Combining Geographic and Social Proximity to Model Urban Domestic and Sexual Violence

Claire Kelling
Penn State University
State College, Pennsylvania
ckelling@vt.edu

Gizem Korkmaz
Biocomplexity Institute of Virginia Tech
Arlington, VA
gkorkmaz@vt.edu

Corina Graif (if desired)

Murali Haran (if desired)

ABSTRACT

In order to understand the dynamics of crime in urban areas, it is important to understand the culture of communities and neighborhoods. Many studies have created analyses of crime over space through spatial statistical models such as spatially weighted regression. However, in order to obtain a complete picture of the proximity between two neighbors, we must not only consider geographic proximity, but also social proximity. If there are strong social ties between two neighborhoods, they are more likely to transfer ideas, customs, and perhaps behaviors. This implies that crime could be transfered along social ties, but crime prevention/interventions can also transfer over these ties. In our modeling framework, we combine geographic and spatial proximity using spatial Generalized Linear Models. The area of our study is Wayne County, Michigan (Detroit). The analysis relies on data from various publicly available data sources such as the Police Data Initiative and Census. We investigate the difference between geographic and social proximity with Bayesian model comparison techniques and draw conclusions about the strengths of these models for this data.

KEYWORDS

crime, spatial GLMM, social proximity, commuting data

ACM Reference Format:
Claire Kelling, Gizem Korkmaz, Corina Graif (if desired), and Murali Haran (if desired). 2018. Combining Geographic and Social Proximity to Model Urban Domestic and Sexual Violence. In *Proceedings of ACM KDD Conference (KDD'18).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

2 RELATED WORK

Insert (more) Literature Review/Motivation Here

Recent literature on crime has focused on modeling the diffusion of crime between neighborhoods or communities [12]. Neighborhoods are often treated as closed systems where there is no transfer of ideas or behaviors inside of a city once you leave some geographic bound. We know this is not the case, as there is a national and international transfer of cultures and customs due to the modern ease of travel and communication. We are interested in both the geographic and social proximity between neighborhoods. Geographic proximity has been studied over many years, mostly focusing on identifying hotspots in certain communities. This has led to many controversial policing strategies, such as predictive policing, which was referenced thoroughly in "Weapons of Math Destruction" [11], where certain neighborhoods are predominantly targeted, mainly based on race and other demographic information.

We are interested in investigating what other ties, such as social, might exist between neighborhoods, in order to create more effective policing interventions. Social proximity is a relatively new topic in the criminology literature. Wang et al. use taxi data in Chicago to establish social proximity between neighborhoods. We will use commuting data, available through the Census, to establish social proximity. We compare models that incorporate purely geographic (spatial) proximity with models that incorporate both social and geographic proximity to evaluate whether commuting data adds information about ties between communities that may affect crime.

3 PROPOSED METHODS

In this paper, we rely on the Conditional Autoregressive model (CAR) model for analyzing our areal data, a popular method in sociology, political science, epidemiology and many more applications. We use the CARBayes package in R, and the descriptions below will serve as a literature review of the vignette which describes the use of this package [7]. We will consider two models that are presented in the vignette, the BYM and Leroux models, and compare their results for geographic proximity to geographic proximity combined with social proximity. We also compare these models to the sparse Spatial Generalized Linear Mixed Model (SGLMM), which controls for potential spatial confounding present in the BYM and Leroux models. All of these models have a common base structure, which is described below.

For this framework, we assume that our study region $\mathcal{S}$ is partitioned into K non-overlapping areal units. In our case, this is our 1,822 block groups in Wayne County. These areal units are linked to set of responses $\mathbf{Y} = (Y_1, ..., Y_K)$, or the aggregated crime counts for each block group. The model referenced in the paper also includes a vector of known offsets $\mathbf{O} = (O_1, ..., O_k)$ that are associated with each areal unit. This is not relevant to our application, so is not included in the model referenced below. Spatial variation in

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
KDD'18, August 2018, London, United Kingdom
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the response is modeled by a matrix of covariates $\mathbf{X} = (x_1, ..., x_k)$ and a spatial structure component $\psi = (\psi_1, ..., \psi_k)$. We see that $\psi = (\psi_1, ..., \psi_k)$ models any spatial autocorrelation that remains after covariate effects have been accounted for. Lee references that these models are a special case of a Gaussian Markov Random Field (GMRF).

We will use a Generalized Linear Mixed Model (GLMM) framework for spatial areal unit data. The framework for the GLMM can be seen in equation 1, with the parameters described above.

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

Due to the fact that our response is count data over the areal units, we will be using a Poisson form of the GLMM. The two other options for this package are Gaussian or Binomial. Over-dispersion may be an issue in our data, but the package currently does not support either a negative binomial or a quasi-poisson model. So, we assume $Y_k \sim Poisson(\mu_k)$ and $ln(\mu_k) = \mathbf{x_k^T}\beta + \psi_{\mathbf{k}}$.

3.1 BYM Model

The first model we consider for the ψ_k parameter for spatial autocorrelation is called the BYM model, named for the initials of the three authors on the paper (Besag, York, and Mollié) [2]. This was the first CAR model to be proposed, and it is also called the intrinsic CAR model. There are two sets of random effects, spatially autocorrelated and independent. The full model specification in the Bayesian framework can be seen in equation 2. This model is equivalent to the multivariate specification $\phi \sim N(0, \tau^2\mathbf{Q}(\mathbf{W})^{-\mathbf{1}})$, where $\mathbf{Q}(\mathbf{W}) = \text{diag}(\mathbf{W1}) - \mathbf{W}$.

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

This model requires two random effects to be estimated at each data point, whereas only their sum is identifiable. This is one of the main reasons why the following model, the Leroux Model, was proposed.

3.2 Leroux Model

Next, we turn to our second model for spatial auto-correlation ψ_k, which was presented by Brian Leroux et al in 2000 [3]. The Bayesian model specification can be found in equation 3. This model uses only a single random effect, ϕ_k and incorporates another spatial autocorrelation parameter. This parameter ρ is a spatial autocorrelation parameter that is added to the model, which no longer includes θ_k. This model is equivalent to the multivariate specification $\phi \sim N(0, \tau^2\mathbf{Q}(\mathbf{W}, \rho)^{-\mathbf{1}})$, where $\mathbf{Q}(\mathbf{W}, \rho) = \rho[\text{diag}(\mathbf{W1}) - \mathbf{W}] + (\mathbf{1} - \rho)\mathbf{I}$ This version of the CAR model has been widely said to be the most appealing CAR model, from both theoretical and practical standpoints [6].

$$\begin{aligned} \psi_k &= \phi_k \\ \phi_k|\phi_{-k}, \mathbf{W}, \tau^2, \rho &\sim N\Big(\frac{\rho\sum_{i=1}^K w_{ki}\phi_i}{\rho\sum_{i=1}^K w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{i=1}^K w_{ki} + 1 - \rho}\Big) \\ \tau^2 &\sim \text{Inverse-Gamma}(a, b) \\ \rho &\sim \text{Uniform}(0, 1) \end{aligned} \tag{3}$$

3.3 Sparse SGLMM

In comparison to the results of the BYM and Leroux model, we also consider the sparse SGLMM model [5]. The Sparse SGLMM model is an extension of the restricted spatial regression (or RHZ model) proposed to control for spatial confounding [8]. In our original model parameterization in equation 1, we see that $g(\mu_k) = x_k^T\beta + \psi_k$. Reich et al. set up their proof by letting $\mathbf{P}$ be the orthogonal projection onto the regression manifold C($\mathbf{X}$), and constructing the eigendecomposition of $\mathbf{P}$ and $\mathbf{I}$-$\mathbf{P}$ to obtain orthonormal bases, such as $\mathbf{K}_{n\times p} and L_{n\times(n-p)}$, for C($\mathbf{X}$) and C($\mathbf{X}$)$^\perp$. They argue that Equation 1 can be rewritten as $g(\mu_k) = x_k^T\beta + k_i'\gamma + l_i'\delta$ where $\gamma_{p\times 1}$ and $\delta_{(n-p)\times 1}$ are random coefficients and K is the source of the confounding. Therefore, Reich et al. proceed by recommending removing K from the model, as it's columns have no scientific meaning. The resulting model, $g(\mu_k) = x_k^T\beta + l_i'\delta$, is the RHZ model.

The Sparse SGLMM model provides a method to reduce the computational complexity of the RHZ model. Hughes implement this model by reducing the number of random effects through the use of Moran operator. We will compare this model to the results of the BYM and Leroux model as a check of spatial confounding.

4 METHODS

4.1 Neighborhood Matrix, W

The spatial structure of the block groups is incorporated into these models through the neighborhood matrix, $\mathbf{W}$, whose elements are the w_{ki} in equations 2 and 3. Therefore, we will spend some time going over the properties of these matrices and how they are constructed.

The neighborhood matrix, $\mathbf{W}$, is a non-negative, symmetric, K×K matrix, where K is the number of areal units. We know that the (k,j)th element of the neighborhood matrix w_{kj} represents spatial closeness between areas $(\mathcal{S}_k, \mathcal{S}_j)$. These elements are positive values denoting geographical closeness and zero values denoting non-closeness (0-1 is the most common structure). The diagonal elements of this matrix, w_{kk}, are 0. The $\mathbf{W}$ matrix forces (ϕ_k, ϕ_j) relating to geographical adjacent areas to be autocorrelated, whereas random effects relating to non-contiguous areal units are conditionally independent given values of remaining random effects [7].

For our purposes, we will use the 0-1 binary structure, where 0 indicates that two areal units are not neighbors and 1 indicates that they are neighbors. In future work, we would like to consider other options, such as including options that are 2 areal units apart, or further. However, the R package *ngspatial* only allows for binary elements in the neighborhood matrix.

We create a neighborhood matrix for both the geographic and social proximity. For geographic proximity, we define a neighbor as any block group that shares a border. For social proximity, we define a neighbor as any block group where there is a certain level of commuting between two block groups. We recognize that commuting is technically a directed activity but we treat it as undirected. We believe that regardless if it is the origin or the destination, it is still establishing a tie between the two block groups, as people will return home after work. Therefore, this creates a symmetric neighborhood matrix. For the purposes of our model, we define a cutoff for the number of commuters in order to establish meaningful social proximity ties between two block groups. In our case, we define the cutoff to be 15 commuters between two block groups in order for a tie to be defined. This includes about 0.89% of the commuters.

In Figure 1, we see the neighborhood structures depicted on the map of the block groups of Detroit. We are able to visualize how we have defined the geographic proximity neighborhood structure quite easily - links are created when block groups share a boundary. In the social proximity case, the edges are less clear as there are many more ties. However, we can see that some folks are commuting across the county from the west to/from the south but the majority of the population are commuting to/from the middle east side of the city. There are clearly social hubs in some areas of the county where many people commute to/from.

4.2 Combining Social and Geographic Proximity

In the existing literature, it is a relatively new technique to combine social proximity and geographic proximity. Wang et al. study crime rates for Chicago, also using areal units. The social proximity data that they use is taxi data, where they establish links between where people get in and out of a taxi. They also have Point of Interest (POI) data that we do not incorporate in this work. The models they use are linear regression, Poisson GLM, and Negative Binomial GLM. They use the structure in equation 4 for the relationship between block groups and crime rate.

$$\vec{y} = \vec{\alpha}^T \vec{x} + \beta^f W^f \vec{y} + \beta^g W^g \vec{y} + \vec{\epsilon} \tag{4}$$

In equation 4, $\vec{x}$ represents nodal features, including demographics and POI distribution, W^f is the taxi flow (or social proximity), and W^g is the spatial matrix which represents geographic adjacency. The authors discuss their "leave-one-out" evaluation where they estimate the crime rate of one geographic region given all of the information of all the other regions. This is an example of spatially weighted regression, where the crime rates closest to given areal unit have a larger effect on the predicted crime rate of the region that is left out.

In our methods, we also include social and geographic proximity in the same model but we attempt to do so through the Bayesian CAR model structure. We believe that this Bayesian CAR model may be a more effective way to model crime over areal units.

Therefore, in order to use the CAR structure, we must determine how to combine the social and geographic neighborhood matrices. There are many ways that these matrices could be combined. This could be done using a weighted average of the two matrices or adding the two matrices. Another approach could be to create links in the geographic neighborhood matrix, where there are no links already present but there are links in the social proximity neighborhood matrix. In other words, if $w_{ij}^{geog} = 0$ but $w_{ij}^{social} = 1$, then replace w_{ij}^{geog} with 1. The resulting matrix will include all ties that are both geographic and social, and still be only include the indicator values 0 and 1. In general, we are also interested in considering different forms of W for geographic proximity as well, such as including second and third neighbors in the structure of the neighborhood matrix, **W**.

For this study, we consider the simple binary matrix. If there was not already a geographic tie between two communities but there is a strong social tie between the two communities (more than 15, than we replace the "0" element in the matrix with "1" and consider them neighbors. In Table 1 we have included a brief summary of the neighborhood matrices below.

5 DATA

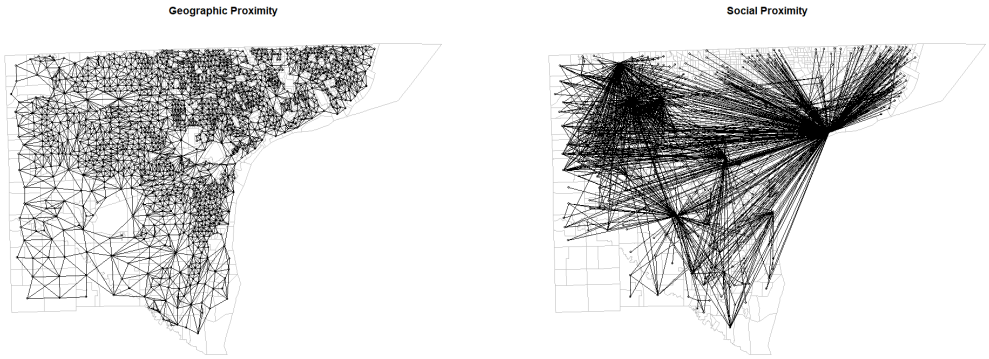
One advantage of this analysis is that we use all publicly available data for our models. We use crime data from the Police Data Initiative as well as commuting and demographic data from the Census. We describe them in some detail below.

5.1 Crime Data: Police Data Initiative

First, we acquire the crime data through the Police Data Initiative. This is a recently popularized data source created by the Police Foundation to support academic research. This project encourages folks in the community to use their data in order to create more effective relationships between law enforcement and local citizens. There are several different kinds of datasets hosted by the Police Data Initiative, including data on accidents/crashes, community engagement, officer-involved shootings, and complaints. We will focus our efforts on the Calls for Service (CFS) category. Each row in these datasets represent an individual call to 911. It is easiest to think of these "calls" as a 911 call, but they can be either officer or call-initiated. For example, traffic stops are almost always officer initiated.

Currently, there are 29 cities that are referenced on the Police Data Initiative's website in the CFS category. These include larger cities like Baltimore, Maryland as well as smaller college towns like Bloomington, Indiana. There is large variation in this data between cities. Some cities include exact locations (lat/long) of the individual crimes while others just include general neighborhoods. Some give no other detail other than the location, while some have information on crime type and response time. We decide to focus our efforts on Detroit, Michigan because crime is known to be a problem in Wayne County, so there is potential for effective interventions. On NeighborhoodScout.com, they list Detroit's "Crime Index" as a 2, when 100 is the safest. They say that Detroit is safer than just 2% of cities [1]. They also have a particularly rich datasets, with over 400,000 crimes from September, 2016 to November, 2017.

There are many crime types available in Wayne County's crime data. In our collaboration with the Arlington, Virginia Police Department, we were pointed to three types of crime that are of particular interest to them: mental health-related, drug-related, and domestic/sexual violence. For our analysis, we chose to use the crime type



(a) Geographic Proximity **(b) Social Proximity**

Figure 1: Plot of Neighborhood Structure

Variable	Geographic	Social	Geographic and Social
Number of regions	1706	600	1706
Number of nonzero links	10,378	2,490	12,634
Percentage nonzero weights	0.36	0.69	0.43
Average # of weights	6.08	4.15	7.41

Table 1: Descriptive Statistics of Neighborhood Structures

of domestic/sexual violence. We used call description codes such as "RAPE IP or JH" or "ASSAULT OR SEX ASSAULT DELTA", where IP means In Progress, JH means Just Happened, and Delta indicates a high severity crime (on the range alpha to echo).

5.2 Demographics: ACS

Next, we collect demographic information for all of the block groups in Wayne County, Michigan through the American Communities Survey (ACS). This is a national survey that is administered every few years through the US Census Bureau and has variables ranging from income information to migration information. The block groups in Wayne County can be seen in Figure 1. There are 1,822 block groups in total for Wayne County, of which only 1,706 have complete publicly available ACS data. The data of the remaining block groups are not released due to concerns of identifiability, a common practice with ACS. There were not high crime counts in these block groups, and it represents a small portion of the data, so this is not a large concern, though this is something we will address in future work. We use data from the 2015 ACS survey in order for our data to be representative of the demographics present with the crime data.

The demographic variables that we consider through ACS are as follows: median income, median age, gender (female) percentage, unemployment rate, total population, and race (white, black/African American, American Indian/Alaska Native, Asian, Native Hawaiian and other Pacific Islander, two or more races, and some other race).

We used the set of variables on race to calculate the Herfindahl Index (HI) [9]. This is a measure of concentration that is often used in demography and sociology studies, but has many other applications. For example, it is often used in finance or economics to show the concentration or diversity of a given sector of the economy. For our purposes, we use HI, shown in equation 5, as a measure of diversity in the block group. This is another effort, in addition to the incorporation of social proximity, to attempt to move away from racial profiling in predictive policing by using a measure of diversity rather than variables associated with each individual race. We see that a HI of 0 means that there is no diversity in the block group, or that the block group is completely made up of one racial category. A HI approaching 1 means more diversity in the block group. For example, if the block group's population is made up of entirely white people, then $HI = 1-(1^2+0+0+\ldots) = 0$.

$$\text{HI} = 1 - \left(\left(\frac{\#\text{ white}}{\text{total pop}} \right)^2 + \left(\frac{\#\text{ black}}{\text{total pop}} \right)^2 + \left(\frac{\#\text{ Asian}}{\text{total pop}} \right)^2 + \ldots \right) \quad (5)$$

5.3 Social Proximity: LODES

Finally, we have collected data on commuting to establish social proximity through another Census data source, LODES. The acronym LODES stands for LEHD Origin-Destination Employment Statistics. LEHD stands for Longitudinal Employer-Household Dynamics. This data is the source for the Census app OnTheMap which shows how many people are commuting into and out of a given geographical area, and how many stay in the area for work. In this dataset that we have created from the LODES, we have the complete set of block groups for Wayne County, as well as all of Michigan. However, for the sake of our analysis, we do not include commuters

that are leaving Wayne County for work. In other words, we treat Wayne County as a closed system. We will address possibilities for changing this in future work.

6 RESULTS

6.1 Model Comparison Metrics

Before we proceed to the results, we will explain how we compare our models and our methods of creating proximity. We use the following criteria for model comparison: the Deviance Information criterion (DIC), and the Percentage Deviance Explained.

We will briefly explain the Deviance Information Criteria. The Deviance Information Criteria **(DIC)** is a measure that combines the "goodness of fit" of a model and the "complexity" of such a model [10]. We measure the **fit** via the deviance, where $D(\theta) = -2logL(data|\theta)$. We measure **complexity** by the estimate for effective # of parameters, or $p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta}) =$ posterior mean deviance - deviance evaluated at the posterior mean of the parameters. So, the DIC is defined as in equation 6.

$$\text{DIC} = D(\bar{\theta}) + 2p_D = \bar{D} + p_D \tag{6}$$

This DIC is commonly used in a Bayesian framework. A model with a smaller DIC is better supported by the data than a model with a larger DIC, just like with AIC that is commonly used in model comparison.

6.2 Spatial Autocorrelation Assessment

Before fitting our models, we want to make sure that it is appropriate to be creating spatial models for this data, based on our two neighborhood matrices. We do this using Moran's I, a common method for assessing spatial processes, which is calculated by the formula seen in equation 7.

$$\text{I} = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{ij} w_{i \neq j})(Y_i - \bar{Y})^2} \tag{7}$$

Moran's I is large when spatial association is large. We know that the statistic seen in equation 8 is asymptotically normal but convergence is very slow, we use permutation tests.

$$\frac{I + 1/(n-1)}{\sqrt{Var(I)}} \xrightarrow{\mathcal{D}} N(0,1) \tag{8}$$

We calculated Moran's I as well as the p-value for this test and summarize our results in Table 2. For Moran's I, the null hypothesis states that the spatial processes promoting the observed pattern of values is random chance or it is randomly distributed among the features in your study area [4]. Therefore, for spatial modeling to be appropriate, we would like the p-value to be less than an α of 0.05 so that we can reject this null hypothesis. We see that spatial modeling is appropriate using the neighborhood matrices for both geographic and combined proximity models because the p-values are less than 0.05.

6.3 Model Comparison

In this section we will compare the six models using the model comparison criteria outlined in section 6.1. We will compare the six models the BYM model, Leroux Model and Sparse SGLMM model, all for both geographic and combined (geographic and social) proximity.

Proximity Type	Moran's I	p-value
Geographic	0.22656	0.000999
Geographic and Social	0.25188	0.000999

Table 2: Moran's I and p-value

In Table 3 we see the summary of our results for our six models, through DIC and Percentage Deviance Explained. We see that when we combine social and geographic proximity and compare it to the results of just the geographic proximity, there is an improvement in every model in terms of percentage deviance explained and DIC. All of the models that incorporate both geographic proximity and social proximity have higher percentage deviance explained and lower DIC, indicating a better model fit. Therefore, we conclude that adding meaningful commuting ties between census blocks improves the notion of ties between neighborhoods.

We would also like to compare model performance between model types. We see that in the case of just geographic proximity, we see that the Leroux model outperforms both the BYM model and the sparse SGLMM model. For the sparse SGLMM model in the geographic case, there is actually a quite low percentage deviance explained. ???? In the case where we combine social and geographic proximity, we see that the Leroux model outperforms the BYM and sparse SGLMM models in terms of DIC and percentage deviance explained.

6.4 Fitted Values

Next, we compare the models to see the fitted values compared to the actual observed counts per block group.

First, in Figure 2 we see that there is a large peak in crimes in the northeast part of the city. Most of the city is pretty evenly distributed other than this peak, and many block groups, ????? out of 1822, have 0 domestic/sexual violence crimes reported during this time.

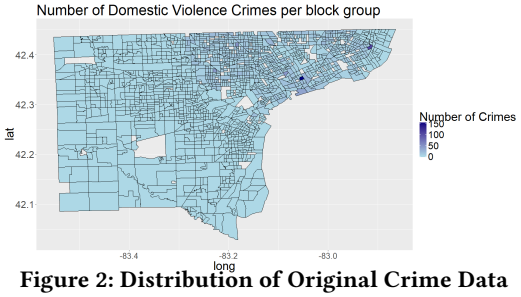


Figure 2: Distribution of Original Crime Data

In Figure 3 we see the fitted values for all six models. It is a bit difficult to see the exact distribution of the fitted values for the models, but we can see from the scale on the axis alone that the ??? model that is picking up the peak visible in Figure 2 ???. Therefore, we conclude that the ??? is the best fit for the data, with these six options.

Criterion	Geog BYM	Combined BYM	Geog Leroux	Combined Leroux	Geographic Sparse SGLMM	Combined Sparse SGLMM
DIC	8923.91823	8845.05473	8029.71980	7005.55019	8522.328	8251.777
% deviance explained	56.92746	61.01017	63.90752	68.11767	38.95984	58.63444

Table 3: Model Comparison

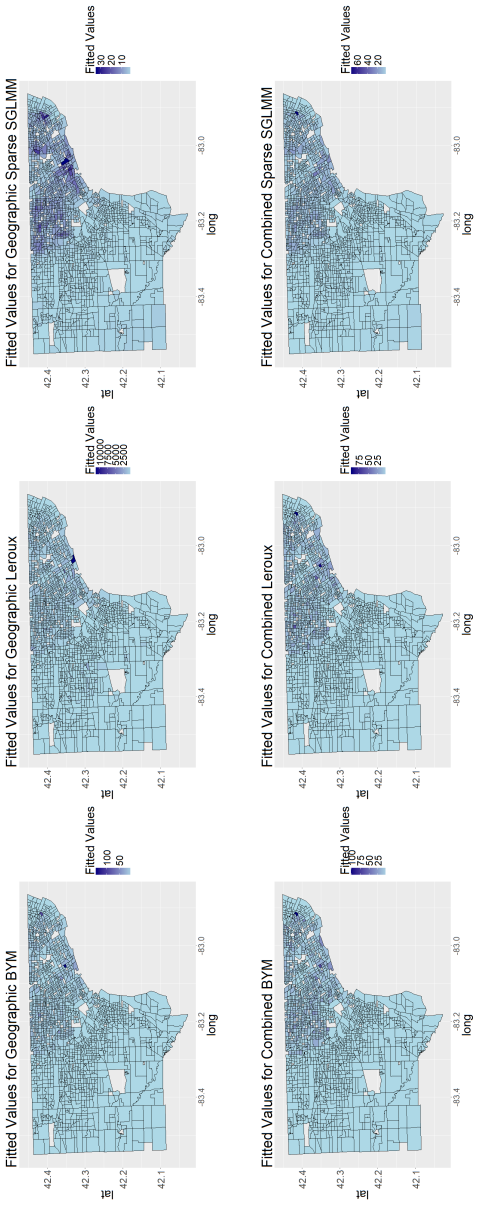


Figure 3: Fitted Values, All Models

6.5 Coefficients

In addition to the accuracy of the estimation of crime and the fit of the model, we are also interested in the estimates of the coefficients of the covariates we have used in our model, or the demographic information. Specifically, we are interested to see if there are any problems with spatial confounding in our case. To assess this, we will see if the estimates for the coefficients and their confidence intervals are similar between the BYM/Leroux Model and the Sparse SGLMM. We see that the combined social and geographic proximity seems to be a better fit than using only the geographic proximity. We also see that the Leroux model seems to out perform the BYM model in our data. Therefore, we will focus on the comparison between the Leroux model and the Sparse SGLMM, the latter of which controls for spatial confounding.

We have included the estimated values of the posterior median as well as the 95% credible intervals for the model parameters in Table 4 for the Combined Leroux Model and in Table 5 for the Combined Sparse SGLMM model. Our covariates are all of the demographic variables that we collected from Census through the American Communities Survey for Wayne County.

	median	credible interval	
Term	0.5	0.025	0.975
Intercept			
Median Income			
Unemployment Rate			
Total Population			
Percentage Male			
Median Age			
Herfindahl Index			

Table 4: Combined Leroux

	median	credible interval	
	0.5	0.025	0.975
Intercept			
Median Income			
Unemployment Rate			
Total Population			
Percentage Male			
Median Age			
Herfindahl Index			

Table 5: Combined Sparse SGLMM

Based on Tables 4 and 5, we see that ????.

7 DISCUSSION/FUTURE WORK

There are many possibilities for future work with this data and analysis.

One of the immediate next steps with this analysis is to incorporate into our modeling framework that Detroit is not a closed system, especially socially. At some point, we have to set a geographic boundary where we use the area as a closed system. However, for this analysis

Also, we would like to conduct the above analysis on other subsets of the crime data, for different crime types, such as violent crime, drug crimes, or crimes related to mental health. We hypothesize that some crime types, such as drug crimes, will be affected more by social and geographic dynamics than other crime types, such as property-related crimes.

We would also like to make additional adjustments to our neighborhood matrices. First, we would like to try different methods of combining the matrices. One of the first methods we are considering is a weighted average of the two matrices. We will also consider simply normalizing the matrix entries so that block groups that are closer receive a higher weight in the model.

As discussed in our data collection process, we had missing ACS data for some of our block groups. This was not a huge concern in our modeling process because it was a small number of block groups affected by this problem. However, in the future we would like to consider perhaps conducting this analysis for census tracts, rather than block groups, to avoid this problem.

Lastly, we would like to conduct our analysis on different cities that provide their data to the Police Data Initiative. If we can show that this sort of analysis provides a good fit for more than just Detroit, and there are consistent results across cities, than this analysis technique has the potential to be more impactful in terms of policy implications.

REFERENCES

[1] [n. d.]. Detroit,. ([n. d.]). https://www.neighborhoodscout.com/mi/detroit/crime

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

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

[4] Arthur Getis and J Keith Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical analysis* 24, 3 (1992), 189–206.

[5] John Hughes. 2014. ngspatial: A Package for Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data. *The R Journal* 6, 2 (2014), 81–95. https://journal.r-project.org/archive/2014/RJ-2014-026/index.html

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

[7] Duncan Lee. 2013. CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors. *Journal of Statistical Software* 55, 13 (2013), 1–24. http://www.jstatsoft.org/v55/i13/

[8] Brian J Reich, James S Hodges, and Vesna Zadnik. 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* 62, 4 (2006), 1197–1206.

[9] Stephen A Rhoades. 1993. The herfindahl-hirschman index. *Fed. Res. Bull.* 79 (1993), 188.

[10] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, 4 (2002), 583–639.

[11] Edward M Spiers. 2000. Weapons of mass destruction. In *Weapons of Mass Destruction*. Springer, 1–18.

[12] Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. 2016. Crime rate inference with big data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 635–644.