A Comparison of Spatial Models Incorporating Nonspatial Information, with a Policing Case Study

Abstract

Incorporating nonspatial variables into spatial models is of increasing importance, especially when studying policing. Experts propose ways of combining spatial and nonspatial data into spatial models, such as the k-Groups model [Quick et al., 2015] and the Two-Stage model [Kelling and Haran, 2022]. We explain existing methods in spatial statistics, and compare and contrast the two selected strategies. Utilizing a police use of force data set from Minneapolis, Minnesota, we use computational tools such as predictive processes and Bayesian computing to estimate model parameters. The k-Groups method models use of force incident intensity as a function of spatial variables differentiating between each combination of categorical nonspatial variables. The Two-Stage method models (1) the spatial use of force incident intensity and (2) the probability of our mark (weapon or no weapon) influenced by nonspatial and spatial variables, and their interaction. Finally, we discuss the benefits and drawbacks of each method.

1 Introduction

Excessive use of force by law enforcement is a cause of great concern across the United States. There is often a lack of transparency and accountability from law enforcement Geller et al. [2021], and the disclosure of comprehensive policing data to the public has been more sought after than ever. Incorporating information about the individuals involved in policing events into spatial models could allow for a greater understanding of police use of force. We define this as 'nonspatial' information, available either by individuals, such as officers/civilians, or by events, as opposed to spatial information that is defined in the space where an event happened. Through the use of this information, we may be able to gain insight into patterns or potential biases in law enforcement interactions. Additionally, nonspatial data can create a more complete picture of police-civilian interactions. For instance, simply examining use of force incidents does not explore whether the frequency of incidents is due to civilian behavior or officer misconduct [Fryer, 2019].

Community characteristics have been investigated to explain policing information. Previous research has found that marginalized communities are more likely to have police present in their neighborhoods [Lersch et al., 2008]. To investigate this relationship, it is necessary to consider spatial information that provides insight into a neighborhood's socioeconomic and diversity status. Spatial modeling is another way to incorporate community characteristics into models describing policing data. For example, Kelling and Haran [2023] analyzed police use of force and police stops in Chicago, estimating the likelihood of observing one of these events across the city based on census tract unemployment rates, measure of diversity, and median age.

In addition, individual-level characteristics have been used when analyzing policing data. When investigating these individual attributes, we find that police departments have a large level of discretion when it comes to how much detail they include in their data released to the public, as well as how often they release this data [Geller et al., 2021]. For example, lethal use of force events are often the type for which data are most consistently collected due to its severity [Geller et al., 2021]. Information regarding lower-level instances of force, which occur more often than officer-involved shootings, is difficult to find [Fryer, 2019]. However, lower-level instances of force can still have a large impact and are worth exploring especially considering their frequency and thus impact on every-day life [Fryer, 2019].

While scholars have raised concerns based on lived experiences about disparities in policing practices by race, there remains little systematic understanding on how criminologists and other experts should analyze the impact of race[Obasogie and Provenzano, 2023]. In Wright and Headley [2020], it was explored whether the race of the civilian and race of the officer had an impact on the level of force used by police. It was found that white officers were more prone to use higher levels of force against Black civilians. Similarly, Smith et al. [2023] discovered that Black civilians encountered greater force severity and a larger total amount of force compared to white civilians within a large county police agency. Moreover, Maksuta et al. [2024] found relationships between changes in racial/ethnic minority proportions in neighborhoods and police-involved homicides, as well as associations with measures of social disorganization. In contrast, it was found that civilian resistance was not a discernible predictor of force, though this contradicts some previously existing findings [Smith et al., 2023]. Other research in the field has examined the relationship between officer race and predicting police use of force and civilian resistance, such as Paoline et al. [2018]. Due to the results of the studies investigating police use of force, there is a strong desire for supervision of law enforcement actions [Fryer, 2019]. Minneapolis has been at the lead of many of these discussions following the events that have occurred over the past few years, especially the murder of George Floyd by Minneapolis police [Smith, 2022].

In this analysis, we build upon the work by Liang and Carlin [2008], which devised new methods for estimating spatial models, by reviewing different statistical methods that incorporate nonspatial information into spatial models. We also analyze data on Minneapolis Policing Districts, Police Use of Force [Open Data, 2024], and information from the American Community Survey [Census Data, 2021] to both generate a thorough analysis and visualization of Minneapolis Police Use of Force data as well as integrate nonspatial information into spatial models. We make initial conclusions regarding the relationship between police use of force events and nonspatial information. Finally, we compare two models, the k-Groups model developed by Quick et al. [2015] and the Two-Stage model proposed by Kelling and Haran [2022], to explore the similarities and differences regarding what can be said by each of the two models, and consider computational challenges that must be overcome in the process.

We will begin with outlining the methodology for modeling point process data, found in Section 2. This includes discussions of complete spatial randomness in Section 2.1 and both the Nonhomogeneous Poisson Process in Section 2.2 and the Log Gaussian Cox Process in Section 2.3 as modeling options that allow the spatial intensity to vary. We then outline the computational details proposed by Liang and Carlin [2008] which include utilizing predictive processes to help alleviate computational challenges and Markov Chain Monte Carlo methods to estimate parameters in Section 2.3.1. We then explore how both the k-Groups model and the Two-Stage model incorporate nonspatial data, found in Section 2.4. Then, we investigate an application of these methods using a police use of force data set from Minneapolis, Minnesota, described in Section 3. Finally, we share the results of our analyses in Section 4, compare the two models in Section 4.4, and discuss limitations and potential next steps in Sections 5.1 and 5.2 respectively.

2 Methodology for Modeling Point Processes

In this section, we will begin with the definition of a point process, as well as complete spatial randomness (CSR). Next, we will outline the types of point process models that we consider, their motivations, and computational considerations. Then, we will describe two ways that past researchers have incorporated nonspatial data into spatial models.

2.1 Complete Spatial Randomness

A point process consists of individual events $s_1, s_2, ...s_n$ (n= total number of events) at specified locations within a spatial domain or window, W [Bivand et al., 2013]. These points can be completely spatially random (CSR) or form other spatial patterns. In order to test whether the data is completely spatially random, we compute and analyze the G and F functions for the point process dataset [Bivand et al., 2013].

The G function can determine whether a point process is completely spatially random by first computing the nearest neighbor distance of each event. The nearest neighbor is defined as the event j with the shortest distance d from event i ($d_{ij} \leq d_{ik} \forall k$). For the G function, we plot a cumulative distribution function (CDF) of these distances as a summary visualization of the nearest neighbor distances [Bivand et al., 2013]. We can test for CSR by comparing the CDF of the police use of force (our dataset considered in this paper) nearest neighbor distances to simulations of point processes with the same number of points and the same spatial domain simulated under CSR. If nearest neighbor distances are smaller for the use of force dataset, as summarized by the G function, compared to the CSR simulations, then there is preliminary evidence of clustering, and against CSR. The CDF of clustered data would thus be to the left of the CSR envelope, which summarizes many CSR simulations [Bivand et al., 2013].

The F function can also determine whether a point process is completely spatially random, but the first step is to define an arbitrary set of points in the spatial window. We then calculate the distance between each arbitrary point and its nearest event (police use of force incident) in the dataset to create a CDF, and compare this to simulations of point processes under CSR [Bivand et al., 2013]. If nearest neighbor distances are larger for the dataset, then there is more blank space in the police use of force dataset, meaning we have evidence that the dataset does not follow CSR but rather shows evidence of clustering. The CDF of clustered data would thus be to the right of the CSR envelope, again summarizing many instances of CSR [Bivand et al., 2013].

If the data exhibits CSR, then an implied model for the data is the Homogeneous Poisson process (HPP), where the number of events in the window with area $|W|$ is Poisson distribution with mean $\lambda * |W|$, and λ represents the **constant** intensity of the point process model [Bivand et al., 2013]. If the data shows evidence against CSR, then this motivates modeling the data with other point process models such as the Nonhomogeneous Poisson Process (NHPP) or the Log Gaussian Cox Process (LGCP), where the intensity is allowed to vary over the spatial domain [Bivand et al., 2013].

2.2 Nonhomogeneous Poisson Process

When the data does not exhibit CSR, this motivates the use of a model that allows the spatial intensity to vary. One preliminary option is a model called the Nonhomogeneous Poisson Process (NHPP), which is a

generalization of the HPP [Bivand et al., 2013]. As illustrated by Bivand et al. [2013], the spatial intensity under this model follows Equation 1.

$$log\left(\lambda(s)\right) = x(s)'\beta \tag{1}$$

In Equation 1, s is the location of an event, $x(s)$ is the vector of spatial covariates at each location s, and β is the vector of coefficients, which are then estimated from the data, for the spatial variables. The assumptions for this model are that the points follow a Poisson distribution with mean $\int_W \lambda(s)ds$, and the n events in W form an independent random sample from the distribution on W with pdf proportional to $\lambda(s)$ [Diggle, 2009].

While it is possible to estimate a spatial intensity of a point process non-parametrically via kernel smoothing, an alternative approach is to estimate the spatial intensity parametrically utilizing the maximization of the likelihood function of an NHPP [Bivand et al., 2013]. The log-likelihood of a realization of independent events from an NHPP is shown in Equation 2, where each s_i is the individual event and $\lambda(s)$ is the spatial intensity of an NHPP, defined above [Bivand et al., 2013].

$$\mathcal{L}(\lambda) \propto \sum_{i=1}^{n} log\left(\lambda(s_i)\right) - \int_W \lambda(s)ds \tag{2}$$

The NHPP model can be useful when considering relatively simple spatial data, but complications arise when working with more complex data due to the model's limiting attribute that the estimated intensity is regionally constant when considering spatial covariates that are defined on areal units [Liang and Carlin, 2008]. For example, the estimated intensity in Equation 1 is regionally constant because the spatial variables, $x(s)$, come from census data, where there is only one value per census unit such as census tracts. In practice, it is rare to find situations where the spatial intensity is regionally constant over an areal unit, such as a census tract, and then immediately changes to a new intensity at a neighboring census tract. A more common scenario is one in which there is a smoother spatial intensity, which can be modeled by the Log Gaussian Cox Process.

2.3 Log Gaussian Cox Process

The spatial intensity of an LGCP, defined in Equation 3, is very similar to the NHPP intensity but includes an additional term, $\omega(s)$ Liang and Carlin [2008], which represents a Gaussian Process (GP).

$$log\left(\lambda(s)\right) = x(s)'\beta + \omega(s) \tag{3}$$

A Gaussian Process, $\omega(s)$ has the distribution defined in Equation 4.

$$\omega(s) \sim MVN(0, \Sigma) \tag{4}$$

The $\omega(s)$ defined in Equation 4 is a column vector, fully represented in Equation 5.

$$\begin{pmatrix} \omega(s_1) \\ \omega(s_2) \\ \vdots \\ \omega(s_n) \end{pmatrix} \tag{5}$$

Existing methods consider many possibilities for modelling the covariance for the Gaussian Process, such as the structure defined in Equation 6 [Stephens, 2020].

$$\Sigma_{ij} = cov(\omega(s_i), \omega(s_j)) = \sigma^2 exp\left(\frac{-|s_i - s_j|}{\phi}\right) \tag{6}$$

The parameter σ^2 represents the scale, ϕ represents the smoothness, and $|s_i - s_j|$ represents the distance between any two events, s_i and s_j. The additional Gaussian Process term in an LGCP allows the spatial intensity to vary within areal units, such as the census tracts, because the GP is continuous.

The formula for the log-likelihood of an LGCP does not differ from the formula for the log-likelihood of an NHPP shown in Equation 2, except the intensity function defined there now includes the GP [Liang and Carlin, 2008]. To maximize the log-likelihood for the LGCP, it is necessary to utilize the computational techniques summarized in the following section.

2.3.1 Computational Details

We utilize a Bayesian approach to estimate the parameters of the log Gaussian Cox process, β and σ. In a Bayesian approach, we start with a prior distribution and then utilize data that is collected to form a posterior distribution. This can be seen in Bayes' theorem in Equation 7

$$P(\beta \mid data) = \frac{P(data \mid \beta) \times P(\beta)}{P(data)} \tag{7}$$

Bayes' theorem allows us to estimate parameters, β and σ in the case of the LGCP, given data. The left side of Equation 7 represents the posterior distribution, which is what we are attempting to estimate. The component $P(data \mid \beta)$ in Equation 7 represents the likelihood. $P(\beta)$ is the prior distribution. The prior distribution is what is assumed before viewing the data. Priors can be relatively uninformative or based on previous information. Still, the denominator of Bayes' theorem can be difficult to compute, so we used Markov Chain Monte Carlo (MCMC) estimation in the NIMBLE R package [de Valpine et al., 2017]. This package allows users to compile code into C++ for fast computation as well as implement the maximization of customized log-likelihood functions easily.

When implementing NIMBLE R code, as done in de Valpine et al. [2017], it is best to fix either σ or ϕ in Equation 6 to due to identifiability concerns in the estimation of the parameters of the GP covariance function. In our application, we chose to fix ϕ as done in Liang and Carlin [2008] such that the 95th percentile of distances have a correlation of 0.05 and the value of the 5th percentile of distances have a correlation of 0.95. We fix ϕ at the average of these two values. The priors for our regression coefficients, β, follow a Normal(0,100) distribution. As done in Liang and Carlin [2008], we use an Inverse-Gamma($\alpha = 2$, $\beta = 0.5$) prior for σ. Finally, we need to estimate the Gaussian Process at each event in the data in order to calculate the first component of Equation 2. However, the dimensions for Σ are too large given that matrix multiplication and inversion are required to estimate a Gaussian Process. Therefore, we use knots, or lower resolution points, to estimate the Gaussian Process at the locations of the data. The choice for the knots used in our application is found in Appendix B. Our prior for these Gaussian Process values at lower resolution points follows a Multivariate Normal Distribution.

Because we estimated the Gaussian Process in a low resolution, we must use predictive processes to transform the low-resolution Gaussian Process values on the knots into a higher resolution [Banerjee et al., 2008]. Equation 8 demonstrates this process, where $\widetilde{\omega}(s)$ is the estimated Gaussian Process over higher resolution points (such as the locations of the data). The matrix $cov(\omega(s), \omega^*)$ is the covariance matrix between the higher and lower resolution points, $var^{-1}(\omega^*)$ is the inverse of the covariance matrix between the lower resolution points, and ω^* is the estimated Gaussian Process values on points in a lower resolution.

$$\widetilde{\omega}(s) = cov(\omega(s), \omega^*) var^{-1}(\omega^*) \omega^* \tag{8}$$

Next, to maximize the log-likelihood function, it is necessary to integrate the intensity over the spatial domain in order to compute $\int_W \lambda(s)ds$. Functionally, this integral must be estimated, so we utilize Monte Carlo integration. Monte Carlo integration is an estimation technique that evaluates integrals through simulations, shown in Equation 9, where $s_{int,i}$ represents randomly generated integration points across the spatial domain W. A discussion of how the integration points were generated for our application can be found in Appendix B. The intensity is calculated at each integration point, and the average intensity is determined by summing the calculated intensity at each integration point and dividing by the total number of integration points. The final step in estimating $\int_W \lambda(s)ds$ through Monte Carlo integration is to multiply the average intensity by the area of the spatial domain. Note that to estimate $\lambda(s_{int,i})$ over the integration points, we also apply a predictive process transformation to estimate the Gaussian Process at the integration points.

$$\widehat{\int_W \lambda(s)ds} = \frac{\text{Area}}{n_{int}} \sum_{i=1}^{n_{int}} \lambda(s_{int,i}) \tag{9}$$

The Markov chain portion of MCMC Bayesian estimation involves maximizing the likelihood function over many iterations of different model values. First, we choose initial values for our model parameters (β's, σ, Gaussian Process on knots), then propose a new value for each parameter based on their respective

(relatively uninformative) prior distributions, then either accept or reject each proposed value based on how probable the proposed value is. More specifically, the process utilizes the Metropolis algorithm when proposing a new value, which involves computing the ratio of the probability of the proposed value given the observed data to the probability of the existing value given the observed data. If this ratio is greater than 1 - or in other words, if the probability of the new value is greater than the probability of the old value - then the new value is accepted. If the ratio is less than 1, the new value is accepted with probability equal to the ratio of the two aforementioned probabilities. Over thousands, hundreds of thousands, or even millions of iterations, the estimates for all parameters should begin to converge, changing less severely and even less frequently as the likelihood slows its increase; the MCMC trace plots for each parameter will reflect this and can be used to visually assess for convergence. After convergence has been established, the period of burnin - the part of the MCMC chain before the estimates converge - can be identified and removed. Once the burnin is removed, credible intervals for each model parameter are calculated by finding the quantile estimates of the posterior distribution after removing burnin, and the mean of these values serve as estimates of the true values.

2.4 Incorporating Nonspatial Data

The study of incorporating nonspatial data into point process analysis is a current and pressing area of research. One of the primary goals of our research is to assess existing methods of incorporating nonspatial data and determine the types of questions that can/cannot be answered with the existing methods. The two main methods we analyzed are the k-Groups method developed by Quick et al. [2015] and the Two-Stage method developed by Kelling and Haran [2022].

2.4.1 *k*-Groups Method

Quick et al. [2015] outlines one method for incorporating categorical nonspatial variables in which there is a separation of observations into k groups based on the characteristics of interest. For example, let the two categorical variables of interest be binary sex (male or female) and binary race (white or non-white). Then, group k_1 would include all observations where the individuals are white male, k_2 would include all observations where individuals are non-white male, k_3 would include all observations where the individuals are white female, and k_4 would include all of the observations where the individuals are non-white female. Following the designation of k groups, Quick et al. [2015] proposes the utilization of an LGCP very similar to the model described in Section 2.3. The equation for the log-likelihood of the realization of independent events used by Quick et al. [2015] is defined in Equation 10, where $log\left(\lambda_k(s)\right) = x_k(s)'\beta_k + w_k(s)$.

$$\mathcal{L}(\lambda_k) \propto \sum_{i=1}^{n} log\left(\lambda_k(s_i)\right) - \int_W \lambda_k(s)ds \tag{10}$$

The β coefficients for the spatial variables of interest are estimated for each k-group as described in Section 2.3.1. To investigate the impact of the nonspatial variables, we compare how the β coefficients differ across the k groups [Quick et al., 2015]. However, it is important to note we used a simplification of the method proposed by Quick et al. [2015] in that we assumed independence between the k-Groups and we only worked with incorporating categorical variables. Information on incorporating dependence between k-Groups can be found in Appendix A.

2.4.2 Two-Stage Method

A second approach that we consider to incorporate nonspatial data into spatial models is developed by Kelling and Haran [2022]. The proposed model does not separate the data by the categorical variables of interest. Instead, the analysis is split into two different stages. The first stage utilizes an LGCP model to estimate the spatial intensity of the events based on spatial variables as described in Section 2.3. The second stage, referred to as the mark determination stage, utilizes regression analysis to determine the distribution of the mark using both spatial and nonspatial data [Kelling and Haran, 2022]. A mark is the nonspatial outcome variable that the researchers wish to explain. For our policing application, we chose a binary mark, which is best modeled using logistic regression. A general logistic regression function for the mark determination

stage is shown in Equation 11, where $\pi(s)$ is the probability of the reference mark, $x(s)$ is the vector of spatial covariates, γ is the vector of coefficients for the spatial variables, ν is the vector of nonspatial variables, α is the vector of coefficients for the nonspatial variables, and δ is the vector of coefficients for interaction terms between the spatial and nonspatial variables.

$$logit(\pi(s)) = x(s)'\gamma + \nu'\alpha + (x(s)'\nu)\delta \tag{11}$$

It is important to note, however, that this is a simplification of the Two-Stage model for two main reasons. First, the type of mark determines the link function for the regression model that is used, which changes the likelihood function. For example, when working with a mark consisting of counts, one might want to utilize Poisson regression instead of logistic regression [Kelling and Haran, 2022]. Second, as described in this section, each stage of the Two-Stage process considered in this paper is treated independently. However, the full Two-Stage model described by Kelling and Haran [2022] allows for dependence between the two stages through the utilization of Gaussian Processes, $\omega_1(s)$ and $\omega_2(s)$, in each stage of the model. See Appendix A for more information on incorporating dependence between the stages.

3 Data

This project focused on data from two sources. First, we utilize data collected by the Minneapolis Open Data Portal, including Minneapolis Policing Districts and Police Use of Force [Open Data, 2024]. The Minneapolis Policing Districts dataset contains the spatial shapefile information for the shapes of the police districts for Minneapolis. Second, we include American Community Survey data collected and published by the Census Bureau [Census Data, 2021].

The original Police Use of Force dataset contained 39,758 observations with 30 variables. These variables included the anonymized latitude and longitude coordinates of police use of force incidents, the date of the incident (beginning in January 2008 and ending in December 2022), the type of force used on the civilian, the type of resistance by the civilian, and the race of the civilian. It is important to note that we are unsure as to how Latine civilians were categorized in the police use of force data. This population may have been categorized as white, Other/Mixed Race, or Unknown, so our analysis is limited. In addition, while we are unsure how the coordinates were anonymized because the Minneapolis Police Department has not formally documented their privatization process, it is likely that coordinates within a block are moved to the midpoint of the block, as supported by visual inspection in Figure 1. An assumption of point process models is that points are non-overlapping. Therefore, we slightly jittered the overlapping points in the dataset so that there were no repeated locations of police use of force incidents.

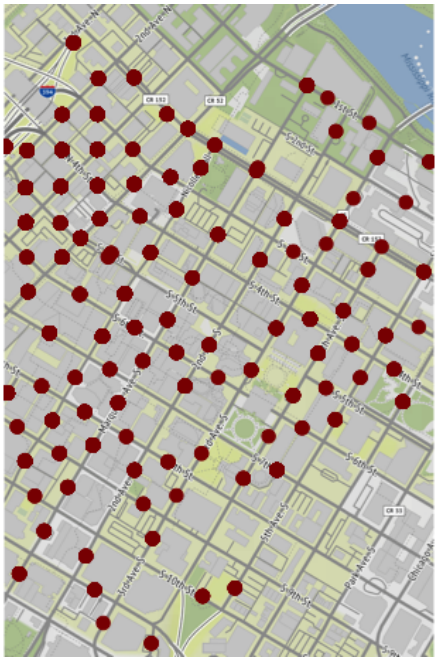


Figure 1: Original Use of Force Incidents at Anonymized Locations

It is necessary to consider how changes in data collection or reporting practices can impact this dataset. Figure 2 shows the total counts of Minneapolis use of force incidents over the entire range of dates, grouped by month. Though quite variable on a month-to-month basis, there appears to be a general trend of higher incident counts in earlier years, followed by slightly decreased counts from around 2012 to 2019, finally followed by a sharp increase in 2020, which is suspected to be a result of increased scrutiny following the murder of George Floyd in May 2020 and changes in data practices [Smith, 2022]. As we did not find this plot to be strongly indicative of a distinct change in practices, we chose to cut off the use of force data so that dates ranged from 2018 onward.

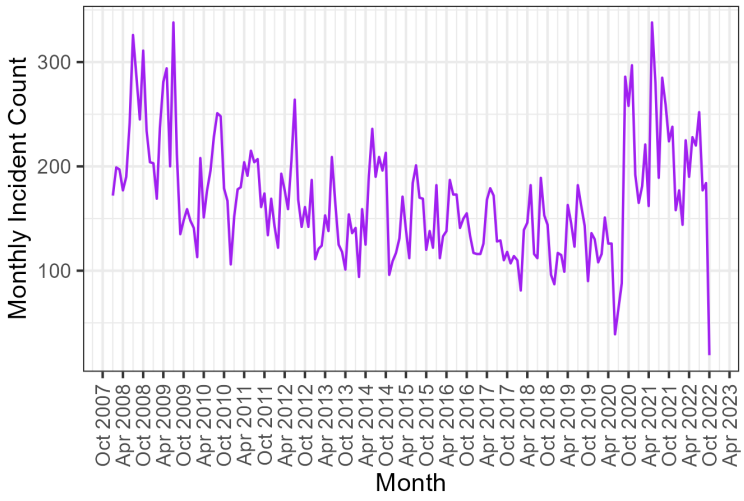


Figure 2: Use of Force Incident Count by Month

Finally, the original American Community Survey dataset contained information about 329 census tracts in Hennepin County, Minnesota. We collected the variables for the total population (shown in Figure 3 at a county level), number of people unemployed, number of people in the civilian labor force, white alone

population, Black alone population, American Indian or Alaskan alone population, Asian alone population, Hawaiian alone population, Other race alone population, and two or more races alone population. We then calculated the unemployment rate as the number of unemployed people divided by the civilian labor force, as well as the Herfindahl-Hirschman index (HHI), calculated as defined below:

$$\text{HHI} = \left(\frac{\text{Race 1 Count}}{\text{Total Population}}\right)^2 + \left(\frac{\text{Race 2 Count}}{\text{Total Population}}\right)^2 + ... + \left(\frac{\text{Race n Count}}{\text{Total Population}}\right)^2$$

The HHI is a measure of diversity, with a value close to 1 corresponding to an areal unit with low diversity. This occurs because the proportion of the race counts relative to total population will have a value close to 1, making its square close to 1. A value close to 0 corresponds to an areal unit with high diversity because the proportion of the race counts relative to total population for each race will all be closer to 0, and adding their squares together will produce a value closer to 0.

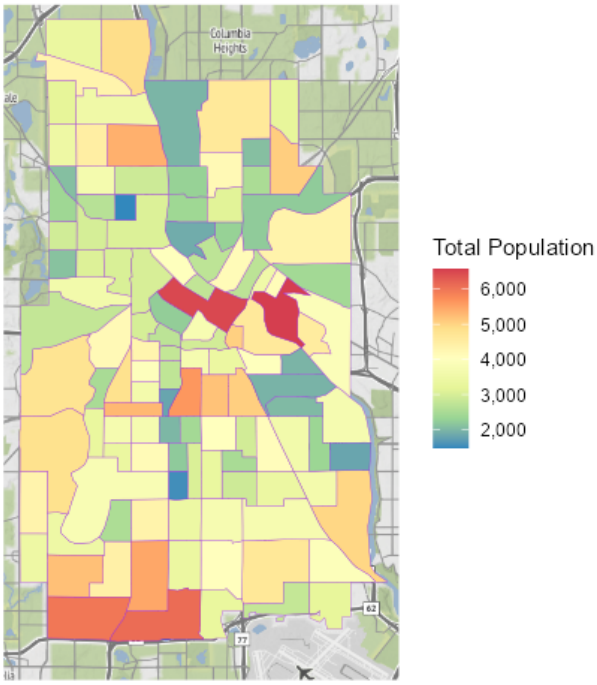


Figure 3: Map of Total Population by Census Tract

3.1 Pre-processing

The first part of processing the data was to narrow down the census tracts to include only census tracts in the police district. However, the census tract boundaries and police district boundaries do not completely align. Therefore, we calculated the percentage of overlap between the census tracts and the police districts. We then removed all census tracts that did not have at least 70 percent coverage by the police districts from our American Community Survey dataset. We were left with 121 census tracts.

Next, we merged the finalized census tracts dataset with our Police Use of Force dataset so that each use of force incident was affiliated with a census tract and its corresponding spatial covariates. We excluded some use of force incidents that were not in the 121 census tracts and were not near the boundary of these tracts. We also moved some points that were near the boundary back into the census tracts because the anonymization process may have changed where the true observation occurred. The most that a point was moved was 427 feet. Figure 4 displays the original and moved points. Our final dataset included 34,260 observations of police use of force.

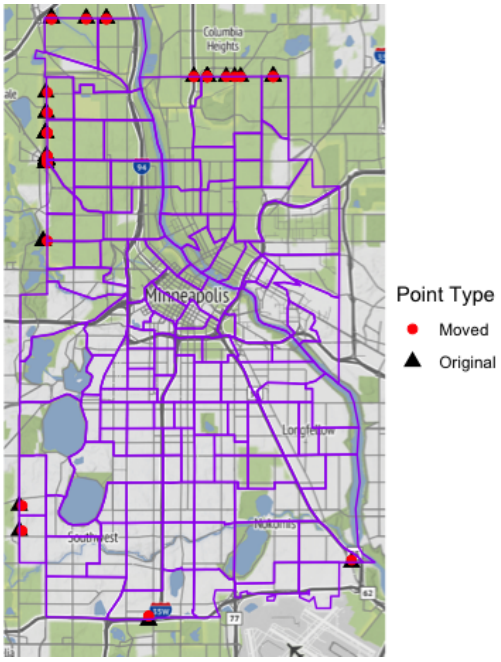


Figure 4: Points Moved Into the Census Tract

3.2 Variable Groupings

We created two binary variables, type of force and race, for our k-Groups analysis. The type of force category included weapon and non-weapon. Non-weapon comprised of acts such as bodily force and maximal restraint technique while weapon comprised of acts involving an object such as a taser, firearm, baton, chemical irritant, police K9 bite, improvised weapon, less lethal projectile, and gun point display [Lee et al., 2010]. Our four k-Groups are thus white weapon, white non-weapon, Black weapon, and Black non-weapon. Table 1 displays the counts of the four groups. The race category included white and Black civilians, leaving out Native American, Asian, Pacific Islander, Other/Mixed Race, Unknown, and Not Recorded. While the experiences of all races when interacting with police are important, they are also different from one another and must be separated [Gorsuch and Rho, 2019]. We elected to narrow our focus to include only white and Black citizens for this project. Figure 5 shows the map of incidents split using the binary race variable, which takes on values of 'White' and 'Black' respectively.

	Weapon	Non-weapon	Total
White	599	2,253	2,852
Black	1,384	5,291	6,675
Total	1,983	7,544	9,527

Table 1: Count of Use of Force Incidents by k-Groups

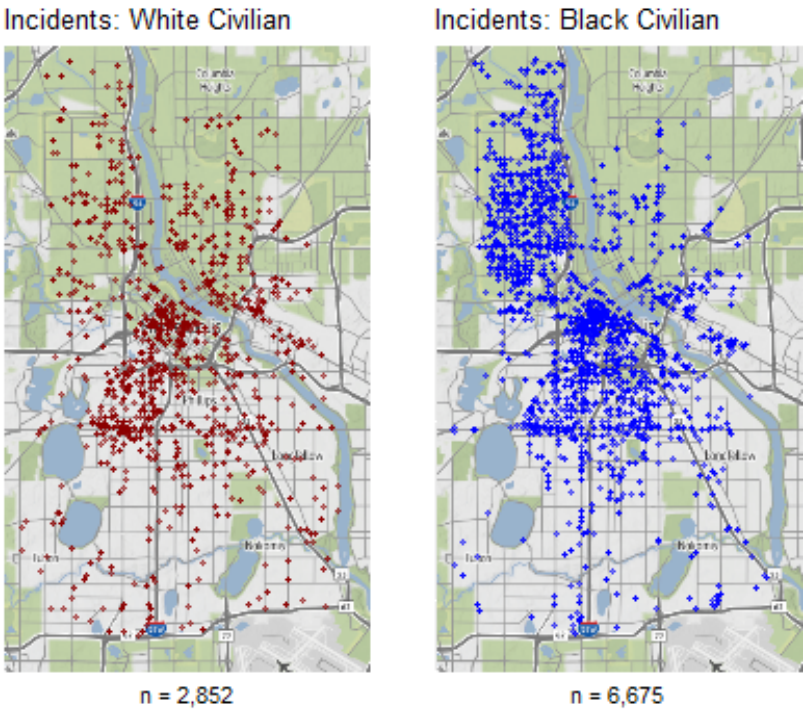


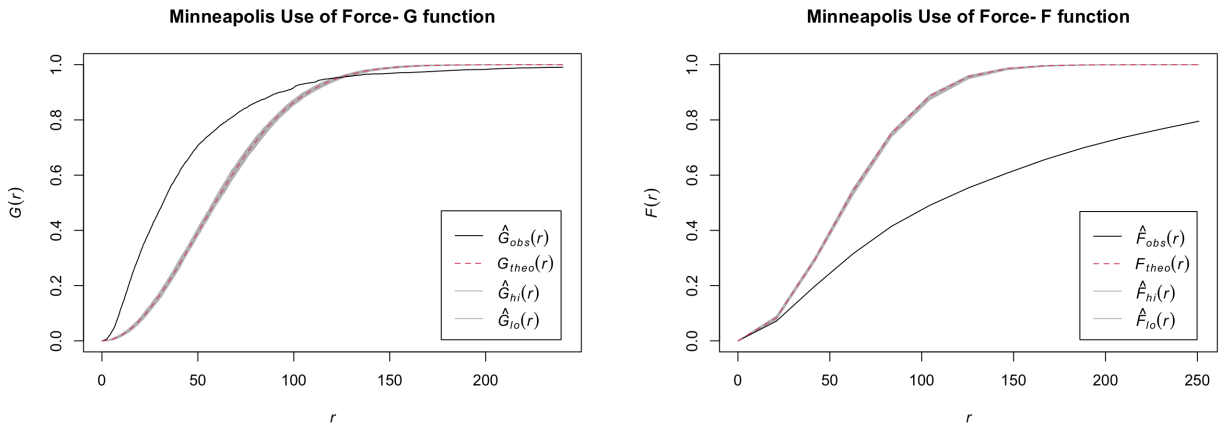
Figure 5: Maps of White vs. Black Civilian Use of Force Incidents

4 Results

The results for testing for CSR and from our application of the methods proposed by Quick et al. [2015] and Kelling and Haran [2022] are displayed below. We begin by reporting our results from the G and F functions to motivate a need for spatial modeling. Next, we display our results for the spatial intensity estimates from both the k-Groups analysis and the Two-Stage analysis. Then, we report our findings from our logistic regression model from the second stage of the Two-Stage method. Finally we compare the fundamental ways in which the results differ between the k-Groups method and the Two-Stage method.

4.1 G and F Functions

As discussed in Section 2.1, the G and F Functions are used to assess whether data exhibits CSR, which justifies a need for spatial modeling. Based on the G and F functions applied to the use of force data, illustrated in Figure 6, there is evidence against the null hypothesis that the police use of force data follows CSR towards the alternative that there is spatial clustering. For the G function in Figure 6a, the nearest neighbor distances between each point are smaller than we would expect to see under the CSR envelope created by 99 simulated datasets following CSR. For the F function in Figure 6b, the nearest nearest neighbor distances between an arbitrary point and the police use of force data are larger than we would expect to see under the CSR envelope created by 99 simulated datasets following CSR, suggesting empty space in the spatial domain. Because there is evidence against CSR for the police use of force dataset for both the G and F function, we are motivated to continue with spatial modeling.



(a) (b)

Figure 6: G and F Functions calculated for the use of force incidents (black line), compared to those generated under CSR (grey shaded region)

4.2 Intensity Estimation

As discussed in Section 2.3, the intensity function from an LGCP model measures how certain spatial variables influence the estimated spatial intensity of the event. The k-Groups model divides the data into k categorical groups based on nonspatial categorical variables, as explained in Section 2.4.1. The estimated intensity equations for each of the 4 k-Groups are in Equations 12, 13, 14, and 15. The notation WW corresponds to the white weapon group, WN corresponds to the white non-weapon group, BW correspods to the Black weapon group, and BN corresponds to the Black non-weapon group.

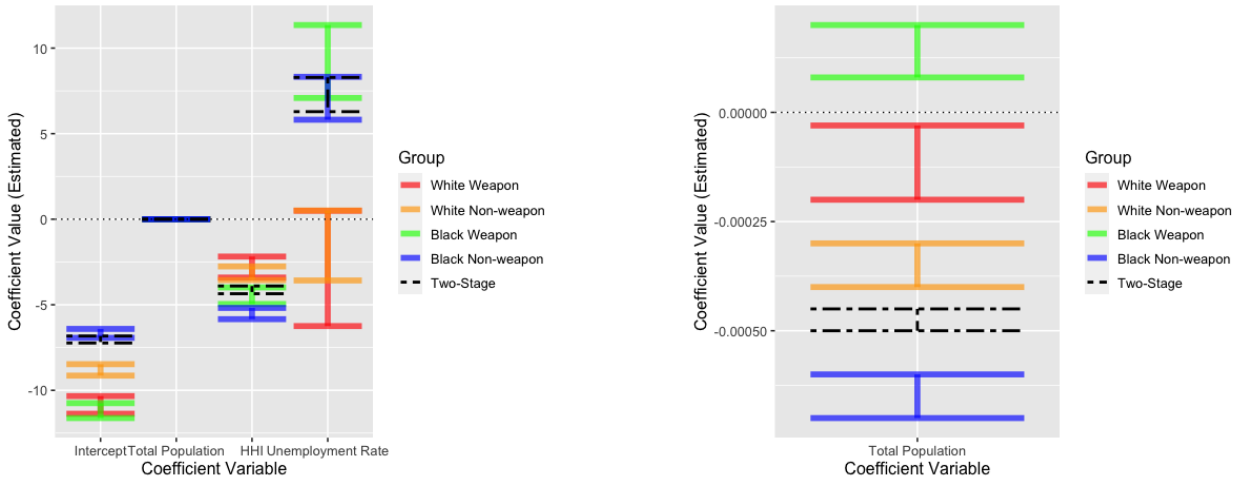
$$log(\lambda_{WW}) = -10.859 - 0.0001(TotalPopulation) - 2.792(HHI) - 2.782(UnemploymentRate) \quad (12)$$

$$log(\lambda_{WN}) = -8.811 - 0.0003(TotalPopulation) - 3.158(HHI) - 1.500(UnemploymentRate) \quad (13)$$

$$log(\lambda_{BW}) = -11.197 + 0.0001(TotalPopulation) - 4.464(HHI) + 9.268(UnemploymentRate) \quad (14)$$

$$log(\lambda_{BN}) = -6.671 - 0.0006(TotalPopulation) - 5.512(HHI) + 7.025(UnemploymentRate) \quad (15)$$

The estimated 95% credible intervals for each of the 4 groups in the k-Groups model are displayed in Figure 7. Beginning with total population, we see in the enlarged portion of the graph that all of the k-groups are estimated to have a negative total population coefficient except for the Black weapon group, which is somewhat surprising. For HHI, the credible intervals yield negative coefficient estimates for all groups. Finally, the credible intervals for unemployment rate yield positive coefficient estimates for both Black groups, while the intervals of both white groups contain 0 and as such we cannot be certain that the effect of this variable on these groups is nonzero. The exact values for the credible intervals can be found in Appendix C.



(a) (b)

Figure 7: Credible Intervals for k-Groups and Two-Stage

The estimated intensity across the spatial domain for each k-Group is plotted in Figure 8. There appears to be the least variability in intensity estimates in the white weapon group, and the most variability in intensity estimates in the Black non-weapon group. There are also similar patterns across groups with higher intensity in parts of Northwest and central Minneapolis, and lower intensity in parts of South, Northeast, and West central Minneapolis.

The estimated intensity equation for the first stage of the Two-Stage method is shown in Equation 16.

$$log(\lambda_{Two-Stage}) = -7.034 - 0.0005(TotalPopulation) - 4.128(HHI) + 7.317(UnemploymentRate) \quad (16)$$

The 95% credible intervals for the Two-Stage estimates - also displayed in Figure 7 - show a negative coefficient estimate for total population, a negative coefficient estimate for HHI, and a positive coefficient estimate for unemployment rate. The exact values for the credible intervals can be found in Appendix C. In comparison to the k-Groups credible interval estimates, the Two-Stage credible intervals tend to end up being closest to the Black non-weapon group, likely due to the group's larger proportional representation in the data.

The estimated intensity is plotted in Figure 9. Similarly to the k-Groups intensity estimates, there is higher intensity in the Northwest and central regions of Minneapolis, while there is lower intensity in the South, Northeast, and West central regions of Minneapolis.

(a)

(b)

(c)

(d)

Figure 8: Intensity Estimates Across Groups

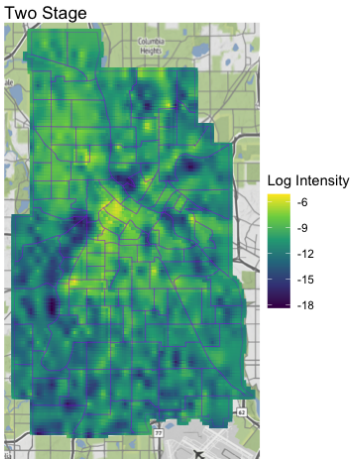


Figure 9: Intensity Estimate of All Points

4.3 Two-Stage Mark Determination

In order to gain insight into the impact of nonspatial variables in the Two-Stage method, we fit a logistic regression equation using both the spatial variables of interest and the non spatial variables of interest to model the probability of the force type being a weapon. The estimated equations of the second stage of the Two-Stage method for both white and Black civilians is shown in Equations 17 and 18, where p_W is the probability that a use of force incident involved a weapon (compared to no weapon) for white civilians, and p_B is the probability that a use of force incident involved a weapon (compared to no weapon) for Black civilians. In these equations, we see a negative estimate for the effect of total population on the log odds ratio of a use of force incident involving a weapon for a white civilian while we see a positive estimate for a Black civilian. For incidents involving either Black or white civilians, the second stage regression equations estimate a positive effect of both HHI and unemployment rate.

$$logit(p_W) = -1.03 - 0.00007(TotalPopulation) + 0.052(HHI) + 0.126(UnemploymentRate) \quad (17)$$

$$logit(p_B) = -1.12 + 0.00007(TotalPopulation) + 0.690(HHI) + 0.500(UnemploymentRate) \quad (18)$$

The confidence intervals for the second stage of the Two-Stage method (the logit portion) are shown in Table 2. In this table, we see that the regression coefficients are found to be statistically discernible for race, total population, and an interaction between the two. For both race and total population, the coefficient estimate is negative, while the estimate for the interaction is positive (where level 0 is equal to white).

Coefficient	**Confidence Interval**
Intercept	(-1.55, -0.51)
Race	(-1.44, -0.27)
Total Population	(-0.0001, -0.00002)
HHI	(-0.75, 0.63)
Unemployment Rate	(-2.65, 2.74)
Race × (Total Population)	(0.00005, 0.0002)
Race × (HHI)	(-0.17, 1.45)
Race × (Unemployment Rate)	(-2.50, 3.40)

Table 2: Confidence Intervals for Two-Stage Regression

4.4 Comparison of *k*-Groups vs Two-Stage Model

Both the k-Groups and Two-Stage methods can incorporate nonspatial variables into spatial models, but each have their strengths and limitations for implementation and inference. From our analysis, the k-Groups model allows us to directly compare the spatial intensities across categorical nonspatial variables. With these direct comparisons in the k-Groups model, the analysis allows us to determine how spatial variables affect the intensity of police use of force incidents across many categorical groups. For example, we can see that the total population positively impacts the intensity of police use of force incidents involving Black civilians and a weapon, but negatively impacts the intensity of police use of force incidents involving Black civilians and no weapon as well as white civilians (involving a weapon or no weapon). However, because the k-Groups method is not modeling a mark, it is thus limited to analyzing only the spatial intensity. Incorporating dependence between the k-Groups as described in Appendix A would allow us to further our analysis by acknowledging that the members of all groups are existing within the same spatial window.

Based on our analysis (where we assumed independence between stages), the Two-Stage analysis only allows us to determine how the spatial variables affect the spatial intensity of all the points. For example, we can only say that total population has a negative impact on the intensity of police use of force incidents without being able to distinguish between intensity estimates for differing races or force types. However, incorporating dependence between the two stages of the method, as discussed in Appendix A, would allow the findings from the spatial intensity stage and the mark determination stage to influence each other. Without dependence, it is not possible for the nonspatial information and spatial intensity to involve each other

because there is no relationship between the mark and estimated intensity. Unlike the k-Groups method, the Two-Stage method can model marks at spatial locations. For instance, total population has a positive effect on the probability that Black civilians are involved in an incident involving a weapon but a negative affect on the probability that white civilians are involved in an incident involving a weapon.

5 Discussion

The k-Groups method models the intensity of use of force incidents as a function of spatial variables differentiating by groups formed with combinations of categorical nonspatial variables. This method has wider credible intervals for coefficient estimates, but it is able to take into consideration the differences between the k categorical groups. The Two-Stage method models (1) the spatial intensity of use of force events (LGCP) and (2) the probability of our mark (weapon or non-weapon) influenced by a combination of our nonspatial variables, spatial variables, and their interaction. This method has narrower credible intervals than the k-Groups method when estimating the intensity, but inherently favors the largest portion of the location data, Black non-weapon.

The most notable part of our model parameter estimates across the five models is that the estimates differ in sign for total population. The estimate for the effect of total population on the intensity of use of force incidents is positive for the "Black, weapon" group, but negative for all other groups in the k-Groups method as well as the Two-Stage model. This implies that among use of force incidents involving a Black civilian and a police weapon, higher population is associated with higher incident intensity, while the opposite is true for all other models. Similarly intriguing are the parameter estimates for unemployment rate: while negative for both "white" groups, the estimates are positive for both "Black" groups, implying that among incidents involving a Black civilian, a higher unemployment rate is associated with higher intensity, unlike incidents involving white civilians. Further, it is important to note that the Two-Stage model agrees more with the parameter estimates of the "Black" groups, likely due to the higher proportion of Black civilians involved in use of force incidents. This is one instance where the Two-Stage model may not capture the nuances that the k-Groups method can using its categorical splits. The consistent negative HHI estimates in the k-Groups model estimates indicate that that among use of force incidents for all groups, more diverse census tracts are associated with higher incident intensity after controlling for other variables.

For the first stage of the Two-Stage model with all the points in the dataset, the total population estimates surprisingly indicate that as total populations increase in census tracts, we expect to see fewer police use of force incidents after controlling for other variables. Not surprisingly, HHI estimates indicate that as diversity increases in census tracts, we expect to see more police use of force incidents after controlling for other variables. Finally, unemployment rate estimates indicate that as unemployment rates increase in census tracts, we expect to see more police use of force incidents after controlling for other variables.

For the second stage of the Two-Stage model, we modeled the probability of the force type being a weapon compared to non-weapon based on our spatial and nonspatial variables with logistic regression. We found that there is a higher probability of the force type being a weapon if the incident involved a Black individual compared to white individuals after controlling for other variables. Additionally, we found an increase in total population results in an increase in the probability of the force type being a weapon compared to a non-weapon for Black individuals holding other variables constant. However, an increase in total population results in a decrease in the probability of the force type being a weapon for white individuals holding other variables constant.

5.1 Limitations

There are some limitations with regard to the data we used in this analysis. Beginning with the Minneapolis Police Use of Force data set, it is important to note that there were missing values present in multiple variables of interest, such as race or type of force. Next, the methods which were used to privatize the specific locations of the incidents are unknown, as the Minneapolis Police Department has not made these details publicly available. As a result, we jitter locations slightly so as to avoid violating assumptions of the point process model. More broadly, this policing data can be prone to human error; this analysis assumes that the use of force data is both accurate and reliable, an assumption which would benefit from further investigation.

Another concern regarding the policing data is the vagueness and subjectivity with which incidents can be classified. For instance, the force type variable has factor values such as "bodily force" and "maximal restraint technique", which can encompass a wide range of actions taken by an officer. Tied to this issue, the groups we selected for the k-Groups model were somewhat limited; as mentioned previously, our variable groupings limit incidents to only those where the civilian was either Black or white, and consolidates force types into a binary variable. The variables which make up the groups could be expanded upon either by including more levels at existing variables or including new variables.

Looking to the American Community Survey data, we note that due to the way the HHI is calculated - in combination with the categorization of "Hispanic" as an ethnicity and not a race - the HHI does not take into account some aspects of measuring diversity, at least in this specific application. Finally, it could be worth taking into account population density within census tracts instead of using total population, especially since some Minneapolis census tracts are smaller in size with a higher population, or some are larger in size with a smaller population. This could have contributed to the surprising results related to the total population variable because a higher total population measurement does not necessarily mean that a region has a higher density of people. Instead, the higher census measurement for total population could just be a result of a larger census tract.

5.2 Next Steps

Applying these methods to other police use of force data sets would be a logical next step. We looked into examining use of force for Atlanta, Chicago, and Dallas, but ultimately chose to focus on Minneapolis for this analysis. In the future, we would be interested to see how the relationship between police use of force events and nonspatial information varies across cities and jurisdictions. In addition, the inclusion of other variables - such as resistance type, median census income, or age of civilian - could prove informative.

Incorporating dependence between k-Groups intensity estimates and both parts of the Two-Stage method would also be important to investigate. When exploring real world applications, including these dependencies would allow for better fit models. For the k-Groups method, it is important to acknowledge that members of all k groups influence one another in the same spatial domain. For the Two-Stage method, it is important to note that the intensity stage and mark determination stage influence one another.

There are a multitude of other potential applications with integrating nonspatial information into spatial models. One application that has been explored is forest fire data. Kelling and Haran explore this application in their two stage paper by determining which elements can be helpful in deciding both how much area the fire burned as well as forest fire location [Kelling and Haran, 2022]. They found that the elements that influenced location and amount burned were not automatically the same [Kelling and Haran, 2022].

Another potential application is looking at the impact of nonspatial variables on cancer as done by Liang and Carlin [2008]. In the study, the researchers investigated if variables had different effects on two types of cancer by parameterizing their results of the bivariate mark model. Kelling and Haran [2022] also anticipate that the Two-Stage model could be applied to this example as well as many other types of data.

Finally, our application of police use of force incidents did not incorporate time into our analyses. Instead, the use of force incidents were analyzed as a cross-section from 2018-2023. Future studies could investigate how the spatial intensity of police use of force incidents change over a temporal domain.

Appendix A Incorporating Dependence

Quick et al. [2015] discusses how one might incorporate dependence between k-Groups into the model. They assert that one option is to specify $w_\lambda = (w'_{\lambda|1}, ..., w'_{\lambda|k})'$ such that $Cov(w_\lambda) = \Psi_\lambda \otimes C_\lambda$, with Ψ_λ representing covariance between surfaces and C_λ for spatial association [Quick et al., 2015]. Incorporating dependence leads to some computational challenges, but expands the types of questions the model can answer Quick et al. [2015]. For instance, the dependence incorporated in Quick et al. [2015] allows range parameters to vary and be estimated using the MCMC process, which could result in more accurate intensity estimates. Additionally, including a correlation between each of the k-Groups intensity surfaces accounts for the coexistence of members of all k groups in the same spatial domain W.

Kelling and Haran [2022] also discuss how to incorporate dependence between both stages of the proposed Two-Stage model. This would include two individual Gaussian Processes, one in each stage of the model. It also includes a cross-covariance structure to incorporate dependence between the two Gaussian Processes. The cross-covariance matrix of the bivariate Gaussian Process has dimension 2n x 2n (n is the number of points) and consists of both a univariate exponential correlation function as defined in Equation 19 (similar to Equation 5), as well as a matrix Λ defined in Equation 20. In the matrix Λ, if $|\rho|$ is close to 1, there would be strong dependence, and if $|\rho|$ is close to 0, there would be weak dependence. The full cross covariance matrix would be $\Sigma_{ij}\Lambda$.

$$\Sigma_{ij} = cov(\omega(s_i), \omega(s_j)) = exp\left(\frac{-|s_i - s_j|}{\phi}\right) \tag{19}$$

$$\Lambda = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{20}$$

Incorporating a bivariate GP is beneficial because it can produce a better fit model than when we assumed independence between the stages. It can introduce dependence in the spatial structure that determines where points occur and the mark at those locations. With a bivariate GP, one can interpret ρ; how close ρ is to 0 will indicate the strength of dependence between the two stages. The sign of ρ will indicate whether there is a positive or negative spatial structure in variables unaccounted for between the first and second stages.

Appendix B Choosing Integration Points and Knots

We chose the integration points for our application such that each census tract has random points generated with an amount proportional to its size. When estimating the integral with the integration points, we want each census tract to have some integration points, and we want the bigger census tracts to have more points than smaller census tracts. This will best find the average intensity over the region W. Overall, we used 9,999 integration points.

In terms of how we chose the knots, we used a grid format with 1,000 points evenly spaced across the window. We chose fewer knots compared to integration points due to the larger computational load involved with a larger number of knots. Also, because we use lower resolution points to estimate the Gaussian Process, a grid suffices.

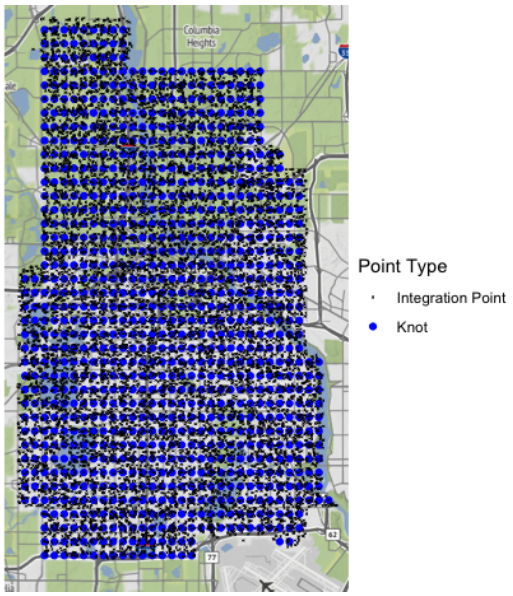


Figure 10: Integration Points and Knots

Appendix C Exact Credible Intervals

Group	Intercept	Total Population	HHI	Unemployment Rate
White, Weapon	(-11.370,-10.338)	(-.0002,-.00003)	(-3.4106, -2.1829)	(-6.2466,0.4942)
White, Non-weapon	(-9.1413, -8.4730)	(-0.0004, -0.0003)	(-3.5575, -2.7584)	(-3.5801, 0.5093)
Black, Weapon	(-11.623,-10.750)	(.00008,.0002)	(-4.9564, -3.9859)	(7.0858,11.3477)
Black, Non-weapon	(-6.9335,-6.4117)	(-0.0007,-0.0006)	(-5.8422,-5.1841)	(5.8140,8.3209)
Two Stage (LGCP)	(-7.2360,-6.8273)	(-0.0005,-0.00045)	(-4.3548, -3.9041)	(6.2855,8.2858)

Table 3: Credible Intervals for k-Groups and Two-Stage

References

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.

Roger S Bivand, Edze Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer New York, NY, 2 edition, 2013.

Census Data, 2021. URL `https://api.census.gov/data/2021/acs/acs5/variables.html`.

Perry de Valpine, Daniel Turek, Christopher Paciorek, Cliff Anderson-Bergman, Duncan Temple Lang, and Ras Bodik. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26:403–413, 2017. doi: 10.1080/10618600.2016.1172487.

Peter J Diggle. *Statistical Analysis of Spatial Point Patterns*. Edward Arnold, 2 edition, 2009.

Roland Fryer. An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy*, 127(3), 2019.

Amanda Geller, P.A. Goff, T. Lloyd, A. Haviland, D. Obermark, and J. Glaser. Measuring Racial Disparities in Police Use of Force: Methods Matter. *Journal of Quantitative Criminology*, 37:1083–1113, 2021.

Marina M Gorsuch and Deborah T Rho. Police stops and searches of Indigenous people in Minneapolis: the roles of race, place, and gender. *The International Indigenous Policy Journal*, 10(3), 2019.

Claire Kelling and Murali Haran. A two-stage cox process model with spatial and nonspatial covariates. *Spatial Statistics*, 51:100685, 2022.

Claire Kelling and Murali Haran. A Shared Component Point Process Model for Urban Policing. *arXiv preprint arXiv:2303.00206*, 2023.

Hoon Lee, Hyunseok Jang, Ilhong Yun, Hyeyoung Lim, and David W Tushaus. An examination of police use of force utilizing police training and neighborhood contextual factors: A multilevel analysis. *Policing: An International Journal of Police Strategies & Management*, 33(4):681–702, 2010.

Kim M Lersch, Thomas Bazley, Thomas Mieczkowski, and Kristina Childs. Police use of force and neighbourhood characteristics: An examination of structural disadvantage, crime, and resistance. *Policing & Society*, 18(3):282–300, 2008.

Shengde Liang and Bradley P Carlin. Analysis Of Minnesota Colon and Rectum Cancer Point Patterns With Spatial and Nonspatial Covariate Information. *The Annals of Applied Statistics*, 3(3):943–962, 2008.

Kyle Maksuta, Yunhan Zhao, and Tse-Chuan Yang. Race, disadvantage, and violence: A spatial exploration of macrolevel covariates of police-involved homicides within and between US counties. *Social Science Research*, 119, 2024.

Osagie Obasogie and Peyton Provenzano. Race, Racism, and Police Use of Force in 21st Century Criminology: An Empirical Examination. *UCLA Law Review*, 69, 2023.

Open Data, 2024. URL `https://opendata.minneapolismn.gov`.

Eugene Paoline, Jacinta Gau, and William Terrill. Race and the Police Use of Force Encounter in the United States. *The British Journal of Criminology*, 58, 2018.

Harrison Quick, Scott H Holan, Christopher K Wikle, and Jerome P Reiter. Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics*, 14:439–451, 2015.

Michael Smith, Rob Tillyer, and Robin Engel. Race and the Use of Force by Police Revisited: Post-Ferguson Findings From a Large County Police Agency. *Police Quarterly*, 26, 2023.

Michael R. Smith. Reimagining the Use of Force by Police in a Post-Floyd Nation. *Police Quarterly*, 25: 228–251, 2022.

Matthew Stephens. Introduction to Gaussian Processes: the OU covariance function, 2020. URL `https://stephens999.github.io/fiveMinuteStats/gaussian_process.html#matern_covariance_function`.

James Wright and Andrea Headley. Police Use of Force Interactions: Is Race Relevant or Gender Germane? *The American Review of Public Administration*, 50, 2020.