Accident Analysis and Prevention 42 (2010) 1273–1282

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson–gamma models

Srinivas Reddy Geedipally[a,*], Dominique Lord[b,1]

[a] Texas Transportation Institute, Texas A&M University, 3135 TAMU, College Station, TX 77843-3135, United States
[b] Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, United States

ARTICLE INFO

Article history:
Received 28 April 2009
Received in revised form 4 February 2010
Accepted 8 February 2010

Keywords:
Single-vehicle crashes
Multi-vehicle crashes
Confidence intervals

ABSTRACT

Crash prediction models still constitute one of the primary tools for estimating traffic safety. These statistical models play a vital role in various types of safety studies. With a few exceptions, they have often been employed to estimate the number of crashes per unit of time for an entire highway segment or intersection, without distinguishing the influence different sub-groups have on crash risk. The two most important sub-groups that have been identified in the literature are single- and multi-vehicle crashes. Recently, some researchers have noted that developing two distinct models for these two categories of crashes provides better predicting performance than developing models combining both crash categories together. Thus, there is a need to determine whether a significant difference exists for the computation of confidence intervals when a single model is applied rather than two distinct models for single- and multi-vehicle crashes. Building confidence intervals have many important applications in highway safety.

This paper investigates the effect of modeling single- and multi-vehicle (head-on and rear-end only) crashes separately versus modeling them together on the prediction of confidence intervals of Poisson–gamma models. Confidence intervals were calculated for total (all severities) crash models and fatal and severe injury crash models. The data used for the comparison analysis were collected on Texas multilane undivided highways for the years 1997–2001. This study shows that modeling single- and multi-vehicle crashes separately predicts larger confidence intervals than modeling them together as a single model. This difference is much larger for fatal and injury crash models than for models for all severity levels. Furthermore, it is found that the single- and multi-vehicle crashes are not independent. Thus, a joint (bivariate) model which accounts for correlation between single- and multi-vehicle crashes is developed and it predicts wider confidence intervals than a univariate model for all severities. Finally, the simulation results show that separate models predict values that are closer to the true confidence intervals, and thus this research supports previous studies that recommended modeling single- and multi-vehicle crashes separately for analyzing highway segments.

Published by Elsevier Ltd.

1. Introduction

Crash prediction models still constitute one of the primary tools for estimating traffic safety. These statistical models play a vital role in various types of safety studies, but they are most often used for estimating the safety performance of various transportation elements or entities (Persaud and Nguyen, 1998; Lord, 2000; Ivan et al., 2000; Lyon et al., 2003; Tarko et al., 2008; Lord and Mannering, in press). In this context, these models have been employed to estimate the number of crashes per unit of time for an entire highway segment or intersection, without distinguishing the influence different sub-groups have on crash risk. A few exceptions have been noted in the literature however. For instance, some researchers have developed distinct predictive models to estimate the safety performance as a function of different categories of vehicles, such as passenger vehicles and truck crashes (Jovanis and Chang, 1986; Miller et al., 1998; Lee and Abdel-Aty, 2005), different time periods (Cercarelli et al., 1992; Mensah and Hauer, 1998; Zador, 1985) or the number of vehicles involved in each crash, i.e., single-vehicle (SV) and multi-vehicle (MV) crashes (Qin et al., 2004; Lord et al., 2005; Griffith, 1999) among others.

Given the differences observed in the characteristics associated with SV and MV crashes, some transportation safety analysts have proposed that distinct crash prediction models should be developed for these two categories of crashes when the objective of the study consists of estimating the safety performance of highway segments (Mensah and Hauer, 1998; Ivan, 2004; Lord et al., 2005; Jonsson et al., 2007; Harwood et al., 2007; Bonneson et al.,

* Corresponding author. Tel.: +1 979 862 1651; fax: +1 979 845 6006.
E-mail addresses: srinivas8@tamu.edu, gsrinivas8@gmail.com (S.R. Geedipally), d-lord@tamu.edu (D. Lord).
[1] Tel.: +1 979 458 3949; fax: +1 979 845 6481.

0001-4575/$ – see front matter. Published by Elsevier Ltd.
doi:10.1016/j.aap.2010.02.004

2007). These researchers noted that developing two distinct models provides better predicting performance than developing models combining both crash categories together. In most cases, the motivation for separating models by the number of vehicles involved in the crash is based on shape of the functional form linking both crash types to the traffic flow variable that has been found to be very different from one another (Ivan, 2004; Lord et al., 2005).

Since these two categories of models have been used to estimate or predict the number of crashes on highway segments (Harwood et al., 2007; Bonneson et al., 2007; Lord et al., 2008), there is a need to determine whether there exist differences in the computation of confidence intervals when a single model is applied rather than two distinct models. Confidence intervals can play a major role in the selection of various highway design alternatives (Lord, 2008) and the identification of hazardous sites (Hauer, 1997) among others. The primary objective of this paper is to investigate if there is an important difference in the prediction of confidence intervals when a unique model is estimated compared to a distinct model for SV and MV crashes. The secondary objective is to examine if there is any difference in the prediction of confidence intervals when a bivariate negative binomial (BNB) model is used rather than a univariate negative binomial (UNB) model for analyzing SV and MV crashes. Confidence intervals were calculated for the Poisson mean, gamma mean and predicted response, respectively (Wood, 2005). Crash data collected on Texas undivided roads from 1997 to 2001 were used for this comparison analysis.

This paper is divided into five sections. The first section provides a discussion on the modeling prospects related to SV and MV crashes. The second section gives a brief description of data used in this study. The third section describes the methodology utilized for the comparison analysis. The fourth section presents the results of this analysis. The last section provides a summary of the work carried out in this research and recommendation for further work.

2. Background

Several researchers have examined the characteristics and the differences associated with SV and MV crashes (direct comparison). For instance, Ostrom and Eriksson (1993) were the first to examine crash characteristics as a function of the number of vehicles involved in a crash. They studied factors influencing crashes involving intoxicated drivers in northern Sweden. These authors reported that the driver's blood alcohol content was more significantly related to SV crash fatalities than those associated with MV crash fatalities.

Shankar et al. (1995) analyzed the safety effects of highway design features, weather and other seasonal variables on different crash types. Using data collected on a 61-km section near Seattle, WA, they analyzed several variables, such as the number of horizontal curves and their spacing, rainfall amount, snowing conditions, and their relationship to different types of SV and MV crashes. The authors concluded that models predicting crashes for different crash types had a greater explanatory power than a single model that combined all crash types together.

Mensah and Hauer (1998) examined the effects of aggregated and disaggregated traffic flow variables on the estimation of predictive models. They used data on two-lane rural highways published in Persaud and Mucsi (1995) for their analysis. Using exploratory data analyses and regression methods, they found that SV and MV crashes have significant different characteristics. They reported that using an aggregated model that combine both SV and MV crashes predicts fewer crashes than combining the output of two separate models for the same two categories of crashes.

Griffith (1999) studied SV and MV crashes caused by alcohol and drug-impaired drivers. This author found that SV run-off-the-road crashes on freeways resulted in a higher number of nonfatal injuries than for MV crashes involving an impaired driver. In contrast, MV crashes on freeways resulted in a higher number of fatal injuries than for SV run-off-the-road crashes.

Ivan et al. (1999) investigated differences in causality factors for SV and MV crashes on two-lane rural highways in Connecticut. They found that contributing factors were different for each category of crashes. For example, SV crashes were negatively associated with an increase in traffic intensity (exposure), shoulder width, sight distance, and level of service (LOS). On the other hand, MV crashes were positively associated with an increase in traffic intensity, shoulder width, truck percentage, and number of traffic signals.

In a subsequent study, Ivan et al. (2000) reported that the time-of-day differently influenced for both categories of crashes. SV crashes occurred mostly during the evening and at night, as expected, whereas MV crashes occur more frequently during daylight and evening peak periods. This was mainly attributed to the higher traffic intensity (or exposure). Driveway density had a mixed effect on SV crashes. Driveways at gas stations and minor road intersections were negatively associated with SV crashes, whereas driveways located adjacent to apartment complexes seemed to be associated with an increase in SV crashes. MV crashes increased for all types of driveways.

Ivan (2004) modeled crashes, also using data collected in Connecticut, according to the manner of collision (i.e., number of vehicles involved and their direction of travel). The study showed that the expected number of SV crashes decreases with the increase in traffic volume, whereas all types of MV crashes increased with the increase in traffic flow during the evening time period. During daytime, SV and MV opposite-direction crashes decreased with an increase in traffic flow, whereas MV same-direction and intersecting-direction crashes increased with an augmentation in traffic flow.

Qin et al. (2004) developed zero-inflated Poisson models for different crash types (see Lord et al., 2005, 2007 for a discussion about the application of such models in highway safety). These authors examined crashes occurring on highway segments in Michigan and concluded that crashes are differently associated with traffic flows for different crash types. They noted, for example, that aggregated crash prediction model ignores significant variation in highway crashes. For SV crashes, the marginal crash rate was found to be high for low traffic flow levels and small for high traffic flow volumes. For all MV crashes (except multi-vehicle intersecting-direction crashes), the marginal crash rate was small at low traffic volumes, but high for large traffic volumes, probably because this type of crash is more likely to occur under short headways.

Lord et al. (2005) evaluated several functional forms for SV and MV crashes, and one that combined both (referred to as *All Model*), as a function of traffic flow, vehicle density and volume over capacity (V/C) ratio on urban freeway segments in Montreal, Canada. Using regression methods, they recommended developing different predictive models for SV and MV crashes rather than developing a common model for both crash categories. The output of the SV model showed that crashes initially increase, peak and then decrease as density or V/C increases. On the other hand, the MV crash model and the model grouping both SV and MV crashes (*All Model*) showed that crashes increased with an increasing vehicle density or V/C ratio, as expected.

Recently, Jonsson et al. (2007) developed distinct models for SV and different types of MV crashes occurring at intersections on rural four-lane highways in California. These authors concluded that the SV crash model loses some of the observed covariate effects through the aggregation of collision types. Different crash type models exhibited dissimilar relationships with traffic flow and other covariates. Furthermore, Jonsson et al. (2007) noted that the distribution by crash severity varied for different crash types.

In summary, numerous studies have shown that MV and SV crashes have vastly different associative relationships with exposure and geometric design features. These differences were noted for various types of highway facilities. Given these observations, some transportation safety analysts have indicated that distinct models should be estimated for SV and MV crashes when the objective consists of estimating the safety performance of highway segments.

3. Methodology

This section describes the probabilistic structure of the Poisson–gamma (aka negative binomial NB) and bivariate models, the functional form used for linking SV and MV crashes to traffic flow, the procedure employed for estimating the confidence intervals, and characteristics of the Monte Carlo simulation study.

3.1. Probabilistic structure of Poisson–gamma models

Poisson and Poisson–gamma models belong to the family of generalized linear models (GLMs). Poisson–gamma models in highway safety applications have been shown to have the following probabilistic structure: the number of crashes at the ith entity (road section, intersections, etc.) and tth time period, Y_{it}, when conditional on its mean μ_{it}, is assumed to be Poisson distributed and independent over all entities and time periods as:

$$Y_{it}|\mu_{it} \sim Po(\mu_{it}) \quad i = 1, 2, \ldots, I \text{ and } t = 1, 2, \ldots, T \tag{1}$$

The mean of the Poisson is structured as:

$$\mu_{it} = f(X; \beta)\exp(e_{it}) \tag{2}$$

where $f(\cdot)$ is a function of the covariates (X); β is a vector of unknown coefficients; e_{it} is a model error independent of all the covariates.

With this characteristic, it can be shown that Y_{it}, conditional on μ_{it} and α, is distributed as a Poisson–gamma random variable with a mean μ_{it} and a variance $\mu_{it} + \alpha{\mu_{it}}^2$, respectively. (*Note*: Other variance functions exist for the Poisson–gamma model, but they are not covered here since they are seldom used in highway safety studies. The reader is referred to Cameron and Trivedi (1998) and Maher and Summersgill (1996) for a description of alternative variance functions.) The probability density function (PDF) of the Poisson–gamma structure described above is given by the following equation:

$$f(y_{it}; \alpha, \mu_{it}) = \frac{\Gamma(y_{it} + \alpha^{-1})}{\Gamma(\alpha^{-1}) y_{it}!} \left(\frac{\alpha^{-1}}{\mu_{it} + \alpha^{-1}}\right)^{\alpha^{-1}} \left(\frac{\mu_{it}}{\mu_{it} + \alpha^{-1}}\right)^{y_{it}} \tag{3}$$

where y_{it} = response variable for observation i and time period t; μ_{it} = mean response for observation i and time period t; and, α = dispersion parameter of the Poisson–gamma distribution.

The variance of the Poisson–gamma random variable is given by:

$$Var(y_{it}) = \mu_{it} + \alpha\mu_{it}^2 \tag{4}$$

Note that if $\alpha \to 0$, the crash variance equals the crash mean and this model converges to the standard Poisson regression model.

The choice of functional form for linking the number of crashes with the model's covariates is an important characteristic associated with the development of statistical relationships. In this study, crash counts were assumed to follow a nonlinear relationship with traffic volume. Although this functional form is very popular among transportation safety analysts, it may not be the most appropriate form to properly capture exposure at the boundary conditions (Lord, 2002; Lord et al., 2005). Furthermore, the segment length was assumed to be directly proportional to the crash frequency, meaning that the segment length has linear relation with the crash occurrence. Thus, the segment length is considered to be an offset rather than as a covariate.

The statistical model considered in this study is similar to that used elsewhere in the safety literature (see Lord and Bonneson, 2007), except that only one independent variable, namely Average Daily Traffic (ADT), was used. These models are often referred to as flow-only models. They are the most popular type of models utilized by transportation safety analysts (Hauer, 1997; Persaud et al., 2001). Furthermore, they are frequently preferred over models that include several covariates because they can be easily re-calibrated when they are developed in one jurisdiction and applied to another (Persaud et al., 2002; Lord and Bonneson, 2005). Although such models will suffer from an omitted variables bias (because many non-flow related factors are known to affect the frequency of crashes), the empirical assessment carried out in this work still provides valuable information for the development of SV and MV predictive models.

The mean of the crashes per year for segment i can be calculated by:

$$\mu_i = \beta_0 L_i {F_i}^{\beta_1} \tag{5}$$

where L_i = length (in miles) of segment i; F_i = Average Daily Traffic (ADT); β_0 = intercept (to be estimated); β_1 = coefficient (to be estimated) associated with ADT.

It is usually assumed that the number of crashes increases at a decreasing rate as the traffic volume increases. This relationship is characterized in predictive models with the coefficient for the traffic volume parameter (β_1) to be below 1.

3.2. Probabilistic structure of bivariate negative binomial models

There has not been a lot of research conducted on jointly modeling crash counts in highway safety. As stated by Park and Lord (2007), a multivariate model treats the correlated crash counts as interdependent variables and leads to more precise estimates for the effects of factors on crash risk, than a univariate model. The few studies that documented the application of multivariate models for analyzing crash data include Maher (1990), Tunaru (2002), Bijleveld (2005), Miaou and Song (2005), Song et al. (2006) and Ma and Kockelman (2006).

The probability function of a bivariate negative binomial (BNB) model is given by the following equation (Subrahmaniam and Subrahmaniam, 1973):

$$f(y_{1it}, y_{2it}) = \frac{(\mu_{1it} - \delta)^{y_{1it}} (\mu_{2it} - \delta)^{y_{2it}} \Gamma(r + y_{1it} + y_{2it})}{y_{1it}! y_{2it}! \Gamma(r)(1 + \mu_{1it} + \mu_{2it} - \delta)^{r + y_{1it} + y_{2it}}} S(y_{1it}, y_{2it}) \tag{6}$$

where y_{1it}, y_{2it} = two response variables for observation i and time period t; $\mu_{1it} \times r$ = mean of the first response variable for observation i and time period t; $\mu_{2it} \times r$ = mean of the second response variable for observation i and time period t; r = combined inverse dispersion parameter:

$$S(y_{1it}, y_{2it}) = \frac{\displaystyle\sum_{j=0}^{\min(y_{1it}, y_{2it})} \tau^j \binom{y_{1it}}{j} \binom{y_{2it}}{j}}{\dbinom{r + y_{1it} + y_{2it} - 1}{j}} \tag{7}$$

with

$$\tau = \frac{\delta(1 + \mu_{1it} + \mu_{2it} - \delta)}{(\mu_{1it} - \delta)(\mu_{2it} - \delta)}; \tag{8}$$

$$\delta = \left(\frac{m_{11}}{r} - \frac{\overline{y_1} \times \overline{y_2}}{r^2}\right) \tag{9}$$

with m_{11} = first mixed central moment.

Table 1
Confidence intervals for μ, m, y (Wood, 2005).

Parameter	Intervals
μ	$\left[\frac{\hat{\mu}}{e^{1.96\sqrt{Var(\hat{\eta})}}}, \hat{\mu}e^{1.96\sqrt{Var(\hat{\eta})}}\right]$
m	$\left[\max\left\{0, \hat{\mu}-1.96\sqrt{\hat{\mu}^2 Var(\hat{\eta})+\frac{\hat{\mu}^2 Var(\hat{\eta})+\hat{\mu}^2}{\phi}}\right\}, \hat{\mu}+1.96\sqrt{\hat{\mu}^2 Var(\hat{\eta})+\frac{\hat{\mu}^2 Var(\hat{\eta})+\hat{\mu}^2}{\phi}}\right]$
y	$\left[0, \left\lfloor \hat{\mu}+\sqrt{19}\sqrt{\hat{\mu}^2 Var(\hat{\eta})+\frac{\hat{\mu}^2 Var(\hat{\eta})+\hat{\mu}^2}{\phi}}+\hat{\mu}\right\rfloor\right]$

Note: $Var(\eta) = XI^{-1}X^T$, where I^{-1} is the variance–covariance matrices and X is a matrix containing observed values in logarithmic form. $\lfloor x \rfloor$ denotes the largest integer less or equal to x.

3.3. Goodness-of-fit statistics

Different methods were used for evaluating the goodness-of-fit (GOF) and predictive performance of the models. The methods used in this research include the following:

Mean Absolute Deviance (*MAD*):
MAD provides a measure of the average mis-prediction of the model (Oh et al., 2003). It is computed using the following equation:

$$\text{Mean Absolute Deviance(MAD)} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \quad (10)$$

Mean Squared Predictive Error (*MSPE*):
MSPE is typically used to assess the error associated with a validation or external data set (Oh et al., 2003). It can be computed using Eq. (11):

$$\text{Mean Squared Predictive Error(MSPE)} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \quad (11)$$

3.4. Estimation of confidence intervals

Confidence intervals can be used for selecting highway design alternatives where the safety performance is used as a screening criterion and for identifying hazardous sites. Wood (2005) has proposed a method for estimating the confidence intervals for the mean response (μ), for the gamma mean (m), and the predicted response (y) at a new site having similar characteristics as the sites used in the original dataset from which the model was developed. The following table gives the equations for calculating the confidence intervals. In this table, η is the logarithm of the estimated mean response μ, while ϕ is the inverse dispersion parameter.

The primary purpose of this study is to evaluate whether or not there will be any significant difference in the prediction of confidence intervals by modeling SV and MV crashes separately. To examine this difference, confidence intervals for the Poisson mean, the gamma mean and the predicted response were calculated for the total number of crashes (ALL) and for the sum of SV and MV crashes predicted by the SV + MV UNB model and the SV + MV BNB model. For the sake of simplicity, a fixed dispersion parameter was considered during the calculation of confidence intervals, although a varying dispersion parameter can reduce the width of the confidence intervals (Geedipally and Lord, 2008). Furthermore, since the database contains more than 1500 observations, it was assumed that the inverse dispersion parameter is properly estimated (see Lord, 2006).

The modeling procedure was accomplished using the following 4-stage process:

(1) In the initial step, the models for total (fatal or killed, injury type A, injury type B, injury type C, property damage only) or (KABCO) crashes, and fatal and severe injury (KAB) crashes were estimated using a fixed dispersion parameter in SAS (SAS, 2002). Individual models for SV crashes, MV crashes and the total number of crashes (ALL) were estimated using PROC GENMOD, whereas a joint model for SV and MV crashes was estimated using PROC NLMIXED in SAS. It should be noted that the number of years and the length for each site were used as an offset.
(2) The confidence interval for Poisson mean μ, gamma mean m, and the prediction interval y at a new site with the same traffic flow characteristics was then calculated using the equations in Table 1 for the ALL and SV + MV BNB models. In the equation, η is the logarithm of the estimated mean (μ) while ϕ is the inverse dispersion parameter estimated during the fitting process. $Var(\eta)$ is calculated using $XI^{-1}X^T$ where I^{-1} represents the variance–covariance matrices and X is the matrix containing the observed values. The variance–covariance matrices were provided by SAS.
(3) The above step was also calculated for the SV + MV UNB model. The Poisson mean is now normally distributed with mean $\mu = \mu_{SV} + \mu_{MV}$ and variance $\sigma_0{}^2 = \mu^2 Var(\eta)$ where $Var(\eta)$ is calculated as $(\mu^2 Var_{SV}(\eta) + \mu^2 Var_{MV}(\eta))/((\mu_{SV} + \mu_{MV})^2)$ (this equation is based on the fact that SV crashes and MV crashes are two independent random variables). Here μ_{SV} and $Var_{SV}(\eta)$ were estimated from the SV crash model and μ_{MV} and $Var_{MV}(\eta)$ were estimated from the MV crash model. The confidence intervals were then calculated with the calculated means and variances.
(4) The difference in the width of confidence intervals calculated by ALL, SV + MV UNB and SV + MV BNB was then estimated.

3.5. Simulation

A Monte Carlo simulation study was used to verify the results produced using the empirical data. Using simulation, we can establish *a priori* intervals and can assess if the proposed models are working well for identifying these intervals.

The simulation design was accomplished using the following process:

(1) The traffic flow and segment length from the Texas dataset are used as the independent variables.
(2) Generate the "true" mean for SV and MV crashes at each site using $\mu_i = \beta_0 L_i F_i^{\beta_1}$. The parameters are defined prior to the simulation. These parameters are directly taken from the UNB model in Table 4 for SV and MV crashes (UNB model is preferred to other two models, as it provided slightly better fit to the data). The "true" crash mean at each site is calculated by summing the true mean of SV and MV crashes. Similarly, using the defined

Table 2
Descriptive statistics of independent variables and crash data.

Variable	Minimum	Maximum	Mean (Std. dev.)	Total
Segment length (miles)	0.1	6.275	0.55 (0.66)	848.29
AADT (vehicles/day)	42	24,800	6684.3 (4104.9)	–
SV				
Incapacitating injury (A)	0	5	0.21 (0.61)	331
Non-incapacitating injury (B)	0	18	0.51 (1.29)	789
Possible injury (C)	0	10	0.39 (0.95)	619
Fatal (K)	0	3	0.06 (0.26)	88
Non-injury (O)	0	34	0.94 (2.12)	1456
Total crashes	0	64	2.12 (4.31)	3283
MV				
Incapacitating injury (A)	0	9	0.29 (0.79)	458
Non-incapacitating injury (B)	0	18	0.62 (1.44)	966
Possible injury (C)	0	20	0.97 (2.04)	1512
Fatal (K)	0	3	0.08 (0.32)	124
Non-injury (O)	0	17	0.85 (1.81)	1320
Total crashes	0	48	2.82 (5.26)	4380
ALL				
Incapacitating injury (A)	0	9	0.51 (1.12)	789
Non-incapacitating injury (B)	0	31	1.13 (2.24)	1755
Possible injury (C)	0	22	1.37 (2.43)	2131
Fatal (K)	0	6	0.14 (0.45)	212
Non-injury (O)	0	42	1.79 (3.14)	2776
Total crashes	0	99	4.94 (7.95)	7663

parameters, generate the "true" confidence intervals with the SV + MV UNB model.

(3) Simulate SV and MV crash counts for each site. After the simulation, the parameters and confidence intervals are re-estimated by all the three models (as described above in the modeling procedure).

(4) Steps (2) and (3) are repeated for 10 times to obtain statistically reliable estimates. The average values for the parameters and confidence intervals are finally used. The average confidence intervals obtained are then compared with the true intervals.

4. Data description

The statistical models were estimated using crash data collected at 1552 Texas undivided four-lane highway segments. The data were obtained from the Texas Department of Public Safety and the Texas Department of Transportation (TxDOT) for the years 1997–2001. The data included information on segment length, ADT, number of intersections, number of horizontal curves and other variables that influence crash risk. The crash data also provided details about the severity as well as the number of vehicles involved in the collision. Only head-on and rear-end collisions were considered for MV crashes. A total of 3283 SV crashes and 4380 MV crashes were extracted. Table 2 provides relevant descriptive statistics for key explanatory variables and the crash data.

Fig. 1 shows the proportion by severity levels for SV, MV and ALL crashes. As seen in this figure, SV crashes have a much larger percentage of non-injury than MV crashes. This is expected since SV crashes on average have a lesser transfer of energy than MV crashes.

Table 3
Correlation between SV and MV crashes.

Severity	Crash type	SV	MV
KABCO	SV	1	0.372
	MV	0.372	1
KAB	SV	1	0.431
	MV	0.431	1

Table 3 shows the correlation matrix between SV and MV crashes for total crashes and fatal and serious injury crashes. As seen in this table, the correlation is larger for KAB crashes than KABCO crashes. Although the crash types are not highly correlated, they cannot be assumed to be independent either. Because of this, a bivariate model needs to be used for analyzing SV and MV crashes simultaneously.

5. Results

This section describes the results of the comparison analysis between SV, MV and ALL crash models using empirical data as well as the output of the Monte Carlo simulation study. The confidence intervals were calculated for total crashes (KABCO) and fatal and serious injury (KAB) crashes for the empirical data and for total crashes for the simulation study.

5.1. Empirical data

Table 4 provides the parameter estimates of the univariate NB models with their associated standard errors. The coefficient (β_1) shows that the SV and ALL crashes increase at a decreasing rate as traffic flow increases, whereas MV crashes increase almost linearly with the traffic flow. The dispersion parameter (α) reveals that the MV crashes are more dispersed than the SV and ALL crashes.

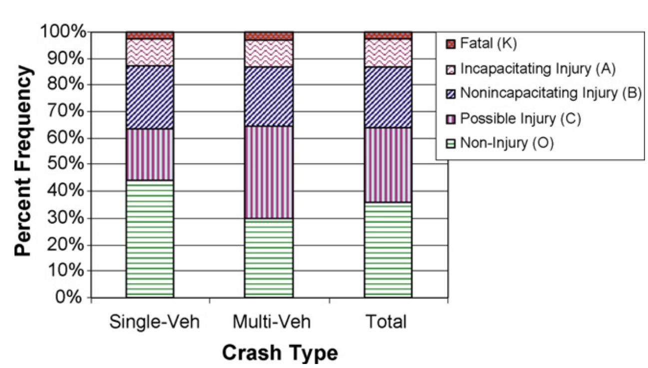

Fig. 1. Percent frequencies of crash severities.

Table 4
Estimates of model coefficients (and standard errors) for univariate NB models.

Crash type	Severity	$\ln(\beta_0)$	β_1	$\alpha\ (1/\phi)$
SV	KABCO	−5.9403 (0.4708)	0.6535 (0.0539)	1.0616 (0.0734)
	KAB	−5.5348 (0.6075)	0.4878 (0.0695)	1.1895 (0.1321)
MV	KABCO	−8.9067 (0.5650)	1.0504 (0.0645)	1.6883 (0.0877)
	KAB	−8.9794 (0.6646)	0.9233 (0.0753)	1.4173 (0.1175)
ALL	KABCO	−6.9613 (0.4075)	0.8839 (0.0467)	0.9981 (0.049)
	KAB	−6.7263 (0.4915)	0.7288 (0.0560)	1.0093 (0.0737)

Table 5
Estimates of model coefficients (and standard errors) for bivariate NB models.

Severity	Crash type	$\ln(\beta_0)$	β_1	$\ln(r)$	δ
KABCO	SV	−3.4303 (0.5919)	0.4156 (0.06389)	−0.1816	−0.1967
	MV	−7.7371 (0.6006)	0.9288 (0.06497)	(0.0581)	(0.0406)
KAB	SV	−5.0467 (0.5918)	0.4525 (0.06619)	−0.0451	−0.0016
	MV	−9.1087 (0.6186)	0.9357 (0.06866)	(0.0746)	(0.0002)

Table 6
Variance–covariance matrices for univariate NB models.

Crash type	Severity		$(\ln\beta_0)$	β_1
SV	KABCO	$\ln\beta_0$	0.22164	−0.02530
		β_1	−0.002530	0.002905
	KAB	$\ln\beta_0$	0.36904	−0.04209
		β_1	−0.04209	0.004828
MV	KABCO	$\ln\beta_0$	0.31925	−0.03637
		β_1	−0.03637	0.004164
	KAB	$\ln\beta_0$	0.44167	−0.04993
		β_1	−0.04993	0.005671
ALL	KABCO	$\ln\beta_0$	0.16609	−0.01897
		β_1	−0.01897	0.002178
	KAB	$\ln\beta_0$	0.24157	−0.02747
		β_1	−0.02747	0.003141

Table 5 gives the parameter estimates (with their associated standard errors) of the bivariate NB models. The parameter 'r' is the combined inverse dispersion parameter of the joint mode and the parameter delta 'δ' characterizes the correlation between SV and MV crashes. For both KABCO and KAB crashes, 'δ' is significantly different from zero, which suggests that there exists a correlation between SV and MV crashes.

Tables 6 and 7 provide the variance–covariance matrices for univariate and bivariate models, respectively. These values are useful in computing the confidence intervals for Poisson mean, gamma mean and predictive response.

Table 8 gives the goodness-of-fit statistics for each type of model. This table shows that there is no clear difference in the prediction of total crashes with each type of the model. The SV + MV bivariate NB model is found to provide a slightly better statistical fit than the other two types of model (based on the MSPE) for KABCO crashes. There is also no significant difference in the prediction of KAB crashes among each type of the model. Overall, the SV + MV (UNB) model fits the data slightly better than the other two models.

Table 8
Goodness-of-fit statistics.

Severity	Crash type	MAD	MSPE
KABCO	SV + MV (UNB)	4.022	48.683
	SV + MV (BNB)	4.114	47.857
	ALL	4.044	48.719
KAB	SV + MV (UNB)	1.512	6.096
	SV + MV (BNB)	1.525	6.159
	ALL	1.522	6.141

Fig. 2 illustrates the relationship between total crashes and traffic flow for the ALL, SV + MV (UNB) and SV + MV (BNB) models, respectively. From this figure, the crash–flow relationships indicate that the ALL model and SV + MV (BNB) model predicts the total number of crashes at a decreasing rate as traffic flow increases. This relationship basically means that there are proportionally less crashes per passing vehicles as the traffic flow increases and thus the crash risk per vehicle diminishes when traffic flow increases. However, the SV + MV (UNB) model shows a linear relationship between total crashes and traffic flow. Fig. 2 also shows that the SV + MV (BNB) model always predicts more crashes than the ALL and SV + MV (UNB) crash model for traffic flows less than 7000 vehicles/day. At higher traffic flows (flow > 15,000), the SV + MV (UNB) predicts more crashes than other two models.

Table 7
Variance–covariance matrices for bivariate NB models.

Severity	Crash type		SV		MV		$\ln(r)$
			$(\ln\beta_0)$	β_1	$(\ln\beta_0)$	β_1	
KABCO	SV	$\ln\beta_0$	0.3503	−0.03755	0.2774	−0.02943	−0.01691
		β_1	−0.03755	0.004082	−0.02942	0.003171	0.001488
	MV	$\ln\beta_0$	0.2774	−0.02942	0.3607	−0.03877	−0.01642
		β_1	−0.02943	0.003171	−0.03877	0.004222	0.001442
	$\ln(r)$		−0.01691	0.001488	−0.01642	0.001442	0.003374
KAB	SV	$\ln\beta_0$	0.3503	−0.03872	0.1745	−0.01885	−0.00934
		β_1	−0.03872	0.004381	−0.0188	0.002117	0.000393
	MV	$\ln\beta_0$	0.1745	−0.01880	0.3827	−0.04204	−0.00953
		β_1	−0.01885	0.002117	−0.04204	0.004715	0.000417
	$\ln(r)$		−0.00934	0.000393	−0.00953	0.000417	0.005559

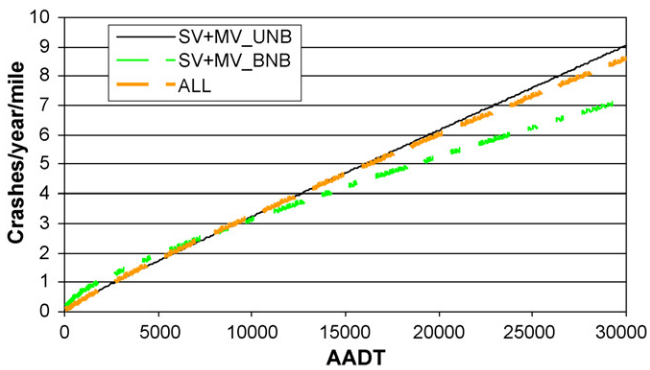

Fig. 2. Crash–flow relationship for total crashes.

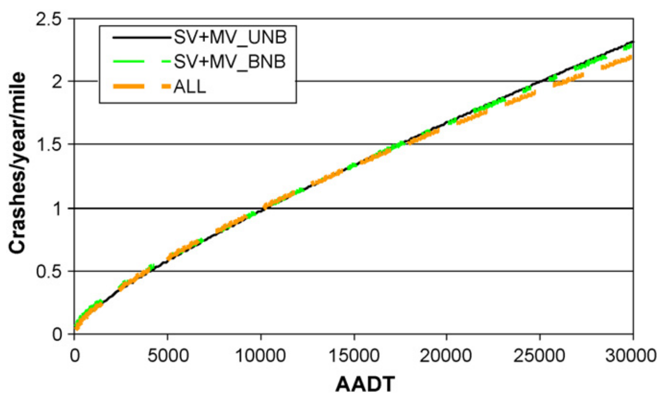

Fig. 3. Crash–flow relationship for KAB crashes.

Fig. 3 shows the relationship between severe crashes (KAB) and traffic flow for the ALL, SV + MV (UNB) and SV + MV (BNB) models. All model types predict crashes at a decreasing rate as traffic flow increases. Both the SV + MV (UNB) and SV + MV (BNB) models predict almost the same number of crashes for the entire range of traffic flow volumes. The ALL crash model predicts fewer crashes than other two models at larger traffic flows.

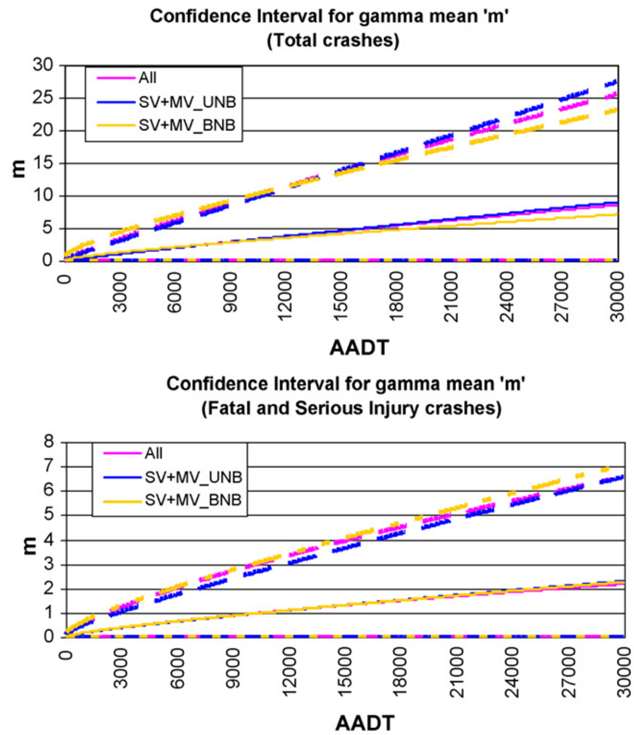

Fig. 5. 95-Percentile confidence intervals for the Gamma mean (m).

Fig. 4 shows the confidence intervals for the Poisson mean 'μ' for KABCO crashes, and KAB (fatal and serious injuries) crashes. This figure illustrates that the width of the interval for Poisson mean 'μ'

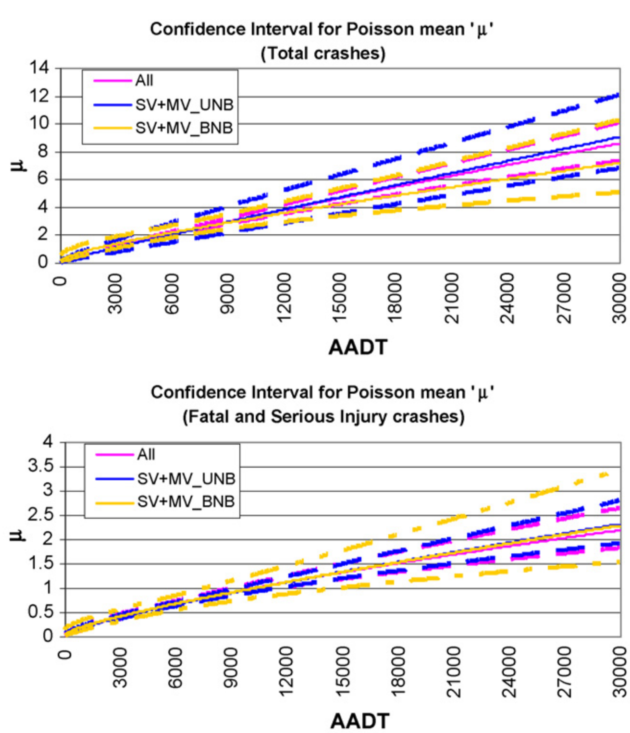

Fig. 4. 95-Percentile confidence intervals for the Poisson mean (μ).

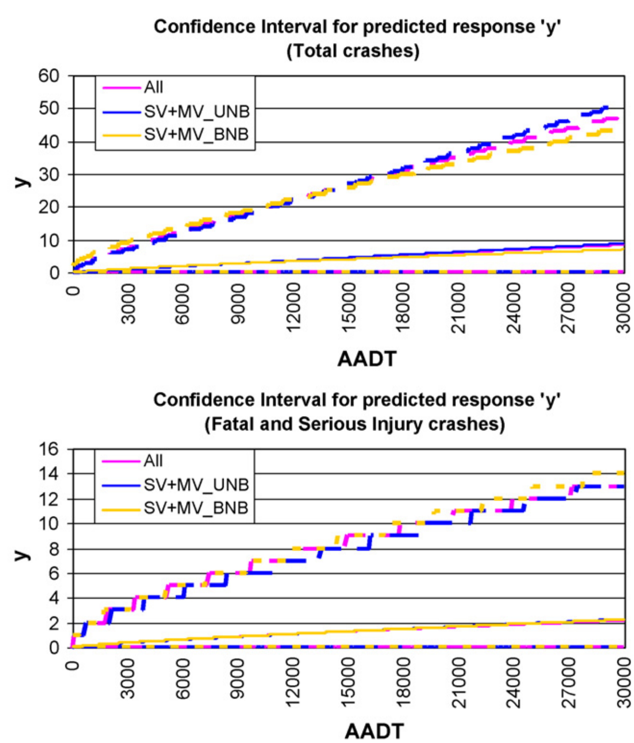

Fig. 6. 95-Percentile confidence intervals for the predictive response (y).

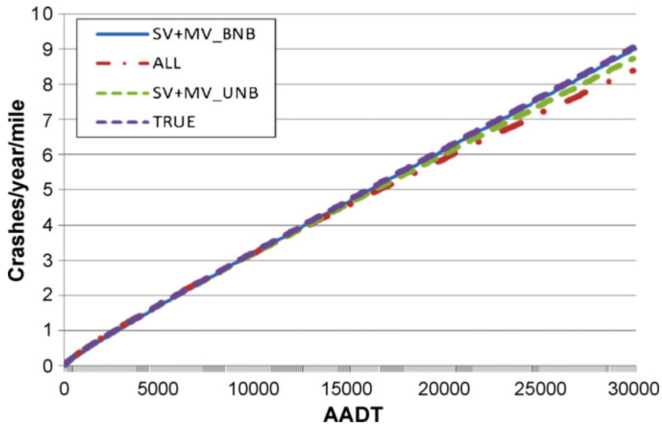

Fig. 7. Comparison of crash means.

for total crashes (i.e., the distance between the mean response and the upper confidence interval boundary) predicted by the SV + MV UNB and SV + MV BNB models is always wider than the width predicted by the ALL crash model for KABCO crashes. (*Note*: The CIs overlap each other for KABCO crashes and this is caused because the difference in predicted values is greater than the one for KAB crashes; when the absolute different between the CI and the mean is examined, the same results are noted for both crash severities.) For fatal and serious injury crashes, the width of the interval for Poisson mean predicted by SV + MV UNB is narrower than the width predicted by the ALL crash model for most of the time. The SV + MV BNB model always predicts wider interval width than the ALL crash model, indicating that the two crash types are correlated. The maximum relative difference in the mean of predicted crashes between SV + MV UNB model and ALL crash model is 21% for KABCO crashes and 25% for KAB crashes. Similarly, the maximum relative difference in the mean of predicted crashes between SV + MV BNB model and ALL crash model is 73% for KABCO crashes and 40% for KAB crashes.

Fig. 5 shows the confidence intervals for the gamma mean '*m*' for KABCO crashes and KAB crashes. The figure illustrates that the confidence interval for the gamma mean '*m*' for KABCO crashes predicted by SV + MV BNB model is narrower than the ALL and SV + MV UNB model for flows less than 15,000 vehicles/day. For KAB crashes, the interval predicted by SV + MV BNB is wider than that predicted by ALL crash model. The SV + MV UNB model predicted

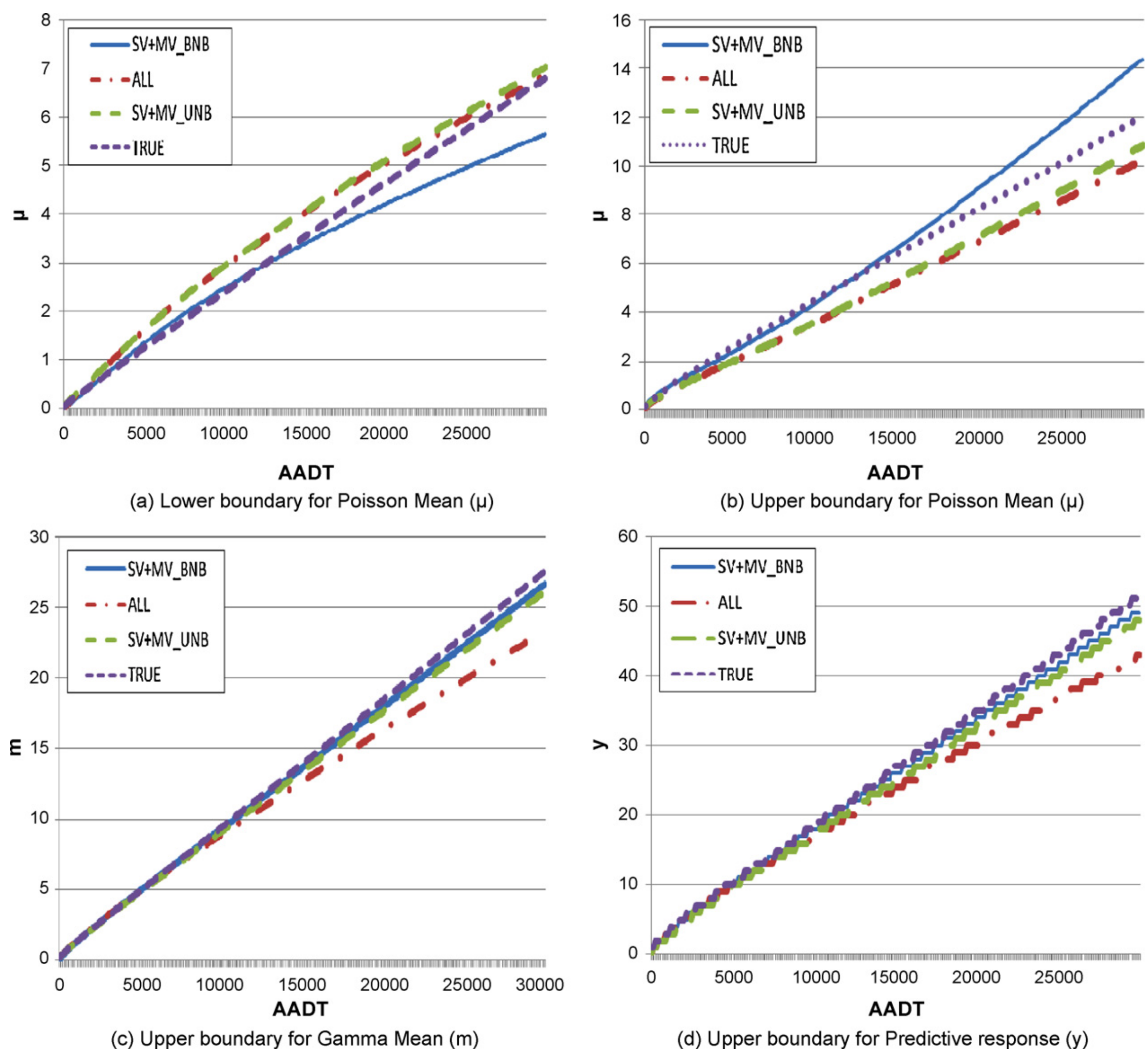

Fig. 8. Comparison of 95-percentile confidence intervals. (a) Lower boundary for Poisson mean (μ). (b) Upper boundary for Poisson mean (μ). (c) Upper boundary for Gamma mean (m). (d) Upper boundary for predictive response (y).

narrower confidence intervals than the other two type of modeling frameworks.

Fig. 6 shows the confidence intervals for the predictive response 'y' for KABCO and KAB crashes. Since the predicted number of crashes must be non-negative integer, the estimated values were rounded to the nearest integer. The figure illustrates that the SV + MV UNB model predicts narrower confidence intervals than the ALL crash model for at least half of the time for KABCO crashes and at least 77% of the time for KAB crashes. On the other hand, the SV + MV BNB model predicts wider confidence intervals than the ALL crash model for at least half of the time for both KABCO crashes and KAB crashes. As discussed by Lord (2008), the predicted response and its associated variance are the most important values to be computed, since most of the time, analysts will apply the model to other datasets (e.g., comparing different highway design alternatives).

In sum, the comparison analysis showed that using two distinct models to predict crashes for highway segments most often provides wider confidence intervals than using one model for all categories of crashes, especially for the Poisson and gamma mean values. Although each subset contains fewer crashes or observations and the functional forms are vastly different, the summation of the variances is higher than the variance estimated from the full dataset by a single model. Overall, a bivariate NB model provides wider confidence intervals than a univariate NB model, as expected. This is attributed to fact that the joint model takes correlation of the dependent variables into effect which increase the variance of the mean values and predicted response.

5.2. Simulation

Fig. 7 shows the comparison between the true crash means and the crash means predicted by the ALL, SV + MV (UNB) and SV + MV (BNB) models. From this figure, it is evident that the SV + MV (BNB) model predicts the values that are much closer to the true crash means. Also, the difference between the values predicted by the ALL model and the true crash means are much larger than the difference predicted by SV + MV (UNB) and SV + MV (BNB) models.

Fig. 8 illustrates the comparative analysis of the 95% confidence intervals. It is clearly seen that the SV + MV models predict intervals that are much closer to true intervals than the ALL model. More specifically, the SV + MV (BNB) model predicts much closer values to the true intervals, at least for gamma mean and predictive response.

The results of the analysis with the empirical and simulated data show that a separate model increases the variance of mean and predicted response. However, the values predicted by a separate model are much closer to the true values. Thus, this study supports previous work, which recommended that SV and MV crashes should be modeled separately. Furthermore, a joint model (which considers the correlation between dependent variables) is recommended when modeling SV and MV crashes, even though the confidence intervals will be larger.

6. Summary and conclusions

This paper documented a research study that examined the potential differences in the prediction of confidence intervals of Poisson–gamma models when SV and MV crashes are modeled separately and together. Recently, some transportation safety analysts have recommended modeling both categories of crashes separately for estimating the safety performance of highway segments. To accomplish the comparison analysis, crash data were collected on four-lane rural highways in Texas. Then, KABCO and KAB models were estimated for SV, MV and ALL crashes along with the confidence intervals for the Poisson mean (μ), gamma mean (m), and predictive response (y), respectively. A simulation study was also conducted to do the comparative analysis.

The following results were obtained from the analysis:

(1) There is a clear difference in the prediction of confidence intervals for the Poisson mean, gamma mean, and predictive response between aggregated crash prediction (ALL) and the summation of distinct models (SV and MV). Overall, the bivariate NB model provides wider confidence intervals than the univariate model.
(2) Combining SV and MV models predicts wider confidence intervals than developing a single model for both total (KABCO) crashes and fatal and serious injuries (KAB) crashes.
(3) Although wider confidence intervals were observed with SV + MV UNB and BNB models, these intervals are closer to the "true intervals." Also, the crash mean predicted by a separate model is closer to the true crash mean. Thus, this paper supports the research done by others that recommended analyzing SV and MV crashes separately. However, given the correlation, a joint NB model should be utilized.

There are several avenues for further work. First, it is recommended to conduct similar analyses using data collected on other types of highways, such as two-lane and divided highways as well as controlled-access facilities. As part of these analyses, the effect of the sample size for each subset should be examined more carefully. This would give us the opportunity to determine whether sample size plays an important role in the computation of the confidence intervals when the ratio between SV and MV crashes is different. Second, since the analysis was carried with only one covariate, namely ADT, one should also evaluate the influence of models containing several covariates on the computation of confidence intervals for a single versus two distinct models. The approach proposed by Wood (2005) is still appropriate for computing confidence intervals for models that includes several covariates (see Lord, 2008). Finally, given the role confidence intervals can play in the identification of hazardous sites (i.e., hot spot identification) and selecting highway design alternatives, it is suggested to compare how well two distinct models can influence both kinds of application versus the use of a single model.

Acknowledgments

The authors would like to thank Dr. Samiran Sinha from the Department of Statistics at Texas A&M University for his help in deriving the equation for computing the variance of the sum of SV and MV crashes. The authors would also like to thank two anonymous reviewers for providing useful comments to improve this paper.

References

Bijleveld, F.D., 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis and Prevention 37 (4), 591–600.
Bonneson, J., Lord, D., Zimmerman, K., Fitzpatrick, K., Pratt, M., 2007. Development of tools for evaluating the safety implications of highway design decisions. In: TTI Report FHWA/TX-07/0-4703-4. Texas Transportation Institute, College Station, TX.
Cameron, A.C., Trivedi, P.K., 1998. Regression analysis of count data, Econometric Society Monograph No. 30, Cambridge University Press.
Cercarelli, L.R., Arnold, P.K., Rosman, D.L., Sleet, D., Thornett, M.L., 1992. Travel exposure and choice of comparison crashes for examining motorcycle conspicuity by analysis of crash data. Accident Analysis and Prevention 24 (4), 363–368.
Geedipally, S.R., Lord, D., 2008. Effects of the varying dispersion parameter of Poisson–gamma models on the estimation of confidence intervals of crash prediction models. Transportation Research Record 2061, 46–54.
Griffith, M.S., 1999. Safety evaluation of rolled-in continuous shoulder rumble strips installed on freeways. Transportation Research Record 1665, 28–35.

Harwood, D.W., Bauer, K.M., Richard, K.R., Gilmore, D.K., Graham, J.L., Potts, I.B., Torbic, D.J., Hauer, E., 2007. Methodology to predict the safety performance of urban and suburban arterials. In: NCHRP Web Document No. 129: Phases I and II. Transportation Research Board, Washington, DC.
Hauer, E., 1997. Observational Before–After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Elsevier Science Ltd., Oxford.
Ivan, J.N., Pasupathy, R.K., Ossenbruggen, P.J., 1999. Differences in causality factors for single and multi-vehicles crashes on two-lane roads. Accident Analysis and Prevention 31, 695–704.
Ivan, J.N., Wang, C., Bernardo, N.R., 2000. Explaining two-lane highway crash rates using land use and hourly exposure. Accident Analysis and Prevention 32 (6), 787–795.
Ivan, J., 2004. New approach for including traffic volumes in crash rate analysis and forecasting. Transportation Research Record 1897, 134–141.
Jonsson, T., Ivan, J., Zhang, C., 2007. Crash prediction models for intersections on rural multilane highways: differences by collision type. Transportation Research Record 2019, 91–98.
Jovanis, P.P., Chang, H.L., 1986. Modeling the relationship of accidents to miles traveled. Transportation Research Record 1068, 42–51.
Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accident Analysis and Prevention 37 (4), 775–786.
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., 2002. Issues related to the application of accident prediction models for the computation of accident risk on transportation networks. Transportation Research Record 1784, 17–26.
Lord, D., 2006. Modeling motor vehicle crashes using Poisson–gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38 (4), 751–766.
Lord, D., 2008. Methodology for estimating the variance and confidence intervals of the estimate of the product of baseline models and AMFs. Accident Analysis and Prevention 40 (3), 1013–1017.
Lord, D., Bonneson, J.A., 2005. Calibration of predictive models for estimating the safety of ramp design configurations. Transportation Research Record 1908, 88–95.
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 and Prevention 37 (1), 35–46.
Lord, D., Bonneson, J.A., 2007. Development of accident modification factors for rural frontage road segments in Texas. Transportation Research Record 2023, 20–27.
Lord, D., Washington, S.P., Ivan, J.N., 2007. Further notes on the application of Zero Inflated models in highway safety. Accident Analysis and Prevention 39 (1), 53–57.
Lord, D., Mannering, F., in press. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research, Part A. doi:10.1016/j.tra.2010.02.001.
Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., van Schalkwyk, I., Ivan, J.N., Lyon, C., Jonsson, T., 2008. Methodology for estimating the safety performance of multilane rural highways. In: NCHRP Web-Only Document 126. National Cooperation Highway Research Program, Washington, DC.
Lyon, C., Oh, J., Persaud, B.N., Washington, S.P., Bared, J., 2003. Empirical investigation of the IHSDM accident prediction algorithm for rural intersections. Transportation Research Record 1840, 78–86.
Ma, J., K.M. Kockelman, 2006. Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity. In Transportation Research Record: Journal of the Transportation Research Board, No. 1950, Transportation Research Board of the National Academies, Washington, DC, pp. 24–34.
Maher, M.J., 1990. A bivariate negative binomial model to explain traffic accident mitigation. Accident Analysis and Prevention 22, 487–498.
Maher, M.J., Summersgill, I., 1996. A comprehensive methodology for the fitting predictive accident models. Accident Analysis and Prevention 28 (3), 281–296.
Mensah, A., Hauer, E., 1998. Two problems of averaging arising in the estimation of the relationship between accidents and traffic flow. Transportation Research Record 1635, 37–43.
Miaou, S.-P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial dependence. Accident Analysis and Prevention 37 (4), 699–720.
Miller, T.R., Levy, D., Spicer, R., Lestina, D., 1998. Allocating the cost of motor vehicle crashes between vehicle types. Transportation Research Record 1635, 81–97.
Oh, J., Lyon, C., Washington, S.P., Persaud, B.N., Bared, J., 2003. Validation of the FHWA crash models for rural intersections: lessons learned. Transportation Research Record 1840, 41–49.
Ostrom, M., Eriksson, A., 1993. Single-vehicle crashes and alcohol: a retrospective study of passenger car fatalities in Northern Sweden. Accident Analysis and Prevention 25 (2), 171–176.
Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1–6.
Persaud, B.N., Mucsi, K., 1995. Microscopic accident potential models for two-lane rural roads. Transportation Research Record 1485, 134–139.
Persaud, B., Nguyen, T., 1998. Disaggregate safety performance models for signalized intersections on Ontario provincial roads. Transportation Research Record 1635, 113–120.
Persaud, B.N., Retting, R., Garder, P., Lord, D., 2001. Observational before-after study of U.S. roundabout conversions using the empirical Bayes method. Transportation Research Record 1751, 1–8.
Persaud, B.N., Lord, D., Palminaso, J., 2002. Issues of calibration and transferability in developing accident prediction models for urban intersections. Transportation Research Record 1784, 57–64.
Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis and Prevention 36 (2), 183–191.
SAS Institute Inc., 2002. Version 9 of the SAS System for Windows. SAS Institute, Cary, NC.
Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometric and environmental factors on rural freeway accident frequencies. Accident Analysis and Prevention 27 (3), 371–389.
Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97, 246–273.
Subrahmaniam, K., Subrahmaniam, K., 1973. On the estimation of the parameters in the bivariate negative binomial distribution. Journal of the Royal Statistical Society 35, 131–146.
Tarko, A.P., Inerowicz, M., Ramos, J., Li, W., 2008. Tool with road-level crash prediction for transportation safety planning. Transportation Research Record 2083, 16–25.
Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics 31 (2–3), 221–229.
Wood, G.R., 2005. Confidence and prediction intervals for generalized linear accident models. Accident Analysis and Prevention 37 (2), 267–273.
Zador, P.L., 1985. Motorcycle headlight-use laws and fatal motorcycle crashes in the US. American Journal of Public Health 75 (5), 543–546.