Modeling of Crime Data to Detect Social and Spatial Proximity

Claire Kelling*

Advisors: Murali Haran*, Corina Graif**

Penn State University
Department of Statistics*, Department of Sociology and Criminology**

STAT 544: Categorical Data Analysis

Topics

Motivation

Geographic proximity has been studied over many years

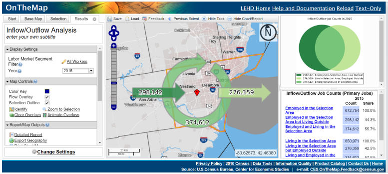
- mostly focusing on identifying hotspots in certain communities
- led to many controversial policing strategies, such as predictive policing, which was referenced repeatedly in "Weapons of Math Destruction" (O'Neill, 2016).

Our Approach:

- Demographics
- Geographic Proximity
- Social Proximity

Data Sources

- Crime: Police Data Initiative
- Demographics: American Communities Survey (Census)
- Social Proximity: LODES (Origin-Destination Employment Statistics)



Variables considered

Response Variable: **crime count** per block group

Covariates of interest:

- median income
- median age
- percentage female
- unemployment rate
- total population
- Herfindahl Index (HI) [1]

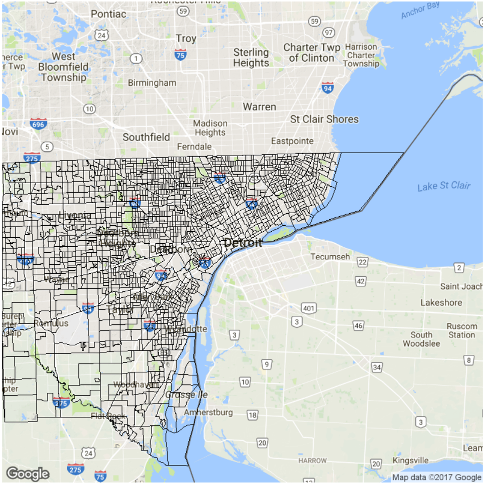
$$HI = 1 - \left(\left(\frac{\#\text{ white}}{\text{total pop}} \right)^2 + \left(\frac{\#\text{ black}}{\text{total pop}} \right)^2 + \left(\frac{\#\text{ Latino}}{\text{total pop}} \right)^2 + \ldots \right)$$

For the HI, 0 means no diversity, and approaching 1 means more diversity.

- Ex: if a population in a block group is completely white, $HI = 1 - (1^2 + 0 + 0 + \ldots) = 0$

Area of Interest

Wayne County (Detroit)



Bayes CAR Model [2] Set Up

CAR = Conditional Autoregressive model for areal data

- study region $\mathcal{S}$ is partitioned into K non-overlapping areal units
- linked to set of responses $\mathbf{Y} = (Y_1, ..., Y_K)$
- spatial variation in the response is modeled by a matrix of covariates $\mathbf{X} = (x_1, ..., x_k)$ and a spatial structure component $\psi = (\psi_1, ..., \psi_k)$
- $\psi = (\psi_1, ..., \psi_k)$ models any spatial autocorrelation that remains after covariate effects have been accounted for

GLMM for spatial areal unit data

$$Y_k|\mu_k \sim f(y_k|\mu_k, \nu^2) \text{ for k} = 1,...,\text{K}$$
$$g(\mu_k) = \mathbf{x_k}^\mathbf{T}\beta + \psi_\mathbf{k}$$
$$\beta \sim N(\mu_\beta, \Sigma_\beta)$$

- Poisson: $Y_k \sim Poisson(\mu_k)$ and $ln(\mu_k) = \mathbf{x}_\mathbf{k}^\mathbf{T}\beta + \psi_\mathbf{k}$

BYM Model [3]

$$\psi_k = \phi_k + \theta_k$$
$$\phi_k | \phi_{-k}, \mathbf{W}, \tau^2 \sim N\left(\frac{\sum_{i=1}^K w_{ki}\phi_i}{\sum_{i=1}^K w_{ki}}, \frac{\tau^2}{\sum_{i=1}^K w_{ki}}\right)$$
$$\theta_k \sim N(0, \sigma^2)$$
$$\tau^2, \sigma^2 \sim \text{Inverse-Gamma}(a, b)$$

- First CAR model to be proposed.
- Two sets of random effects, spatially autocorrelated and independent
- Called the intrinsic CAR model
- Requires two random effects to be estimated at each data point, whereas only their sum is identifiable
- this model is equivalent to the multivariate specification $\phi \sim N(0, \tau^2\mathbf{Q}(\mathbf{W})^{-1})$, where $\mathbf{Q}(\mathbf{W}) = \text{diag}(\mathbf{W1}) - \mathbf{W}$

Leroux Model [5]

$$\psi_k = \phi_k$$

$$\phi_k | \phi_{-k}, \mathbf{W}, \tau^2, \rho \sim N\left(\frac{\rho \sum_{i=1}^K w_{ki}\phi_i}{\rho \sum_{i=1}^K w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{i=1}^K w_{ki} + 1 - \rho}\right)$$

$$\tau^2 \sim Inverse - Gamma(a, b)$$

$$\rho \sim Uniform(0, 1)$$

- Uses only single of random effects
- ρ is a spatial autocorrelation parameter
- $\mathbf{W}$ is the neighborhood matrix
- this model is equivalent to the multivariate specification $\phi \sim N(0, \tau^2 \mathbf{Q}(\mathbf{W}, \rho)^{-1})$, where $\mathbf{Q}(\mathbf{W}, \rho) = \rho[\text{diag}(\mathbf{W1}) - \mathbf{W}] + (\mathbf{1} - \rho)\mathbf{I}$
- Widely said to be the most appealing CAR model, from both theoretical and practical standpoints [4]

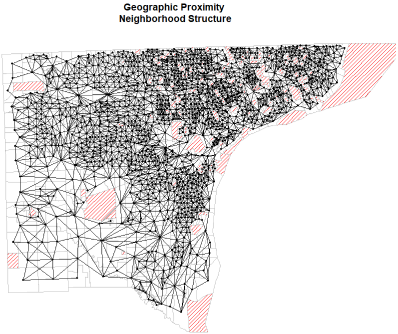
Neighborhood Matrix, $\mathbf{W}$

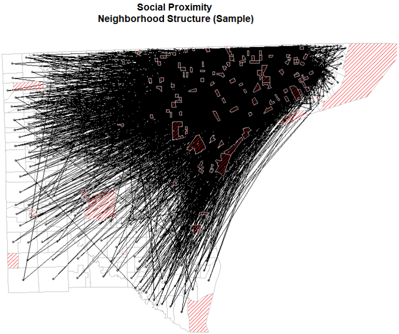
Some details on neighborhood matrix, $\mathbf{W}$

- non-negative, symmetric, K$\times$K
- (k,j)th element of the neighborhood matrix w_{kj} represents spatial closeness between areas $(\mathcal{S}_k, \mathcal{S}_j)$
- positive values denoting geographical closeness and zero values denoting non-closeness (0-1 is the most common structure)
- $w_{kk} = 0$

Neighborhood Structure

Establishing neighbors based on geographic or social proximity:





Irregular Block group size

What about irregular block group sizes and shapes?

- This model only depends on the structure of $\mathbf{W}$, the neighborhood matrix
- It does not account for the shape/size of the spatial areal unit
- The neighborhood matrix accounts for a bit of the irregularity through adjacency (some units have 4 neighbors, some have 10, etc)
- We can somewhat account for the size of the block group in our model through population size

Evaluation of Spatial Autocorrelation

Moran's I

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{ij} w_{i \neq j})(Y_i - \bar{Y})^2}$$

- Asymptotically normal but convergence is very slow $\rightarrow$ permutation tests
- Calculated for our data to make sure spatial modeling is appropriate:
  - Geographic Proximity: 0.55932 (p-value = 0.000999)
  - Social Proximity: 0.084956 (p-value = 0.0006662)

Model Comparison Criteria

Deviance Information Criteria **(DIC)** ="goodness of fit" + "complexity"[6]

Measure **fit** via the deviance

- $D(\theta) = -2logL(data|\theta)$

Measure **complexity** by the estimate for effective # of parameters

- $p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta}) =$ posterior mean deviance - deviance evaluated at the posterior mean of the parameters

DIC $= D(\bar{\theta}) + 2p_D = \bar{D} + p_D$

- Models with smaller DIC are better supported by the data (just like with AIC)

Model Comparison Criteria

We use the following criteria for model comparison

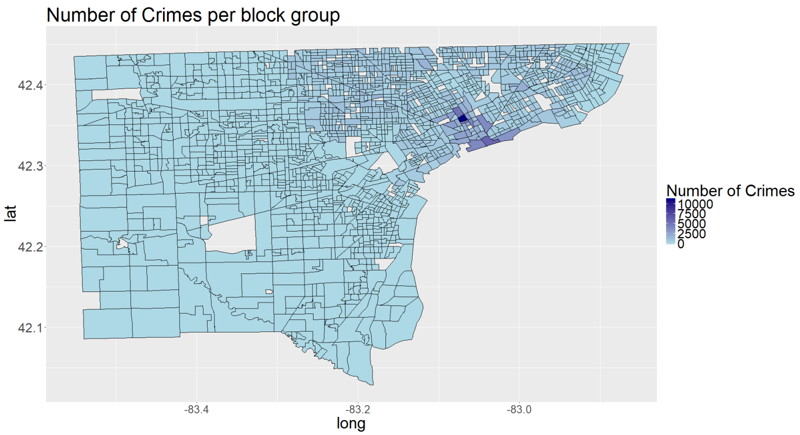
- Deviance Information criterion (DIC)
- corresponding estimated effected number of parameters (p.d)
- Percentage Deviance Explained

Model Comparison

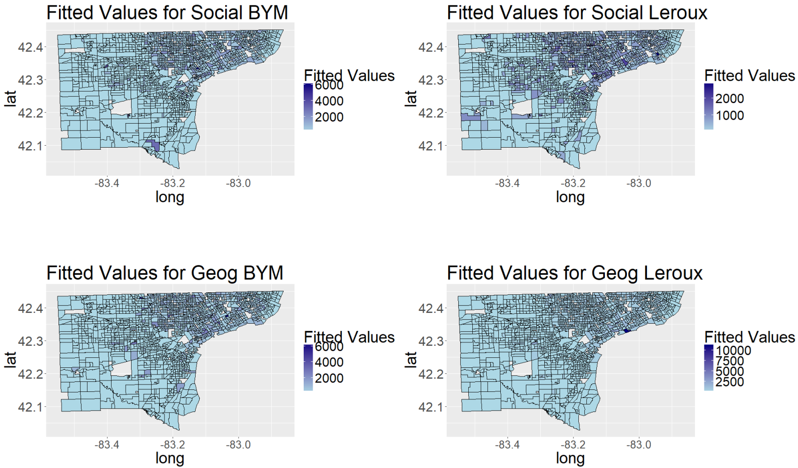
Criterion	Social BYM	Geog BYM	Social Leroux	Geog Leroux
DIC	671,819	564,584	704,606	410,674
p.d	1,978	1,190	660	672
% deviance explained	35	45	32	60

Number of Crimes

Recall the distribution of the number of crimes:



Fitted Plots



Next Steps

- Find a way to fit both the social and geographic proximity into the same model
- Improve on the current literature's approach: including both as weighted regression using leave-one-out
- Try the above analysis on a subset of the crime data, for different crime types (ex: domestic violence, drug crimes, crimes related to mental health)
- Try the above analysis on other cities that are listed on the Police Data Initiative's website

References

[1] Stephen A Rhoades. The herfindahl-hirschman index. *Fed. Res. Bull.*, 79:188, 1993.

[2] Duncan Lee. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24, 2013.

[3] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

[4] Duncan Lee. A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2(2):79–89, 2011.

[5] G Leroux Brian, Xingye Lei, Norman Breslow, M Halloran, and Berry Donald Elizabeth. Estimation of disease rates in small areas: a new mixed model for spatial dependence. *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191, 2000.

[6] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit.