Fall 2017 Research Summary

Claire Kelling

December 13, 2017

Abstract

In this project, we model crime data using two different kinds of spatial modeling techniques: areal and point-referenced. The focus 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 create areal models of crime counts using two different versions of a Bayesian conditional autoregressive (CAR) model. 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. We also will give our first attempt at combining the social and geographic proximity information into one model, using additive and binary approaches. We compare these models with the models that only include social or geographic proximity. In the second part, we study point-referenced data, where each individual crime data point is also affiliated with a response time, which we seek to model. We create two different models for this part: one model where we recreate a model in the homework for STAT 597 and the other with the spBayes R package.

Contents

1 Introduction

In the Summer of 2017, I worked as a Graduate Research Fellow in the Social and Decision Analytics Lab in Arlington, Virginia. I worked closely with Dr. Josh Goldstein, a recent Statistics PhD graduate from Penn State, and Dr. Gizem Korkmaz, a research faculty member in the lab. The three of us, with other students and faculty, partnered with the local Arlington Police Department to model crime data to see if there was any spatial correlation in crime, and to see if certain events affected crime rates, both in time and space. This opportunity served as my introduction to this work, and I will continue collaborating with the lab throughout the next year.

Arlington's police data is governed by several data usage agreements. Therefore, the main source of data for this project will be the Police Data Initiative, which houses publicly accessible crime data for many large cities across the United States. The structure of this data is very similar to Arlington's data, where each crime is recorded with a lat/long coordinate and a Response Time, as well as some other covariate information. We will first focus on Detroit, but will plan to complete our analysis in future work for more than one major city. The focus of this analysis will be quite different than the project this summer, as we are aggregating by spatial units rather than modeling point-referenced data. We are also doing a much more complex Bayesian GLMM approach, rather than the basic CAR model considered this summer. Also, we are interested in response times to crime rather than fires for the second part of the project.

2 Social and Spatial Proximity

2.1 Motivation

The motivation of our work is mainly being drawn from Dr. Corina Graif's recent work in criminology [1]. She is interested in modeling the diffusion of crime between neighborhoods. In the current literature, neighborhoods are often treated as closed systems, even though we know this is not the case. 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" [2], where certain neighborhoods are predominantly targeted, mainly based on race and other demographic information.

We are interested in investigating what other ties might exist between neighborhoods, in order to create more effective policing interventions. Social proximity is a relatively new topic in the criminology literature, and Dr. Graif has experience with commuting data that we will use to establish social ties between communities. We compare social and spatial proximity in our neighborhood matrices, using two different models, to see which models may best characterize crime in block groups. We also include two attempts to combine the two matrices. We will also consider demographics in our modeling efforts.

2.2 Data

2.2.1 Crime Data: Police Data Initiative

We consider many data sources for our analysis. 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 research just like this project. They are trying to encourage people 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. 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. They also have a particularly rich datasets, with over 400,000 crimes from September, 2016 to November, 2017. Also, some of the other major cities that are present in the database have been studied exhaustively in previous literature. Something interesting to note is that Chicago's crime data is publicly available and is studied quite a bit in the crime literature, either because of or resulting in the national media attention, but the data is not listed through the Police Data Initiative. In this project, we will focus on the aggregated data, by block group, so the data structure is not too interesting other than what I have described above.

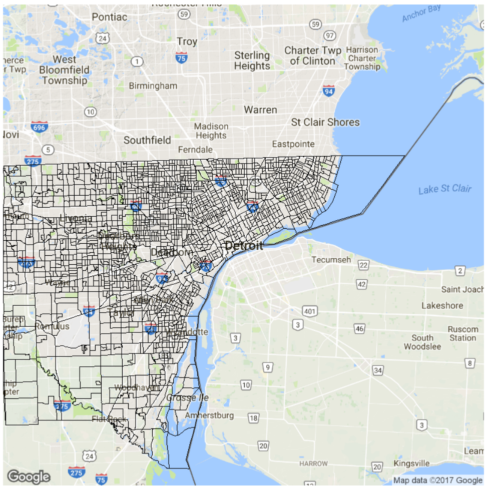


Figure 1: Wayne County (Detroit) Block Groups

2.2.2 Demographics: ACS

Next, we collect demographic variables 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 on 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 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 we don't think this represents 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.

Through our collaborations in sociology and demography, we have narrowed down the vast set of potential demographic variables to the following variables:

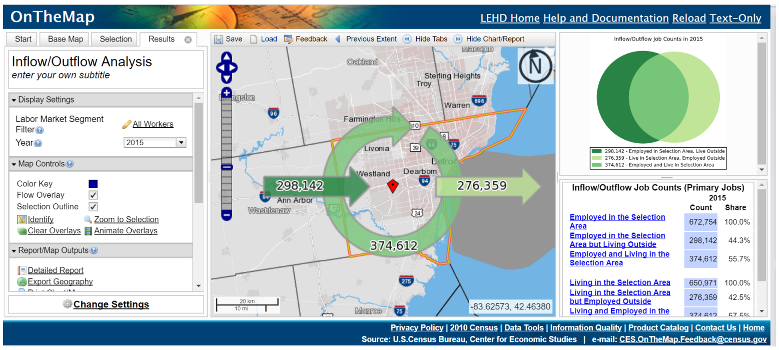
- median income
- median age
- percentage female
- unemployment rate
- total population
- race counts (white, black/African American, American Indian/Alaska Native, Asian, Native Hawaiian and other Pacific Islander, two or more races, and some other race)

We used the last set of race variables to calculate the Herfindahl Index (HI) [3]. 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 1, 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 $HI = 1 - (1^2 + 0 + 0 + ...) = 0$.

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

2.2.3 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. It uses data from 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. The structure of this Census app can be seen in Figure 2. 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. The details of the social ties will be given in the methods section.


OnTheMap
LEHD Home Help and Documentation Reload Text-Only
Start Base Map Selection Results
Save Load Feedback Previous Extent Hide Tabs Hide Chart/Report
Display Settings
Map Controls
Report/Map Outputs
Identify Zoom to Selection
Clear Overlays Animate Overlays
Detailed Report
Export Geography
Change Settings
Oakland
Macomb
Livingston
Sterling Heights
Troy
Warren
Farmington Hills
Livonia
Westland
Dearborn
Detroit
Ann Arbor
Washtenaw
Monroe
298,142
276,359
374,612
20 km
10 mi
-83.62573, 42.46380
Privacy Policy | 2010 Census | Data Tools | Information Quality | Product Catalog | Contact Us | Home
Source: U.S.Census Bureau, Center for Economic Studies | e-mail: CES.OnTheMap.Feedback@census.gov

Inflow/Outflow Analysis
enter your own subtitle

Labor Market Segment Filter: All Workers
Year: 2015
Color Key: ■
Flow Overlay: ☑
Selection Outline: ☑

Inflow/Outflow Job Counts in 2015

- 298,142 - Employed in Selection Area, Live Outside
- 276,359 - Live in Selection Area, Employed Outside
- 374,612 - Employed and Live in Selection Area

Inflow/Outflow Job Counts (Primary Jobs)

	2015 Count	Share
Employed in the Selection Area	672,754	100.0%
Employed in the Selection Area but Living Outside	298,142	44.3%
Employed and Living in the Selection Area	374,612	55.7%
Living in the Selection Area	650,971	100.0%
Living in the Selection Area but Employed Outside	276,359	42.5%
Living and Employed in the	374,612	57.5%

Figure 2: OnTheMap Census App

In summary, as seen in Table 1, we have three main data sources: ACS, LODES, and the Police Data Initiative. We will also note that geographic proximity is established through adjacency (whether block groups share a border or not) and this is done using shape-files that were also acquired through Census, and are depicted in Figure 1.

1. Crime Data (Police Data Initiative)
2. Demographics (American Communities Survey, from Census)
3. Geographic proximity (established through adjacency, shapefiles from ACS)
4. Social proximity (LODES Data, from Census)

Description	Data Source
Crime Data	Police Data Initiative
Demographics	American Communities Survey, Census
Geographic proximity	established through adjacency, shapefiles from Census
Social proximity	LODES, Census

Table 1: Data Descriptions and Sources

2.3 Methods

In this section, we rely on the Conditional Autoregressive model (CAR) model for analyzing our areal data, an extremely popular method in sociology, political science, epidemiology and many more applications. We use the CARBayes package in R [4], and the descriptions below will serve as a literature review of the vignette which describes the use of this package. We will consider two models that are presented in the vignette and compare their results for social and geographic proximity, and the two proximities combined. Both of these models have a common base structure, described below.

In this model, 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. In both CAR models discussed below, spatial variation in the response is modeled by a matrix of covariates $\mathbf{X} = (x_1, ..., x_k)$ and a spatial structure component $\psi = (\psi_1, ..., \psi_k)$. We see that $\psi = (\psi_1, ..., \psi_k)$ models any spatial autocorrelation that remains after covariate effects have been accounted for. The paper 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 2, with the parameters described above.

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

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 $ln(\mu_k) = \mathbf{x_k^T}\beta + \boldsymbol{\psi_k}$.

2.3.1 BYM Model

The first model we consider for the ψ_k parameter for spatial auto-correlation is called the BYM model, named for the initials of the three authors on the paper (Besag, York, and Mollié) [5]. 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 3. This model is equivalent to the multivariate specification $\phi \sim N(0, \tau^2\mathbf{Q}(\mathbf{W})^{-1})$, where $\mathbf{Q}(\mathbf{W}) = \text{diag}(\mathbf{W1}) - \mathbf{W}$.

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

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.

2.3.2 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 [6]. The Bayesian model specification can be found in equation 4. 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 $\phi \sim N(0, \tau^2\mathbf{Q}(\mathbf{W}, \rho)^{-1})$, 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 [7].

$$\begin{aligned} \psi_k &= \phi_k \\ \phi_k|\phi_{-k}, \mathbf{W}, \tau^2, \rho &\sim N\Big(\frac{\rho\sum_{i=1}^K w_{ki}\phi_i}{\rho\sum_{i=1}^K w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{i=1}^K w_{ki} + 1 - \rho}\Big) \\ \tau^2 &\sim \text{Inverse-Gamma}(a, b) \\ \rho &\sim \text{Uniform}(0, 1) \end{aligned} \tag{4}$$

Both of these models heavily rely on the neighborhood matrix, **W**, whose elements are the w_{ki} used in the conditional distribution of ϕ_k in both models. Therefore, we will give a bit more detail about the formation of our **W** matrix in the following section.

2.3.3 Neighborhood Matrix, W

The main spatial element of these two models is our neighborhood matrix, **W**, whose elements are the w_{ki} in equations 3 and 4. Therefore, we will spend some time going over the properties of these matrices and how they are constructed.

The neighborhood matrix, **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. According to the vignette narrative, the **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.

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. We recognize that the structure of **W** does influence our model, but we have decided to use the simple structure for our first modeling attempt.

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 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. Therefore, this creates a symmetric neighborhood matrix.

In Figure 3, 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.

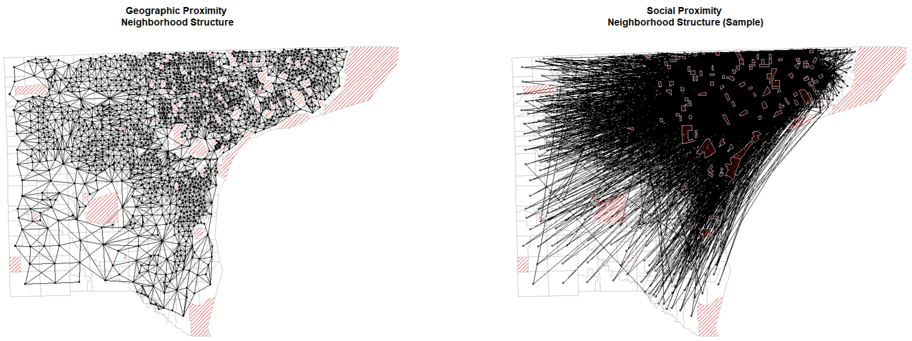


Figure 3: Depiction of W, the neighborhood matrix, based on geographic or social proximity

In Table 2 we have included a brief summary of the neighborhood matrices below, calculated in R. We see that neighborhood structure is quite different between geographic and social proximity. One of the most striking differences is found in the average number of links, which is 189 for social proximity and 6 for geographic proximity. We have also included the summary for the combined matrices. Notice that these summary statistics only include ”nonzero” metrics, so the summary is the same for both of the additive and binary combined models. We notice that the models that combine social and geographic proximity only have slightly more links than the social neighborhood matrix, as expected, because many folks probably commute to their neighboring block groups. We hypothesize that these striking differences may have an impact on the modeling process.

Variable	Geographic	Social	Combined Models
Number of regions	1706	1706	1706
Number of nonzero links	10,378	322,720	327,620
Percentage nonzero weights	0.36	11.09	11.26
Average # of weights	6	189	192

Table 2: Data Descriptions and Sources

2.3.4 Combining Social and Geographic Proximity

In the existing literature, it is a new technique to combine social proximity and geographic proximity. The paper *Crime Rate Inference with Big Data* is one of the first to tackle this challenge [1]. In this paper, they 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. They models they use are linear regression, Poisson GLM, and Negative Binomial GLM. They use the structure in equation 5 for the relationship between block groups and crime rate.

$$\vec{y} = \vec{\alpha}^T \vec{x} + \beta^f W^f \vec{y} + \beta^g W^g \vec{y} + \vec{\epsilon} \tag{5}$$

In equation 5, $\vec{x}$ 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. In section 6 of this paper, the authors discuss their ”leave-one-out” evaluation. In other words, they estimate the crime rate of one geographic region given all of the information of all the other regions. We have been unable to replicate the exact model but it was explained to us by the author as a 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 is an effective way to model crime over areal units. So, our next step in this process is to consider different methods to perhaps combine the two proximity matrices to create a new model that relies on a new neighborhood matrix.

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**. However, for the purposes of this project, we will keep the structure of our separate proximity matrices and try including them together in a model.

In our results section, we will include two simple techniques at combining our matrices, both of which were discussed above.

1. **Additive:** First, we will consider the simple sum of the two proximity. This will result in three possibilities for entries w_{ki} in the neighborhood matrix: 0,1,2. The entry 0 occurs if it is not a neighbor geographically or socially, 1 occurs if it is either but not both, and 2 occurs if it is both.
2. **Binary:** This will result in the same form as the additive neighborhood matrix above but keep the binary property, where the elements are either 0 if there is no geographic or social tie, and 1 if there is a geographic or social tie.

2.3.5 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:

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

We will briefly explain the Deviance Information Criteria, as this is the major way to compare these models, in addition to percentage deviance explained, and is not as intuitive. The Deviance Information Criteria **(DIC)** is a measure that combines the "goodness of fit" of a model and the "complexity" of such a model [8]. 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.

2.4 Results

2.4.1 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. This is calculated by the form 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, so as we discussed in lecture, we will use the code in R to 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 3. 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 [9]. 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 spatial proximity because the p-values are less than 0.05. We also notice that the p-value for the binary combined model is also less than 0.05. However, we unfortunately we cannot compute Moran's I for the additive combined model because this test expects a 0-1 structure for the neighborhood matrix, $\mathbf{W}$. We expect that the Moran's I statistic for the additive model will be close to the binary model, and the binary model's p-value is much less than 0.05. Therefore, we proceed with our analyses.

Proximity Type	Moran's I	p-value
Geographic	0.55932	0.000999
Social	0.08496	0.000662
Binary	0.08665	0.000995

Table 3: Moran's I and p-value

2.4.2 Model Results

2.4.3 Model Comparison

In this section we will compare the four models using the model comparison criteria outlined in section 2.3.4. The four models include both BYM models, with social and geographic proximity, and both Leroux models, with social and geographic proximity. For the scope of this project, we will consider social and geographic proximity separately as a means of comparison. However, in the future we would like to combine the two types of proximity measures.

We would like to note that we did assess the mixing of the MCMC procedure for our parameters. There are a large number of parameters for these models so we have not included the ACF and trace plots here but there was adequate mixing in these models.

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

Table 4: Model Comparison

In Table 4 we see the summary of our results for our four models. It is interesting to note that the geographic proximity has a lower DIC, and therefore better fit, than the social proximity for both the BYM and Leroux Model. We see that for social proximity, the BYM model outperforms the Leroux model both in terms of DIC and in terms of percentage deviance explained, while the Leroux model is better than the BYM model for the geographic proximity models. We see that the model with geographic proximity and the Leroux model has by far the lowest DIC and the highest percentage deviance explain, at 60%.

Next, we include our results for our two combined matrices, for both the BYM and Leroux Models in Table 5. We notice that the fit for all of these models is better than the Social

Leroux Model, and most are better than the Social BYM model. Therefore, we see that this is an improvement to the social proximity alone. However, we also notice that the DIC's are all drastically higher than the DIC of approximately 410,000 for the model that incorporates geographic proximity alone. This is similar in the percentage deviance explained, where none of these combined models come close to the percentage explained by the Geographic Leroux model, which had 60% deviance explained. We discuss possibilities for future work in the discussion section, but generally speaking we believe that the social proximity matrix still has potential but not in its current form.

Criterion	Additive BYM	Binary BYM	Additive Leroux	Binary Leroux
DIC	650,200	678,316	650,411	643,091
p.d	1,909	1,775	601	556
% deviance explained	37	35	37	38

Table 5: Model Comparison

2.4.4 Fitted Values

Next, we compare the models to see the fitted values compared to the actual observed counts per block group. First, in Figure 4 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. We will discuss concerns with this distribution of crimes in the next section as future work.

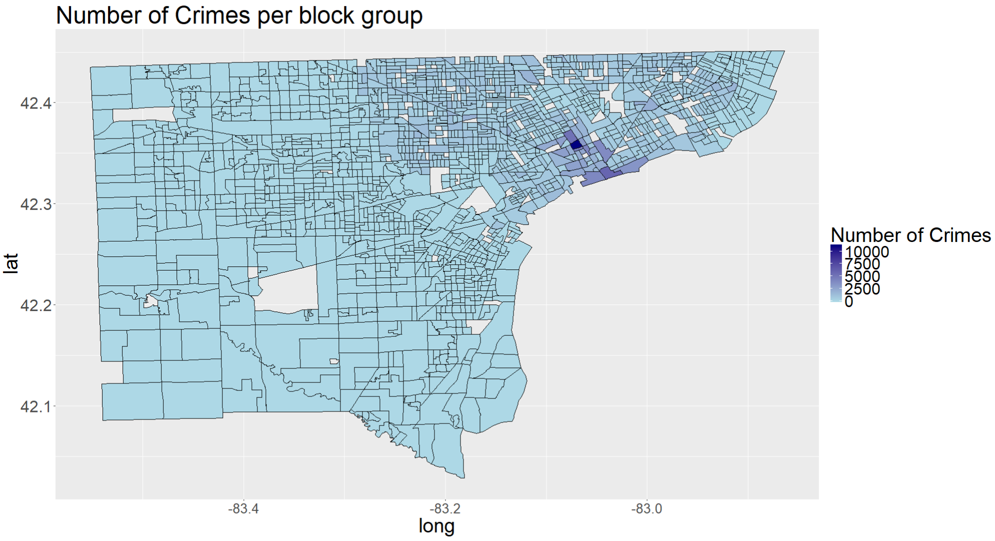


Figure 4: Distribution of Original Crime Data

In Figure 5 we see the fitted values for all four 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 only model that is picking up the peak visible in Figure 4 is the Leroux model with geographic proximity, which was said earlier to be the best fit for this data. Some of the other models have higher levels of crime in this general area, but they do not come as close to fitting the large value (around 10,000) for this peaked area. Therefore, we conclude that the Leroux model with geographic proximity is the best fit for the data, with these four options.

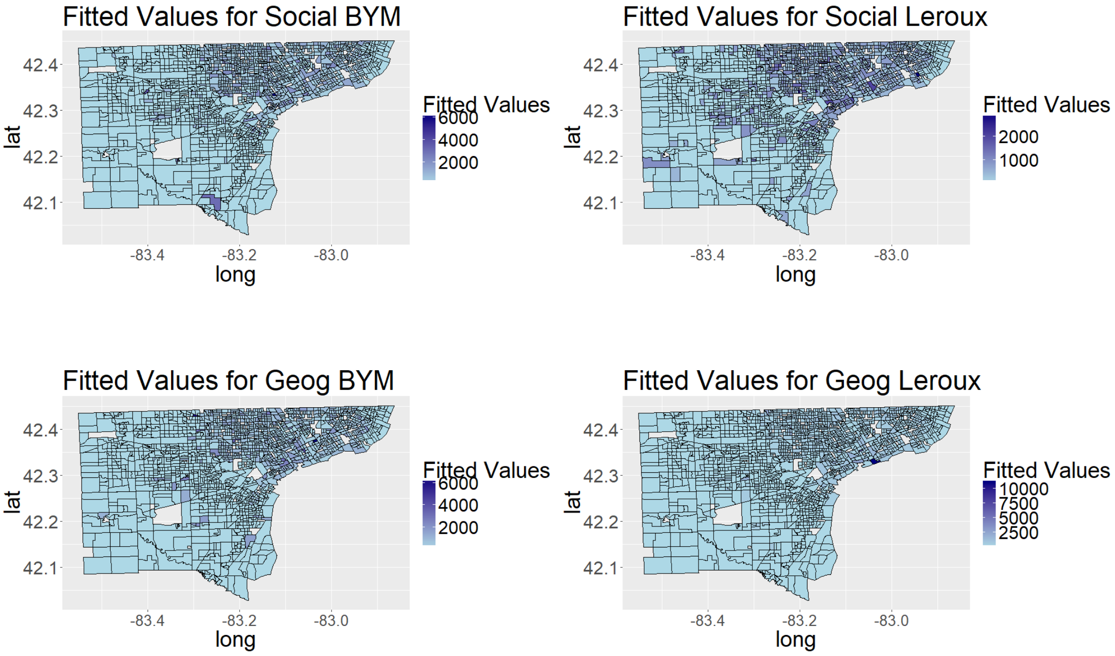


Figure 5: Fitted Values, Geographic and Social Proximity

Next, in Figure 6 we examine the fitted values for the combined models. Once again, as with 3 out of the four models in Figure 5, we notice that these models are not picking up on the high values in the Northeast of the city. We notice that once again the geographic proximity Leroux model seems to be the best fit for this data, which confirms the model comparison analysis.

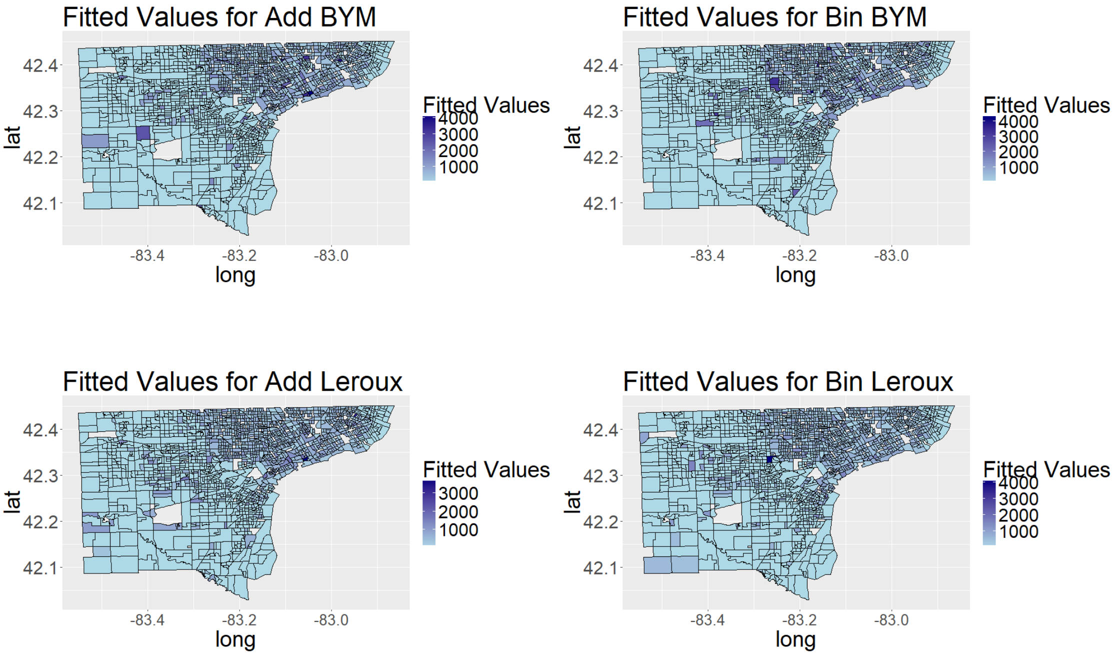


Figure 6: Fitted Values, Additive and Binary Combined Models

We have included the values posterior median as well as the 95% credible intervals for the model parameters in the appendix, section 4.1, for each model incorporating only geographic or social proximity and in section 8.2 for the models that combine the two matrices. In these tables, we see that there are not very many covariates that have consistently non-zero estimates and credible intervals. In the Leroux Geographic model, which we have argued to provide the best fit, the credible intervals do not include 0 for the covariates of Herfindahl Index and percentage male. These are also generally non-zero in their credible intervals for other models.

2.5 Discussion/Future Work

There are many possibilities for future work with this project which will be summarized in this section, though some have been referenced in the above sections. One of the next steps that we would like to consider is we would like to conduct the above analysis on a subset of the crime data, for different crime types, like violent crime, domestic violence, 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.

Another immediate next step is that we would like to make many 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 thinning the social proximity model. This is extremely important because as of now the matrix includes all commuters in Wayne County. We will consider perhaps setting a cutoff where we will only define a social link between two communities if there is a large number of people commuting between those two block groups. Unfortunately, the distribution of this data is quite odd in that most of the links are associated with very few people. However, if we only include a subset of the links, we may be able to create a more meaningful social proximity matrix.

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 districts, 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.

3 Response Time Modeling

3.1 Motivation

This project is also motivated by my work in the Social and Decision Analytics Lab (SDAL) in Arlington, Virginia. Last summer, we were interested in modeling both travel time and turnout time to structure fires. We define Total Response Time, Travel Time, and Turnout Time as in Figure 6, following the National Fire Protection Association (NFPA) definitions. Over the summer, we modeled turnout time and travel time separately, using linear and spatial Gaussian Process models respectively. For the first section of this part of the project, we follow the structure of Homework 1 to model a Gaussian Process and create estimates for β by hand. For the second section, we rely on the spBayes package to fit a slightly more complex model using knots and covariate information.

We ran into a couple problems in the implementation of this problem, as compared to the homework. First, in the homework, we had elevation data available for all of the grid points that we predicted for. In this problem, we do have some potential covariates available in the data, but we do not have any of the covariate information for the grid points. For example, the data includes the "priority" of the call and if the call was officer initiated or not. These could be two potentially interesting covariates, but they are obviously not available for all of the grid points across Wayne County, as is elevation, due to the nature of this problem.

Therefore, we proceed in this part of the project by creating an initial model that just incorporates spatial information (Latitude and Longitude) as covariates. We follow the structure of Homework 1 otherwise for the first part of this analysis. Then, we create a model using spBayes by creating a Gaussian process which incorporating covariates.

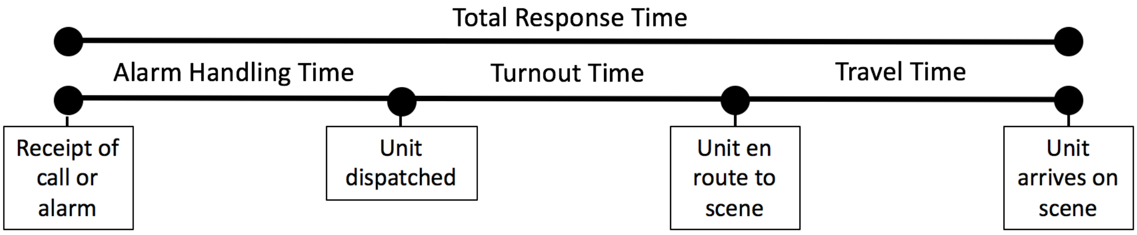


Figure 7: NFPA Response Time and Component Definitions

3.2 Basic Gaussian Process

3.2.1 Methods

We consider the same model for response times as we did for temperature in Homework 1, shown in equation (8) below, where $e(s_i; \sigma^2, \rho, \tau)$ is a zero mean stationary Gaussian process with exponential covariance function.

$$Y_i = \mu(s_i; \beta) + e(s_i; \sigma^2, \rho, \tau) \text{ where } \mu(s; \beta) = \beta_0 + \beta_1 \text{Longitude}(s) + \beta_2 \text{Latitude}(s) \tag{9}$$

As we saw in Homework 1, another way of writing this is as

$$Y_i = \mu(s_i; \beta) + Z(s_i; \sigma^2, \rho) + \epsilon_i$$

where now Z is a mean zero Gaussian process like e but without the nugget term, and the ϵ_i are iid $N(0, \tau^2)$, independent of Z.

As one of the first steps in our modeling process, we fit the variogram with a nonparametric estimate and a fitted variogram. Next, we fit this model specified above and calculate the GLS estimate of β by hand. We recall from Homework 1 that the *gls* function doesn't handle longitude and latitude well, so that is why we do it by hand. Finally, we use our estimates to plot the EBLUP of $\mu + Z$ at the grid locations that I have created for Wayne County. We also calculate and plot the estimated standard error of Z at each prediction location.

To review, we know from the class notes that the non-parametric estimate of the isotropic variogram is used to estimate a parametric model for $\gamma(h)$ using weighted least squares. We see $\hat{\theta}_{WLS}$ minimizes equation 10, where $n_u = \#\{s_i - s_j \in H_u\}$ and the function $\gamma(h) = \frac{1}{2}Var[Y(s+h) - Y(s)]$ is the semivariogram (or variogram).

$$\sum_u \frac{n_u}{\gamma(h_u;\theta)^2}[\hat{\gamma}(h_u) - \gamma(h_u;\theta)]^2 \tag{10}$$

3.2.2 Results

As mentioned above, the first step of this analysis is to create an ordinary least squares model. Due to the limitations of the prediction grid, we only have included Latitude and Longitude as covariates in this preliminary model. We have included my estimates for β through ordinary least squares in Table 5, as well as a color plot of the residuals in Figure 8. There is not extremely obvious spatial structure of the residuals, but this may be due to the dense nature of the data. It is also quite difficult to see the high values in the plot of the residuals.

	β_{linear}
(Intercept)	-10117.38
Longitude	-35.27
Latitude	170.27

Table 6: Linear Model Initial β Estimates

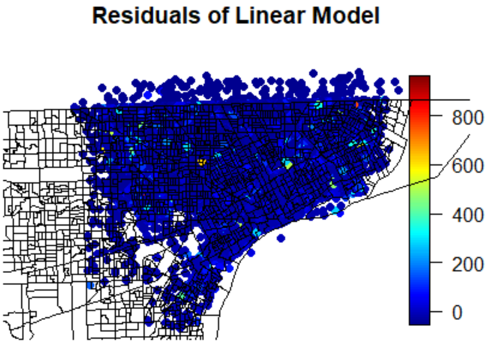


Figure 8: Residuals of Linear Model

Our next step in this analysis is to create a nonparametric and fitted variogram, in order to get estimates of the model parameters σ^2, ρ, and τ. We chose which variograms to use by using

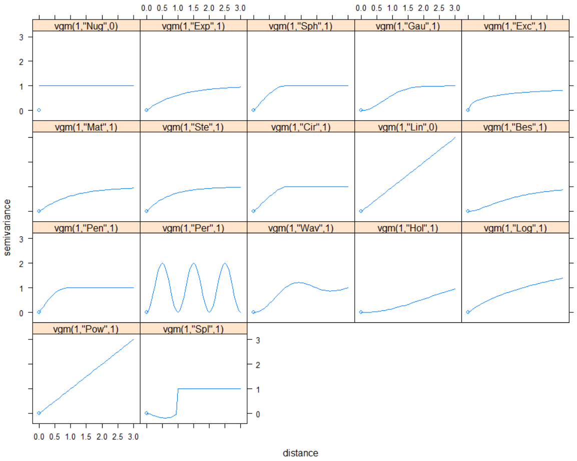


Figure 9: Variogram options available in R

the *show.vgms()* function in R. After seeing our nonparametric estimate and the options shown in Figure 9, we choose to fit both the Spherical and Wave variograms. We see in Figure 10 that the points on the plot represent the nonparametric variogram estimate and the line represents the fitted variogram, using weighted least squares. There seems to be some structure to the variogram and it seems that both the Spherical and Wave variogram models seem to provide a decent fit. Therefore, we proceed by creating models under both variograms and comparing the outputs.

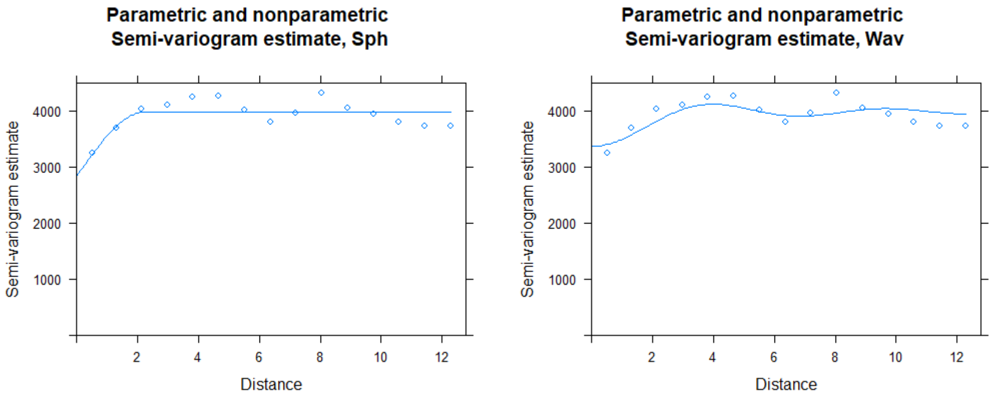


Figure 10: Variograms for Spherical and ”Wav” Variogram Fit

In Table 6 we have included my estimates of σ^2, ρ, and τ under both the spherical and wave variograms. We notice that they are pretty similar. Next, we form the estimates of β_{GLS} by hand, as discussed above. We know that $\hat{\beta}_{GLS} = (X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}Y$. By performing these calculations in R, we calculate the estimates and include these estimates for $\hat{\beta}_{GLS}$ below in Table 7.

Parameter	Wav	Sph
$\hat{\sigma}^2$	253.89	1354.36
$\hat{\rho}$	4.48	2.52
$\hat{\tau}^2$	4097.39	3004.09

Table 7: Estimates of variogram parameters

Parameter	Wav	Sph
Intercept	-3433.82	-8408.97
Long	-20.47	-47.54
Lat	41.69	106.01

Table 8: Estimates for Sph and Wav Variograms, $\hat{\beta}_{GLS}$

Lastly, we use these estimates to make prediction at unknown locations. In order to do this, we first have to create a grid of locations in order to predict over. We do this for Wayne County, and the grid points can be seen in Figure 11.

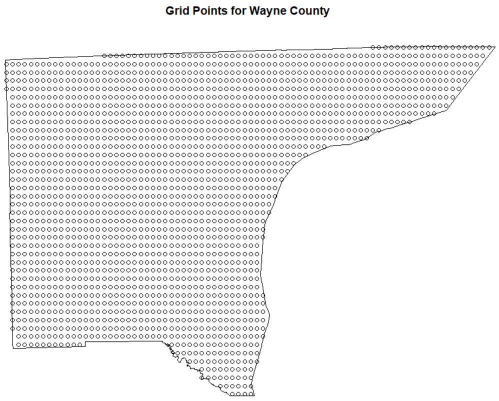


Figure 11: Grid over Wayne County

We use our estimates for β to apply the kriging predictor $\hat{Y}(s_0) = X^T\hat{\beta}_{GLS} + \gamma^T\Sigma^{-1}(Y - X^T\hat{\beta}_{GLS})$. As shown in the class notes, this gives us the EBLUP and a naive estimate of our prediction error. We have also included the estimated standard error of Z at each prediction location, where $Var[\hat{Y}(s_0) - Y(s_0)] = \sigma^2 - \gamma^T\Sigma^{-1}\gamma + b^T(X^T\Sigma^{-1}X)^{-1}b$ and where $b = x_0 - X^T\Sigma^{-1}\gamma$. We also see that the standard error ($se = \sqrt{variance}$) also has some spatial dependence.

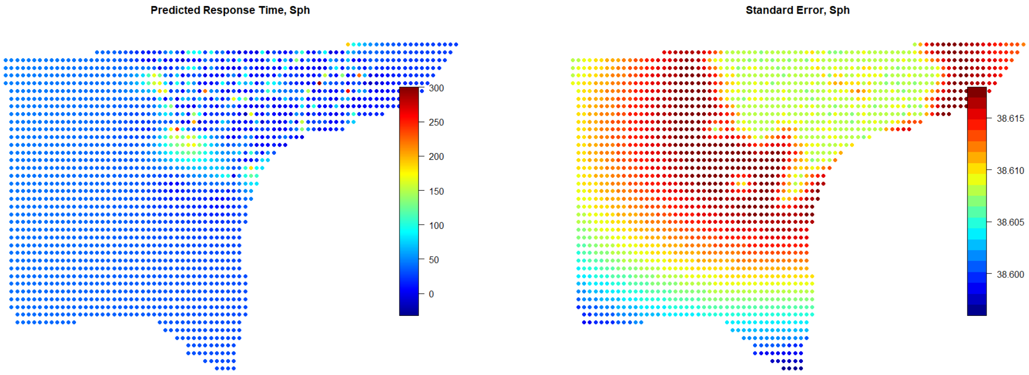


Figure 12: Predicted Response Times and Std Errors for Spherical Variogram Fit

In Figure 12, we see the EBLUP estimates for the Spherical Variogram and in Figure 13, we see the EBLUP estimates for the Wav Variogram. We see that the estimates are quite similar in terms of the EBLUP of $\mu + Z$. The range is similar and they are predicting the same kind of spatial structure in the middle of the county. However, other than this spot in the middle of the county, there is very little spatial structure in the estimate of $\mu + Z$. We expect this may be because of the lack of covariates other than Longitude and Latitude, and we improve on this in the model in the next section.

Although the estimates for $\mu + Z$ are quite similar and almost identical between the Spherical and Wave variograms, there is an interesting difference in the estimates of the standard error of Z. For both variograms models, we notice that there is a relatively low standard error where the observations lie, then it is higher around the observation points, and then it is lower at the border of Wayne County, especially towards the south. However, we don't believe that this estimate of the standard error is accurate because there are no observations at this point. Therefore, we should focus on the relatively low standard error where the observations are, and the higher standard error around these points. We see where the observations lie in Figure 14.

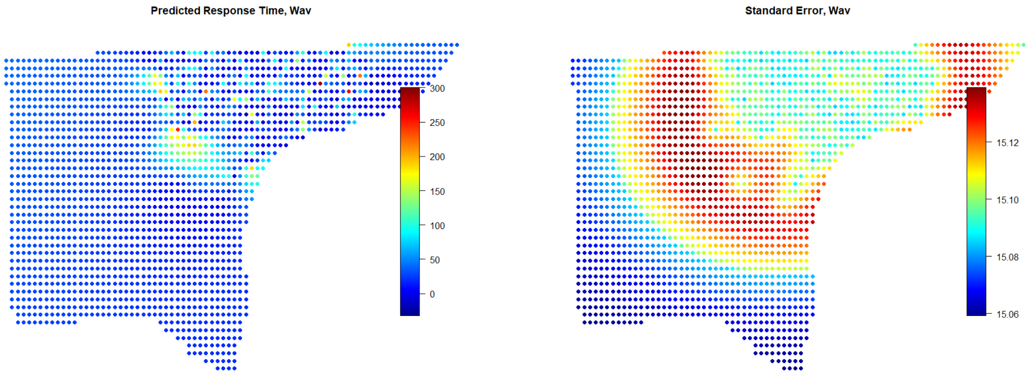


Figure 13: Predicted Response Times and Std Errors for Wav Variogram Fit

The interesting difference between the two variogram models lies in the range of the standard error estimates. We notice that for the spherical variogram, the range of the image is very small, around 36. We notice for the wave variogram that it has a similar structure but that the range is very small, around 15, instead of 36. Therefore, we conclude that the wave variogram might be a better model to use, rather than the spherical variogram.

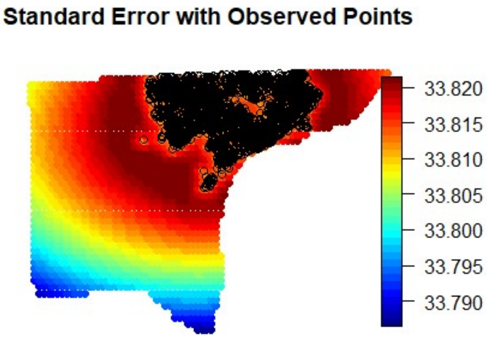


Figure 14: Std Error with Observations

In the last image, we have included the observation points on the image. Once again, we see that the standard error is relatively low where the observations lie and higher around the observations. The low part at the bottom of the county we believe is unreliable because of the lack of data.

3.3 Gaussian Process with Covariates, spBayes

3.3.1 Methods

Now, we fit a model that allows us to incorporate covariates by using the spBayes package. Let $\boldsymbol{Y} = \{Y(s_1), ..., Y(s_n)\}$ be observations over spatial locations $\boldsymbol{s} = \{s_1, ..., s_n\}$. By fitting a Gaussian process to $\boldsymbol{Y}$ we can interpolate a spatial surface, predict the surface at new locations $Y(s_0)$ and estimate uncertainties. Spatial dependence between nearby locations is accounted for by using a parametric covariance function. The Gaussian process assumption implies any multidimensional draw from a Gaussian process $\boldsymbol{Y}$ follows a multivariate normal density

$$\boldsymbol{Y} \sim N\left(\boldsymbol{\mu}(\boldsymbol{s}), \Sigma(\Theta)\right).$$

The mean is $\boldsymbol{\mu}(\boldsymbol{s}) = X(s)\beta$, which depends upon covariates $X(s)$ that can vary spatially. For our model, we include the covariates of "Officer Initiated" and "Priority". The first variable is an indicator variable, with 1 indicating that the officer did initiate the call. The latter is coded as 1, 2, or 3, where the first priority is the most urgent kind of call. The covariance is given by a function $\Sigma_{ij}(\Theta) = k(s_i, s_j, \Theta)$, with $\Theta = \left(\sigma^2, \phi, \tau^2\right)$. A commonly used class of covariance functions is the Matérn family[10] with smoothness parameter ν. A special case when $\nu = 1/2$ is the exponential covariance, seen in equation (10).

$$k(s_i, s_j, \Theta) = \sigma^2 \exp\left(-\frac{||s_i - s_j||}{\phi}\right), \tag{11}$$

which captures the strength (σ^2) and effective range (ϕ) of the spatial dependence. A 'nugget' effect (τ^2) can be added to explain micro-scale variability.

We fit a spatial Gaussian Process model to total response times in Wayne County. The total response time Y at a location s_0 is given by equation (11), where $\mu(s_0) = X(s_0)\beta$, $\epsilon(s_0) \sim N(0, \tau^2)$, and $w(\boldsymbol{s})$ is a Gaussian process with exponential covariance.

$$Y(s_0) = \mu(s_0) + w(s_0) + \epsilon(s_0) \tag{12}$$

We use MCMC to conduct inference, and find the starting values for the algorithm using a variogram technique, as in the previous model.

MCMC inference is done using the Gaussian predictive process of [11]. Flat priors are assumed for mean parameters β. Inverse gamma priors are used for σ^2 and τ^2 with shape parameter 2 and scale parameter chosen from the empirical semivariogram. A uniform prior is chosen for ϕ with a support that allows the process to vary from low to high spatial dependency. This model structure is very similar to the structure of my model for travel time in structure fire data in my research over the summer, which will hopefully be published soon.

3.3.2 Results

Before we can fit this model, we took a subset of the data. Due to computation issues, it was not possible to complete the model on the full data. This may result in some results that are not representative of the full data. We took a random subset of the data, and analyzed 10,000 out of the total of 400,000 data points.

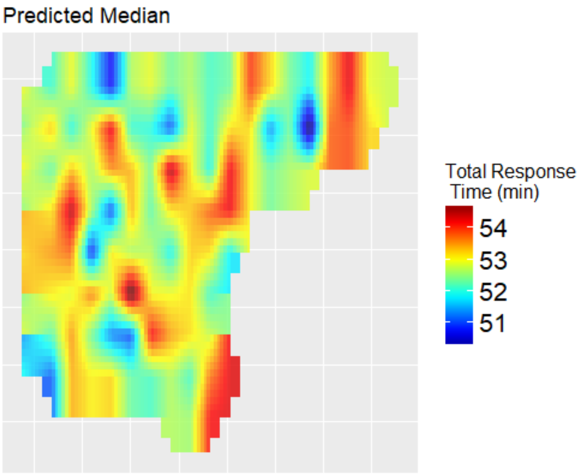


Figure 15: Predicted Median, Wayne County

After fitting this model to the subset of the data, we see some initial results. We would like to run the MCMC for longer but due to extensive computation time, we include our initial results with 5000 MCMC iterations. Figure 15 shows the predicted median that is smoothed across the spatial locations using *raster* and *disaggregate* in R. We wee that the predicted median values seem to be rather high where we observed many crimes. This might be due to the sheer volume of crime in these areas. Also, it is interesting to note the range of the predicted median is in the mid-50's. We believe that this is quite high for Total Response Time's and that it might be skewed by the high values in the data. It also may be due to the subset of the data that we took.

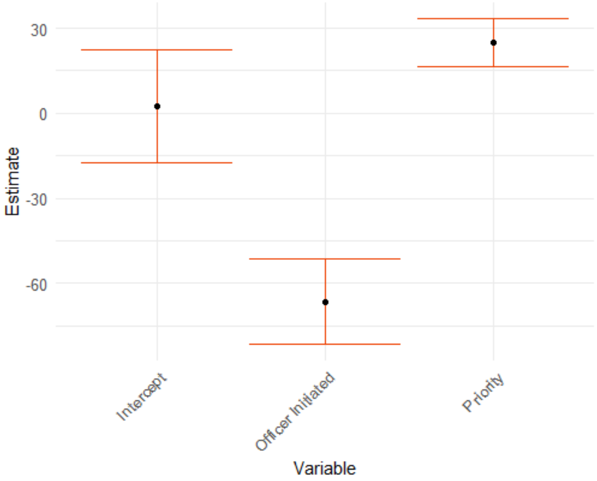


Figure 16: Covariates Credible Intervals and Median

Next, we compute the median, and 95% credible intervals for all 3 of our covariates. We see that the intercept is actually not significant, as the credible interval includes 0. However, the credible interval for Officer Initiated seems to suggest that this has a negative effect on Total Response Time. In other words, if the call was officer initiated, the Total Response Time would be longer, controlling for other factors. This may be due to the fact that the officers do not accurately report when they have completed their responses, or that officer initiated calls tend to be due to traffic stops or less serious crimes, which can have longer response times. Lastly, we see that Priority has a positive effect on Total Response Time. This would make sense because Priority 1 is the most urgent kind of call, followed by Priority 2 and then Priority

3. Therefore, as a call gets more urgent, the Total Response Time would decrease, holding everything else constant. This intuitively makes sense.

3.4 Discussion/Future Work

There is a lot of potential for future work in this part of the project. For the project that I did over the summer, we were interested in response times to structure fires (not crimes). We created a model to see if certain fire stations had a significantly higher or lower turnout time than other stations. The dataset that was made available to us for this project was from the Arlington County Fire Department, and many other potential kinds of fire that we could use for the project, but we focused on structure fires because these kinds of fire often result in the highest damage. Therefore, for this project we would like to consider including only a subset of the crime data, such as violent crime as with the first part of this project. Unfortunately, there are 206 crime types in the dataset, so this is not an easy task to include this in the model or to subset the data as the categories are not intuitively coded.

In the future, we would like to also spend some time investigating the data further or talking with the data providers. We have many questions about the data. For example, we are curious about what some of the categories for crime mean. However, our main concern comes in the distribution of the Total Response Time. In the data, some of the Total Response Times are negative, though this is a very small number (12). We assume these are errors and just remove them from the data. However, there are also concerns at the high end of the distribution. We see that there are many values over 900 for Total Response Time. Unfortunately, the unit of Total Response Time is not given on the website that the data is hosted on. We hypothesize that the unit is minutes, because there are many values that are between 5 and 15, which would be almost impossible for seconds. Therefore, these response times at the high tail of the distribution are in the range of 15-16 hours, which is quite concerning. It also definitely does not follow the guidelines of the NFPA in terms of quick Response Times. Therefore, we would like maybe eliminate some of the higher times because of data entry errors or know that we are justified in including them in the model.

4 Appendix

4.1 β Posterior Medians and Credible Intervals, Social and Geographic

	0.5	0.025	0.975
Intercept	2.18	1.93	2.49
Median Income	-0.00	-0.00	-0.00
Unemployment Rate	0.43	0.24	0.57
Total Population	0.00	0.00	0.00
Percentage Male	0.43	0.36	0.50
Median Age	-0.00	-0.00	-0.00
Herfindahl Index	0.80	0.77	0.83

Table 9: BYM Geographic

	0.5	0.025	0.975
Intercept	2.63	2.40	2.97
Median Income	-0.00	-0.00	-0.00
Unemployment Rate	0.41	0.20	0.55
Total Population	0.00	0.00	0.00
Percentage Male	0.09	-0.02	0.17
Median Age	-0.00	-0.00	0.00
Herfindahl Index	0.64	0.61	0.68

Table 10: BYM Social

	0.5	0.025	0.975
Intercept	2.41	2.20	2.69
Median Income	-0.00	-0.00	-0.00
Unemployment Rate	-0.01	-0.16	0.10
Total Population	0.00	0.00	0.00
Percentage Male	0.71	0.62	0.78
Median Age	-0.00	-0.00	-0.00
Herfindahl Index	0.43	0.40	0.45

Table 11: Leroux Geographic

	0.5	0.025	0.975
Intercept	3.41	3.30	3.50
Median Income	-0.00	-0.00	-0.00
Unemployment Rate	-0.23	-0.27	-0.19
Total Population	0.00	0.00	0.00
Percentage Male	-0.00	-0.06	0.05
Median Age	-0.00	-0.00	-0.00
Herfindahl Index	0.47	0.45	0.49

Table 12: Leroux Social

4.2 β Posterior Medians and Credible Intervals, Additive and Binary

	0.5	0.025	0.975
1	2.70	2.46	3.04
2	-0.00	-0.00	-0.00
3	0.22	0.01	0.37
4	0.00	0.00	0.00
5	0.05	-0.06	0.15
6	0.00	0.00	0.01
7	0.84	0.81	0.87

Table 13: BYM Additive

	0.5	0.025	0.975
1	2.43	2.19	2.77
2	-0.00	-0.00	-0.00
3	0.58	0.37	0.72
4	0.00	0.00	0.00
5	-0.23	-0.32	-0.14
6	0.00	0.00	0.01
7	0.84	0.81	0.87

Table 14: BYM Binary

	0.5	0.025	0.975
1	3.39	3.28	3.49
2	-0.00	-0.00	-0.00
3	-0.05	-0.09	-0.01
4	0.00	0.00	0.00
5	-0.28	-0.33	-0.23
6	-0.00	-0.00	-0.00
7	0.45	0.43	0.47

Table 15: Leroux Additive

	0.5	0.025	0.975
1	3.24	3.13	3.34
2	-0.00	-0.00	-0.00
3	-0.07	-0.12	-0.02
4	0.00	0.00	0.00
5	0.30	0.25	0.35
6	-0.00	-0.00	-0.00
7	0.26	0.24	0.28

Table 16: Leroux Binary

5 References

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

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

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

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

[5] 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.

[6] 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.

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

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

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

[10] Mark S Handcock and Michael L Stein. A Bayesian analysis of kriging. *Technometrics*, 35(4):403–410, 1993.

[11] Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.