Bayesian machine learning and spatial count data models: Advances in estimation and specification

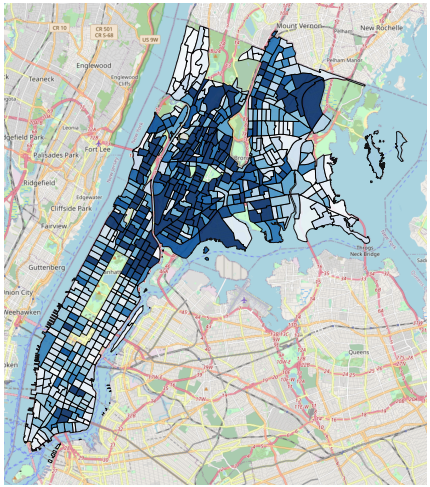
Rico Krueger

Research & teaching associate,
Ecole Polytechnique Fédérale de Lausanne (EPFL)

Research seminar, DTU

Spatial count data models

- → explain and predict frequencies of phenomena such as traffic accidents in geographically distinct entities (e.g. census tracts).
- Spatial count data may exhibit...
  - spatial dependence due to systematic correlation in unobserved factors.
  - spatial heterogeneity due to spatially varying effects of covariates on dependent variable.
- Accounting for spatial effects...
  - complicates model estimation.
  - precludes extensive specification searches.



Bayesian machine learning (Bishop, 2013)

Probabilistic modelling

Bayesian point-of-view

Approximate Bayesian inference

Outline of this talk

1. Fast Bayesian estimation of spatial count data models[1]
2. A new spatial count data model with Bayesian additive regression trees (BART) for accident hot spot identification[2]

Literature:

[1] Bansal, P.*, **Krueger, R.***, Graham, D. J. (2021): Fast Bayesian Estimation of Spatial Count Data Models. *Computational Statistics & Data Analysis*, 157, 107152.

[2] **Krueger, R.***, Bansal, P.*, Buddhavarapu, P. (2020): A New Spatial Count Data Model with Bayesian Additive Regression Trees for Accident Hot Spot Identification. *Accident Analysis & Prevention*, 144, 105623.

* joint first authorship.

Fast Bayesian estimation of spatial count data models

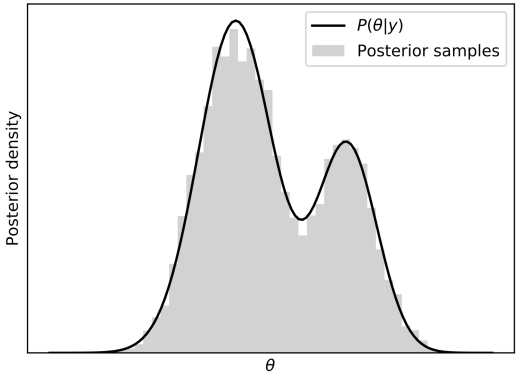
Motivation

Challenges

- Ignoring spatial effects leads to **biased estimates** and **inaccurate inference**.
- Accounting for spatial effects **complicates model estimation**.
- Datasets are growing in size and are becoming available in continuous streams.
- **Scalable** and **computationally-efficient** estimation methods are needed.

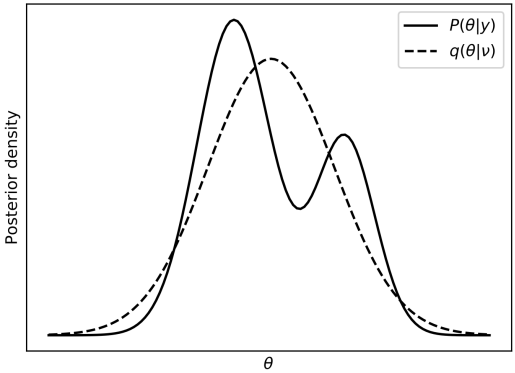
Markov chain Monte Carlo (MCMC)

- Approximate posterior $P(\theta|\boldsymbol{y})$ numerically through samples from a Markov chain.
- Use Metropolis-Hastings algorithm and Gibbs sampling to construct Markov chain.
- **Issues:**
  - computation times,
  - storage of posterior draws,
  - convergence assessment,
  - serial correlation.



Variational Bayes (VB)

- Recast Bayesian estimation as an optimisation problem.
- Approximate posterior $P(\theta|\boldsymbol{y})$ using a parametric variational distribution $q(\theta|\boldsymbol{\nu})$.
- **Advantages:**
  - Reduced storage requirements,
  - Straightforward convergence assessment,
  - Serial correlation no longer a concern,
  - Scalable—stochastic optimisation.



Model formulation

Negative binomial likelihood

$$y_i \sim \text{NB}(r, p_i), \quad p_i = \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}, \quad \psi_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}_i + \phi_i, \quad i = 1, \ldots, N$$

Hierarchical prior for link function parameters to accommodate spatial heterogeneity

$$\boldsymbol{\beta}_i \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, N$$

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

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

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

Issue

- Negative binomial distribution does NOT have a **conjugate prior**.
- Thus, the conditional posterior 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.

Variational Bayes (e.g. Blei et al., 2017)

- VB seeks to minimise the KL divergence (probability distance) between an approximating variational distribution $q(\boldsymbol{\theta})$ and the posterior of interest $P(\boldsymbol{\theta}|\boldsymbol{y})$:

$$q^*(\boldsymbol{\theta}) = \arg\min_q \left\{ \mathrm{KL}\left(q(\boldsymbol{\theta}) || P(\boldsymbol{\theta}|\boldsymbol{y}) \right) \right\}.$$

- q must be selected by the analyst. Its expressiveness determines the **quality** of the variational approximation and the **complexity** of the estimation problem.

Mean-field variational Bayes (MFVB)

- Impose posterior independence between parameter blocks $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_J$:

$$q(\boldsymbol{\Theta}) = \prod_{j=1}^{J} q(\boldsymbol{\Theta}_j)$$

- For **conditionally-conjugate** models, MFVB leads to a coordinate ascent algorithm.
- τ and σ^2 are recovered poorly under MFVB assumption.

Integrated nonfactorised variational Bayes (INFVB; Han et al., 2013)

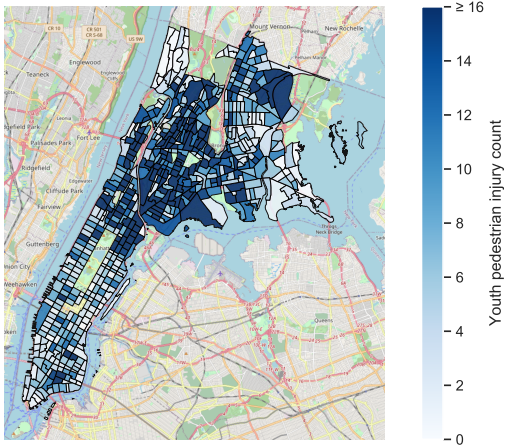
- Decompose parameters $\boldsymbol{\Theta}$ into two disjoint subsets $\{\boldsymbol{\Theta}_c, \boldsymbol{\Theta}_d\}$ to introduce dependencies in variational distribution ($\boldsymbol{\Theta}_d = \{\tau, \sigma^{-2}\}$):

$$q_{\text{INFVB}}(\boldsymbol{\Theta}) = q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d)q(\boldsymbol{\Theta}_d)$$

- Direct minimisation of KL divergence between q_{INFVB} and posterior is challenging.
- Conditional on $\boldsymbol{\Theta}_d$, INFVB involves a simple coordinate ascent algorithm.
- Define a grid with points $\boldsymbol{\Theta}_d^{(g)}$, and run MFVB separately for each grid point. Then, compute weight of each grid point using the Boltzmann distribution.
- Embarrassingly parallel architecture.

Case study

- Benchmark INFVB against MCMC.
- Youth pedestrian injury counts in 603 census tracts of the Bronx and Manhattan from 2005–14 (Morris et al., 2019).



Sample description (N = 603)

Variable	Mean	Std.	Min.	Max.
Youth pedestrian injury count, 2005-14	9.69	8.35	0.00	44.00
Prop. of households with poverty status, 2012-16	0.24	0.15	0.00	0.57
Prop. of black or African-American alone population, 2012-16	0.24	0.22	0.00	0.91
No. of workers per km^2 in 1000, 2012-16	17.96	37.34	0.02	260.40
Social fragmentation index	2.02	2.73	-4.50	18.67
Avg. annual daily traffic (AADT) in 10k, 2015	4.45	4.68	0.21	27.65
Private vehicle commute mode share, 2010-14	0.19	0.15	0.00	0.76

Estimation time and goodness of fit comparison

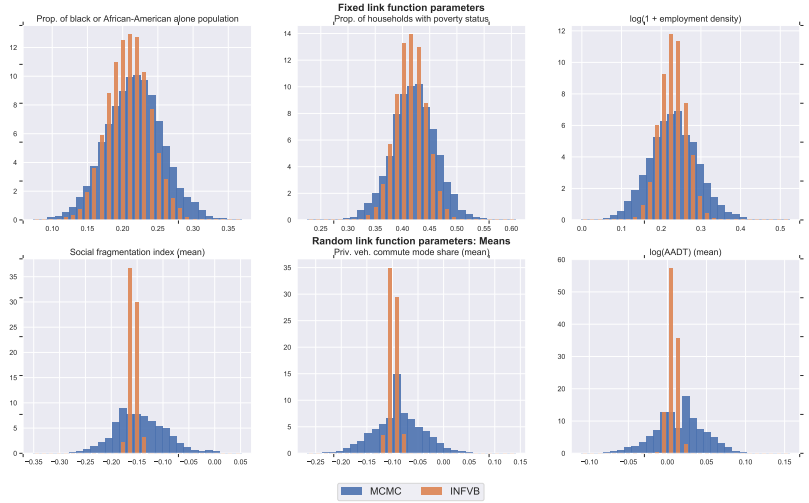
Estimation times:

- MCMC: 135.9 minutes
- INFVB: 2.9 minutes

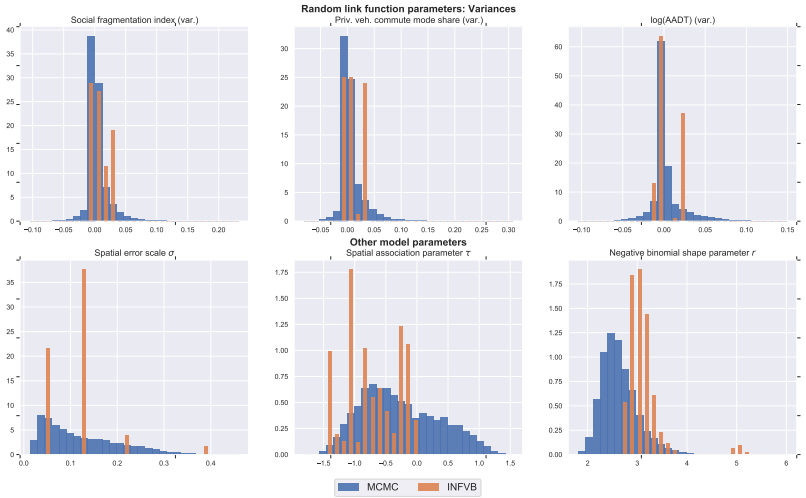
Goodness of fit:

	MCMC			INFVB		
Score	**Mean**	**[2.5%;**	**97.5%]**	**Mean**	**[2.5%;**	**97.5%]**
Log	1846.3	[1785.2;	1878.1]	1832.5	[1770.8;	1855.7]
Dawid-Sebastiani	2762.3	[2588.0;	2864.7]	2720.2	[2552.0;	2796.3]
Ranked probability	2159.6	[1953.9;	2275.4]	2102.5	[1858.2;	2192.0]

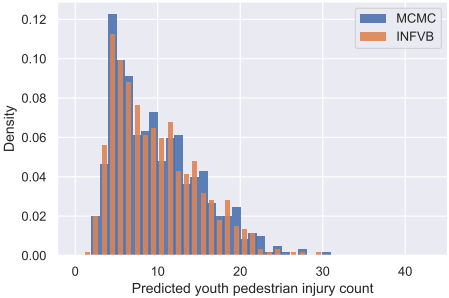
Marginal posterior distributions I

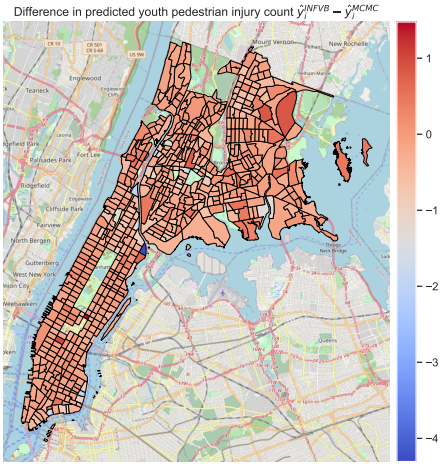


Marginal posterior distributions II



Predicted youth pedestrian injury counts





Conclusion

Key points

- In application, INFVB is $\approx$ 47 times faster than MCMC on a regular eight core computer.
- INFVB can be made 20 times faster due to its embarrassingly parallel architecture.
- INFVB offers similar estimation and predictive accuracy as MCMC.

Future research directions

- Extension to models with spatiotemporal dependencies.
- Online estimation using stochastic variational inference.

A new spatial count data model with Bayesian additive regression trees for accident hot spot identification (NB-BART)

Motivation

Crash count data models

$\Downarrow$

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

Traditional crash count data models:

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

Motivation

- Linear-in-parameters link function is suitable for interpretability, but **predictive performance** is of paramount importance.
- Accounting for spatial dependence precludes extensive specification searches.
- Modern machine learning (ML) methods **automate specification searches**:
  - 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 dependence between observations.
- Do not offer straightforward ways to construct confidence or credible intervals.
- Fully nonparametric, with no easy way to include interpretable aspect in link function.

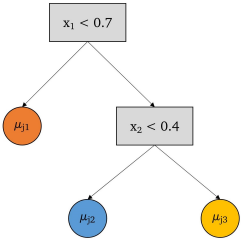
Research objective

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

- Specifies non-parametric aspect of 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 dependence.
- 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

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

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

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

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

SAME AS IN PART I

Issue

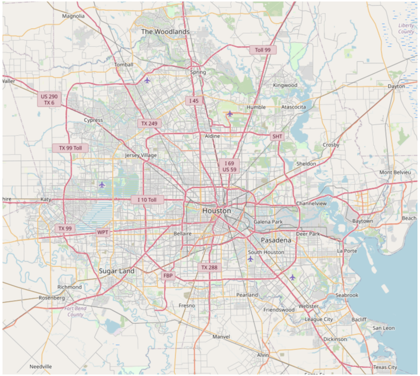
- Negative binomial distribution does NOT have a **conjugate prior**.
- Thus, the conditional posterior distributions of the link function parameters do not constitute known distributions.

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 real **crash frequency data**
- 1,158 contiguous road segments of 11 highway facilities in the Greater Houston metro 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**

Goodness of fit

Method	2007		2008		2009		2010	
	Log	RMSE	Log	RMSE	Log	RMSE	Log	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**

Log = log score; 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

Site ranking

- Sites are ranked by their posterior mean probability to belong to the top 5% most hazardous sites in the network.
- $H_{\alpha,t}$: the set sites identified as hazardous in time period t at risk level α (i.e., top 5%)

Site ranking consistency test (Cheng and Washington, 2008)

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

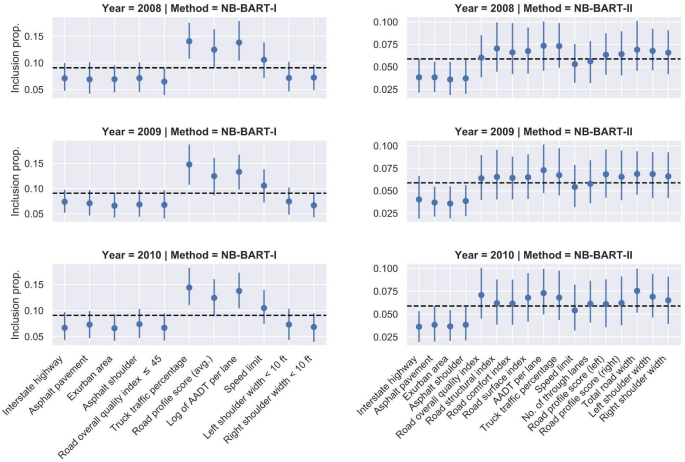
$$\mathcal{T}_{\mathrm{SC},t} = \frac{1}{|H_{\alpha,t}|} \sum_{h \in H_{\alpha,t}} \hat{y}_{h,t+1}, \quad \text{(A larger value is better).}$$

Site ranking consistency: Test scores

Method	2007	2008	2009
NB-fixed	54.95	46.94	60.87
NB-random	56.35	47.48	61.10
NB-BART-I	57.52	49.80	61.84
NB-BART-II	**61.05**	**53.73**	**69.09**

For each reference period, the best-performing hot spot identification method is highlighted in bold font.

Variable importance



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.

Conclusion

Key points

- NB-BART flexibly accounts for interactions and non-linear relationships between predictors.
- NB-BART outperforms state-of-the-art methods.

Future research directions

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

Thank you

References I

Bishop, C. M. (2013). Model-based machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20120222.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

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.

References II

Han, S., Liao, X., and Carin, L. (2013). Integrated non-factorized variational inference. In *Advances in Neural Information Processing Systems*, pages 2481–2489.

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.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., and DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the besag york mollié model in stan. *Spatial and spatio-temporal epidemiology*, 31:100301.

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.