Analyzing Highly Dispersed Crash Data Using the Sichel Generalized Additive Models for Location, Scale and Shape

By

Yajie Zou
Ph.D. Candidate
Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: 979/595-5985, fax: 979/845-6481
E-mail: yajiezou@tamu.edu

Dominique Lord*, Ph.D.
Associate Professor and Zachry Development Professor I
Zachry Department of Civil Engineering
Texas A&M University, TAMU 3136
College Station, TX 77843-3136
Phone: (979) 458-3949, fax: (979) 845-6481
E-mail: d-lord@tamu.edu

Yunlong Zhang, Ph.D.
Associate Professor
Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: 979/845-9902, fax: 979/845-6481
E-mail: yzhang@civil.tamu.edu

Paper submitted for publication

December 5th, 2011

*Corresponding author

ABSTRACT

This paper documents the application of the Sichel (SI) generalized additive models for location, scale and shape (GAMLSS) for modeling highly dispersed crash data. The Sichel distribution is a compound Poisson distribution, which mixes the Poisson parameter with the generalized inverse Gaussian distribution. This distribution is particularly useful as a model for highly dispersed count data and has provided satisfactory fits in many cases where other models have proved to be inadequate. The objectives of this study were to evaluate the application of the Sichel GAMLSS for analyzing highly dispersed crash data and compare the results with the traditional Negative Binomial (NB) generalized linear model (GLM). To accomplish the objectives of the study, the NB, zero-inflated NB (ZINB) and SI GAMLSS were developed and compared using two highly dispersed crash datasets. The first dataset contains the crash data collected on 338 rural interstate road sections in Indiana. The second dataset consists of vehicle crash data that occurred on undivided 4-lane rural roadway segments in Texas. Several goodness-of-fit metrics were used to assess the statistical fit of the models. The results show that the Sichel GAMLSS can always have a better fitting performance than the NB and ZINB for the crash datasets examined in this study. Thus, the SI GAMLSS may offer a viable alternative to the traditionally used NB GLMs for analyzing highly dispersed crash datasets.

Key Words: Negative binomial, sichel, dispersion, generalized additive models for location, scale and shape (GAMLSS), crash data.

1. INTRODUCTION

As documented in the literature (Lord and Mannering, 2010), one notable characteristic associated with crash data is that the variance usually exceeds the mean of the crash counts. The large amount of zeros and a long/heavy tail observed in crash data generally create high dispersion. As described in previous research, the Poisson (PO) distribution can potentially result in biased and inconsistent parameter estimates because the PO distribution restricts the mean and variance of crash frequency to be equal. To overcome these limitations, the Negative Binomial (NB) distribution has often been used as an alternative because the NB model relaxes the assumption that the mean equals the variance. Within the framework of generalized linear model (GLM), a large number of analysis models for overcoming the over-dispersion problem have been proposed by transportation safety analysts (Lord and Mannering, 2010). Miaou (1994) and Poch and Mannering (1996) noted that if the over-dispersion in crash data is found to be moderate or high, the NB GLM could be used. More recently, several researchers have proposed different models, under the framework of the GLM, for analyzing over-dispersed data. These models, such as the Conway-Maxwell-Poisson (Lord et al., 2008), the Random Parameters (Anastasopoulos and Mannering, 2009), the Generalized Additive Model (GAM) (Xie and Zhang, 2008), and the Finite Mixture/Markov Switching (Park and Lord, 2009; Malyshkina et al., 2009), have often been found to perform better than the NB both in terms of fit or predictive capabilities.

These new models however may still provide erroneous or biased estimates when they are developed using data characterized by a large amount of zeros and/or with a very large dispersion (i.e., long tail). For instance, it has been documented that the NB distribution cannot be used efficiently for datasets that simultaneously have a skewness coefficient greater than two and a mode greater than zero (Stein et al., 1987). The same problem can be observed for datasets that contain a large number of zeros. Zero-inflated models have been proposed to handle such datasets, but they have been found to suffer from important methodological problems, at least for crash data (Lord et al., 2005 & 2007). So far, very few models that have been proposed or evaluated could be suitable for analyzing highly dispersed data. A negative binomial-Lindley (NB-L) distribution was recently introduced by Lord and Geedipally (2011) for analyzing data characterized by a large number of zeros. In a subsequent study, Geedipally et al. (2012) 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. Currently, only the NB-L model works well for highly dispersed data. Even with a large number of statistical models used in highway safety research, there might still exist other statistical models that can fully analyze highly dispersed crash data. Such models are particularly needed when the observed high dispersion cannot be efficiently handled by the NB model.

With the aim of finding a model that can handle highly dispersed crash data, the generalized additive models for location, scale and shape (GAMLSS) are introduced (Rigby and Stasinopoulos, 2005). The GAMLSS can be considered as the extension of GAMs and GLMs. In the GAMLSS, the exponential family distribution assumption for the response variable, y, is relaxed and replaced by a general distribution family, including highly skewed continuous and discrete distributions. Thus, the GAMLSS are suited to model a highly dispersed count response variable. The GAMLSS are a general framework for univariate regression analysis that allows

modeling not only the mean but all other parameters (including dispersion parameter) of the distribution of y as linear and/or nonlinear parametric and/or additive non-parametric functions of explanatory variables (Stasinopoulos and Rigby, 2007). Recently, the GAMLSS have been applied in various fields including phenological research (Hudson et al., 2008), medical studies (Visser et al. 2009), etc. However, the GAMLSS have not been applied to crash data analysis to date. The advantages of the GAMLSS for analyzing the crash data are that: (1) the GAMLSS are especially suited to model the crash data with very high dispersion. (2) the GAMLSS can be a useful tool to explore flexible functional forms between the mean/dispersion parameter and explanatory variables.

Within the framework of the GAMLSS, this paper introduces a new distribution that can handle highly dispersed count data. The Sichel (SI) distribution is a compound Poisson distribution, which mixes the Poisson distribution with the generalized inverse Gaussian distribution. Previous studies (Sichel, 1982; Stein et al., 1987) have shown that the resultant mixture of the Poisson and generalized inverse Gaussian distributions is particularly useful as a model for highly dispersed count data and has provided satisfactory fits in many cases where other models proved to be inadequate. The application of a Sichel distribution for analyzing traffic crash data is documented in this study.

2. METHODOLOGY

This section describes the characteristics of the GAMLSS and the Sichel distribution.

2.1. THE GENERALIZED ADDITIVE MODELS FOR LOCATION, SCALE AND SHAPE

The GAMLSS were introduced by Rigby and Stasinopoulos (2005) as a way of overcoming some of the limitations associated with the popular GLM and GAM. A GAMLSS model assumes that for i= 1, 2, …, n, independent observations y_i have a distribution D with probability density function $f(y_i \mid \theta^i)$ conditional on $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$ a vector of four distribution parameters, each of which can be a function to the explanatory variables. The $\mu_i, \sigma_i, \nu_i, \tau_i$ are referred to as the distribution parameters of distribution D. μ_i and σ_i are defined as location and scale parameters. ν_i and τ_i are characterized as shape parameters.

Let y_i, i= 1, 2, …, n, be the n length vector of the response variable. For k= 1, 2, 3, 4, let $g_k(\cdot)$ be a known monotonic link function relating the distribution parameter θ_k to explanatory variables and random effects through an additive model given by (Rigby and Stasinopoulos, 2005):

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk} \tag{1}$$

For example,

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1} \tag{2}$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2}\mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2} \tag{3}$$

$$g_3(\mathbf{v}) = \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3}\mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3} \tag{4}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4}\mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4} \tag{5}$$

Where,

$\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{v}, \boldsymbol{\tau}$ and $\boldsymbol{\eta}_k$ = vectors of length n;

$\boldsymbol{\beta}_k^T$ = vector of length $J_k^{'}$;

$\mathbf{X}_k$ = known design matrix of order $n \times J_k^{'}$;

$\mathbf{Z}_{jk}$ = fixed known $n \times q_{jk}$ design matrix;

$\boldsymbol{\gamma}_{jk}$ = q_{jk} dimensional random variable and is assumed to follow $\boldsymbol{\gamma}_{jk} \sim N_{q_j k}(0, \mathbf{G}_{jk}^{-1})$;

$\mathbf{G}_{jk}^{-1}$ = generalized inverse of a $q_{jk} \times q_{jk}$ symmetric matrix;

J_k and $J_k^{'}$ = number of explanatory variables considered; and,

q_{jk} = dimension of the random effect vector.

There are several important sub-models of the GAMLSS. First, let $\mathbf{Z}_{jk} = \mathbf{I}_n$, where $\mathbf{I}_n$ is an $n \times n$ identity matrix, and $\gamma_{jk} = h_{jk}(\mathbf{x}_{jk})$ for all combinations of j and k, the semi-parametric additive GAMLSS model is given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{6}$$

Where for k =1, 2, 3, 4,

$\boldsymbol{\theta}_k$ = distribution parameter vectors $\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{v}$ and $\boldsymbol{\tau}$;

h_{jk} = unknown function of the explanatory variable; and,

$\mathbf{x}_{jk}$ = vectors of length n.

If there are no additive terms in any of the distribution parameters, we can have the parametric linear GAMLSS model,

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k \tag{7}$$

In equations (6) and (7), replace $\mathbf{X}_k\boldsymbol{\beta}_k$ with $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k)$, where h_k for k =1, 2, 3, 4 are non-linear functions, then the new equations (8) and (9) are the non-linear semi-parametric additive and non-linear parametric GAMLSS models, respectively.

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{8}$$

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) \tag{9}$$

In the framework of the GAMLSS, the logarithm of expected crash frequency can be modeled as either a linear parametric or an additive non-parametric function of explanatory variables. When adding the additive terms in equation (7), the resulting semi-parametric additive GAMLSS model provides the potential for a better fit with the data than the parametric linear GAMLSS model. However, the semi-parametric additive GAMLSS model may run the risk of relaxing the actual relationship between the expected crash frequency and explanatory variables, perhaps at the expense of interpretability of results. Thus, in order to reasonably interpret the modeling results and avoid the risk of overfitting, the parametric linear GAMLSS model (equation (7)) was adopted as the functional form in the study.

Once the GAMLSS model is determined, the parametric vectors $\boldsymbol{\beta}_k$ and the random effects parameters γ_{jk} in equation (1) are estimated by maximizing a penalized likelihood function ℓ_p:

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_{jk}\gamma_{jk}^{'}\mathbf{G}_{jk}\gamma_{jk} \tag{10}$$

where $\ell = \sum_{i=1}^{n}\log f(y_i \mid \theta^i)$ is the log-likelihood function. Since the parametric linear GAMLSS model has been selected, only $\boldsymbol{\beta}_k$ for k =1, 2, 3, 4 are unknown parameters. The penalized likelihood function ℓ_p reduces to ℓ and the objective is to maximize the likelihood function ℓ. Two algorithms can be used to maximize the likelihood function. The first is the CG algorithm which is a generalization of the Cole and Green algorithm. The second algorithm is a generalization of the algorithm used by Rigby and Stasinopoulos for fitting mean and dispersion additive models, and this algorithm is named the RS algorithm. In this study, the RS algorithm was selected and has been successfully used. More details about the two algorithms are given in Stasinopoulos and Rigby's paper (2007).

2.2. CHARACTERISTICS OF THE SICHEL DISTRIBUTION

As discussed above, the Sichel distribution (also known as the Poisson-generalized inverse Gaussian distribution) is a compound Poisson distribution where the mixing distribution of the Poisson rate is a generalized inverse Gaussian distribution. This mixed distribution works very well when the data is highly dispersed. In other situations, it works similar to the NB distribution.

Before deriving the probability density function of the Sichel distribution, we first need to define the NB distribution. The number of crashes y during some time period is assumed to be Poisson distributed, which is defined by:

$$p(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!} \tag{11}$$

Where λ = mean response of the observation.

The NB distribution can be viewed as a mixture of Poisson distributions where the Poisson rate is gamma distributed. For the complete derivation of the NB, readers are referred to Lord and Mannering (2010). The probability density function of the NB is defined as follows:

$$f(y \mid \mu, \sigma) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y+1)} (\frac{\sigma\mu}{1+\sigma\mu})^{y} (\frac{1}{1+\sigma\mu})^{1/\sigma} \tag{12}$$

Where,

y = response variable;

μ = mean response of the observation; and,

σ = dispersion parameter.

The dispersion parameter σ of the NB is traditionally assumed to be a fixed value for the entire crash dataset. However, within the GAMLSS framework, both the mean and dispersion parameters of the NB can be modeled using the explanatory variables with parametric linear functional forms. This will be the subject of a subsequent paper.

The Sichel distribution arises if we let λ take a generalized inverse Gaussian distribution, with the probability density function (pdf) given by

$$f(\lambda \mid \gamma, \theta, \alpha) = \frac{[2(1-\theta)^{1/2} / \alpha\theta]^{\gamma}}{2K_{\gamma}[\alpha(1-\theta)^{1/2}]} \lambda^{\gamma-1} e^{-(\frac{1}{\theta}-1)\lambda - (\frac{\alpha^2\theta}{4\lambda})} \tag{13}$$

For $\lambda > 0$, the distribution parameters are $-\infty < \gamma < \infty$, $0 \le \theta \le 1$, $\alpha \ge 0$ and $K_{\gamma}(t)$ is the modified Bessel function of the third kind of order γ with argument t. Note that the gamma is a limiting distribution of the generalized inverse Gaussian distribution obtained by letting $\alpha \to 0$ and $\gamma > 0$.

The number of crashes y is given by

$$p(y \mid \gamma, \theta, \alpha) = \int_{0}^{\infty} p(y \mid \lambda) f(\lambda \mid \gamma, \theta, \alpha) d\lambda \tag{14}$$

A Sichel distribution can be derived by solving this convolution integral. The pdf of the Sichel distribution with distribution parameters γ, θ and α is defined as:

$$p(y \mid \gamma, \theta, \alpha) = \frac{(1-\theta)^{\gamma/2}}{K_{\gamma}(\alpha\sqrt{1-\theta})} \frac{(\alpha\theta/2)^{y}}{y!} K_{y+\gamma}(\alpha) \tag{15}$$

To overcome some estimation problems associated with the above formulation, Stein et al. (1987)

reparameterized equation (15) using $\xi = \alpha\theta/(2(1-\theta)^{1/2})$ in place of θ and derived the new probability density function:

$$p(y \mid \gamma, \xi, \alpha) = \frac{\xi^y K_{y+\gamma}(\alpha)}{K_\gamma(w) y!(\alpha / w)^{y+\gamma}} \tag{16}$$

Where $w = (\xi^2 + \alpha^2)^{1/2} - \xi$, and $K_\gamma(t) = \frac{1}{2}\int_0^\infty x^{\gamma-1} \exp(-\frac{1}{2}t(x + x^{-1})\}dx$ is the modified Bessel function of the third kind.

Stein et al. (1987) treated ξ, α and γ as the parameters, and consequently their parameterization cannot be expressed and interpreted as a multiplicative Poisson random effects model and their location parameter is not the mean of y (Rigby et al., 2008). Later, Rigby et al. (2008) solved this problem by using $\xi = \mu / c$ and $w = 1/\sigma$. Thus, the final formulation of the Sichel distribution, $SI(\mu, \sigma, \gamma)$, is given by,

$$p(y \mid \mu, \sigma, \gamma) = \frac{(\mu / c)^y K_{y+\gamma}(\alpha)}{K_\gamma(1/\sigma) y!(\alpha\sigma)^{y+\gamma}} \tag{17}$$

For $y = 0,1,2,...,\infty$, where $c = \frac{K_{\gamma+1}(1/\sigma)}{K_\gamma(1/\sigma)}$ and $\alpha^2 = \sigma^{-2} + 2\mu(c\sigma)^{-1}$. The mean of y is $E[y] = \mu$ and variance is $V(y) = \mu + \mu^2[2\sigma(\gamma+1)/c + 1/c^2 - 1]$. For $\gamma = -1/2$, the Sichel distribution can be reduced to the Poisson-inverse Gaussian distribution with mean μ and variance $\mu + \mu^2\sigma$. Note that it can be also shown that the Sichel distribution converges to the NB distribution when $\sigma \to \infty$ and $\gamma > 0$. For more information about the formulation in equation (17) and parameter c, interested readers are referred to Rigby et al. (2008).

3. DATA DESCRIPTION

The characteristics of the two datasets used in this study are described in this section. The first dataset contains crash data collected on rural interstate road sections in Indiana. The second dataset consists of vehicle crash data that occurred on 4-lane undivided rural segments in Texas.

3.1. INDIANA DATA

The first dataset used for this study contains crash data collected on 338 rural interstate road sections in Indiana over a five-year period from 1995 to 1999. The data have been investigated in some previous studies (Anastasopoulos et al., 2008; Geedipally et al., 2012). Geedipally et al. (2012) used this dataset to develop NB and NB-L regression models. To be consistent with their work, the same explanatory variables are used in this study and summarized in Table 1. During the five-year study period, there were 5,737 crashes happened on 218 out of 338 highway segments, and the other 120 segments (36%) did not have any reported crashes. As shown in

Table 1, the observed crash frequency ranges from 0 to 329, and the mean frequency is 16.97 with a standard deviation of 36.30. Note that the variance to mean ratio is 77.6. For a complete list of variables in this dataset, interested readers can consult (Washington et al., 2011).

Table 1. Summary statistics of characteristics for the Indiana data.

Variable	Minimum	Maximum	Mean(SD†)	Sum
Number of crashes (5 years) X_1*	0	329	16.97 (36.30)	5737
Average daily traffic over the 5 years (ADT) X_2	9442	143422	30237.6 (28776.4)	
Minimum friction reading in the road section over the 5-year period (FRICTION) X_3	15.9	48.2	30.51 (6.67)	
Pavement surface type (1: asphalt, 0: concrete) (PAVEMENT) X_4	0	1	0.77 (0.42)	
Median width (in feet) (MW) X_5	16	194.7	66.98 (34.17)	
Presence of median barrier (1: present, 0: absent) (BARRIER) X_6	0	1	0.16 (0.37)	
Interior rumble strips (RUMBLE) X_7	0	1	0.72 (0.45)	
Segment length (in miles) (L) X_8	0.009	11.53	0.89 (1.48)	300.09

* X_1 is the serial number of variable number of crashes. † Standard deviation.

3.2. TEXAS DATA

The second crash dataset was collected at 4-lane undivided rural segments in Texas. This dataset contains crash data collected on 1499 undivided rural segments in Texas over a five-year period from 1997 to 2001. The data were collected as a part of NCHRP 17-29 research project (Lord et al., 2008). 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 over the five-year period, and a total of 4,253 crashes occurred on 946 segments. The mean of crashes was 2.84, with a variance of 32.4 and the variance to mean ratio is 11.4. Table 2 provides the summary statistics for the Texas data.

Table 2. Summary statistics of characteristics for the Texas data.

Variable	Minimum	Maximum	Mean(SD†)	Sum
Number of crashes (5 years) X_9*	0	97	2.84(5.69)	4253
Average daily traffic over the 5 years (ADT) X_{10}	42	24800	6613.61 (4010.01)	
Lane Width (LW) X_{11}	9.75	16.5	12.57(1.59)	
Total Shoulder Width (SW) X_{12}	0	40	9.96(8.02)	
Curve Density (CD) X_{13}	0	18.07	1.43 (2.35)	
Segment Length (L) (miles) X_{14}	0.1	6.28	0.55(0.67)	830.49

* X_9 is the serial number of variable number of crashes. † Standard deviation.

4. MODELING RESULTS

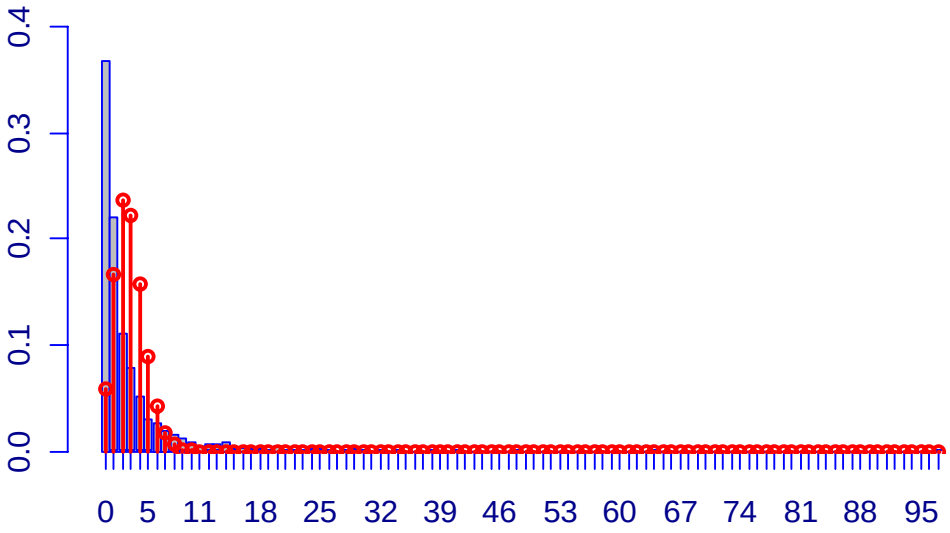
This section describes the modeling results of the NB, ZINB and SI models. The section is divided into three parts. The first part presents the goodness-of-fit analysis results for PO, NB, ZINB and SI distributions. The second and third parts provide the modeling results for the Indiana data and the Texas data, respectively. In this study, the GAMLSS models were estimated using GAMLSS package in the software R.

4.1. GOODNESS-OF-FIT COMPARISONS

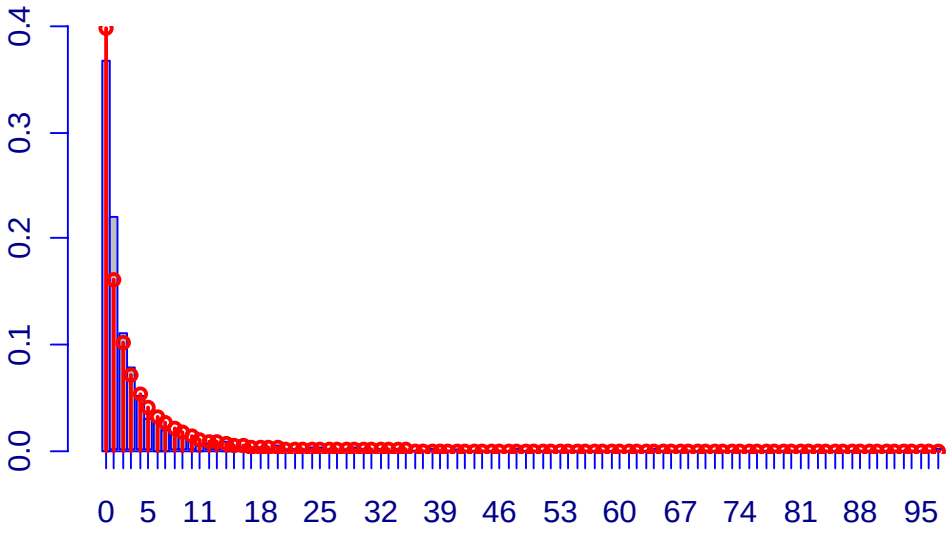
The Sichel distribution is very useful for modeling highly dispersed count data and has been shown to be the case in many studies (Sichel, 1985). To examine the applicability of the SI to crash count data, the goodness-of-fit comparisons were performed using the Texas data and the results are provided in Table 3. Compared with the PO, NB and ZINB distributions, the Sichel distribution has the smallest deviance, Akaike information criterion (AIC), and Bayesian information criterion (BIC) values. This indicated that the Sichel distribution can improve the goodness-of-fit of dispersed crash data. As shown in Figure 1, it can be observed that the trend of the crash count histogram is well captured by the Sichel distribution. However, the Sichel distribution only works very well when the count data is highly dispersed with a long tail. Based on additional examined crash datasets (not shown here), when the crash data are less dispersed with a short tail, the Sichel distribution can at least provide a performance that is equal to that of the NB distribution.

Table 3. Goodness-of-fit statistics for the Texas data.

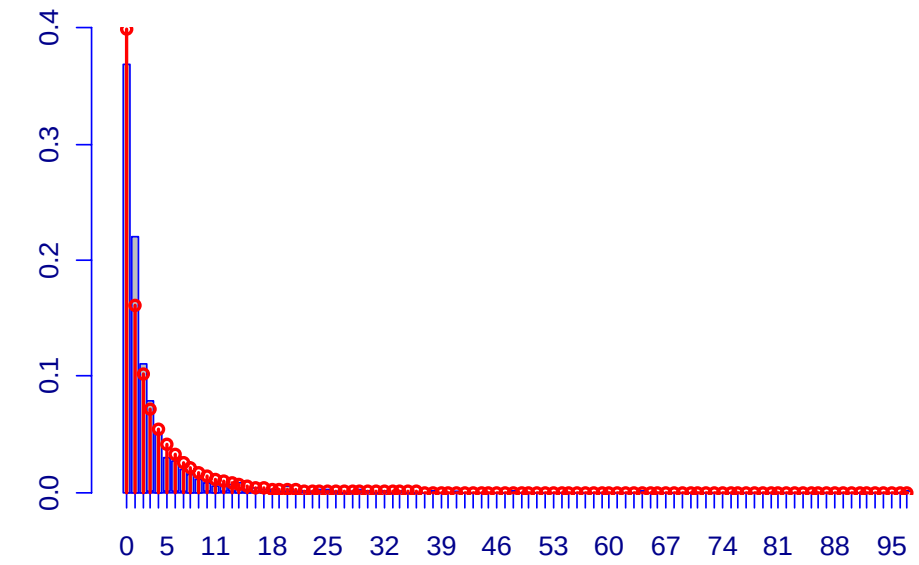
	PO	NB	ZINB	SICHEL
Deviance	11618.3	6354.79	6354.79	**6276.87**
AIC	11620.3	6358.79	6360.79	**6282.87**
BIC	11625.7	6369.41	6376.73	**6298.81**



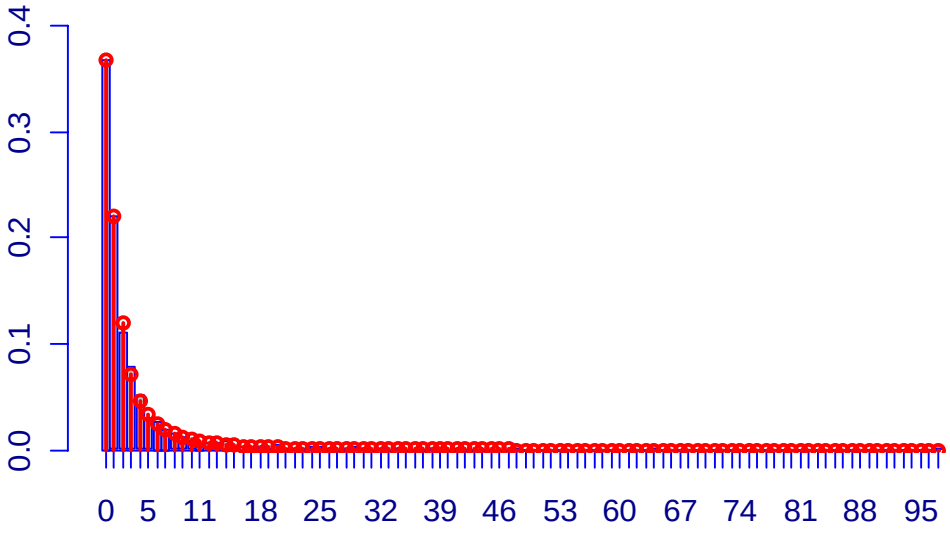
(a) Texas crash data with fitted Poisson distribution



(b) Texas crash data with fitted negative binomial distribution



(c) Texas crash data with fitted zero-inflated negative binomial distribution



(d) Texas crash data with fitted Sichel distribution

Figure 1. Texas crash data with fitted distributions.

4.2. INDIANA DATA

This section describes the modeling results for the NB, ZINB and SI GAMLSS using the Indiana data. When analyzing the crash data, we consider the segment length as an offset term, which means that the number of crashes is linearly proportional to the segment length. As mentioned previously, the GAMLSS can be seen as an extension of the GLM which allows using both parametric and nonparametric methods. In this study, the parametric linear GAMLSS is used to model the crash frequency, and a parametric linear GAMLSS has almost the same formulation as a GLM except for a few differences. The main differences are: first, for the GAMLSS, the exponential family distribution assumption for the response variable, y, is relaxed and replaced by a general distribution family; second, within the framework of the GAMLSS, each of the distribution parameters can be a function of the explanatory variables. In this study, only the mean (or location) parameter is modeled as a function of the explanatory variables.

The modeling results for the Indiana data are provided in Table 4. Although zeros only account for 36% of this crash dataset, the ZINB model was estimated in order to compare its fitting performance to that of the SI model. Note that for the NB and ZINB GAMLSS, the coefficients of variables and goodness-of-fit statistics are exactly the same as that of the NB and ZINB GLMs (the NB GLM was fitted using MASS package and the ZINB GLM was fitted using pscl package in the software R). For the SI model, the coefficient for the variable flow is smaller than that of the other two models, which means that when traffic flow increases, the estimated crash rate using the SI model increases at a slower rate than that of the NB and ZINB models. The coefficients between the models have the same sign, but their estimated values may differ significantly. For example, the coefficient of median barrier is -5.2573 for the SI model, which indicates that the crash risk will decrease significantly with the presence of a median barrier. All goodness-of-fit statistics show that the SI model has better fitting results than the NB and ZINB models. And the ZINB model provides a slightly better fit than the NB model.

Table 4. Modeling results for the Indiana data.

	NB		ZINB		SI	
Variable	Value	SE	Value	SE	Value	SE
INTERCEPT	-4.4556	1.2920	-7.8840	1.1969	-3.7672	1.1603
Ln(ADT)	0.6878	0.1202	1.0165	0.1049	0.6434	0.1028
FRICTION	-0.0267	0.0101	-0.0182	0.0093	-0.0284	0.0096
PAVEMENT	0.4296	0.1850	0.2718	0.1938	0.4710	0.1798
MW	-0.0052	0.0019	-0.0029	0.0012	-0.0076	0.0019
BARRIER	-3.0263	0.2825	-1.9751	0.1992	-5.2573	0.3227
RUMBLE	-0.3976	0.1800	-0.3801	0.1718	-0.1677	0.1621
Deviance	1884.51		1869.87		**1821.64**	
AIC	1900.51		1887.87		**1839.64**	
BIC	1931.1		1922.28		**1874.04**	

4.3. TEXAS DATA

We also applied the NB, ZINB and SI GAMLSS to the Texas data, and the segment length was considered as an offset term. The modeling results for the Texas data are provided in Table 5.

The SI model shows that crashes increase almost linearly with the increase in flow. Between the three models, the estimated coefficients of explanatory variables are very close to each other. Note that the coefficient values of the ZINB model are almost the same as that of the NB model. Moreover, the AIC and BIC values are even worse for the ZINB model and this is because AIC and BIC penalize the number of parameters in the model. Since the Texas data contain a moderate percentage of zeros (37%), the ZINB model may actually reduce to the NB model by letting the parameter v (probability of roadway section exists in the zero-crash state) equal zero. Thus, the results can be seen as the evidence that the ZINB model may not be an appropriate model for the Texas data. The goodness-of-fit statistics show that the SI model can work better than the NB and ZINB models. Compared to the fitting results for the Indiana data, the improvement of fitting performance of the SI model is not significant in this case. The possible reason is that the Texas data are less dispersed than the Indiana data and the Texas data have a relatively shorter tail (the maximum number of crashes for the Texas data is 97 while the maximum number of crashes for the Indiana data is 329). Recently, Cheng et al. (2011) fitted both the Indiana and Texas data by using a Poisson-Weibull model. Compared with the results in their study, the SI model can provide a better fit than the Poisson-Weibull model. Since the SI distribution is a mixture of Poisson distributions with three parameters and some other mixtures of Poisson distributions (Poisson-gamma and Poisson-inverse Gaussian) are special cases of the SI distribution, it is suspected that the SI distribution can outperform other common mixtures of Poisson distributions (Poisson-gamma, Conway–Maxwell–Poisson, Poisson-lognormal, Poisson-Weibull, etc.) used in highway safety studies. Obviously, more work needs to be done to verify this hypothesis.

Table 5. Modeling results for the Texas data.

	NB		ZINB		SI	
Variable	Value	SE	Value	SE	Value	SE
INTERCEPT	-7.9489	0.4233	-7.9482	0.3287	-7.9984	0.3863
Ln(ADT)	0.9749	0.0453	0.9748	0.0361	0.9926	0.0403
LW	-0.0533	0.0167	-0.0533	0.0158	-0.0600	0.0173
SW	-0.0100	0.0033	-0.0100	0.0032	-0.0100	0.0033
CD	0.0675	0.0120	0.0675	0.0110	0.0627	0.0121
Deviance	5122.772		5122.772		**5100.453**	
AIC	5134.772		5136.772		**5114.453**	
BIC	5166.647		5173.96		**5151.64**	

5. DISCUSSION

In this paper, the modeling results are very interesting and deserve further discussion. Based on the modeling results in this study, the following conclusions can be made: first, the Sichel distribution works better than the NB distribution; second, the SI GAMLSS can provide a better fit than the NB GAMLSS for highly dispersed crash datasets, at least for those two datasets. Note that the NB distribution is a special case of the SI distribution. Based on other examined datasets (although not shown in the paper), the SI GAMLSS can perform at least as well as the NB GAMLSS. Thus, for the highly dispersed crash data, transportation safety researchers are recommended to consider the SI GAMLSS.

One characteristic associated with crash data is that usually a large number of zeros are observed in the collected database. To examine the applicability of the SI model to the datasets that contain a large amount of zeros, the San Antonio crash data (about 88% of the segments in the data have zero crash) with 1903 observations used by Geedipally et al. (2012) and the Michigan data (zeros account for about 70% of the crash data) used by Qin et al. (2004) were modeled using the NB, ZINB and SI models. Although not shown here, the results indicate that SI models provide a slightly better fit than NB and ZINB models for those two datasets. For the Indiana and Michigan data, Geedipally et al. (2012) used a NB-L model and concluded that the NB-L GLM was much better than the ZINB and NB GLMs. Compared to the reported goodness-of-fit statistics in their study, it shows that the NB-L model performs better than the SI model for the crash data contain a large number of zeros or the crash data are highly dispersed. Although the NB-L model can perform better than the SI model, the SI model still offers an alternative to the traditionally-used NB models for analyzing highly dispersed datasets. Furthermore, the SI GAMLSS model could also be useful to better understand the characteristics of the dispersion, similar to the work done by Park and Lord (2009) and Anastasopoulos and Mannering (2009) on this topic.

Basically, the GAMLSS consist of four different formulations: the semi-parametric additive model (equation 6), the parametric linear model (equation 7), the non-linear semi-parametric additive model (equation 8) and the non-linear parametric model (equation 9). In this study, the parametric linear formulation was adopted and the NB, ZINB and SI GAMLSS were applied to crash frequency analysis. Previous studies (see, e.g., Xie and Zhang, 2008) have indicated that the relationship between crash frequency and explanatory variables may not be limited to linear or logarithm. Thus, to explore more flexible functional forms, the semi-parametric additive model was also examined in this paper. Taking the cubic spline as the additive terms for the mean function (Xie and Zhang, 2008), the semi-parametric NB and SI models were used to model the Indiana data. The functional form of mean function was selected as follows, $\mu = \beta_1 X_8 \exp(\sum_{i=3}^{7} \beta_i X_i + f(X_2))$, where $f(X_2)$ is the cubic spline term and X_2 is the variable Ln(ADT). The modeling results (not shown here) indicate that the fitting performance of the NB and SI models are improved when adding the additive terms. Although the semi-parametric additive model with a cubic spline term can provide better fits to the data than the parametric linear model, Lord and Mannering (2010) pointed out that the generalized additive models with spline functions are more difficult to interpret than traditional count models. Thus, the parametric linear GAMLSS model is recommended for crash data analysis unless a nonlinear relationship between logarithm of crash frequency and explanatory variables is clearly determined.

The Sichel distribution has three distribution parameters. When letting $\gamma = -\frac{1}{2}$, the Sichel distribution can reduce to a Poisson-inverse Gaussian distribution (PIG). The PIG is obtained by considering the mean of Poisson distribution as a random variable with an inverse-Gaussian probability. The probability density function of the PIG is defined as follows:

$$f(y \mid \mu, \sigma) = (\frac{2\alpha}{\pi})^{\frac{1}{2}} \frac{\mu^y e^{1/\sigma} K_{y-1/2}(\alpha)}{(\alpha\sigma)^y y!} \quad (18)$$

Where $\mu > 0$, $\sigma > 0$, $\alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu}{\sigma}$ and $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp(-\frac{1}{2}t(x + x^{-1}))\}dx$. The NB model is commonly used to model heavy-tailed count data, and most transportation safety analysts are not familiar with PIG models, as it has not used for analyzing crash data. As a special case of the SI, the PIG has a larger range of skewness than that of the NB (Zhu and Joe, 2009). The PIG model was also applied to the crash data used in this study. Although the modeling results are not given here, the authors compared the PIG model with the NB and SI models. Interestingly, for all four datasets examined in this study (Indiana data, Texas data, San Antonio data and Michigan data), the modeling results indicate that the PIG model can work better than the NB model, although not as well as the SI model. For NB and PIG models, the PIG model seems to be preferred over the NB model for modeling the crash data based on the goodness-of-fit statistics. However, it is known that there is not one single model that provides a perfect fit to a given data set, particularly in the absence of knowledge of the mechanism generating the responses. It is recommended that transportation safety analysts should fully investigate the advantages and the disadvantages of the PIG model.

The computation times for the NB, ZINB, PIG and SI models were recorded in this study. Compared to the NB model, the SI model can improve the goodness-of-fit of highly dispersed count data while the increase in computational effort is not remarkable. Thus, the computation time of the SI model is not a concern for transportation safety analysts to use this model. Even if the allowed computation time is limited, since the PIG model requires less computation time than the SI model, the PIG model might be used as an alternative to model dispersed crash data and can provide a better fit than the NB 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 SI model should be examined. Second, when the data are suspected to belong to different groups, a finite mixture of SI models should be used and compared to the finite mixture of PO and NB models (see Park and Lord, 2009, as an example). Third, an empirical Bayes modeling framework can be developed for the SI model. Finally, recent studies in transportation safety have shown that the dispersion parameter of NB models can be potentially dependent upon the explanatory variables and NB models with a varying dispersion parameter can provide better statistical fitting performance (Heydecker and Wu, 2001; El-Basyouny and Sayed, 2006) or help describing the characteristics of the dispersion (Miaou and Lord, 2003). Interestingly, the proposed GAMLSS are very flexible and allow modeling not only the mean but all other parameters (including dispersion parameter) as linear and/or nonlinear parametric and/or additive non-parametric functions of explanatory variables. Therefore, within the framework of the GAMLSS, it would also be interesting to see the results of using explanatory variables to model other distribution parameters (such as the dispersion parameter). In addition, previous studies (Geedipally et al., 2009) assumed a linear relationship between the logarithm of the dispersion parameter and explanatory variables. This linear relationship assumption might not be the best assumption. Thus, given the flexibility and strong nonlinear modeling ability of the GAMLSS, the GAMLSS can be a useful tool to explore functional forms other than the linear one that can better describe the relationship between the logarithm of the dispersion parameter and explanatory variables.

6. SUMMARY AND CONCLUSIONS

This paper has described the application of the SI GAMLSS for analyzing crash data. The proposed model was evaluated using two highly dispersed crash datasets. Traditional NB GLMs that have been proposed for analyzing highly dispersed count data are found to suffer from two methodological problems: first, NB models cannot handle high dispersion with a long tail efficiently; second, if the highly dispersed count response variable does not follow an exponential family distribution, the GLM cannot be used. The newly introduced SI GAMLSS offer the advantage of being able to model crash data with high dispersion. Moreover, in the GAMLSS, the exponential family distribution assumption is relaxed and replaced by a general family distribution. The goodness-of-fit test showed that the Sichel distribution works very well when the count data is highly dispersed with a long tail. The modeling results showed that the SI GAMLSS provided better fitting performances than the NB and ZINB GAMLSS for the crash datasets examined in this study. In conclusion, it is believed that the Sichel distribution and the SI GAMLSS may offer a viable alternative to the traditionally used NB GLMs for analyzing highly dispersed crash datasets.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Fred Mannering from Purdue University and Dr. Xiao Qin from South Dakota University for graciously providing us with the 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.

Anastasopoulos, P., Tarko, A., Mannering, F., 2008. Tobit analysis of vehicle accident rates on interstate highways. Accident Analysis and Prevention 40(2), 768-775.

Cheng, L., Geedipally, S.R., and Lord, D., 2011. Examining the Poisson-Weibull Generalized Linear Model for Analyzing Crash Data. Paper to be presented at the 91st Annual Meeting of the Transportation Research Board, Washington, D.C.

El-Basyouny K. and Sayed T., 2006. Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. In Transportation Research Record: Journal of the Transportation Research Board, No. 1950, 9–16.

Geedipally, S.R., Lord D., and Park B.-J., 2009. Analyzing Different Parameterizations of the Varying Dispersion Parameter as a Function of Segment Length. In Transportation Research Record: Journal of the Transportation Research Board, No. 2103, 108-118.

Geedipally S.R., Lord, D. and Dhavala S.S., 2012. The negative binomial-Lindley generalized linear model: characteristics and application using crash data. Accident Analysis and Prevention, In Press.

Heydecker, B.G., and 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.

Hudson, I. L., Rea, A., and Dalrymple, M. L., 2008. Climate impacts on sudden infant death syndrome: a GAMLSS approach, Proceedings of the 23rd international workshop on statistical modeling.

Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., van Schalkwyk, I., Ivan, J.N., Lyon, C., Jonsson, T., 2008. Methodology for estimating the safety performance of multilane rural highways. NCHRP Web-Only Document 126, National Cooperation Highway Research Program, Washington, D.C.

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

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., Washington, S.P., Ivan, J.N., 2007. Further notes on the application of zero inflated models in highway safety. Accident Analysis & Prevention 39(1), 53-57.

Lord, D., and 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. and 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.

Mayshkina, N.V., Mannering, F.L., 2009. Zero-state Markov switching count-data models: An empirical assessment. Accident Analysis & Prevention 41(1), 122-130.

Miaou, S.P., 1994. The Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson versus Negative Binomial Regressions. Accident Analysis and Prevention 26(4), 471–482.

Miaou, S.-P., Lord, D., 2003. Modeling Traffic-Flow Relationships at Signalized Intersections: Dispersion Parameter, Functional Form and Bayes vs Empirical Bayes. In Transportation Research Record: Journal of the Transportation Research Board, No. 1840, 31-40.

Park, B.J. and Lord D., 2009. Application of Finite Mixture Models for Vehicle Crash Data Analysis. Accident Analysis and Prevention 41(4), 683–691.

Poch, M and Mannering, F. L., (1996). Negative binomial analysis of intersection accident frequency. Journal of Transportation Engineering 122, 105-113.

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

Rigby, R.A. and Stasinopoulos, D.M., 2005. Generalized Additive Models for Location, Scale and Shape. Applied Statistics 54, 507–554.

Rigby, R.A., Stasinopoulos, D.M., Akantziliotou, C., 2008. A framework for modeling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. Computational Statistics and Data Analysis 53, 381–393.

Sichel, H. S., 1982. Repeat-buying and the generalized inverse Gaussian-Poisson distribution. Applied Statsitics 31, 193–204.

Sichel, H. S., 1985. A bibliometric distribution which really works. Journal of the American society for information science. 36(5), 314-321.

Stasinopoulos, D.M. and Rigby, R.A., 2007. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. Journal of Statistical Software 23, 7.

Stein, G., Zucchini, W., and Juritz J., 1987. Parameter Estimation for the Sichel Distribution and Its Multivariate Extension. Journal of the American Statistical Association 82 (399), 938–944.

Visser, G.H., Eilers, P.H., Elferink-Stinkens, P.M., Merkus, H.M., and Wit, J.M., 2009. New Dutch reference curves for birthweight by gestational age. Early Human Development, 85(12), 737–744.

Washington, S., Karlaftis, M. and Mannering, F., 2011. Statistical and Econometric Methods for Transportation Data Analysis. Second edition, Chapman and Hall/CRC, Boca Raton, FL.

Xie, Y. and Zhang, Y., 2008. Crash Frequency Analysis with Generalized Additive Models. In Transportation Research Record: Journal of the Transportation Research Board, No. 2061, 39–45.

Zhu R. and Joe H., 2009. Modelling heavy-tailed count data using a generalized Poisson-inverse Gaussian family. Statistics and Probability Letters 79, 1695-1703.