A New Spatial Count Data Model with Bayesian Additive Regression Trees for Accident Hot Spot Identification

Rico Krueger[1,*], Prateek Bansal[2,*], Prasad Buddhavarapu[3]

[1]Ecole Polytechnique Fédérale de Lausanne, [2]Imperial College London,[3]The University of Texas at Austin
[*] *Equal contribution*

2nd BTR Online Conference

Motivation i

Crash count data models

$\Downarrow$

Model-based rankings of hazardous sites, and predict crash counts at hot spots

Traditional crash count data models:

- Poisson log-normal or negative binomial regression models.
- Account for spatial correlation between spatial units.
- Linear-in-parameter link function.

Motivation ii

- Linear-in-parameters link function is suitable for interpretability, but **predictive performance** is of paramount importance.
- Exhaustive 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 **automate the specification search**:
  - Kernel-based regression (Thakali, 2016).
  - Neural networks (Chang, 2005; Huang et al., 2016).
  - Support vector machine (Dong et al., 2015; Li et al., 2008).
  - Deep learning architectures (Cai et al., 2019; Dong et al., 2018).

Research gaps

Limitations of modern ML methods:

- Do not account for spatial correlations between observations.
- Do not offer straightforward ways to construct confidence or credible intervals.
- Fully nonparametric, with no easy way to include interpretable part in link function.
- Unfair comparison – benchmark performance against simplistic parametric models.

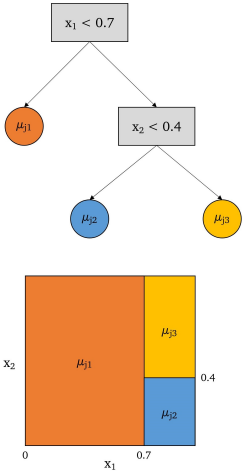
Research objectives

A negative binomial regression model with a **semi-parametric** link function, which:

- Specifies non-parametric part of the link function using Bayesian additive regression trees (BART)
- Induces endogenous partition of predictor space.
- Retains interpretability by allowing linear link component.
- Can account for random parameters and spatial correlations.
- Is embedded in Bayesian inferential framework.

Bayesian additive regression trees (BART)

- BART specifies the regression function as sum of trees.
- An example tree of BART:
  - Splitting rules at decision nodes are $x_1 < 0.7$ and $x_2 < 0.4$.
  - Terminal leaf nodes are $\boldsymbol{M}_j = \{\mu_{j1}, \mu_{j2}, \mu_{j3}\}$.
  - Induces three partitions $\mathcal{P}_{jt}$ and accounts for interaction.
  - $g(\boldsymbol{X}; \boldsymbol{T}_j, \boldsymbol{M}_j) = \mu_{jt}$ if $\boldsymbol{X} = \{x_1, x_2\} \in \mathcal{P}_{jt} \quad \forall t \in \{1, 2, 3\}$.
  - Tree is a step function and BART is sum of step functions.
- BART endogenously partitions the predictor space, accounts for nonlinearities, main and interaction effects



Model formulation

Negative Binomial (NB) likelihood

$$y_i \sim \text{NB}(r, p_i), \quad p_i = \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}, \quad i = 1, \ldots, N$$

Semi-parametric link function specification with sum-of-trees

$$\psi_i = G_i(\boldsymbol{X}_i; \boldsymbol{T}, \boldsymbol{M}) + \boldsymbol{F}_i^\top \gamma + \phi_i, \quad G_i(\boldsymbol{X}_i; \boldsymbol{T}, \boldsymbol{M}) = \sum_{j=1}^{m} g_i(\boldsymbol{X}_i; \boldsymbol{T}_j, \boldsymbol{M}_j), \quad i = 1, \ldots, N$$

Matrix exponential spatial specification (MESS; LeSage and Pace, 2007) of spatial dependence

$$\boldsymbol{S}\phi = \exp(\tau \boldsymbol{W})\phi = \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{Normal}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$$

Pólya-Gamma data augmentation (Polson et al., 2013)

Issue

- Negative binomial distribution does not have a conjugate prior.
- Thus, the conditional distributions of the link function parameters do not constitute known distributions.

Remedy

- Introduce Pólya-Gamma-distributed auxiliary variables.
- Conditional on the auxiliary variable, the likelihood of observed counts is translated into a heteroskedastic Gaussian likelihood.

Case study

- Test model on crash frequency data from 1,158 contiguous road segments of 11 highway facilities in the Greater Houston metropolitan area.
- Observation period covers 4 consecutive calendar years in the period from 2007 to 2010.

Sample description (1,158 road segments; Houston, USA)

	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)	✗	✗	✓	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

Model specifications

- Restricted predictor space (I) with some continuous predictors converted to dummy variables:
  - **NB-BART-I:** proposed NB regression model with spatial error terms and BART-based link function.
  - **NB-fixed:** NB regression model with spatial error terms, linear-in-parameters link function, and all fixed parameters.
  - **NB-random:** NB regression model with spatial error terms, linear-in-parameters link function, and combination of fixed and random parameters.
- Unrestricted predictors space (II) with all predictors in their original form:
  - **NB-BART-II**

Assessment of model fit

Log point-wise predictive density (LPPD; Gelman et al., 2014)

$$\text{LPPD} = \sum_{i=1}^{N} \log \left(\int P(y_i|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i|\boldsymbol{y}) d\boldsymbol{\theta}_i \right),$$

A larger value of LPPD indicates superior goodness of fit.

Root mean square error (RMSE)

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

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 in **bold** font.

Assessment of site ranking ability

We rank sites by their posterior mean probability to belong to the top 5% most hazardous sites in the network.

Important notations:

$H_{\alpha,t}$: a set of hazardous sites in time period t at risk level α (i.e., top 5%).
$R_{h,t}$: the rank of site h in period t.

Site ranking consistency tests (Cheng and Washington, 2008)

Site consistency test

Average of the predicted accident counts $\hat{y}_{h,t+1}$ for period $t+1$ of all sites $h \in H_{\alpha,t}$:

$$\mathcal{T}_{\mathrm{SC},t} = \frac{1}{|H_{\alpha,t}|} \sum_{h \in H_{\alpha,t}} \hat{y}_{h,t+1}, \quad \text{(A larger value is better).}$$

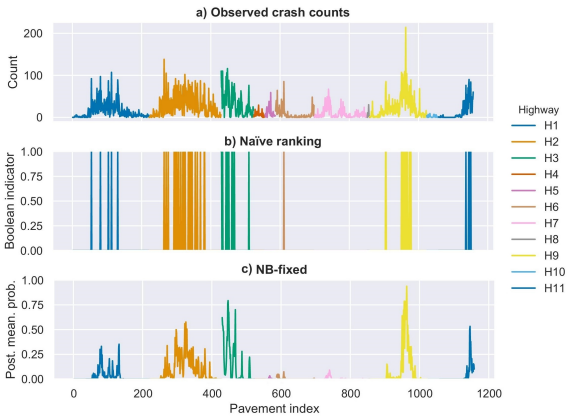
Method consistency test

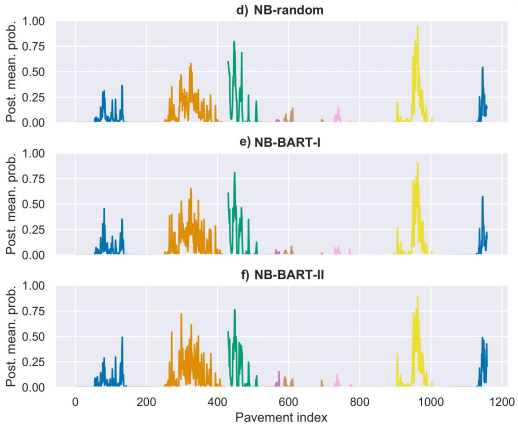
$$\mathcal{T}_{\mathrm{MC},t} = \left| H_{\alpha,t} \cap H_{\alpha,t+1} \right|, \quad \text{(A larger value is better).}$$

Total rank differences test

$$\mathcal{T}_{\mathrm{TRD},t} = \frac{1}{|H_{\alpha,t}|} \sum_{h \in H_{\alpha,t}} \left| R_{h,t+1} - R_{h,t} \right|, \quad \text{(A smaller value is better).}$$

Site ranking results



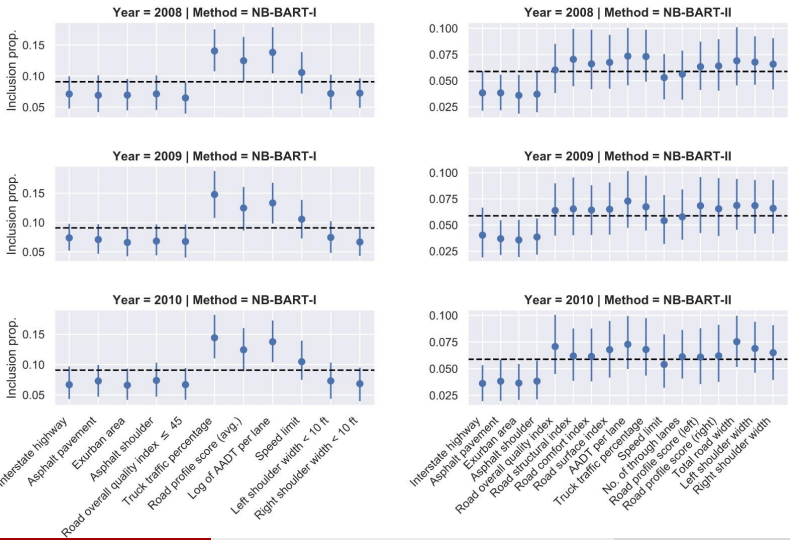


Site ranking consistency

	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.

Variable importance



Conclusions

Key takeaways

- The proposed NB-BART endogenously partitions the predictor space.
- Flexibly accounts for interactions and non-linear relationships between predictors.
- NB-BART performs as well as or better than the state-of-the-art models.
- If predictive performance is paramount, NB-BART may be better than NB models with linear-in-parameters link function and random parameters.

Avenues for future research

- Multivariate NB-BART for the joint modelling of crash counts by crash type.
- Accommodate spatiotemporal heterogeneity.

Thank you

References I

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. and Washington, S. (2008). New criteria for evaluating methods of identifying hot spots. *Transportation Research Record*, 2083(1):76–85.

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., 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.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016.

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.

References II

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.

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.

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

Comparison of Posterior Estimates