



Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis

Yajie Zou¹, Yunlong Zhang*, Dominique Lord²

Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, United States

ARTICLE INFO

Article history:

Received 17 February 2012

Received in revised form 23 July 2012

Accepted 5 August 2012

Keywords:

Negative binomial
Finite mixture model
Dispersion
Weight parameter
Crash data

ABSTRACT

Recently, a finite mixture of negative binomial (NB) regression models has been proposed to address the unobserved heterogeneity problem in vehicle crash data. This approach can provide useful information about features of the population under study. For a standard finite mixture of regression models, previous studies have used a fixed weight parameter that is applied to the entire dataset. However, various studies suggest modeling the weight parameter as a function of the explanatory variables in the data. The objective of this study is to investigate the differences on the modeling and fitting results between the two-component finite mixture of NB regression models with fixed weight parameters (FMNB-2) and the two-component finite mixture of NB regression models with varying weight parameters (GFMNB-2), and compare the group classification from both models. To accomplish the objective of this study, the FMNB-2 and GFMNB-2 models are applied to two crash datasets. The important findings can be summarized as follows: first, the GFMNB-2 models can provide more reasonable classification results, as well as better statistical fitting performance than the FMNB-2 models; second, the GFMNB-2 models can be used to better reveal the source of dispersion observed in the crash data than the FMNB-2 models. Therefore, it is concluded that in many cases the GFMNB-2 models may be a better alternative to the FMNB-2 models for explaining the heterogeneity and the nature of the dispersion in the crash data.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

As documented in the literature (Lord et al., 2005), over-dispersion in vehicle crash data arises from the nature of the crash process. This process dictates that the over-dispersion is the result of Bernoulli trials with unequal probability of independent events (this is also known as Poisson trials). To overcome the over-dispersion problem, a variety of ways have been proposed within the negative binomial (NB) modeling framework (Poch and Mannering, 1996; Hauer, 2001; Miaou and Lord, 2003; Lord and Mannering, 2010). More recently, several researchers have proposed different models for analyzing over-dispersed data. Malyshkina et al. (2009) found out that the two-state Markov switching models can perform better than the traditional NB model in terms of goodness-of-fit and the capability of capturing unobserved heterogeneity. Anastasopoulos and Mannering (2009) applied random-parameter count models to vehicle crash data by employing a normal error term in the coefficients to allow them

to vary. A negative binomial-Lindley (NB-L) model was recently introduced by Lord and Geedipally (2011) and Geedipally et al. (2012). They showed that the NB-L GLM works much better than the traditional NB model when datasets contain a large number of zeros or datasets are highly dispersed. Besides these models, a finite mixture model has been proposed as an alternative to address the over-dispersion problem (Park and Lord, 2009; Park et al., 2010). This approach used a flexible semi-parametric model that allows one to capture unobserved heterogeneity through a small number of simple regression models such as Poisson or NB regression models. According to Park et al. (2010), the finite mixture model has two advantages over the traditional NB regression model. First, it can effectively account for heterogeneity without imposing strong distributional assumptions on the mixing variable. Second, while most traditional models constrain the coefficients to be fixed across observations, the finite mixture model allows the vector of regression coefficients to vary from component to component.

Given the potential advantages of finite mixture models in describing the heterogeneity in crash data, there is a need to conduct further research on improving their performance. It is useful to understand what characteristics make a particular site more prone to fall into one or the other sub-population (i.e., one sub-population consists of accident-prone sites and the other contains low-risk sites). For a standard finite mixture of regression models, previous studies (Park and Lord, 2009; Park et al.,

* Corresponding author. Tel.: +1 979 845 9902; fax: +1 979 845 6481.

E-mail addresses: yajiezou@tamu.edu (Y. Zou), yzhang@civil.tamu.edu (Y. Zhang), d-lord@tamu.edu (D. Lord).

¹ Tel.: +1 979 595 5985; fax: +1 979 845 6481.

² Tel.: +1 979 458 3949; fax: +1 979 845 6481.

2010; El-Basyouny and Sayed, 2010; Chang and Kim, 2012) use a fixed weight parameter that is applied to the entire dataset. However, various studies (Frühwirth-Schnatter, 2006; Frühwirth-Schnatter et al., 2011) suggest modeling the weight parameter as a function of the explanatory variables in the data. This is sensible whenever the span of the explanatory variables is different between different groups (in other words, sub-populations). For example, Frühwirth-Schnatter and Kaufmann (2008) modeled the probabilistic structure of the group indicators and found that the covariate helped predict group membership. In this respect, the objective of this study is to investigate the differences on the modeling and fitting results between the mixture models with fixed weight parameters and the mixture models with varying weight parameters, and compare the group classification from both models. Specifically, we mainly examine the modeling results and group classification from the two-component finite mixture of negative binomial regression models (termed as the FMNB-2 model)³, both for fixed and varying weights. This is because the FMNB-2 model was considered more useful than the finite mixture of Poisson regression models (termed as the FMP-g model; g is the number of components). Since the FMP-g models cannot handle the extra-variation within components, they often produce too many components. To accomplish the objective of this study, the FMNB-2 models with both fixed and varying weight parameters are applied to two crash datasets (one intersection dataset and one roadway segment dataset).

2. Background

This section describes the characteristics of the finite mixture of NB regression models and the parameter estimation method.

2.1. Finite mixture of NB regression models

A finite mixture model allows for extremely flexible modeling of heterogeneous data because it incorporates a combination of discrete and continuous representation of population heterogeneity. If the empirical distribution of data $\mathbf{y} = (y_1, y_2, \dots, y_n)$ exhibits multimodality, skewness or excess kurtosis, we can assume that the data are independent realizations of a random variable Y from a finite mixture distribution (Frühwirth-Schnatter, 2006; Zou and Zhang, 2011). In this paper, it is assumed that the data are independent and identically distributed realizations from a random variable that follows a g -component mixture distribution. The density can be formulated as follows:

$$f_Y(\mathbf{y}|\boldsymbol{\Theta}) = \sum_{j=1}^g w_j f_j(\mathbf{y}|\boldsymbol{\theta}_j) \quad (1)$$

where w_j is the weight of component j , with $w_j > 0$, and $\sum_{j=1}^g w_j = 1$,

$\boldsymbol{\theta}_j$ are vectors of parameters for the component j , $f_j(\mathbf{y}|\boldsymbol{\theta}_j)$ is the component density for component j ($j = 1, 2, \dots, g$), g is the number of components and $\boldsymbol{\Theta} = ((w_1, \dots, w_g), \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_g)$ is the vector of all unknown parameters. In this study, it is assumed that all components are NB distributions.

For the g -component finite mixture of negative binomial regression models (termed as the FMNB-g model), it is assumed that the

marginal distribution of y_i follows a mixture of NB distributions,

$$f_Y(y_i|\mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{j=1}^g w_j \text{NB}(\mu_{ij}, \phi_j) = \sum_{j=1}^g w_j \left[\frac{\Gamma(y_i + \phi_j)}{\Gamma(y_i + 1)\Gamma(\phi_j)} \left(\frac{\mu_{ij}}{\mu_{ij} + \phi_j} \right)^{y_i} \left(\frac{\phi_j}{\mu_{ij} + \phi_j} \right)^{\phi_j} \right] \quad (2)$$

$$E(y_i|\mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{j=1}^g \mu_{ij} w_j \quad (3)$$

$$\text{Var}(y_i|\mathbf{x}_i, \boldsymbol{\Theta}) = E(y_i|\mathbf{x}_i, \boldsymbol{\Theta}) + \left(\sum_{j=1}^g w_j \mu_{ij}^2 (1 + 1/\phi_j) - E(y_i|\mathbf{x}_i, \boldsymbol{\Theta})^2 \right) \quad (4)$$

where $\mu_{ij} = \exp(\mathbf{x}_i \boldsymbol{\beta}_j)$, μ_{ij} is the mean rate of component j , \mathbf{x}_i is a vector of covariates, $\boldsymbol{\beta}_j$ is a vector of the regression coefficients for component j ; $\boldsymbol{\Theta} = \{(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_g), (\phi_1, \dots, \phi_g), \mathbf{w}\}$ and $i = 1, 2, \dots, n$. For the FMNB-g model, the variance of y_i is always greater than the mean. When ϕ_j in each component goes to infinity, the FMNB-g model is reduced to the FMP-g model. Thus, the FMNB-g models allow for additional heterogeneity within components not captured by the explanatory variables. If additional heterogeneity is present within components, the FMP-g model would be a misspecification.

In the previous studies (Park and Lord, 2009; Park et al., 2010), the weight parameter w_j for the FMNB-g models was treated as a constant variable to simplify the estimation process. The constant weight model can be extended to a more generalized model by parameterizing the weight parameter as a function of covariates (see Guo and Trivedi, 2002; Frühwirth-Schnatter, 2006). Wang et al. (1998) used a finite mixed Poisson regression model that incorporates covariates in the mixing probabilities to explain overdispersion with respect to a Poisson regression model. In this paper, the FMNB-g model with a varying weight parameter w_{ij} is considered (henceforth defined as generalized FMNB-g model or GFMNB-g model). The GFMNB-g model has the same probability density function shown in Eq. (2) and estimates the number of crashes of each site, similar to the FMNB-g model. However, instead of estimating a fixed weight parameter, the varying weight parameter w_{ij} is modeled as a function of covariates:

$$\frac{w_{ij}}{w_{ig}} = \exp(\boldsymbol{\alpha}_j^T \mathbf{z}_i) \quad (5)$$

where \mathbf{z}_i is a vector of secondary covariates that might help classify the sites (not necessarily the same as the covariates in estimating the mean function μ_{ij}), $\boldsymbol{\alpha}_j$ is a vector of regression coefficients for component j , and $\boldsymbol{\alpha}_g = \mathbf{0}$.

With Eq. (5), the new parameterization allows each observation to have a different weight that is dependent on the sites' attributes (i.e., covariates), similar to the application of the varying dispersion parameter for the standard negative binomial model (see Lord and Park, 2008). This new parameterization can yield important insights into the factors that determine the group classification. If the included secondary covariates fail to improve the resulting classification, then the weight parameter should be a fixed value, resulting in a FMNB-g model.

2.2. Parameter estimation method

The maximum likelihood estimation with expectation maximization (EM) algorithm and Bayesian estimation are the most widely applied methods to estimate the parameters of mixture models. Assuming the number of components is known, Bayesian

³ In fact, there may be more than two components but, for illustration purposes, the FMNB-2 is considered in this paper. Adding more components in the mixture model would significantly complicate the modeling process.

approach can be implemented with data augmentation and Markov Chain Monte Carlo (MCMC) estimation procedure using Gibbs sampling techniques. However, one of the main drawbacks of MCMC procedures is that they are generally computationally demanding, and it can be difficult to diagnose convergence. Furthermore, the label switching (the term label switching describes the invariance of the mixture likelihood function under relabeling the components of a mixture model) is another difficulty and has to be addressed explicitly when using a Bayesian approach to conduct parameter estimation and clustering (McLachlan and Peel, 2000; Frühwirth-Schnatter, 2006). For the maximum likelihood estimation, the goal is to find one of the equivalent modes of the likelihood function, so the label switching is of no concern (Frühwirth-Schnatter, 2006).

In this study, the EM algorithm is used to maximize the likelihood function

$$L = L(\Theta, \mathbf{y}, \delta) = \prod_{i=1}^n \left\{ \prod_{j=1}^g [f_j(y_i)^{\delta_{ij}} w_{ij}^{\delta_{ij}}] \right\} \quad (6)$$

with respect to $\Theta = \{(\theta_1, \dots, \theta_g), (\alpha_1, \dots, \alpha_g)\}$. We first treat all the component indicator variables $\delta_i = (\delta_{i1}, \delta_{i2}, \dots, \delta_{ig})'$ as missing variables. The starting values $\delta_i^{(0)}$ are specified by randomly dividing the data into g groups corresponding to the g components of the mixture model. The weight parameter is modeled using a multinomial logistic model where δ_i is a single draw from a multinomial distribution with probability vector \mathbf{w} . The EM algorithm alternates between the E-step and the M-step until convergence. Iteration $r+1$ of the EM algorithm is summarized as follows: E-step: compute the expected complete-data log likelihood given a current estimate of $\Theta^{(r)}$; M-step: the complete data log likelihood is maximized with respect to Θ . For more details about the EM algorithm, interested readers should see Dempster et al. (1977) and Rigby and Stasinopoulos (2010) (pp. 153–158).

As explained in the previous paragraph, the EM algorithm starts with random starting values of $\Theta^{(0)}$. Some studies have showed that different starting strategies and stopping rules can lead to quite different estimates in the context of fitting mixtures of exponential components via the EM algorithm (McLachlan and Peel, 2000). If the initial value $\Theta^{(0)}$ is poorly specified, the EM algorithm can converge very slowly. Moreover, the likelihood function of mixture models might have multiple roots corresponding to local maxima. In order to be sure that we achieved the global (rather than a local) maximum, we repeat the fitting process 20 times using different random starting values. Additional discussion about the local maximum problem is described in Section 5.

3. Data description

To test the applicability of the GFMNB-2 model to different types of crash datasets, intersection crash data and roadway segment crash data were used in this study. The first dataset contains crash data collected at signalized intersections in Toronto, Canada. The second dataset consists of vehicle crash data that occurred on 4-lane undivided rural segments in Texas. The characteristics of the two datasets are described in this section.

3.1. Toronto data

The Toronto dataset contains crash data collected in 1995 at urban 4-legged signalized intersections in Toronto, Canada. The data have been investigated in some previous studies (Miao and Lord, 2003; Park and Lord, 2009). Park and Lord (2009) used this dataset to develop FMP-g and FMNB-g models. This dataset is selected for two main reasons. First, by using the same dataset,

we can compare the modeling results in this study with that of the previous work. Second, this dataset has been extensively used and is found to be of relatively good quality. The explanatory variables are summarized in Table 1. During the one-year study period, there were 10,030 crashes on 852 out of 868 intersections, and the other 16 intersections (2%) did not have any reported crashes. As shown in Table 1, the observed crash frequency ranges from 0 to 54, and the mean frequency is 11.56 with a standard deviation of 10.02. Note that the variance to mean ratio (VMR) is 8.69.

3.2. Texas data

The Texas dataset was collected at 4-lane undivided rural segments in Texas. This dataset contains crash data collected on 1499 undivided rural segments over a five-year period from 1997 to 2001. The segment length ranged from 0.10 to 6.275 miles, with an average of 0.55 miles. During the study period, 553 out of the 1,499 (37%) segments did not have any reported crashes, and a total of 4,253 crashes occurred on 946 segments. The mean of crashes was 2.84, with a standard deviation of 5.69 and a VMR of 11.4. Table 2 provides the summary statistics for the Texas data.

4. Modeling results

This section describes the modeling results of the GFMNB-2 models for the Toronto data and the Texas data, respectively. In this study, the GFMNB-2 models were estimated using GAMLSS package in software R (R Development Core Team, 2006).

4.1. Toronto data

Similar to the previous research (Park and Lord, 2009), the mean functional form of the GFMNB-2 models for the Toronto data for each component is as follows:

$$\mu_{j,i} = \beta_{j,0} F_{1i}^{\beta_{j,1}} F_{2i}^{\beta_{j,2}} \quad (7)$$

where $\mu_{j,i}$ is the estimated numbers of crashes at intersection i for component j ; F_{1i} is the daily entering flows from the major approaches at intersection i ; F_{2i} is the daily entering flows from the minor approaches at intersection i ; and $\beta = (\beta_{1,0}, \beta_{1,1}, \beta_{1,2}, \beta_{2,0}, \beta_{2,1}, \beta_{2,2})'$ are the estimated coefficients for components 1 and 2.

So far, no functional form has been proposed for estimating the weight parameter. Since only entering traffic flows (from major and minor intersecting roads) are used as covariates, the three combinations of these two covariates are considered to explore different functional forms of the weight parameter. The considered functional forms are listed below:

$$\text{Model 1: } \frac{w_{2,i}}{1 - w_{2,i}} = \exp(\gamma_0 + \gamma_1 F_{1i} + \gamma_2 F_{2i}) \quad (8)$$

$$\text{Model 2: } \frac{w_{2,i}}{1 - w_{2,i}} = \exp(\gamma_0 + \gamma_1 F_{1i}) \quad (9)$$

$$\text{Model 3: } \frac{w_{2,i}}{1 - w_{2,i}} = \exp(\gamma_0 + \gamma_2 F_{2i}) \quad (10)$$

where $w_{2,i}$ is the estimated weight of component 2 at intersection i ; and $\gamma = (\gamma_0, \gamma_1, \gamma_2)'$ are the estimated coefficients. The three equations above offer all possible functional forms. It should be pointed out that the traffic flows are directly used in the equations instead of taking the logarithm of the flows. In the following section, Model 1 is denoted as GFMNB-2(1), Model 2 is defined as GFMNB-2(2), and Model 3 is notated as GFMNB-2(3).

The results for the FMNB-2 models and those from the standard NB model are provided in Table 3. For the FMNB-2 model, the parameter estimation results are within the 95% credible intervals of the results in Park and Lord's paper (2009). The difference

Table 1
Summary statistics of characteristics for the Toronto data.

Variable (symbol)	Minimum	Maximum	Mean (SD ^b)	Sum
Number of crashes (μ)	0	54	11.56 (10.02)	10030
Major-approach AADT ^a (F_1)	5469	72178	28044.81 (10660.39)	
Minor-approach AADT (F_2)	53	42644	11010.18 (8599.40)	

^a Annual average daily traffic.

^b Standard deviation.

Table 2
Summary statistics of characteristics for the Texas data.

Variable (symbol)	Minimum	Maximum	Mean (SD ^a)	Sum
Number of crashes in 5 years (μ)	0	97	2.84(5.69)	4253
Average daily traffic over the 5 years (F)	42	24800	6613.61 4010.01	
Total Shoulder Width (SW)	0	40	9.96(8.02)	
Curve Density (CD)	0	18.07	1.43 (2.35)	
Segment Length (L) (miles)	0.1	6.28	0.55(0.67)	830.49

^a Standard deviation.

Table 3
Modeling results of the FMNB-g models for the Toronto data.

FMNB-g	Weight	$\ln(\beta_0)$	β_1	β_2	ϕ
Single component ^a					
Estimate	1.00	−10.2458	0.6207	0.6853	7.1564
Std. error		0.4790	0.0470	0.0215	0.9148
Two component					
Component 1					
Estimate	0.2241 ^b	−10.8716	0.9066	0.4609	27.5224
Std. error		0.7330	0.0727	0.0292	0.7248
Component 2					
Estimate	0.7759	−10.0893	0.5005	0.7923	7.3082
Std. error		0.5491	0.0537	0.0256	0.9024

^a FMNB-1 model is the standard NB model.

^b The standard error of the estimated weight is not provided in the model output.

is probably caused by different parameter estimation approaches. The results for the GFMNB-2 models are provided in Table 4. All coefficients of variables have plausible values. For the GFMNB-2(1) model, note that the coefficients of major entering flow and minor entering flow for the weight parameter are 4.77E-05 and −1.48E-04, respectively. The results indicate that as the major entering flow increases, the probability that the selected site will belong to

component 2 increases, while as the minor entering flow increases, the probability that the selected site will belong to component 2 decreases.

The goodness-of-fit statistics are provided in Table 5. Compared to the standard NB model, the FMNB-2 and GFMNB-2 models all have the smaller deviance and Akaike information criterion (AIC) values; however the standard NB model has a smaller Bayesian

Table 4
Modeling results of the GFMNB-2 models for the Toronto data.

GFMNB-2	$\ln(\beta_0)$	β_1	β_2	ϕ	γ_0	γ_1	γ_2
Model 1							
Component 1					−1.5744 ^a	4.77E-05	−1.48E-04
Estimate	−10.6331	0.6378	0.7121	10.9902			
Std. error	0.4656	0.0469	0.0228	0.8955			
Component 2							
Estimate	−6.5530	0.6383	0.2037	3.3872			
Std. error	1.5395	0.1462	0.0649	0.8315			
Model 2					0.888925	1.08E-05	
Component 1							
Estimate	−11.6387	0.9864	0.4531	25.5593			
Std. error	0.73729	0.0731	0.02962	0.7303			
Component 2							
Estimate	−9.9917	0.4892	0.7952	7.2790			
Std. error	0.55201	0.05392	0.02563	0.9023			
Model 3					−0.63958		1.60E-04
Component 1							
Estimate	−5.9415	0.4233	0.3639	12.8713			
Std. error	0.93364	0.08542	0.04463	0.7032			
Component 2							
Estimate	−9.5519	0.6508	0.5913	13.7357			
Std. error	0.47294	0.04664	0.02252	0.8846			

^a The standard errors of the estimated coefficients associated with weight parameter are not provided in the model output.

Table 5

Goodness-of-fit statistics for the Toronto data.

	Standard NB	FMNB-2	GFMNB-2(1)	GFMNB-2(2)	GFMNB-2(3)
Deviance	5069.26	5048.75	5036.02	5048.62	5039.48
AIC	5077.26	5066.75	5058.02	5068.62	5059.48
BIC	5096.33	5109.65	5110.45	5116.29	5107.15

information criterion (BIC) value. This is because although both AIC and BIC penalize the number of parameters in the model, BIC generally penalizes free parameters more strongly than AIC does. As the results indicated, not all GFMNB-2 models can significantly improve the goodness-of-fit of the FMNB-2 model. On the one hand, for the GFMNB-2(2) model (with the weight parameter modeled by using the major entering flow), only the deviance value is slightly better than that of the FMNB-2 model, and AIC and BIC values are even worse. This suggests that the GFMNB-2(2) model did not make any improvement in terms of fit by including the major entering flow as the covariate in the functional form. On the other hand, for the GFMNB-2(3) model (with the weight parameter modeled by using the minor entering flow), the deviance, AIC and BIC values are all smaller than those of the FMNB-2 model. This indicates that the GFMNB-2(3) model can improve the goodness-of-fit by using minor traffic flow as the covariate. The goodness-of-fit results support the conclusion that the minor entering flow plays a significant role for modeling the weight parameter and the major entering flow is relatively insignificant for this dataset. Between the FMNB-2 and GFMNB-2 models, although the GFMNB-2(1) model has the smallest deviance and AIC values, its BIC value is slight larger. Only the GFMNB-2(3) model has the smaller deviance, AIC and BIC values than those of the FMNB-2 model. In summary, although the GFMNB-2 models did not consistently perform better based on AIC and BIC, the parameter estimate results (Tables 3 and 4) suggest that there is something going on in the dataset that remains unexplained by the FMNB and standard NB models. Therefore, the GFMNB-2 model should be the recommended model since it provides additional information about the source of the dispersion in the dataset. This is described in the following paragraph.

In order to further explore the nature of the dispersion in the data, based on the FMNB-2 and GFMNB-2 models, the crash data were classified into two groups by assigning each site to the component with the highest posterior probability. The summary statistics of variables for each group are provided in Table 6. Note that the VMRs for crashes are also calculated for each group. Since the sites

with similar characteristics are generally classified into the same group, it is likely to observe less variation (in other words, lower dispersion) within the same group. Thus, the VMR can be used to reasonably measure the homogeneity in each group. For the FMNB-2 model, the probabilities of components 1 and 2 are 7% and 93% respectively. Compared with the results in Park and Lord's paper (2009), there are significant differences in the summary statistics of the variables. This is probably caused by different estimation methods. We also calculated the VMRs for the groups in their paper. The values are 5.31 for the minor component and 9.61 for the major component. Note that these values are larger than the corresponding VMRs for the two components in the FMNB-2 model from this study. And this may indicate that the group classification in this study is more reasonable. Between the FMNB-2 and GFMNB-2 models, the two groups in the GFMNB-2(3) model have the least VMRs. The VMR for component 1 is only 1.12 (the crash data in component 1 have almost equal variance and mean), and the VMR for component 2 is 6.19. Moreover, for the two components in the GFMNB-2(3) model, there is a striking difference in the mean value of the minor entering flow. Fig. 1 shows that many of low minor entering flow is assigned to component 1 resulting in a low average value, and most of high minor entering flow is associated with component 2; while there is no significant difference in the distribution of major entering flow between two components. Based on the modeling results of the GFMNB-2(3) model, it can be seen that the minor entering flow helps classify the intersections. Therefore, it can be concluded that the minor entering flow is a more significant factor for determining the mean and standard deviation of the number of crashes in each group, and thus the variability in minor entering flow is an important source of dispersion in the data.

4.2. Texas data

The modeling results for the Texas data are provided in this section. When analyzing the Texas data, we consider the segment

Table 6

Summary Statistics of each component for the Toronto data.

FMNB-2	Component 1 (7%)				Component 2 (93%)		
	Crashes	F1	F2		Crashes	F1	F2
Mean	14.05	31863.76	4604.50	Mean	11.38	27771.35	11468.86
SD	8.32	9613.10	3303.99	SD	10.11	10684.56	8679.27
VMR ^a	4.93			VMR	8.98		
GFMNB-2(1)	Component 1 (93%)				Component 2 (7%)		
	Crashes	F1	F2		Crashes	F1	F2
	Mean	12.05	27390.62		Mean	4.76	37014.98
	SD	10.11	10261.21		SD	5.03	12005.27
VMR	8.49			VMR	5.32		
GFMNB-2(2)	Component 1 (6%)				Component 2 (94%)		
	Crashes	F1	F2		Crashes	F1	F2
	Mean	13.67	31481.60		Mean	11.42	27825.80
	SD	8.57	8994.13		SD	10.09	10725.25
VMR	5.37			VMR	8.92		
GFMNB-2(3)	Component 1 (33%)				Component 2 (67%)		
	Crashes	F1	F2		Crashes	F1	F2
	Mean	3.26	26594.37		Mean	15.69	28768.78
	SD	1.91	10781.60		SD	9.86	10534.17
VMR	1.12			VMR	6.19		

^a Variance to mean ratio.

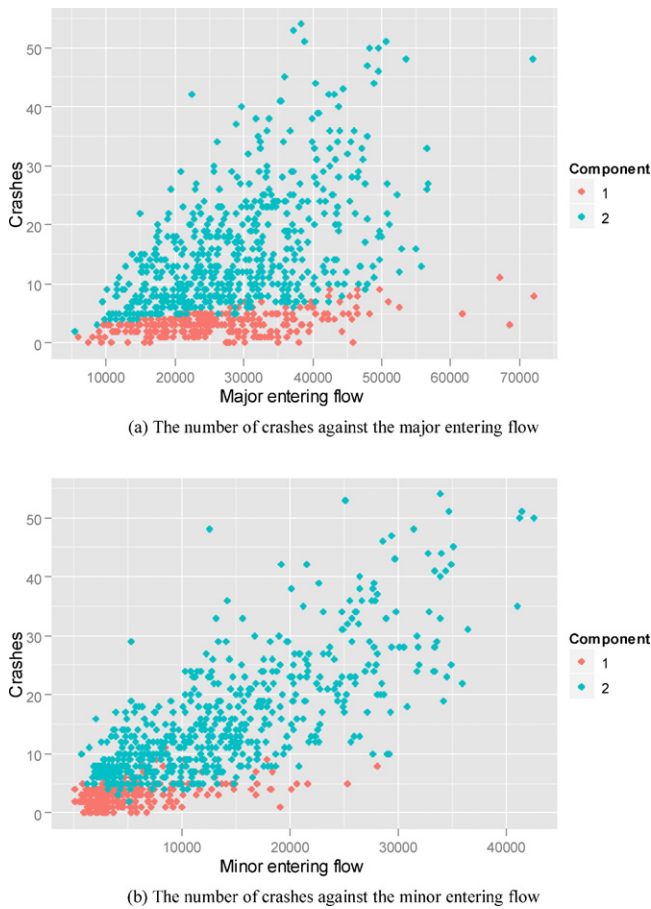


Fig. 1. Scatter plots of the two components for the Toronto data (GFMNB-2(3)).

length as an offset term (Eq. (11)), which means that the number of crashes is linearly proportional to the segment length. Compared to the Toronto data, more variables are included in the modeling process for the Texas data. The mean functional form for each component is adopted as follows:

$$\mu_{j,i} = \beta_{j,0} L_i F_i^{\beta_{j,1}} e^{\beta_{j,2} * SW_i + \beta_{j,3} * CD_i} \quad (11)$$

where $\mu_{j,i}$ is the estimated numbers of crashes at segment i for component j ; L_i is the segment length in miles for segment i ; F_i is the flow (average daily traffic over five years) traveling on segment i ; SW_i is the total shoulder width in feet for segment i ; CD_i is the curve density (curves per mile) for segment i ; and $\beta = (\beta_{1,0}, \beta_{1,1}, \beta_{1,2}, \beta_{1,3}, \beta_{2,0}, \beta_{2,1}, \beta_{2,2}, \beta_{2,3})'$ are the estimated coefficients for components 1 and 2.

Table 7
Modeling results of the FMNB-g models for the Texas data.

FMNB-g	Weight	$\ln(\beta_0)$	β_1	β_2	β_3	ϕ
Single component ^a						
Estimate	1.00	−8.4090	0.9530	−0.0117	0.0690	2.4918
Std. error		0.3962	0.0448	0.0033	0.0120	0.9225
Two component						
Component 1						
Estimate	0.4007 ^b	−7.2777	0.8531	−0.0153	0.1061	1.8145
Std. error		0.5874	0.0669	0.0053	0.0184	0.8966
Component 2						
Estimate	0.5993	−10.0554	1.1174	−0.0085	0.0226 ^c	8.1173
Std. error		0.4779	0.0536	0.0036	0.0152	0.8016

^a FMNB-1 model is the standard NB model.

^b The standard error of the estimated weight is not provided in the model output.

^c Not significant at 5% significance level.

Given the specific objectives of this study, instead of developing the models with the best statistical fit considering every possible combination of covariates, only important explanatory variables (flow and segment length) were used as the covariates to model the weight parameter. The considered functional forms are listed below:

$$\text{Model 1 : } \frac{w_{2,i}}{1 - w_{2,i}} = \exp(\gamma_0 + \gamma_1 F_i) \quad (12)$$

$$\text{Model 2 : } \frac{w_{2,i}}{1 - w_{2,i}} = \exp(\gamma_0 + \gamma_2 L_i) \quad (13)$$

$$\text{Model 3 : } \frac{w_{2,i}}{1 - w_{2,i}} = \exp(\gamma_0 + \gamma_1 F_i + \gamma_2 L_i) \quad (14)$$

where $w_{2,i}$ is the estimated weight of component 2 at segment i ; and $\gamma = (\gamma_0, \gamma_1, \gamma_2)'$ are the estimated coefficients. For the above three equations, note that the explanatory variables flow and segment length are directly used in the equations. In the following section, Model 1 is denoted as GFMNB-2(1), Model 2 is defined as GFMNB-2(2), and Model 3 is notated as GFMNB-2(3).

The modeling results for the FMNB-2 and GFMNB-2 models are provided in Tables 7 and 8, respectively. All coefficients of variables have plausible values that are consistent with previous work on this topic. For the GFMNB-2(3) model, the results indicate that as flow and segment length increase, the probability that the selected site will belong to component 2 decreases.

The goodness-of-fit statistics for the Texas data are provided in Table 9. Similar to the Toronto data, the FMNB-2 model has the smaller deviance and AIC values than the standard NB model. In addition, the GFMNB-2(2) and GFMNB-2(3) models are preferred over the standard NB model based on deviance, AIC and BIC values. Note that not all GFMNB-2 models can significantly improve the goodness-of-fit of the FMNB-2 model. For example, the GFMNB-2(1) model (with the weight parameter modeled by using flow) performs even worse than the FMNB-2 model in terms of fit. This may suggest that flow is an insignificant covariate. For the GFMNB-2(2) model (with the weight parameter modeled by using segment length), the improvement of the fitting performance indicates that segment length is significant. Thus it is recommended to include segment length as a covariate in modeling the weight parameter for the Texas data. Between the GFMNB-2(2) and GFMNB-2(3) models, the GFMNB-2(3) model barely improves the fitting performance by adding flow in the functional form.

We also classified the Texas data into two groups based on the FMNB-2 and GFMNB-2 models. The summary statistics of variables and the VMRs for each group are provided in Table 10. For the FMNB-2 model, the VMRs of components 1 and 2 are 13.76 and 6.84, respectively, and these ratios are larger than the corresponding ratios for the two components in the GFMNB-2 models, thus indicating that the group classification based on the

Table 8
Modeling results of the GFMNB-2 models for the Texas data.

GFMNB-2	$\ln(\beta_0)$	β_1	β_2	β_3	ϕ	γ_0	γ_1	γ_2
Model 1								
Component 1						0.22925 ^a	1.18E-05	
Estimate	-7.3868	0.8641	-0.0154	0.1045	1.8201			
Std. error	0.5767	0.0657	0.0052	0.018	0.898			
Component 2								
Estimate	-10.2069	1.1334	-0.0082	0.0217 ^b	8.6798			
Std. error	0.4849	0.0543	0.0036	0.0154	0.793			
Model 2								
Component 1					3.221774		-5.35905	
Estimate	-6.1239	0.687	-0.0155	0.1873	4.4017			
Std. error	0.4461	0.051	0.0042	0.0197	0.8757			
Component 2								
Estimate	-11.805	1.3227	-0.0013 ^b	0.023 ^b	1.7832			
Std. error	0.6664	0.0745	0.0049	0.0166	0.8921			
Model 3								
Component 1					3.469476	-5.62E-05	-5.26571	
Estimate	-6.0831	0.6821	-0.0148	0.1851	4.4327			
Std. error	0.4405	0.0502	0.0041	0.0190	0.8776			
Component 2								
Estimate	-11.8556	1.3300	-0.0011 ^b	0.0201 ^b	1.7109			
Std. error	0.6828	0.0766	0.0050	0.0170	0.8907			

^aThe standard errors of the estimated coefficients associated with weight parameter are not provided in the model output.

^bNot significant at 5% significance level.

Table 9
Goodness-of-fit statistics for the Texas data.

	Standard NB	FMNB-2	GFMNB-2(1)	GFMNB-2(2)	GFMNB-2(3)
Deviance	5132.86	5105.08	5105.06	5062.49	5061.94
AIC	5142.86	5127.08	5129.06	5086.49	5087.94
BIC	5169.42	5185.52	5192.81	5150.24	5157.00

Table 10
Summary statistics of each component for the Texas data.

FMNB-2	Component 1 (14%)					Component 2 (86%)				
	Crashes	F	L	SW	CD	Crashes	F	L	SW	CD
Mean	8.38	6544.21	0.84	8.97	2.04	Mean	1.94	6624.85	0.51	10.12
SD	10.74	4224.78	1.00	7.73	2.74	SD	3.64	3975.72	0.58	8.05
VMR ^a	13.76					VMR	6.84			2.27
GFMNB-2(1)										
Component 1 (15%)										
Crashes F L SW CD										
Mean	7.97	6346.10	0.84	8.97	1.91	Mean	1.91	6661.84	0.50	10.14
SD	10.44	4166.69	1.01	7.83	2.67	SD	3.61	3980.88	0.57	8.04
VMR	13.67					VMR	6.83			2.28
GFMNB-2(2)										
Component 1 (25%)										
Crashes F L SW CD										
Mean	7.51	7144.17	1.33	8.56	1.27	Mean	1.25	6433.43	0.29	10.44
SD	9.19	4309.56	0.94	7.58	1.42	SD	2.20	3888.65	0.14	8.11
VMR	11.26					VMR	3.86			2.59
GFMNB-2(3)										
Component 1 (27%)										
Crashes F L SW CD										
Mean	7.39	7571.92	1.30	8.64	1.29	Mean	1.19	6267.19	0.29	10.44
SD	9.03	4471.95	0.94	7.64	1.43	SD	2.12	3772.04	0.14	8.10
VMR	11.03					VMR	3.78			2.60

^a Variance to mean ratio.

GFMNB-2 models is more appropriate. Between the GFMNB-2 models, both the GFMNB-2(2) and GFMNB-2(3) models have smaller VMRs than that of the GFMNB-2(1) model. For the two components in the GFMNB-2(2) and GFMNB-2(3) models, there is a remarkable difference in the mean values of segment length. For example, the ratio of the means of segment lengths between two components in the GFMNB-2(2) model is 1.33/0.29 = 4.59, while that ratio in FMNB-2 model is 0.84/0.51 = 1.65. The difference in the mean value of other variables is not so noticeable in the GFMNB-2(2) and GFMNB-2(3) models. Fig. 2(b) shows that most of the small segment lengths are assigned to component 2, resulting in a low average value, and most of the long segment lengths are associated

with component 1. For Fig. 2(d), there is a clear difference in the distributions of points between two components. Moreover, although not shown here, the scatter plot between crashes and the number of curves also supports that the number of curves has different effects on crashes between two components. Overall, since there is no significant difference in the distributions of other explanatory variables between two components, we can thus conclude that the variabilities in segment length and curve density are two important sources of dispersion observed in the data. Previously, Hauer (2001) and Geedipally et al. (2009) have discussed the relationship between segment length and dispersion in crash modeling.

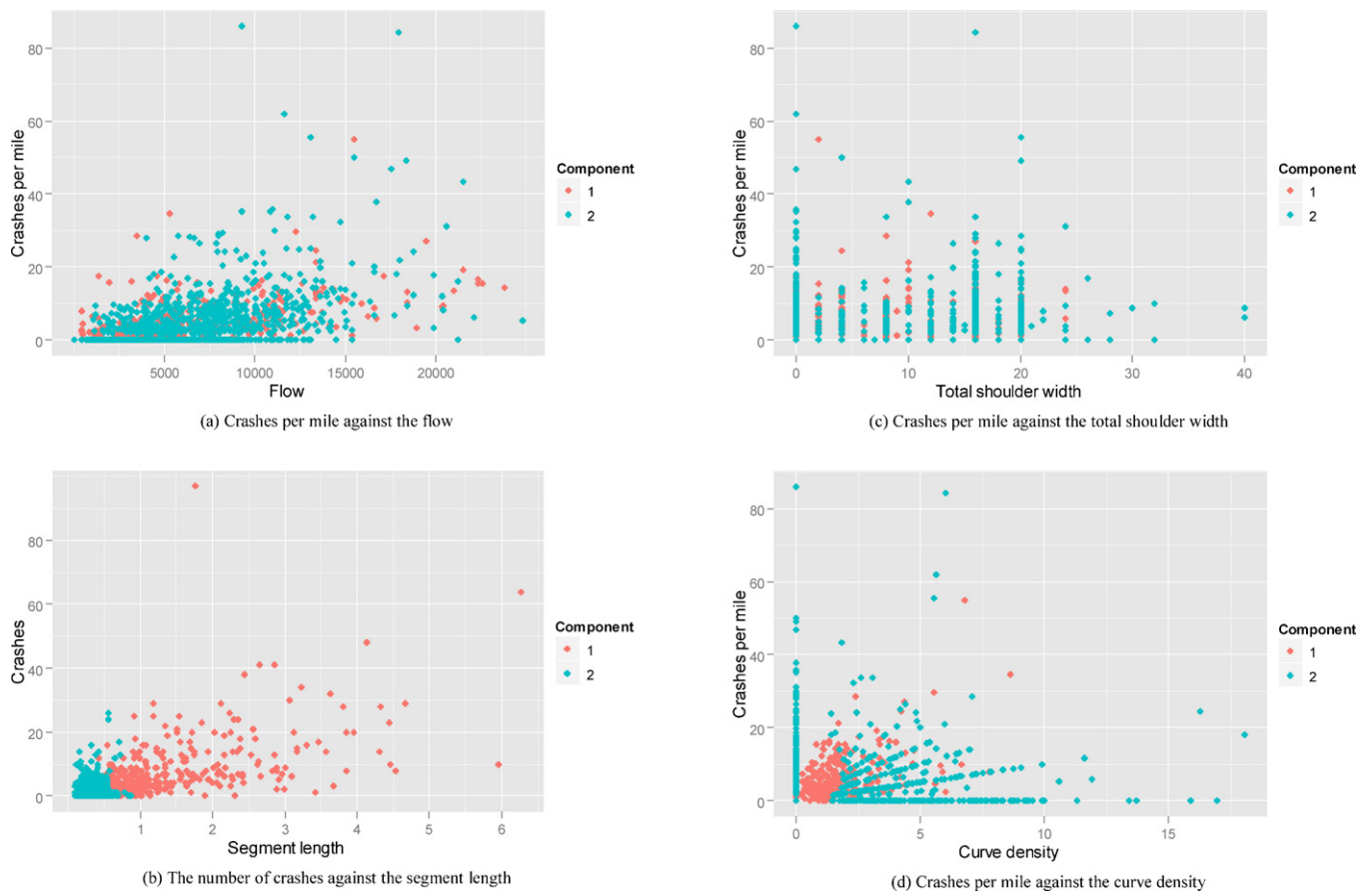


Fig. 2. Scatter plots of the two components for the Texas data (GFMNB-2(2)).

5. Discussion

In this paper, the modeling results are very interesting and deserve some further discussion. Based on the results in this study, the following conclusions can be made: the GFMNB-2 models can perform better than the FMNB-2 models in terms of fitting performance; and more importantly, if the GFMNB-2 model includes significant explanatory variables as the covariates for modeling the weight parameter, then these covariates can help classify the sites and improve the resulting classification. Thus, the GFMNB-2 model can be used to better reveal the source of dispersion than the FMNB-2 model. Based on the grouped sites from the GFMNB-2 models, it can be seen that the crash data may actually be generated from two distinct sub-populations, with each population having its own regression coefficients and degrees of dispersion (one sub-population consists of accident-prone sites and the other contains low-risk sites). For the Toronto data, from the analysis of GFMNB-2 models, we have determined that the minor entering flow is the major source of the dispersion and the value of minor entering flow can significantly affect the level of the dispersion. For the Texas data, compared with other explanatory variables, segment length and curve density are two important sources of dispersion observed in the data.

Despite the popularity of the EM algorithm, one problem with mixture models is that in general the likelihood equation can potentially have multiple roots corresponding to local maxima. When estimating the mixture models in this study, we also experienced this problem that the roots of the likelihood equation vary between a few values (each value corresponds to a maximum of the likelihood function). According to McLachlan and Peel (2000), in the scenario that the information of any known consistent estimator

of Θ is absent, an obvious choice for the root of the likelihood equation is the one corresponding to the largest of the local maxima located. Thus, in order to ensure a global maximum has been found, many different random starting values are applied with the EM algorithm and we select the optimal root that corresponds to the largest likelihood value. As more explanatory variables are included as covariates to model the mean and weight parameters, the estimation process becomes more complex and the EM algorithm gradually suffers from slow convergence. To make sure that the convergent criterion for the EM algorithm is satisfied before the maximum number of iterations is reached, we let the maximum number of iterations equal to 2000 and the modeling results show that all models have met the convergent criterion.

One important point that should be noted is that only a few explanatory variables are included in the mean functional forms (for example, the traffic flow-only model, or Eq. (7), is used for the Toronto data). There are two reasons for selecting the simplified mean functional forms for this study. First, due to the age of the datasets, we are not able to get access to other important explanatory variables (i.e., traffic signal phasing schemes, pavement friction values, weather information, etc.). More importantly, adding more variables in the mean functional forms would significantly complicate the modeling process. However, one problem associated with the simplified mean functional form is that leaving out important explanatory variables can result in biased parameter estimates and suspicious inferences (see Washington et al., 2011). As discussed by Mitra and Washington (2007), when sufficient explanatory variables are used to model the mean, the modeling of the dispersion parameter generally becomes insignificant. Similarly, considering the mean structure used in this study, it is possible that if a well-defined mean

functional form is used, the difference in group classification results between FMNB-2 and GFMNB-2 models could be slight. In that case, the modeling of the weight parameter may become unnecessary.

Another point worth noting is that transportation safety analysts are recommended to evaluate different functional forms describing the weight parameter. The selection of the best functional form should be based on not only the goodness-of-fit statistics, but also the resulting classification. For the Toronto data, the results show that the minor entering flow plays a more significant role for modeling the weight parameter than the major entering flow. By using the minor entering flow in the functional form (Eq. (10)), the classified groups in the corresponding GFMNB-2 model have the least VMRs. If both the major and minor entering flows are included in the functional form (Eq. (8)), the corresponding GFMNB-2 model yields a worse result in terms of group classification. Interestingly, for the Texas data, segment length (used as an offset term in the mean functional form) is found to be the most significant variable for modeling the weight parameter. If segment length is used together with other explanatory variables (i.e., flow (Eq. (14))), the corresponding GFMNB-2 model can provide a better result in terms of group classification. The modeling results based on the two datasets suggest that the significant variable for determining the weight parameter should always be included in the functional form. Moreover, adding other less significant explanatory variables in the functional form may positively or negatively affect the grouping outcome. Thus, more work needs to be done to fully investigate the effect of different functional forms on the modeling results.

The Texas data used in this study are highly dispersed (the VMR is 11.4). Interestingly, Zou et al. (submitted for publication) also examined the Texas data using a Sichel model and they showed that the Sichel model works very well when the crash data are highly dispersed. When comparing the goodness-of-fit statistics of the two models, it is shown that the FMNB-2 model and the Sichel model are favored by different criteria. The FMNB-g model is recommended if the high dispersion is caused by the heterogeneity in the crash data (the crash data are suspected to consist of observations from several distinct sub-populations). For example, following the same procedure adopted in Park et al. (2010), we generated data with high dispersion based on FMNB-2 models (not shown here). Although the Sichel model can provide a satisfactory fitting performance for the data, it failed to consider the population heterogeneity and cannot reasonably capture the heterogeneous impact of the covariates. As a result, it gave a poor prediction performance compared to the FMNB-2 model.

For future work, first, since crash data characterized by small size and low sample-mean values can cause estimation problems, the robustness of the GFMNB-2 models should be examined. Second, when the highly dispersed crash data are suspected to belong to different groups, a finite mixture of Sichel models should be used and compared to the finite mixture of NB models (both for fixed and varying weights). Third, the Bayesian method can be used to estimate the GFMNB-2 models, and their influence on the modeling results should be investigated. Fourth, recent studies in transportation safety have shown that NB models with a varying dispersion parameter can provide better statistical fitting performance (Heydecker and Wu, 2001; El-Basyouny and Sayed, 2006) or help describe the characteristics of the dispersion (Miaou and Lord, 2003; Geedipally et al., 2009). Thus, the GFMNB-2 models with the varying dispersion parameter can be developed to analyze crash data. Finally, since only limited explanatory variables are used to model the weight parameter and the selection of the functional form could significantly affect the modeling results, it is useful to further evaluate various functional forms for the weight parameter.

6. Summary and conclusions

This paper has described the application of the GFMNB-2 model for analyzing crash data. The proposed models were evaluated using two crash datasets and the important findings can be summarized as follows: first, the GFMNB-2 models can provide more reasonable classification results, as well as better statistical fitting performance than the FMNB-2 models for both roadway segment crash data and intersection crash data; second, the GFMNB-2 models can be used to better reveal the source of dispersion observed in the crash data than the FMNB-2 models. In addition, transportation safety analysts should evaluate different functional forms describing the weight parameter when using the GFMNB-2 models, since the selection of the functional form for weight parameter has a significant impact on the grouping. In conclusion, it is believed that in many cases the GFMNB-2 models may be a better alternative to the FMNB-2 models for explaining the heterogeneity and the nature of the dispersion in the crash data.

References

- Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis and Prevention* 41 (1), 153–159.
- Chang, I., Kim, S., 2012. Modelling for identifying accident-prone spots: Bayesian approach with a Poisson mixture model. *KSCE Journal of Civil Engineering* 16 (3), 441–449.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39, 1–38.
- El-Basyouny, K., Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models. *Transportation Research Record: Journal of the Transportation Research Board* 1950, 9–16.
- El-Basyouny, K., Sayed, T., 2010. A method to account for outliers in the development of safety performance functions. *Accident Analysis and Prevention* 42 (4), 1266–1272.
- Frühwirth-Schnatter, S., 2006. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics, Springer, New York.
- Frühwirth-Schnatter, S., Kaufmann, S., 2008. Model-based clustering of multiple time series. *American Statistical Association Journal of Business & Economic Statistics* 26 (1), 78–89.
- Frühwirth-Schnatter, S., Pamminger, C., Weber, A., Winter-Ebmer, R., 2011. Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *Journal of Applied Econometrics*.
- Geedipally, S.R., Lord, D., Park, B.-J., 2009. Analyzing different parameterizations of the varying dispersion parameter as a function of segment length. *Transportation Research Record: Journal of the Transportation Research Board* 2103, 108–118.
- Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-Lindley generalized linear model: characteristics and application using crash data. *Accident Analysis and Prevention* 45, 258–265.
- Guo, J.Q., Trivedi, P.K., 2002. Flexible parametric models for long-tailed patent count distributions. *Oxford Bulletin of Economics & Statistics* 64 (1), 63–82.
- Hauer, E., 2001. Overdispersion in modeling accidents on road sections and in empirical Bayes estimation. *Accident Analysis and Prevention* 33 (6), 799–808.
- Heydecker, B.G., Wu, J., 2001. Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. *Advances in Engineering Software* 32, 859–869.
- Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention* 37 (1), 35–46.
- Lord, D., Park, P.Y.-J., 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. *Accident Analysis and Prevention* 40 (4), 1441–1457.
- Lord, D., Mannering, F.L., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A* 44 (5), 291–305.
- Lord, D., Geedipally, S.R., 2011. The negative binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accident Analysis and Prevention* 43 (5), 1738–1742.
- Malyskhina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. *Accident Analysis and Prevention* 41 (2), 217–226.
- McLachlan, G., Peel, D., 2000. *Finite Mixture Models*. John Wiley & Sons Inc., New York.
- Miaou, S.-P., Lord, D., 2003. Modeling traffic-flow relationships at signalized intersections: dispersion parameter, functional form and Bayes vs Empirical Bayes. *Transportation Research Record: Journal of the Transportation Research Board* 1840, 31–40.

- Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. *Accident Analysis and Prevention* 39 (3), 459–468.
- Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis and Prevention* 41 (4), 683–691.
- Park, B.J., Lord, D., Hart, J.D., 2010. Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis. *Accident Analysis and Prevention* 42 (2), 741–749.
- Poch, M., Mannering, F.L., 1996. Negative binomial analysis of intersection accident frequency. *Journal of Transportation Engineering* 122, 105–113.
- R Development Core Team. 2006. R: A language and environment for statistical computing. R. Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.r-project.org/>
- Rigby, R.A., Stasinopoulos, D.M., 2010. A flexible regression approach using GAMLSS in R. <http://gamlss.org/images/stories/papers/book-2010-Athens.pdf> (accessed April 2012).
- Wang, P.M., Cockburn, I.M., Puterman, M.L., 1998. Analysis of patent data – a mixed Poisson regression model. *Journal of Business and Economic Statistics* 16 (1), 27–41.
- Washington, S., Karlaftis, M.G., Mannering, F., 2011. *Statistical and Econometric Methods for Transportation Data Analysis*, second edition. Chapman and Hall/CRC, Boca Raton, FL.
- Zou, Y., Zhang, Y., 2011. Use of skew-normal and skew-t distributions for mixture modeling of freeway speed data. *Transportation Research Record: Journal of the Transportation Research Board* 2260, 67–75.
- Zou, Y., Lord, D., Zhang Y. Analyzing highly dispersed crash data using the Sichel generalized additive models for location, scale and shape, submitted for publication.