Spatial Negative Binomial Bayesian Additive Regression Trees for Accident Hot Spot Identification

15 February 2020

Rico Krueger
Transport and Mobility Laboratory
Ecole Polytechnique Fédérale de Lausanne, Switzerland
rico.krueger@epfl.ch

Prateek Bansal
Transport Strategy Centre, Department of Civil and Environmental Engineering
Imperial College London, UK
prateek.bansal@imperial.ac.uk

1 Introduction

The identification of accident-prone locations, so-called hot spots, in road networks represents a core task of road safety management (Cheng et al., 2020; Huang et al., 2009). Crash frequency models are used to produce model-based rankings of hazardous sites and to predict crash counts at hot spots under counterfactual traffic flow and road design scenarios (Deacon et al., 1975).

Crash counts are typically modelled using Poisson log-normal or negative binomial regression models (Lord and Mannering, 2010). Accommodating flexible representations of unobserved heterogeneity in model parameters and accounting for correlations between spatial units are central themes of the crash count modelling literature in the past decade (Cheng et al., 2020; Dong et al., 2016; Mannering et al., 2016; Ziakopoulos and Yannis, 2020). However, these flexible representations of unobserved heterogeneity are achieved at the cost of a restrictive linear specification of the link function. Whilst linear-in-parameters link functions are appealing from an interpretability perspective, an over-simplification of the relationship between predictors may negatively affect the predictive performance of a model (Li et al., 2008; Huang et al., 2016).

Since the predictive performance of a model is of paramount importance in site-ranking and hot spot identification applications, the specification of the link function should be carefully selected. However, in practice, the space of possible link function specification is prohibitively large, precluding exhaustive specification searches. Modern machine learning (ML) methods offer a remedy to this challenge, as they enable automatic specification searches. A few studies have adopted kernel-based regression (Thakali, 2016), neural networks (Chang, 2005; Huang et al., 2016; Xie et al., 2007; Zeng et al., 2016a,b), support vector machine (Dong et al., 2015; Li et al., 2008), and deep learning architectures (Cai et al., 2019; Dong et al., 2018) for crash count modeling.

Whereas these ML methods are shown to surpass the traditional count models in terms of predictive accuracy, they succumb to four limitations. First, except Dong et al. (2015), none of the existing ML studies account for spatial correlations between observations. It is important to note that a non-linear link function specification does not inherently account for spatial correlations, and ignoring these correlations can deteriorate the robustness of predictions (Dong et al., 2015). Second, unlike traditional count models, ML methods do not provide a quantification of estimation uncertainty. Third, ML methods are fully non-parametric, with no easy ways to integrate interpretable components in link functions. In other words, if a user is interested in estimation and inference on the relationship between selected explanatory variables and crash counts, there is no straightforward way to specify a semi-parametric link function in the above-mentioned ML-based count models. Fourth, ML-based crash count studies benchmark the performance of their methods against simplistic parametric models which do not account for unobserved and spatial heterogeneity, which is not a fair comparison. More specifically, none of the previous studies address an important question – whether a count model with a non-parametric link function can outperform a model with a linear link function that also accounts for unobserved heterogeneity.

We emphasise that the before-mentioned ML methods adopt classical inference approaches. However, the Bayesian approach facilitates accounting for various sources of uncertainty. For instance, posterior draws of site rankings at each iteration of the Gibbs sampler can directly provide ranking estimates with credible intervals (Miaou and Song, 2005). Hence, the fully Bayesian approach has emerged as the workhorse method in the site ranking literature.

Along the same lines, we propose a Bayesian negative binomial regression model, which not only addresses the first three limitations of the above-discussed ML methods by retaining all advantages of the statistical crash count models (interpretability, inference, accounting for random parameters and spatial correlations) but also allows for an additional non-parametric component in the link function with an endogenous selection of the specification. This non-parametric part is specified as the sum-of-trees (Bayesian Additive Regression Trees (BART), Chipman et al., 2010). The sum-of-trees specification inherently partitions the support of each explanatory variable during the estimation, resulting in a sum of step functions of individual predictors. Furthermore, if a tree depends only on one predictor, then this specification captures the main effect of the predictors; it represents an interaction effect, if a tree depends on more than one predictor. This process is equivalent to creating categorical variables from continuous variables and inherently accounting for interaction effects between predictors, but cut-offs and functional forms of interaction effects are endogenously selected during estimation based on predictive accuracy. This is particularly relevant in the context of the site ranking where continuous predictors like speed limit and shoulder width are often converted into categorical variables by manually selecting cut-offs before entering into linear link functions because such explanatory variables are unlikely to have a constant marginal effect over the entire support. Moreover, finding the optimal functional form of interaction effects of predictors is infeasible in practice.

In the Gibbs sampler for the proposed model, we adopt the Pólya-Gamma augmentation method to deal with the non-conjugacy of the negative binomial likelihood (Polson et al., 2013). The key idea of this data augmentation is to translate the negative binomial likelihood into a Gaussian likelihood by introducing auxiliary Pólya-Gamma-distributed random variables into the model. Finally, we address

the last limitation of the recent ML studies by providing a fair comparison of the proposed model with its linear-link counterpart, while also incorporating random parameters and spatial random effects.

2 Model formulation and estimation

2.1 Model formulation

Let y_i denote a crash count for road segment $i = \{1, 2, \ldots, N\}$. We consider the distribution of y_i to be negative binomial, with the probability parameter p_i and the dispersion parameter r. We specify the link function $\left(\psi_i = \log \frac{p_i}{1-p_i}\right)$ in terms of road-segment-specific attributes $\boldsymbol{F}_i$ and $\boldsymbol{X}_i$. Whereas $\boldsymbol{F}_i$ enters in the link function in a linear interpretable form, the effect of $\boldsymbol{X}_i$ is specified using sum of trees $\sum_{j=1}^{m} g_i(\boldsymbol{X}_i; \boldsymbol{T}_j, \boldsymbol{M}_j)$ (Chipman et al., 2010). Here, $\boldsymbol{T}_j$ is a binary regression tree and $\boldsymbol{M}_j$ denotes the parameters associated with its terminal nodes. To account for spatial correlations between road segments, we include spatial random effect ϕ_i. $\boldsymbol{\phi}$ follows a matrix exponential spatial specification (MESS) of dependence that ensures exponential decay of influence over space (LeSage and Pace, 2007). Here, $\boldsymbol{W}$ is an $N \times N$ non-negative spatial weight matrix and τ captures the magnitude of spatial dependence. MESS is not only comparable to other spatial autoregressive models in terms of prediction and estimation (Strauss et al., 2017) performance, but also has analytical and computational advantages (LeSage and Pace, 2007). The proposed model is summarized below:

$$y_i \sim \text{NB}(r, p_i), \qquad i = 1, \ldots, N \tag{1}$$

$$p_i = \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}, \qquad i = 1, \ldots, N \tag{2}$$

$$\psi_i = \boldsymbol{F}_i^{\top} \boldsymbol{\gamma} + G_i(\boldsymbol{X}_i; \boldsymbol{T}, \boldsymbol{M}) + \phi_i, \qquad i = 1, \ldots, N \tag{3}$$

$$G_i(\boldsymbol{X}_i; \boldsymbol{T}, \boldsymbol{M}) = \sum_{j=1}^{m} g_i(\boldsymbol{X}_i; \boldsymbol{T}_j, \boldsymbol{M}_j), \qquad i = 1, \ldots, N \tag{4}$$

$$\boldsymbol{S}\boldsymbol{\phi} = \exp(\tau \boldsymbol{W})\boldsymbol{\phi} = \boldsymbol{\epsilon} \tag{5}$$

$$\boldsymbol{\epsilon} \sim \text{Normal}(0, \sigma^2 \boldsymbol{I}_N) \tag{6}$$

Equation 3 can be rewritten in the vector form as follows:

$$\begin{gathered} \boldsymbol{\psi} = \boldsymbol{F}\boldsymbol{\gamma} + \boldsymbol{G}(\boldsymbol{X}; \boldsymbol{T}, \boldsymbol{M}) + \boldsymbol{\phi} \\ P(\boldsymbol{\psi} | \boldsymbol{\gamma}, \boldsymbol{T}, \boldsymbol{M}, \sigma^2, \tau) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{[\boldsymbol{\psi} - \boldsymbol{F}\boldsymbol{\gamma} - \boldsymbol{G}]^{\top} \tilde{\boldsymbol{\Omega}} [\boldsymbol{\psi} - \boldsymbol{F}\boldsymbol{\gamma} - \boldsymbol{G}]}{2} \right), \text{ where } \tilde{\boldsymbol{\Omega}} = \frac{\boldsymbol{S}^{\top}\boldsymbol{S}}{\sigma^2} \end{gathered} \tag{7}$$

2.2 Pólya-Gamma data augmentation

Traditional Bayesian inference in models with negative binomial likelihood does not provide conjugate posterior updates. To handle this challenge, we adopt Polson et al. (2013)'s data augmentation technique which translates the negative binomial likelihood into the Gaussian likelihood, after

conditioning on the auxiliary Pólya-Gamma random variables ω_i.

$$\begin{aligned} \omega_i &\sim \text{PG}(y_i + r, 0) \\ P(y_i|\psi_i, r, \omega_i) &\propto \exp\left(-\frac{\omega_i}{2}\left[\psi_i - \frac{y_i - r}{2\omega_i}\right]^2\right) \\ P(\boldsymbol{y}|\boldsymbol{\psi}, r, \boldsymbol{\omega}) &\propto \exp\left(-\frac{1}{2}[\boldsymbol{\psi} - \boldsymbol{Z}]^\top \boldsymbol{\Omega}[\boldsymbol{\psi} - \boldsymbol{Z}]\right) \end{aligned} \tag{8}$$

where

$$\boldsymbol{Z} = \begin{bmatrix} \frac{y_1 - r}{2\omega_1} \\ \vdots \\ \frac{y_N - r}{2\omega_N} \end{bmatrix}_{N\times 1} \quad \boldsymbol{\Omega} = \begin{bmatrix} \omega_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \omega_N \end{bmatrix}_{N\times N}$$

$$\boldsymbol{Z} = \boldsymbol{\psi} + \boldsymbol{\varrho} = \boldsymbol{F\gamma} + \boldsymbol{G}(\boldsymbol{X}; \boldsymbol{T}, \boldsymbol{M}) + \boldsymbol{\phi} + \boldsymbol{\varrho}, \quad \boldsymbol{\varrho} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}) \tag{9}$$

2.3 Prior specification

We adopt the strategy used by Chipman et al. (2010) to specify priors on $\boldsymbol{T}_j$ and $\boldsymbol{M}_j|\boldsymbol{T}_j$. Other prior distributions are summarized below:

$$\boldsymbol{\gamma} \sim \text{Normal}(\boldsymbol{\zeta}_\gamma, \boldsymbol{\Delta}_\gamma) \tag{10}$$

$$\tau \sim \text{Normal}(\zeta_\tau, \sigma_\tau^2) \tag{11}$$

$$\sigma^{-2} \sim \text{Gamma}(b_{\sigma^2}, c_{\sigma^2}) \tag{12}$$

$$r \sim \text{Gamma}(r_0, h) \tag{13}$$

$$h \sim \text{Gamma}(b_0, c_0) \tag{14}$$

Here $\{\boldsymbol{\zeta}_\gamma, \boldsymbol{\Delta}_\gamma, \zeta_\tau, \sigma_\tau^2, b_{\sigma^2}, c_{\sigma^2}, r_0, b_0, c_0\} \cup \{\text{Tree Hyper-parameters}\}$ is a set of hyper-parameters and $\boldsymbol{\Theta} = \{\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{T}, \boldsymbol{M}, \sigma^2, \boldsymbol{\omega}, r, h, \tau\}$ is a set of latent variables of the models. Thus, the joint distribution of observed and latent variables is:

$$\begin{aligned} P(\boldsymbol{y}, \boldsymbol{\Theta}) = P(\boldsymbol{y}|r, \boldsymbol{\omega}, \boldsymbol{\gamma}, \boldsymbol{T}, \boldsymbol{M}, \boldsymbol{\phi}) P(\boldsymbol{\phi}|\sigma^2, \tau) P(r|r_0, h) P(h|b_0, c_0) P(\sigma^{-2}|b_{\sigma^2}, c_{\sigma^2}) P(\tau|\zeta_\tau, \sigma_\tau^2) \dots \\ \dots P(\boldsymbol{\gamma}|\boldsymbol{\zeta}_\gamma, \boldsymbol{\Delta}_\gamma) P(\boldsymbol{T}, \boldsymbol{M}|\{\text{Tree Hyper-parameters}) \left(\prod_{i=1}^{N} P(\omega_i|r)\right) \end{aligned} \tag{15}$$

2.4 Posterior updates

To construct the posterior distributions of latent variables, we generate a Markov chain by taking samples from their conditional distributions. One iteration of the Gibbs sampler is described below.

- Update $\boldsymbol{\phi}$ by sampling $\boldsymbol{\phi} \sim \text{Normal}\left((\boldsymbol{\Omega} + \tilde{\boldsymbol{\Omega}})^{-1}\boldsymbol{\Omega}(\boldsymbol{Z} - \boldsymbol{G} - \boldsymbol{F\gamma}), (\boldsymbol{\Omega} + \tilde{\boldsymbol{\Omega}})^{-1}\right)$.
- Update $\boldsymbol{\gamma}$ by sampling from
  $\boldsymbol{\gamma} \sim \text{Normal}\left((\boldsymbol{\Delta}_\gamma^{-1} + \boldsymbol{F}^\top\boldsymbol{\Omega F})^{-1}(\boldsymbol{F}^\top\boldsymbol{\Omega}(\boldsymbol{Z} - \boldsymbol{G} - \boldsymbol{\phi}) + \boldsymbol{\Delta}_\gamma^{-1}\boldsymbol{\zeta}_\gamma), (\boldsymbol{\Delta}_\gamma^{-1} + \boldsymbol{F}^\top\boldsymbol{\Omega F})^{-1}\right)$
- Update σ^{-2} by sampling $\sigma^{-2} \sim \text{Gamma}\left(b_{\sigma^2} + \frac{N}{2}, c_{\sigma^2} + \frac{\boldsymbol{\phi}^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{\phi}}{2}\right)$.
- Update ω_i by sampling from $\omega_i \sim \text{PG}(y_i + r, \psi_i)$.

- Update r by sampling from a Gamma distribution (see section 4.1.1 of Zhou et al. (2012)).
- Update h by sampling from $h \sim \text{Gamma}(r_0 + b_0, r + c_0)$.
- Update τ using slice sampling where $P(\tau|\cdot) \propto \exp\left(-\frac{(\tau-\zeta_\tau)^2}{2\sigma_\tau^2}\right)\exp\left(-\frac{\boldsymbol{\phi}^\top\tilde{\boldsymbol{\Omega}}\boldsymbol{\phi}}{2}\right)$.
- Update $\boldsymbol{T}$ and $\boldsymbol{M}$ as illustrated for the heteroskedastic BART by Bleich and Kapelner (2014), where the dependent variable is $\boldsymbol{Z} - \boldsymbol{\phi} - \boldsymbol{F}\boldsymbol{\gamma}$ and the error covariance matrix is $\boldsymbol{\Omega}^{-1}$.
- Compute $Z_i = \frac{y_i - r}{2\omega_i}, \quad i = 1, \ldots, N$.

3 Empirical analysis

3.1 Data

We empirically validate the site-ranking and estimation performance of the proposed model framework using real crash frequency data. The considered data set consists of 1,158 adjacent pavement segments of eleven highways in a metropolitan area in the south of the USA.[1] A spatial weight matrix is constructed based on the neighbourhood structure of the road segments. Table 1 enumerates summary statistics of the considered data set.

Variable	**Mean**	**Std.**	**Min.**	**Max.**
Crash count	16.5	22.5	0.0	214.0
Interstate highway (dummy)	0.5	0.5	0.0	1.0
Asphalt pavement (dummy)	0.2	0.4	0.0	1.0
Rural area (dummy)	0.3	0.4	0.0	1.0
Asphalt shoulder (dummy)	0.6	0.5	0.0	1.0
Road condition score > 90	0.5	0.5	0.0	1.0
Truck traffic percentage	10.6	6.5	2.6	34.3
International Roughness Index (IRI)	115.5	35.2	35.2	319.0
Logarithm of Annual Avg. Daily Traffic (AADT)	9.5	0.6	7.7	10.8
Speed limit in MPH	61.3	4.9	55.0	70.0
Left shoulder width < 10 ft	0.5	0.5	0.0	1.0
Right shoulder width < 10 ft	0.4	0.5	0.0	1.0

Table 1: **Description of data set**

3.2 Evaluation of site ranking performance

Numerous criteria for the ranking of hazardous sites have been proposed (see Buddhavarapu, 2015, for a review). In the current application, we rank sites based on the posterior probability that a site belongs to the top 5% most hazardous sites (see Schmidt, 2012). The proposed negative binomial Bayesian additive regression trees (NB-BART) models is benchmarked against negative binomial regression models with spatial error terms and linear-in-parameters link function specification. We consider one model with fixed parameters (NB with fixed parameters) as well as a model whose

[1] Due to data privacy issues, the exact location of the data collection cannot be disclosed.

linear-in-parameters link function contains both fixed and random parameters (NB with random parameters). Furthermore, we contrast the site ranking performance of the proposed model against a naive ranking approach based on raw accident counts.

3.3 Preliminary results

In Table 2, we report model fit statistics for the estimated models. It can be seen that NB-BART provides a better fit to the data that NB with fixed parameters. However, NB-BART is outperformed by NB with random parameters. In Figure 1, we visualise the site ranking performance of the estimated models along with the observed accident counts and the naive ranking approach. We observe that the model-based approaches perform consistently on the considered data set.

Model	**LPPD**
NB with fixed parameters	-3810.82
NB with random parameters	-3733.73
NB-BART	-3758.52

Note: NB = negative binomial, LPPD = log pointwise predictive density

Table 2: **Comparison of model fit**

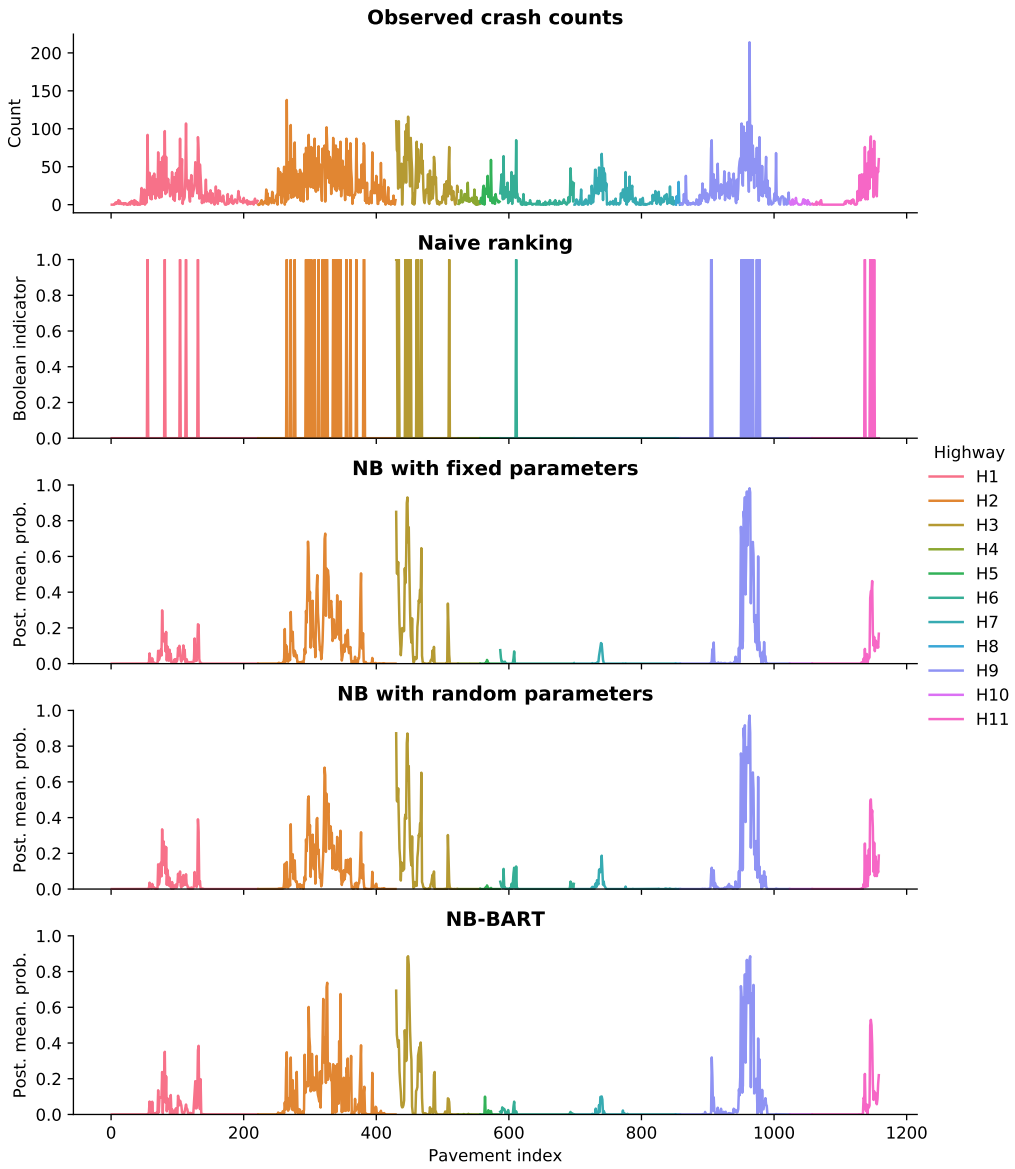

Figure 1: **Comparison of site ranking performance (posterior probability of a site to belong to the top 5% most hazardous sites)**

4 Conclusion

In this paper, we propose a spatial negative binomial Bayesian additive regression trees (NB-BART) model for the identification of accident hot spots in road networks. The sum-of-trees formulation enables a highly flexible specification of the link function. We apply the proposed model to a real data set of crash counts. Our preliminary results indicate that the proposed model framework performs at least as well as established models at identifying accident hot spots using standard settings for the BART priors. Future work will be directed at calibrating the regularisation prior for the leaf parameters and at evaluating the consistency of the site-ranking performance across multiple time periods (see Cheng et al., 2020).

References

Bleich, J. and Kapelner, A. (2014). Bayesian additive regression trees with parametric models of heteroskedasticity. *arXiv preprint arXiv:1402.5397*.

Buddhavarapu, P. N. V. S. R. (2015). *On Bayesian estimation of spatial and dynamic count models using data augmentation techniques: application to road safety management*. PhD thesis.

Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., and Yuan, J. (2019). Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. *Transportation research part A: policy and practice*, 127:71–85.

Chang, L.-Y. (2005). Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety science*, 43(8):541–557.

Cheng, W., Gill, G. S., Zhang, Y., Vo, T., Wen, F., and Li, Y. (2020). Exploring the modeling and site-ranking performance of bayesian spatiotemporal crash frequency models with mixture components. *Accident Analysis & Prevention*, 135:105357.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Deacon, J. A., Zegeer, C. V., and Deen, R. C. (1975). Identification of hazardous rural highway locations. *Transportation Research Record*, (543).

Dong, C., Shao, C., Li, J., and Xiong, Z. (2018). An improved deep learning model for traffic crash prediction. *Journal of Advanced Transportation*, 2018.

Dong, N., Huang, H., Lee, J., Gao, M., and Abdel-Aty, M. (2016). Macroscopic hotspots identification: a bayesian spatio-temporal interaction approach. *Accident Analysis & Prevention*, 92:256–264.

Dong, N., Huang, H., and Zheng, L. (2015). Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. *Accident Analysis & Prevention*, 82:192–198.

Huang, H., Chin, H. C., and Haque, M. M. (2009). Empirical evaluation of alternative approaches in identifying crash hot spots: Naive ranking, empirical bayes, full bayes methods. *Transportation Research Record*, 2103(1):32–41.

Huang, H., Zeng, Q., Pei, X., Wong, S., and Xu, P. (2016). Predicting crash frequency using an optimised radial basis function neural network model. *Transportmetrica A: transport science*, 12(4):330–345.

LeSage, J. P. and Pace, R. K. (2007). A matrix exponential spatial specification. *Journal of Econometrics*, 140(1):190–214.

Li, X., Lord, D., Zhang, Y., and Xie, Y. (2008). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis & Prevention*, 40(4):1611–1618.

Lord, D. and Mannering, F. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation research part A: policy and practice*, 44(5):291–305.

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.

Miaou, S.-P. and Song, J. J. (2005). Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis & Prevention*, 37(4):699–720.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

Schmidt, K. (2012). Modeling crash frequency data.

Strauss, M. E., Mezzetti, M., and Leorato, S. (2017). Is a matrix exponential specification suitable for the modeling of spatial correlation structures? *Spatial statistics*, 20:221–243.

Thakali, L. (2016). *Nonparametric Methods for Road Safety Analysis*. PhD thesis, University of Waterloo.

Xie, Y., Lord, D., and Zhang, Y. (2007). Predicting motor vehicle collisions using bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention*, 39(5):922–933.

Zeng, Q., Huang, H., Pei, X., and Wong, S. (2016a). Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Analytic methods in accident research*, 10:12–25.

Zeng, Q., Huang, H., Pei, X., Wong, S., and Gao, M. (2016b). Rule extraction from an optimized neural network for traffic crash frequency modeling. *Accident Analysis & Prevention*, 97:87–95.

Zhou, M., Li, L., Dunson, D., and Carin, L. (2012). Lognormal and gamma mixed negative binomial regression. In *Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access.

Ziakopoulos, A. and Yannis, G. (2020). A review of spatial approaches in road safety. *Accident Analysis & Prevention*, 135:105323.