Mortality Mapping with R-INLA Under Bayesian Framework

With Application to Lung Cancer Mortality in PA

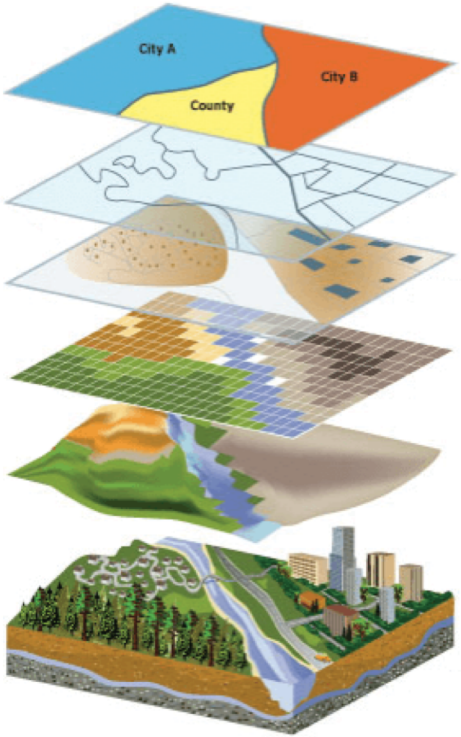
Xiaoqing Tan, Anran Liu, Liwen Wu
4/15/2019

Overview

- Spatial Data and Analysis
- Bayesian Approach in Spatial Modelling
- Integrated Nested Laplace Approximation (INLA)
- Real-world Data Application
- Conclusions
- Limitations

Spatial Data

- Inherently contains geometric and topological properties.
- Location: defined by coordinates in a specific reference system.
- Temporality: existence and change over time of features.
- Generalization: relates to the level of scale and details associated with the object.
- Spatial dependency: highly correlated between neighbors.



Spatial Analysis

- A type of geographical analysis which seeks to explain relationships of human behavior and its spatial patterns.
- Highly effective for evaluating the geographic suitability of certain locations for specific purposes.

Map by Dr. John Snow, showing clusters of cholera cases in the 1854 Broad Street cholera outbreak.

Bayesian Approach in Spatial Modelling

- Advantages:
  - MCMC methods are very general, can be effectively applied to “any” model and give exact inference in theory.
  - Standard MCMC samplers are easy to program and available in many software (e.g. WinBUGS, JAGS).
- Disadvantages:
  - Requires much longer computational time, especially for large data or complex structure (e.g. hierarchical models).

Integrated Nested Laplace Approximations (INLA)

- INLA is a fast alternative to MCMC for the general class of latent Gaussian models (LGMs).
- A wide class of statistical models can be re-cast to look like LGMs, such as:
  - GLM(M)s, GAM(M)s, spatio-temporal models, survival analysis, spline smoothing, etc.
- To understand the idea of what INLA is doing, we need to be familiar with:
  - Bayesian Inference
  - Latent Gaussian models (LGMs)
  - Gaussian Markov Random Fields (GMRFs)
  - Laplace Approximations

Latent Gaussian Models

- Generally speaking, the class of LGM can be represented by a hierarchical structure containing three stages.
  - First stage (conditionally independent likelihood function):

$$\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi} \sim p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta}, \boldsymbol{\psi})$$

  - Second stage (assume the parameters are described by GMRF):

$$\boldsymbol{\theta} \mid \boldsymbol{\psi} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\psi}))$$
$$\theta_l \perp\!\!\!\perp \theta_m \mid \boldsymbol{\theta}_{-lm}$$

  - Third stage (prior distribution assigned to the hyperparameters):

$$\boldsymbol{\psi} \sim p(\boldsymbol{\psi})$$

LGMs as a general framework

- A very general way of specifying the problem is by modelling the mean for the i-th unit by means of an additive linear predictor, defined on a suitable scale (e.g. log for poisson data)

$$\eta_i = \alpha + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li})$$

where

- α is the intercept;
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$ quantify the effect of $\boldsymbol{x} = (x_1, \ldots, x_M)$ on the response;
- $\boldsymbol{f} = \{f_1(\cdot), \ldots, f_L(\cdot)\}$ is a set of functions defined in terms of some covariates $\boldsymbol{z} = (z_1, \ldots, z_L)$

and then assume

$$\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \boldsymbol{f}) \sim \text{GMRF}(\boldsymbol{\psi})$$

Upon varying the form of the functions $f_l(\cdot)$, this formulation can accommodate a wide range of models

- Standard regression
  - $f_l(\cdot) = \text{NULL}$
- Hierarchical models
  - $f_l(\cdot) \sim \text{Normal}(0, \sigma_f^2)$ (Exchangeable)
    $\sigma_f^2 \mid \psi \sim$ some common distribution
- Spatial and spatio-temporal models
  - Two components: $f_1(\cdot) \sim \text{CAR}$ (Spatially structured effects)
    $f_2(\cdot) \sim \text{Normal}(0, \sigma_{f_2}^2)$ (Unstructured residual)
- Spline smoothing
  - $f_l(\cdot) \sim \text{AR}(\phi, \sigma_\varepsilon^2)$
- Survival models / logGaussian Cox Processes
  - More complex specification in theory, but relatively easy to fit within the INLA framework!

Laplace approximation

$$p(\theta_i|y) = \int p(\psi|y)p(\theta_i|\psi, y)d\psi$$

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)} \propto \frac{p(\psi)p(\theta|\psi)p(y|\theta, \psi)}{p(\theta|\psi, y)}$$

- A Laplace approximation is used to estimate any distribution (PDF) with a normal distribution using the first 3 terms of the taylor series expansion.
- The idea is simple but powerful: Approximate the target with a Gaussian, matching the mode and the curvature at the mode.

$$\approx \frac{p(\psi)p(\theta|\psi)p(y|\theta, \psi)}{\tilde{p}(\theta|\psi, y)}|_{\theta=\theta^*(\psi)} =: \tilde{p}(\psi|y)$$

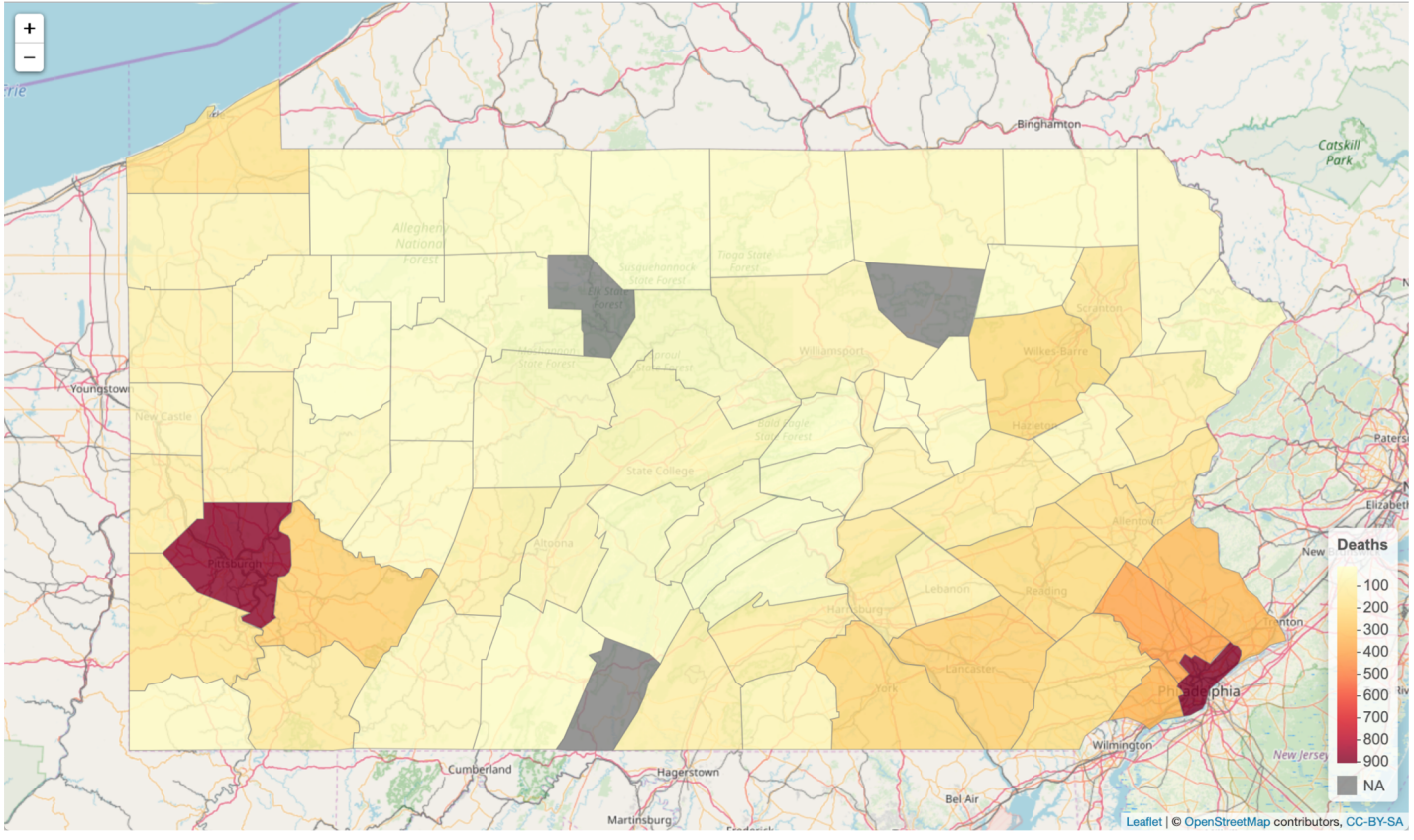
Real Data Application
PA lung cancer mortality data

Data

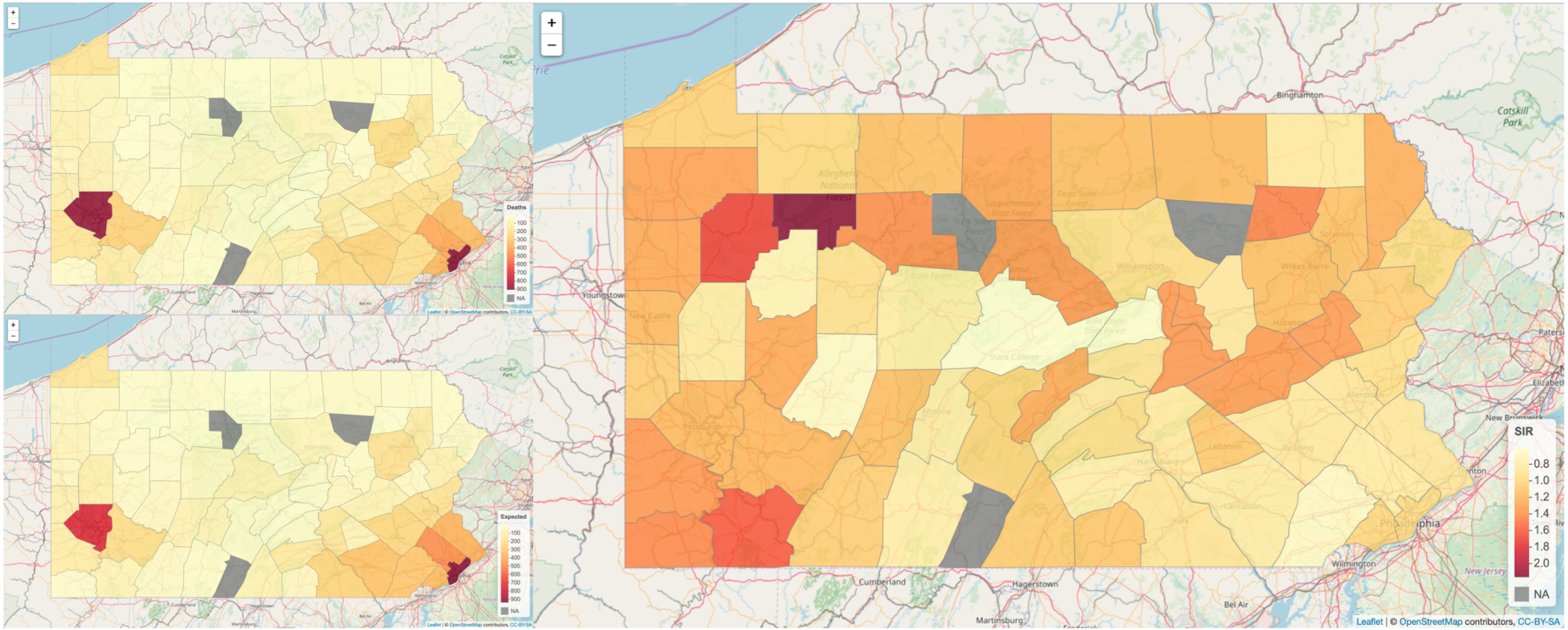
Lung cancer mortality: 1989-2016 PA lung cancer mortality data at county level from Pitt Public Health's Mortality Information and Research Analytics (MOIRA) System
(Courtesy of Kendra Bobby and Dr. Saumyadipta Pyne)

Covariates:
Radon level: PA Department of Environmental Protection (DEP) Radon Test Results Data from 1989 to 2017
Smoking percentage: 2002 county specific smoking percentage data obtained from the pennLC dataset in the SpatialEpi package in R.

PA lung cancer mortality (2012)



Observed/Expected mortality outcomes



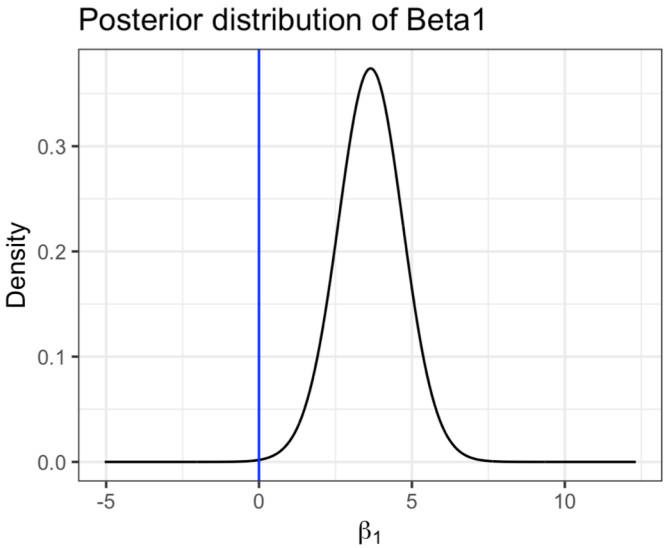
Model Specification

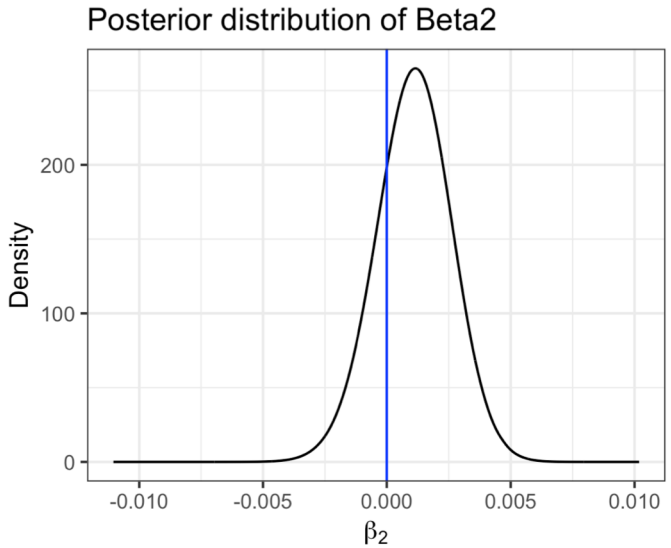
- For our lung cancer mortality application in PA:
  - $Y_i \sim Poisson(\lambda_i E_i)$
  - $\log(\lambda_i) = \beta_0 + \beta_1 * smoke + \beta_2 * radon + U_i + V_i$
  - $U_i \sim CAR$: $U_i | U_{j \neq i} \sim N(n_i^{-1} \sum_{N_i} U_j , n_i^{-1} \sigma_U^2)$
  - $V_i \sim_{iid} N(0, \sigma_V^2)$
- R-INLA default priors are used
  - Fixed effect: $\beta \sim Normal(0, 0.001)$
  - Random effect: $\theta_U = log\ \sigma_U^{-2}$, $\theta_U \sim logGamma(1, 0.00005)$

$$\theta_V = log\ \sigma_V^{-2}, \quad \theta_V \sim logGamma(1, 0.00005)$$

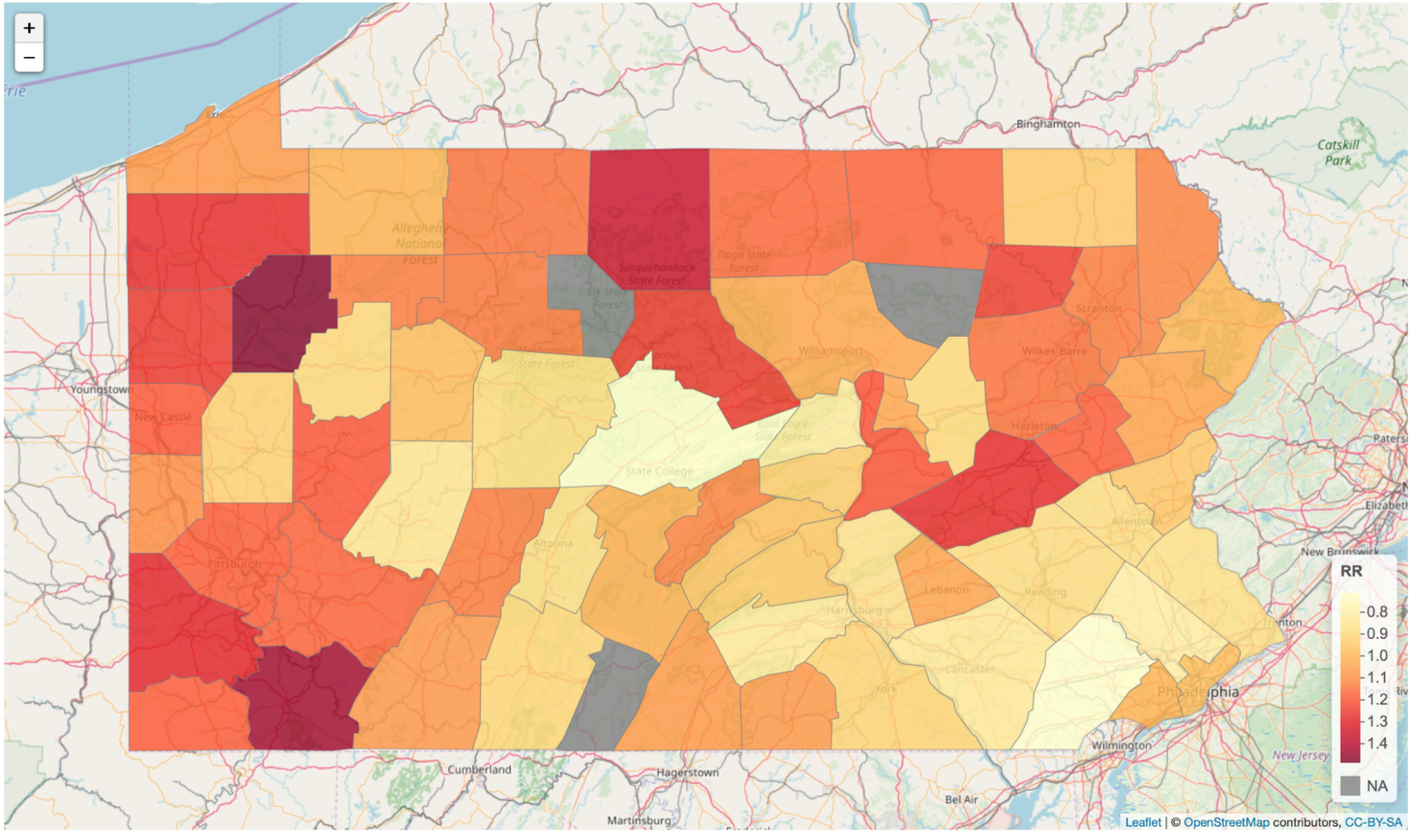
Inference based on INLA

Parameters	Posterior Median	95% Credible Interval
Smoking	3.6433	(1.4917, 5.7660)
Radon	0.0011	(-0.0020, 0.0040)





Map of Relative Risk for Lung Cancer Mortality



Conclusion

- We modeled the risk of lung cancer mortality in PA with a Bayesian spatial model. The spatial correlation was accounted by a conditional independent autoregressive structure for neighboring counties. A weakly-informative prior was used in INLA for model inference.
- Both the proportion of smokers and radon level in each county is positively associated with lung cancer mortality.
- Counties in western PA in general show a higher risk compared to eastern counties.

Limitations

- Data quality
  - Smoking data was estimated by a very small sample of population in each county (ie. hundreds)
  - Under-reporting in mortality
- Potentially important demographic variables are not included
- Temporal effect
  - We naively created a 10 year time lag between exposure variables and outcome, but this time effect could also be investigated if we could have more information on exposure status.

References

- Moraga, P. (2018). Small Area Disease Risk Estimation and Visualization Using R. *The R Journal*, *10*(1), 495-506.
- Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, *4*, 395-421.
- Ferkingstad, E., Held, L., & Rue, H. (2017). Fast and accurate Bayesian model criticism and conflict diagnostics using R-INLA. *Stat*, *6*(1), 331-344.
- http://www.statistica.it/gianluca/Talks/INLA.pdf
- https://www.precision-analytics.ca/blog-1/inla

Supplemental slides

Model criticism with Node-splitting analysis

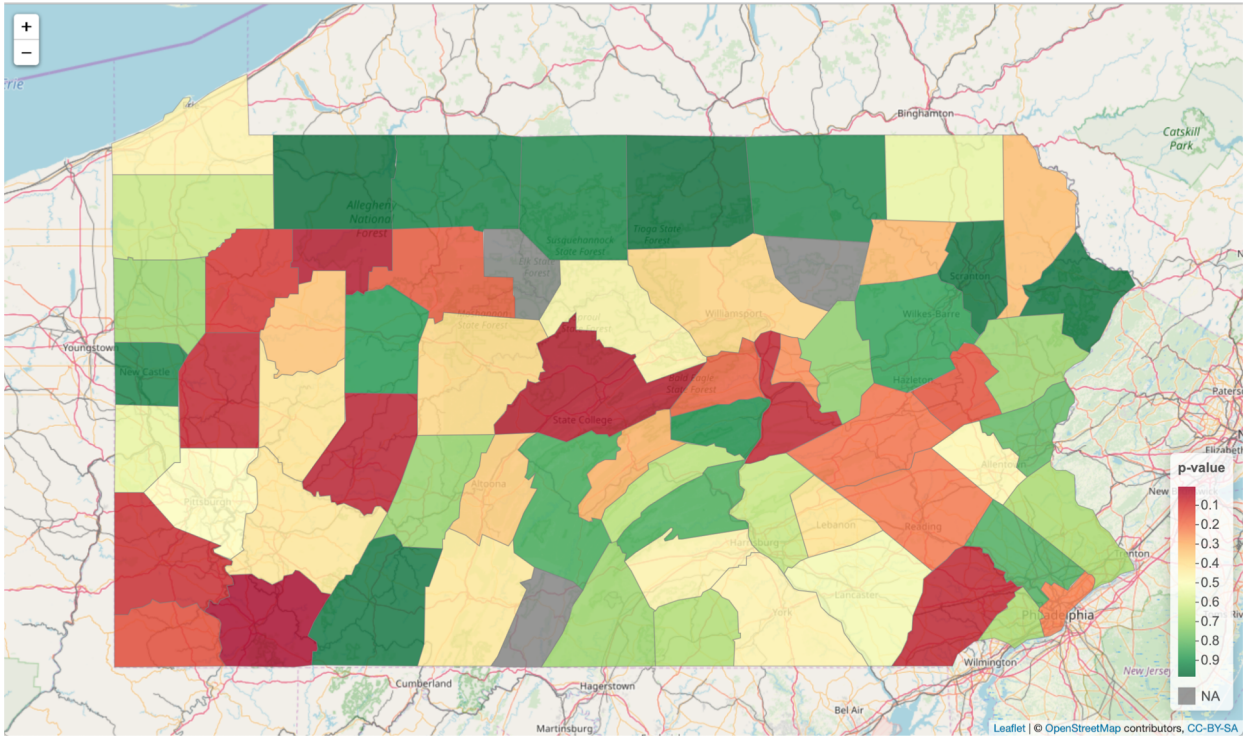
An intuitive way to check the spatial assumption is to exam how much the count for a specific county is different from what would be estimated from all the other (surrounding) counties.

Ferkingstad et al. (2017) introduces a simple test $H_0: \delta_j = 0$ for each county j

$$\delta_j = \theta_j^{-j} - \theta_j^{j} \qquad \text{where } \theta_j^{-j} \sim \pi\left(\theta_j \middle| y_{-j}\right) \text{ and } \theta_j^{-j} \sim \pi\left(\theta_j \middle| y_{-j}\right)$$

P-values are calculated as the probability of rejecting H_0 based on Wald-type tests

Model criticism with Node-splitting analysis



County	p-value
Fayette	0.0067
Forest	0.0267
Centre	0.0289
Indiana	0.0390
Northumberland	0.0484

Possible solutions

- More complex model specification
  - Blocking (Allocate the parameters into blocks and update them separately.)
  - Overparameterization
- More complex sampling schemes
  - Hamiltonian Monte Carlo
  - No U-turn sampling (e.g. stan — a variant of Hamiltonian Monte Carlo)
- Alternative methods of inference
  - Approximate Bayesian Computation (ABC)
  - Integrated Nested Laplace Approximations (INLA)

Thank you for listening!