Bayesian Poisson Hierarchical Models for Crash Data Analysis: Investigating the Impact of Model Choice on Site-Specific Predictions

S. Hadi Khazraee, Ph.D.*
Safety Research Scientist
Uber Technologies, Inc.
San Francisco, CA, 94103
Tel. (866) 576-1039
Email: hadi@uber.edu

Valen Johnson, Ph.D.
Professor and Department Head
Department of Statistics
Texas A&M University
College Station, TX, 77843-3143
Tel. (979) 862-7583
Email: vjohnson@stat.tamu.edu

And

Dominique Lord, Ph.D.
Professor
Zachry Department of Civil Engineering
Texas A&M University
College Station, TX, 77843-3136
Tel. (979) 458-3949
Email: d-lord@tamu.edu

January 20, 2018

*Corresponding author

ABSTRACT

The Poisson-gamma (PG) and Poisson-lognormal (PLN) regression models are among the most popular means for motor vehicle crash data analysis. Both models belong to the Poisson-hierarchical family of models. While numerous studies have compared the overall performance of alternative Bayesian Poisson-hierarchical models, little research has addressed the impact of model choice on the expected crash frequency prediction at individual sites. This paper sought to examine whether there are any trends among candidate models predictions e.g., that an alternative model's prediction for sites with certain conditions tends to be higher (or lower) than that from another model. In addition to the PG and PLN models, this research formulated a new member of the Poisson-hierarchical family of models: the Poisson-inverse gamma (PIGam). Three field datasets (from Texas, Michigan and Indiana) covering a wide range of over-dispersion characteristics were selected for analysis.

This study demonstrated that the model choice can be critical when the calibrated models are used for prediction at new sites, especially when the data are highly over-dispersed. For all three datasets, the PIGam model would predict higher expected crash frequencies than would the PLN and PG models, in order, indicating a clear link between the models predictions and the shape of their mixing distributions (i.e., gamma, lognormal, and inverse gamma, respectively). The thicker tail of the PIGam and PLN models (in order) may provide an advantage when the data are highly over-dispersed. The analysis results also illustrated a major deficiency of the Deviance Information Criterion (DIC) in comparing the goodness-of-fit of hierarchical models; models with drastically different set of coefficients (and thus predictions for new sites) may yield similar DIC values, because the DIC only accounts for the parameters in the lowest (observation) level of the hierarchy and ignores the higher levels (regression coefficients).

Key Words: Poisson hierarchical, Bayesian model, site-specific prediction, model choice, Poisson-inverse gamma

INTRODUCTION

Over the past three decades, highway safety researchers have been concerned with developing statistical models to predict motor vehicle crashes. The purpose of such models is to relate the frequency of crashes to the traffic, geometrical, and/or environmental characteristics of highway entities. A mass of published work in the field of crash data modeling has focused on evaluating the application of countless statistical models by comparing their goodness-of-fit (GOF) to field data sets. These studies seek to answer the question of "which model performs better?" Measuring the GOF of statistical models is not a straightforward task; researchers use many different methods for assessing GOF. These methods can yield contradictory results in determining which model performs best.

Even if there was a consensus among researchers in employing a unique method for GOF assessment, the performance of different models would depend on the characteristics of the data set to which they are fitted. Based on the model structure and application to a limited number of datasets, some studies have suggested that certain models are expected to perform better than other specified models for data with certain characteristics. For example, Geedipally et al. (2012) noted that the heavy-tailed negative binomial-Lindley (NB-L) distribution is expected to outperform the negative binomial model for datasets with abundant zero crash observations and a long/heavy tail. While rules of thumb like the aforementioned may be valid, it is difficult to predict with certainty which model performs better before the models are actually fitted to the data. Therefore, one must be very cautious when making conclusions of the type "one model is better than another" based on a comparison of the models' GOF to a few data sets.

GOF analyses in the literature of crash data modeling have been focused on the overall fit of a model to an entire dataset. An important question that has often been overlooked is "how are

the site-specific predicted crash frequencies affected by model choice?" Can we find a trend in the difference between the predicted crash frequencies of different models; e.g., that a model's prediction for sites with certain conditions is higher (or lower) than that from a different model? This paper conducts such a microscopic analysis for three hierarchical-Poisson regression models, including the two most popular models in crash data analysis, the Poisson-gamma and Poisson-lognormal models. We also examine a new model, the Poisson-Inverse Gamma (PIGam) model; note: the PIGam should not be confused with the Poisson-Inverse Gaussian (PIG) model, as detailed in Zha et al. (2016). The PIGam model was formulated and added to this study to investigate the potential benefit of having a long/heavy-tailed mixing distribution (i.e., inverse Gamma) in handling data sets with unusually high crash count observations. Three crash data sets with a range of dispersion characteristics were used to compare the performance of the candidate models.

This research is concerned with the Bayesian estimation of Poisson-hierarchical models. The Bayesian structure of the models and the practical interpretation of the model parameters are described in detail and distinguished from the traditional frequentist models. The models' predictions are compared at two levels: a) for sites in the dataset where an observed crash count is available, and b) for hypothetical new sites without an observed crash count. Rather than considering several candidate models and employing the best one (based on GOF comparisons) for prediction, this study aims to add depth to our understanding of how the magnitude of predictions is influenced if each of the three considered models is selected.

BACKGROUND

Because crashes are random events and typically independent of one another, the Poisson distribution is the intuitive means to describe the randomness in crash counts. In practice, however, the Poisson is rarely an appropriate distribution because the crash data are often over-dispersed (Poch and Mannering, 1996; Hauer, 2001; Mitra and Washington, 2006), meaning that the conditional variance of observed crash counts is greater than the mean, whereas the Poisson distribution is equi-dispersed (i.e., the mean equals the variance).

The over-dispersion in crash count data is mainly attributed to the heterogeneity among the different sites (highway segments, intersections, etc.) where crash data is collected (Hauer, 2001; Washington et al., 2010; Lord and Mannering, 2010; Mannering and Bhat, 2014; Mannering et al., 2016). Although the effect of some important factors (such as traffic volume) on the expected number (mean) of crashes is usually accounted for by a regression model, some degree of heterogeneity is always believed to remain unobserved due to factors either unknown or known but hard to collect and include in the model. In addition, Lord et al. (2005) showed that the theory behind the fundamental crash process itself gives rise to over-dispersion. They argued that a crash count is the sum of a series of Bernoulli random variables, with each trial representing a vehicle/driver going through a site with an un-equal probability of success (crash). These assumptions lead to an over-dispersed distribution of crash counts.

To accommodate for possible over-dispersion in crash data, researchers have extensively used mixed-Poisson models, where crash counts are assumed to have a Poisson distribution with a variable mean that follows an underlying distribution often referred to as the mixing distribution. Mixed-Poisson models are indeed hierarchical, where at the lowest level of hierarchy, conditional on the mean, the observed crash counts are mutually independent and

Poisson-distributed while at the second level, the unobservable mean of crash counts varies across sites with an assumed probability distribution. Importantly, hierarchical Poisson models particularly suit the theoretical nature of crash frequency data. Such conceptual suitability, often referred to as "goodness-of-logic", has been sacrificed for a better statistical goodness-of-fit (GOF) in a number of regression models proposed for crash data analysis (Miaou and Lord, 2003).

Two types of mixed-Poisson models have gained extensive attention among highway safety researchers: Poisson-gamma, and Poisson-lognormal:

1) *Poisson-Gamma (PG)*

When the Poisson parameter is assumed to have a gamma probability distribution, the mixed-Poisson distribution will have a closed form probability density function (pdf) which turns out to be that of the Negative Binomial (NB). The over-dispersion parameter (α) of the NB distribution captures the unmodeled heterogeneity (Miaou and Lord, 2003). The value of this parameter defines the relationship between the distribution mean and variance as $Var(y_i) = E[y_i] + \alpha E[y_i]^2$. The simplicity of model fitting and parameter interpretation has made the Poisson-gamma/NB the most popular model in crash data analysis. In its classical applications, modelers estimated the parameters using the Maximum Likelihood Estimate (MLE) method (e.g. Maycock and Hall, 1984; Hauer et al., 1989; Bonneson and McCoy, 1993; Vogt and Bared, 1998). Recently, the Poisson-gamma models have also been estimated using Bayesian methods (Schluter et al., 1997; Miaou and Lord, 2003; Miaou and Song, 2005; Lord and Miranda-Moreno, 2008).

2) *Poisson-Lognormal (PLN)*

The Poisson-lognormal model results when the Poisson parameter is assumed to follow a lognormal distribution. Unlike the Poisson-gamma model, the marginal distribution of the Poisson-lognormal model does not have a closed form and the model parameters cannot be estimated analytically using the Maximum Likelihood Estimates (MLE) method. Despite the availability of Hinde's (1982) numerical integration method to approximate the MLE parameters, the Poisson-lognormal safety performance models in the published studies have all been estimated using the Bayesian approach. The Poisson-lognormal model is potentially more flexible than the Poisson-gamma (Lord and Mannering, 2010) and the model has become increasingly popular in crash data analysis over the past few years (Miaou et al., 2003; Aguero-Valverde and Jovanis, 2008; Aguero-Valverde, 2013). Highway safety researchers have also used multivariate Poisson-lognormal models to jointly model crash frequency by severity while accounting for the correlation among different severity levels (Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009).

The MLE method has traditionally been used to estimate safety performance model parameters. Nonetheless, the MLE method cannot be used straightforwardly when the marginal likelihood function is difficult to characterize, as in the case of the Poisson-lognormal model. Nonetheless, the development of such models and others with complex functional forms was greatly facilitated after the rediscovery of Markov Chain Monte Carlo (MCMC) simulation methods (Tanner and Wong, 1987; Gelfand and Smith, 1990) for model estimation from the Bayesian perspective. Unlike the traditional frequentist approach, the Bayesian approach to statistics aims to estimate the probability distribution of model parameters using the information in the observed data as well as the prior knowledge about model parameters. The drastic growth

in the processing speed of personal computers and availability of Bayesian software programs such as WinBUGS (Spiegelhalter et al., 2003) and MLwiN (Yang et al., 1999) has helped significantly in the increasing popularity of Bayesian models.

The fully (hierarchical) Bayesian (FB) estimation of safety performance models has been explored only in the last two decades (Schluter et al., 1997; Davis and Yang, 2001; Miaou and Lord, 2003; Carriquiry and Pawlovich, 2004; Miaou and Song, 2005; Park and Lord, 2007; Daziano et al., 2013). Prior to FB models, however, the empirical Bayes (EB) method was introduced into the highway safety literature (Hauer and Persaud, 1983; Hauer, 1986, 1992). The EB method basically uses the Bayes rule to combine the information from some reference population (or results of a regression model) with the observed crash counts at a certain site to estimate the expected (long-term) mean of crashes. There is extensive documentation and application of the EB method in highway safety, especially in before-after studies to estimate the effect of safety countermeasures (Hauer, 1997; Persaud, 1998; Harwood et al., 2002; Persaud and Lyon, 2007; Fitzpatrick and Park, 2009). EB methods have also been used for identification of hotspot locations (Persaud et al., 1999; Heydecker and Wu, 2001; Miranda-Moreno et al., 2005; Lord and Park, 2008).

The FB approach has a fundamental advantage over the MLE and EB methods; it takes into account the uncertainty associated with model parameters and provides exact measures of uncertainty (Miaou and Lord, 2003). In the MCMC method, this is carried out by sampling from the posterior distribution of model parameters. The MLE and EB methods, on the other hand, ignore this uncertainty and thus overestimate the model precision (Carriquiry and Pawlovich, 2004; Goldstein, 2010; Park et al., 2010). This advantage of the FB approach is especially important when the sample size is relatively small (Miaou and Lord, 2003).

The performance of safety prediction models is often evaluated and compared using statistical GOF measures. In the Bayesian paradigm, measures such as the Bayes factor, Bayesian information criterion (BIC), Watanabe-Akaike information criterion (WAIC), and others are used to assess and compare the performance of fitted models (Gelman et al. 2013). However, the GOF of Bayesian hierarchical regression models is most commonly compared using the deviance information criterion (DIC). Proposed by Spiegelhalter et al. (2002), the DIC is a Bayesian generalization of the Akaike Information Criterion (AIC), and is defined as:

$$DIC = \bar{D} + P_D \quad (1)$$

where $\bar{D} = \mathrm{E}(-2\log(\Pr(y|\theta))$ is the expectation of the model deviance under the posterior distribution of the model parameters (collectively denoted as θ), and P_D is the effective number of parameters, defined as:

$$P_D = \bar{D} - D(\bar{\theta}) \quad (2)$$

where $D(\bar{\theta})$ is the deviance under the posterior expectation of parameters.

$\bar{D}$ is a classical estimate of fit; a smaller $\bar{D}$ indicates a better fit as it corresponds to a greater log-likelihood. P_D is indeed a penalty for model complexity and ensures a fair comparison between competing models with different degrees of complexity. DIC is particularly useful when the posterior distributions of model parameters are obtained via MCMC simulation, which is the case for the models in this study.

It should be pointed out that the DIC can be very sensitive to the structure of the hierarchical Bayesian model. Geedipally et al. (2013) noted that when different hierarchical Bayesian models are compared, the DIC can be used as a model selection criterion only if the

likelihood structure remains the same across all the models under consideration. This requirement was met for the model comparisons performed in this work.

MODEL SPECIFICATION

This section reviews the characteristcs of the models evaluated in this paper, which all belong to the Bayesian Poisson hierarchical family of models. First, the common structure of these models is described and compared with the traditional frequentist models (such as the negative binomial). Next, the alternative Bayesian Poisson hierarchical models selected for this study are listed and their mathematical formulation is presented. Finally, the fundamental properties of the models' mixing distribution (i.e., the distinctive element of the alternative models) are characterized.

Common Model Structure

Let y_i denote the number of crashes observed at the i'th site (road segment, intersection, etc.) during the study period. In Poisson hierarchical models, y_i's, when conditional on their mean λ_i, are assumed to be Poisson distributed:

$$y_i \mid \lambda_i = \text{Poisson}(\lambda_i) \qquad i = 1, 2, \ldots, n \qquad (3)$$

In the second level of hierarchy, the mean of the Poisson distribution is variable according to an underlying mixing distribution: gamma, lognormal, or inverse gamma. In the most common type of crash prediction regression models, the site-specfic mean of the mixing distribution, μ_i, is modeled as a log-linear function of the prevailing traffic, geometric, and/or environmental variables:

$$\text{E}(\lambda_i \mid \beta) = \mu_i = \exp(X_i \beta) \qquad i = 1, 2, \ldots, n \qquad (4)$$

where X_i is the vector of covariates for site i, and β is the vector of unknown coefficients. This functional formulation is adopted in this paper.

At this point, it is critical to emphasize the distinction between the frequestist and Bayesian approaches to hierarchical regression modeling. The frequentist approach is based on the assumption that the model parameters are fixed and the observed data is a repeatable random sample. Therefore, λ_i's are constant values from a mixing distribution with mean μ_i. Once the coefficients are estimated based on the data, the expected crash frequency at site i is simply calculated as:

$$\mathrm{E}(y_i) = \mathrm{E}(\lambda_i | \beta) = \exp(X_i \beta) \qquad i = 1, 2, \ldots, n \qquad (5)$$

In contrast, the Bayesian paradigm views the parameters probabilistically. Every parameter is assumed to be drawn from a specified *prior* distribution (representing the modeler's prior knowledge), which is combined with the information in the data to produce *posterior* distributions using Bayes rule. In the hierarchical framework under study here, the λ_i's are assumed *a priori* to follow an undelying mixing distribution $p(\lambda_i)$ with mean $\exp(X_i\beta)$, where the hyper-parameter β itself follows a prior distribution (a.k.a. *hyper-prior*). However, the posterior distribution of λ_i (for every i) neither follows the specified mixing distribution (gamma, lognormal, or inverse gamma) nor has an expectation necessarily equal to μ_i.

In the Bayesian model formulation, it is critical to understand the difference between μ_i and λ_i. μ_i is the unconditional mean that depends only on the covariates. This parameter captures the effects of site characteristics on the expected crash frequency through the regression model and is therefore referred to in this paper as the "regression mean." The regression mean can be interpreted as the expected crash frequency for sites with the same covariate values (i.e., traffic, geometric, and/or environmental conditions) as those of site i but without an observed crash

count. λ_i, on the other hand, is the site-specific expected crash frequency after μ_i is adjusted by the observed crash frequency through the Bayesian model.

In the Bayesian hierarchical models in this paper, each λ_i represents a model parameter. Therefore, for a dataset with *n* sites, a Bayesian Poisson hierarchical model will have *n* parameters more than its frequentist counterpart. However, these parameters are not regarded as being independent, but are instead assumed to be drawn from a common distribution. As the variance of this common distribution approaches 0, the frequentist, fixed-effect model is obtained as a limiting case. As the case is made in this paper, the greater inherent complexity of the Bayesian approach demands greater attention when the performance of alternative Bayesian models are evaluated and compared.

Alternative Models

This research thoroughly evaluates and compares the crash frequency predictions of the following hierarchical Poisson regression models using a full Bayesian approach:

1) Poisson-Gamma (PG)
2) Poisson-Lognormal (PLN)
3) Poisson-Inverse Gamma (PIGam)

The first two models are commonly used in highway safety analyses, whereas the latter is new and its appropriateness for crash data modeling is to be examined. As mentioned before, the Poisson-Inverse Gamma model is abbreviated as “PIGam” to avoid confusion with the Poisson-Inverse Gaussian (PIG) model, whose application to crash data has been investigated by Zha et al. (2016).

Every mixing distribution selected for this study has two parameters. Below, these distributions are reparametrized in terms of their mean (μ_i) and a remaining hyper-parameter (shape or scale) to structure the regression models:

1) Poisson-Gamma (PG): the Poisson parameter (mean) follows a gamma distribution with shape parameter φ and scale parameter θ_i:

$$\Pr(\lambda_i|\varphi,\theta_i)=\frac{1}{\Gamma(\varphi)\theta_i^{\varphi}}\lambda_i^{\varphi-1}\exp(\frac{-\lambda_i}{\theta_i}) \qquad \varphi,\ \theta_i>0 \qquad (6)$$

$$\mathrm{E}(\lambda_i|\beta)=\mu_i=\varphi\theta_i=\exp(X_i\beta) \qquad (7)$$

$$\Pr(\lambda_i|\varphi,X_i,\beta)=\frac{\varphi^{\varphi}}{\Gamma(\varphi)\exp(\varphi.(X_i\beta))}\lambda_i^{\varphi-1}\exp(\frac{-\lambda_i\varphi}{\exp(X_i\beta)}) \qquad (8)$$

2) Poisson-Lognormal (PLN): the Poisson parameter follows a lognormal distribution with location parameter υ_i and shape parameter σ^2:

$$\Pr(\lambda_i|\upsilon_i,\sigma^2)=\frac{1}{\lambda_i\sigma\sqrt{2\pi}}\exp(-\frac{(\ln(\lambda_i)-\upsilon_i)^2}{2\sigma^2}) \qquad \sigma^2>0 \qquad (9)$$

$$\mathrm{E}(\lambda_i|\beta)=\mu_i=\exp(\upsilon_i+\sigma^2/2)=\exp(X_i\beta) \qquad (10)$$

$$\Pr(\lambda_i|\sigma^2,X_i,\beta)=\frac{1}{\lambda_i\sigma\sqrt{2\pi}}\exp(-\frac{(\ln(\lambda_i)-X_i\beta+\sigma^2/2)^2}{2\sigma^2}) \qquad (11)$$

3) Poisson-Inverse Gamma (PIGam): The Poisson parameter follows an inverse gamma distribution with shape parameter φ and scale parameter θ_i:

$$\Pr(\lambda_i|\varphi,\theta_i)=\frac{\theta_i^{\varphi}}{\Gamma(\varphi)}\lambda_i^{-\varphi-1}\exp(\frac{-\theta_i}{\lambda_i}) \qquad \varphi,\ \theta_i>0 \qquad (12)$$

$$\mathrm{E}(\lambda_i|\beta)=\mu_i=\frac{\theta_i}{\varphi-1}=\exp(X_i\beta) \qquad \text{for } \varphi>1 \qquad (13)$$

$$\Pr(\lambda_i|\varphi, X_i, \beta) = \frac{[(\varphi-1)\exp(X_i\beta)]^{\varphi}}{\Gamma(\varphi)}\lambda_i^{-\varphi-1}\exp(\frac{-(\varphi-1)\exp(X_i\beta)}{\lambda_i}) \quad \text{for } \varphi > 1 \qquad (14)$$

To distinguish between the parameters of the PG and PIGam models, the shape parameters are augmented by subscripts denoting the model name: φ_{PG} for the Poisson-gamma and φ_{PIGam} for the Poisson-inverse gamma.

The similar structure and number of parameters provides for a reasonable comparison between the predictions of the three models. No model can be presumed to take advantage of a higher number of parameters for its flexibility to fit the data. Site-specific λ_i's and μ_i's under each model can be directly and meaningfully compared, as carried out later in this paper.

Mixing Distribution Properties

Given the common structure of the alternative Poisson hierarchical models, the distinctive performance of the models is attributable to the fundamental characteristics of their mixing distributions (i.e., prior distribution for λ_i's).

Figure 1 compares the shape of the gamma, lognormal, and inverse gamma probability distribution functions (pdf's) for four different combinations of mean and variance. The inverse gamma distribution is the distribution of the reciprocal of a variable that follows a gamma distribution. Since the gamma distribution is light-tailed, the inverse of a gamma-distributed variable has a very low probability in the vicinity of zero (similar to the NB-L, see Shirazi et al., 2016). It is interesting to note that the lognormal distribution lies between the gamma and inverse gamma distributions almost in every important aspect including the probability density in vicinity of zero, probability density for very large values (i.e., tail thickness), mode, probability at mode, skewness, etc. As we move from the gamma distribution to lognormal and then to inverse gamma, near-zero values for expected crash frequencies (λ_i's) become less likely, and extremely high values for the expected crash frequency become more likely.

Unlike the gamma distribution, the lognormal and inverse gamma distributions are classified as heavy-tailed because their right tails are not exponentially bounded. Moreover, the inverse gamma distribution has a thicker tail than the lognormal distribution and, therefore, the lognormal distribution lies between the gamma and inverse-gamma distribution in terms of the tail thickness. The thicker tail, which is representative of greater probabilities for higher expected crash frequencies, can theoretically be beneficial when modeling data with occasional unusually high crash count observations. Please note that Figure 1 cannot demonstrate the aforementioned relation between the tail thickness of the distributions; as $x \rightarrow \infty$ the tails become too thin and their different thicknesses are unrecognizable.

This notion is evident in Figure 1, which compares the shape of the gamma, lognormal, and inverse gamma probability distribution functions (pdf's) for four different combinations of mean and variance. It is interesting to note that the lognormal distribution lies between the gamma and inverse gamma distributions almost in every important aspect including the probability density in vicinity of zero, probability density for very large values (i.e., tail thickness), mode, probability at mode, skewness, etc. As we move from the gamma distribution to lognormal and then to inverse gamma, near-zero values for expected crash frequencies (λ_i's) become less likely, and extremely high values for the expected crash frequency become more likely.

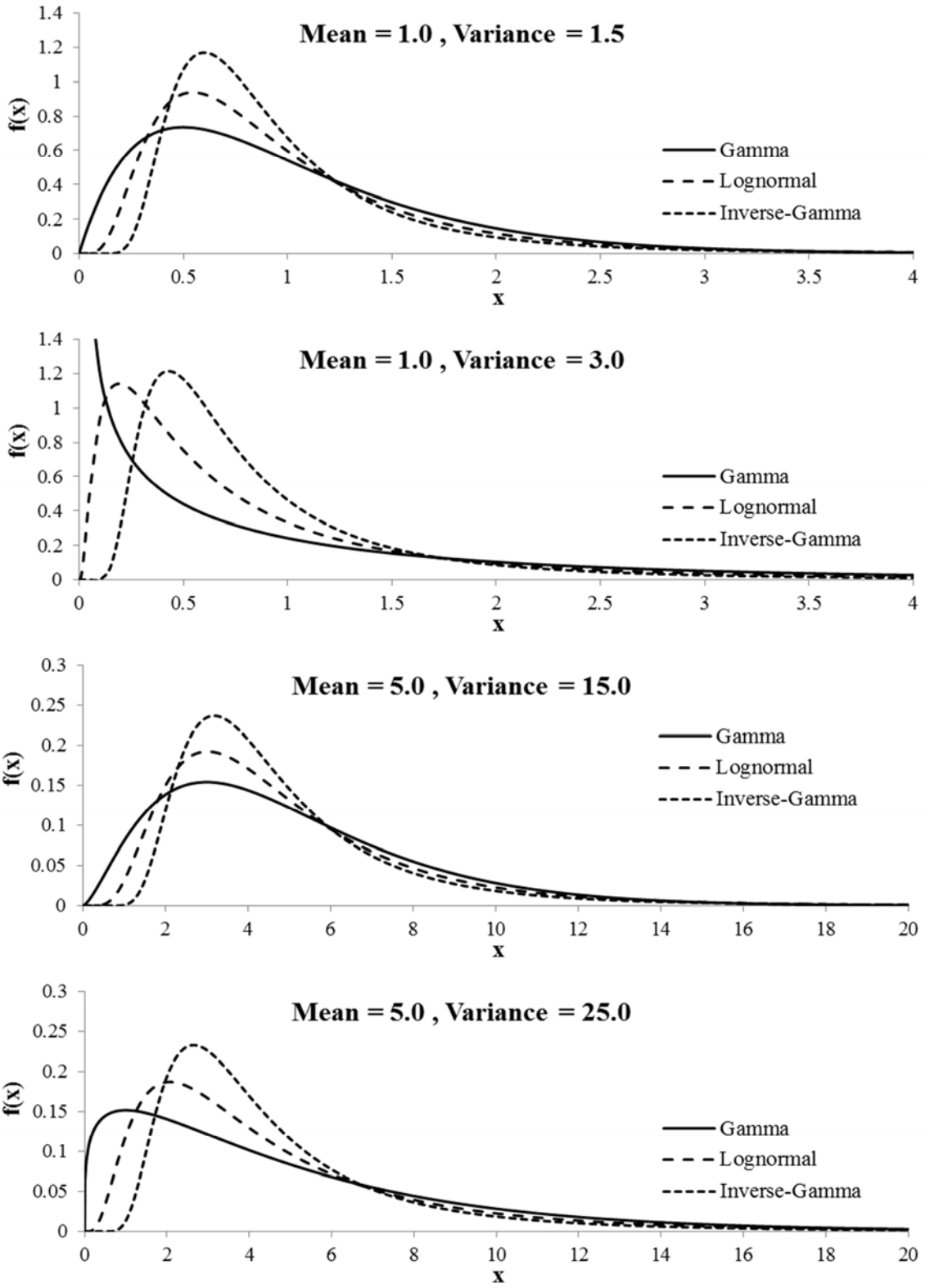

Figure 1. Probability distribution function of the gamma, lognormal, and inverse gamma distributions for different mean-variance combinations

DATA DESCRIPTION

Three relatively large crash data sets from highway segments in Texas, Michigan, and Indiana are used to compare the predictions of the selected models. Table 1 presents the summary statistics of the variables in these datasets. The source and main characteristics of the datasets are described below.

The Texas dataset is a randomly selected sample of 500 rural four-lane undivided highway segments in Texas. The crash counts indicate the total number of crashes observed during a five-year period from 1997 to 2001. The original dataset included a total of 1,499 segments and was used by Lord et al. (2008) to develop safety prediction models for inclusion in the first edition of the Highway Safety Manual (HSM) (AASHTO, 2011).

The Michigan dataset contains the single-vehicle crashes that occurred on rural two-lane highways in Michigan in 2006. Similar to the Texas dataset, a randomly selected sample of 500 highway segments was taken from the original data and used for analysis. The data were collected for the Federal Highway Administrations's (FHWA) Highway Safety Informations System (HSIS) and originally included a total of 33,970 segments. In this dataset, about 67% of segments experienced no crashes in 2006. The selection of this dataset provides an opportunity to examine the performance of different models with data characterized by a large number of zeros, which is specifically common in road safety analyses.

The Indiana dataset contains data collected for a five-year period (1995-1999) from 338 rural interstate freeway segments in Indiana. The Indiana data is the most highly over-dispersed dataset considered in this study, which includes a relatively large proportion of zero crash counts (about 36%) along with very high crash counts observed at some other sites.

Table 1. Summary statistics of the data sets in this study

Variable	Min.	Max.	Average	(std. dev)	Total
Texas Data					
Crash count	0	41	2.83	(5.06)	1415
Segment length (mi)	0.10	4.54	0.54	(0.61)	270.87
AADT*(veh/day)	420	23800	6548.5	(41.1)	--
Lane width (ft) (LW)	10	16.5	12.63	(1.57)	--
Shoulder width† (ft) (SW)	0	32	10.15	(7.80)	--
Curve density (curve/mi) (CD)	0	16.95	1.37	(2.30)	--
Michigan Data					
Crash count	0	30	0.77	(2.16)	383
Segment length (mi)	0.001	4.21	0.19	(0.35)	93.61
AADT(veh/day)	270	19990	4699.0	(3487.2)	--
Right shoulder width (ft) (RSW)	0	12	8.50	(2.80)	--
Lane width (ft)	9	14	11.23	(0.74)	--
Speed limit (mph) (SL)	25	55	53.01	(5.71)	--
Left shoulder width (ft) (LSW)	0	12	8.46	(2.87)	--
Indiana Data					
Crash count	0	329	16.97	(36.30)	5737
Segment length (mi)	0.009	11.53	0.89	(1.48)	300.09
AADT(veh/day)	9,442	143422	30237.6	(28776.4)	--
Minimum friction reading in the road segment (FRICTION)	15.9	48.2	30.51	(6.67)	--
Median width (ft) (MW)	16	194.7	66.98	(34.17)	--
Pavement surface type (PAVEMENT) (1 if asphalt, 0 if concrete)	0	1	0.77	(0.42)	--
Presence of median barrier (BARRIER) (1 if present, 0 if absent)	0	1	0.16	(0.37)	--
Presence of interior rumble strips (RUMBLE) (1 if present, 0 if absent)	0	1	0.72	(0.45)	--

*annual average daily traffic †right shoulder width + left shoulder width

Functional Form

The following functional form is adopted for all models and datasets in this study:

$$\mu_i = \beta_0 L F_i^{\beta_1} \exp(\sum_{j=2}^{p} \beta_j x_j) \quad (15)$$

where F_i is the AADT through the highway segment, L is the segment length, and x_j's are the additional covariates listed in Table 1. In this functional form, the segment length is treated as an offset (i.e., its power is kept fixed at 1) because theoretically, the crash risk is expected to have a linear relationship with the segment length.

Given the selected functional form, the datasets cover a relatively wide range of dispersion characteristics. The overdispersion parameter ($\alpha = 1/\varphi$) of the traditional negative binomial (NB) model (estimated using the MLE method) is considered as a rough measure of conditional dispersion in the data (i.e., dispersion of the crash counts conditional on the modeled mean). The NB model overdispersion parameter (α_{NB}) is 0.361, 0.496, and 0.888 for the Texas, Michigan, and Indiana data, respectively.

MODEL ESTIMATION

This section describes the assumptions used to fit the models to the datasets and presents the results of parameter estimation.

Choice of hyper-priors

A critical step in Bayesian modeling is the selection of hyper-prior distributions (Gelman et al. 2013; Lee 2012). In order to control for the potential influence of prior information on the performance of the alternative models, non-informative uniform distributions are considered for the models' hyperparameters. However, for large data sets such as the ones selected here, the

choice of hyper-priors will not make a significant change in the parameter estimates. The following prior distributions are used:

- Regression coefficients (all models): $\beta_j \sim$ Uniform $(-\infty,+\infty)$ for $j = 0,1, 2, \ldots, p$
- φ_{PG} (Poisson-gamma): $\varphi_{\text{PG}} \sim$ Uniform $(0,+\infty)$
- φ_{PIGam} (Poisson-inverse gamma): $\varphi_{\text{PIGam}} \sim$ Uniform $(1,+\infty)$
- σ (Poisson-lognormal): $\sigma \sim$ Uniform $(0,+\infty)$

MCMC Simulation for Posterior Inference

Posterior distribution of model parameters were estimated using basic Markov Chain Monte Carlo (MCMC) simulation methods. Sampling from all parameters was carried out using the Metropolis-Hastings algorithm. The first 20,000 samples were discarded to diminish the influence of the starting values, and the posterior distribution of each parameter was constructed using the next 1,000,000 MCMC iterations. Gelman's $\hat{R}$ factor (Gelman et al., 2013) and visual inspection were utilized to ensure sufficient convergence.

Parameter Estimation Results

Table 2-4 summarize the posterior inference with the mean and standard deviation of the parameters. The results indicate that the difference between the models' coefficients intensifies as the over-dispersion in the data increases (from the Texas data to the Indiana data). The number of model covariates (increasing from the Texas models to the Indiana models) may also be a contributing factor. Variation in coefficient estimates across alternative models (especially for the Indiana dataset) may result in distinctive predictions by the alternative models, as demonstrated in the next section.

Table 2. Posterior Estimates of Model Parameters for Texas Data

Parameter	Respective Variable	**PG** mean (std dev.)	**PLN** mean (std dev.)	**PIGam** mean (std dev.)
$\ln(\beta_0)$	Intercept	-7.3986 (0.6683)	-7.4276 (0.7069)	-7.4725 (0.7413)
β_1	AADT	0.9636 (0.0692)	0.9752 (0.0737)	0.9812 (0.0753)
β_2	LW	-0.0869 (0.0290)	-0.0913 (0.0300)	-0.0909 (0.0296)
β_3	SW	-0.0122 (0.0057)	-0.0131 (0.0058)	-0.0132 (0.0056)
β_4	CD	0.1040 (0.0207)	0.1038 (0.0209)	0.1022 (0.0203)
φ_{PG}		2.8169 (0.4746)		
σ			0.6060 (0.0502)	
φ_{PIGam}				3.9323 (0.5741)

Table 3. Posterior Estimates of Model Parameters for Michigan Data

Parameter	Respective Variable	**PG** mean (std dev.)	**PLN** mean (std dev.)	**PIGam** mean (std dev.)
$\ln(\beta_0)$	Intercept	-7.3166 (1.6729)	-7.5750 (1.6709)	-7.3671 (1.7026)
β_1	AADT	0.5521 (0.0949)	0.5477 (0.0982)	0.5430 (0.0903)
β_2	RSW	0.0733 (0.0695)	0.0710 (0.0755)	0.0654 (0.0731)
β_3	LW	0.2366 (0.1009)	0.2569 (0.1068)	0.2553 (0.1024)
β_4	SL	0.0193 (0.0192)	0.0204 (0.0201)	0.0165 (0.0211)
β_5	LSW	-0.0285 (0.0631)	-0.0246 (0.0680)	-0.0109 (0.0745)
φ_{PG}		2.4660 (1.1162)		
σ			0.6898 (0.1001)	
φ_{PIGam}				3.4994 (1.1192)

Table 4. Posterior Estimates of Model Parameters for Indiana Data

Parameter	Respective Variable	**PG** mean (std dev.)	**PLN** mean (std dev.)	**PIGam** mean (std dev.)
$\ln(\beta_0)$	Intercept	-4.5245 (1.4117)	-3.5477 (1.4432)	-3.6814 (1.4259)
β_1	AADT	0.6974 (0.1314)	0.6231 (0.1335)	0.6459 (0.1223)
β_2	FRICTION	-0.0269 (0.0106)	-0.0301 (0.0109)	-0.0283 (0.0102)
β_3	PAVEMENT	0.4222 (0.1886)	0.4902 (0.2106)	0.4818 (0.2048)
β_4	MW	-0.0052 (0.0019)	-0.0073 (0.0022)	-0.0077 (0.0021)
β_5	BARRIER	-3.0591 (0.3028)	-4.1406 (0.4952)	-5.4245 (0.5866)
β_6	RUMBLE	-0.4050 (0.1825)	-0.3138 (0.2038)	-0.1574 (0.1924)
φ_{PG}		1.1143 (0.1421)		
σ			0.9665 (0.0661)	
φ_{PIGam}				1.8844 (0.2104)

COMPARISON OF MODEL PREDICTIONS

This section compares the predictions of the alternative models on two levels: 1) crash frequency expectation at sites in the dataset (i.e., posterior λ_i's), and 2) regression mean (μ) function across the models.

It is crucial to understand the practical importance of each level of comparison. λ_i's, the actual site-specific crash frequency expectations, are used in before-after studies (to represent the before conditions) and to rank sites based on their crash-proneness.

The μ function, on the other hand, not only affects the λ_i's estimates, but also has additional applications by itself. For instance, it is common for safety analysts to use calibrated crash prediction models to predict the expected crash frequency at planned facilities that have not yet been constructed. Here, the main objective is to estimate the safety impact of each design feature (e.g., lane width, type of median barrier, signal phasing, etc.) and help decision making in the design/planning stage. Also, the observed crash frequency may not be available at existing sites for reasons such as limited resources for data collection, etc. For such applications, the models' prediction is not affected by the site-specific observed crash frequency and, as shown later, the pattern of the difference between the models' predictions is quite different from that for the sites with observed crash counts. Therefore, the two comparison cases are presented in separate subsections.

In the comparisons hereafter, λ and μ variables are superscripted with the abbreviated name of the model with which they are associated; for example, λ_i^{PLN} is the λ of the i'th site when the Poisson-lognormal model is used.

Crash Frequecy Expectation ($\lambda_i | y$)

At this level of comparison, we focus on the posterior mean of λ_i's (i.e., $E(\lambda_i | y)$ for $i = 1, 2, \ldots, n$) and how they are affected by the model choice. Since the variations of λ_i's across the alternative models are relatively small (compared to the magnitude of λ_i's) , a traditional scatter plot may conceal possible trends. Therefore,

Figure 2-4 are constructed to magnify the differences between the models' predictions and observe their relationship. For each dataset, $E(\lambda_i | y)$'s from every pair of models are compared in a separate scatter plot. In all these plots, the horizontal axis represents the value of $E(\lambda_i | y)$ predicted by one of the two models being compared, whereas the vertical axis indicates the difference between the two models' predicted $E(\lambda_i | y)$.

The plots do not reveal a strong relationship between the alternative models' predictions for the site-specific expected crash frequencies. The only noticeable trend is that the most severe differences between the predictions are typically related to the sites where $E(\lambda_i^{PIGam} | y) > E(\lambda_i^{PLN} | y) > E(\lambda_i^{PG} | y)$. The relationship between the $E(\lambda_i | y)$'s of alternative models was further investigated and documented by Khazraee (2016); more complicated trends were identified when the regression mean and observed crash frequency were also considered in the analysis. However, from a practical standpoint, the main takeaway of the analysis is that the positive and negative differences between the predictions of different models balance and none of the models predicts an overall higher or lower crash frequency compared to the others. As Table 5 indicates, the total expected crash frequency (over all sites combined) is virtually equal across the three models and very close to the total observed crash frequency, indicating that none of the models predicts significantly biased $E(\lambda_i | y)$'s for any of the datasets.

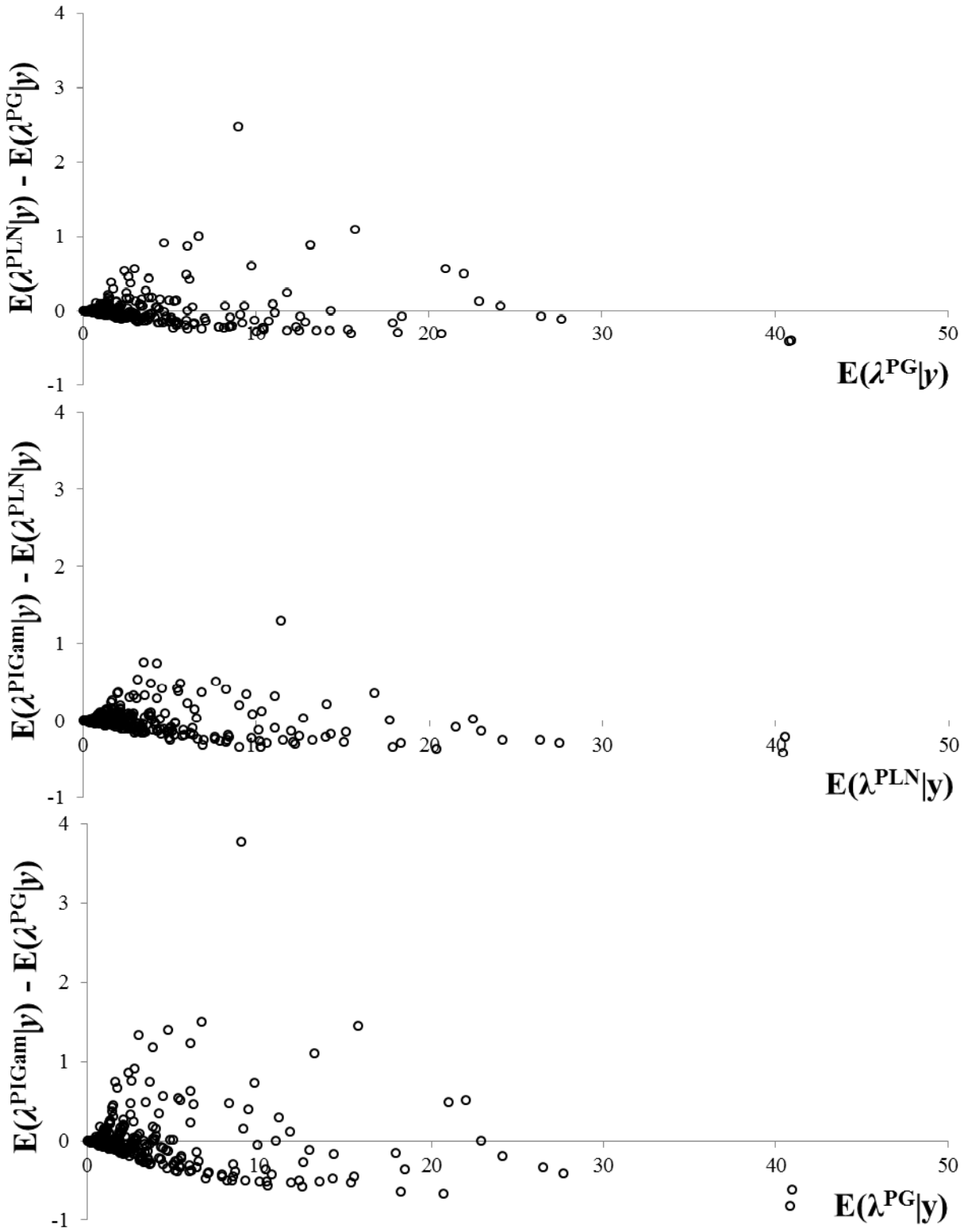

Figure 2. Difference between Texas dataset E($\lambda_i | y$)'s estimated using any two of the candidate models as a function of the E($\lambda_i | y$) from one of the models considered

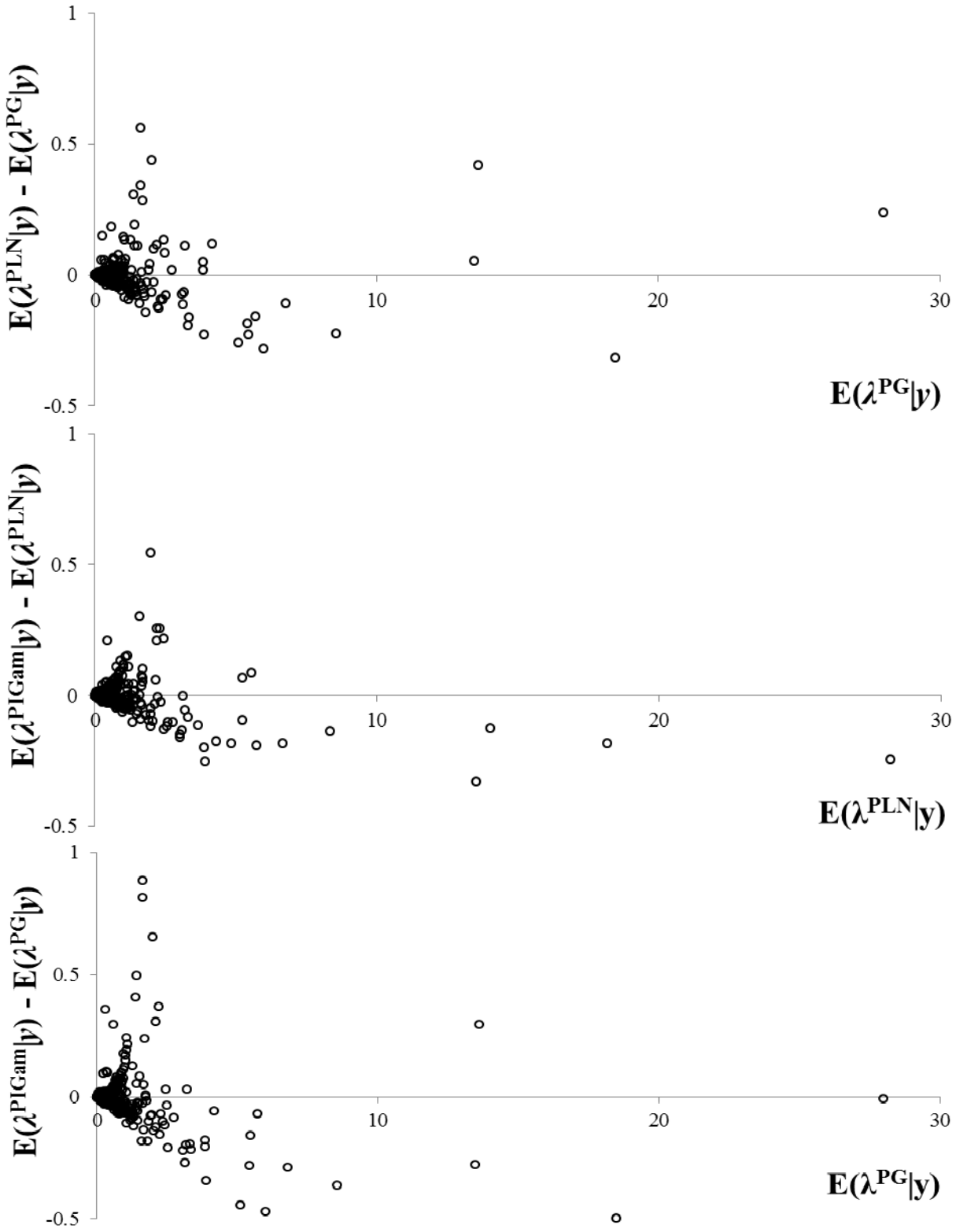

Figure 3. Difference between Michigan dataset $E(\lambda_i | y)$'s estimated using any two of the candidate models as a function of the $E(\lambda_i | y)$ from one of the models considered

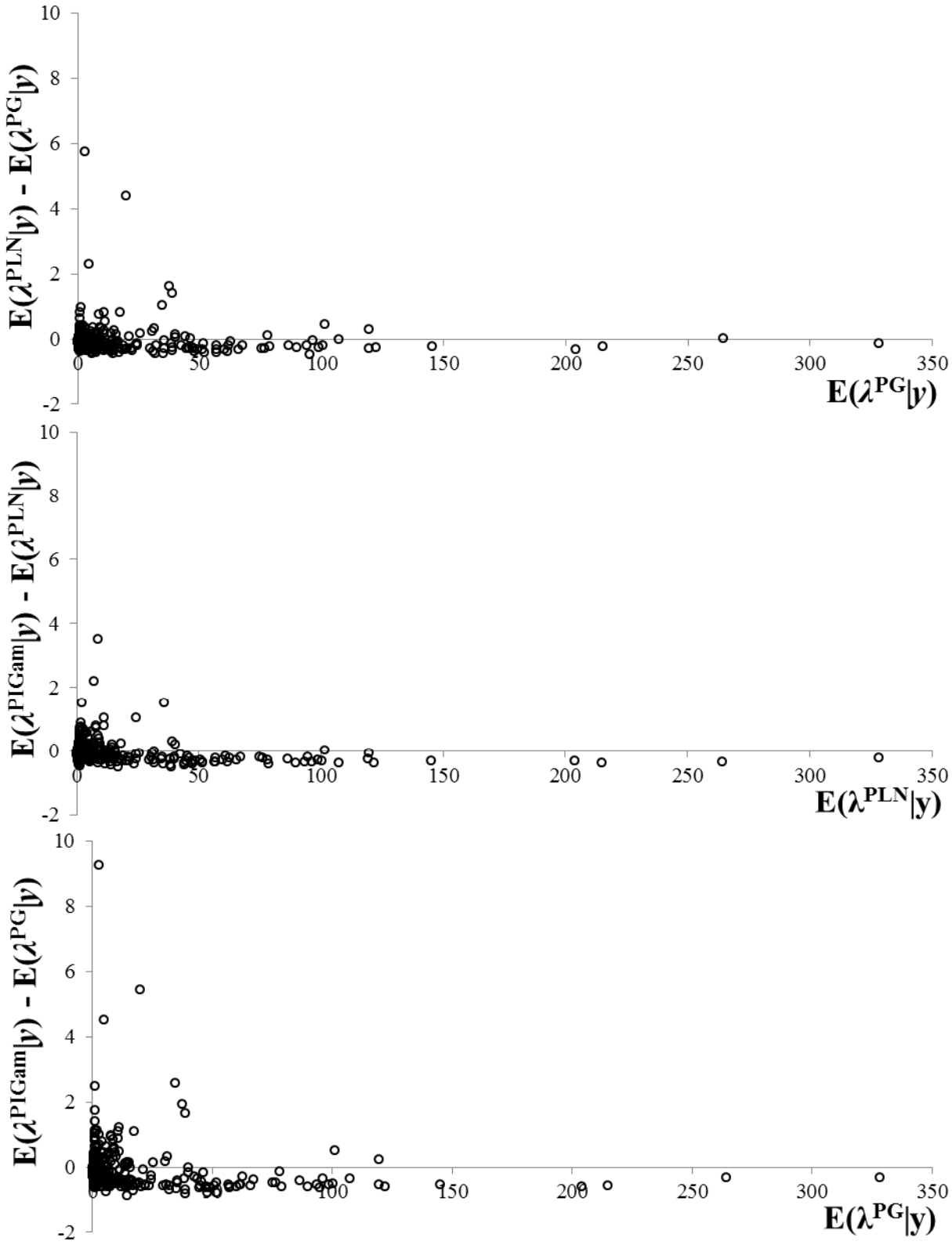

Figure 4. Difference between Indiana dataset E($\lambda_i | y$)'s estimated using any two of the candidate models as a function of the E($\lambda_i | y$) from one of the models considered

Table 5. Total expected crash frequency ($E(\lambda_i | y)$) over all sites predicted by each model for each dataset

Dataset	Total Expected Crash Frequency, $\sum_{i=1}^{n} E(\lambda_i \mid y)$			Total Observed Crash Frequency, $\sum_{i=1}^{n} Y_i$
	PG Model	PLN Model	PIGam Model	
Texas	1413.5	1413.5	1414.7	1415
Michigan	384.0	383.5	383.3	383
Indiana	5737.3	5736.4	5737.2	5737

Regression Mean (μ)

Crash prediction models are calibrated with historic crash data from a sample of roadway sites, and applied to sites without observed crash data to predict the expected crash frequency. Let j denote the index for a new roadway site to be constructed. In the absence of an observed crash count at site j, the expected crash frequency will follow a model-specific distribution (gamma, lognormal, or inverse gamma) with mean $\mu_j = \exp(X_j\beta)$. Differences in posterior β of the alternative models, and therefore their respective μ's, lead to different predictions for expected crash frequency at new sites.

To illustrate the implication of the model choice, we should compare μ_j across the alternative models for new sites with a range of characteristics. In the interest of convenience, we consider a hypothetical set of new sites with exactly silimar features (covariate vector values) as those in the dataset used for fitting the models. The expected crash frequency at each new site will be equal to $E(\mu_i | y)$ of the site in the calibration data set that it is replicating. With this hypothestical new set of sites, comparison between expected crash frequencies will be equivalent to simply comparing $E(\mu_i | y)$'s across the alternative models. This comparison is carried out in Figure 5-7 using the same scatter plots as those used earlier for comparing $E(\lambda_i | y)$'s.

One might expect that the differences between the values of each models' $E(\mu_i \mid y)$ predictions would increase with the increase in the magnitude of $E(\mu_i \mid y)$. That is, the difference between the alternative models' coefficients for traffic volume causes a larger difference between the models' $E(\mu_i \mid y)$'s for sites with greater traffic volume and thus greater $E(\mu_i \mid y)$'s. However, Figure 5-7 clearly indicate that the PLN model tends to predict larger μ_i's compared to the PG model, and the PIGam model tends to predict larger μ_i's compared to PLN model. The figures also indicate that the $E(\mu_i^{PG} \mid y) < E(\mu_i^{PLN} \mid y) < E(\mu_i^{PIGam} \mid y)$ relationship becomes more pronounced for larger $E(\mu_i \mid y)$'s. The aforementioned relationship is an important finding because, unlike the λ_i predictions for sites with observed crash frequency (see Table 5), it results in each model over- or under-predicting μ_i's (collectively, over all sites combined) relative to the other two models (as indicated in Table 6).

Table 6. Total ($E(\mu_i \mid y)$) over all sites predicted by each model for each dataset

Dataset	$\sum_{i=1}^{n} E(\mu_i \mid y)$			Total Observed Crash Frequency, $\sum_{i=1}^{n} Y_i$
	PG Model	PLN Model	PIGam Model	
Texas	1412.0	1428.9	1444.8	1415
Michigan	387.8	394.0	399.3	383
Indiana	5326.0	6298.2	7430.9	5737

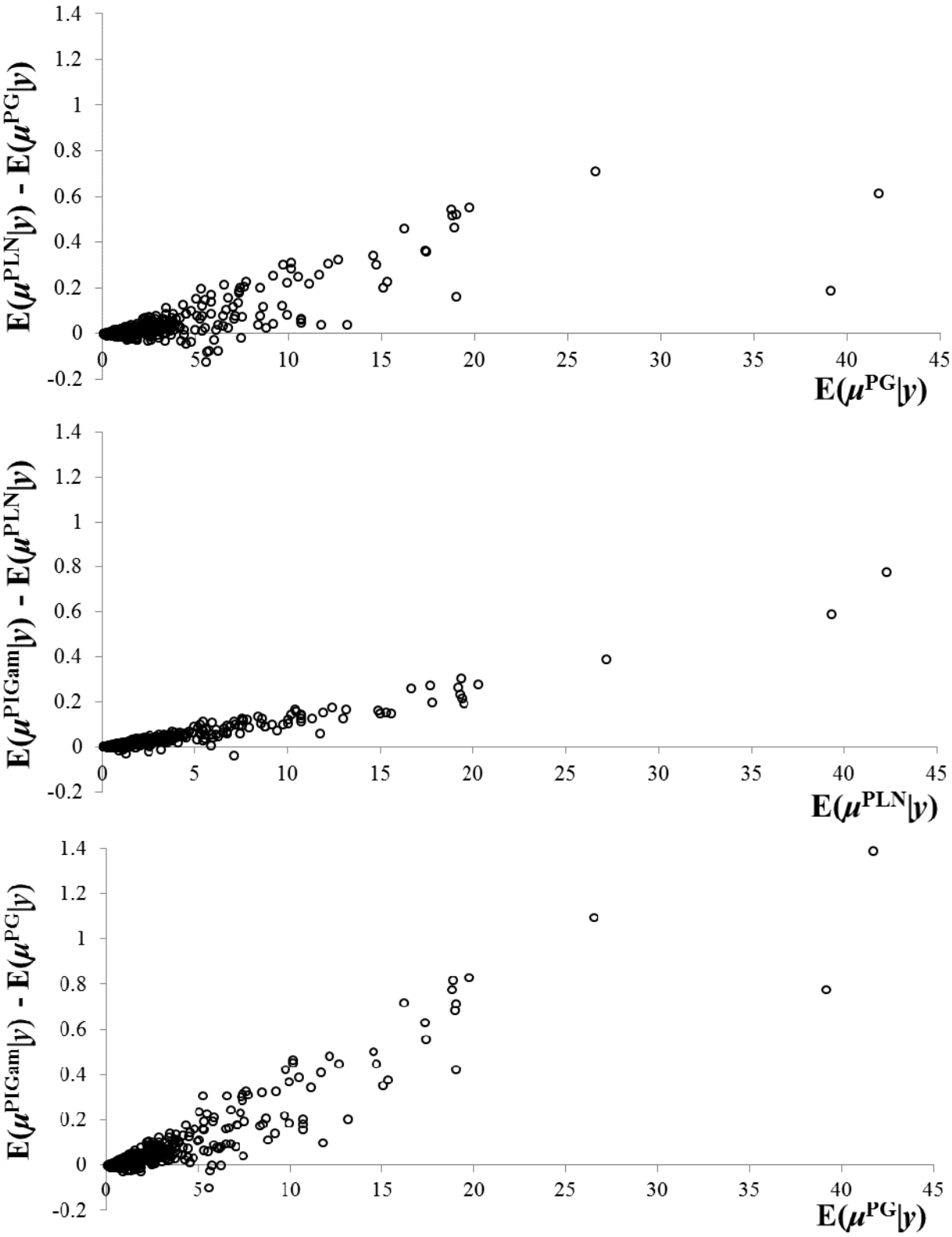

Figure 5. Difference between Texas dataset E($\mu_i | y$)'s estimated using any two of the candidate models as a function of the E($\mu_i | y$) from one of the models considered

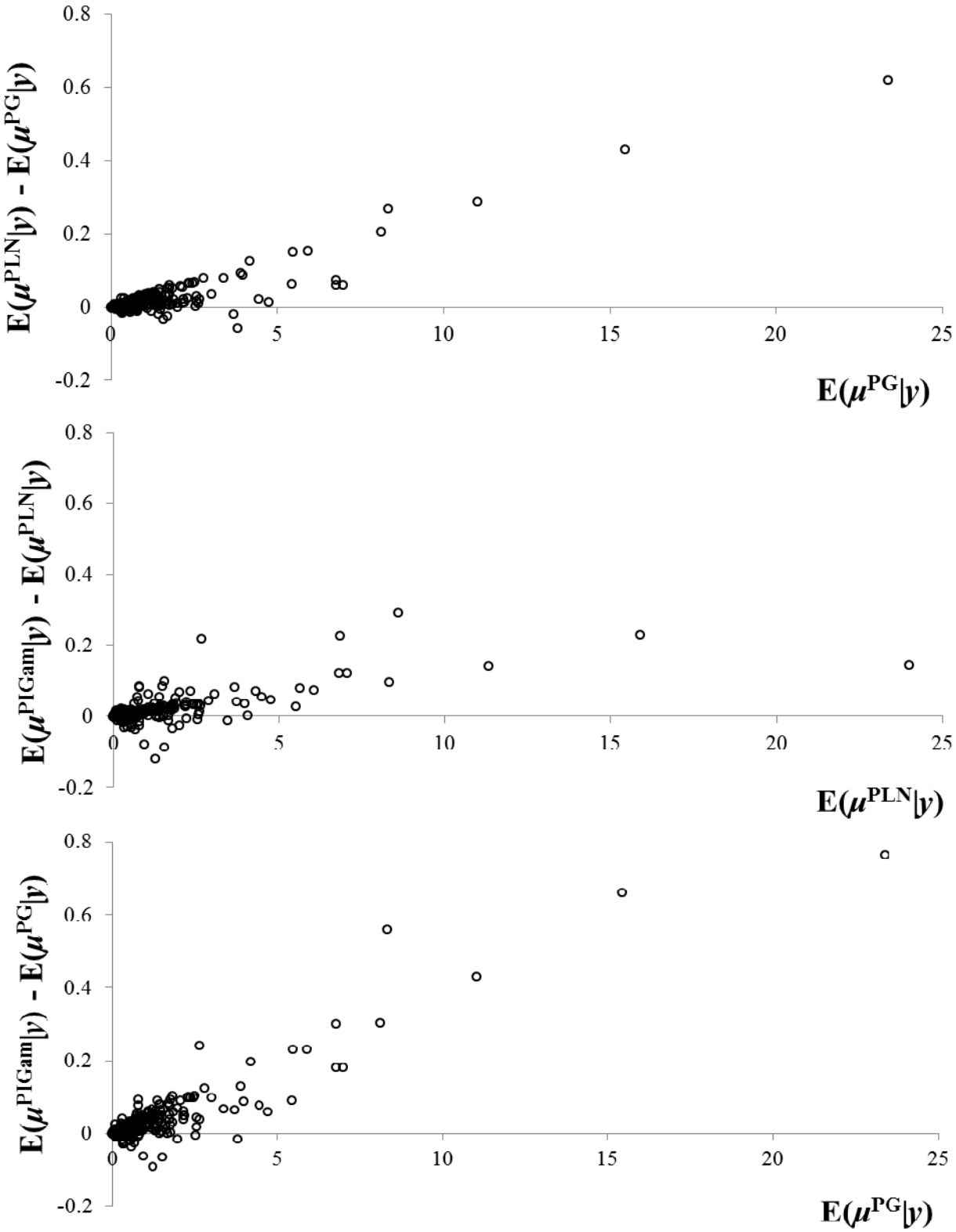

Figure 6. Difference between Michigan dataset E($\mu_i | y$)'s estimated using any two of the candidate models as a function of the E($\mu_i | y$) from one of the models considered

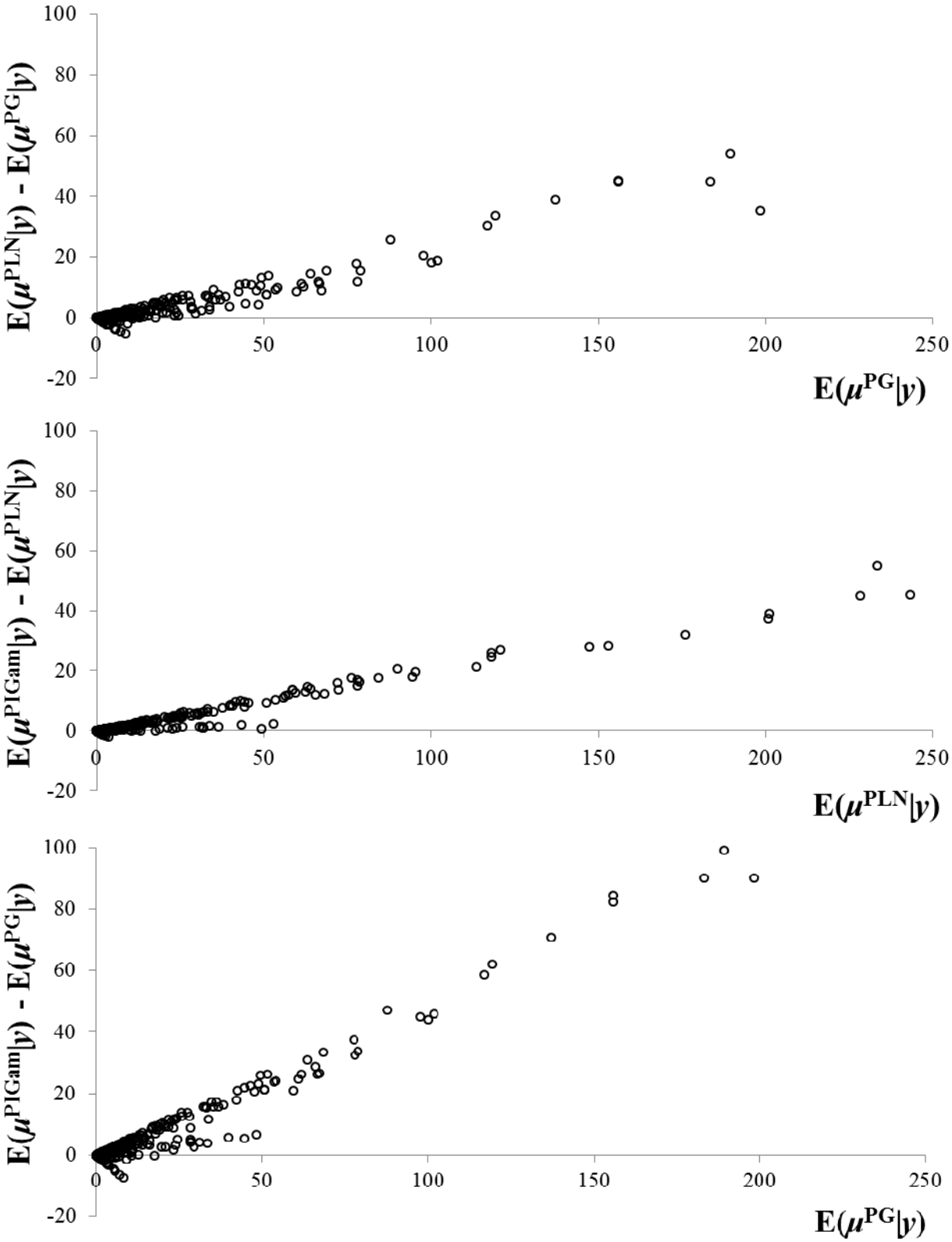

Figure 7. Difference between Indiana dataset E($\mu_i | y$)'s estimated using any two of the candidate models as a function of the E($\mu_i | y$) from one of the models considered

It is evident from Table 6 that the PIGam model predicts higher posterior regression means (for all sites combined) than the PLN model and the PG model, in order. It is interesting to note that most E($\mu_i \mid y$)'s from the PLN model are in between those from the PG and PIGam models, indicating an apparent link between the shape properties of the three mixing distributions (as explained earlier) and the predictions of their respective Poisson-hierarchical models.

Another important observation is that the degrees of variation between models' μ_i predictions increases as the data become more over-dispersed. As implied by Figure 5-7 and Table 6, the proportional (rather than absolute) difference between predicted E($\mu_i \mid y$)'s become more variant across the models in the following order of the datasets: Texas, Michigan, and Indiana, the same ranking of the datasets with respect to their NB model over-dispersion parameter. For the Indiana dataset, which is much more over-dispersed than the other two datasets, the alternative models result in drastically different regression means for sites with high traffic volume (AADT). For such datasets, the model choice is critical for any safety analysis.

GOODNESS OF FIT ANALYSIS

As explained earlier, the GOF of alternative Bayesian hierarchical models are most commonly compared using the DIC. In the context of Poisson-hierarchical models, the deviance (D) is typically calculated using the parameters in the lowest level of the hierarchy (λ_i's):

$$\bar{D} = \sum_{i=1}^{T} -2\log\left(\frac{\lambda_i^{y_i} e^{\lambda_i}}{y_i!}\right) \tag{16}$$

where T is the total number of MCMC samples used for posterior inference. While the model coefficients (and regression means, in turn) affect the λ_i estimates, they do not directly enter the

DIC calculation formula. Table 7 presents the DIC values obtained for each model-dataset.

Table 7. Deviance Information Criterion (DIC) for each model-dataset

Dataset	DIC		
	PG Model	PLN Model	PIGam Model
Texas	1608.9	1614.2	1624.6
Michigan	883.3	891.8	903.8
Indiana	1490.6	1465.3	1468.2

According to the DIC's in Table 7, for the two moderately over-dispersed datasets in this study (Texas and Michigan), the PG model fits the data better than the PLN model, which itself fits better than the PIGam model (the PLN model again falling in between the PG and PIGam models). For the highly over-dispersed Indiana dataset, however, the PLN and PIGam models provide a better fit than the PG model. Please note that the small difference between the DIC's of the PLN and PIGam models does not indicate a meaningful difference between the two models quality of fit to the Indiana dataset; as a rule of thumb, only a DIC difference of more than 10 constitutes a disparity between the models' fit to the data (MRC Biostatistics Unit, 2016).

The better performace of the PG model (compared to the PLN and PIGam models, in order) for the less over-dispersed datasets and its worse performance for a highly-overdispersed dataset may be attributable to the greater density near zero and thinner tail of the gamma distribution (resulting from higher skewness) compared to that of the lognormal and inverse gamma distributions, in order. The pattern hints to the possibility that the newly-proposed PIGam model might be the best-fitting model for data—even for more skewed and over-dispersed data than the Indiana dataset. The advantages of long/heavy distribution tails for fitting highly over-dispersed data have previously been demonstrated for the NB-L (Geedipally et al.,

2012; Shirazi et al., 2016) and Sichel (Zou et al., 2012) regression models. Furthermore, the results shown here support the recent work on using heuristic methods for selecting models based on the characteristics of the data (Shirazi et al., 2017a, 2017b). However, based on the experience of the authors in fitting the three models to several different datasets (only three of which are included in this paper), the relative GOF of the models (judged based on DIC) is far too complicated to be reliably predicted based on a simple measure of over-dispersion (such as α_{NB}) for the given dataset. In other words, one should not simply assume that a heavy-tailed model will perform better than a light-tailed model as over-dispersion in data increases; models have to be fitted before their GOF characteristics can be reliably compared.

The modeling results for the Indiana dataset reveal a very important limitation of the DIC for model selection. Virtually similar DIC's from the PLN and PIGam models implies that either model may be used for predicive applications without considerable difference. However, when the calibrated PIGam model is used to predict the expected crash frequency for new sites, for the majority of site conditions (i.e, covariate vector) it will predict a higher expectation than would the PLN model. As Figure 7 illustrated, for sites with high traffic flow (and thus great crash frequency) the practical difference between the two models will be drastic. Therefore, it is by no means reasonable to conclude (based on the DIC's) that the two models are performing equally well.

The aforementioned limitation of the DIC arises from the notion that it is solely calculated based on the conditional likelihood of the parameters in the lowest (i.e., observation) level of the hierarchy. As previously described in detail, in these models each λ_i is a parameter whose posterior expectation is not necessarily equal to that of μ_i. Therefore, alternative models may have similar (or slightly different) λ_i's while having considerably different regression mean

functions; for the Indiana dataset, the difference between λ_i^{PLN}'s and λ_i^{PIGam}'s is far smaller than that between μ_i^{PLN}'s and μ_i^{PIGam}'s.

This weakness of the DIC has extensively been documented by statisticians (Spiegelhalter et al. 2002; Celeux et al. 2006; Carlin, 2006; Meng and Vaida, 2006; and Plummer, 2006). In the context of hierarchcial bayesian models for over-dispersed count data (the typical case for crash count data), Miller (2009) warned against the naïve use of the "condtional DIC" and reported major inconsistenies between the DIC and more robust measures of fit such as the Bayes factor. The analysis in this paper illustrates the shortcoming of the DIC by presenting an exmple (Indiana data) where the DIC fails to differentiate between alternative models that result in drastically different predictions (for μ_i's).

Alternatives to the DIC are abundant (see e.g., Celeux, 2006). The most familiar and theoretically robust is the Bayes factor, which is defined as the ratio of the marginal densities of the data under two alternative models:

$$K = \frac{\Pr(D|M_1)}{\Pr(D|M_2)} = \frac{\int \Pr(\theta_1|M_1)\Pr(D|\theta_1, M_1)d\theta_1}{\int \Pr(\theta_2|M_2)\Pr(D|\theta_2, M_2)d\theta_2} \quad (17)$$

where D denotes the observed data, M_1/M_2 the alternative Model 1 and 2, and θ_1/θ_2 the collective set of parameters for Model 1 and 2, respectively. Unlike DIC, the Bayes factor accounts for the marginal probability structure of the competing models. In addition, the value of the Bayes factor is easily interpretable. For example, a Bayes factor of 0.2 means that model 2 is 5 times more likely than model 1. These likelihood ratios can be used to mix the alternative models and obtain more accurate mixed models.

The major drawback of the Bayes factor is the computational intensiveness due to the high-dimension integral over the parameter space. Researchers have proposed a plothera of methods for approximating the marginal density (e.g., Gelman and Meng, 1998; Han and Carlin, 2001; Sinharay and Stern, 2005). Determination of the marginal density for the PG model is straightforward due to the closed form of the marginal likelihood function. Miller (2009) used a complicated method developed by Chib (1995) to approximate the marginal density for the PLN model. The marginal density for the PIGam model is expectedly more complicated and difficult to estimate.

Given the complexity of estimating Bayes factors for the models in this paper, reviewing the existing methods, adjusting and implementing them can be the subject of another research srtudy. The initial goal of this paper was to evaluate and compare Poisson-hierarchical models in terms of the relative magnitude of their predictions (rather than GOF) and the resulting practical implications of using one model instead of another. The methodology used for analysis of predictions revealed a known (but often neglected) weakness of evaluating the GOF using the DIC.

SUMMARY AND CONCLUSIONS

This research investigated the model selection dilemma for crash prediction models from a new perspective. The methodology deviated from the conventional model selection studies in the following sense: rather than ranking the alternative models based on their overall GOF and predictive performance, the research focused on the magnitude of models predictions for the expected crash frequency at individual sites. Focusing on an important family of count regression

models, Bayesian Poisson-hierarchical, the study found important trends among predictions of three alternative models: Poisson-gamma, Poisson-lognormal, and Poisson-inverse gamma.

The common structure of the selected candidate models facilitated comparison between their predictions and the similar number of parameters provided for a fair comparison. The distinction between the regression mean (μ_i) and the site-specific crash count expectation (λ_i) was well explained; posterior λ_i is the actual expected crash frequency at site i after the regression mean (μ_i) is adjusted for the observed crash frequency (y_i). μ_i may be interpreted as the expected crash frequency at a hypothetical new site with the same modeled characteristics as those of site i but without any observed crash count.

Fitting the candidate models to three relatively large datasets (from Texas, Michigan, and Indiana), the following outcomes were observed:

1) The differences between the λ_i's predicted by the three models grew larger as the over-dispersion in the data increased. While the differences between posterior λ_i's across models were significant at a few sites (especially in the highly over-dispersed Indiana dataset), the positive and negative differences balanced off and none of the models predicted an overall higher or lower crash frequency compared to another. Thus, the model choice would not make a practically important difference for predicitng the crash frequency expectation at sites in the dataset (with an available observed crash count).
2) The differences between the μ_i's predicted by the three models were quite significant and grew larger as the over-dispersion in the data increased. The importance of the differences between the μ_i's was that they resulted in the PIGam model predicting higher posterior regression means (for all sites combined) compared to the PLN model and the PG model, in order. Thus, the model choice can make a critical difference when the calibrated models are

used for predictions at new sites without observed crash counts. There is a clear link between the μ_i predictions of alternative models and the shape of their respective mixing distribution; the lognormal distribution falls in between the gamma and inverse gamma distribution in terms of its important shape characteristics (e.g., tail thickness, skewness, etc.) and so do the PLN model's μ_i predictions fall in between those of the PG and PIGam models.

3) According to the most commonly used measure of fit for Bayesian hierarchical models, the DIC, the PIGam and PLN models provided a better fit to the highly over-dispersed Indiana dataset (compared to the PG model) but not to the other less over-dispersed datasets. The better performance of the PLN and PIGam models may have arisen from their thicker tails. However, the analysis illustrated a major deficiency of the DIC in comparing hierarchical models. For Indiana data, the PLN and PIGam models yielded very similar DIC values while predicting drastically different μ_i's, with the PIGam model predicting a higher μ_i for most sites in the dataset. It was explained that this shortcoming of the DIC is due to the notion that its calculation is based solely on the conditional likelihood of the parameters in the lowest (i.e., observation) level of the hierarchy, ignoring the higher level(s). Consequently, the authors advise against using DIC (as the sole measure of GOF) for comparison of hierarchical models when the coefficients are highly dependent on the choice of model's distribution (often the case for highly over-dispersed datasets such as Indiana data).

As described throughout this paper, the research did not seek to advise a technique for selecting the best performing model out of a pool of candidates, but rather to demonstrate the impact on the magnitude of predictions when one model is selected instead of another. The analysis rejected the suggestion that a simple measure of dispersion can be used to reliably predict the best-performing model. Research using simulated data and a multitude of field data

with a wide range of characteristics is needed for such purpose. At the time of publication for this paper, new research was identified which has addressed this matter using heuristic methods (Shirazi et al., 2017, 2018).

In addition, further research is needed to assess the implications of this research's findings on different highway safety analysis applications, such as identification of hazardous sites and before-after analyses that employ FB models. Similar assessment can also be performed for recently introduced multi-parameter models such as the NB-L (Geedipally et al., 2012), the NB-generalized exponential (NB-GE) (Vangala et al., 2015), or the NB- Dirichlet Process (NB-DP) (Shirazi et al., 2016).

REFERENCES

AASHTO, 2010. Highway Safety Manual, 1st Edition. American Association of State and Highway Transportation Officials, Washington, D.C.

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record 2061, 55–63.

Aguero-Valverde, J., 2013. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. Accident Analysis & Prevention 50, 289-297.

Bonneson, J.A., McCoy, P.T., 1993. Estimation of safety at two-way stop-controlled intersections on rural highways. Transportation Research Record 1401, 83-89.

Carlin, B. P., 2006. Comment on article by Celeux et al. Bayesian Analysis 1, 675–676.
Carriquiry, A., Pawlovich, M.D., 2004. From empirical Bayes to full Bayes: methods for analyzing traffic safety data. http://www.iowadot.gov/crashanalysis/pdfs/eb_fb_ comparison_whitepaper_october2004.pdf (accessed March 7th, 2014).

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., 2006. Deviance information criteria for missing data models (with discussion). Bayesian Analysis 1, 651–706.

Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.

Davis, G.A., Yang, S., 2001. Bayesian identification of high-risk intersections for older drivers via Gibbs sampling. Transportation Research Record 1746, 84-89.

Daziano, R.A., Miranda-Moreno, L., Heydari, S., 2013. Computational Bayesian statistics in transportation modeling: from road safety analysis to discrete transport reviews 33 (5), 570-592.

El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate poisson-lognormal regression. Accident Analysis & Prevention 41, 820–828.

Fitzpatrick, K., Park, E.S., 2009. Safety effectiveness of HAWK pedestrian treatment. Transportation Research Record 2140, 214-223.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-Lindley generalized linear model: characteristics and application using crash data. Accident Analysis & Prevention 45(2), 258–265.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2013. A caution about using the Deviance Information Criterion while modeling traffic crashes. Safety Science, Vol. 62, pp. 495-498.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398-409.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data analysis. Third edition. Chapman and Hall/CRC, New York.

Gelman, A., Meng, X.-L., 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Statistical Science 13, 163–185.

Goldstein, H, 2010. Multilevel statistical models, 4th Edition, John Wiley & Sons, West Sussex, England.

Han, C., Carlin, B. P., 2001. Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. Journal of the American Statistical Association 96, 1122–1132.

Harwood, D.W., Bauer, K.M., Potts, I.B., Torbic, D.J., Richard, K.R., Kohlman Rabbani, E.R., Hauer, E., Elefteriadou, L., 2002. Safety effectiveness of intersection left- and right-turn lanes, Report No. FHWA-RD-02-089. Federal Highway Administration (FHWA), Washington, D.C.

Hauer, E., Persaud, B.N., 1983. A common bias in before and after accident comparisons and its elimination. Transportation Research Record 905, 164-174.

Hauer, E., 1986. On the estimation of the expected number of accidents. Accident Analysis & Prevention 18, 1-12.

Hauer, E., Ng, J.C.N., Lovell, J., 1989. Estimation of safety at signalized intersections. Transportation Research Record 1185, 48-61.

Hauer, E., 1992. Empirical Bayes approach to the estimation of unsafety: the multivariate regression approach. Accident Analysis & Prevention 24 (5), 456-478.

Hauer, E., 1997. Observational before–after studies in road safety: estimating the effect of highway and traffic engineering measures on road safety. Pergamon Press, Elsevier Science, Ltd., Oxford, United Kingdom.

Hauer, E., 2001. Overdispersion in modeling accidents on road sections and in empirical Bayes estimation", Accident Analysis and Prevention 33(6), pp. 799-808.

Heydecker B.G., Wu, J., 2001. Identification of sites for accident remedial work by Bayesian statistical methods: An example of uncertain inference. Advances in Engineering Software 32, 859-869.

Hinde, J., 1982. Compound Poisson regression models, in R. Gilchrist, ed., GLIM 82: Proceedings of the International Conference on Generalized Linear Models, New York, Springer-Verlag.

Khazraee S.H., 2016. Full Bayesian Poisson-hierarchical models for crash data analysis: investigating the impact of model choice on site-specific predictions. PhD Dissertation. Department of Civil Engineering, Texas A&M University, College Station, Texas.

Lee, P., 2012. Bayesian statistics: an introduction. 4th Edition, John Wiley & Sons, West Sussex, England.

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., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44, 291–305.

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Safety Science 46 (5), 751–770.

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

Lord, D., Washington, S.P., Ivan, J.N. 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis & Prevention 37(1), pp. 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 and Prevention 39(1), 53–57.

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

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods in Accident Research 1, 1–22.

Mannering, F. L., Shankar, V., and Bhat, C. R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research, 11, 1-16.

Maycock, G., Hall, R.D., 1984. Accidents at four-arm roundabouts. Laboratory Report LR 1120, Transport Research Laboratory, Crowthorne, Berkshire, U.K.

Meng X.-L., Vaida, F., 2006. What's missing for DIC with missing data? (Comment on article by Celeux et al.). Bayesian Analysis 1, 687–698.

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

Miaou, S.-P., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: a space time modeling approach. Journal of Transportation and Statistics 6 (1), 33–57.

Miaou, S.-P., Song J.J., 2005. Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion and spatial dependence. Accident Analysis and Prevention, Vol. 37, No. 4, pp. 699-720.

Miller, 2009. Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. Biometrics 65, 962-969.

Miranda-Moreno, L.F., Fu, L., Saccomanno, F.F., Labbe, A, 2005. Alternative risk models for ranking locations for safety improvement. Transportation Research Record 1908, pp 1-8.

Mitra, S., Washington, S., 2006. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39, 459–468.

MRC Biostatistics Unit, Cambridge Biomedical Campus. DIC: Deviance Information Criterion. <http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-dic>. Accessed April 3rd, 2016.

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

Park, E.S., Park, J., Lomax, T.J., 2010. A fully Bayesian multivariate approach to before-after safety evaluation. Accident Analysis and Prevention 42, 1118-1127.

Persaud, B.N., 1988. Do traffic signals affect safety? Some methodological issues. Transportation Research Record 1185, 37–47.

Persaud, B., Lyon, C., Nguyen, T., 1999. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transportation Research Record 1665, 7-12.

Persaud, B., Lyon, C., 2007. Empirical Bayes before–after safety studies: lessons learned from two decades of experience and future directions. Accident Analysis & Prevention 39, 546–555. Plummer, M., 2006. Comment on article by Celeux et al. Bayesian analysis 1, 681–686.

Poch, M., Mannering, F.L., 1996. Negative binomial analysis of intersection accident frequency. Journal of Transportation Engineering 122(2), pp. 105-113.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL <http://www.R-project.org>

Schluter, P.J., Deely, J.J., Nicholson, A.J., 1997. Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. The Statistician 46, 293–316.

Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017. A methodology to design heuristics for model selection based on characteristics of data: application to investigate when negative

binomial Lindley (NB-L) is preferred over negative binomial (NB). Accident Analysis & Prevention, Vol. 107, pp. 186-194. (http://dx.doi.org/10.1016/j.aap.2017.07.002).

Shirazi, M, Lord, D., 2018, Characteristics Based Heuristics to Select a Logical Distribution between the Poisson Gamma and the Poisson Log-Normal. Paper presented at the 97th Annual Meeting of the Transportation Research Board, Washington, D.C.

Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over dispersed count data with a heavy tail: characteristics and applications to crash data. Accident Analysis & Prevention 91, 10-18.

Sinharay, S., Stern, H.S., 2003. Posterior predictive model checking in hierarchical models. Journal of Statistical Planning and Inference 111, 209–221.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B 64, 583–640.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Lun, D., 2003. WinBUGS Version 1.4.3 User Manual. MRC Biostatistics Unit, Cambridge. <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf>

Tanner, M., Wong, W., 1987. The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association 82, 528-550.

Vangala, P., Lord, D., Geedipally, S.R., 2015. Exploring the application of the negative binomial-generalized exponential model for analyzing traffic crash data with excess zeros. Analytic Methods in Accident Research 7, 29-36.

Vogt, A., Bared, J., 1998. Accident models for two-lane rural segments and intersections. Transportation Research Record 1635, 18-29.

Washington, S.P., Karlaftis, M., Mannering, F.L., 2010. Statistical and econometric methods fo`r transportation data analysis. 2nd Ed. Chapman and Hall, Boca Raton.

Yang, M., Rasbash, J., Goldstein, H., Barbosa, M., 1999. MLwiN macros for advanced multilevel modelling. Version 2.0a. Multilevel Models Project, Institute of Education, University of London, UK. <http://www.bris.ac.uk/cmm/software/mlwin/download/d-1-10/advmacma.pdf>

Zha, L., Lord, D., Zhou, Y., 2016. The Poisson inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data. Journal of Transportation Safety and Security 8 (1), 18-35.

Zou, Y., Lord, D., Zhang, Y., 2012. Analyzing highly dispersed crash data using the Sichel generalized additive models for location, scale and shape. Working Paper. Zachry Department of Civil Engineering, Texas A&M University, College Station, TX. < https://ceprofs.civil.tamu.edu/dlord/Papers/Zou_et_al_Sichel_GAMLSS.pdf>