A flexible discrete density random parameters model for count data: Embracing unobserved heterogeneity in highway safety analysis

Shahram Heydari, PhD (Corresponding Author)
Lecturer (Assistant Professor)
Transportation Research Group
Department of Civil, Maritime and Environmental Engineering
Faculty of Engineering and Physical Sciences
University of Southampton
United Kingdom
S.Heydari@soton.ac.uk

ABSTRACT

In traffic safety studies, there are almost inevitable concerns about unobserved heterogeneity. As a feasible alternative to current methods, this article proposes a novel crash count model that can address asymmetry and multimodality in the data. Specifically, a Bayesian random parameters model with flexible discrete densities for the regression coefficients is developed, employing a Dirichlet process prior. The approach is illustrated on the Ontario Highway 401, which is one of the busiest North American highways. The results indicate that the proposed model better captures the underlying structure of the data compared to conventional models, improving predictive power examined based on pseudo Bayes factors. Interestingly, the model can identify sites (highway segments, intersections, etc.) with similar site characteristic (risk factor) profiles, those that manifest similarity in the heterogeneous effects of their risk factors (e.g., traffic flow) on traffic safety, providing useful insight towards designing effective countermeasures.

1. Introduction

Crash data are often limited since many unobserved or unmeasured factors that affect crash likelihood may not be available (e.g., driver behaviour, environmental conditions, etc.), causing unobserved heterogeneity. In road safety studies it is usually necessary to account for unobserved heterogeneity to draw more reliable statistical inferences, which in turn help decision-makers to plan safety improvement programs more effectively. Due to multiple sources of heterogeneity in crash data sets, more complex statistical models seem inevitable.

In fact, various efforts have been underway to mitigate the unobserved heterogeneity problem in traffic safety analysis, which could be summarized as follows: (1) the random effects (intercepts) approach (Shankar et al., 1998; Kim al., 2007; Aguero-Valverde, 2013; Naznin et al., 2016; Sarwar et al., 2017b); (2) the random parameters (also referred to as random slopes or random coefficients) approach (Anastasopoulos and Mannering, 2009; El-Basyouny and Sayed, 2009; Venkataraman et al., 2014; Wu et al., 2013; Anastasopoulos, 2016; Sarwar et al., 2017a; Alarifi et al., 2017; Bogue et al., 2017; Chen et al., 2017; Bhat et al., 2017, Fountas and Anastasopoulos, 2017; Shaon et al., 2018; Fountas et al., 2018; Cai et al., 2018; Heydari et al., 2018); (3) the finite mixture approach (Park and Lord, 2009; Park et al., 2016; Yasmin and Eluru, 2016; Zou et al., 2017); (4) the finite mixture random parameters approach (Xiong and Mannering, 2013; Li et al., 2018); and recently (5) the Bayesian semiparametric Dirichlet process approach (Heydari et al., 2016a and 2016b; Shirazi et al., 2016; Yu et al., 2016; Heydari et al., 2017; Cheng et al., 2018), which has been applied only in random effects model settings in traffic safety research. A detailed discussion relating to various statistical models used in traffic safety research and unobserved heterogeneity can be found in Lord and Mannering (2010), Mannering and Bhat (2014), and Mannering et al. (2016).

The random parameters approach — which is the focus of this article — is perhaps the most commonly used method to address unobserved heterogeneity in the crash literature. It can be unrealistic to assume that the effect of explanatory variables is fixed across observations or groups of observations since the effect of covariates (e.g., site characteristics) could vary from one site (highway segment, intersection, region, etc.) to another due to several unobserved or unmeasured factors. Therefore, random parameters (slopes) models are in general expected to better address unobserved heterogeneity relative to random intercepts (effects) models that only allow the constant term to vary across observations (Washington et al., 2011; Mannering et al., 2016). It should be noted, however, that this does not necessarily rule out random intercepts models as they may adequately approximate the distribution of a regression coefficient if the variation between sites is small enough. The use of random parameters models is important when the intrinsic interest lies in investigating the range of a parameter or in capturing unobserved heterogeneity more fully.

Although random parameters models have been extensively used in traffic safety studies, a major limitation is in their inherent parametric assumptions (Mannering and Bhat, 2014). To draw more robust statistical inferences, while providing more insight, a limited number of traffic safety studies have recently employed models with heterogeneity in means and variances (Venkataraman et al., 2014; Behnood and Mannering, 2017a and 2017b; Seraneeprakarn et al., 2017; Xin et al., 2017; Heydari et al., 2018). Anastasopoulos (2016) addressed concerns relating to parametric assumptions for random parameters by testing different densities such as lognormal, Weibull and normal to find the most appropriate distribution. This choice, however, usually needs to be done via model selection, an exhaustive task when there are many regression parameters, and even then, none of the candidate densities may be optimal. In addition, model selection criteria such as the AIC or the DIC have limitations, for example, due to excessive sensitivity to different parameterizations (Washington et al., 2011; Geedipally et al., 2014).

Using a Bayesian nonparametric approach, Heydari et al. (2016a and 2016b) discuss sensitivity to parametric assumptions in random intercepts models, showing that flexible densities can address unobserved heterogeneity better than restrictive parametric densities in random intercepts models. Another limitation to standard random parameters models is that the analyst often decides groupings based on apparent data features but what if there are hidden groupings? Those groupings not known to the analyst could pose significant challenge in addressing unobserved heterogeneity properly (Mannering and Bhat, 2014). Finite mixture modeling is a viable approach that can side-step the above issues (asymmetry, multimodality, and hidden groupings), but similar to choosing the best density in random parameter modeling one needs to select the optimal number of components through model selection and statistical fit.

1.1. The current paper

This paper proposes an alternative approach to current methods: a new crash count random parameters model developed based on flexible Bayesian discrete densities. The proposed model does not require model selection neither to choose the proper density for random regression parameters as in random parameters modeling nor to choose the optimal number of components as in finite-mixture modeling. The proposed model assumes an unspecified discrete density for regression parameters in a count model with both fixed and random regression parameters — inferring the shape of random parameters densities from the data. Specifically, a Bayesian semiparametric Poisson lognormal model is developed using a stick-breaking algorithm.

In the statistical literature, a similar approach can be found in Dhavala et al. (2010) who, using a polya-urn scheme (Escobar and West, 1998), proposed a Bayesian semiparametric zero inflated count model of massively parallel signature sequencing for gene expression profiling. Carota and Parmigiani (2002) provide a general discussion of Bayesian semiparametric Dirichlet process models for count data. Other applications can be found, for example, in Guindani et al. (2014) and Canale and Prunster (2017). A useful by-product of the model developed in this paper is that it allows the analyst to identify sites (highway segments, intersections, etc.) that manifest similarity in the heterogeneous effects of their site characteristics (e.g., median shoulder width) on safety (here, crash frequencies). This could be particularly useful for decision-makers in designing effective countermeasures.

2. Methodological approach

This paper employs a Poisson lognormal model. Similar to negative binomial (Poisson gamma) models, Poisson lognormal models allow for overdispersion, often encountered in count data. The Poisson-lognormal approach is appealing from both theoretical and practical perspectives. Winkelmann (2008) provides a discussion in this regard. A number of traffic safety studies have used and discussed various aspects of Poisson lognormal models (Lord and Miranda-Moreno, 2008; El-Basyouny and Sayed, 2009; Aguero-Valverde, 2013; Heydari et al., 2016a; Khazraee et al, 2018; Heydari et al., 2018). This section first discusses the standard random intercepts and random parameters Poisson lognormal models before extending them to include flexible discrete densities for random regression coefficients. Note that the ideas discussed in this paper could be used to extend other models (e.g., random parameters Poisson gamma models) as well.

2.1. Standard random intercepts model

Let y_i and θ_i denote the observed and the expected crash counts for sites ($i = 1,2,\ldots, N$). Let $\boldsymbol{X} = (X_1, X_2,\ldots, X_k)$ be the vector of covariates (i.e., site characteristics) with the corresponding regression parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2,\ldots, \gamma_k)$, excluding the constant term η. A generic Poisson lognormal model can be specified as

$$\begin{gathered} y_i|\boldsymbol{X_i}, \boldsymbol{\gamma}, \varepsilon_i, \eta \ \sim Poisson\ (\theta_i) \\ \theta_i = \lambda_i * e^{\epsilon_i} \\ log(\lambda_i) = \eta + \boldsymbol{\gamma X_i} \\ log(\theta_i) = \eta + \boldsymbol{\gamma X_i} + \epsilon_i \\ \varepsilon_i|v_\varepsilon \sim normal(0, v_\varepsilon) \end{gathered} \qquad (1)$$

In the above model specification, e^{ε} follows a lognormal density (instead of a gamma density as in Poisson gamma models), meaning that ε follows a normal density with the mean 0 and the variance following a low information prior. By including the random variation ε_i in the constant term η, the above model can be written as

$$\begin{gathered} y_i|\boldsymbol{X_i}, \boldsymbol{\gamma}, \varepsilon_i, \eta_i \ \sim Poisson\ (\theta_i) \\ log(\theta_i) = \ \eta_i + \boldsymbol{\gamma X_i} \\ \eta_i|\mu_\eta, v_\eta \sim normal(\mu_\eta, v_\eta) \\ \boldsymbol{\gamma} \sim normal(0, 1000) \\ \mu_\eta \sim normal(0,1000) \\ \sigma_v \sim uniform(0,20) \\ v_\eta = \sigma_v^2 \end{gathered} \qquad (2)$$

where random intercepts η_i are usually assumed to follow a normal density, but other densities are possible as well. The vector of fixed parameters $\boldsymbol{\gamma}$ and the mean and the variance of the random intercepts are assumed to follow low information priors as indicated in (2).

2.2. Standard random parameters model

One can allow the above model to also include a set of covariates with varying effects, $\boldsymbol{Z} = (Z_1, Z_2,\ldots, Z_m)$, across observations with their corresponding random parameters $\boldsymbol{\beta} = (\beta_1, \beta_2,\ldots, \beta_m)$, including varying intercepts. Given the above notation, a generic random parameters Poisson lognormal model can be specified as

$$\begin{aligned} y_i|\boldsymbol{X_i},\boldsymbol{Z_i},\boldsymbol{\gamma},\boldsymbol{\beta_i} &\sim Poisson\,(\theta_i) \\ log(\theta_i) &= \boldsymbol{\beta_i}\boldsymbol{Z_i} + \boldsymbol{\gamma}\boldsymbol{X_i} \\ \boldsymbol{\beta_i}|\boldsymbol{\mu_\beta},\boldsymbol{v_\beta} &\sim normal(\boldsymbol{\mu_\beta},\boldsymbol{v_\beta}) \end{aligned} \qquad (3)$$

In the above formulation, $\boldsymbol{\beta}_i$ are often assumed to be normally distributed. Note that one can assume other densities such as lognormal or Weibull for random regression parameters $\boldsymbol{\beta}$, so the normality assumption is not a requirement. However, the normal density is often the first and the only choice due to convenience. Low information priors can be specified for the vector of fixed parameters $\boldsymbol{\gamma}$ and the means $\boldsymbol{\mu}_i$ and variances $\boldsymbol{\nu}_i$, as in (2).

2.3. Extension to a flexible discrete density random parameters model

The proposed approach is an extension of the method, a flexible random intercepts model, discussed in Ohlssen et al. (2007) and is rooted in Bayesian nonparametrics (Escobar and West, 1998; Walker et al., 1999; Neal, 2000; Muller and Quintana, 2004; Hjort et al., 2010; Ladouceur et al., 2011). Specifically, this paper introduces a Bayesian semiparametric model that places a unique Dirichlet process prior on two or multiple random regression parameters. Doing so, random parameters are tied together allowing the analyst to identify observations (sites) that are similar according to the heterogeneous effects of their corresponding site characteristics on safety. If a single Dirichlet process prior is placed on all regression coefficients, one can then identify similar profiles of an outcome of interest (e.g., crash frequencies) in the data.

A major challenge in statistical modeling, in general, is choosing an appropriate level of complexity for a model. As discussed in Gresham and Blei (2012), a valuable property of Bayesian nonparametrics is that the number of parameters can vary according to data complexity. This number is then decided by the model (given the data) based on a rigorous mathematical algorithm. The Bayesian nonparametric approach is flexible as the analyst can test whether a parametric assumption (normal, lognormal, etc.) holds. If a parametric assumption does not hold, the proposed model fits a flexible density. Otherwise, the model approximates that parametric density, for example, confirming that there is no multimodality or skewness in the data. Heydari et al. (2016a) showed this explicitly for the vector of random intercepts using a simulated data set.

Conventional statistical models under both frequentist and Bayesian frameworks (e.g., the model specified in Section 2.2) assume that each of the random coefficients $\boldsymbol{\beta}$ follows a specific "known" continuous density function G_β(.) with a "known" (and finite) number of "unknown" parameters; for example, a normal distribution with two parameters: the mean and the variance to be inferred from the data. Assuming a parametric density (whether normal or other densities) is the standard approach. The Bayesian nonparametric methods (which constitute the idea behind the proposed random parameters model), however, assume that G_β is an unspecified discrete density with an "unknown" number of "unknown" parameters to be inferred from the data, reflecting the lack of knowledge about a parameter or data set of interest more realistically. The Bayesian nonparametric approach then places a Dirichlet process (DP) prior (Freedman, 1963; Ferguson, 1973) on G_β (corresponding to $\boldsymbol{\beta}$) to infer its density:

$$\begin{aligned}&\boldsymbol{\beta}_i \sim DP(\alpha G_{0\beta})\\&G_{0\beta} \sim normal(0, v_\beta)\\&\alpha \sim uniform(.)\end{aligned} \tag{4}$$

where G_0 is a Dirichlet baseline density, a prior density selected for the unspecified random density G_β, and α is the Dirichlet precision (concentration) parameter that indicates the degree of similarity between G_β and G_0. This parameter is usually assumed to follow a gamma or a uniform density. Small values of α imply larger departures from the baseline G_0 compared to larger values of α. If G_0 is selected to be a normal density, the model indicates departures from the normality assumption. One may specify another density for G_0 as well, then the model measures departures from that specified density. This paper specifies a normal density for the baseline with mean zero and the variance v_β to be inferred from the data for each random regression parameter. This variance follows a low information prior as specified in (2). In the model formulation (4) and according to a full Dirichlet process, each random coefficient $\beta_1, \beta_2, \ldots, \beta_m$ is an infinite random object being in the form of a mixture of multiple points; that is, a discrete density with a flexible shape that do not necessarily follow any standard parametric density. It can, therefore, accommodate skewness and multimodality, situations where parametric densities may not fit well and hence may result in misleading statistical inferences.

Note that the location of the points (atoms) in the resulting discrete density is obtained based on the pre-specified baseline density G_0. In this paper, probabilities associated with these atoms are estimated based on the stick-breaking algorithm (Ishwaran and James, 2002). The idea is to have a set of sequentially generated random probabilities that sum to one. In this regard, the beta distribution is a convenient density function as it is defined on the interval [0,1]. The precision parameter α comes into play at this point as described below. The above procedures to define the locations of the atoms and their respective probabilities $p_1, p_2, \ldots$ is as follows (Ohlssen et al., 2007):

(i) draw a set of random variables $\theta_1, \theta_2, \ldots$ from G_0;

(ii) draw a set of random variables $\xi_1, \xi_2, \ldots$ from a $Beta(1, \alpha)$;

(iii) allocate probabilities $p_1 = \xi_1$, $p_2 = (1 - \xi_1)\xi_2$, $p_3 = (1 - \xi_1)(1 - \xi_2)\xi_3$, ... to $\theta_1, \theta_2, \theta_3, \ldots$, respectively.

To summarize, given the above discussion, the density $G_\beta \sim DP(\alpha G_0)$ corresponding to each of the m random parameters $\beta_1, \beta_2, \ldots, \beta_m$ can be written as

$$G_\beta = \sum_{j=1}^{\infty} p_j I_{\theta_j}, \quad \theta_j \sim G_0 \tag{5}$$

where I is a measure corresponding to θ_j. Note that G_β is a discrete random density function rather than being a continuous density as universally used in parametric approaches. Drawing inference with respect to such an infinite random measure is computationally cumbersome; therefore, a truncation is usually considered to limit the number of mixtures. Setting J ($J <= n$; where n is the total number of observations) as the maximum number of mass points in (5), one can then write

$$G_\beta = \sum_{j=1}^{J} p_j I_{\theta_j} \approx G_\beta = \sum_{j=1}^{\infty} p_j I_{\theta_j}, \quad \theta_j \sim G_0 \tag{6}$$

Using a truncated (finite) Dirichlet process that approximates the full Dirichlet process reduces the computational burden while retaining a sufficiently flexible model. In this paper, a *uniform*(0.3, 10) prior for the Dirichlet precision parameter with J=50 was chosen. This means that in the proposed model formulation each regression coefficient $\beta_1, \beta_2, \ldots, \beta_m$ is generated from a discrete density that can have up to 50 different values instead of 418 values (corresponding to the total number of sites in the data set) generated from a common continuous density in conventional random effects/parameters models. We will see in the section of results that the specified uniform prior and the selected value of 50 for J are proper choices in the empirical setting of this paper. Note that the proposed model reduces to a finite mixture model. From a practical standpoint, the proposed model could be considered to be at the intersection of finite mixtures and random parameters models. Therefore, it was also compared to a conventional finite mixture negative binomial model (Park and Lord, 2009), based on pseudo Bayes factors discussed in Section 2.5.

2.4. Simulation of posterior densities

The flexible random parameters model specified in Section 2.3 leads to tractable posterior updates for which standard Markov chain Monte Carlo (MCMC) methods can be used. To draw posterior inferences, MCMC simulations are needed because, due to the presence of high dimensional integrals, the model is intractable analytically. WinBUGS (Lunn et al., 2000) was used for MCMC simulations running two chains each containing 150,000 iterations, with a thinning of 5. The first 40,000 iterations were discarded for convergence requirements; therefore, the posterior inferences are based on the final 110,000 of 150,000 total iterations. It was made sure that this number of iterations was satisfactory by checking the Gelman-Rubin statistic (Gelman and Rubin, 1992), history plots, and Monte Carlo errors.

2.5. Predictive performance

With respect to model selection, note that once the range or shape of a parameter is not supported by a parametric assumption, the model corresponding to that inaccurate assumption should be ruled out, and there is no need for a model selection exercise in terms of statistical fit. For a similar discussion in the context of Bayesian nonparametric methods, see Gershman and Blei (2012). This article, however, compares the above models in terms of their predictive power, using a more robust approach compared to the conventional cross-validation, to verify how predictive performance could be affected.

2.5.1. Leave-one-out cross-validation

In conventional cross-validation a data set is randomly grouped into two samples. A model is calibrated based on one of the samples, then its accuracy is validated using the other sample. However, the accuracy of inferences could be sensitive to the choice of these samples (Ntzoufras, 2009). In other words, repeating the above procedure may lead to different conclusions. Conventional cross-validation is also wasteful of data since not all data are used for inferences. Leave-out-one cross-validation avoids this by leaving out only one observation each time. The leave-out-one cross-validation is the base idea for estimating conditional predictive ordinates (CPOs), which is then used to estimate pseudo Bayes factors as described in the subsequent section (Ntzoufras, 2009).

Let Y_i be the i^{th} observation, $\boldsymbol{\psi}$ denote the vector of all model parameters, and t = (1,2,…, T) denote iterations in the MCMC simulations. One can write

$$CPO_i = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{f(Y_i|\psi^{(t)})}\right)^{-1} \tag{7}$$

2.5.2. Pseudo Bayes factors

The Bayes factor is the ratio of marginal likelihoods (the marginal probability of the data y given a model) of two different models, say, M_1 and M_2 (Kass and Raftery, 1995). While the computation of Bayes factors is not straightforward for complex models, other versions of Bayes factors (e.g., pseudo Bayes factors) can be estimated, for example, based on conditional predictive ordinates obtained from (7) (Gelfand, 1996; Ntzoufras, 2009). The log pseudo marginal likelihoods (LPMLs) can be obtained for the entire data set by computing the product of CPOs as in (8). Consequently, log pseudo Bayes factors (LPBFs) for models M_1 and M_2 can be obtained from (9).

A log pseudo Bayes factor of greater than 5 indicates an important difference between M_1 and M_2 in terms of statistical fit (Kass and Raftery, 1995).

$$LPML = log(\prod_{i=1}^{n} CPO_i) = \sum_{i=1}^{n} log(CPO_i) \tag{8}$$

$$LPBF = LPML_{M_1} - LPML_{M_2} \tag{9}$$

2.6. Estimating overall mean and variance for random parameters

Since each mass point of the flexible discrete density has a different probability p assigned to it, the overall (population) mean (μ_β) and variance (V_β) for each regression coefficient β (from the vector of random parameters $\boldsymbol{\beta}$) is calculated at each iteration of the MCMC simulations:

$$\begin{aligned} \mu_\beta &= \sum_{j=1}^{J} p_j \,.\, \beta_j \\ V_\beta &= \left(\sum_{j=1}^{J} p_j \,.\, {\beta_j}^2 \right) - {\mu_\beta}^2 \end{aligned} \tag{10}$$

After running a total of T iterations in the MCMC simulations, posterior mean and variance densities of the random parameters of interest can be obtained from the entire sample using

$$\begin{aligned} \hat{\mu}_\beta &= T^{-1} \sum_{t=1}^{T} \mu_\beta \\ \hat{V}_\beta &= T^{-1} \sum_{t=1}^{T} V_\beta \end{aligned} \tag{11}$$

2.7. An algorithm for identifying sites with similar risk factors

An important advantage of the developed flexible random parameters model is providing inferences about how different sites may be similar in the effect of their risk factors (site characteristics) on safety. This is particularly valuable for an in-depth study of safety mechanisms among different sites, which could also be useful for designing countermeasures more properly. It is important to highlight that identifying sites with similar performance in the effect of their site characteristics on safety does not provide direct insights relating to reasons for such similarities. Nevertheless, this reveals the underlying structure of the data, indicating the need to open up a new line of inquiry that aim at explaining latent data patterns. Further discussion is provided in Section 4.1.2.

Let *i* = {*1, 2,…, N*} and *J* = {*1, 2,…, j*} index, respectively, a set of sites (observations) and a set of clusters among these sites; and L_i denotes an allocation variable that assigns sites *i* to clusters *j*. At each of the *T* iterations of the MCMC simulations and for each pair of sites *i* and *i*′ (*N*x*N* combinations), one can verify whether L_i and $L_{i'}$ are equal, meaning that sites *i* and *i*′ belong to the same cluster *j*. Suppose *I* is an indicator variable, one can write

$$I_{ii'} = \begin{cases} 1 & if \quad L_i = L_{i'} \\ 0 & if \quad L_i \neq L_{i'} \end{cases} \tag{12}$$

Averaging over *T* iterations of the MCMC simulations, an *N*x*N* similarity matrix *SIM* can be obtained. The cells of the resulting matrix are basically pairwise probabilities of similarities of sites (here, highway segments) in the crash data.

$$SIM = T^{-1} \sum_{t=1}^{T} I_{ii'} \tag{13}$$

In this research, since a unique Dirichlet process for all random parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_m)$ is used, then, if $L_i = L_{i'}$, their respective random parameters are similar; i.e., $(\beta_{i1}, \beta_{i2}, \dots, \beta_{im}) = (\beta_{i'1}, \beta_{i'2}, \dots, \beta_{i'm})$. The above algorithm is fully probabilistic; hence, it better reflects real data scenarios.

Based on (12), one could also identify outlying sites (i.e., those performing very differently from the rest of the sample in the effect of their characteristics on safety), estimating the total number of sites with similar risk factors profiles. Obviously, when *i*=*i*′, the pairwise probability of similarly is 1. One can thus write

$$Total = \sum\nolimits_{i \neq i'} I_{ii'} \tag{14}$$

As in (13), the statistic obtained in (14) is averaged over all iterations. A value close to one indicates an outlying site as there are no other sites that are similar to this outlier (Ohlssen et al., 2007).

2.8. Marginal effects

Marginal effects can be used to interpret the association between crash frequencies and site characteristics. With respect to model interpretation, this research investigates the impact of model formulation on the estimated marginal effects, which indicate the effect of one unit change in an independent variable on the outcome of interest; e.g., cash frequencies. In this paper, the average marginal effects can be obtained for the m^{th} random parameters $\boldsymbol{\beta}_i$ and fixed parameters γ from (15) and (16), respectively.

$$\frac{\partial E(y|\boldsymbol{Z_i}, \boldsymbol{X_i}, \boldsymbol{\beta_i}, \boldsymbol{\gamma})}{\partial (Z_m)} = \frac{1}{N}\sum_{i=1}^{N} \beta_{im} exp(\,\boldsymbol{\beta_i}\boldsymbol{Z_i} + \,\boldsymbol{\gamma}\boldsymbol{X_i}) \tag{15}$$

$$\frac{\partial E(y|\boldsymbol{Z_i}, \boldsymbol{X_i}, \boldsymbol{\beta_i}, \boldsymbol{\gamma})}{\partial (X_m)} = \frac{1}{N}\sum_{i=1}^{N} \gamma_{m} exp(\,\boldsymbol{\beta_i}\boldsymbol{Z_i} + \,\boldsymbol{\gamma}\boldsymbol{X_i}) \tag{16}$$

Under a Bayesian framework, the posterior distribution of the marginal effect can be inferred, for example, for random parameters $\boldsymbol{Z}$ by averaging the calculated values at each iteration t of the MCMC simulations:

$$\left\{\frac{\partial E(y|\boldsymbol{Z_i}, \boldsymbol{X_i}, \boldsymbol{\beta}_i^{(t)}, \boldsymbol{\gamma}^{(t)})}{\partial (Z_m)}\right\}_{t=1}^{T} \tag{17}$$

Therefore, Bayes estimates of marginal effects could be obtained in the form of posterior densities, providing a fuller picture compared to point estimates often obtained from classical estimates.

3. Empirical setting

This paper illustrates the ideas discussed above with an example data set (provided by the Ontario Ministry of Transportation) containing 418 highway segments in Ontario, Canada. Specifically, the crash data were obtained from Highway 401, which is one of the busiest North American highways, from 2006 to 2008. This highway connects the Ontario-Quebec border to the Ontario-Michigan border, passing through the Greater Toronto Area. In the aforementioned three-year period, the data set recorded 29,148 crashes, which result in huge amount of monetary and non-monetary costs. Besides crash counts, a number of operational and geometric segment characteristics such as average annual daily traffic, median (inside) shoulder width, outside shoulder width, and average horizontal curve degree curvature per km were available as well. The outcome of interest is the total crash frequency during a three-year period. Descriptive statistics are provided in Table 1. A histogram of the crash frequency is displayed in Fig. 1.

Several combinations were tested and the final model reported in the section of results was the best. Since a relatively limited site characteristics appear to be statistically important in the model, the omitted variable issue may arise. Although it was attempted to include the most significant variables that have an important effect on safety, having a limited number of variables in the model could be a limitation of this study, and perhaps several other previous traffic safety studies. A discussion in this regard is provided in previous research; for example, see Jovanis et al. (2011), Mitra and Washington (2012), Mannering and Bhat (2014), and Wu et al. (2015). It is discussed in the crash literature that statistical techniques such as random parameters models could mitigate adverse consequences of the omitted variables problem (Mannering and Bhat, 2014; Heydari et al., 2016b). However, whether these techniques are satisfactory remains uncertain and the problem may not be fully addressed. Therefore, it is important to recognize the importance of the omitted variables bias in traffic safety research. Also, note that temporal instability could play a role in the results as the effect of explanatory variables may vary by time during the study period. Therefore, accounting for temporal instability could better address unobserved heterogeneity (Mannering, 2018). While the general importance of addressing temporal instability in traffic safety research is recognized, the topic is beyond the scope of this paper. It was not possible to include some site characteristics in the model at the same time because of high co-linearity.

4. Results and discussion

This section reports the results obtained from the estimation of a standard random intercepts model, a standard random parameters model, and the proposed flexible discrete density random parameters model. Non-informative priors were used for model parameters to minimize the effect of prior information on the final posterior estimates. Sensitivity to prior choice is particularly important for the Dirichlet precision parameter, which measures the similarity between an unknown density and its baseline as discussed in Section 2.3. A sensitivity analysis was conducted for this parameter using a uniform prior on the range (0, 20) and did not result in any important variation in the results. With respect to conventional finite mixture modeling, a 2-component negative binomial model was found to be more appropriate than a 3-component finite mixture model. However, based on the data set analyzed here, pseudo Bayes factors ruled out the finite mixture model with 2 components (which provides a log pseudo marginal likelihood of -1705), so the finite mixture model will not be discussed further.

The results are summarized in Table 2. The results obtained from the conventional random parameters model support varying effects for the intercepts, median shoulder width, and average horizontal curve degree curvature. This means that the effects of some site characteristics vary across the sample, indicating heterogeneity across highway segments. The predictive performance of the model improves relative to the random intercepts model, considering the log pseudo marginal likelihoods reported in Table 2. In fact, a log pseudo Bayes factor of 5.95 provides support for the random parameters model.

While the conventional random parameters model employed in this paper assumes that random parameters follow a normal density, the proposed model relaxes this assumption, fitting flexible discrete densities to these random parameters. The posterior density of the Dirichlet precision parameter (displayed in Fig. 2) is peaked away from 0 and 10 indicating that the normality assumption is not suitable here. The estimated posterior mean of the precision parameter is 2.65 with a 95% credible interval of [1.20, 4.74] based on a *uniform*(0.3, 10) prior. The estimated expected median number of mass points for the fitted discrete densities is 13 with a 95% credible interval of [10, 16]. The estimated standard deviations for the baseline densities (in the flexible discrete density model) are 13.6, 2.34, and 1.49 for random parameters associated with intercept, median shoulder width, and horizontal curve, respectively. The predictive power improves significantly under the flexible random parameters model as the log pseudo marginal likelihood increases to -1607.64 from -1617.71 in the conventional random parameters model. This leads to a log pseudo Bayes factor of around 10, providing strong support for the flexible model.

It should be noted that the normality assumption for random parameters may cause undue shrinkage towards the overall mean for some sites; however, our flexible model prevents this problem. This is particularly important in the presence of outliers (sites performing very differently from the rest of the data), meaning that outlying sites can be accommodated in the analysis without the need to removing them from the data. As the potential presence of outlying sites is often ignored in road safety studies, the proposed approach could constitute a more robust statistical approach by accommodating them without compromising the results. A similar discussion in this regard is provided by Ohlssen et al. (2007) who used a flexible random intercepts model in the context of medical research.

In general, the posterior means of regression coefficients are less or more similar among the two random parameters models except for the median shoulder width. However, a major difference is that the flexible model accounts for a much greater spread of the data, capturing unobserved heterogeneity more fully. Fig. 3 displays the posterior masses of the random parameters $p(\boldsymbol{\beta}|y)$ obtained from the standard and the flexible random parameters models, highlighting their differences in inferring the shape and the range of the random parameters. Fig. 3 implies that the conventional random parameters model falls short in covering the range of the parameters, only partly capturing unobserved heterogeneity. For example, Fig. (3a) shows that the random parameter associated with horizontal curvature varies from around -13 to +7 under the flexible model; however, this parameter varies from around -5 to +4 under the conventional model. This variation is much larger for the vector of intercepts.

The posterior mean estimates of the variances associated with random parameters under the flexible model are larger than those obtained from the standard random parameters model (see Table 2). In accordance with the above discussion regarding the estimated Dirichlet precision parameter, asymmetry and multimodality of random regression parameters (see the left-hand side of Fig. 3) indicate that, based on the data analyzed in this paper, parametric assumptions may be limited in reflecting the underlying structure of crash data sets. As it is discussed in the next section, this limitation could have practical implications such as affecting the effectiveness of countermeasures. One should however take into account that the proposed model may become computationally intensive compared to many current methods commonly used in traffic safety research. It is also important to emphasize that methods such as standard random parameters models should not be ruled out as they may approximate the density of a random regression parameter properly, specifically when for example non-normal densities are considered.

4.1. Practical implications

An important difference observed here is in the range of random parameters although the posterior means indicate that the overall mean of the population is negative for median shoulder with and average horizontal curve degree curvature. This difference has interesting practical implications, for example, with respect to selecting appropriate countermeasures for different sites. According to the standard random parameters model, which erroneously assumes a normally distributed random density for median shoulder width, 15.04% of the sites in the data manifest a positive association between crash frequencies and median shoulder width (this is the portion of density that is greater than zero). However, according to the flexible discrete density random parameters model, 34.02% of the sites manifest such positive association. The difference between these two estimates is relatively large; i.e., 18.98%. Suppose a safety treatment consisting in modifying median shoulder width is to be implemented. Obviously, this discrepancy could distort the overall expected benefits.

With regard to average horizontal curve degree curvature, the proportions of the sites having a coefficient greater than zero are 14.94% and 23.70% for the random parameters model and the flexible random parameters model, respectively. The overall negative sign for this variable could be justified based on an increased driver awareness in relatively sharper curves. This is in accordance with previous research (see, for example, Anastasopoulos and Mannering, 2009). Relative to the random intercepts model that only supports a negative sign for this variable across the entire sample, however, both random parameters models draw a more realistic picture, revealing that, for some sites, crash frequencies could increase as horizontal curve degree curvature increases. Note that, when discussing the range of random parameters, a more holistic approach would be to consider a third group of intervals on the interval [-∞, ∞] that includes a region around zero where a covariate of interest does not have an important effect on safety. However, this requires specifying two arbitrary values around zero by the analyst, creating an interval that is considered as the non-significance region.

4.1.1. Magnitude of association between site characteristics and safety

To interpret the magnitude of association between the expected crash frequency and site characteristics, average marginal effects were computed according to Section 2.8. The posterior summary of marginal effects is reported in Table 3. One advantage of the data set used in this research is that the average crash frequency is relatively large so that the differences between different models (and site characteristics) could be better highlighted as estimated marginal effects are large as well. Apparently, AADT has the largest impact on crash frequencies under the three models examined here. Relative to both random parameters model, the random intercepts model slightly underestimates the average effect of traffic exposure on crash frequencies by around 2 crashes in a 3-year period. This difference is much larger for average horizontal curve degree curvature: around 22 crashes in a 3-year period relative to the proposed flexible discrete density model. Marginal effects are more similar among the two random parameters models, but the effect of horizontal curve is underestimated in the conventional random parameters model relative to the flexible model (-43.99 vs. -52.54).

A large difference can be observed in the standard deviation estimates of the marginal effects (Table 3). For both non-varying regression coefficients (those associated with AADT and segment length), the flexible discrete density model provides smaller standard deviations. In contrast, the latter model provides larger standard deviations for random parameters (those associated with median shoulder with and horizontal curve). This can be explained by the fact that the conventional models considered in this study may be limited in accommodating the range of random parameters. For the non-varying parameters, standard deviations are small relative to their posterior means, suggesting a positive association between these parameters and safety for all highway segments. However, relatively large standard deviations of marginal effects for random parameters indicate that the effect of median shoulder width and horizontal curve can vary considerably in the sample.

4.1.2. Sites performing similarly

An interesting advantage of the proposed model is its ability to identify sites with similar covariate effects (risk factors). This allows the identification of sites that could be affected similarly by implementing similar countermeasures. As discussed in Section 2.7, pairwise probabilities of similarity can be estimated to identify highway segments having similar random regression parameters profiles (here, intercept, median shoulder width, and horizontal curve degree curvature). Recall that if a Dirichlet process prior is placed on all regression coefficients, sites exhibiting similar expected crash frequency profiles can be identified. A pairwise probability plot for a sample of sites is displayed in Fig. 4. This plot implies that, for example, segments 127 and 149 across Highway 401 are similar in the effects of their random parameters with a probability of 74%. As previously discussed in Section 2.7, identifying sites with similar covariate effects stimulates further investigations that could lead to find reasons for which, say, some highway segments are similar in the effect of traffic flow on crash frequencies of a specific type. One may find that these segments are also similar in some other features that were initially unnoticed. For example, they may be similar in the formation of microclimates among them or in the presence of entrance/exit ramps as well. This means that such variables should be tested in the model in the future since they may have a bearing on safety. Therefore, identifying sites performing similarly could also provide insights relating to future data collection activities.

5. Summary and conclusions

In traffic safety research, there has been considerable attention devoted of late to the use of random parameters models to overcome unobserved heterogeneity. In the random parameters approach, it is often necessary to confirm the accuracy of parametric assumptions (e.g., normality of random parameters), which may otherwise compromise the estimation of the effect of explanatory variables on safety. This paper introduces a novel data-driven random parameters approach, rooted in Bayesian nonparametric literature, to model count data. The proposed model overcomes sensitivity to distributional assumptions by fitting a discrete density (instead of a pre-specified continuous density) to random parameters based on a Dirichlet process mixing approach.

The methodology is applied to a highway segment data set from Ontario, Canada. The proposed model is compared with some of the commonly used approaches in traffic safety research such as a random intercepts model and a random parameters (slopes) model. In this work, while heterogeneity across highway segments is mostly ignored by the random intercepts model, it is partially captured by the conventional random parameters model. It is shown that the proposed flexible discrete density model accommodates unobserved heterogeneity more properly relative to the standard random parameters model. The proposed discrete density model provided the best statistical fit according to the estimated pseudo Bayes factors, derived from a more robust leave-out-one cross-validation exercise (compared to conventional cross-validation).

The fitted models were compared in terms of the estimated range of the random parameters and their posterior masses of marginal effects. In this research, in general, the random intercepts model provided the least accurate estimates; for example, underestimating the effect (and the standard deviation of the effect) of horizontal curve degree curvature on highway safety. The results indicate that the standard random parameters models can also compromise the accuracy of estimates if parametric assumptions do not hold. Marginal effects were also affected by different model formulations. Besides coping with situations where parametric assumptions are inappropriate (e.g., due to asymmetry or multimodality), the flexible random parameters model identifies sites that manifest similarity in the effect of their risk factors, providing new guidance that could be useful in designing safety treatments. A further in-depth study can be conducted to reveal the reasons for such similarities or dissimilarities among sites in the data. This research suggests that the proposed discrete density approach, as a feasible alternative to current approaches, provides promise in modeling count data and addressing unobserved heterogeneity by drawing important and interesting insights from the data.

Acknowledgments

The author would like to thank the Ministry of Transportation Ontario for providing the data. The author is grateful to Professor Lawrence Joseph of McGill University, anonymous reviewers, and the Editor-in-Chief for their valuable comments.

References

Alarifi, S., Abdel-Aty, M., Lee, J., Park, J., 2017. Crash modeling for intersections and segments along corridors: a Bayesian multilevel joint model with random parameters. Analytic Methods in Accident Research 16, 48-59.

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 and Prevention 50, 289-297.

Anastasopoulos, P., 2016. Random parameters multivariate tobit and zero-inflated count data models: addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. Analytic Methods in Accident Research 11, 17-32.

Anastasopoulos, P., Mannering, F., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153-9.

Behnood, A., Mannering, F., 2017a. The effect of passengers on driver-injury severities in single-vehicle crashes: a random parameters heterogeneity-in-means approach. Analytic Methods in Accident Research 14, 41-53.

Behnood, A., Mannering, F., 2017b. Determinants of bicyclist injury severities in bicycle-vehicle crashes: a random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research 16, 35-47.

Behnood, A., Roshandeh, A., Mannering, F., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Analytic Methods in Accident Research 3–4, 56–91.

Bhat, C., Astroza, S., Lavieri, P., 2017. A new spatial and flexible multivariate random-coefficients model for the analysis of pedestrian injury counts by severity level. Analytic Methods in Accident Research 16, 1-22.

Bogue, S., Paleti, R., Balan, L., 2017. A modified rank ordered logit model to analyze injury severity of occupants in multivehicle crashes. Analytic Methods in Accident Research 14, 22-40.

Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., Wang, X., 2018. Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. Analytic Methods in Accident Research 19, 1-15.

Canale, A., Prunster, I., 2017. Robustifying Bayesian nonparametric mixtures for count data. Biometrics 73, 174–184.

Carota, C., Parmiggiani, G., 2002. Semiparameteric regression for count data. Biometrics 89 (2), 265-281.

Chen, S., Saeed, T., Labi, S., 2017. Impact of road-surface condition on rural highway safety: a multivariate random parameters negative binomial approach. Analytic Methods in Accident Research 16, 75-89.

Cheng, W., Gill, G., Vo, T., Zhou, J., Sakrani, T., 2018. Use of bivariate Dirichlet process mixture spatial model to estimate active transportation-related crash counts. Transportation Research Record. https://doi.org/10.1177/0361198118782797

Dhavala, S., Mallick, B., Carroll, R., Datta, S., Khare, S., Lawhon, S., Adams, L., 2010. Bayesian modeling of MPSS data: gene expression analysis of bovine salmonella infection. Journal of the American Statistical Association 105 (491), 956-967.

El-Basyouny, K., Sayed, T., 2009. Accident predicyion models with random corridor parameters. Accident Analysis and Prevention 41 (5), 1118-1123.

Escobar, M., West, M., 1998. Computing nonparametric hierarchical models. In: Dey, D., Müller, P., Sinha, D., eds. Practical nonparametric and semiparametric Bayesian statistics. Springer New York, New York, NY, 1-22.

Ferguson, T., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1 (2), 209-230.

Fountas, G., Anastasopoulos, P., 2017. A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. Analytic Methods in Accident Research 15, 1-16.

Fountas, G., Anastasopoulos, P., Abdel-Aty, M., 2018. Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. Analytic Methods in Accident Research 18, 57-68.

Freedman, D., 1963. On the asymptotic behavior of Bayes estimates in the discrete case. Annals of Mathematical Statistics 34 (4), 1386-1403.

Geedipally, S., Lord, D., Dhavala, S., 2014. A caution about using deviance information criterion while modelling traffic crashes. Safety Science 62, 495-498.

Gelfand, A., 1996. Model determination using sampling-based methods, in Gilks, W., Richardson, S., Spiegelhalter, D., eds., Markov Chain Monte Carlo in Practice, Chapman and Hall, Suffolk.

Gelman, A., Rubin, D., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7 (4), 457-472.

Gershman, S., Blei, D., 2012. A tutorial on Bayesian nonparametric models. Journal of Mathematical Psychology 56 (1), 1-12.

Guindani, M., Sepulveda, N., Paulino, C., Muller, P., 2014. A Bayesian semiparametric approach for the differential analysis of sequence counts data. Journal of the Royal Statistical Society, Series C (Applied Statistics) 63, 385–404.

Heydari, S., Fu, L., Joseph, L., Miranda-Moreno, L., 2016a. Bayesian nonparametric modeling in transportation safety studies: applications in univariate and multivariate settings. Analytic Methods in Accident Research 12, 18-34.

Heydari, S., Fu, L., Lord, D., Mallick, B., 2016b. Multilevel Dirichlet process mixture analysis of railway grade crossing crash data. Analytic Methods in Accident Research 9, 27-43.

Heydari, S., Fu, L., Miranda-Moreno, L., Joseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: a joint analysis of pedestrian and cyclist injuries. Analytic Methods in Accident Research 13, 16-27.

Heydari, S., Fu, L., Thakali, L., Joseph, L., 2018. Benchmarking regions using a heteroskedastic grouped random parameters model with heterogeneity in mean and variance: applications to grade crossing safety analysis. Analytic Methods in Accident Research 19, 33-48.

Hjort, N., Holmes, C., Müller, P., Walker, S., 2010. Bayesian nonparametrics: principles and practice. Cambridge University Press.

Ishwaran, H., James, L., 2002. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96 (453), 161-173.

Jovanis, P., Aguero-Valverde, J., Wu, K.-F, Shankar, V., 2011. Analysis of naturalistic driving event data: omitted-variable bias and multilevel modeling approaches. Transportation Research Record: Journal of the Transportation Research Board 2236, 49-57.

Kass, R., Raftery, A., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773-795.

Khazraee, H., Johnson, V., Lord, D., 2018. Bayesian Poisson hierarchical models for crash data analysis: investigating the impact of model choice on site-specific predictions. Accident Analysis and Prevention 117, 181-195.

Kim, D., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accident Analysis and Prevention 39 (1), 125-134.

Ladouceur, M., Rahme, E., Belisle, P., Scott, A., Schwartzman, K., Joseph, L., 2011. Modeling continuous diagnostic test data using approximate Dirichlet process distributions. Statistics in Medicine 30, 2648-2662.

Li, Z., Chen, C., Wu, Q., Zhang, G., Liu, C., Prevedouros, P., Ma, D., 2018. Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. Analytic Methods in Accident Research 20, 1-14.

Lord, D., Mannering, F., 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., Miranda-Moreno, L., 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, 751-770.

Lunn, D., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10 (4), 325-337.

Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research 17, 1-13.

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

Mannering, F., Shankar, V., Bhat, C., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1-16.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accident Analysis and Prevention 49, 439-448.

Mukhopadhyay, S., Gelfand, A., 1997. Dirichlet process mixed generalized linear models. Journal of the American Statistical Association 92 (438), 633-639.

Müller, P., Quintana, F., 2004. Nonparametric Bayesian data analysis. Statistical Science 19 (1), 95–110.

Naznin, F., Currie, G., Logan, D., Sarvi, M., 2016. Application of a random effects negative binomial model to examine tram-involved crash frequency on route sections in Melbourne, Australia. Accident Analysis and Prevention 92, 15-21.

Neal, R., 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9 (2), 249-265.

Ntzoufras, I., 2009. Bayesian Modeling Using WinBUGS. Wiley Series in Computational Statistics, Hoboken, USA.

Ohlssen, D., Sharples, L., Spiegelhalter, D., 2007. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. Statistics in Medicine 26 (9), 2088-2112.

Park, B., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention 41 (4), 683-691.

Park, B., Lord, D., Wu, L., 2016. Finite mixture modeling approach for developing crash modification factors in highway safety analysis. Accident Analysis and Prevention 97, 274-287.

Seraneeprakarn, P., Huang, S., Shankar, V., Mannering, F., Venkataraman, N., Milton, J., 2017. Occupant injury severities in hybrid-vehicle involved crashes: a random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research 15, 41-55.

Sarwar, M., Anastasopoulos, P., Golshani, N., Hulme, K., 2017a. Grouped random parameters bivariate probit analysis of perceived and observed aggressive driving behavior: a driving simulation study. Analytic Methods in Accident Research 13, 52-64.

Sarwar, M., Fountas, G., Anastasopoulos, P., 2017b. Simultaneous estimation of discrete outcome and continuous dependent variable equations: a bivariate random effects modeling approach with unrestricted instruments. Analytic Methods in Accident Research 16, 23-34.

Shankar, V., Albin, R., Milton, J., Mannering, F., 1998. Evaluating median crossover likelihoods with clustered accident counts an empirical inquiry using the random effects negative binomial model. Transportation Research Recor: Journal of the Transportation Research Board 1635, 44-48.

Shaon, M., Qin, X., Shirazi, M., Lord, D., Geedipally, S., 2018. Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data. Analytic Methods in Accident Research 18, 33-44.

Shirazi, M., Lord, D., Dhaval, S., Geedipally, S., 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 and Prevention 91, 10-18.

Venkataraman, N., Ulfarsson, G., Shankar, V., Deptuch, D., 2014. A heterogeneity-in-means count model for evaluating the effects of interchange type on heterogeneous influences of interstate geometrics on crash frequencies. Analytic Methods in Accident Research 2, 12-20.

Walker, S., Adrian, F.., Damien, P., Laud, P., 1999. Bayesian nonparametric inference for random distributions and related functions. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61 (3), 485-527.

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

Winkelmann, R., 2008. Econometric Analysis of Count Data, 5th edition. Springer-Verlag Berline Heidelberg.

Wu, Z., Sharma, A., Mannering, F., Wang, S., 2013. Safety impacts of signal-warning flashers and speed control at high-speed signalized intersections. Accident Analysis and Prevention 54, 90–98.

Wu, L., Lord, D., Zou, Y., 2015. Validation of crash modification factors derived from cross-sectional studies with regression models. Transportation Research Record: Journal of the Transportation Research Board 2514, 88-96.

Xin, C., Guo, R., Wang, Z., Lu, Q., Lin, P., 2017. The effects of neighborhood characteristics and the built environment on pedestrian injury severity: a random parameters generalized ordered probability model with heterogeneity in means and variances. Analytic Methods in Accident Research 16, 117-132.

Xiong, Y., Mannering, F., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: a finite-mixture random-parameters approach. Transportation Research Part B 49, 39-54.

Yasmin, S., Eluru, N., 2016. Latent segmentation based count models: analysis of bicycle safety in Montreal and Toronto. Accident Analysis and Prevention 95, 157-171.

Yu, R., Wang, X., Yang, K., Abdel-Aty, M., 2016. Crash risk analysis for Shanghai urban expressways: a Bayesian semi-parametric modeling approach. Accident Analysis and Prevention 95, 495-502.

Zou, Y., Ash, J., Park, B., Lord, D., Wu, L., 2017. Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety. Journal of Applied Statistics, DOI: 10.1080/02664763.2017.1389863.

Table 1. Summary statistics for the Ontario Highway 401 data (418 observations).

Variables	Mean	Std. Dev.	Min	Max
AADT all vehicles (vehicles per day)	80369.420	95760.440	14499.940	442900.300
AADT commercial vehicles (vehicles per day)	14383.640	6890.880	4864.000	42075.500
Percentage of commercial vehicles	29.027	12.300	3.100	49.100
Segment length (km)	1.952	2.061	0.206	12.703
Number of lanes	5.445	2.428	4.000	12.000
Median (inside) shoulder width (m)	1.598	1.194	0.000	5.190
Median width (m)	11.106	6.147	0.600	30.500
Outside shoulder width (m)	3.135	0.285	2.600	4.000
Lane width (m)	3.707	0.301	1.830	5.625
Average horizontal curve degree curvature per km	0.945	1.864	0	16.592
Paved outside shoulder (1 if paved; 0 otherwise)	0.586	0.493	0.000	1.000
Surface type (1 if HCB[1]; 0 otherwise)	0.526	0.500	0.000	1.000
Crash frequency (3-year period)	69.732	138.975	0	1098

[1] HCB stands for high class bituminous pavement.

Table 2. Posterior estimation summary of model coefficients

	Posterior		95% Credible intervals	
	Mean	Std. Dev.	2.50%	97.50%
Random intercepts model				
Random intercepts mean	-10.340	0.479	-11.180	-9.236
Variance	*0.517*	*0.043*	*0.437*	*0.609*
ln(AADT)	1.248	0.043	1.150	1.322
ln(length)	0.802	0.050	0.705	0.899
Median shoulder width (in tenths of meters)	-0.643	0.318	-1.257	-0.008
Average horizontal curve degree curvature per km	-0.442	0.121	-0.680	-0.208
Model fit (log pseudo marginal likelihood)	-1623.66	-	-	-
Random parameters model				
Random intercepts mean	-10.550	0.472	-11.510	-9.625
Variance	*0.393*	*0.056*	*0.295*	*0.516*
ln(AADT)	1.270	0.042	1.187	1.353
ln(length)	0.792	0.047	0.702	0.887
Median shoulder width (in tenths of meters)	-0.681	0.323	-1.323	-0.038
Variance	*0.873*	*0.037*	*0.007*	*2.911*
Average horizontal curve degree curvature per km	-0.663	0.161	-0.985	-0.356
Variance	*0.755*	*0.286*	*0.268*	*1.387*
Model fit (log pseudo marginal likelihood)	-1617.71	-	-	-
Flexible discrete density random parameters model				
Random intercepts mean	-10.700	0.325	-11.360	-10.090
Variance	*4.001*	*3.324*	*0.909*	*12.770*
ln(AADT)	1.278	0.027	1.228	1.332
ln(length)	0.772	0.020	0.731	0.813
Median shoulder width (in tenths of meters)	-0.430	0.309	-1.110	0.077
Variance	*2.959*	*3.205*	*0.036*	*9.687*
Average horizontal curve degree curvature per km	-0.658	0.179	-1.024	-0.319
Variance	*1.426*	*0.919*	*0.186*	*3.851*
Model fit (log pseudo marginal likelihood)	-1607.64	-	-	-

Table 3. Posterior summary of average marginal effects

	Posterior Mean	Std. Dev.
Random intercepts model		
ln(AADT)	86.96	2.86
ln(length)	55.89	3.44
Median shoulder width (in tenths of meters)	-45.43	21.94
Average horizontal curve degree curvature per km	-30.80	8.43
Random parameters model		
ln(AADT)	88.55	2.96
ln(length)	55.25	3.29
Median shoulder width (in tenths of meters)	-42.88	22.38
Average horizontal curve degree curvature per km	-43.99	12.56
Flexible discrete density random parameters model		
ln(AADT)	89.12	1.96
ln(length)	53.80	1.46
Median shoulder width (in tenths of meters)	-43.18	28.31
Average horizontal curve degree curvature per km	-52.54	20.96

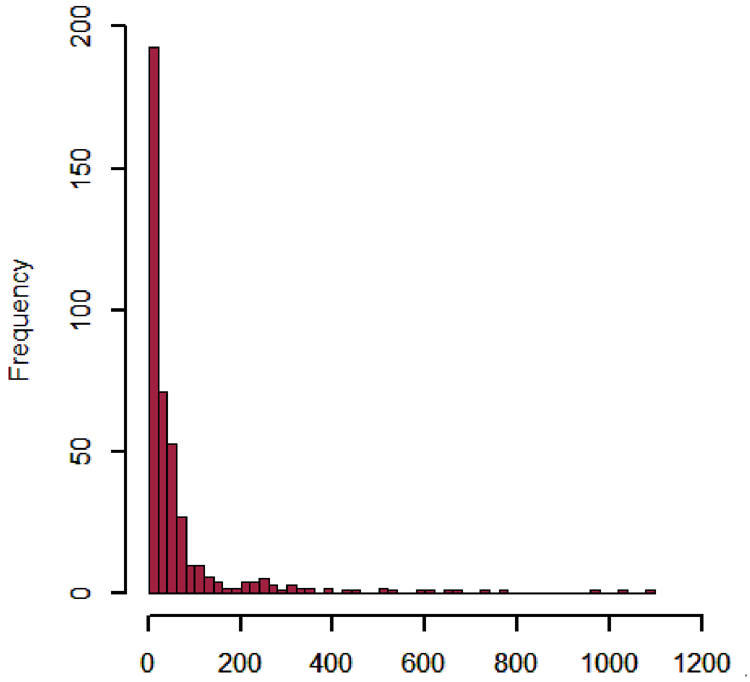

Figure 1. Histogram of observed crash data

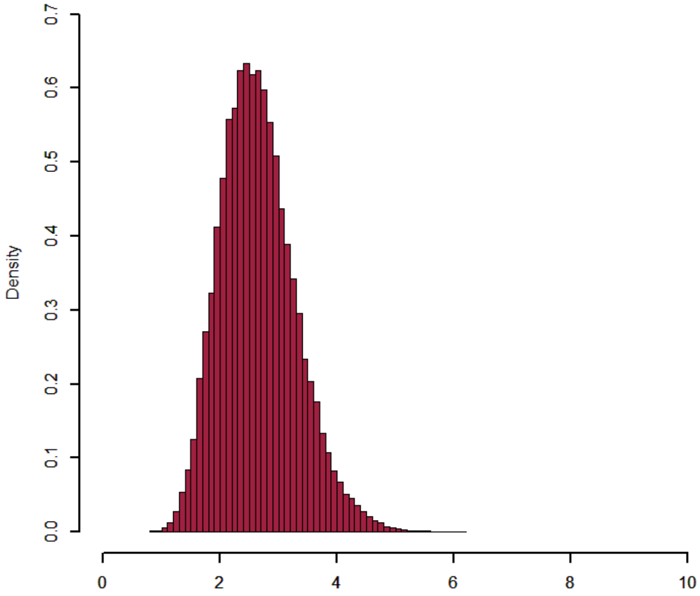

Figure 2. Histogram of posterior density for Dirichlet precision parameter

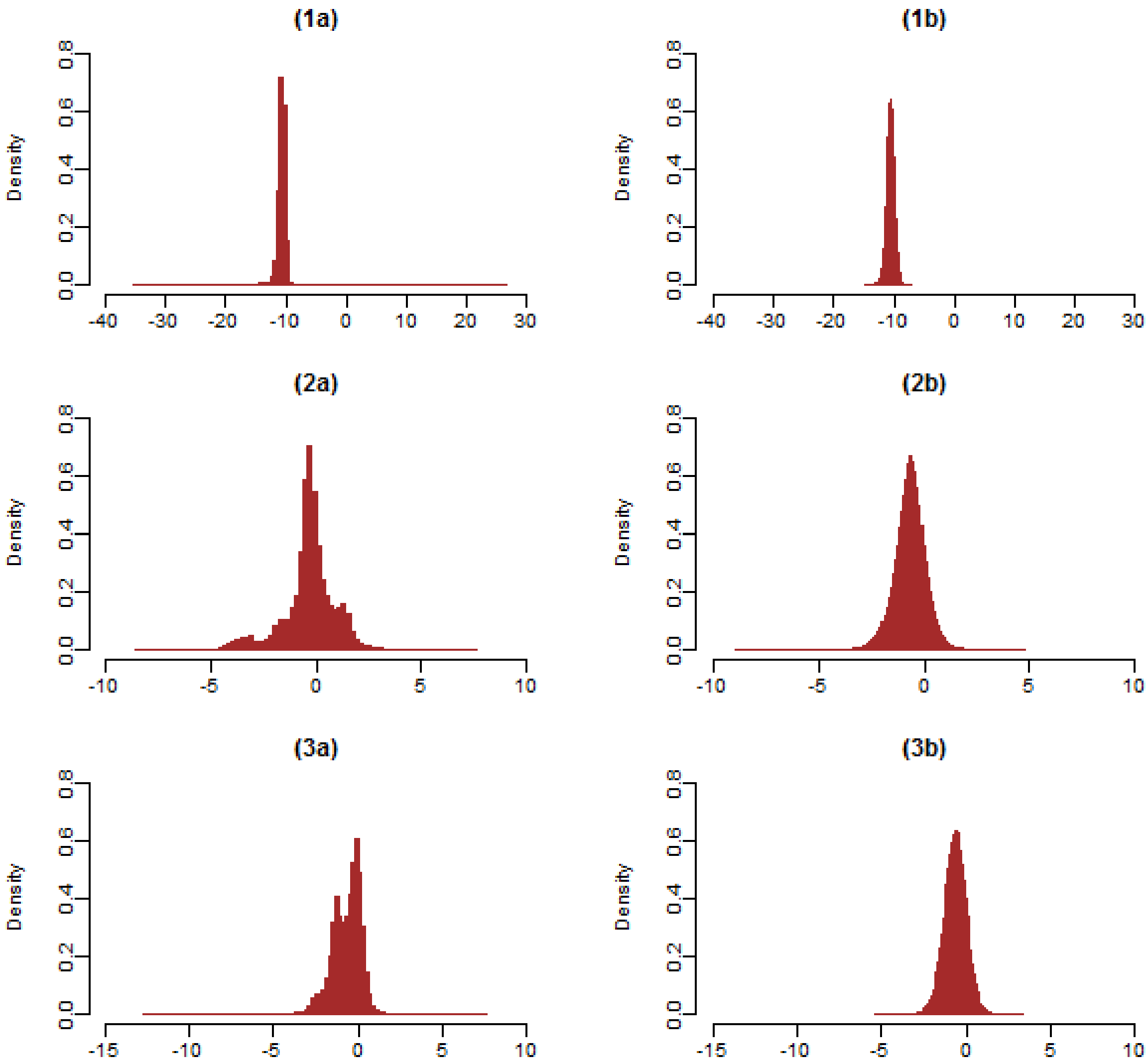

Figure 3. Histograms of posterior masses:
(1) varying intercepts; (2) median shoulder width; and (3) average horizontal curve degree curvature for (a) flexible discrete density random parameters model and (b) standard random parameters model

Highway segment ID

Highway segment ID	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139
138	0.02	0.00	0.00	0.04	0.00	0.06	0.00	0.06	0.00	0.00	0.10	0.00	0.88	1.00	0.01
139	0.06	0.30	0.02	0.25	0.36	0.02	0.02	0.00	0.11	0.02	0.00	0.38	0.01	0.01	1.00
140	0.24	0.24	0.34	0.19	0.23	0.08	0.24	0.22	0.44	0.15	0.14	0.00	0.00	0.00	0.06
141	0.18	0.22	0.48	0.26	0.09	0.14	0.41	0.35	0.32	0.29	0.26	0.00	0.00	0.00	0.04
142	0.17	0.25	0.04	0.14	0.29	0.06	0.07	0.00	0.11	0.07	0.00	0.38	0.01	0.01	0.42
143	0.25	0.25	0.27	0.17	0.29	0.06	0.19	0.17	0.46	0.12	0.11	0.01	0.00	0.00	0.09
144	0.17	0.18	0.55	0.29	0.02	0.17	0.51	0.41	0.27	0.35	0.30	0.00	0.00	0.00	0.01
145	0.23	0.30	0.20	0.17	0.36	0.04	0.13	0.12	0.41	0.07	0.08	0.05	0.00	0.00	0.19
146	0.02	0.01	0.04	0.04	0.00	0.03	0.00	0.03	0.00	0.04	0.06	0.02	0.33	0.31	0.02
147	0.15	0.01	0.00	0.09	0.00	0.00	0.00	0.07	0.09	0.00	0.01	0.38	0.00	0.00	0.04
148	0.34	0.06	0.11	0.05	0.00	0.58	0.44	0.09	0.09	0.48	0.15	0.00	0.01	0.01	0.04
149	0.04	0.25	0.74	0.40	0.04	0.02	0.50	0.59	0.28	0.35	0.45	0.00	0.00	0.00	0.01
150	0.09	0.30	0.01	0.22	0.34	0.00	0.00	0.00	0.08	0.00	0.00	0.42	0.00	0.00	0.64
151	0.08	0.22	0.68	0.37	0.02	0.09	0.51	0.55	0.26	0.35	0.42	0.00	0.00	0.00	0.01
152	0.21	0.12	0.32	0.17	0.02	0.35	0.42	0.27	0.15	0.41	0.25	0.01	0.02	0.02	0.02

Figure 4. Pairwise probabilities of similarities according to random regression coefficients profiles for a sample of highway 401 segments – darker cells indicate higher probabilities