A new spatial count data model with Bayesian additive regression trees for accident hot spot identification

Rico Krueger[a,1,*], Prateek Bansal[b,1], Prasad Buddhavarapu[c]

[a] *Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Switzerland*
[b] *Department of Civil and Environmental Engineering, Imperial College London, UK*
[c] *Department of Civil Architectural and Environmental Engineering, The University of Texas at Austin, United States*

ABSTRACT

The identification of accident hot spots is a central task of road safety management. Bayesian count data models have emerged as the workhorse method for producing probabilistic rankings of hazardous sites in road networks. Typically, these methods assume simple linear link function specifications, which, however, limit the predictive power of a model. Furthermore, extensive specification searches are precluded by complex model structures arising from the need to account for unobserved heterogeneity and spatial correlations. Modern machine learning (ML) methods offer ways to automate the specification of the link function. However, these methods do not capture estimation uncertainty, and it is also difficult to incorporate spatial correlations. In light of these gaps in the literature, this paper proposes a new spatial negative binomial model which uses Bayesian additive regression trees to endogenously select the specification of the link function. Posterior inference in the proposed model is made feasible with the help of the Pólya-Gamma data augmentation technique. We test the performance of this new model on a crash count data set from a metropolitan highway network. The empirical results show that the proposed model performs at least as well as a baseline spatial count data model with random parameters in terms of goodness of fit and site ranking ability.

1. Introduction

The identification of accident-prone locations (so-called hot spots) is a core task of road safety management (Cheng et al., 2020; Huang et al., 2009; Lee et al., 2020). Of the various approaches for accident hot spot identification (see Wang et al., 2018; Zahran et al., 2019, for a comparison), crash frequency analysis is the most widely employed method. Crash count data 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). Recent work also uses multivariate analysis for the joint modelling of road fatality and injury counts (Besharati et al., 2020), single and multi-vehicle crashes (Wang and Feng, 2019), and crash frequency by travel modes (Huang et al., 2017).

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 recent crash count modelling literature (Cai et al., 2019a; Cheng et al., 2020; Dong et al., 2016; Heydari 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 and the explained variable 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 hot spot identification and site ranking applications, the specification of the link function should be carefully selected. However, in practice, the space of possible link function specifications is prohibitively large, which precludes 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., 2019b; Dong et al., 2018) for crash

count modelling.

Modern ML methods are shown to surpass the traditional count data models in terms of predictive accuracy but succumb to four limitations. First, with exception of the work by 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 data models, ML methods do not provide a quantification of estimation uncertainty and thus do not offer straightforward ways to construct confidence or credible intervals. Third, ML methods are fully nonparametric, with no easy ways to integrate interpretable components in link functions. In other words, if a user is interested in inferring the relationship between selected explanatory variables and crash counts, there is no straightforward way to specify a semiparametric link function in the above-mentioned ML-based count data 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 data model with a nonparametric 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, which yield point estimates of parameters of interest. On the contrary, the Bayesian approach facilitates accounting for various sources of uncertainty in model formulation and inference. 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, this paper proposes 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 nonparametric component in the link function with an endogenous selection of the specification. This nonparametric component is specified as sum of trees using the Bayesian Additive Regression Trees (BART) framework (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, as it inherently accounts for interaction effects between predictors, while 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.

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 approach 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 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. The proposed approach is a full Bayes (FB) method and the superiority of FB

over empirical Bayes (EB) methods is well established in the literature (Guo et al., 2019). Nonetheless, we still benchmark our approach against EB in site ranking analysis.

The remainder of this paper is organised as follows: In Section 2, we introduce the formulation of the proposed model framework and in Section 3, we outline the estimation approach. In Section 4, we present the empirical analysis and in Section 5, we conclude and discuss avenues of future research.

2. Model formulation

Let y_i denote an observed crash count on road segment $i = \{1, 2, ...,N\}$. We consider the distribution of y_i to be negative binomial (NB) with probability parameter p_i and shape parameter r. The link function $\psi_i = \log \frac{p_i}{1 - p_i}$ depends on road-segment-specific attributes $\boldsymbol{F}_i$ and $\boldsymbol{X}_i$. While $\boldsymbol{F}_i$ enters in the link function in a linear, interpretable form, the effect of $\boldsymbol{X}_i$ is specified using a 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 also include a spatial random effect ϕ_i into the link function. The collection of spatial random effects $\boldsymbol{\phi} = (\phi_1, ...\phi_N)^\top$ follows a matrix exponential spatial specification (MESS) of dependence that ensures exponential decay of influence over space (LeSage and Pace, 2007). Consequently, we have $\boldsymbol{S\phi} = \exp(\tau \boldsymbol{W})\boldsymbol{\phi} = \varepsilon$ with $\varepsilon \sim \text{Normal}(0, \sigma^2 \boldsymbol{I}_N)$, where $\boldsymbol{W}$ is an $N \times N$ non-negative spatial weight matrix, τ captures the magnitude of spatial association, and σ is an error scale.

The proposed model is summarised below:

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

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

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

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

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

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

Equation (3) can be rewritten in vector form. We have

$$\boldsymbol{\psi} = \boldsymbol{F\gamma} + \boldsymbol{G}(\boldsymbol{X}; \boldsymbol{T}, \boldsymbol{M}) + \boldsymbol{\phi} \tag{7}$$

such that

$$P(\boldsymbol{\psi}|\boldsymbol{\gamma}, \boldsymbol{T}, \boldsymbol{M}, \sigma^2, \tau) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{[\boldsymbol{\psi} - \boldsymbol{F\gamma} - \boldsymbol{G}]^\top \tilde{\boldsymbol{\Omega}} [\boldsymbol{\psi} - \boldsymbol{F\gamma} - \boldsymbol{G}]}{2}\right) \tag{8}$$

with $\tilde{\boldsymbol{\Omega}} = \frac{\boldsymbol{S}^\top \boldsymbol{S}}{\sigma^2}$.

2.1. Bayesian additive regression trees (BART)

Chipman et al. (2010) introduce BART as a nonparametric prior over a regression function to capture non-linear relationships and interaction effects between predictors. BART has been widely applied to a wide range of regression and classification problems (see Hill et al., 2020) but the current study presents the first application of BART in spatial count data modelling.

BART specifies the regression function as a sum of trees. Each tree $\boldsymbol{T}_j$ consists a group of decision nodes with splitting rules and a set of terminal leaf nodes. Fig. 1 illustrates one such tree, whose splitting rules at the decision nodes are $x_1 < 0.7$ and $x_2 < 0.4$, and whose leaf nodes are $\boldsymbol{M}_j = \{\mu_{j1}, \mu_{j2}, \mu_{j3}\}$. A tree and its associated decision rules create partitions of the predictor space such that each partition corresponds to
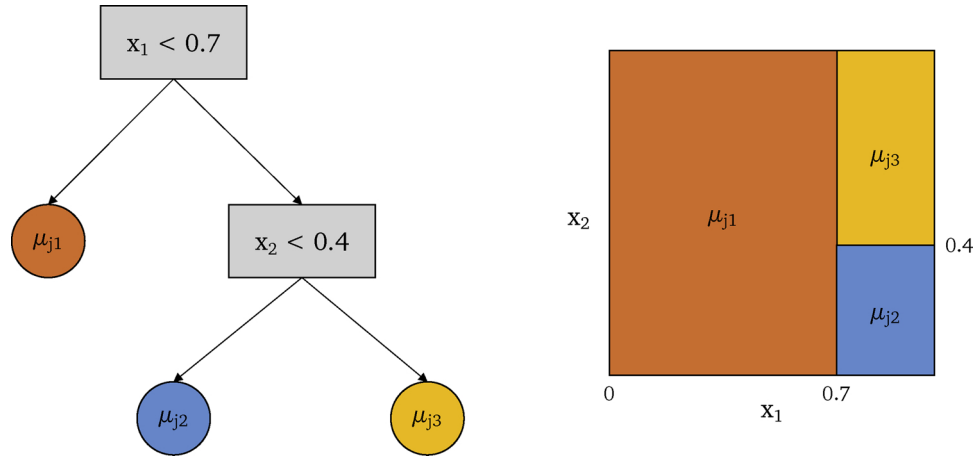
Fig. 1. Illustration of a single binary tree T_j and its corresponding partition of the predictor space . Internal tree nodes are marked with their splitting rules; leaf nodes are marked with their leaf parameters.

a leaf node. In the illustrative example, three partitions are induced: $\mathcal{P}_{j1} = \{x_1 < 0.7\}$, $\mathcal{P}_{j2} = \{x_1 > 0.7, x_2 < 0.4\}$, $\mathcal{P}_{j3} = \{x_1 > 0.7, x_2 > 0.4\}$. The illustrative tree inherently accounts for the interaction between x_1 and x_2. However, if a tree depends only on one predictor, the main effect of that predictor is captured. The function for the illustrative tree is:

$$g(X; T_j, M_j) = \mu_{jt} \quad \text{if} \quad X = \{x_1, x_2\} \in \mathcal{P}_{jt} \quad \forall\, t \in \{1, 2, 3\}. \tag{9}$$

$g(X; T_j, M_j)$ is a step function of a subset of predictors and thus, BART is a sum of step functions. The key idea of BART is to regularise the fit by keeping the effect of each individual tree small such that each of tree can explain a different portion of the regression function. In other words, each tree is constrained to be a weak learner and BART is an ensemble of weak learners. The sum of trees model is a flexible specification, which results in an excellent predictive accuracy (Chipman et al., 2010). Furthermore, splitting rules and leaf nodes are estimable parameters in BART, which in turn facilitates automated specification searches through endogenous partitioning of the predictor space.

2.2. Spatial error dependence

In this study, we adopt the matrix exponential spatial specification (MESS; LeSage and Pace, 2007) to model the unobserved dependence among spatial units. MESS is an appealing specification because it offers computational advantages by simplifying the log-likelihood computations (LeSage and Pace, 2009). MESS also has close correspondence with spatial autoregressive (SAR) and conditional autoregressive (CAR) specifications (Strauss et al., 2017). Whereas SAR and CAR assume geometric decay of spatial correlation, MESS relies on exponential decay (LeSage and Pace, 2007). Readers can refer to Debarsy et al. (2015) for a detailed derivation of the large sample properties of MESS.

3. Estimation

3.1. Pólya-Gamma data augmentation

Irrespective of the link function specification, conditional posterior distributions of the parameters of the negative binomial model do not belong to a known family of distributions (Buddhavarapu et al., 2016; Park and Lord, 2009). In such situations, data augmentation techniques, which involve intermediate or additional latent random variables, are used to derive tractable conditional posteriors of model parameters (Van Dyk and Meng, 2001). BART was originally developed for Gaussian models and therefore, its application in non-Gaussian models can be operationalised by transforming the original model into a Gaussian model via data augmentation. For example, Chipman et al. (2010)

extend BART for a binary probit model with a flexible link function and estimated it by adopting the data augmentation technique suggested by Albert and Chib (1993). Similarly, Kindo et al. (2016) use BART to specify indirect utilities in a multinomial probit model and estimated it by augmenting indirect utilities as latent Gaussian random variables.

In this study, we adopt the Pólya-Gamma data augmentation technique which transforms the negative binomial model into a heteroskedastic Gaussian model (Polson et al., 2013). This augmentation is not only appropriate for any negative binomial specification, as shown by Buddhavarapu et al. (2016), but also enables seamless integration of the Gibbs sampler of the heteroskedastic BART (Bleich and Kapelner, 2014) into that of the proposed model.

Pólya-Gamma data augmentation introduces auxiliary Pólya-Gamma random variates ω_i into the model. Conditional on ω_i, the negative binomial likelihood is translated into a heteroskedastic Gaussian likelihood, while closed-form posteriors for the augmentation variables are retained (see Polson et al., 2013, for a detailed proof).

$$\omega_i \sim \text{PG}(y_i + r, 0) \tag{10}$$

$$P(y_i | \psi_i, r, \omega_i) \propto \exp\left(-\frac{\omega_i}{2}\left[\psi_i - \frac{y_i - r}{2\omega_i}\right]^2\right) \tag{11}$$

$$P(y | \psi, r, \omega) \propto \exp\left(-\frac{1}{2}[\psi - Z]^\top \Omega [\psi - Z]\right) \tag{12}$$

where

$$Z = \begin{bmatrix} \dfrac{y_1 - r}{2\omega_1} \\ \vdots \\ \dfrac{y_N - r}{2\omega_N} \end{bmatrix}_{N \times 1} \quad \Omega = \begin{bmatrix} \omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_N \end{bmatrix}_{N \times N} \tag{13}$$

$$Z = \psi + \varrho = F\gamma + G(X; T, M) + \phi + \varrho, \quad \varrho \sim \text{Normal}(0, \Omega^{-1}) \tag{14}$$

3.2. Prior specification

We adopt the strategy used by Chipman et al. (2010) to specify priors on T_j and $M_j | T_j$. Other prior distributions are summarised below:

$$\gamma \sim \text{Normal}(\zeta_\gamma, \Delta_\gamma) \tag{15}$$

$$\tau \sim \text{Normal}(\zeta_\tau, \sigma_\tau^2) \tag{16}$$

$$\sigma^{-2} \sim \text{Gamma}(b_{\sigma^2}, c_{\sigma^2}) \tag{17}$$

$$r \sim \text{Gamma}(r_0, h) \tag{18}$$

Table 1
Sample description ($N = 1158$).

	Predictor space			2007		2008		2009		2010	
	I	I (random)	II	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Crash count	N.A.	N.A.	N.A.	19.15	25.76	15.17	20.60	14.17	19.21	17.44	23.68
Interstate highway (dummy)	✓	×	✓	0.45	0.50	0.45	0.50	0.45	0.50	0.45	0.50
Exurban area (dummy)	✓	×	✓	0.27	0.45	0.27	0.45	0.27	0.45	0.20	0.40
Asphalt pavement (dummy)	✓	×	✓	0.17	0.38	0.17	0.37	0.14	0.35	0.12	0.33
Asphalt shoulder (dummy)	✓	×	✓	0.60	0.49	0.58	0.49	0.58	0.49	0.57	0.50
Total road width	×	×	✓	54.51	15.17	55.00	15.27	55.97	15.62	56.27	15.43
Left shoulder width [ft]	×	×	✓	8.61	2.78	8.66	2.78	8.36	3.33	8.52	3.23
Right shoulder width [ft]	×	×	✓	9.02	2.31	9.06	2.30	9.75	2.07	9.60	2.26
Left shoulder width < 10 ft	✓	✓	×	0.53	0.50	0.53	0.50	0.52	0.50	0.51	0.50
Right shoulder width < 10 ft	✓	✓	×	0.44	0.50	0.44	0.50	0.25	0.43	0.27	0.45
Road overall quality index	×	×	✓	35.40	20.11	36.66	18.08	36.03	19.32	37.89	17.92
Road overall quality index ≤ 45	✓	×	×	0.50	0.50	0.50	0.50	0.51	0.50	0.56	0.50
Road comfort index	×	×	✓	34.48	5.70	34.95	5.57	34.57	5.78	35.13	5.57
Road structural index	×	×	✓	41.80	14.95	42.52	13.71	42.69	13.97	43.56	12.76
Road surface index	×	×	✓	0.61	1.27	0.62	1.28	1.82	1.47	1.06	1.54
Speed limit [MPH]	✓	✓	✓	61.17	5.05	61.25	4.92	61.35	4.85	61.37	4.79
No. of through lanes	×	×	✓	3.13	0.99	3.16	1.01	3.15	1.03	3.18	1.01
Road profile score (avg.)	✓	×	×	117.45	35.76	114.40	34.47	116.89	36.10	113.11	34.19
Road profile score (left)	×	×	✓	117.15	35.11	107.67	34.38	114.86	34.51	112.04	32.91
Road profile score (right)	×	×	✓	117.88	37.55	121.32	37.50	119.06	38.67	114.32	36.44
Annual average daily traffic (AADT) [10k veh.] per lane	×	×	✓	1.52	0.84	1.54	0.85	1.56	0.86	1.50	0.85
Logarithm of AADT per lane	✓	×	×	9.46	0.61	9.47	0.61	9.48	0.61	9.45	0.61
Truck traffic percentage	✓	×	✓	10.49	6.73	10.36	6.47	10.69	6.57	10.79	6.36

$$h \sim \mathrm{Gamma}(b_0, c_0) \tag{19}$$

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

$$P(y, \Theta) = P(y|r, \omega, \gamma, T, M, \phi)P(\phi|\sigma^2, \tau)P(r|r_0, h)P(h|b_0, c_0)P(\sigma^{-2}|b_{\sigma^2}, c_{\sigma^2})P(\tau|\zeta_\tau, \sigma_\tau^2)...$$

$$... P(\gamma|\zeta_\gamma, \Delta_\gamma)P(T, M|\{\mathrm{TreeHyper - parameters}\})\left(\prod_{i=1}^{N} P(\omega_i|r)\right)$$

$$\tag{20}$$

3.3. Posterior updates

To infer the posterior distributions of the model parameters of interest, we construct Markov chains by generating samples from the conditional distributions of individual coordinates of the parameters space. One iteration of the Gibbs sampler is described below:

- Update ϕ by sampling from $\phi \sim \mathrm{Normal}((\Omega + \tilde{\Omega})^{-1}\Omega(Z - G - F\gamma), (\Omega + \tilde{\Omega})^{-1})$.
- Update γ by sampling from $\gamma \sim \mathrm{Normal}((\Delta_\gamma^{-1} + F^\top \Omega F)^{-1}(F^\top \Omega(Z - G - \phi) + \Delta_\gamma^{-1}\zeta_\gamma), (\Delta_\gamma^{-1} + F^\top \Omega F)^{-1})$
- Update σ^{-2} by sampling from $\sigma^{-2} \sim \mathrm{Gamma}\left(b_{\sigma^2} + \frac{N}{2}, c_{\sigma^2} + \frac{\phi^\top S^\top S\phi}{2}\right)$.
- Update ω_i by sampling from $\omega_i \sim \mathrm{PG}(y_i + r, \psi_i)$.
- Update r by sampling from a Gamma distribution (see Zhou et al., 2012, Section 4.1.1).
- Update h by sampling from $h \sim \mathrm{Gamma}(r_0 + b_0, r + c_0)$.
- Update τ using the Metropolis-Hastings algorithm where $P(\tau|\cdot) \propto \exp\left(-\frac{(\tau - \zeta_\tau)^2}{2\sigma_\tau^2}\right)\exp\left(-\frac{\phi^\top \tilde{\Omega}\phi}{2}\right)$.
- Update T and M as illustrated for the heteroskedastic BART by Bleich and Kapelner (2014), where the dependent variable is $Z - \phi - F\gamma$ and the error covariance matrix is Ω^{-1}. The updates of BART parameters are based on an iterative Bayesian backfitting algorithm (Hastie et al., 2000), where one tree is updated conditional on all other trees.
- Compute $Z_i = \frac{y_i - r}{2\omega_i}, \quad i = 1, ..., N.$

In this study, we present the estimation procedure for a specification with only non-random parameters in the interpretable part of the link function. However, a modeller might expect unobserved heterogeneity in some elements of γ. For completeness, we note that the proposed estimation procedure can be extended for such specifications of unobserved heterogeneity to facilitate flexibility at no compromise in interpretability. Thanks to Pólya-Gamma data augmentation, the extension is as simple as deriving a Gibbs sampler for a linear mixed-effects model.

4. Empirical analysis

4.1. Data

The proposed model framework is empirically validated using detailed crash frequency data from 1158 contiguous road segments of 11 highway facilities in the Greater Houston metropolitan area in the United States of America. The data were compiled by geographically fusing information retrieved from accident and road databases for an observation period covering four consecutive calendar years in the period from 2007 to 2010. For each road segment, the considered data include the annual crash count aggregated over all accident types and severity levels as well as other segment-specific characteristics, namely the type of highway facility the segment in question is associated with, the traffic volume, the truck traffic percentage, the road condition and roadway design attributes. The identified Table 1 enumerates summary statistics of the annual crash counts and the segment features for the considered road segments for each year of the observation period. More details about the data collection and processing are presented in Buddhavarapu (2015). In the subsequent analysis, we treat the data from each year as a separate sample and evaluate the empirical performance of the proposed model across the different years of the observation period.

4.2. Methodology

4.2.1. Model specifications

We contrast the performance of four different model specifications. The proposed negative binomial Bayesian additive regression trees (*NB-*

BART-I, henceforth) model is benchmarked against negative binomial regression models with spatial error terms and linear-in-parameters link functions. We consider one model with fixed parameters (*NB-fixed*, henceforth) as well as a model whose linear-in-parameters link function contains both fixed and random parameters (*NB-random*, henceforth). The specifications *NB-BART-I*, *NB-fixed* and *NB-random* use a restricted predictor space, as indicated in column "Predictor space: I" of Table 1. In the restricted predictor space, some continuous predictors are converted into dummy variables, because assuming a constant marginal effect over the entire support of the predictors is not meaningful. For example, left shoulder width is a continuous variable but its support is partitioned at 10 feet. Column "Predictor space: I (random)" in the same table indicates the predictors that are associated with fully-correlated random parameters in the link function. We conducted an extensive specification search, during which we closely monitored model tractability, to determine which predictors to associate with random parameters. We also test another specification of the proposed model (*NB-BART-II*, henceforth), which considers all predictors in their original form, and allow BART to automatically partition the predictor space. Column "Predictor space: II" of Table 1 indicates which predictors are included in specification *NB-BART-II*.

For the definition of the spatial weight matrix W, we exploit the inherent neighbourhood structure of the road segments. First, we create a $N \times N$ contiguity-based proximity matrix C, which is informed by the neighbourhood relationships of the road segments. An element C_{ij} of C is given by

$$C_{ij} = \begin{cases} \dfrac{1}{d_i(j)} & \text{for } d_i(j) \in \{1, ..., k^*\}, \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where $d_i(j)$ takes a value of k if i and j are kth-order neighbours and zero otherwise. $k^* \in \{x \in \mathbb{N} | x \in [1, N]\}$ is a truncation point, which is set to 3 in the current application. The intuition underlying the definition of C is that distant road segments exhibit weaker spatial correlation than proximal ones. Ultimately, we obtain W by row-normalising C, i.e. $W_{ij} = \frac{C_{ij}}{\sum_j C_{ij}}$.

4.2.2. Implementation and estimation practicalities

We implement the MCMC algorithm outlined in Section 3 by writing our own Python code.[2] Our implementation of the heteroskedastic BART updates follows the documentation provided by Bleich and Kapelner (2014) and Kapelner and Bleich (2016). For the generation of the Pólya-Gamma random variates, we rely on an existing Python implementation (Linderman et al., 2015, 2016a,b) of the appropriate sampling methods (Polson et al., 2013; Windle et al., 2014).

For each of the considered model specifications, the MCMC algorithm is executed with four parallel Markov chains and 10,000 draws for each chain, whereby the initial 5000 draws of each chain are discarded for burn-in. After the burn-in period, a thinning factor of 2 is applied. The step size of the Metropolis-Hastings algorithm for the generation of the posterior draws of the spatial association parameter τ is adaptively tuned with a target acceptance rate of 44%, which is the recommended acceptance ratio for a one-dimensional target density (see Roberts et al., 1997). Convergence of the MCMC simulation is monitored with help of the potential scale reduction factor (Gelman et al., 1992).

4.2.3. Assessment of model fit

We evaluate the goodness of fit of the considered methods using the log pointwise predictive density (LPPD; Gelman et al., 2014) and the root mean square error (RMSE):

---

[2] The Python code is publicly available at https://github.com/RicoKrueger/nb_bart.

(22) LPPD is a strictly proper scoring rule (Gneiting and Raftery, 2007), which corresponds to the logarithm of the pointwise likelihood integrated over the posterior distribution of the relevant model parameters. It is given by

$$\text{LPPD} = \sum_{i=1}^{N} \log\left(\int P(y_i | \theta_i) p(\theta_i | \mathbf{y}) d\theta_i \right), \tag{22}$$

where $\theta_i = \{\psi_i, r\}$. A larger value of LPPD indicates superior goodness of fit.

(23) RMSE is informed by the discrepancy between the predicted accident count $\widehat{y}_i$ and the observed accident count y_i, whereby the former is given by the posterior mean of the conditional expectation of the crash count at site i. It is given by

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\widehat{y}_i - y_i)^2}. \tag{23}$$

4.2.4. Assessment of site ranking ability

Numerous techniques for the model-based identification of accident hot spots are presented in the literature (see e.g. Aguero-Valverde and Jovanis, 2009; Geedipally and Lord, 2010; Miaou and Song, 2005; Miranda-Moreno et al., 2007; Washington et al., 2014). In this study, we use the probabilistic ranking method proposed by Schmidt (2012) to identify hazardous sites. To be specific, we rank sites by their posterior mean probability to belong to the top 5% most hazardous sites in the network. Formally, we proceed as follows: In each MCMC iteration d, we calculate the probability that a site belongs to the top m most hazardous sites, i.e. $P_d(R(\lambda_i) \le m)$, where $R(\lambda_i)$ denotes the rank of site i as a function of the expected accident count λ_i. The posterior mean of the crash statistic $\mathbb{E}(R_i \le m | \mathbf{y}, \mathbf{X})$ is then obtained by averaging the respective posterior draws, and in the current application, we use $m = \alpha \cdot N$ with $\alpha = 5\%$. Schmidt's (2012) probabilistic ranking method robustifies the Bayesian ranking procedures introduced by Miaou and Song (2005) against heterogeneity in the posterior variance of the underlying decision parameter.

Different hot spot identification methods can be compared by assessing the temporal consistency of the produced site rankings (Cheng and Washington, 2008; Montella, 2010). In this study, we employ the site consistency, method consistency and total rank differences tests of Cheng and Washington (2008) to evaluate site ranking consistency We use $H_{\alpha,t}$ to denote the set of sites that are identified as hazardous in time period t at risk level α and define the site ranking consistency tests as follows:

(24) The site consistency test ($\mathcal{T}_{SC}$) measures the ability of a method to consistently identify a site as high risk over two consecutive time periods. It is based on the assumption that an unsafe site should remain hazardous provided that no safety improvements are implemented. $\mathcal{T}_{SC}$ for period t is obtained by averaging the predicted accident counts $\widehat{y}_{h,t+1}$ for period $t + 1$ of all sites $h \in H_{\alpha,t}$:

$$\mathcal{T}_{SC,t} = \frac{1}{|H_{\alpha,t}|} \sum_{h \in H_{\alpha,t}} \widehat{y}_{h,t+1}. \tag{24}$$

A larger value of $\mathcal{T}_{SC,t}$ indicates superior site ranking consistency.

(25) The method consistency test ($\mathcal{T}_{MC}$) measures the ability of a method to consistently identify the same hazardous sites over two consecutive time periods. $\mathcal{T}_{MC}$ for period t is given by the cardinality of the intersection of $H_{\alpha,t}$ and $H_{\alpha,t+1}$, i.e.

$$\mathcal{T}_{MC,t} = |H_{\alpha,t} \cap H_{\alpha,t+1}|. \tag{25}$$

A larger value of $\mathcal{T}_{MC,t}$ indicates superior site ranking consistency.

(26) Finally, the total rank differences test ($\mathcal{T}_{TRD}$) measures the average absolute differences in ranks of hazardous sites over two

Table 2
Goodness of fit.

Method	2007		2008		2009		2010	
	LPPD	RMSE	LPPD	RMSE	LPPD	RMSE	LPPD	RMSE
NB-fixed	−3784.34	14.74	−3545.04	11.66	−3545.89	11.73	−3627.46	12.80
NB-random	−3726.01	13.77	−3483.54	10.85	−3478.76	10.93	−3564.49	12.03
NB-BART-I	−3734.88	13.82	−3490.67	10.69	−3510.13	10.97	−3558.87	11.46
NB-BART-II	**−3664.14**	**12.15**	**−3437.68**	**9.84**	**−3407.94**	**9.39**	**−3510.43**	**10.53**

Note: LPPD = log pointwise predictive density; RMSE = root mean square error. For each observation period and goodness fit measures, the best-performing method is highlighted in bold font.

consecutive time periods. $\mathcal{T}_{\text{TRD}}$ for period t is given by

$$\mathcal{T}_{\text{TRD},t} = \frac{1}{|H_{\alpha,t}|} \sum_{h \in H_{\alpha,t}} |R_{h,t+1} - R_{h,t}|, \tag{26}$$

where $R_{h,t}$ denotes the rank of site h in period t. A smaller value of $\mathcal{T}_{\text{TRD},t}$ indicates superior site ranking consistency.

4.3. Results

4.3.1. Model fit

Table 2 gives the goodness of fit measures of the considered model specifications across the four years of the observation period. It can be seen that NB-BART-II with an unrestricted predictor space outperforms the competing model specifications by a significant margin in all four samples. For example, for year 2010, NB-BART-II yields LPPD and RMSE values of −3510.43 and 10.53, respectively, whereas the next-best model specification NB-BART-I yields LPPD and RMSE values of −3558.87 and 11.46, respectively. Furthermore, we observe that NB-BART-I closely matches the performance of NB-random in the considered samples. As expected, NB-fixed exhibits the poorest performance among the four considered model specifications in all four samples due to its rigid link function specification.

In sum, Table 2 shows that the endogenous partitioning of the predictor space induced by the BART component in NB-BART-II results in a substantial improvement in goodness of fit. We acknowledge that a modeller could exogenously partition the support of the continuous predictors in manifold ways to eventually achieve a comparable or better goodness-of-fit. However, the computation time of such an iterative hit-and-trial specification search would be prohibitive, as the space of possible specifications is enormously large. NB-BART-II automates this search process and achieves the optimal partitioning of the predictor space and thus obviates the need for expensive specification searches. Consequently, NB-BART-II with an unrestricted predictor space should be preferred over NB-BART-I with a restricted predictor space.

4.3.2. Site ranking performance

Next, we compare the site ranking ability of the different model specifications. Our analysis covers all four years of the observation period but to avoid clutter, we restrict the visual display of the site ranking results to year 2010. Subplot (a) of Fig. 2 shows the observed crash count for each pavement segment. Subplot (b) of the same figure displays the results of the naïve ranking approach, which involves assigning a value of 1 to the top 5% sites with the highest accident count and a value of 0 to all other sites. Subplot (c) shows the expected accident counts calculated via the empirical Bayes (EB) approach (Hauer et al., 2002) in conjunction with a negative binomial model and maximum likelihood estimation. Subplots (d) to (g) correspond to the four NB-BART specifications; each of the plots shows the posterior mean probabilities of each road segment to belong to the top 5% most hazardous sites in the network.

Fig. 2 illustrates that the probabilistic ranking approaches can better capture the inherent uncertainty about the safety of a site, as the results differ markedly from the observed accident counts. In contradistinction,

the results of the naïve approach and the EB method closely mimic the observed accident counts and thus provided limited additional information. Furthermore, it can be seen that the posterior mean probability plots corresponding to the specifications NB-fixed, NB-random, NB-BART-I, NB-BART-II are virtually indistinguishable from each other. This insight corroborates our earlier finding that the automated selection of an optimal link function specification in BART obviates the need for random parameters in a link function. Notwithstanding significant differences in goodness-of-fit (see Table 2), Fig. 2 suggests that model specifications NB-BART-I and NB-BART-II compare favourably to each other as well as to NB-fixed and NB-random in terms of their site ranking abilities.

Next, we contrast the site ranking ability of the EB approach as well as of NB-fixed, NB-random, NB-BART-I and NB-BART-II using the site consistency ($\mathcal{T}_{\text{SC}}$), method consistency ($\mathcal{T}_{\text{MC}}$) and total rank differences ($\mathcal{T}_{\text{TRD}}$) tests of Cheng and Washington (2008). Table 3 shows the results of the site ranking consistency tests for three reference period. We observe that NB-BART-II, followed by NB-BART-I, performs best in respect to site consistency for all reference periods. In terms of method consistency and total rank differences, none of the considered methods display superior performance across the three reference periods and the relative difference in the test score of the methods are not substantial. Therefore, we conclude that NB-BART affords similar site ranking consistency as NB-fixed and NB-BART. In addition, we note the EB approach underperforms in terms of site consistency but remains competitive with respect to the other methods in terms of method consistency and total rank differences.

4.3.3. Variable importance and parameter estimates

As a byproduct of the estimation, BART provides a statistic about the variable importance (see Bleich et al., 2014). In the context of BART, the notion of "importance" is different from elasticity. To be precise, variable importance of BART predictors denotes the proportion of times a variable of the predictor space is included in a splitting rule. Fig. 3 shows the inclusion proportions of the predictors used in the model specifications NB-BART-I and NB-BART-II for the whole observation period. For both NB-BART specifications, we observe that the importance of the different variables is stable across the different years of the observation period. This insight shows that BART can consistently identify the variation explained by each predictor. Furthermore, the importance of several predictors (e.g. road quality index and left shoulder width) is higher in NB-BART-II than in NB-BART-I. This is because these variables are included as dummy variables in NB-BART-I, and variables with binary support only admit at most one split. In contrast, these variables are included as continuous predictors in NB-BART-II, and variables with continuous support inherently provide more ways to be incorporated into decision rules. This analysis suggests that the exogenous conversion of a continuous predictor into a dummy variable may reduce its importance. Yet, the endogenous partitioning of the predictor space in NB-BART-II allows to bypass the shortcomings of a manual specification selection.

Furthermore, Fig. 3 shows the estimated posterior distribution of the negative binomial shape parameter r (whose inverse is also referred to as the dispersion parameter) and the spatial association parameter τ
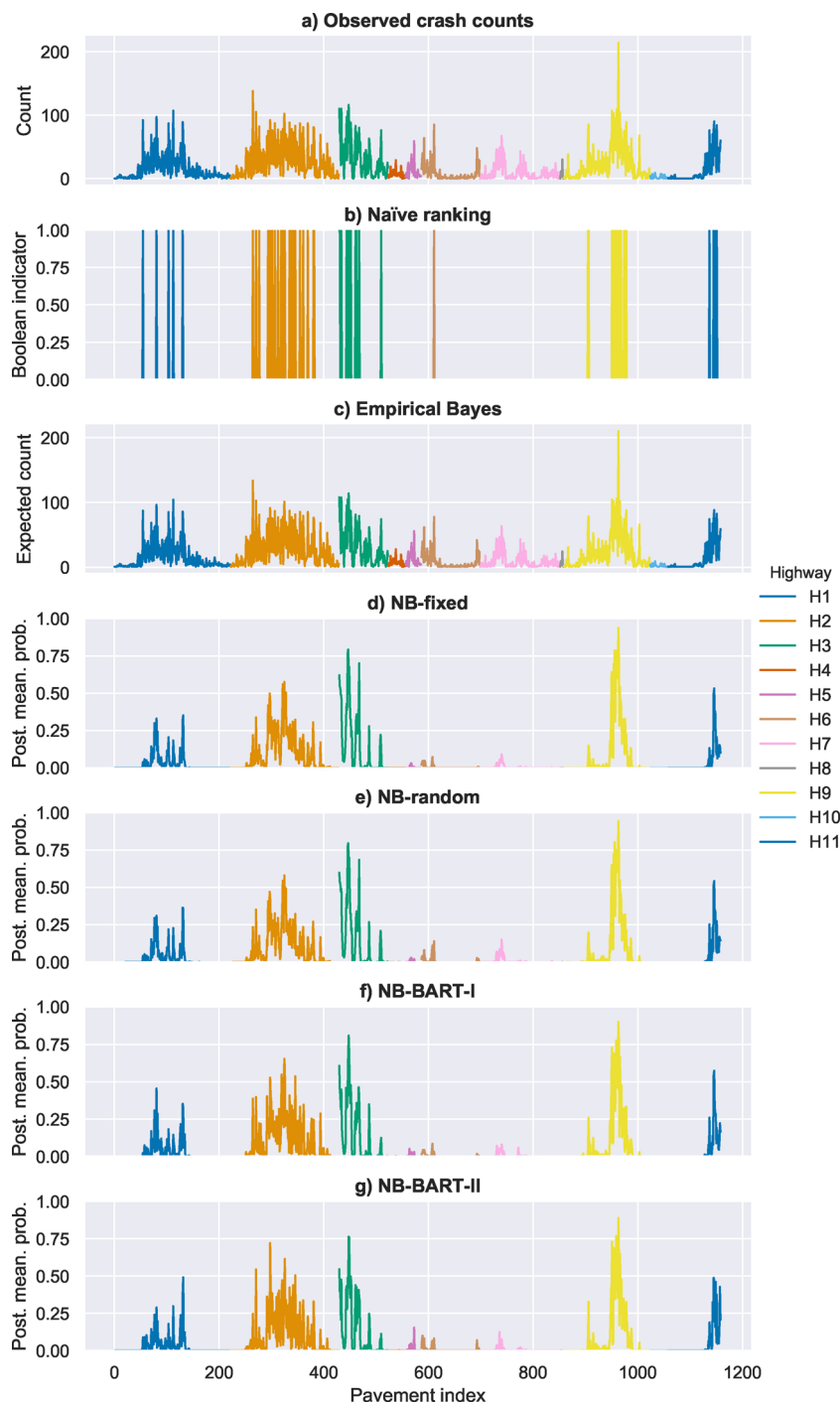
Fig. 2. Site ranking results for year 2010. The naïve ranking approach consists of identifying the top 5% most hazardous sites based on the observed accident counts. For the probabilistic methods NB-fixed, NB-random, NB-BART-I and NB-BART-II, the plot shows the posterior mean probability that a site belongs to the top 5% most hazardous sites. The highways are given arbitrary names for proprietary reasons.

for year 2010 of the observation period for each of the considered model specifications. For both model parameters, it can be seen that the posterior distributions produced by the different model specifications closely overlap. For the spatial association parameter τ, we observe that the posterior distributions produced by NB-BART-I and -II are slightly shifted to the right relative the posterior distributions produced by the other model specifications. A possible explanation for this right shift is that the BART component absorbs some of the spatial dependence that is left unexplained by the other model specifications.

Finally, Table 4 displays the parameters estimates for NB-fixed and NB-random for all four years of the observation period. Along with the

variable inclusion proportions shown in Fig. 3, Table 4 suggests that the importance of the predictors is stable across time periods. Consequently, temporal instability (Mannering, 2018) is not a serious concern in the current application (Fig. 4).

5. Conclusions

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 contribution of our work is threefold:

Table 3
Site ranking consistency.

| | 2007 | | | 2008 | | | 2009 | | |
	$\mathcal{T}_{SC}$	$\mathcal{T}_{MC}$	$\mathcal{T}_{TRD}$	$\mathcal{T}_{SC}$	$\mathcal{T}_{MC}$	$\mathcal{T}_{TRD}$	$\mathcal{T}_{SC}$	$\mathcal{T}_{MC}$	$\mathcal{T}_{TRD}$
EB	28.27	**41**	25.57	28.62	38	**26.34**	30.99	27	48.45
NB-fixed	54.95	39	27.03	46.94	39	30.10	60.87	**31**	49.17
NB-random	56.35	40	**23.60**	47.48	**41**	29.67	61.10	30	48.41
NB-BART-I	57.52	33	29.83	49.80	40	33.72	61.84	30	59.40
NB-BART-II	**61.05**	37	38.19	**53.73**	40	38.47	**69.09**	30	**45.93**

Note: $\mathcal{T}_{SC}$ = site consistency test; $\mathcal{T}_{MC}$ = method consistency test; $\mathcal{T}_{TRD}$ = total rank differences test. For each test and reference period, the best-performing hot spot identification method is highlighted in bold font.

- First, the proposed model re-conceptualises the specification of the link function in count data models. Since NB-BART specifies the link function as a sum-of-step-functions, it can flexibly account for interactions and non-linear relationships between predictors. This flexibility comes at no expense in interpretability, because a modeller can still specify a semiparametric link function in combination with a BART component and random parameters in a linear component. Furthermore, a modeller does not require to exogenously create dummy variables from continuous covariates before entering them into the link function, because BART endogenously partitions the predictor space and automatically extracts the maximum possible information from the predictors.

- Second, we derive a Gibbs sampler for NB-BART by employing the state-of-the-art Pólya-Gamma data augmentation technique. The sampler ensures conjugacy of the conditional posteriors of all non-BART parameters with the exception of a scalar spatial correlation parameter, which can be updated in a Metropolis-Hastings step. The Bayesian inference approach allows for the construction of credible intervals and other derived quantities such as site rankings.

- Third, we benchmark the performance of NB-BART against a baseline negative binomial regression model with a linear link function and spatial correlations. Goodness-of-fit and site ranking measures indicate that NB-BART performs as well as or better than the baseline model with random parameters in the linear-in-parameters link function. Our results suggest that if the objectives of a study are centred around prediction, a modeller may be better off considering a flexible link function with a BART component in lieu of a linear-in-parameters link function with random parameters. Nonetheless, if heterogeneity in some parameters is of interest to the modeller, it can be incorporated in NB-BART with a semiparametric link function. At the same time, we acknowledge that it is possible that a count data model with a sophisticated mixing distribution such as a finite mixture-of-normals or Dirichlet process mixture-of-normals mixing distribution (see e.g. Buddhavarapu et al., 2016; Cheng et al., 2020; Heydari et al., 2016) could effectively control for non-linearities in the predictors and could be more competitive with NB-BART in certain empirical applications. However, such models are time-consuming to estimate and necessitate post-hoc model selection for determining the dimensionality and complexity of the mixing distribution, whereas NB-BART endogenously partitions the predictor space.
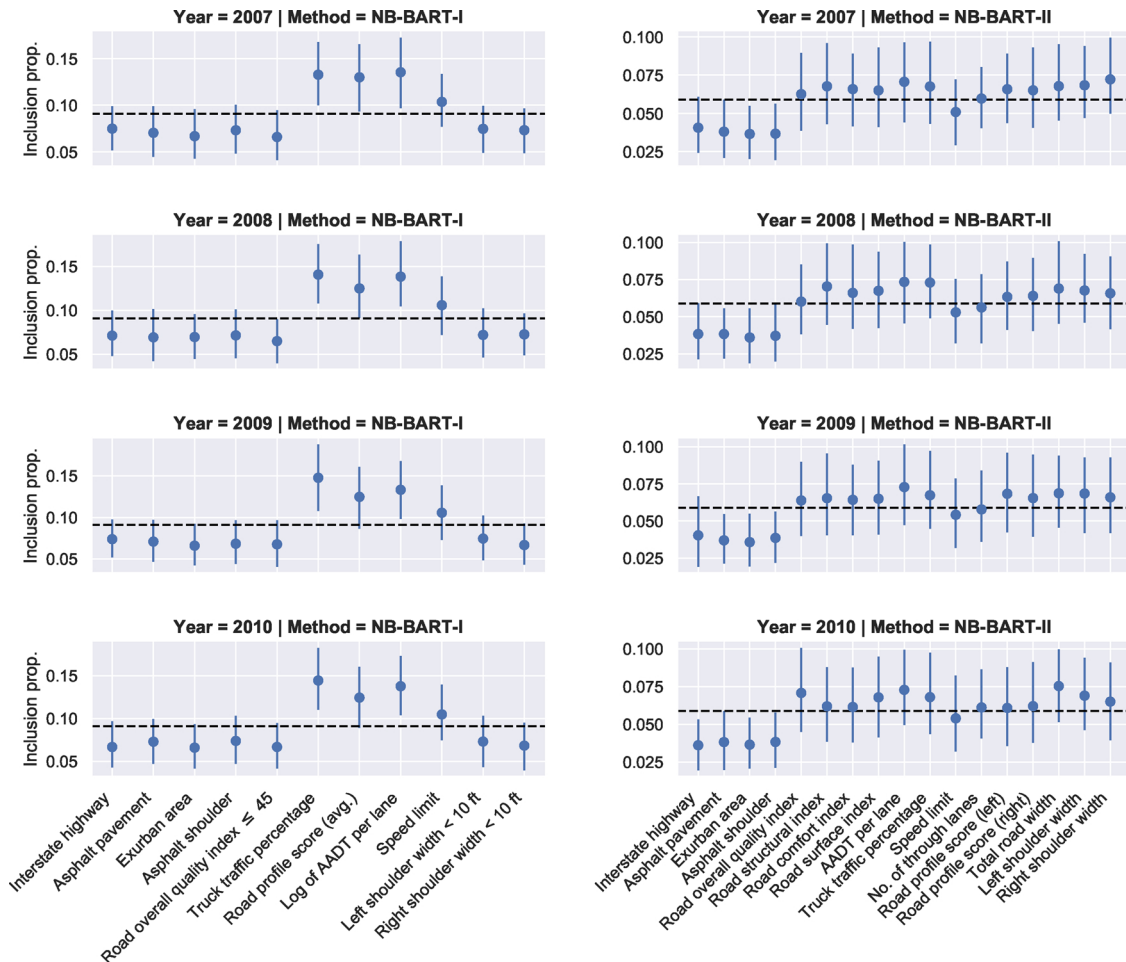

Fig. 3. Variable inclusion proportions. The dots mark the posterior means; the vertical error bars mark the 95% credible intervals; the dashed horizontal lines indicate the equal importance inclusion proportions.

Table 4

Estimation results for the negative binomial model with only fixed link function parameters (NB-fixed) and the negative binomial model with both fixed and random link function parameters (NB-random).

Parameter	NB-fixed				NB-random			
	2007	2008	2009	2010	2007	2008	2009	2010
Intercept	2.00*	1.70*	1.69*	1.72*	1.90*	1.59*	1.58*	1.61*
Interstate highway	0.44*	0.42*	0.35*	0.30	0.44*	0.40*	0.34*	0.29
Asphalt pavement	0.07	0.08	0.12	0.10	0.06	0.09	0.11	0.09
Exurban area	−1.03*	−0.98*	−0.87*	−0.83*	−1.03*	−0.96*	−0.86*	−0.83*
Asphalt shoulder	0.02	0.10	0.04	0.01	0.03	0.11	0.03	0.01
Road overall quality index ≤ 45	0.16	0.07	0.19	0.13	0.15	0.06	0.18	0.13
Truck traffic percentage	−0.16	−0.26*	−0.25*	−0.19*	0.42*	0.40*	0.44*	0.53*
Road profile score (avg.)	0.15*	0.12*	0.11	0.24*	−0.16	−0.26*	−0.25*	−0.18
Log of AADT per lane	0.41*	0.39*	0.43*	0.52*	0.13	0.10	0.10	0.24*
Speed limit								
Loc.	−0.06	−0.05	−0.02	0.01	−0.07	−0.05	−0.01	0.00
Scale					0.11*	0.10*	0.15*	0.15*
Left shoulder width < 10 ft								
Loc.	−0.05	−0.13	−0.16	−0.40*	−0.05	−0.11	−0.20	−0.41*
Scale					0.26*	0.25*	0.26*	0.20*
Right shoulder width < 10 ft								
Loc.	−0.31	−0.26	−0.28*	−0.10	−0.34*	−0.37*	−0.37*	−0.14
Scale					0.20*	0.32*	0.43*	0.29*
Negative binomial shape parameter r	1.47*	1.64*	1.58*	1.91*	1.60*	1.82*	1.79*	2.12*
Spatial association parameter τ	−1.50*	−1.43*	−1.42*	−1.45*	−1.52*	−1.47*	−1.47*	−1.47*

* At least 95% of the posterior mass exclude zero.
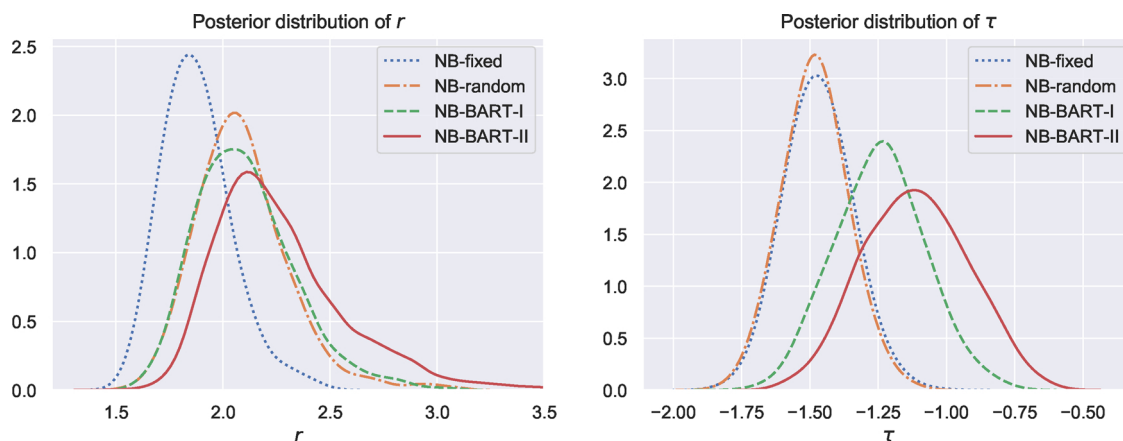

Fig. 4. Posterior distributions of the negative binomial shape parameter r and the spatial association parameter τ for year 2010 by model specification.

The hierarchical nature of NB-BART offers opportunities for extensions in two directions. First, a multivariate extension of NB-BART can facilitate the joint modelling of multiple crash frequency variables such as crash counts by crash type, severity level, and vehicle type (Dong et al., 2014; Yasmin et al., 2018). Second, the proposed NB-BART formulation accounts for only spatial random effects but can be extended to accommodate temporal variation in model parameters as well as spatiotemporal random effects (Li et al., 2019; Liu and Sharma, 2017).

Recent developments in BART can also be incorporated in the above-discussed extensions. First, in the case of sparse and large predictors spaces, the original BART framework is vulnerable to the curse of dimensionality. Recently proposed soft trees can be adopted in order to handle this challenge (Linero and Yang, 2018). Second, Bayesian regression trees have recently been used for observational causal inference (Hahn et al., 2020), which is a critical avenue for future research in the accident analysis literature (Mannering et al., 2020).

Author statement

R Krueger: conception and design, method derivation and implementation, data analysis, manuscript writing and editing.

P Bansal: conception and design, method derivation, data analysis, manuscript writing and editing.

P Buddhavarapu: data acquisition and processing, manuscript editing.

References

Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. Transp. Res. Rec. 2136 (1), 82–91.

Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. 88 (422), 669–679.

Besharati, M.M., Kashani, A.T., Li, Z., Washington, S., Prato, C.G., 2020. A bivariate random effects spatial model of traffic fatalities and injuries across provinces of iran. Accid. Anal. Prevent. 136, 105394.

Bleich, J., Kapelner, A., 2014. Bayesian Additive Regression Trees with Parametric Models of Heteroskedasticity. arXiv preprint arXiv:1402.5397.

Bleich, J., Kapelner, A., George, E.I., Jensen, S.T., 2014. Variable selection for Bart: an application to gene regulation. Ann. Appl. Stat. 1750–1781.

Buddhavarapu, P., Scott, J.G., Prozzi, J.A., 2016. Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. Transp. Res. Part B: Methodol. 91, 492–510.

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., Lee, J., Huang, H., 2019a. Integrating macro-and micro-level safety analyses: a Bayesian approach incorporating spatial interaction. Transp. A: Transp. Sci. 15 (2), 285–306.

Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., Yuan, J., 2019b. Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. Transp. Res. Part A: Policy Pract. 127, 71–85.

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

Cheng, W., Gill, G.S., Zhang, Y., Vo, T., Wen, F., Li, Y., 2020. Exploring the modeling and site-ranking performance of Bayesian spatiotemporal crash frequency models with mixture components. Accid. Anal. Prev. 135, 105357.

Cheng, W., Washington, S., 2008. New criteria for evaluating methods of identifying hot spots. Transp. Res. Rec. 2083 (1), 76–85.

Chipman, H.A., George, E.I., McCulloch, R.E., et al., 2010. Bart: Bayesian additive regression trees. Ann. Appl. Stat. 4 (1), 266–298.

Deacon, J.A., Zegeer, C.V., Deen, R.C., 1975. Identification of hazardous rural highway locations. Transp. Res. Rec.(543).

Debarsy, N., Jin, F., Lee, L.-F., 2015. Large sample properties of the matrix exponential spatial specification with an application to FDI. J. Econom. 188 (1), 1–21.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections. Accid. Anal. Prev. 70, 320–329.

Dong, C., Shao, C., Li, J., Xiong, Z., 2018. An improved deep learning model for traffic crash prediction. J. Adv. Transp. 2018.

Dong, N., Huang, H., Lee, J., Gao, M., Abdel-Aty, M., 2016. Macroscopic hotspots identification: a Bayesian spatio-temporal interaction approach. Accid. Anal. Prev. 92, 256–264.

Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. Accid. Anal. Prev. 82, 192–198.

Geedipally, S.R., Lord, D., 2010. Identifying hot spots by modeling single-vehicle and multivehicle crashes separately. Transp. Res. Rec. 2147 (1), 97–104.

Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for bayesian models. Stat. Comput. 24 (6), 997–1016.

Gelman, A., Rubin, D.B., et al., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7 (4), 457–472.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102 (477), 359–378.

Guo, X., Wu, L., Zou, Y., Fawcett, L., 2019. Comparative analysis of empirical Bayes and Bayesian hierarchical models in hotspot identification. Transp. Res. Rec. 2673 (7), 111–121.

Hahn, P.R., Murray, J.S., Carvalho, C.M., et al., 2020. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. Bayesian Anal.

Hastie, T., Tibshirani, R., et al., 2000. Bayesian backfitting (with comments and a rejoinder by the authors. Stat. Sci. 15 (3), 196–223.

Hauer, E., Harwood, D.W., Council, F.M., Griffith, M.S., 2002. Estimating safety by the empirical Bayes method: a tutorial. Transp. Res. Rec. 1784 (1), 126–131.

Heydari, S., Fu, L., Joseph, L., Miranda-Moreno, L.F., 2016. Bayesian nonparametric modeling in transportation safety studies: applications in univariate and multivariate settings. Anal. Methods Accid. Res. 12, 18–34.

Hill, J., Linero, A., Murray, J., 2020. Bayesian additive regression trees: a review and look forward. Annu. Rev. Stat. Appl. 7.

Huang, H., Chin, H.C., Haque, M.M., 2009. Empirical evaluation of alternative approaches in identifying crash hot spots: Naive ranking, empirical Bayes, full Bayes methods. Transp. Res. Rec. 2103 (1), 32–41.

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

Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. Anal. Methods Accid. Res. 14, 10–21.

Kapelner, A., Bleich, J., 2016. Bartmachine: machine learning with Bayesian additive regression trees. J. Stat. Softw. Articles 70 (4), 1–40.

Kindo, B.P., Wang, H., Pe na, E.A., 2016. Multinomial probit Bayesian additive regression trees. Stat 5 (1), 119–131.

Lee, J., Chung, K., Papakonstantinou, I., Kang, S., Kim, D.-K., 2020. An optimal network screening method of hotspot identification for highway crashes with dynamic site length. Accid. Anal. Prev. 135, 105358.

LeSage, J., Pace, R.K., 2009. Introduction to Spatial Econometrics. Chapman and Hall/ CRC.

LeSage, J.P., Pace, R.K., 2007. A matrix exponential spatial specification. J. Econom. 140 (1), 190–214.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accid. Anal. Prev. 40 (4), 1611–1618.

Li, Z., Chen, X., Ci, Y., Chen, C., Zhang, G., 2019. A hierarchical Bayesian spatiotemporal random parameters approach for alcohol/drug impaired-driving crash frequency analysis. Anal. Methods Accid. Res. 21, 44–61.

Linderman, S., Adams, R.P., Pillow, J.W., 2016a. Bayesian latent structure discovery from multi-neuron recordings. Advances in Neural Information Processing Systems. pp. 2002–2010.

Linderman, S., Johnson, M.J., Adams, R.P., 2015. Dependent multinomial models made easy: stick-breaking with the pólya-gamma augmentation. Advances in Neural Information Processing Systems. pp. 3456–3464.

Linderman, S.W., Miller, A.C., Adams, R.P., Blei, D.M., Paninski, L., Johnson, M.J., 2016b. Recurrent Switching Linear Dynamical Systems. arXiv preprint arXiv:1610.08466.

Linero, A.R., Yang, Y., 2018. Bayesian regression tree ensembles that adapt to smoothness and sparsity. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 80 (5), 1087–1110.

Liu, C., Sharma, A., 2017. Exploring spatio-temporal effects in traffic crash trend analysis. Anal. Methods Accid. Res. 16, 104–116.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A: Policy Pract. 44 (5), 291–305.

Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13.

Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Anal. Methods Accid. Res. 25, 100113.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

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

Miranda-Moreno, L.F., Labbe, A., Fu, L., 2007. Bayesian multiple testing procedures for hotspot identification. Accid. Anal. Prev. 39 (6), 1192–1201.

Montella, A., 2010. A comparative analysis of hotspot identification methods. Accid. Anal. Prev. 42 (2), 571–581.

Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. 41 (4), 683–691.

Polson, N.G., Scott, J.G., Windle, J., 2013. Bayesian inference for logistic models using pólya-gamma latent variables. J. Am. Stat. Assoc. 108 (504), 1339–1349.

Roberts, G.O., Gelman, A., Gilks, W.R., et al., 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. Ann. Appl. Probab. 7 (1), 110–120.

Schmidt, K., 2012. Modeling Crash Frequency Data. Iowa State University PhD thesis.

Strauss, M.E., Mezzetti, M., Leorato, S., 2017. Is a matrix exponential specification suitable for the modeling of spatial correlation structures? Spat. Stat. 20, 221–243.

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

Van Dyk, D.A., Meng, X.-L., 2001. The art of data augmentation. J. Comput. Graph. Stat. 10 (1), 1–50.

Wang, K., Zhao, S., Ivan, J.N., Ahmed, I., Jackson, E., 2018. Evaluation of hot spot identification methods for municipal roads. J. Transp. Saf. Secur. 1–19.

Wang, S., Feng, M., 2019. Freeway single and multi-vehicle crash safety analysis: influencing factors and hotspots. Accid. Anal. Prev. 132, 105268.

Washington, S., Haque, M.M., Oh, J., Lee, D., 2014. Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots. Accid. Anal. Prev. 66, 136–146.

Windle, J., Polson, N.G., Scott, J.G., 2014. Sampling Pólya-Gamma Random Variates: Alternate and Approximate Techniques. arXiv preprint arXiv:1405.0506.

Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis. Accid. Anal. Prev. 39 (5), 922–933.

Yasmin, S., Momtaz, S.U., Nashad, T., Eluru, N., 2018. A multivariate copula-based macro-level crash count model. Transp. Res. Rec. 2672 (30), 64–75.

Zahran, E.-S.M.M., Tan, S.J., Tan, E.H.A., Mohamad'Asri Putra, N.A.B., Yap, Y.H., Abdul Rahman, E.K., 2019. Spatial analysis of road traffic accident hotspots: evaluation and validation of recent approaches using road safety audit. J. Transp. Saf. Secur. 1–30.

Zeng, Q., Huang, H., Pei, X., Wong, S., 2016a. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Anal. Methods Accid. Res. 10, 12–25.

Zeng, Q., Huang, H., Pei, X., Wong, S., Gao, M., 2016b. Rule extraction from an optimized neural network for traffic crash frequency modeling. Accid. Anal. Prev. 97, 87–95.

Zhou, M., Li, L., Dunson, D., Carin, L., 2012. Lognormal and gamma mixed negative binomial regression. Proceedings of the International Conference on Machine Learning, vol. 2012 1343.

Ziakopoulos, A., Yannis, G., 2020. A review of spatial approaches in road safety. Accid. Anal. Prev. 135, 105323.