On the nature of over-dispersion in motor vehicle crash prediction models

Sudeshna Mitra [a], Simon Washington [b]

[a] *Civil & Environmental Engineering Department, Cal Poly State University, San Luis Obispo, CA 93407-0353, United States*
[b] *Department of Civil & Environmental Engineering, Arizona State University, PO Box 875306, Tempe, AZ 85287-53006, United States*

Abstract

Statistical modeling of traffic crashes has been of interest to researchers for decades. Over the most recent decade many crash models have accounted for extra-variation in crash counts—variation over and above that accounted for by the Poisson density. The extra-variation – or dispersion – is theorized to capture unaccounted for variation in crashes across sites. The majority of studies have assumed fixed dispersion parameters in over-dispersed crash models—tantamount to assuming that unaccounted for variation is proportional to the expected crash count. Miaou and Lord [Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transport. Res. Rec. 1840, 31–40] challenged the fixed dispersion parameter assumption, and examined various dispersion parameter relationships when modeling urban signalized intersection accidents in Toronto. They suggested that further work is needed to determine the appropriateness of the findings for rural as well as other intersection types, to corroborate their findings, and to explore alternative dispersion functions.

This study builds upon the work of Miaou and Lord, with exploration of additional dispersion functions, the use of an independent data set, and presents an opportunity to corroborate their findings. Data from Georgia are used in this study. A Bayesian modeling approach with non-informative priors is adopted, using sampling-based estimation via Markov Chain Monte Carlo (MCMC) and the Gibbs sampler. A total of eight model specifications were developed; four of them employed traffic flows as explanatory factors in mean structure while the remainder of them included geometric factors in addition to major and minor road traffic flows. The models were compared and contrasted using the significance of coefficients, standard deviance, chi-square goodness-of-fit, and deviance information criteria (DIC) statistics. The findings indicate that the modeling of the dispersion parameter, which essentially explains the extra-variance structure, depends greatly on how the mean structure is modeled. In the presence of a well-defined mean function, the extra-variance structure generally becomes insignificant, i.e. the variance structure is a simple function of the mean. It appears that extra-variation is a function of covariates when the mean structure (expected crash count) is poorly specified and suffers from omitted variables. In contrast, when sufficient explanatory variables are used to model the mean (expected crash count), extra-Poisson variation is not significantly related to these variables. If these results are generalizable, they suggest that model specification may be improved by testing extra-variation functions for significance. They also suggest that known influences of expected crash counts are likely to be different than factors that might help to explain unaccounted for variation in crashes across sites.

Keywords: Over-dispersion; Crash prediction; Bayesian method; Intersection safety

1. Introduction and background

Research on highway and traffic safety has been of great interest to engineers and planners for decades. Major factors known to affect safety are driver characteristics, vehicle features, exposure to risk (traffic volumes), traffic control, weather conditions, and roadway design characteristics. To predict the safety of transportation system traffic engineers model crash rate or frequency as a function of the above mentioned factors. Of course these measurable factors do not completely explain accident occurrence and so the models typically used are stochastic models including a disturbance or error term.

Ideally, the mathematical relationship between crashes and various explanatory factors is specified correctly so as to reveal the underlying effects on safety and to enable useful insights into the underlying crash process. A significant proportion of past studies have been focused on modeling accidents at roadway intersections, due primarily to the relative share of accidents at intersections. While the majority of these studies report on urban

signalized intersections, there is considerable work on rural and unsignalized intersections as well. Relatively early work on this topic is reported by Chapman (1972), Satterthwaite (1981), and Hauer et al. (1988). Further development and continuation of the previous crash modeling research is reported by Bonneson and McCoy (1993), Belanger (1994), Maher and Summersgill (1996), Persaud and Nguyen (1998), Lord and Persaud (2000), Wang and Nihan (2004), Persaud et al. (2002), and Miaou and Lord (2003). A main focus of prior studies has been to identify a defensible statistical relationship between crash counts and exposure. Researchers not only considered the effect of total traffic flow, but have also accounted for various turning movements, pedestrian flows, and even bicycle flows in some cases to model different crash outcomes. Statistical modeling of crashes remains a significant area of ongoing research.

During the past decade or so researchers have begun to extensively use the negative binomial (NB) model for modeling crashes. The NB model arises mathematically (and conveniently) by assuming that unobserved crash heterogeneity (variation) across sites (intersections, road segments, etc.) is gamma distributed, while crashes within sites are Poisson distributed (Washington et al., 2003). The Poisson, Poisson-Gamma (NB), and other related models are collectively called generalized linear models (GLM). In the majority of NB modeling, covariates are used to forecast the Poisson mean, while the gamma heterogeneity is assumed constant. The functional form of the NB model is as follows:

$$\mu_i = \exp(\beta X_i)\exp(\varepsilon_i) \tag{1}$$

where μ_i is the Poisson mean, X_i's are the vector of various covariates that include geometric characteristics and exposure, β's the vector of unknown fixed-effect parameters and $\exp(\varepsilon_i)$ is the gamma distributed error with mean 1 and variance $1/\alpha$ (where α is the inverse dispersion parameter and α is greater than 0). This negative binomial formulation provides flexible properties where mean = Poisson mean or $\exp(\beta X_i)$ and variance = $\exp(\beta X_i)[1 + \exp(\beta X_i)/\alpha]$, a simple function of mean. While GLM employs a simple variance structure as a function of the mean structure, statisticians like Dey et al. (1997) commented that in some applications, heterogeneity in the sample is too great to be explained by the simple variance function implicit in the GLM. This is particularly important in the presence of important omitted variable that influence the mean structure. To deal with this issue they performed a Bayesian modeling with ship-accident count data, previously used by other statisticians (McCullagh and Nelder, 1983) using GLM and found that the variance function has a particular structure. While modeling traffic crash-flow relationship for intersections, Miaou and Lord (2003) also challenged the assumption of fixed dispersion parameter. Due to complexity and interaction of traffic flow in and around intersections, they suspected that the unmodeled heterogeneity of the mean of crash counts would be spatially structured. This means that the variance of NB model is not a simple function of mean as explained before but contains a dispersion function that depends on site-specific characteristics such as major and minor road traffic flows. Following this the dispersion function they suggested is as follows:

$$\alpha_i = \exp\left(\frac{\eta_0 + \eta_1 F_{1,i} + \eta_2 F_{2,i} + \eta_3 F_{2,i}}{F_{1,i}}\right) \tag{2}$$

where F_1 and F_2 are respectively major and minor road flow and η's are the vector of covariates. This leads to the modeling of gamma heterogeneity in terms of included covariates which will be discussed later in more detail.

While modeling urban intersection crashes using this functional form for variance structure, Miaou and Lord (2003) found an improvement in model estimation. In addition they mentioned that treatment of dispersion parameter as a fixed parameter can seriously undermine the goodness of the estimate for individual sites by up to about 35%. Further, they suggested that more work is needed to determine the appropriateness of the findings for rural as well as other intersection types, to corroborate their findings, and to explore alternative dispersion functions.

This study builds upon the work of Miaou and Lord (2003), with exploration of additional dispersion functions, the use of an independent data set from rural intersections of Georgia, and an opportunity to corroborate findings. In addition to the exposure data, information about geometric design and lighting conditions of the intersections are considered in this study. A Bayesian modeling approach with non-informative prior is adopted over the classical methodology to check the structure of gamma heterogeneity. While the intent of using Bayesian methodology over the classical methods is never to show any superiority of a Bayesian approach, it is still important to point out some key reasons as to why it was preferred over the classical framework. As correctly specified by Dey et al. (1997), a Bayesian estimation method using non-informative prior is something where we let the data decide the inference. Hence, this is an approximate likelihood-based inference but with more reliable estimates of variability for small sample sizes. Hence, it can be concluded that for large sample sizes this inference will be closer to that of maximum likelihood and for smaller samples this estimate of variability should be more appropriate than asymptotic estimates from maximum likelihood. In the following sections, discussion about the Bayesian modeling approach with sampling-based estimation and Markov Chain Monte Carlo (MCMC) techniques, such as the Gibbs sampler are given first. This is then followed by the methodology of the study and the goodness-of-fit criteria used. Finally, summary of the study findings and conclusions are presented.

2. Bayesian modeling approach

While the Bayesian statistical analysis have been widely used in the field of health science and social sciences, the application in the field of transportation engineering is few if not absent (Hauer, 1992; Heydecker and Wu, 2001; and Miaou and Lord, 2003). Although the argument about the superiority between the Bayesian and classical method existed for decades, this study chose to use the Bayesian approach for some specific reasons as described in the following paragraphs. Also, brief descriptions about the Bayesian method along with Markov Chain

Monte Carlo (MCMC) technique and Gibbs sampling are given here.

As described by Congdon (2003), the determination of parameters in the traditional classical estimation processes is aimed at finding a single optimum estimate using asymptotic normality based on large samples, at a particular confidence interval around its mode, whereas the Bayesian estimation determines posterior density for each parameter under consideration. This density estimation is the product of a process where a long run or a series of long runs of samples are taken from the posterior density based on the prior information about the parameter as well as the data. As a result a Bayesian modeling approach provides a considerable interpretive advantage because posterior estimates reflect the probabilities that the analyst is primarily interested in, the probability of the null hypothesis being true (called a posterior credible interval or credible set) (Washington et al., 2005). In contrast, classical confidence intervals on parameter estimates provide the probability of observing data given that a parameter(s) takes on a specific value. "Bayesians" argue that this distinction provides a considerable philosophical and practical advantage over classical estimation methods.

Instead of the difference in approach, the classical maximum likelihood estimation (MLE) and Bayesian analysis are closely related. As mentioned before, inference in MLE is based on the likelihood of the data alone, whereas in Bayesian models, the likelihood of the observed data x given parameters θ, is used to modify the prior beliefs $\pi(\theta)$, with the updated knowledge summarized in a posterior density $\pi(\theta|x)$. An example using a linear regression model with form:

$$\hat{y}_k = \beta_0 + \beta_1 x_k \tag{3}$$

in which a prior experience or evidence-based research may provide analyst with information regarding the expected values of some or all of the β values. However, in the absence of existing knowledge, it is difficult to form a so-called "informative prior" and the application of non-informative or flat prior comes into play. For example, if we consider the slope parameter of the regression equation is distributed such that $\beta_1 \sim N(\mu,\tau)$, and both μ and τ are known, then a lower value of τ would indicate a higher variance of the distribution, i.e. a lower precision of the coefficient estimate, while larger values of τ would be associated with very low variance which would be indicated as informative prior. In contrast a large value of $\sigma = 1/\tau$, is considered for a non-informative or diffuse prior. However, as mentioned in Section 1, even using non-informative prior we would get a better estimate compared to MLE in case of small sample sizes due to the limitation of the asymptotic normality assumption of MLE. Also as described by experts like Congdon (2003), in case of discrete outcomes, a sampling-based approach to estimation can be beneficial to screen out marginally important predictors, or in comparing between non-nested models. Hence, with the prior knowledge about the distribution of the parameter, $\pi(\theta)$, as well as the data x we get the posterior information in a Bayesian framework as

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) \tag{4}$$

In this context, it is worth mentioning that for the prior information in Bayesian analysis, the location of the parameter (mean and mode) and the precision (the reciprocal of variance) of the prior are more critical than its actual shape in terms of conveying prior information (Walsh, 2004). The shape of the prior distribution is often chosen to facilitate calculation of the prior through the use of conjugate priors for the posterior distribution. However, in the absence of conjugate priors, much difficulty arises in terms of summarizing the marginal posterior of a particular parameter of θ, if the dimension of θ is really large. In that case it is necessary to integrate out the joint distribution of Eq. (4), excluding the one parameter of particular interest. While calculation of such an integral was computationally demanding in the past, with today's desktop computer, the sampling of the posterior distribution using a Markov Chain and Monte Carlo technique becomes relatively easy.

Markov Chain Monte Carlo (MCMC) methods are used to repeatedly sample from the joint posterior distribution. A Markov chain operates in discrete time intervals to produce a sequence of evolving random variables, with the probability of transition dependent on its current time. Chains are generated from a transition kernel, a conditional probability density function. The resulting chains enjoy very strong stability with desirable properties (Washington et al., 2005). In particular, a stationary probability distribution exists by construction of the chain and convergence to the limiting or stationary distribution occurs almost surely, or after the chain converges.

The Gibbs sampler is an application of the MCMC process that allows the user to generate random variables indirectly from the marginal density function without actually computing the density itself (Casella and George, 1992). The key to the Gibbs sampler is that only univariate conditional distributions are considered: the distribution when all of the random variables but one is assigned fixed values (Walsh, 2004). Thus, one simulates n random variables sequentially from the n univariate conditionals rather than generating a single n dimensional vector in a single pass using the full joint distribution. The general principal of the Gibbs sampler as given by Robert and Casella (1997), involves successive sampling from the complete conditional densities:

$$f(\theta_k|x, \theta_1, \theta_2, \ldots, \theta_{k+1}, \ldots, \theta_p) \tag{5}$$

which condition both the data and the other parameters. The Gibbs sequence converges to a stationary distribution that is independent of starting values and it converges to a stationary sampling distribution of the posterior density (Congdon, 2003). Greater details of the MCMC and Gibbs sampling can be obtained from Norris (1997), Robert and Casella (1997) and Casella and George (1992).

3. Methodology and analytical approach

Extensive work by researchers like Joshua and Garber (1990), Miaou et al. (1992), Miaou and Lum (1993a,b) established that accident phenomenon is best expressed by count models such as Poisson regression process. The Poisson process with parameter

μ is given as

$$Pr(n_i|\mu_i) = \frac{\exp(-\mu_i)\mu_i^n}{n_i!} \tag{6}$$

where $Pr(n_i|\mu_i)$ is the probability of n accidents occurring at intersection i and μ_i is the expected number of accidents at intersection i. Now if X_i is a vector of covariates which describes the geometric, traffic control and regulatory characteristics of an intersection i, and β is a vector of estimable coefficients, then μ can be estimated by the equation:

$$\ln \mu_i = \beta X_i \tag{7}$$

This functional form of Poisson process assumes that all the covariates in X_i are sufficient in explaining the mean and variance of crash occurrence. However, this is a limitation of Poisson model as also described by Cox (1983) and Dean and Lawless (1989), that the variance of the data is constrained to be equal to the mean, i.e.:

$$\text{var}(n_i) = E(n_i) = \mu_i \tag{8}$$

In case of rare events like traffic accident variance is significantly greater than mean which is well known as over-dispersion in accident modeling field. Hence, to deal with this problem of extra-variation, researchers proposed the inclusion of a gamma-distributed error term in the parent Poisson model such that:

$$\ln \mu_i = \beta X_i + \varepsilon_i \tag{9}$$

As explained in Section 1, the $\exp(\varepsilon_i)$ is gamma distributed and takes the unobserved heterogeneity of the Poisson mean into account. The unobserved heterogeneity is caused by various factors such as model misspecification, unavailability of important but immeasurable covariates, as well as omission of relevant but measurable factors. These factors can result in over-dispersed count models where a substantial portion of the mean is not captured by included covariates and instead is captured through the gamma distributed error term and the over-dispersion parameter. This dispersion parameter is proportional to model uncertainty and is aptly used in the calculation of overall model fit and standard errors of estimated parameters and confidence intervals. To better understand the importance of the over-dispersion parameter in model assessment is a prime motivation for this study. Whether the over-dispersion parameter is fixed or variable plays an important role in this assessment.

To examine the structure of gamma heterogeneity, it is first necessary to check the mean structure. This includes both the functional form of the mean and the covariates considered. Till date, researchers postulated many functional forms that differ by the type of covariates included and the form of variable transformation. While for intersection crash models almost all studies considered the traffic volume from major and minor road, other studies took the effect of geometric factors in addition to the exposure variables to develop the mean structure. For example, Bauer and Harwood (2000), Vogt and Bared (1998), Oh et al. (2003), Lyon et al. (2003) considered geometric factors in addition to traffic volume. In all of these studies they adopted a logarithmic transformation of the major road and minor road traffic flow along with a linear effect of common geometric factors. A similar functional form has been adopted by interactive highway safety design model (IHSDM) as well. For the purpose of this research different transformation of traffic volume variables were tested. However, the logarithmic transformation resulted in improved model fit. Also, this form ensures that: (1) in the absence of traffic there are no expected crashes and (2) a nonlinear relationship describes the relationship between traffic flow and accidents. Guided by prior research the structure of the mean for this study is

$$\ln \mu_i = \beta_0 + \beta_1 \log(F_1) + \beta_2 \log(F_2) + \beta_j X_j \tag{10}$$

where F_1 and F_2 are respectively major and minor road flow, X_j's are j different types of geometric properties of the intersection and β's are the vector of covariates. As mentioned before, this study compares and contrasts the findings from Miaou and Lord (2003), so some possible relationship between dispersion parameter and the traffic flows are considered here. The following section describes the Bayesian approach of model specification and estimation.

3.1. Bayesian framework of the model

The Bayesian framework to obtain the NB model for accident observation has been described by Hauer (1992) and Heydecker and Wu (2001). While these studies use an empirical Bayesian method to obtain the posterior mean of the crash occurrences, a similar concept has been utilized but using a full Bayes method to find the posterior mean of the coefficient estimates for the mean as well as variance structure of the NB model. By following this procedure, the NB model in Eq. (1) is rewritten as

$$\mu_i = v_i \lambda_i \tag{11}$$

where μ_i is the Poisson parameter but is assumed to be random, $v_i = \exp(\beta X_i)$ and $\lambda_i = \exp(\varepsilon_i)$. The model specifies that λ_i is gamma distributed with a mean of the distribution = 1 and variance = $1/\alpha$, i.e. with shape parameter = scale parameter for the distribution such that

$$\lambda_i \sim \Gamma(\alpha_i, \alpha_i) \tag{12}$$

To test the variance structure, four different functional form of the dispersion parameter α are investigated. The models are given below:

- Model 1 considers a fixed dispersion parameter so that the Eq. (2) can be written as $\log \alpha_i = \eta_1$.
- In model 2, in addition to the fixed parameter major road traffic volumes are considered so that:

$$\log \alpha_i = \eta_1 + \eta_2 (\log F_{1,i} - \text{mean}(\log F_1))$$

- In model 3, the effect of both major road and minor road traffic volumes are taken into account:

$$\log \alpha_i = \eta_1 + \eta_2 (\log F_{1,i} - \text{mean}(\log F_1)) + \eta_2(\log F_{2,i} - \text{mean}(\log F_2))$$

- In model 4, in addition to the model 3 parameters, a major and minor road volume interaction term as proposed by Miaou and Lord (2003) is considered and given as

$$\log \alpha_i = \eta_1 + \eta_2(\log F_{1,i} - \text{mean}(\log F_1)) + \eta_3(\log F_{2,i} - \text{mean}(\log F_2)) + \eta_4(\log(F_{2,i}/F_{1,i}) - \text{mean}(\log(F_{2,i}/F_1)))$$

where F_1 and F_2 are major and minor road traffic volumes. In models 2, 3 and 4 the term mean(log F) is addititive to capture the absolute deviation of the traffic volume of a particular site from the mean traffic volumes of all sites considered in the study.

3.2. Goodness-of-fit

Much has been written on MCMC convergence issues via MCMC sampling techniques. In this section first the model convergence criteria used in this study is described followed by some of the goodness of fit measures used for model selection. A total of five different measures of goodness of fit are computed to select the most parsimonious model and these are: mean deviance, chi-square measure, sum of model deviance, R^2-like measure of fit and deviance information criteria. The Bayesian modeling was conducted and goodness-of-fit estimates was obtained using the WinBUGS modeling software.

3.2.1. Model convergence

Model convergence is concerned with the time of producing sequence of draws from the posterior distribution. This is particularly important because (1) it ensures that the posterior has been 'found' and (2) it indicates when sampling of parameters should begin. A common methodology to check the convergence is by tracking the Gelman-Rubin convergence statistic as modified by Brooks and Gelman (1998). A Gelman–Rubin statistic under 1.2 indicates approximate convergence and it is used to assess when convergence occurred.

3.2.2. Chi-square measure

Another measure of goodness of fit is the χ^2-statistic distributed with the degrees of freedom equal to the difference in the numbers of coefficients in the restricted and unrestricted models. The χ^2-statistic that is based on standardized residuals, is given as

$$\sum \left[\frac{(n_i - \mu_i)^2}{V(n_i)} \right] \tag{13}$$

3.2.3. Sum of model deviance

Theoretically, if the sum of model deviance, or G^2, is equal to zero, then the model is regarded to have a perfect fit. This is a theoretical lower bound because the observed values are integer values and the predicted values are continuous (Washington et al., 2003). The G^2 statistic is given by

$$G^2 = 2\sum_{i=1}^{n} n_1 \ln\left(\frac{n_i}{\mu_i}\right) \tag{14}$$

The model with the lowest G^2 value is therefore regarded as the model with the better fit.

3.2.4. R^2-like measure of fit

As a result of heteroscedasticity in the regression and the non-linearity of the conditional mean the R^2 value used in ordinary least squares linear regression is not available. Based on standardized residuals a similar statistic can be calculated where the residual sum of squares forms the numerator and the total sum of squares forms the denominator (Washington et al., 2003):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[(n_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i}]^2}{\sum_{i=1}^{n}[(n_i - \bar{n})/\sqrt{\bar{n}}]^2} \tag{15}$$

In Eqs. (13)–(15) n_i is the observed number and μ_i is the expected number of accidents occurring at intersection i, $\hat{\mu}_i$ is the predicted expected value and $\bar{n}$ is the average observed number of accident.

3.2.5. Deviance information criteria

Another criterion for assessing model goodness of fit is deviance information criteria or DIC as proposed by Spiegelhalter et al. (2002). DIC is a generalized and the Bayesian version of Akaike's information criterion (AIC), and is a penalized fit measure (larger parameter models are penalized). The DIC for model m are calculated as follows:

$$\text{DIC}_\text{m} = \bar{\Delta}_\text{m} + d_\text{em} \tag{16}$$

where $\bar{\Delta}_\text{m} = -2\log L_\text{m}$ also known as "Dbar", represents the posterior mean of the deviance of un-standardized model while L_m is the mean of the model log likelihood; d_em, also known as "pD" can be calculated as follows $d_\text{em} = \bar{\Delta}_\text{m} - \Delta(y|\bar{\theta})$ and represents the penalty for the number of effective model parameters. The term $\Delta(y|\bar{\theta})$ is also known as "Dhat", which is a point estimate of deviance obtained by substituting in the posterior means $\bar{\theta}$ of θ. The model that provides the best short-term predictions will have the lowest DIC value. It is also important to keep in mind that DIC values can only be compared between models that were developed using the same set of data although the models need not be nested.

4. Description of the data

Accident data for this study were obtained from rural intersections in 38 counties in the state of Georgia for the years 1996 and 1997. Road characteristic (RC) files, aerial photographs, and geographic information system (GIS) roadmaps were used to find various geometric characteristics of the intersections. Digital Orthophotography Quarter-Quadrangles (DOQQs) aerial photos were used from 1994 and 2000 to extract information regarding intersection angle and degree of horizontal curvature of selected intersections by overlapping with GIS roadmaps. A

Table 1
Variable list used in this study

Variables	Definition
Dependent variables	
TOTACC	Number of total crashes
Independent variables	
AADTMAJ (F_1)	AADT on major road
AADTMIN (F_2)	AADT on minor road
MDWDMAJ	Median width on major road in feet
MDWDMIN	Median width on minor road in feet
SHLWDMAJ	Shoulder width on major road in feet
SHLWDMIN	Shoulder width on minor road in feet
SIGNAL	Intersection type (0 if non-signalized intersection, 1 if signalized intersection)
RTLMAJ	Right-turn lane indicator (1 if at least one right-turn lane on the major road, 0 otherwise)
LTLMAJ	Left-turn lane indicator (1 if at least one left-turn lane on the major road, 0 otherwise)
RTLMIN	Right-turn lane indicator (1 if at least one right-turn lane on the minor road, 0 otherwise)
LTLMIN	Left-turn lane indicator (1 if at least one left-turn lane on the minor road, 0 otherwise)
HZRATMAJ	Roadside hazard rating on major road (from 1, least hazardous case, to 7, most hazardous case)
HZRATMIN	Roadside hazard rating on minor road
DRWYMAJ	Number of driveways on major road within 250 ft of the intersection center
DRWYMIN	Number of driveways on minor road within 250 ft of the intersection center
LIGHTMAJ	Lighting indicator (1 if lighting exists on the major road, 0 otherwise)
LIGHTMIN	Lighting indicator (1 if lighting exists on the minor road, 0 otherwise)
TERNMAJ	Terrain on major road (0 = flat, 1 = rolling, 2 = mountainous)
TERNMIN	Terrain on minor road (0 = flat, 1 = rolling, 2 = mountainous)
SPDLIMAJ	Speed limit on major road in mph
SPDLIMIN	Speed limit on minor road in mph
SDMAJ	Sight distance on major road in feet
SDMIN	Sight distance on minor road in feet
VIMAJ/VIMIN	Sum of absolute change of grade in percent per hundred feet for each curve on major road or minor road within 250 ft of the intersection center, divided by the number of such curves
HAU	Intersection angle variable in degrees

description of various independent variables used in the analysis is provided in Table 1.

The data included 165 rural intersections on two-lane roads: 51 were signalized and 114 non-signalized. The sample included a total of 837 accidents (345 at non-signalized and 492 at signalized intersections). Intersection crashes were defined as any accident that occurred at the intersection or occurred within 250 ft (76 m) from the intersection on either the major or the minor road. For the purpose of this study total crashes rather than crash outcomes are used to find the overall effect.

5. Results and discussion

For the purpose of analysis all the eight models are estimated using three chains taken to 100,000 iterations. The convergence in all of these models was obtained after 4000 iterations and this convergence is assured by checking the Gelman–Rubin statistics as mentioned in Section 3.2.1. Consequently, the samples for posterior analysis have been taken after 4000 burn-ins. The priors for regression coefficients of mean as well as variance structure are taken as $N(0,0.001)$ in WINBUGS modeling software where 0.001 is the precision of the normal distribution which indicates that the variance is high, i.e. they are non-informative or flat priors.

Table 2
Estimation results for total crashes using classical maximum likelihood method (negative binomial regression model)

Variables	Estimated coefficient	t-Statistic	p-Value
Constant	−4.4552	−6.490	0.0000
AADTMAJ (F_1)	0.4356	5.974	0.0000
AADTMIN (F_2)	0.3196	3.692	0.0002
MDWDMAJ	−0.0757	−3.006	0.0026
RTLMAJ	0.7408	3.330	0.0009
DRWYMAJ	0.1168	3.015	0.0026
LIGHTMAJ	−0.4785	−2.326	0.0200
A (dispersion parameter)	0.4139	5.064	0.0000
Number of observations	165		
Log-likelihood at zero	−464.54		
Log-likelihood at convergence	−394.52		
ρ^2	0.15		

The results from this analysis are compared with a likelihood-based estimation done by Kim et al. (2006) using the same dataset but following a classical estimation approach and is given in Table 2. The results from the four models where traffic volume and geometric design factors are considered in the mean function are given in Table 3.

From the model findings (Table 3) it is clear that over-dispersion is significant with α significantly greater than 1. In terms of the coefficient estimates, the results from four modeling strategies did not have much variation which shows a similar trend as that of previous researches (Miaou and Lord, 2003; Maher and Summersgill, 1996). However, there is a difference in parameter estimates between the best fitted model (i.e. Model 1) from Bayesian analysis and that from maximum likelihood estimate, but it is less than 2%. This could be attributed to the small size of data that violate the implicit assumption of asymptotic normality of the maximum likelihood estimate of regression coefficients.

A closer comparison between these two models based on posterior credible intervals and model parameters shows that in all the four models the coefficient estimates for the constant as well as other predictors are very similar to that of classical estimate except for beta5, which identifies the effect of major road right turn lane on crash occurrence. The estimation of this variable in the Bayesian framework is about 0.55, whereas the same in maximum likelihood estimate is 0.74, which clearly indicates an enhanced effect of the right turn lane. To check the variance structure and the parameterization of α, the significance and the estimations of η_1, η_2, η_3 and η_4 from the four models are compared. The hypothesis that α is a function of flow is very weakly supported as η_2, η_3 and η_4s are consistently found not

Table 3
Bayesian estimation results from four different models for total crashes

Variables	Model 1			Model 2			Model 3			Model 4		
	Mean	S.D.	Median (2.5–97.5%)	Mean	S.D.	Median (2.5–97.5%)	Mean	S.D.	Median (2.5–97.5%)	Mean	S.D.	Median (2.5–97.5%)
Constant	−4.941	0.6002	−4.893 (−6.234, −3.798)	−4.819	0.6719	−4.721 (−6.453, −3.696)	−5.028	0.6453	−4.277 (−5.618, −2.928)	−5.047	0.7524	−5.028 (−6.593, −3.602)
AADTMAJ (F_1)	0.4309	0.0652	0.4279 (0.2962, 0.552)	0.3899	0.0845	0.3973 (0.2326, 0.5500)	0.4163	0.0978	0.4158 (0.2248, 0.5835)	0.4316	0.0851	0.4293 (0.2701, 0.6113)
AADTMIN (F_2)	0.2909	0.0764	0.2955 (0.1253, 0.4435)	0.3213	0.0827	0.3182 (0.1517, 0.4843)	0.3193	0.0853	0.3154 (0.158, 0.4842)	0.304	0.0614	0.3011 (0.1894, 0.4295)
MDWDMAJ	−0.0768	0.0229	−0.0767 (−0.1214, −0.0314)	−0.0786	0.0248	−0.0788 (−0.1265, −0.0300)	−0.0783	0.0237	−0.0775 (−0.1245, −0.0309)	−0.0755	0.0236	−0.0750 (−0.1221, −0.0282)
RTLMAJ	0.5803	0.1254	0.5773 (0.3204, 0.823)	0.5775	0.1357	0.5791 (0.3052, 0.8473)	0.5612	0.1237	0.5615 (0.3076, 0.8292)	0.5536	0.1293	0.5557 (0.2992, 0.8053)
DRWYMAJ	0.1195	0.0403	0.1194 (0.0417, 0.1194)	0.1242	0.0399	0.1236 (0.0447, 0.2026)	0.1298	0.0399	0.1274 (0.0488, 0.2064)	0.1276	0.0398	0.1273 (0.0492, 0.2075)
LIGHTMAJ	−0.4421	0.1982	−0.4528 (−0.8373, −0.0652)	−0.4274	0.2065	−0.4314 (−0.8254, −0.0217)	−0.4526	0.2051	−0.4612 (−0.8432, −0.0578)	−0.4624	0.2023	−0.4685 (−0.842, −0.0487)
Deviance	159.4	18.24	158.7 (125.5, 196.9)	164.9	19.55	164.1 (128.9, 205.5)	161.7	18.8	161.0 (126.8, 200.3)	161.5	18.93	160.8 (126.4, 200.5)
Chi-square	76.02	14.46	74.49 (52.26, 108.6)	83.34	19.96	80.01 (54.43, 132.8)	79.86	17.51	77.44 (53.3,120.7)	79.52	18.23	76.97 (52.78, 121.1)
G^2	178.1	60.54	177.7 (60.59, 298.0)	182.7	61.04	182.2 (63.87, 303.3)	180.2	60.6	179.8 (62.74, 300.0)	179.4	60.83	179.1 (61.32, 299.3)
R^2	0.1947	0.0097	0.1947 (0.1757, 0.2138)	0.1937	0.0097	0.1937 (0.1745, 0.213)	0.1939	0.0098	0.1939 (0.1747, 0.213)	0.1939	0.0098	0.1939 (0.1748, 0.213)
η_1	0.8894	0.198	0.8486 (0.4829, 1.249)	0.9602	0.2536	0.9389 (0.5225, 1.52)	0.9521	0.2545	0.9327 (0.5069, 1.518)	7.359	9.486	4.191 (−6.454, 27.22)
η_2				−0.1251	0.3093	−0.1005 (−0.8089, 0.4116)	−0.1846	0.3271	−0.1635 (−0.9063, 0.3914)	−0.8548	1.173	−0.5196 (−3.4, 0.838)
η_3							0.1506	0.2257	0.1481 (−0.2889, 0.6004)	0.936	1.317	0.5456 (−1.013, −3.768)
η_4										−6.792	11.51	2.967 (−30.83, 9.936)
DIC	720.988			724.664			725.385			723.411		

to be significant. In comparison the model 1 where α is a fixed dispersion parameter showed better overall model fit. Although the effect of the parameter estimates for η_1, η_2, η_3 and η_4 from models 2, 3 and 4 are in the same direction with what Miaou and Lord (2003) obtained, the parameters η_2, η_3 and η_4 are not significantly different from zero. What this means is that even without the flow-varying part in parameterization of α, it is possible to capture the variation and resulted in a better model fit. This supports a comment from Winkelmann (2003), who did extensive research on count data and explained in his book that it is very hard to justify that some variables affect the variance but not the mean. Hence, the finding supports the fact that if the mean function is well defined, a simple variance as a function of mean and a fixed dispersion parameter is sufficient to capture the unobserved heterogeneity for traffic crash occurrences in rural intersections. A comparison of several goodness-of-fit statistics such as mean deviance, chi-square statistic, R^2-like measure of fit also supported the appropriateness of the fixed dispersion parameter model which can be seen in Table 3. Another important goodness of fit statistic in the case of Bayesian models is the DIC, where a lower value suggests a better fit. From this study it was observed that the DIC value of model 1 is 720.9 and that of models 2, 3 and 4 are 724.6, 725.3 and 723.4, respectively. Hence, in terms of significance of coefficient as well as goodness-of-fit measures a flow dependent dispersion function did not result in better model.

While the results seemed to be plausible, a different finding than previous studies triggered a greater investigation of the process under consideration. As a first step, the mean functions from these two studies are compared and as mentioned before they are not same in terms of functional form as well as included covariates. Miaou and Lord (2003) considered mainly the effect of major and minor road traffic flow with various functional forms all of which was supported by previous research findings. However, in the present study the effect of various geometric factors are considered in addition to the major and minor road traffic volume and some of the included variables are found to be significant. So under this scenario of two different mean functions, a direct comparison of the variance function is not possible. Also it is important to remember that the unexplained part of the mean structure goes into the variance structure and thus influences the over-dispersion. This means that in the presence of a well defined mean function the structure of the variance function would vanish. Hence, any structural or covariate changes in the mean function would affect the structure and coefficient estimates of the variance function.

To test the effect of a different mean function another estimation has been undertaken. The explanatory covariates for this second estimation are restricted to only major and minor road AADT, both in mean as well as variance structure. All the four different modeling strategies as explained in Section 3.1 have been adopted in this case too.

The results from these models are given in Table 4. The table shows a dramatic change in findings from the previous results shown in Table 3. In the case of a mean function that deals mainly with traffic flows, there is a distinct structure present in the variance function. In other words, while major and minor road traffic

Table 4
Bayesian estimation results from four different models for total crashes including only major and minor road traffic flows

Variables	Model 1			Model 2			Model 3			Model 4		
	Mean	S.D.	Median (2.5–97.5%)	Mean	S.D.	Median (2.5–97.5%)	Mean	S.D.	Median (2.5–97.5%)	Mean	S.D.	Median (2.5–97.5%)
Constant	−5.772	0.6148	−5.818 (−6.744, −4.533)	−5.539	0.6672	−5.547 (−6.724, −4.241)	−5.952	0.97	−5.841 (−8.025, −4.38)	−6.058	0.7119	−6.068 (−7.544, −4.793)
AADTMAJ (F_1)	0.5219	0.0819	0.5227 (0.3648, 0.6795)	0.4889	0.0729	0.4841 (0.3559, 0.6494)	0.509	0.1086	0.5083 (0.3391, 0.7658)	0.524	0.1013	0.5115 (0.3319, 0.7137)
AADTMIN (F_2)	0.3424	0.0698	0.3424 (0.2174, 0.4885)	0.3451	0.0843	0.3479 (0.1519, 0.4870)	0.3823	0.0776	0.3725 (0.2552, 0.5521)	0.3811	0.0767	0.3894 (0.2281, 0.5141)
Deviance	154.1	17.93	153.5 (120.8, 1191.1)	162.7	19.72	161.9 (126.4, 203.6)	163.8	19.75	163.0 (127.3, 204.6)	160.7	18.64	160.0 (126.2, 199. 2)
Chi-square	65.23	13.54	63.61 (44.76, 94.62)	81.78	24.8	76.56 (49.25, 144.0)	82.01	24.63	76.86 (49.51,143.5)	77.66	22.68	75.1 (51.84, 118.8)
G^2	171.9	60.46	171.5 (54.47, 291.6)	180.9	61.24	180.4 (62.07, 302.1)	180.8	60.96	180.5 (62.49, 301.7)	177. 9	60.56	177.6 (59. 94, 297.9)
R^2	0.1931	0.0097	0.1931 (0.174 0.2122)	0.1929	0.0098	0.1929 (0.1736, 0.2121)	0.1924	0.0098	0.1924 (0.1733, 0.2116)	0.1931	0.0097	0.1931 (0.1741, 0.2122)
η_1	0.5644	0.1659	0.5608 (0.2491, 0.8986)	1.275	0.3823	0.7312 (0.326, 1.312)	1.245	0.3715	1.23 (0. 5515, 2.052)	2.782	0.7805	2.726 (1.316, 4.467)
η_2				−1.06E−4	0.0005	−0.3272 (−0.9701, 0.1066)	−1.11E−4	0.0005	−0.001 (−0.002, −0.0002)	−2.66E−4	8.29E−4	−2.62E−4 (−0.004, −0.001)
η_3							2.64E−5	8.30E−5	0.0003 (−0.001, 0.001)	3.48E−4	1.43E−4	3.48E−4 (0.0005, 0.006)
η_4										−3.027	1.175	−3.021 (−5.341, −0.7244)
DIC	722.638			722.95			724.399			719.869		

volumes solely explain the mean structure, the variance is no longer a function of the mean and a constant dispersion function, but has a specific structural form. Also, among the four models, the model where variance is structured as a function of the major, minor and interaction with major and minor road flow yielded the greatest explanatory power. These findings agree with findings from Miaou and Lord (2003) based on the analysis of data from urban intersections of Toronto. The goodness-of-fit of model 4 is also improved compared to the rest of the models in Table 4. Interesting enough, the DIC obtained from this model (model 4) is almost the same as that of the Model 1 in Table 3 with additional covariates in the mean structure and simple variance structure. This result corroborates the findings from Miaou and Lord (2003), and emphasizes the fact that variance structure is dependent on the mean structure. If a limited number of covariates are considered in the model, the dispersion parameter is indeed structured, resulting in significant mis-specification if α is assumed to be fixed (Miaou and Lord, 2003). The result also suggests that in case of a well-specified mean structure (proper form and no significant omitted variables) the variance is proportional to the expected crash count. However, the question still remains as to which functional form is most appropriate: a simple mean structure with flow-dependent variance structure (to avoid additional data collection) or the inclusion of additional geometric variables. To address the question, it is important to compare the results and inferences from these two specifications and to consider theoretical appeal. As mentioned previously, the structure of the dispersion parameter affects confidence interval estimation, and a comparison between the results obtained from these two mean structures reveals similar outcomes. The estimated confidence intervals for major and minor road AADT in Table 3 are narrower than the confidence intervals of the same variables in Table 4, which suggests that the addition of variables is beneficial on statistical grounds. In addition, theoretically it is more appealing to include geometric variables (and perhaps environmental and traffic) in accounting for between-site variation rather than allowing for additional random variation.

The results described here are empirical and not theoretical. As such, it may be possible that both datasets examined in this study and Miaou and Lord's study performed similarly by chance. It also leaves the possibility that the variance could in fact be explained through explanatory variables not included in the mean function with other datasets, although the likelihood of this result is now diminished. If this were the case, however, it would suggest that the mean and variance functions are related to mutually exclusive sets of predictors—one set helping to explain expected crash counts and the other set helping to explain unaccounted for variation across crash sites or locations. The evidence for this outcome based on this study (which in turn builds on Miaou and Lord), however, is lacking.

If the results described in this paper are generalizeable – i.e. corroborated by numerous researchers – then a couple of important implications arise. First, modeling the variance as a function of covariates may be a reliable way to provide modeling feedback as to model mis-specification. The guidance would be that significant variables in the variance function had heretofore been mis-specified in the mean function. Second, the assumption of

an unstructured variance function is reasonable, leaving the standard negative binomial density as a reasonable approximation of crash distributions across sites. Finally, it should be remembered that consideration of a more complex model – i.e. a dispersion function rather than a fixed dispersion parameter – involves additional model parameters and thus should be compared to the 'base' model using penalized fit criterion. In other words, we should expect relatively significant improvements in fit in order to justify the additional model complexity and parameters.

6. Conclusions and recommendations

While the study was motivated to corroborate the findings of Miaou and Lord (2003) regarding the variance structure in overdispersed crash models, during the course of research some interesting findings emerged. The research re-emphasized the basic theory of count data models and focused on how the unobserved heterogeneity can be effectively considered in crash prediction models through over-dispersion, with the negative binomial density particularly well-suited. Using crash data from two-lane rural intersections of Georgia, this research emphasized the importance of a well-defined mean structure that accommodates all the relevant covariates in explaining crash occurrence with a simple variance structure as a function of mean and a constant dispersion parameter. However, in the presence of small number of explanatory variables, researchers are cautioned about using the standard NB count models where the dispersion parameter does not vary among different sites—there appears to be a tradeoff between these two functions with an overlapping set of predictors. Hence, the main finding of this study is to closely judge the functional form of crash prediction models in the light of available data in hand and to emphasize that the mean function must be correctly specified (functional form correct and no significant omitted variables) in order to reduce omitted variable bias in the variance function.

An important extension of this research is the investigation of crash data for other kinds of rural as well as urban intersections and/or road segments to obtain further corroboration. In any case, this research contributes to the existing state of the knowledge regarding the nature of over-dispersion in motor vehicle crash models and how over-dispersion should ideally be modeled.

References

Bauer, K.M, Harwood, D.W., 2000. Statistical Models of At-Grade Intersection Accidents—Addendum. Federal Highway Administration Report-RD-99-094, Kansas City.

Belanger, C., 1994. Estimation of safety of four-legged unsignalized intersections. Transport. Res. Rec. 1467, 23–29.

Bonneson, J.A., McCoy, P.T., 1993. Estimation of safety at two-way stop-controlled intersections on rural highways. Transport. Res. Rec. 1401, 83–89.

Brooks, S.P., Gelman, A., 1998. Alternative methods for monitoring convergence of iterative simulations. J. Computat. Graph. Stat. 7, 434–455.

Casella, G., George, E.I., 1992. Explaining the Gibbs Sampler. Am. Statistician 46 (3), 167–174.

Chapman, R.A., 1972. The concept of exposure. Accid. Anal. Prev. 5, 95–110.

Congdon, P., 2003. Bayesian Statistical Modeling. John Wiley and Sons, Inc., New York.

Cox, D.R., 1983. Some remarks on overdispersion. Biometrika 70 (1), 269–274.

Dean, C., Lawless, J.F., 1989. Tests for detecting overdispersion in Poisson regression models. J. Am. Stat. Assoc. 84 (406), 467–472.

Dey, D., Gelfand, A., Peng, F., 1997. Overdispersed generalized linear models. J. Stat. Plan. Inference 64, 93–107.

Hauer, E., Ng, J.C.N., Lovell, J., 1988. Estimation of safety at signalized intersections. Transport. Res. Rec. 1185, 48–61.

Hauer, E., 1992. Empirical Bayes approach to the estimation of unsafety: the multivariate regression approach. Accid. Anal. Prev. 24, 456–478.

Heydecker, B.G., Wu, J., 2001. Identification of sites for road accident remedial works by Bayesian statistical methods: an example of uncertain inference. Adv. Eng. Softw. 32, 859–869.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. Transport. Plan. Technol. 15 (1), 41–58.

Kim, D., Washington, S., Oh, J., 2006. Modeling crash outcomes: New insights into the effects of covariates on crashes at rural intersections. ASCE, J. Transport. Eng. 132 (4), 282–292.

Lord, D., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transport. Res. Rec. 1717, 102–108.

Lyon, C., Oh, J., Persaud, B., Washington, S., Bared, J., 2003. Empirical investigation of interactive highway safety design model accident prediction algorithm: rural intersections. Transport. Res. Rec. 1840, 78–86.

Maher, J., Summersgill, I., 1996. A comprehensive methodology for the fitting of predictive accident models. Accid. Anal. Prev. 28 (3), 281–296.

McCullagh, P., Nelder, J.A., 1983. Generalized Linear Models. Chapman and Hall, London.

Miaou, S.P., Hu, P.S., Wright, T., Rathi, A.K., Davis, S.C., 1992. Relationships between truck accidents and highway geometric design: a Poisson regression approach. In: Presented at 71st Annual Meeting of the Transportation Research Board, Washington, DC.

Miaou, S.P., Lum, H., 1993a. Modeling vehicle accidents and highway geometric design relationships. Accid. Anal. Prev. 25 (6), 689–709.

Miaou, S.P., Lum, H.A., 1993b. Statistical evaluation of the effects of highway geometric design on truck accident involvements. Transport. Res. Rec. 1407, 11–23.

Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transport. Res. Rec. 1840, 31–40.

Norris, J.R., 1997. Markov Chains. Cambridge University Press, Cambridge.

Oh, J., Lyon, C., Washington, S., Persaud, B., Bared, J., 2003. Validation of the FHWA crash models for rural intersections: lessons learned. Transport. Res. Rec. 1840, 41–49.

Persaud, B., Nguyen, T., 1998. Disaggregate safety performance models for signalized intersections on Ontario provincial roads. Transport. Res. Rec. 1635, 113–120.

Persaud, B., Lord, D., Palmisano, J., 2002. Calibration and transferability of accident prediction models for urban intersections. Transport. Res. Rec. 1784, 57–64.

Robert, C., Casella, G., 1997. Monte Carlo Statistical Methods. Springer, New York.

Satterthwaite, S.P., 1981. A survey of research into relationships between traffic accidents and traffic volumes. Supplementary Report 692. Crowthor, Berks, U.K. Transportation Research Laboratory.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van-der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. 64 (B), 583–640.

Vogt, A., Bared, J., 1998. Accident prediction models for two-lane rural roads: segments and intersections. FHWA-RD-98-133, Washington, D.C.

Walsh, 2004. Bruce Walsh's Homepage. University of Arizona, Department of Life Sciences, Tucson. Accessed July 22, 2004: http://nitro.biosci.arizona.edu/.

Wang, Y., Nihan, N.L., 2004. Estimating the risk of collision between c bicycles and motor vehicles at signalized intersections. Accid. Anal. Prev. 36 (3), 313–321.

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

Washington, S., Congdon, P., Karlaftis, M., Mannering, F., 2005. Bayesian multinomial logit models: exploratory assessment of transportation applications. In: Presented in Transportation Research Board Annual Conference, TRB, Washington, D.C.

Winkelmann, R., 2003. Econometric Analysis of Count Data. Springer-Verlag, Berlin.