, 2008; Lord et al. 2008a). Table 4.4 summarizes the comparison results of the DP GLM, NB model, and COM-Poisson GLM using the Toronto data.

Table 4.4 Comparison of results between the DP GLM, NB model, and COM-Poisson GLM using the Toronto data

Estimated Parameters and Standard Errors						
	DP		**NB[a]**		**COM-Poisson[b]**	
	Estimate	**Std Error**	**Estimate**	**Std Error**	**Estimate**	**Std Error**
$Ln(\beta_0)$	-10.2342	0.4518	-10.2458	0.465	-11.53	0.4159
β_1	0.6029	0.0458	0.6207	0.046	0.635	0.04742
β_2	0.7038	0.0223	0.6853	0.0211	0.795	0.03101
Shape Parameter	0.3944	0.0189	0.1398	0.0122	0.3408	0.02083
Goodness-of-fit Statistics						
AIC	5066		5077.3		--	
DIC	--		--		4953.7	
MAD	4.138		4.142		4.129	
MSPE	32.600		32.699		33.664	

[a] Based on the modeling results for NB model documented in Lord et al. (2008a).
[b] Based on the modeling results for COM-Poisson GLM documented in Lord et al. (2008a).

As shown in Table 4.4, the DP GLM gives a slightly better fit than the NB model since all the GOF statistics for the DP GLM are smaller than those for the NB model. The COM-Poisson GLM fits the data slightly better than the other two since it has lowest value of DIC (it is assumed that DIC is equivalent to AIC) and MAD. Similar to

what we have found for the Texas data, the DP GLM provides the smallest standard errors of almost all estimated coefficients. The coefficients estimated by the DP GLM are very similar to those by the NB model, while the coefficients estimated by the COM-Poisson are somewhat different than those for the other two. This difference could be explained by the fact that the coefficients for the COM-Poisson are directly linked to the mode rather than the mean as the case with the DP GLM and NB model (see the link function for the COM-Poisson GLM in Section 4.2). It should be also noted that the parameters of the COM-Poisson GLM were estimated using the Bayesian framework whereas the estimated parameters for the other two models were developed based on MLE.

Figure 4.7 shows the distributions of crash numbers (means) for the observed and predicted crashes of the three models for the Toronto data. The predicted crash means of the COM-Poisson GLM were obtained using the method we mentioned at the end of Section 4.2. The observed crashes are more scattered than the predicted crashes.

Figure 4.8 presents the scatter plot for the predicted vs. observed crashes for each site. All the three models give a good fit to the data with their data points in the scatter plot fall evenly along the Y=X reference line. It can be clearly seen that when the observed crashes is larger than 20, the DP GLM tends to predict more crashes than the NB model but less crashes than the COM-Poisson GLM. The fact that the DP GLM always gives a higher prediction than the NB model is consistent with that for the Texas data (see Figure 4.2). Moreover, the difference on each prediction for the three models is considerably minor, especially for the DP GLM and the NB model.

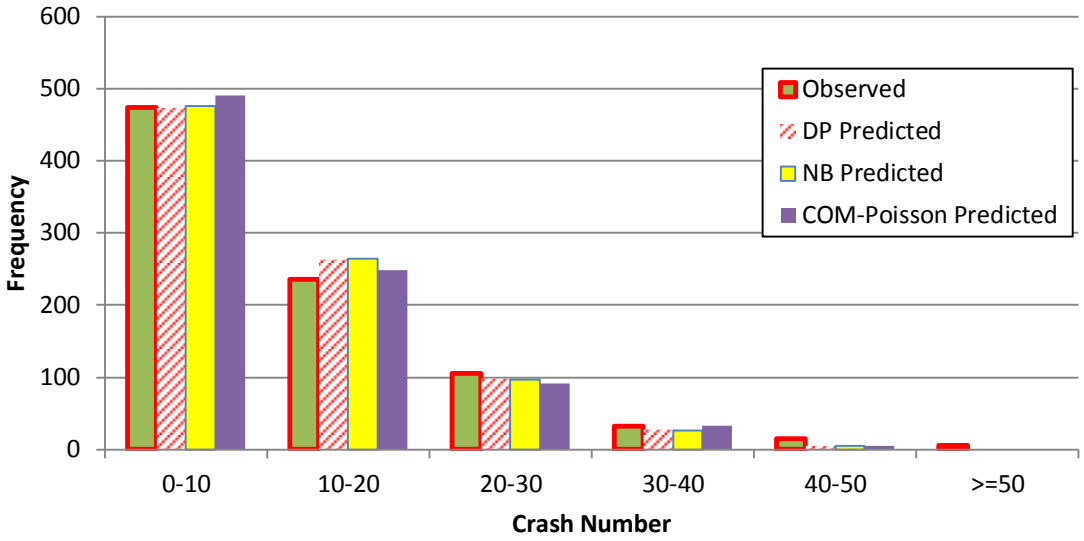


Figure 4.7 Frequencies of observed and predicted crashes for the Toronto Data

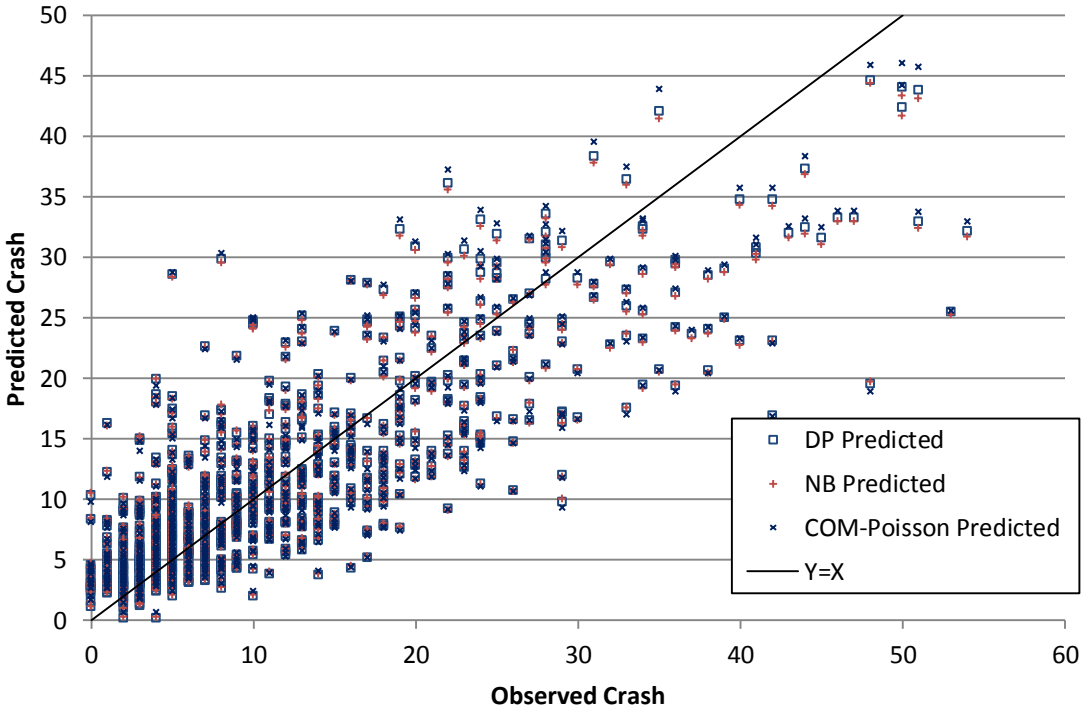
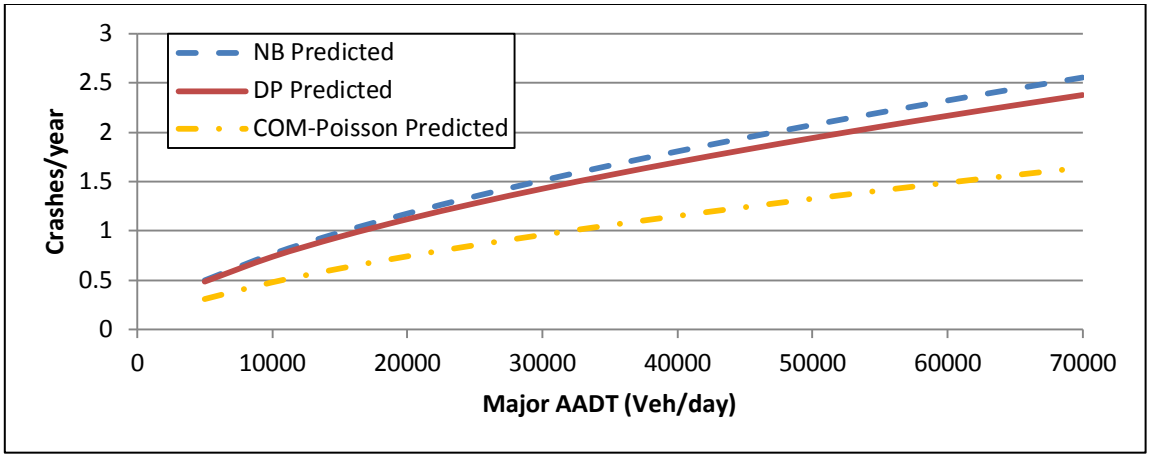


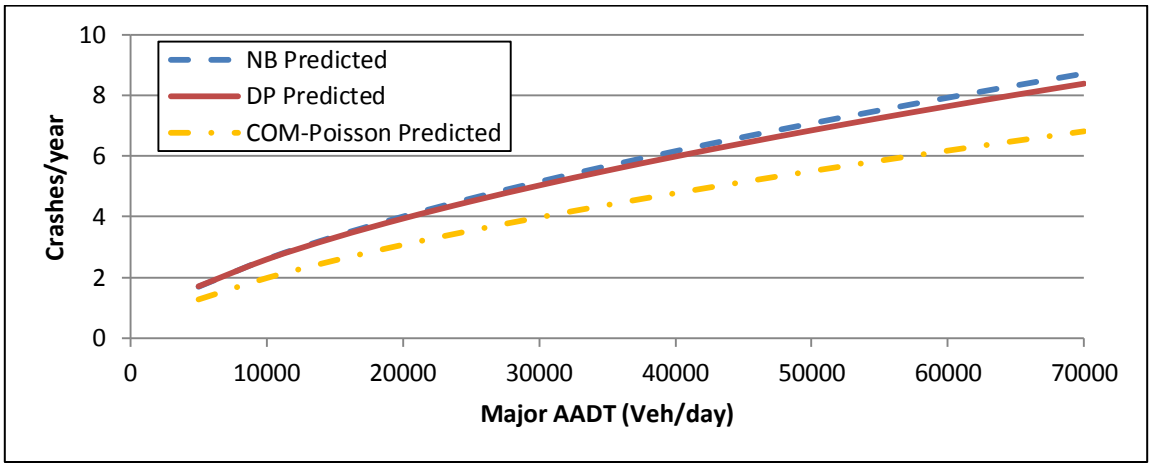
Figure 4.8 Predicted vs. observed crashes for the Toronto data

Figures 4.9 and 4.10 demonstrate the predicted crashes of the three models against the variable Major AADT and Minor AADT respectively. In Figure 4.9, three models have very similar trend to predict crashes. The DP GLM always has higher prediction than the COM-Poisson GLM but lower prediction than the NB model. The difference of the predictions of the DP GLM compared to the NB model is much less pronounced than those compared to the COM-Poisson GLM. The absolute differences among the three models increase when the Major AADT gets larger, whereas the relative differences among the three models decrease when the Minor AADT gets larger.

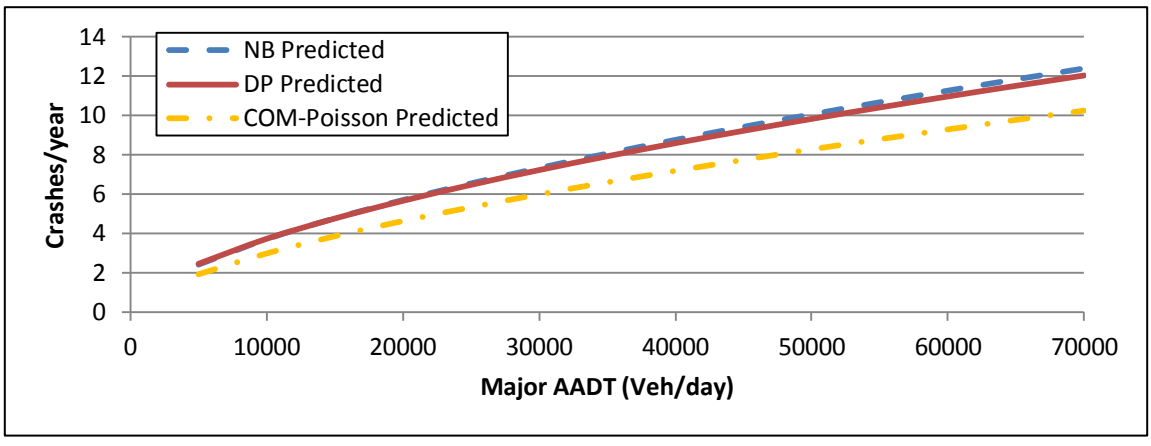
In Figure 4.10, three models give very close predicted crashes on the whole spectrum of the Minor AADT in all three scenarios of the Major AADT. The DP GLM always has higher prediction than the other two models. It should be noted that in Figures 4.9 and 4.10, the predicted crashes of the COM-Poisson GLM are meant for the mode, rather than the mean as the case with the DP GLM and NB model. The curves of three models would have been closer if the posterior mean value were used for the COM-Poisson instead of the mode (i.e. using the method introduced at the end of the Section 4.2).



a) Minor AADT = 500 Veh/day

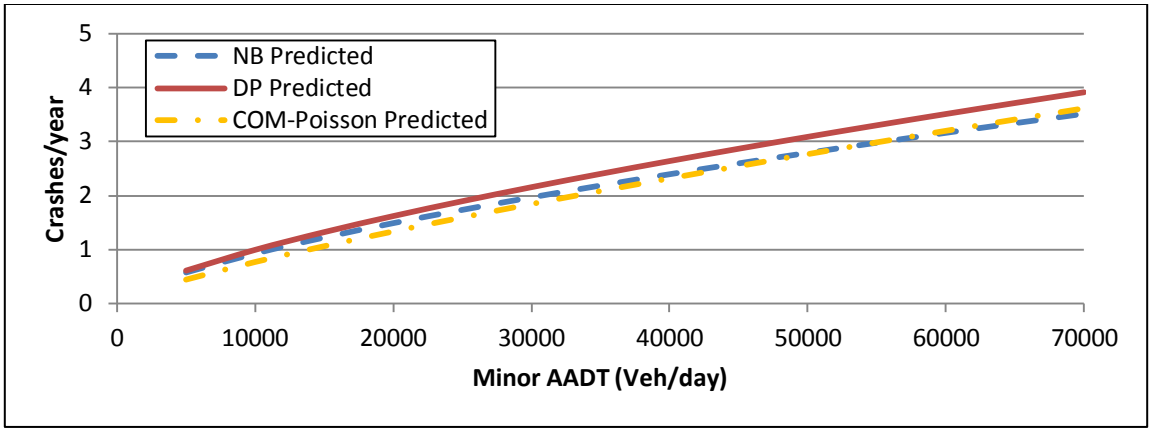


b) Minor AADT = 3000 Veh/day

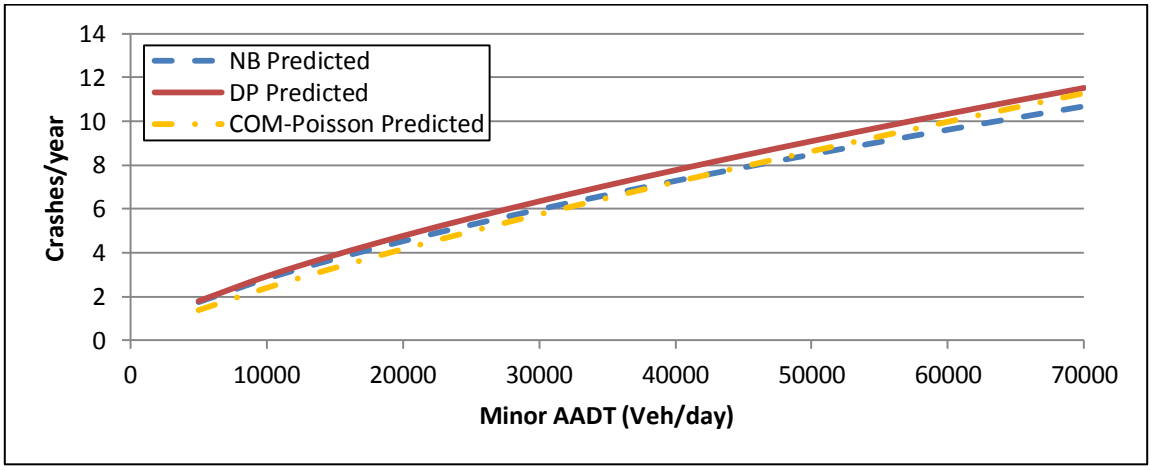


c) Minor AADT = 5000 Veh/day

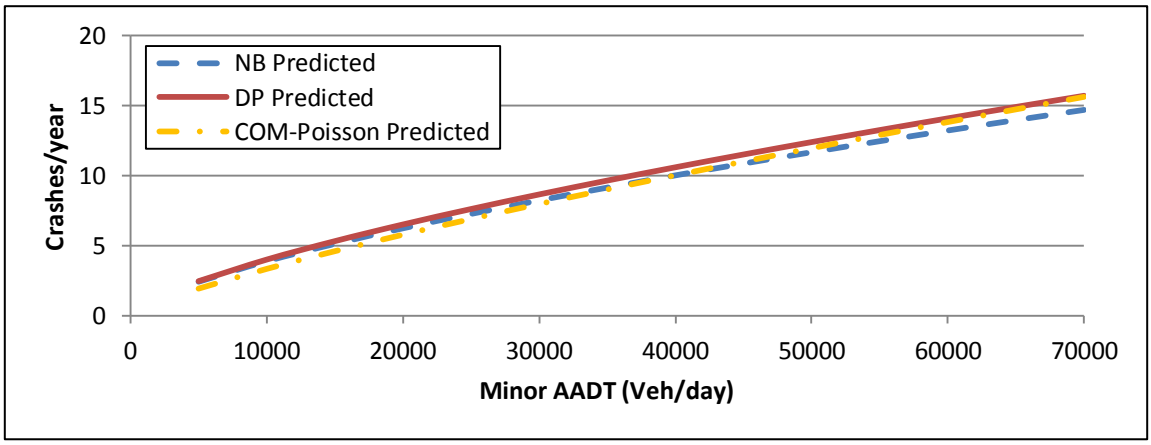
Figure 4.9 Estimated values for the Toronto data (against Major AADT)



a) Major AADT = 500 Veh/day



b) Major AADT = 3000 Veh/day

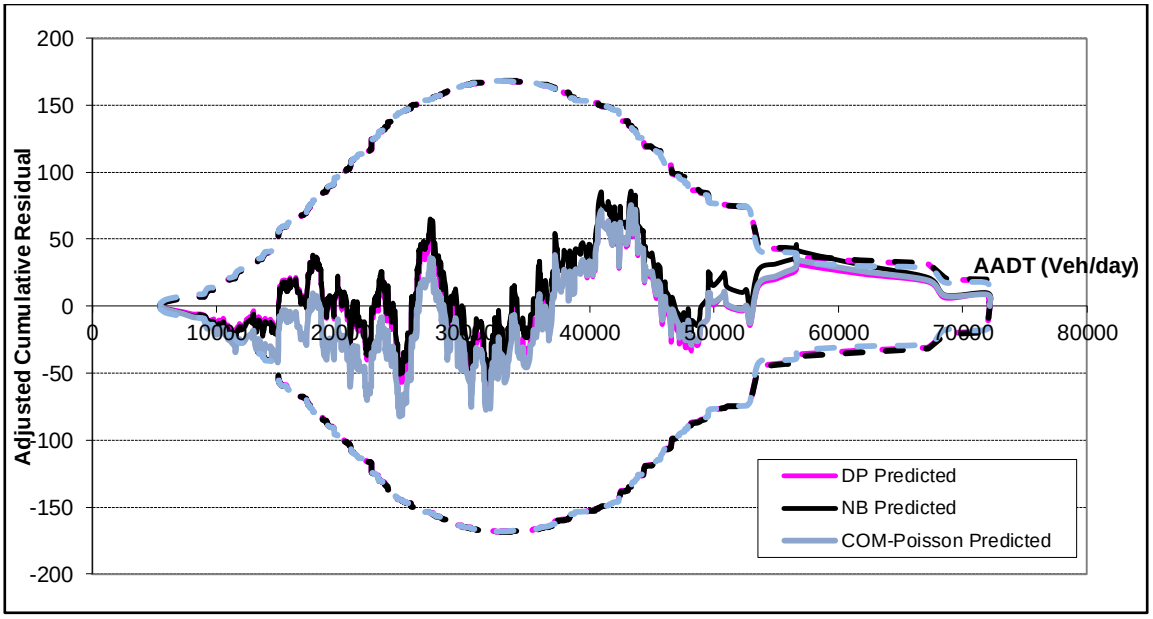


c) Major AADT = 5000 Veh/da

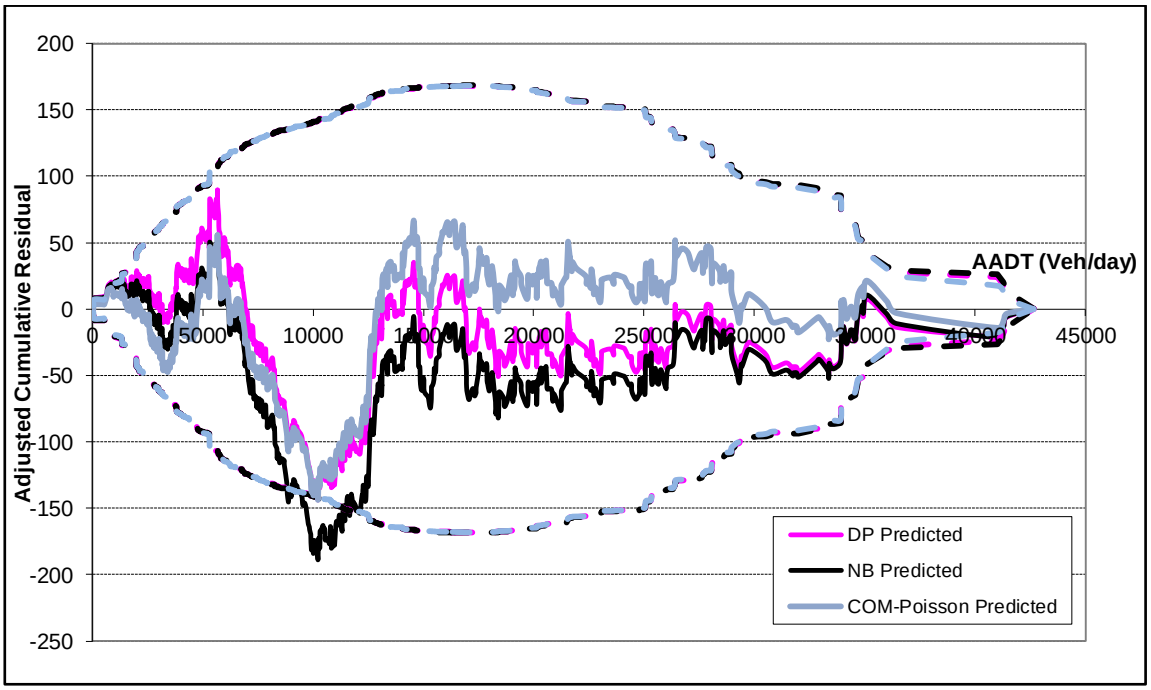
Figure 4.10 Estimated values for the Toronto data (against Minor AADT)

Figures 4.11a and 4.11b show the adjusted CURE plots for the Major AADT and Minor AADT respectively. In Figure 4.11a, the three models give almost equally good fit to the data. The difference between the DP GLM and the NB model is slightly smaller than that when the NB is compared to the other two models. In Figure 4.11b, the DP GLM fits the data as good as the COM-Poisson model, with curves of both models oscillating closely around the X axis for the most part of the range of Minor AADT variable. The performance of the NB model is worse than the other two models, with most part of its curve are the farther away from the X axis.

Figure 4.12 presents the comparison on the variances predicted by the three models. The way the predicted crash variance is calculated here is the same as that for the Texas data in Figures 4.5 and 4.6. For the COM-Poisson GLM, the predicted value obtained directly from its link function is the mode (μ_i). The predicted mean of the COM-Poisson for each site was obtained by taking the mean of 100,000 simulated samples with the estimated μ_i and v from the output of the COM-Poisson GLM. The theoretical relationship between the mean and variance for the COM-Poisson distribution is $E(Y) = \nu \times VAR(Y) + 1/(2\nu) - 1/2$, which is based on the Equations (2.28) and (2.29). We can see from the figure that the data points of the NB presents a quadratic relationship pattern between its variance and its mean, while the predicted variance of the DP and COM-Poisson is related to its mean in a linear pattern.

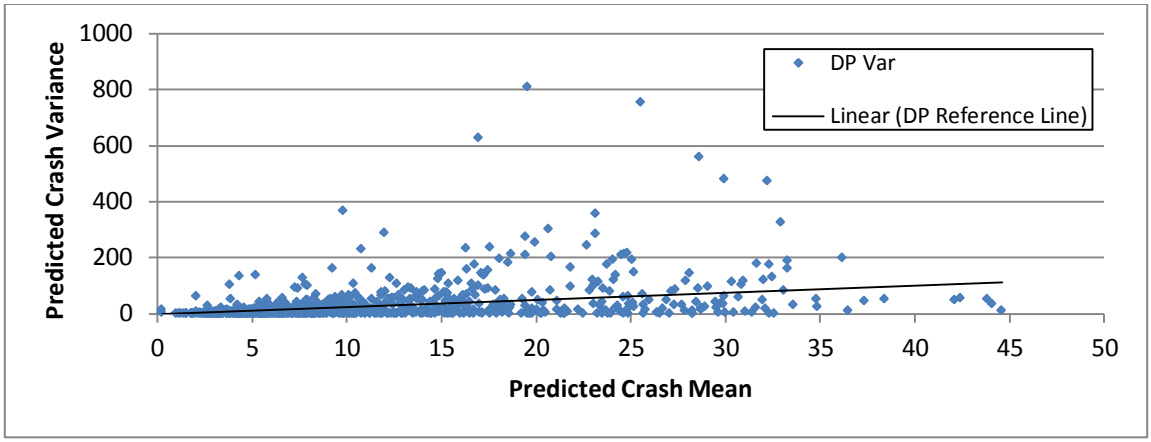


a) Against variable Major AADT

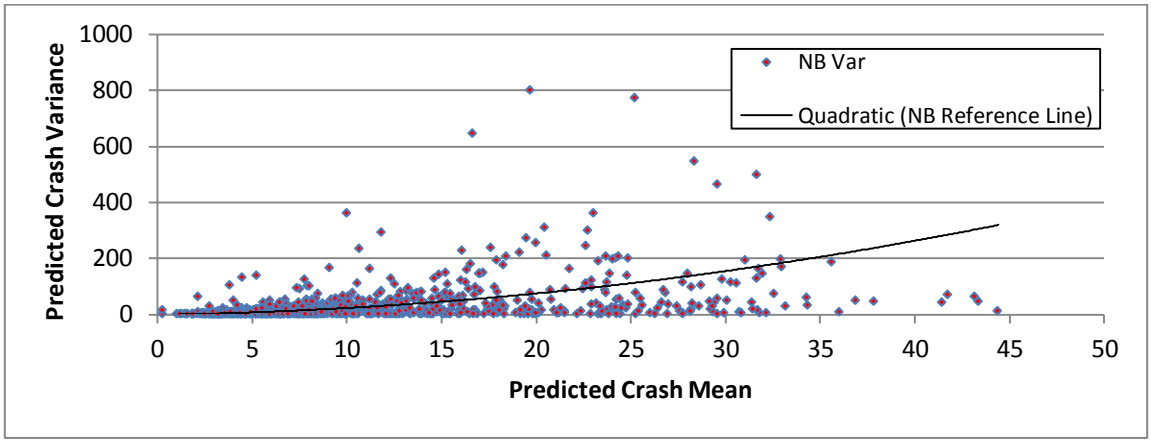


b) Against Variable Minor AADT

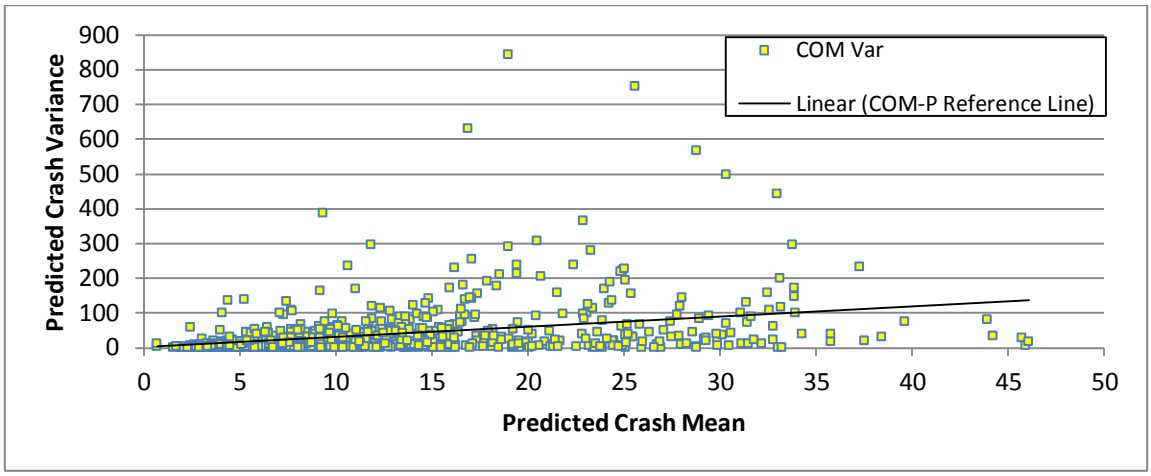
Figure 4.11 Cumulative residual plots for the Toronto data
Note: Dotted lines represent ±2 standard deviance



a) DP GLM



b) NB Model



c) COM-Poisson GLM

Figure 4.12 Predicted crash variance vs. predicted crash mean for the Toronto data

4.5.3 DP GLM with or without the normalizing constant

As documented in Section 2.5, Equation (2.34) demonstrates the expression and approximate solution to the normalizing constant (NC) in the DP distribution. Since the exact NC is an infinite sum, the inclusion of the NC in the PMF of the DP increases the non-linearity and poses computational challenges. It has been found that even the inclusion of the closed form approximation to the NC incurs a convergence issue on solving for the MLE in SAS in many cases. For all the datasets investigated in this study, only the Toronto data converged successfully when incorporating the approximate solution to the NC in the PMF of DP distribution.

Given the computational challenges from the inclusion of the NC, all the DP GLMs in this thesis were developed based on the PMF without the NC. Efron (1986) also pointed out that the PMF without the NC is highly accurate for the case $\mu = 10$. The objective of the study in this subsection is to verify if including the close-formed approximate solution to the NC in the PMF of the DP would improve the DP GLMs. The dataset and the corresponding link function are those based on which we developed the DP GLMs for the Toronto data. Likewise, the GOF statistics and parameter estimation method documented in Sections 4.3 and 4.4 respectively are used.

Table 4.5 summarizes the results of the DP GLM with and without the NC for the Toronto data. The NC here refers to its closed form approximate solution. The two models provide very similar estimated coefficients. The standard errors of the estimated parameters in the DP GLM with the NC are slightly smaller than those without the NC. In terms of the GOF, it is interesting to see that all the GOF statistics for the DP GLM

without the NC is slightly smaller than those with the NC, indicating the DP GLM without the NC performs better than that with the NC. The difference between the two models is very minor though, which is also confirmed by the scatter plot for the predicted vs. observed crashes of the two models as shown in Figure 4.13. In Figure 4.13, the data point of each observation for one model is very close to that for the other model and the smaller the observed crash, the smaller the difference of predictions made by the two models.

Table 4.5 Comparison between the DP with and without normalizing constant using the Toronto data

Estimated Parameters and Standard Errors				
	DP without NC		**DP with NC**	
	Estimate	**Std Error**	**Estimate**	**Std Error**
$Ln(\beta_0)$	-10.2342	0.4518	-10.0000	0.4248
β_1	0.6029	0.0458	0.5988	0.0431
β_2	0.7038	0.0223	0.6843	0.0207
Shape Parameter	0.3944	0.0189	0.4334	0.0189
Goodness-of-fit Statistics				
AIC	5066.0		5115.8	
MPB	-2.75765E-16		0.074	
MAD	4.138		4.152	
MSPE	32.600		32.685	

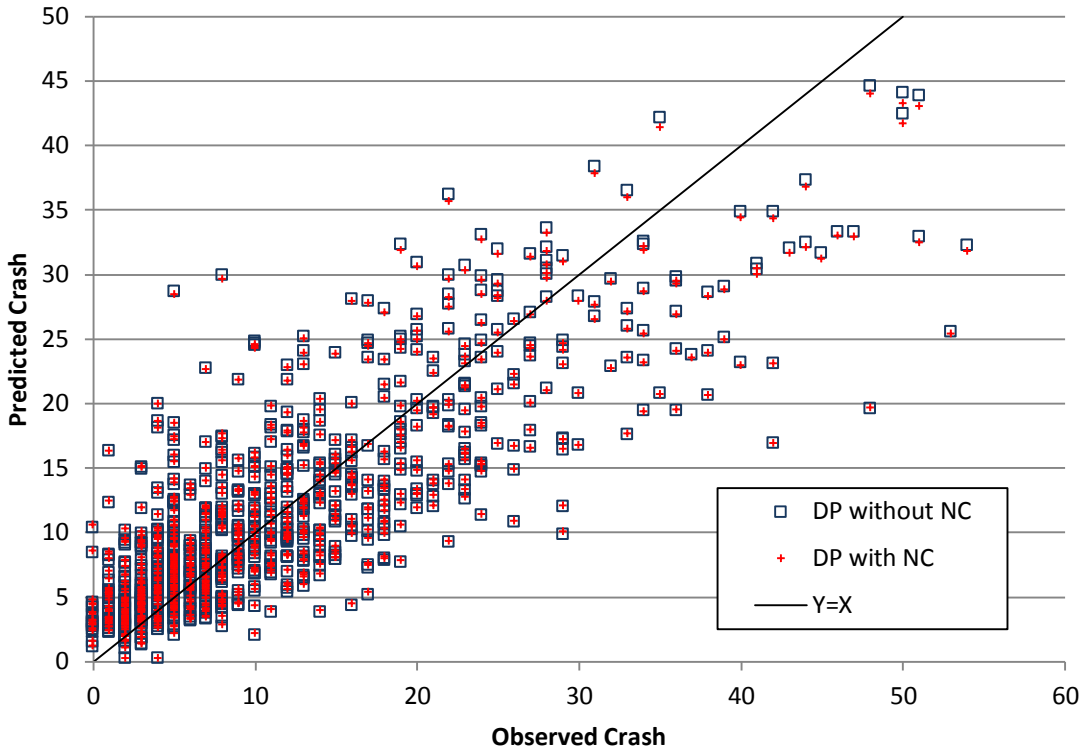


Figure 4.13 Predicted vs. Observed Crashes for the DP with and without normalizing Constant

It is concluded that for this dataset, including the closed form approximate solution to the NC in the PMF does not improve the performance of the DP GLM. The possible explanation might be that the approximate solution to the NC in Equation (2.34) is not close enough to the true value under certain circumstances.

It is important to note that other than the closed form approximation proposed by Efron (1986), the NC might be also approximated by truncation of its infinite sum. This approximation method has the potential to improve the DP GLM since the sum is similar to the Poisson sum and thus expected to converge quickly. If the sum only requires a

small number of summations to achieve high accuracy, it would not pose great computation challenges and could expand the application of the DP GLMs from a practical perspective. Therefore, it is worth further investigating on the accurate approximation of the NC.

4.6 Discussion

This subsection provides a detailed discussion on the overall performances of the DP GLM, COM-Poisson GLM and NB model for analyzing over-dispersed data based on the previous results.

- Goodness of fit

  The DP GLM fits the over-dispersed data almost the same as the NB model and COM-Poisson GLM. In the first dataset the DP GLM gives a slightly worse fit than the NB GLM while for the second dataset, the former is slightly better than the latter. Given that the differences in GOF statistics between the two models are not pronounced, we conclude that the DP GLM can provide as good fit to the over-dispersed data as the NB model. Given that the comparison results are very similar for the two different datasets using different link functions, the above conclusion seems to hold in spite of the form of the link function. Meanwhile, it has been found that the COM-Poisson GLM performs as well as the NB model (Guikema and Coffelt, 2008; Lord et al. 2008a) in terms of the GOF statistics and predictive performance, we further conclude that the DP GLM fits the over-dispersed almost the same as the NB model and COM-Poisson GLM.

In this section, it has also been found that all the three models tend to over-predict crashes for smaller mean values and under-predict crashes for larger mean values. The investigation on the use of the normalizing constant in the PMF of the DP distribution indicates that the inclusion of the closed form approximation to the normalizing constant proposed by Efron (1986) does not improve the DP GLM in terms of GOF.

- Estimated parameters and standard errors

  The results of both datasets indicate that the DP GLM can detect the presence of over-dispersion with its estimated shape parameter being significantly less than 1 (for the DP, $Var(Y) \approx \mu / \theta$), which is similar to the shape parameter of the NB model. The DP GLM also gives very similar estimates for those coefficients as the NB model. One advantage the DP GLM over the NB model and COM-Poisson GLM is that the DP GLM tends to always give the smallest standard errors for its estimated coefficients.

Besides, the parameter estimation method in this study is very similar between the DP GLM and NB model, since they both estimate their parameters based on MLE and they both use the software SAS to code their mathematical functions. To be exact, the MLE for the DP GLM in this study is an approximation since it uses the approximate PMF in which the normalizing constant was not considered. Although Sellers and Shmueli (2010) developed the code for the COM-Poisson

GLM using MLE, its GLM framework was based the parameter λ as shown in Equations (2.18) and (2.19), which cannot provide a clear centering parameter (Guikema and Coffelt, 2008). Although the re-parameterization of the COM-Poisson proposed by the Guikama and Coffelt (2008) was based on a more clearer centering parameter which is the mode μ, it is still not intuitive to interpret as the use of the mean in the DP GLM and COM-Poisson. In this section, the COM-Poisson GLMs estimate their parameters μ and v based on the Bayesian method in the software WinBUGS (Guikema and Coffelt, 2008).

- Mathematical relationship

  All the three models we discussed in this section have two parameters: a centering parameter and shape parameter (or dispersion parameter). As we mentioned before, the COM-Poisson model first used the parameter λ in its parameterization and it cannot provide a clear centering parameter as the mean of the DP GLM and NB model. Although the results of the COM-Poisson GLM in this section come from another parameterization of the COM-Poisson which was recently proposed and based on the mode (Guikema and Coffelt, 2008), its corresponding GLM framework was directly linked to the mode rather than the mean, and this framework was established under the Bayesian framework which costs much time to calculate the estimated parameters (discussion on the computational time presented in next subsection).

For the NB model, it could theoretically handle under-dispersion by setting its shape parameter as negative ($Var(Y) = \mu + (-\alpha)\mu^2$). However, doing that would lead to a misspecification of its PDF (Clark and Perry, 1989; Saha and Paul, 2005) and unreliable parameter estimates (Lord et al, 2010). Therefore, the shape parameter of the NB model is only limited to measure the over-dispersion, whereas the shape parameter of the DP model is very flexible and it can handle under-, equi-, and over-dispersion ($Var(Y) \approx \mu/\theta$, $\theta > 1$ for under-dispersion; $\theta = 1$ for equi-dispersion; $\theta < 1$ for over-dispersion). The capability of the DP GLM of handling under-dispersed data is presented in the next section.

- Computational time

  As we discussed earlier, both the DP GLMs and NB models are developed based on MLE using SAS. The computational time for generating the results in those two models is within ten seconds. The COM-Poisson GLMs in this study, however, normally take several hours to generate the outputs in WinBUGS which is based on the Bayesian method (Guikema and Coffelt, 2008, Lord at al, 2008). As mentioned earlier, although recently the MLE code of the COM-Poisson GLM has been developed in R (Sellers and Shmueli, 2010; Francis et al. 2012) and the computational time has been greatly reduced, its GLM framework was established based on the parameter λ as the response variable (in next section, the MLE based COM-Poisson GLM will be compared with the DP GLM). The use of the parameter λ in establishing GLM is not intuitive as the

use of mean, and its corresponding estimated coefficients cannot compare directly with those estimated with mean as the response variable.

4.7 Summary

This section evaluated the application of the DP GLM in analyzing motor vehicle crash data characterized by over-dispersion. In most cases, crash data are over-dispersed. The NB is the most widely applied distribution in handling over-dispersed crash data. The focus of the study in this section is to compare the performance of the DP GLM with that of the NB model in handling over-dispersed crash data. Previous research has found that the COM-Poisson GLM can fit the over-dispersed data as well as the NB model. Thus, the results of the COM-Poisson GLM were also provided as a reference.

Two datasets, the Texas data and Toronto data were used to develop those modes. The crashes in the Texas data occurred on the roadway segments (non-intersection related crash data) whereas the crashes in the Toronto data occurred on the intersections (intersection related data). Correspondingly, two commonly used link functions were employed in this section, one for modeling roadway segment crashes and the other for modeling intersection crashes. For the Texas data, two response variables were used (KABCO crashes and KAB crashes). Several measures of GOF were used to compare the performances of the models that can handle over-dispersed data.

The comparison results for both the datasets indicate that the DP GLM fits the over-dispersed data almost the same as the NB GLM and COM-Poisson GLM. The DP GLM provides very similar estimates for those coefficients as the NB GLM. However,

the standard errors of estimated parameters for the DP GLM are smaller than those for the other two models. The mathematics to manipulate the DP GLM is very simple and very similar to the NB model, since both of the two models use the mean as the centering parameter and have a shape parameter to handle the presence of dispersion. The shape parameter of the NB model was only intended for the over-dispersed data while the shape parameter for the DP GLM can handle over-, equi-, and under-dispersion (the performance of the DP GLM handling under-dispersed data is presented in next section). Moreover, both the DP GLM and NB model were developed based on the MLE, which was not the case with the COM-Poisson GLM developed on the Bayesian framework in this section. Although the MLE recently becomes available for the COM-Poisson, its corresponding GLM framework is linked to the parameter which cannot serve as a clear centering parameter as the mean in the DP GLM and NB model. The computational time for the DP GLM and NB GLM using MLE framework was substantially quicker than that for the COM-Poisson GLM using Bayesian framework.

Thus, the overall performance of the DP GLM is better than that of the COM-Poisson GLM and the same or slightly better than that of the NB model (the DP GLM always gives the smallest standard errors) in handling the over-dispersed crash data. It should be noted that all the DP GLMs in this thesis were developed based on the approximate PMF and thus the MLE for the DP GLMs is an approximation. Considering the shape parameter of the DP distribution can handle under-, equi-, and over-dispersed data, it is of great interest to examine the performance of the DP GLM in handling

under-dispersed data. The next section will investigate the applicability of the DP GLMs in handling under-dispersed data.

5. APPLICATION OF THE DOUBLE-POISSON GLM TO CRASH DATA CHARACTERIZED BY UNDER-DISPERSION

On rare occasions, traffic crash data are characterized by under-dispersion (Lord and Mannering, 2010). The presence of data characterized by under-dispersion makes the most commonly used models such as the Poisson and NB unable or very difficult to handle those traffic crash data (Clark and Perry, 1989; Saha and Paul, 2005). Recently, the gamma count model (Oh et al., 2006) and the COM-Poisson model (Kadane et al., 2006; Sellers and Shmueli, 2010; Guikema and Coffelt, 2008; Geedipally et al., 2008) has been applied to handle the under-dispersed data. Particularly, the COM-Poisson distribution and GLM has been found very flexible to handle under-dispersed data.

The study in this section aims to examine the performance of the DP GLM for analyzing traffic crash data characterized by under-dispersion. Pairwise comparisons are first conducted between the DP GLM with other two models (the COM-Poisson GLM and gamma count model) that can handle under-dispersed data. Then an overall comparison among the three models is provided. All the comparisons conducted in this section are based on the same dataset, with which Oh et al. (2006) developed the gamma count model to examine the safety effects of railway-highway crossing elements.

5.1 Data Description

The dataset used for modeling under-dispersion were collected at railway-highway crossings in Korea (Korea data). This dataset were found to be under-dispersed and were

used for establishing the gamma count models to examine factors associated with railroad crossing crashes (Oh et al., 2006). Traffic accident records were recorded at a total of 162 railway-highway crossings in the 5-year period from 1998 to 2002. A total of 56 continuous and categorical explanatory variables including average daily traffic (ADT) were collected through site visits and investigations. Tables 5.1 and 5.2 summarize the key continuous and categorical variables of that dataset respectively.

Table 5.1 Summary statistics of continuous variables for Korea data

Variables	Min.	Max.	Average	Std. Dev	Obs
Crashes*	0	3	0.33	0.60	162
Number of Tracks	1	2	1.38	0.49	162
ADT (vehicles per day)	10	61199	4617.00	10391.57	162
Average daily railway traffic (trains per day)	32	203	70.29	-37.34	162
Gradient of Road	-20	10.5	-1.30	3.73	162
Train detector distance (m)	0	1329	824.50	328.38	162
Time duration between the activation of warning signals and gates (s)	0	232	25.46	25.71	162

*Response variable.

Table 5.2 Summary statistics of categorical variables for Korea data

Variables	Coding	Frequency	Percentage
Presence of commercial area	1 (yes)	149	91.98%
	0 (no)	13	8.02%
Presence of track circuit controller	1 (yes)	134	82.72%
	0 (no)	28	17.28%
Presence of guide	1 (yes)	113	69.75%
	0 (no)	49	30.25%
Presence of speed hump	1 (yes)	126	77.78%
	0 (no)	36	22.22%

5.2 Link Function

This subsection describes how the response variables are linked to explanatory variables. The functional form used by Oh et al. (2006) in developing gamma count models was applied to the DP GLM and gamma count model (for gamma count, μ_i is the same as the λ_i in Equation (2.13)) by the following equation:

$$\mu_i = \exp(\beta_0 + \beta_1 \ln(F_i) + \sum_{j=1}^{n} \beta_j x_{ij}) \quad (5.1)$$

where,

μ_i = the mean of crashes in 5 years for the crossing i;

F_i = average daily vehicle or ADT traffic on crossing I (vehicles/day);

x_{ij} = the jth covariate for crossing i;

β_j = estimated coefficients across covariates j=1,…,n.

It should be noted the link function used for the COM-Poisson GLM in this section (MLE-based) is different with that in the last section (Bayesian-based). In this section, the link function for the COM-Poisson was established based on the parameter λ (Sellers and Shmueli, 2010):

$$\lambda_i = \exp(\beta_0 + \beta_1 \ln(F_i) + \sum_{j=1}^{n} \beta_j x_{ij}) \quad (5.2)$$

where,

λ_i = approximately the mean of crashes in 5 years for the crossing i;

F_i = average daily vehicle or ADT traffic on crossing i (vehicles/day);

x_{ij} = the jth covariate for crossing i;

β_j = estimated coefficients across covariates j=1,…,n.

This method proposed by Sellers and Shmueli (2010) was used to establish a single-link COM-Poisson GLM based on MLE. Recall that the link function of the COM-Poisson GLM in last section was based on Bayesian framework (Guikema and Coffelt, 2008).

5.3 Goodness-of-fit

The GOF statistics of the DP GLM for fitting under-dispersed data are the same with those for fitting over-dispersed data. Details are presented in Section 4.

5.4 Parameter Estimation Method

For the DP GLM, the way how parameters were estimated in this section (for under-dispersed data) was exactly the same as that in the Section 4 (for over-dispersed data). Different with the Section 4, the parameters for the COM-Poisson GLM in this section were estimated using the R code developed by Sellers and Shmueli (2010) which is based on MLE. It should be noted that the estimated coefficients for the DP GLM and those for the COM-Poisson GLM in this section cannot be compared directly due to the difference on the response variable between their link functions as shown in Equations 5.1 and 5.2.

5.5 Comparison Results

In this subsection, results on pairwise comparisons (DP GLM vs. gamma count model and DP GLM vs. COM-Poisson GLM) will be first presented. Then results on an overall comparison among the three models will be provided.

5.5.1 Pairwise comparison

Given that this under-dispersed dataset were originally used for establishing the gamma count model to analyze railway-highway crashes, this study first compared the DP GLM with the gamma count model using the variables originally found significant in the gamma count model (it should be noted that the variables found significant in the gamma count model are not necessarily significant in the DP GLM).

Table 5.3 shows the comparison results between the DP GLM and gamma count model using the Korea railway-highway crossing crash data. In Table 5.3, the DP GLM provides smaller values for all the GOF statistics compared with the gamma count model. It can be inferred that the DP GLM fits the under-dispersed data much better than the gamma count model albeit only including the variables found significant in the gamma count model. Meanwhile, the estimated coefficients indicate that the marginal effect for each variable in the DP GLM tends to be larger than that in the gamma count model with the exception of the variable presence of speed hump.

Table 5.3 Comparison between the DP GLM and gamma count model using Korea data

Estimated Parameters and Standard Errors		
Variables	**Gamma** [a]	**DP**
Constant	-3.438 (1.008) [b]	**-5.3360 (0.7489)** [c]
Ln(ADT)	0.230 (0.076)	**0.3443 (0.0668)**
Average daily railway traffic	0.004 (0.0024)	**0.0052 (0.0027)**
Presence of commercial area	0.651 (0.287)	**0.9666 (0.3010)**
Train detector distance	0.001 (0.0004)	**0.0014 (0.0004)**
Time duration between the activation of warning signals and gates	0.004 (0.002)	**0.0047 (0.0026)**
Presence of speed hump	-1.58 (0.859)	**-0.8530 (0.3464)**
Shape parameter	2.062 (0.758)	**1.5033 (0.1670)**
Goodness-of-fit Statistics		
AIC	211.38	205.9
MPB	0.179	-1.2963E-06
MAD	0.459	0.378
MSPE	0.308	0.260

[a] Based on the modeling results for gamma count model documented in Oh et al. (2006).

[b] Values in parentheses are the standard errors for the estimated parameters.

[c] Bolded value indicates the related variable in the DP GLM is significant at the significance level of 0.10.

Table 5.4 shows the comparison results on the estimated parameters (coefficients and shape parameters) and standard errors for the variables that found to be significant in the COM-Poisson model (Lord et al., 2010). Again, the variables found significant in the COM-Poisson GLM are not necessarily significant in the DP GLM. But coincidently, all those variables happen to be significant in the DP GLM at the significance level of 0.10. The shape parameters in both models are significantly larger than 1, both of which confirm the data are under-dispersed.

Table 5.4 Comparison between the DP GLM and Com-Poisson GLM using Korea data

Estimated Parameters and Standard Errors		
Variables	**COM-Poisson**	**DP**
Constant	-6.657 (1.206)	**-5.2031 (0.7363)** [c]
Ln(AADT)	0.648 (0.139)	**0.4393 (0.07231)**
Presence of commercial area	1.474 (0.513)	**0.8725 (0.3023)**
Train detector distance	0.0021 (0.0007)	**0.0017 (0.0005)**
Presence of track circuit controller	-1.305 (0.431)	**-0.7121 (0.2491)**
Presence of guide	-0.998 (0.512)	**-0.5852 (0.3184)**
Presence of speed hump	-1.495 (0.531)	**-1.0977 (0.3476)**
Shape parameter	2.349 (0.634)	**1.5344 (0.1705)**
Goodness-of-fit Statistics		
AIC	210.7	202.5
MPB	-0.007	1.60748E-11
MAD	0.348	0.363
MSPE	0.236	0.253

[a] Based on the modeling results for the COM-Poisson GLM documented in Lord et al. (2010).
[b] Values in parentheses are the standard errors for the estimated parameters.
[c] Bolded value indicates the related variable in the DP GLM is significant at the significance level of 0.10.

As for the GOF statistics, the two models are almost the same. The DP GLM has smaller values for AIC and MPB while the COM-Poisson GLM has slightly smaller values of MAD and MSPE. So at this time, there is no conclusions made on one model is better than the other. It is important to note that all the estimated coefficients in the COM-Poisson GLM are for its centering parameter "λ" (Sellers and Shmueli, 2010) and not for the mean, as in the case of the DP GLM. So the coefficients for the two models are somewhat different and they cannot be compared directly. But this does not affect

the comparison on the GOF, since the mode μ in the COM-Poisson GLM can be calculated from the predicted λ and ν (recall that $\mu = \lambda^{1/\nu}$) and then the predicted mean of the COM-Poisson can be obtained by the method we documented in Section 4.2.

5.5.2 Overall comparison

Given that in the above comparisons the DP GLM only includes the variables that make the gamma count model or the COM-Poisson GLM to be optimal, the DP GLM still has the potential to be better by adding or deleting other explanatory variables. Thus, there is a need to compare the models each of which is at their optimal. By penalizing models with number of parameters, AIC was used to select the most parsimonious model that can best explain the data with a minimum of parameters. The model with a certain combination of variables that achieves the smallest value of AIC was deemed as its optimal. It should be noted that all the models in this study are the main effects model and interaction effects are not considered in this study. To avoid multicollinearity, the correlation matrix between variables was analyzed to delete redundant variables in those models. The level of significance for variable selection was 0.10.

Table 5.5 shows that different models achieve at their own optimal with different significant variables. The optimal DP GLM, COM-Poisson GLM and gamma count model have a total of 8, 6 and 6 significant variables respectively. The variables that are significant in all of the three models are: Ln(ADT), presence of commercial area, train detector distance, and presence of speed hump. The variables that are only significant in the DP GLM are: number of tracks and gradient of road. The variables found to be only

significant in the gamma count model are: average daily railway traffic and time duration between the activation of warning signals and gates. All the variables found to be significant in the COM-Poisson are also significant in the DP model.

Table 5.5 Significant variables in three different models

Variables	**COM-Poisson**	**Gamma**	**DP**
Number of Tracks	-	-	+
Ln(ADT)	+	+	+
Average daily railway traffic	-	+	-
Gradient of Road	-	-	+
Presence of commercial area	+	+	+
Train detector distance	+	+	+
Time duration between the activation of warning signals and gates	-	+	-
Presence of track circuit controller	+	-	+
Presence of guide	+	-	+
Presence of speed hump	+	+	+

Note: "+" for "significant" and "-" for "not significant" at the significance level of 0.10.

Table 5.6 shows the comparison results among all the three models when each model achieves their optimal. In Table 5.6, the DP GLM is shown to better fit the data than any other two models since it has the smallest value in all the GOF statistics. Another interesting point is the variable *number of tracks* which are found to be only significant in the DP GLM are highly significant and has an great marginal effect. According to the coefficients of the variable *number of tracks* in the DP GLM, increasing one track would lead to 2.55 (=exp(0.9343)) times increase in the expected

crash number. According to the signs of the estimated parameters found significant in all of the three models, the increase in *AADT* and *train detector distance* would enhance the expected crash number. The presence of commercial area would increase the expected crash number whereas the presence of speed hump would decrease the expected crash number.

Table 5.6 Comparison among three models when each model at their optimal

Estimated Parameters and Standard Errors			
Variables	**COM-Poisson** [a]	**Gamma** [b]	**DP**
Constant	-6.657 (1.206) [c]	-3.438 (1.008)	-7.1506 (0.8921)
Number of Tracks	-	-	0.9343 (0.2466)
Ln(ADT)	0.648 (0.139)	0.230 (0.076)	0.5225 (0.0748)
Average daily railway traffic	-	0.004 (0.0024)	-
Gradient of Road	-	-	-0.0657 (0.0310)
Presence of commercial area	1.474 (0.513)	0.651 (0.287)	1.0354 (0.3063)
Train detector distance	0.0021 (0.0007)	0.001 (0.0004)	0.0015 (0.0005)
Time duration between the activation of warning signals and gates	-	0.004 (0.002)	-
Presence of track circuit controller	-1.305 (0.431)	-	-0.6554 (0.2314)
Presence of guide	-0.998 (0.512)	-	-0.5338 (0.3702)
Presence of speed hump	-1.495 (0.531)	-1.58 (0.859)	-1.2149 (0.3359)
Shape parameter	2.349 (0.634)	2.062 (0.758)	1.6835 (0.1875)
Goodness-of-fit Statistics			
AIC	210.7	211.38	191.5
MPB	-0.007	0.179	2.98357E-11
MAD	0.348	0.459	0.334
MSPE	0.236	0.308	0.234

[a] Based on the modeling results for COM-Poisson GLM documented in Lord et al. (2010).
[b] Based on the modeling results for gamma count model documented in Lord et al. (2010).
[c] Values in parentheses are the standard errors for the estimated parameters.

Figure 5.1 shows the distributions of crash numbers for the observed and predicted crashes of the three models each at their optimal for Korea data. The observed crashes are more scattered than the three types of predicted crashes.

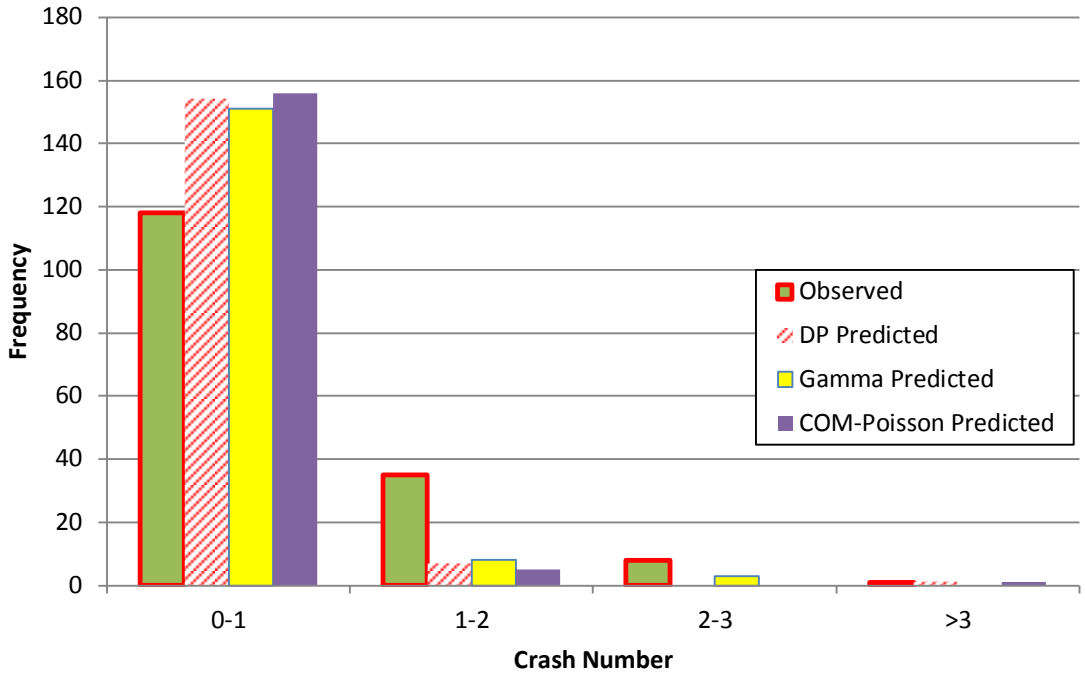


Figure 5.1 Frequencies of observed and predicted crashes for the Korea Data

Figure 5.2 illustrates the scatter plot for the predicted vs. observed crashes for each site. All the three models have shown to over-predict the crashes when observed crash number is equal to 0 and under-predict the crashes when observed crash number is equal to 2. Since most of the data points are with the range of the observed crashes equal to 0, 1 and 2, there is much overlapping between the data predicted by the three models and

thus we cannot draw conclusions on the comparison among those models from Figure 5.2.

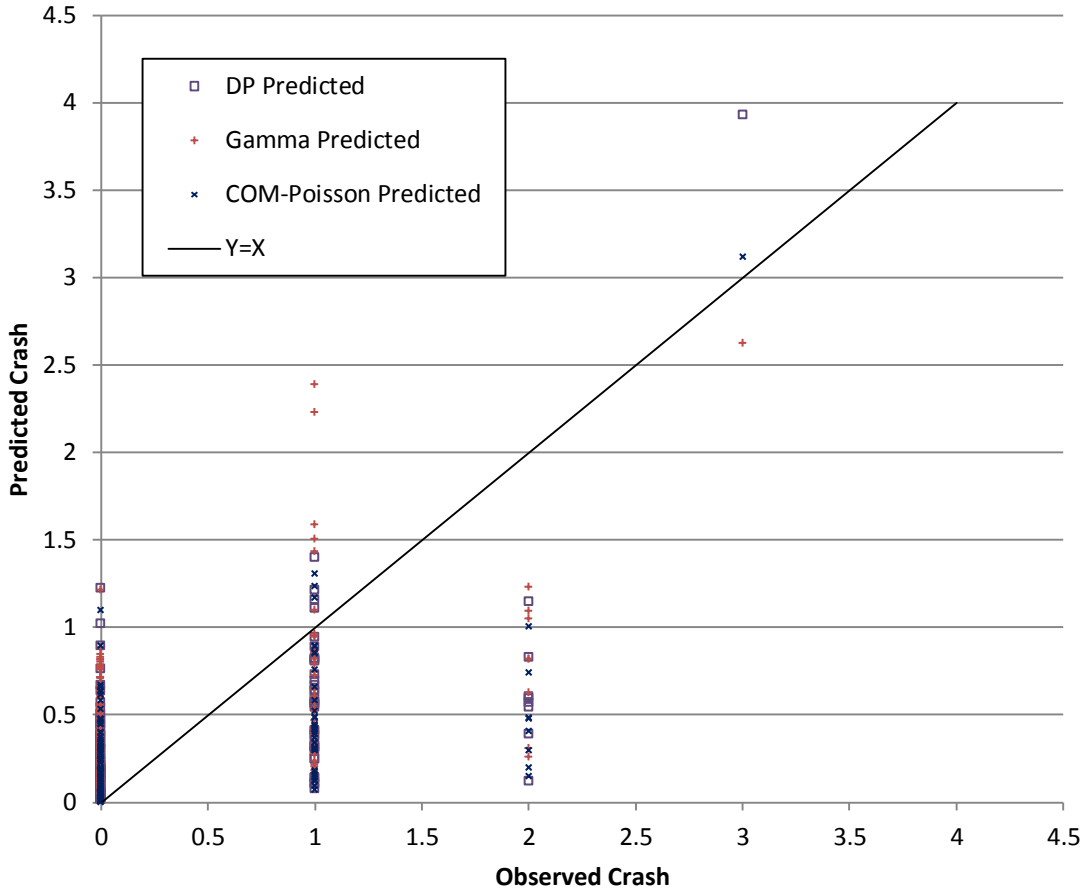


Figure 5.2 Predicted vs. observed crashes for the Korea Data

Figure 5.3 demonstrates the predicted crashes of the three models each at their optimal against the variable ADT when controlling other significant variables at their average. Thus, a direct comparison cannot be conducted since each model uses different input variables. Figure 5.3 shows that the three models share a similar trend on how the predicted crashes vary with the increase of the ADT.

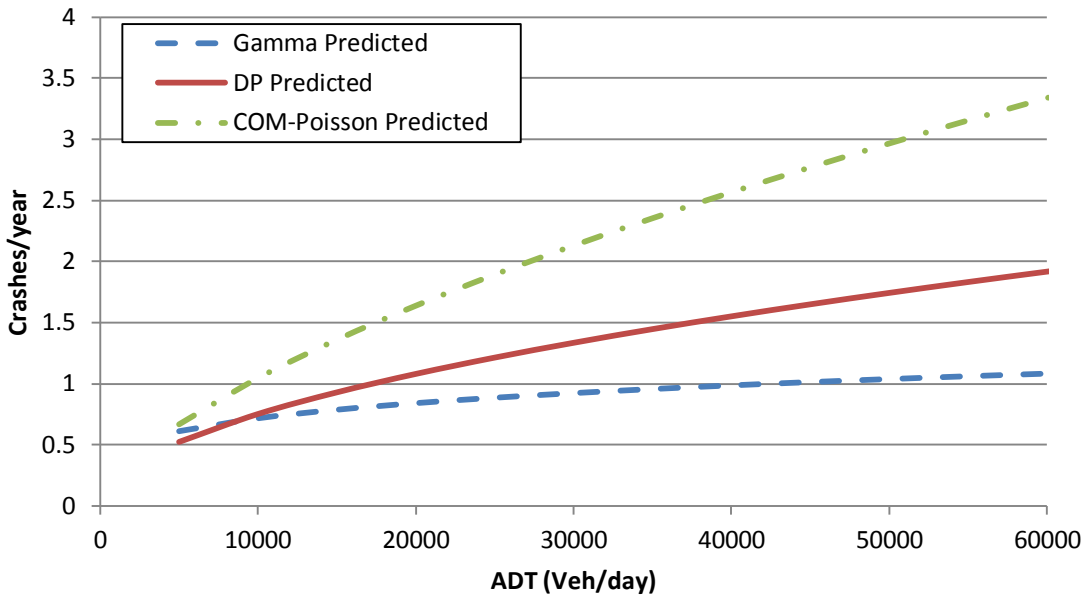


Figure 5.3 Estimated values for the Korea data (against ADT variable)

Figure 5.4 shows the adjusted CURE plot for the variable ADT. The DP GLM and the COM-Poisson GLM provide a similarly good fit to the data with curves of both models oscillating closely around the X axis. The performance of the gamma count model has been found to be worse than the other two models, since the difference

between the predicted and observed crashes for the gamma count model is almost always negative when Ln(ADT) is less than 6 and positive when Ln(ADT) is larger than 6.

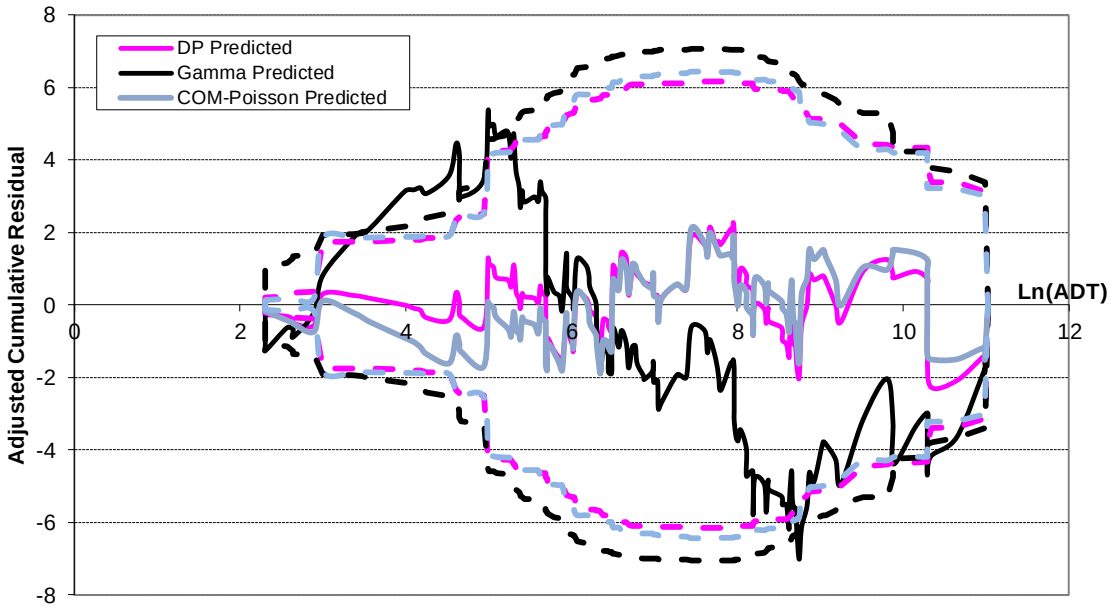
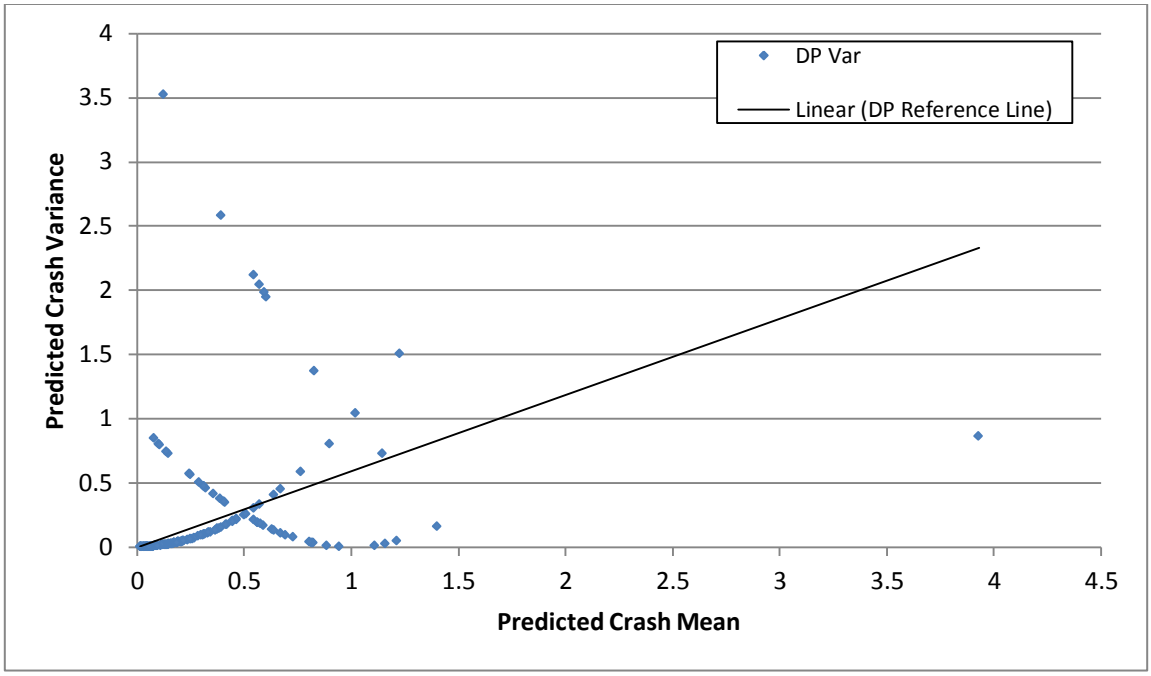
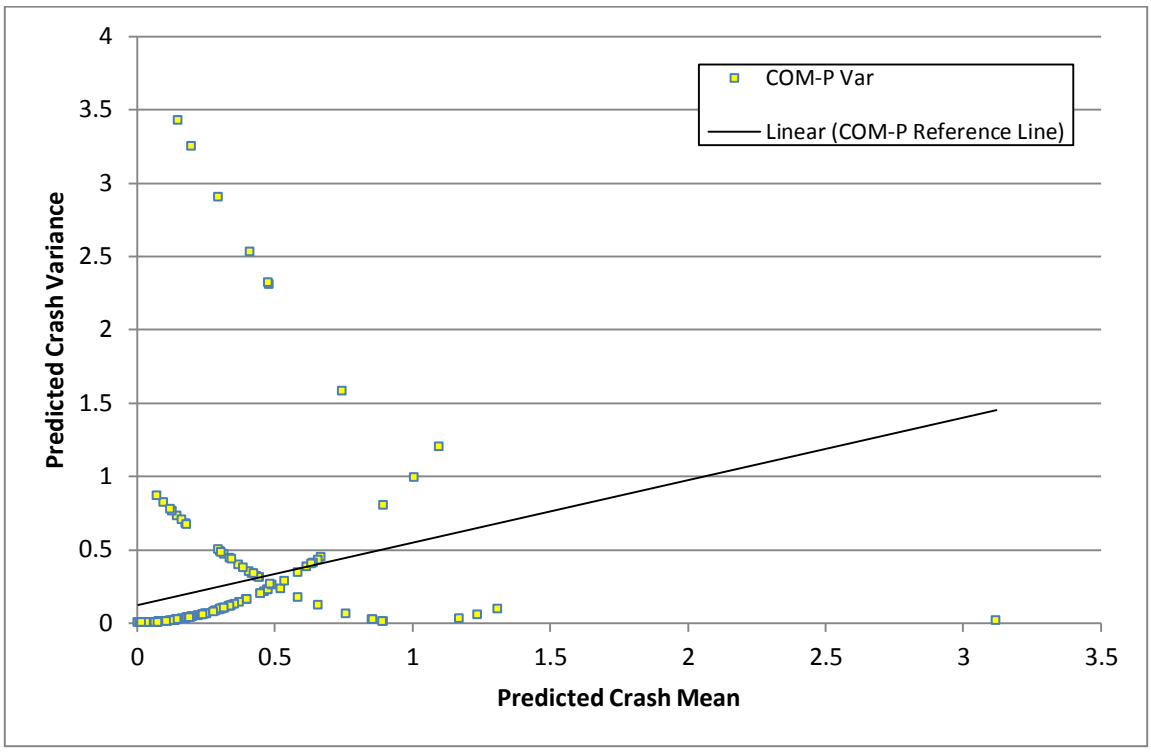


Figure 5.4 Cumulative residual plots for the Korea data (against ADT variable)
Note: Dotted lines represent ±2 standard deviance

Figure 5.5 presents the comparison on the variances predicted by the DP GLM and the COM-Poisson GLM. Due to the limitation of the sample size, however, we cannot see if the predicted crash variance vs. predicted mean follow the pattern of their theoretical relationship.



a) DP GLM



b) COM-Poisson GLM

Figure 5.5 Predicted crash variance vs. predicted crash mean for the Korea data

5.6 Discussion

This subsection will give a detailed discussion on the overall performances of the DP GLM, COM-Poisson GLM and gamma count model for analyzing under-dispersed data based on the previous results.

- Goodness-of-fit

  A pairwise comparison between the DP GLM and the gamma count model using the variables originally found significant in the gamma count model (Oh et al., 2006) was first conducted. It shows that the DP GLM is found to provide a better fit than the gamma count model. From another similar pairwise comparison between the DP GLM and COM-Poisson GLM using the variables originally found significant in the COM-Poisson GLM (Lord et al., 2010), the DP GLM fits the data as well as the COM-Poisson GLM. Given that in the above two pairwise comparisons the DP GLM is not at its optimal as the case with the other two models, we further conduct an overall comparison among the three models in which each model achieves at their optimal. The overall comparison shows that the DP GLM fits the data much better than the other two models with its smallest values in all the different GOF statistics. It should be noted that in the overall comparison, the difference in the GOF statistics between the DP GLM and COM-Poisson GLM is minor with the exception of the AIC value.

- Estimated Parameters and Standard Errors

  The results of the estimated parameters for the DP GLMs indicate that the DP GLM can detect the presence of under-dispersion with its estimated shape parameter being significantly bigger than 1 ($Var(Y) \approx \mu / \theta$). The DP GLMs in this section estimate their parameters based on MLE and use the software SAS to code their mathematical functions. As we mentioned in Section 4, MLE of the DP GLM was an approximation and it was achieved based on the approximate PMF of the DP distribution.

  The COM-Poisson GLM in this section was developed based on MLE in R (Sellers and Shmueli, 2010) which is different with that in Section 4. However, its use of the parameter λ rather than the mean in the link function is not intuitive and much more complex.

  In this section and the last section, it has been found that the DP GLM tends to provide smaller standard errors of the estimated coefficients than other models. This makes the DP GLM more prone to threat variables as being significant. The dominance of the DP GLM in this section might be related to its inclusion of more significant variables than the other models. For those variables only significant in the DP GLM, adding them very likely reduces the AIC value more substantially than does for the other GOF statistics. This might explain why the difference in the AIC value between the DP GLM and COM-Poisson GLM is

much larger than the other GOF statistics. It should be noted that we cannot directly compare the estimated parameters and standard errors of the DP GLM to those of the COM-Poisson GLM due to the different choices of centering parameter and corresponding link function in the two models in this section.

- Mathematical Relationship

  All the three models we discussed in this section have two parameters: a centering parameter and shape parameter (or dispersion parameter). For the DP GLM, it uses the mean as its centering parameter. The shape parameter of the DP GLM has been found to be very flexible and it can handle under-, equi-, and over-dispersion ($Var(Y) \approx \mu / \theta$, $\theta > 1$ for under-dispersion; $\theta = 1$ for equi-dispersion; $\theta < 1$ for over-dispersion).

  As we mentioned before, the COM-Poisson model in this section used the parameterization which cannot provide a clear centering parameter as the mean of the DP model. For the gamma count model, it assumes the observations are dependent where the observation at time t-1 directly influences the observation at time t. This assumption is possible for some datasets but not realistic for most datasets. For instance, a crash occurred at time t cannot directly influence another one that will occur six months after the first event.

- Computational Time

  The computational time for the DP GLM in this section (under-dispersed data) is consistent with that in the last section (over-dispersed data). With the use of MLE-based GLM framework, the computational time for the COM-Poisson GLM in this section was greatly reduced compared with that under the Bayesian-based framework in Section 4. However, it still takes several minutes to converge and much longer than the DP GLM which generates the result in less than ten seconds.

5.7 Summary

This section evaluated the application of the DP GLM in analyzing motor vehicle crash data characterized by under-dispersion. In rare occasions, the traffic crash data are under-dispersed. The presence of data characterized by under-dispersion makes the most commonly used models such as the Poisson and NB unable to handle those traffic crash data. The objective of the study in this section is to compare the performance of the DP GLM with other models in handling data characterized by under-dispersion. Comparison analysis was conducted between the three models (the DP GLM, COM-Poisson GLM and gamma count model) that can handle under-dispersed data. The under-dispersed dataset comes from Oh et al.'s study (2006), in which the data were used for establishing the gamma count model. Several GOF statistics were used to compare the performances of those models.

In pairwise comparisons, the DP GLM only includes the variables found significant in the models to which DP GLM compares. The pairwise comparison results indicate that the DP GLM fits the data much better than the gamma count model and almost the same with the COM-Poisson GLM. For the overall comparison in which the three models are put together and each achieves their own optimal, the DP GLM has been found to give a much better fit than the other two.

Considering all the comparison results obtained from this section, we can conclude that the DP GLM fits the data better than the gamma count model and COM-Poisson GLM. The DP GLM can detect the presence of under-dispersion. The shape parameter of the DP GLM has been found to be very flexible and it can handle under-, equi-, and over-dispersion ($Var(Y) \approx \mu / \theta$, $\theta > 1$ for under-dispersion; $\theta = 1$ for equi-dispersion; $\theta < 1$ for over-dispersion). The centering parameter in the COM-Poisson distribution based on which the COM-Poisson GLM is developed is not intuitive as that for the DP distribution. The gamma count model on the other hand suffers greatly from its theoretical background. The computational time for the DP GLM was a little bit quicker than that for the COM-Poisson in this section.

Thus, the overall performance of the DP GLM is much better than that of the COM-Poisson GLM and gamma count model in handling the under-dispersed crash data.

6. SUMMARY AND CONCLUSIONS

In traffic safety analysis, a large number of distributions have been proposed to analyze the number of vehicle crashes. Among those distributions, the traditional Poisson and Negative Binomial have been the most commonly used probabilistic structures of models. Although the Poisson and NB models possess desirable statistical properties, their application on modeling motor vehicle crashes are associated with limitations. In practice, traffic crash data are often over-dispersed. On rare occasions, they have shown to be under-dispersed. The over-dispersed and under-dispersed data would lead to the inconsistent standard errors of parameter estimates using the traditional Poisson distribution. Although the NB has been found to be able to model over-dispersed data, it cannot handle under-dispersed data.

In light of the difficulties raised by the Poisson and NB models, many new statistical methods have been proposed to handle the over-dispersed and under-dispersed count data. Among those distributions proposed to handle under-dispersed data, the COM-Poisson and DP distributions are particularly noteworthy with each distribution being capable of handling data characterized by under-, equi- and over-dispersion. The COM-Poisson distribution and its GLM have been found to be very flexible to handle count data. While for the DP, its distribution and GLM framework has seldom been investigated and applied since its first introduction 25 years ago.

Therefore, the primary objectives of this research were to: 1) examine the applicability of the DP distribution and its regression model for analyzing crash data

characterized by over- and under-dispersion, and 2) compare the performances of the DP distribution and DP GLM with those of the COM-Poisson distribution and COM-Poisson GLM in terms of GOF and theoretical soundness.

This section first presents the summary of work in this research and then discusses the possible directions for the future work.

6.1 Summary of Work

This subsection briefly describes how the research was conducted and highlights the main findings in this research.

6.1.1 Evaluation of the performance of the DP distribution

The first part of the research work on this thesis was documented in Section 3 and related to the evaluation of the performance of the DP distribution with no covariates considered. As discussed in Sections 1 and 2, very few researchers have applied or used the DP distribution or its regression model for analyzing count data since its introduction. For a new distribution like the DP, it is important to first evaluate the distribution under the wide variety of situations before dealing with the regression model.

This part of research work was accomplished using simulated data for nine different mean-dispersion relationships (or scenarios). Five runs each with 2,000 observations were simulated under each scenario by different distributions. The under-dispersed data were simulated by the COM-Poisson distribution; the equi-dispersed data were simulated by the COM-Poisson and Poisson distributions; and the over-dispersed

data were simulated by the COM-Poisson and NB distributions. For each scenario, different distributions were fitted based on their characteristics of handling dispersion. All scenarios of data were fitted using the DP and COM-Poisson. The gamma count, Poisson and NB were only employed to fit under-dispersed data, equi-dispersed data and over-dispersed data, respectively. Four different GOF statistics were used to evaluate and compare the performances of different distributions.

It was found that the COM-Poisson performs better than the DP for all nine scenarios. It should be noted that the comparison on their performance of handling under-dispersed count data is yet to be determined since all the under-dispersed data in this section were simulated by the COM-Poisson and the COM-Poisson may be expected to generate better results than other distributions. Another main finding is that the DP works better for high mean scenarios independent of the type of dispersion. The lack of fit for the DP in low mean scenarios is due to its inadequacy for fitting "zero" observations.

6.1.2 Comparison of GLM performance for over-dispersed data

The second part of research work on this thesis was documented in Section 4. It was related to the application of the DP GLM for analyzing motor vehicle crash data characterized by over-dispersion, and the comparison analysis between the DP GLM, NB model and COM-Poisson GLM. This study was motivated by the fact that no model was able to replace the NB models for analyzing over-dispersed data. Previous research has found that the COM-Poisson GLM can fit the over-dispersed data as well as the NB

model. Thus, the results of the COM-Poisson GLM were also provided as a reference. Although the results in Section 3 has demonstrated that the DP distribution can handle over-, equi-, and under-dispersed count data, it is of more interest to see how good the DP GLM can link the crash data to the variables that affect traffic safety and how much influence those variables can affect the expected crashes.

Two observed over-dispersed datasets along with two different and commonly used link functions were used to establish the GLMs in order to eliminate the potential bias of using only one dataset or one link function. Several measures of GOF were used to compare performances of the DP GLM, NB model and COM-Poisson GLM.

It was found that the DP GLM fits the over-dispersed data almost the same as the NB model and COM-Poisson GLM. The DP GLM provides very similar estimates for those coefficients as the NB model. The mathematics to manipulate the DP GLM is very simple and very similar to the NB model, since both of the two models use the mean as the centering parameter and have a shape parameter to handle the presence of dispersion.

The advantage of the DP GLM over the NB model lies in the smaller standard errors of its estimated coefficients and the flexibility of its shape parameter. The standard errors of estimated parameters for the DP GLM were found to be smaller than those for the other two models, indicating the DP GLM can more precisely describe the effects of the interested covariates. Moreover, the shape parameter of the NB model was only intended for the over-dispersed data ($Var(Y) = \mu + \alpha\mu^2$) while the shape parameter for the DP GLM can handle under-, equi-, and over-dispersion ($Var(Y) \approx \mu / \theta$) (the performance of the DP GLM handling under-dispersed data is presented in Section 5).

The advantage of the DP GLM over the COM-Poisson is found to be related to the parameter estimation method. The DP GLM and NB model were developed based on the MLE, which was not the case with the COM-Poisson GLM developed on the Bayesian framework in this section. The computational time for the DP GLM and NB GLM using MLE framework was substantially quicker than that for the COM-Poisson GLM using Bayesian framework. Although the MLE recently becomes available and the computational time has been greatly reduced for the COM-Poisson, its corresponding GLM framework is linked to the parameter which cannot serve as a clear centering parameter as the mean in the DP GLM and NB model.

6.1.3 Comparison of GLM performance for under-dispersed data

The third part of research work on this thesis was documented in Section 5. It was related to the application of the DP GLM for analyzing motor vehicle crash data characterized by under-dispersion, and the comparison analysis between the DP GLM, COM-Poisson GLM and gamma count model. This research was motivated by the fact that the DP and COM-Poisson models can overcome the difficulty of handling under-dispersed data raised by the NB model.

The under-dispersed dataset comes from Oh et al.'s study (2006), in which the data were used for establishing the gamma count model. Several GOF statistics were used to compare the performances of those models. Pairwise comparisons were first conducted between the DP GLM with other two models (the COM-Poisson GLM and gamma count model) that can handle under-dispersed data. In pairwise comparisons, the DP GLM

only includes the variables found significant in the models to which the DP GLM compares. Then an overall comparison among the three models was provided. In the overall comparison, the three models are put together and each achieves their own optimal with different input variables.

It was found from the pairwise comparisons that the DP GLM fits the data much better than gamma count model and almost the same with the COM-Poisson GLM. For the overall comparison in which each model achieves their own optimal, the DP GLM was found to provide a much better fit than the other two. Considering all the comparison results in Section 5, a conclusion was made that the DP GLM fits the data better than the gamma count model and COM-Poisson GLM.

It is important to note that the DP GLM also has its theoretical appeal. The shape parameter of the DP GLM was found to be very flexible and it can handle under-, equi-, and over-dispersion ($Var(Y) \approx \mu / \theta$, $\theta > 1$ for under-dispersion; $\theta = 1$ for equi-dispersion; $\theta < 1$ for over-dispersion). The centering parameter of the DP GLM is the mean, and it is more intuitive to interpret than the mode as the centering parameter in the COM-Poisson GLM. The gamma count model on the other hand suffers from its theoretical background.

6.2 Future Research Areas

According to the limitations of this research, the following recommendations are provided for future research:

- As discussed in Section 3, the lack of fit for the DP distribution in low mean scenarios is due to its inadequacy for fitting "zero" observations. Further research should be conducted on explaining the effects of data characterized by low sample mean on the DP model. Due to the wide existence of large number of zeros in crash data, it is important to investigate how much the GOF, parameter estimates and confidence intervals might be biased when fitting the DP model to the data characterized by the low sample mean.

- As mentioned in Section 3, in the approximate PMF of the DP distribution (see Equation (2.32)), the denominator is zero for observation equal to zero, which is not solvable. To circumvent this problem, the author calculated the limits of the likelihood when observation value approached zero in writing the thesis. The validity and accuracy of this approach worth further investigation

- The difference of the exact and approximate DP distribution lies in if the normalizing constant (see Equation (2.34)) is used. The normalizing constant was not considered in developing the DP GLM in Sections 4 and 5 due to the increased non-linearity in the PMF when using it. Since it has been found that the inclusion of the approximation to the normalizing constant proposed by Efron (1986) does not improve the model in Section 4, it is recommended to conduct research on the more accurate approximation methods for the normalizing constant and evaluate their effects on the DP GLM.

- Bayesian method has been commonly used for crash models to estimate their parameters. The research on the DP models could be extended by developing the DP GLM based on the Bayesian framework. Since the Bayesian estimation method is capable of handling very complex models, it has the potential to overcome the difficulty in estimating the parameters of the exact DP distribution, whose PMF incorporates the normalizing constant and poses computation challenges when using MLE in SAS.

- In this study, the GLM of each distribution is only linked to the centering parameter. Further research could be conducted on assessing the performance of the double-link DP model, in which both the centering parameter and shape parameter are linked to covariates.

- The NLMIXED procedure in SAS has an interface to include the random effects in the regression model. Corresponding research could be conducted on assessing the performance of the random effects DP model.

REFERENCES

AASHTO, 2010. Highway Safety Manual. American Association of State and Highway Transportation Officials, Washington, D.C.

Abdelwahab, H.T., Abdel-Aty, M.A., 2002. Artificial neural networks and logit models for traffic safety analysis of toll plazas. Transportation Research Record 1784, 115-125.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on accident frequencies and severities. Accident Analysis and Prevention 33 (1), 99-109.

Casella, G., Berger, R.L., 1990. Statistical Inference. Wadsworth Brooks/Cole, Pacific Grove, CA.

Castillo, J and Perezcasany,M.,2005. Over-dispersed and under-dispersed Poisson generalizations. Journal of Statistical Planning and Inference 134 (2), 486-500.

Clark, S.J., Perry, J.N., 1989. Estimation of the negative binomial parameter k by maximum quasi-likelihood. Biometrics 45, 309-316.

Consul, P.C., 1989. Generalized Poisson Distributions: Properties and Applications, Marcel Dekker, New York.

Conway, R.W., Maxwell, W.L., 1962. A queuing model with state dependent service rates. Journal of Industrial Engineering 12, 132-136.

Daniels, S., Brijs, T., Nuyts, E., Wets, G., 2010. Explaining variation in safety performance of roundabouts. Accident Analysis and Prevention 42 (2), 393-402.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis & Prevention 40 (4), 1257-1266.

Efron, B., 1986. Double exponential families and their use in generalized linear Regression. Journal of the American Statistical Association 81 (395), 709-721.

Famoye, F., 1993. Restricted generalized Poisson regression model. Communications in Statistics - Theory and Methods 22 (5), 1335-1354.

Francis, R.A., Geedipally, S.R., Guikema, S.D., Dhavala, S.S., Lord, D., and Larocca, S., 2012. Characterizing the performance of the Conway-Maxwell-Poisson generalized linear model. Risk Analysis 32 (1), 167-183.

Geedipally, S.R., 2008. Examining the Application of Conway-Maxwell-Poisson Model for Analyzing Traffic Crash Data. Ph.D. Dissertation, Department of Civil Engineering, Texas A&M University, College Station, Texas.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The Negative-Binomial-Lindley generalized linear model: characteristics and application using Crash Data. Accident Analysis & Prevention 45 (2), 258-265

Greene, W.H., 2002. A Review of LIMDEP 8.0: A Powerful and Versatile Package for Econometric Analysis, New York.

Guikema, S.D., Coffelt, J.P., 2008. A flexible count data regression model for risk analysis. Risk Analysis 28 (1), 213-223.

Hauer, E., 1997. Observation Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Elsevier Science Ltd., Oxford.

Hilbe, J.M., 2007. Negative Binomial Regression (Second Edition). Cambridge University Press, Cambridge, UK.

Kadane, J.B., Shmueli, G., Minka, T.P., Borle, S., Boatwright, P., 2006. Conjugate analysis of the Conway–Maxwell–Poisson distribution. Bayesian Analysis 1, 363–374.

Lao, Y., Wu, Y.-J., Corey, J., Wang, Y., 2011. Modeling animal-vehicle collisions using diagonal inflated bivariate Poisson regression. Accident Analysis & Prevention 43 (1), 220-227.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis & Prevention 40 (4), 1611-1618.

Lord, D., 2000. The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models. Ph.D. dissertation, Department of Civil Engineering, University of Toronto, Toronto, Ontario.

Lord, D., Geedipally, S.R., 2011. The Negative Binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. Accident Analysis and Prevention 43 (5), 1738-1742.

Lord, D., Geedipally, S.R., Guikema, S., 2010. Extension of the application of Conway-Maxwell-Poisson models: analyzing traffic crash data exhibiting under-dispersion. Risk Analysis, 30 (8) 1268-1276.

Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., van Schalkwyk, I., Ivan, J.N., Lyon, C., Jonsson, T., 2008b. Methodology for estimating the safety performance of multilane rural highways. NCHRP Web-Only Document 126, National Cooperation Highway Research Program, Washington, D.C. (http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp w126.pdf, retrieved on February 2011).

Lord, D., Guikema, S.D., Geedipally, S.R., 2008a. Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis and Prevention 40 (3), 1123-1134.

Lord, D., Mannering, F.L., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice 44 (5), 291-305.

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Safety Science 46, 751-770.

Lord, D., Park, P.Y-J., 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. Accident Analysis and Prevention 40 (4), 1441-1457.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis & Prevention 37 (1), 35-46.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964-975.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. Transportation Research Record 1840, 31-40.

Oh, J., Washington, S.P., Nam, D., 2006. Accident prediction model for railway-highway interfaces. Accident Analysis & Prevention 38 (2), 346-356.

Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention 41 (4), 683-691.

Park, B.-J., Lord, D., 2008. Adjustment for the maximum likelihood estimate of the negative binomial dispersion parameter. Transportation Research Record 2061, 9-19.

Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1-6.

Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis & Prevention 36 (2), 183-191.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. Retrieved July 2011 from http://www.R-project.org.

SAS Institute Inc., 2002. Version 9 of the SAS System for Windows. Cary, NC.

Saha, K., Paul, S., 2005. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. Biometrics 61 (3), 179-185.

Sellers, K. F., Borle, S., Shmueli, G., 2011. The COM-Poisson model for count data: a survey of methods and applications. Applied Stochastic Models in Business and Industry 28 (2).

Sellers, K.F. and Shmueli, G., 2010. A flexible regression model for count data. Annals of Applied Statistics 4 (2), 943-961.

Shankar, V., Milton, J., Mannering, F.L., 1997. Modeling accident frequency as zero-altered probability processes: an empirical inquiry. Accident Analysis & Prevention 29 (6), 829-837.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P., 2005. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. Journal of the Royal Statistical Society, Part C 54, 127-142.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Lun, D., 2003. WinBUGS Version 1.4.1 User Manual. MRC Biostatistics Unit, Cambridge. Available from: <http://www.mrcbsu.cam.ac.uk/bugs/welcome.shtml>.

The MathWorks Inc, 2011. MATLAB 7.12.0 R2011a, Natick, MA.

Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics 31 (2&3), 221-229.

Washington, S.P., Karlaftis, M., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall, Boca Raton, FL.

Wedagama, D.M., Bird, R.N., Metcalf, A.V., 2006. The influence of urban landuse on non-motorised transport casualties. Accident Analysis & Prevention 38 (6), 1049-1057.

Winkelmann, R., 1995. Duration dependence and dispersion in count-data models. Journal of Business Economic Statistics 13 (4), 467-474.

Winkelmann, R., 2008. Econometric Analysis of Count Data (Fifth Edition). Springer-Verlag, Berlin.

Winkelmann, R., Zimmermann, K., 1995. Recent development in count data modeling: theory and applications. Journal of Economic Surveys 9 (1), 1-24.

Xie, Y., Lord, D., Zhang,Y., 2007. Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis. Accident Analysis and Prevention 39 (5), 922-933.

APPENDIX

THE SIMULATION RESULTS FOR COMPARING DIFFERENT DISTRIBUTIONS

TABLE A.1 Results of Under-dispersion and High Mean Scenario (COM-Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters*		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	4.856	3.773	0.777	DP	4.856	1.225	12.7	13.1	-4148.9	9	1.41
				COM-P	4.979	1.304	10.7	10.8	-4147.8	9	1.19
				Gamma	--	--	10.3	10.4	-4147.7	9	1.15
Run2	4.856	3.922	0.808	DP	4.856	1.186	7.0	7.3	-4179.4	9	0.78
				COM-P	4.962	1.262	5.3	5.4	-4178.4	9	0.59
				Gamma	--	--	5.2	5.2	-4178.4	9	0.58
Run3	4.880	3.850	0.789	DP	4.880	1.223	10.5	11.4	-4157.3	9	1.17
				COM-P	4.962	1.301	10.6	10.8	-4156.9	9	1.17
				Gamma	--	--	9.7	10.1	-4156.5	9	1.08
Run4	4.905	3.936	0.803	DP	4.905	1.199	6.5	6.6	-4182.1	9	0.72
				COM-P	5.018	1.271	5.9	6.0	-4181.7	9	0.66
				Gamma	--	--	6.0	6.1	-4181.7	9	0.66
Run5	4.885	3.8317	0.784	DP	4.885	1.222	7.4	7.4	-4158.2	9	0.82
				COM-P	5.011	1.304	5.6	5.6	-4157.2	9	0.63
				Gamma	--	--	5.4	5.4	-4157.1	9	0.60
Average	--	--	--	DP	--	--	8.8	9.2	-4165.2	9	0.98
				COM-P	--	--	7.6	7.7	-4164.4	9	0.85
				Gamma	--	--	**7.3**†	**7.4**	**-4164.3**	**9**	**0.81**

*For the DP, the estimated parameters are μ and θ; for COM-Poisson, the estimated parameters are μ and ν. (Note: same for all other tables.)

†Bold value indicates best goodness-of-fit. (Note: same for all other tables.)

TABLE A.2 Results of Under-dispersion and Medium Mean Scenario (COM-Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	1.896	1.501	0.792	DP	1.896	1.095	22.3	24.7	-3177.9	4	5.58
				COM-P	2.046	1.374	5.0	5.0	-3167.8	4	1.24
				Gamma	--	--	5.0	5.0	-3167.8	4	1.26
Run2	1.838	1.564	0.851	DP	1.838	1.023	22.2	23.6	-3198.5	5	4.45
				COM-P	1.956	1.263	1.3	1.3	-3187.9	4	0.31
				Gamma	--	--	1.3	1.3	-3187.9	4	0.31
Run3	1.890	1.540	0.815	DP	1.890	1.070	22.4	24.8	-3194.4	5	4.48
				COM-P	2.031	1.338	5.7	5.5	-3184.6	4	1.42
				Gamma	--	--	5.8	5.7	-3184.7	4	1.45
Run4	1.828	1.509	0.826	DP	1.828	1.049	23.7	24.1	-3170.1	4	5.91
				COM-P	1.960	1.312	3.8	3.7	-3159.1	4	0.96
				Gamma	--	--	3.9	3.8	-3159.1	4	0.98
Run5	1.855	1.521	0.820	DP	1.855	1.043	24.2	26.5	-3189.3	5	4.84
				COM-P	1.984	1.304	2.8	2.8	-3177.8	4	0.69
				Gamma	--	--	2.6	2.6	-3177.7	4	0.65
Average	--	--	--	DP	--	--	23.0	24.7	-3186.0	4.6	4.99
				COM-P	--	--	**3.7**	**3.7**	**-3175.4**	**4.0**	**0.92**
				Gamma	--	--	**3.7**	**3.7**	**-3175.4**	**4.0**	**0.93**

TABLEA.3 Results of Under-dispersion and Low Mean Scenario (COM-Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	0.383	0.363	0.950	DP	0.383	1.112	3.6	3.4	-1583.4	1	3.65
				COM-P	0.454	1.183	2.3	2.3	-1582.9	1	2.32
				Gamma	--	--	2.1	2.0	-1582.7	1	2.09
Run2	0.393	0.363	0.924	DP	0.393	1.115	3.2	3.2	-1600.2	1	3.17
				COM-P	0.533	1.363	0.4	0.4	-1598.8	1	0.38
				Gamma	--	--	0.4	0.4	-1598.8	1	0.37
Run3	0.361	0.342	0.948	DP	0.361	1.137	5.1	4.6	-1529.6	1	5.11
				COM-P	0.453	1.237	2.6	2.4	-1528.5	1	2.56
				Gamma	--	--	2.5	2.3	-1528.4	1	2.46
Run4	0.401	0.361	0.902	DP	0.401	1.120	4.3	4.3	-1612.0	1	4.35
				COM-P	0.575	1.470	0.1	0.1	-1609.7	1	0.06
				Gamma	--	--	0.0	0.0	-1609.7	1	0.01
Run5	0.391	0.3713	0.950	DP	0.391	1.101	1.6	1.6	-1604.3	1	1.63
				COM-P	0.477	1.208	0.3	0.3	-1603.6	1	0.28
				Gamma	--	--	0.2	0.2	-1603.6	1	0.23
Average	--	--	--	DP	--	--	3.6	3.4	-1585.9	1.0	3.58
				COM-P	--	--	1.1	1.1	-1584.7	1.0	1.12
				Gamma	--	--	**1.0**	**1.0**	**-1584.6**	**1.0**	**1.03**

TABLE A.4 Results of Equi-dispersion and High Mean Scenario (COM-Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters*		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	5.025	5.241	1.043	DP	5.025	0.927	12.9	13.2	-4445.5	11	1.17
				COM-P	5.003	0.966	10.0	9.9	-4444.1	10	1.00
				Poisson	5.025	--	11.4	11.3	-4444.6	11	1.03
Run2	4.897	5.062	1.034	DP	4.897	0.915	8.9	9.3	-4425.4	11	0.81
				COM-P	4.874	0.954	6.4	6.4	-4424.0	10	0.64
				Poisson	4.897	--	8.2	8.1	-4424.8	11	0.74
Run3	4.995	4.951	0.991	DP	4.995	0.973	18.7	20.0	-4393.5	10	1.87
				COM-P	4.997	1.016	17.1	17.4	-4392.5	10	1.71
				Poisson	4.995	--	17.2	17.9	-4392.6	11	1.56
Run4	4.971	4.972	1.000	DP	4.971	0.958	12.9	13.1	-4400.9	10	1.29
				COM-P	4.964	1.001	10.9	10.6	-4399.7	10	1.09
				Poisson	4.971	--	10.8	10.6	-4399.7	11	0.98
Run5	4.986	5.089	1.021	DP	4.986	0.951	6.1	6.3	-4413.1	10	0.61
				COM-P	4.980	0.995	3.8	3.7	-4411.9	10	0.38
				Poisson	4.986	--	3.9	3.7	-4411.9	11	0.35
Average	--	--	--	DP	--	--	11.9	12.4	-4415.7	10.4	1.14
				COM-P	--	--	**9.6**	**9.6**	**-4414.4**	**10.0**	**0.96**
				Poisson	--	--	**10.3**	**10.3**	**-4414.7**	**11.0**	**0.93**

*For Poisson, the estimated parameter is the mean λ. All other variables are, as described in the previous tables. (Note: same for all other tables below.)

TABLE A.5 Results of Equi-dispersion and Medium Mean Scenario (COM-Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	2.022	2.127	1.052	DP	2.022	0.830	18.0	19.1	-3471.3	6	2.99
				COM-P	1.969	0.917	3.5	3.6	-3463.4	5	0.69
				Poisson	2.022	--	7.4	7.2	-3465.1	6	1.24
Run2	1.998	2.129	1.066	DP	1.998	0.838	24.4	25.3	-3455.2	5	4.87
				COM-P	1.944	0.920	8.5	8.8	-3446.6	5	1.70
				Poisson	1.998	--	12.4	12.1	-3448.1	6	2.06
Run3	2.000	2.049	1.025	DP	2.000	0.867	25.5	26.5	-3430.3	5	5.10
				COM-P	1.984	0.975	9.4	9.3	-3421.6	5	1.87
				Poisson	2.000	--	9.8	9.6	-3421.7	6	1.64
Run4	2.020	2.024	1.002	DP	2.020	0.892	24.6	26.1	-3423.4	5	4.92
				COM-P	2.023	1.012	8.8	8.4	-3414.7	5	1.75
				Poisson	2.020	--	8.7	8.5	-3414.7	6	1.46
Run5	2.083	1.9841	0.953	DP	2.083	0.935	19.1	20.1	-3420.3	5	3.83
				COM-P	2.124	1.080	2.6	2.6	-3411.8	5	0.53
				Poisson	2.083	--	6.2	6.4	-3413.6	6	1.04
Average	--	--	--	DP	--	--	22.3	23.4	-3440.1	5.2	4.29
				COM-P	--	--	**6.5**	**6.5**	**-3431.6**	**5.0**	**1.31**
				Poisson	--	--	8.9	8.8	-3432.6	6.0	1.49

TABLEA.6 Results of Equi-dispersion and Low Mean Scenario (COM-Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	0.504	0.474	0.942	DP	0.504	1.027	4.0	4.0	-1842.8	1	4.03
				COM-P	0.605	1.232	1.4	1.4	-1841.5	1	1.41
				Poisson	0.504	--	5.6	5.7	-1843.7	2	2.82
Run2	0.518	0.524	1.012	DP	0.518	0.974	0.3	0.3	-1894.6	1	0.31
				COM-P	0.482	0.937	0.3	0.3	-1894.6	1	0.32
				Poisson	0.518	--	0.6	0.6	-1894.7	2	0.31
Run3	0.502	0.498	0.994	DP	0.502	0.993	1.1	1.2	-1857.2	1	1.13
				COM-P	0.484	0.974	1.3	1.3	-1857.3	1	1.31
				Poisson	0.502	--	1.1	1.1	-1857.2	2	0.55
Run4	0.491	0.482	0.983	DP	0.491	1.005	1.0	1.0	-1831.9	1	0.96
				COM-P	0.516	1.056	0.8	0.8	-1831.8	1	0.83
				Poisson	0.491	--	1.0	1.0	-1831.9	2	0.50
Run5	0.515	0.529	1.028	DP	0.515	0.973	0.2	0.2	-1891.1	1	0.21
				COM-P	0.454	0.893	0.5	0.5	-1891.1	1	0.50
				Poisson	0.515	--	0.2	0.2	-1891.3	2	0.11
Average	--	--	--	DP	--	--	1.3	1.3	-1863.5	1.0	1.33
				COM-P	--	--	**0.9**	**0.9**	**-1863.3**	**1.0**	**0.87**
				Poisson	--	--	**1.7**	**1.7**	**-1863.8**	**2.0**	**0.86**

TABLEA.7 Results of Equi-dispersion and High Mean Scenario (Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	4.968	5.309	1.069	DP	4.968	0.894	11.0	11.4	-4462.3	11	1.00
				COM-P	4.926	0.930	9.0	8.8	-4460.9	11	0.82
				Poisson	4.968		14.5	13.1	-4463.0	11	1.32
Run2	4.962	5.008	1.009	DP	4.962	0.948	21.8	23.1	-4410.7	10	2.18
				COM-P	4.954	0.991	20.0	20.5	-4409.4	10	2.00
				Poisson	4.962		20.1	20.5	-4409.5	11	1.83
Run3	5.077	4.918	0.969	DP	5.077	0.984	8.1	8.7	-4400.5	10	0.81
				COM-P	5.090	1.032	6.7	6.8	-4399.5	10	0.67
				Poisson	5.077		7.1	7.5	-4399.9	11	0.65
Run4	4.936	4.976	1.008	DP	4.936	0.954	6.2	6.8	-4400.2	10	0.62
				COM-P	4.936	0.994	4.4	4.5	-4399.0	10	0.44
				Poisson	4.936		4.4	4.6	-4399.0	11	0.40
Run5	5.096	4.987	0.979	DP	5.096	0.979	6.6	6.8	-4409.2	10	0.66
				COM-P	5.108	1.023	5.4	5.3	-4408.2	10	0.54
				Poisson	5.096		5.5	5.6	-4408.4	11	0.50
Average				DP			10.8	11.4	-4416.6	10.2	1.05
				COM-P			**9.1**	**9.2**	**-4415.4**	**10.2**	**0.89**
				Poisson			10.3	10.3	-4416.0	11	0.94

TABLE A.8 Results of Equi-dispersion and Medium Mean Scenario (Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	4.968	5.309	1.069	DP	4.968	0.894	11.0	11.4	-4462.3	11	1.00
				COM-P	4.926	0.930	9.0	8.8	-4460.9	11	0.82
				Poisson	4.968		14.5	13.1	-4463.0	11	1.32
Run2	4.962	5.008	1.009	DP	4.962	0.948	21.8	23.1	-4410.7	10	2.18
				COM-P	4.954	0.991	20.0	20.5	-4409.4	10	2.00
				Poisson	4.962		20.1	20.5	-4409.5	11	1.83
Run3	5.077	4.918	0.969	DP	5.077	0.984	8.1	8.7	-4400.5	10	0.81
				COM-P	5.090	1.032	6.7	6.8	-4399.5	10	0.67
				Poisson	5.077		7.1	7.5	-4399.9	11	0.65
Run4	4.936	4.976	1.008	DP	4.936	0.954	6.2	6.8	-4400.2	10	0.62
				COM-P	4.936	0.994	4.4	4.5	-4399.0	10	0.44
				Poisson	4.936		4.4	4.6	-4399.0	11	0.40
Run5	5.096	4.987	0.979	DP	5.096	0.979	6.6	6.8	-4409.2	10	0.66
				COM-P	5.108	1.023	5.4	5.3	-4408.2	10	0.54
				Poisson	5.096		5.5	5.6	-4408.4	11	0.50
Average				DP			10.8	11.4	-4416.6	10.2	1.05
				COM-P			**9.1**	**9.2**	**-4415.4**	**10.2**	**0.89**
				Poisson			10.3	10.3	-4416.0	11	0.94

TABLEA.9 Results of Equi-dispersion and Low Mean Scenario (Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	0.489	0.499	1.021	DP	0.489	0.989	0.8	0.8	-1838.9	1	0.80
				COM-P	0.413	0.867	0.5	0.5	-1838.9	1	0.55
				Poisson	0.489		1.1	1.0	-1839.0	2	0.54
Run2	0.494	0.478	0.968	DP	0.494	1.013	0.4	0.4	-1834.0	1	0.45
				COM-P	0.542	1.104	0.0	0.0	-1833.7	1	0.01
				Poisson	0.494		0.7	0.7	-1834.2	2	0.36
Run3	0.503	0.499	0.993	DP	0.503	0.991	0.1	0.1	-1860.6	1	0.05
				COM-P	0.495	0.992	0.2	0.2	-1860.6	1	0.18
				Poisson	0.503		0.1	0.1	-1860.6	2	0.06
Run4	0.491	0.487	0.993	DP	0.491	0.998	0.8	0.8	-1835.5	1	0.83
				COM-P	0.476	0.977	0.9	0.9	-1835.5	1	0.90
				Poisson	0.491		0.8	0.8	-1835.5	2	0.42
Run5	0.486	0.477	0.982	DP	0.486	1.013	1.0	1.0	-1820.7	1	0.99
				COM-P	0.507	1.052	1.0	1.0	-1820.6	1	1.02
				Poisson	0.486		1.1	1.2	-1820.8	2	0.57
Average				DP			0.6	0.6	-1837.9	1	0.62
				COM-P			**0.5**	**0.5**	**-1837.9**	**1**	**0.53**
				Poisson			**0.8**	**0.8**	**-1838.0**	**2**	**0.39**

TABLEA.10 Results of Over-dispersion and High Mean Scenario (COM-Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters*		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	5.523	10.018	1.814	DP	5.523	0.530	26.1	27.7	-5040.8	14	1.86
				COM-P	4.967	0.496	11.8	11.7	-5032.8	14	0.84
				NB	6.785	0.449	13.5	13.2	-5033.9	15	0.90
Run2	5.706	10.194	1.786	DP	5.706	0.535	25.9	27.3	-5067.3	15	1.73
				COM-P	5.176	0.506	12.5	13.0	-5061.7	14	0.89
				NB	7.255	0.440	26.5	26.0	-5066.2	15	1.77
Run3	5.556	10.242	1.844	DP	5.556	0.517	24.4	26.4	-5065.9	14	1.74
				COM-P	4.967	0.481	11.5	11.8	-5058.6	14	0.82
				NB	6.586	0.458	21.9	21.7	-5063.3	15	1.46
Run4	5.523	10.018	1.814	DP	5.523	0.530	26.1	27.7	-5040.8	14	1.86
				COM-P	5.079	0.557	22.2	21.9	-4962.1	14	1.59
				NB	8.701	0.388	38.4	37.9	-4970.4	14	2.75
Run5	5.426	9.348	1.723	DP	5.426	0.542	23.2	25.5	-4994.9	14	1.66
				COM-P	4.919	0.515	15.3	16.3	-4990.3	14	1.09
				NB	7.504	0.420	36.9	37.1	-5001.0	14	2.63
Average	--	--	--	DP	--	--	25.1	26.9	-5041.9	14.2	1.77
				COM-P	--	--	**14.7**	**14.9**	**-5021.1**	**14.0**	**1.05**
				NB	--	--	27.4	27.2	-5027.0	14.6	1.88

*For NB, the estimated parameters are the inverse dispersion parameter ϕ and the probability of success p. (Note: same for all other tables)

TABLE A.11 Results of Over-dispersion and Medium Mean Scenario (COM-Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	2.597	4.070	1.567	DP	2.597	0.582	15.5	16.4	-4027.1	8	1.93
				COM-P	2.028	0.498	9.3	9.9	-4023.7	8	1.17
				NB	4.580	0.362	21.6	22.8	-4029.5	9	2.40
Run2	2.515	4.152	1.651	DP	2.515	0.567	20.8	21.0	-4010.1	8	2.60
				COM-P	1.852	0.459	5.1	5.1	-4002.3	8	0.64
				NB	3.862	0.394	6.9	7.7	-4002.9	9	0.76
Run3	2.539	4.096	1.613	DP	2.539	0.573	23.2	23.6	-4012.4	8	2.90
				COM-P	1.914	0.475	11.5	11.5	-4006.3	8	1.43
				NB	4.142	0.380	18.2	18.2	-4009.2	9	2.03
Run4	2.541	3.906	1.538	DP	2.541	0.599	19.9	20.8	-3985.3	8	2.49
				COM-P	2.021	0.522	8.7	9.0	-3979.4	8	1.09
				NB	4.725	0.350	12.7	13.2	-3981.7	8	1.59
Run5	2.583	3.976	1.539	DP	2.583	0.596	14.2	14.8	-4008.2	8	1.78
				COM-P	2.058	0.518	4.4	4.5	-4003.1	8	0.54
				NB	4.789	0.350	10.2	10.1	-4006.6	8	1.28
Average	--	--	--	DP	--	--	18.7	19.3	-4008.6	8.0	2.34
				COM-P	--	--	**7.8**	**8.0**	**-4003.0**	**8.0**	**0.98**
				NB	--	--	13.9	14.4	-4006.0	8.6	1.62

TABLE A.12 Results of Over-dispersion and Low Mean Scenario (COM-Poisson Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	0.926	1.176	1.270	DP	0.926	0.756	7.2	7.2	-2609.7	3	2.39
				COM-P	0.526	0.540	3.6	3.7	-2607.3	3	1.20
				NB	3.426	0.213	3.7	3.6	-2606.4	4	0.92
Run2	0.934	1.306	1.399	DP	0.934	0.703	11.6	11.7	-2651.5	3	3.87
				COM-P	0.316	0.383	1.1	1.1	-2645.7	4	0.27
				NB	2.343	0.285	1.3	1.2	-2645.9	4	0.32
Run3	0.964	1.226	1.272	DP	0.964	0.744	4.5	4.7	-2661.8	3	1.50
				COM-P	0.549	0.533	4.1	3.9	-2660.1	4	1.02
				NB	3.543	0.214	3.4	3.3	-2660.1	4	0.84
Run4	0.936	1.190	1.272	DP	0.936	0.744	4.9	5.0	-2627.3	3	1.63
				COM-P	0.515	0.524	1.2	1.2	-2625.6	3	0.41
				NB	3.442	0.214	1.9	1.9	-2626.2	4	0.47
Run5	0.941	1.240	1.318	DP	0.941	0.735	3.7	3.8	-2641.0	3	1.25
				COM-P	0.450	0.474	5.5	5.6	-2638.0	4	1.37
				NB	2.960	0.241	4.1	4.6	-2637.1	4	1.04
Average	--	--	--	DP	--	--	6.4	6.5	-2638.3	3.0	2.13
				COM-P	--	--	3.1	3.1	-2635.4	3.6	0.86
				NB	--	--	**2.9**	**2.9**	**-2635.1**	**4.0**	**0.72**

TABLEA.13 Results of Over-dispersion and High Mean Scenario (NB Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	4.973	7.921	1.593	DP	4.973	0.601	21.9	23.1	-4813.9	13	1.68
				COM-P	4.581	0.580	9.7	9.4	-4807.0	13	0.75
				NB	8.387	0.372	8.4	8.3	-4807.0	13	0.65
Run2	5.089	7.414	1.457	DP	5.089	0.658	22.0	22.1	-4759.8	12	1.84
				COM-P	4.802	0.653	15.6	14.2	-4755.8	12	1.30
				NB	11.140	0.314	12.9	12.3	-4755.4	13	0.99
Run3	5.013	7.887	1.573	DP	5.013	0.624	37.9	41.4	-4799.8	13	2.92
				COM-P	4.656	0.603	22.5	22.3	-4790.3	12	1.88
				NB	8.745	0.364	7.9	7.9	-4783.5	13	0.60
Run4	5.009	8.007	1.599	DP	5.009	0.609	36.7	39.4	-4817.3	13	2.83
				COM-P	4.631	0.588	23.0	21.3	-4808.2	13	1.77
				NB	8.369	0.374	12.2	12.0	-4804.2	13	0.94
Run5	5.026	8.322	1.656	DP	5.026	0.592	39.1	41.7	-4845.8	13	3.00
				COM-P	4.604	0.567	22.5	21.3	-4835.6	13	1.73
				NB	7.661	0.396	7.9	7.9	-4829.0	13	0.61
Average				DP			31.5	33.6	-4807.3	12.8	2.46
				COM-P			18.7	17.7	-4799.4	12.6	1.48
				NB			**9.9**	**9.7**	**-4795.8**	**13**	**0.76**

TABLEA.14 Results of Over-dispersion and Medium Mean Scenario (NB Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	1.988	2.570	1.293	DP	1.988	0.699	17.8	18.1	-3573.5	6	2.97
				COM-P	1.702	0.669	7.2	7.2	-3568.1	6	1.21
				NB	6.786	0.227	8.1	8.0	-3568.4	6	1.34
Run2	2.013	2.740	1.361	DP	2.013	0.682	34.3	34.1	-3615.1	6	5.72
				COM-P	1.667	0.624	16.7	16.5	-3606.3	6	2.78
				NB	5.572	0.265	10.4	10.3	-3603.5	6	1.73
Run3	2.015	2.739	1.359	DP	2.015	0.687	32.8	32.9	-3611.3	6	5.47
				COM-P	1.678	0.631	15.7	15.4	-3602.7	6	2.62
				NB	5.604	0.264	9.7	9.7	-3599.0	6	1.62
Run4	2.019	2.810	1.392	DP	2.019	0.650	10.8	11.0	-3637.8	6	1.80
				COM-P	1.607	0.578	3.0	3.0	-3633.8	6	0.50
				NB	5.150	0.282	7.6	7.6	-3636.1	7	1.09
Run5	2.057	2.791	1.357	DP	2.057	0.673	21.2	22.4	-3642.6	6	3.54
				COM-P	2.058	0.518	4.4	4.5	-4003.1	8	0.54
				NB	5.763	0.263	10.9	11.3	-3636.6	7	1.55
Average				DP			23.4	23.7	-3616.1	6.0	3.90
				COM-P			9.4	9.3	-3682.8	6.4	1.47
				NB			**9.3**	**9.4**	**-3608.7**	**6.4**	**1.46**

TABLEA.15 Results of Over-dispersion and Low Mean Scenario (NB Simulated Data)

Run #	Characteristics of Simulated Data			Distributions	Estimated Parameters		Goodness-of-Fit			DF	Chi-Sq/DF
	Mean	Var	Var/Mean				Chi-Sq	LR	LogL		
Run1	0.505	0.571	1.131	DP	0.505	0.926	10.3	9.9	-1899.5	2	5.13
				COM-P	0.227	0.564	1.4	1.4	-1895.7	2	0.70
				NB	3.849	0.116	1.7	1.6	-1895.9	2	0.84
Run2	0.485	0.581	1.198	DP	0.485	0.918	11.9	12.0	-1869.9	1	11.95
				COM-P	0.123	0.434	1.0	1.1	-1861.6	2	0.51
				NB	2.449	0.165	0.9	0.9	-1861.2	2	0.45
Run3	0.520	0.618	1.188	DP	0.520	0.899	17.9	17.0	-1940.8	2	8.94
				COM-P	0.128	0.418	5.7	6.0	-1934.6	2	2.84
				NB	2.762	0.158	5.1	5.0	-1934.0	2	2.55
Run4	0.515	0.609	1.184	DP	0.515	0.906	13.1	11.8	-1927.8	2	6.57
				COM-P	0.188	0.505	0.8	0.8	-1921.5	2	0.38
				NB	2.798	0.155	0.4	0.4	-1921.2	2	0.22
Run5	0.497	0.596	1.201	DP	0.497	0.916	16.3	15.6	-1893.3	1	16.29
				COM-P	0.111	0.410	7.8	7.9	-1885.6	2	3.90
				NB	2.470	0.167	5.4	5.4	-1884.4	2	2.71
Average				DP			13.9	13.2	-1906.3	1.6	8.69
				COM-P			3.3	3.4	-1899.8	2.0	1.67
				NB			**2.7**	**2.7**	**-1899.4**	**2.0**	**1.35**

VITA

Name: Yaotian Zou

Address: 1600 Southwest Pkwy,
Cedar Ridge Apartments #404,
College Station, TX 77840

Email Address: zouyaotian@hotmail.com

Education: B.E., Transportation Engineering, Southeast University, 2010

M.S., Civil Engineering, Texas A&M University, 2012