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

Claire Kelling
Department of Statistics
Pennsylvania State University
University Park, PA
cek32@psu.edu

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

Corina Graif
Department of Sociology and Criminology
Pennsylvania State University
University Park, PA
corina.graif@psu.edu

Murali Haran
Department of Statistics
Pennsylvania State University
University Park, PA
muh10@psu.edu

ABSTRACT

In order to understand the dynamics of crime in urban areas, it is important to investigate the socio-demographic attributes of the communities as well as the interactions between neighborhoods. If there are strong social ties between two neighborhoods, they may be more likely to transfer ideas, customs, and behaviors between them. This implies that not only crime itself but also crime prevention and interventions could be transferred along these social ties. Most studies on crime rate inference use spatial statistical models such as spatially weighted regression to take into account spatial correlation between neighborhoods. However, in order to obtain a more flexible model for how crime may be related across communities, one must take into account social proximity in addition to geographic proximity. In this paper, we develop techniques to combine geographic and social proximity in spatial generalized linear mixed models in order to estimate domestic and sexual violence in Detroit, Michigan and Arlington County, Virginia. The analysis relies on combining data from local and federal data sources such as the Police Data Initiative and American Community Survey. By comparing three types of CAR models, we conclude that adding information on social proximity to spatial models, we create more accurate estimation of crime in communities.

KEYWORDS

crime, spatial GLMM, social proximity, commuting data, police data initiative, American Community Survey, CAR

ACM Reference Format:
Claire Kelling, Gizem Korkmaz, Corina Graif, and Murali Haran. 2018. Combining Geographic and Social Proximity to Model Urban Domestic and Sexual Violence. In *Proceedings of ACM KDD Conference (KDD'18).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Crime in the US has been on a downward trajectory nationally since the mid 1990's but some cities, such as Detroit, Michigan, remain at much higher levels than others. Also, domestic and sexual violence have remained great problems. Studies show that intimate partner violence accounts for 15 percent of all violent crime [26], 1 in 3 female murder victims are killed by intimate partners [12], and 1 in 5 women in the United States is raped during her lifetime [6]. Victims of domestic and sexual violence experience increased risk of a range of physical and mental health problems, including depression and suicide [29]. Children exposed to domestic violence also suffer emotional and behavioral problems [43].

Urban crime is also known to concentrate in some neighborhoods more than others. The differential distribution of crime across neighborhoods within a city has been the focus of research since the early 1900's [36]. While the focus has been predominantly on internal socio-demographic forces, more recently, studies have highlighted the importance of geographic proximity to capture spatial dependencies [3, 4, 31]. However, social proximity has been emerging as a possibly significant factor as well [13, 25, 41]. For instance, research has begun to show evidence that co-offending ties and gang conflicts connect neighborhoods that are not necessarily geographically proximate [30, 35]. Moreover, people have been shown to move a great deal between non-nearby areas as a result of commuting for work, changing their home address, traveling for recreation, shopping, or other routine activities [7, 8, 15, 17, 23, 24, 28, 34, 42]. Less is known however about the consequences of such population mobility, though indications exists that access to external resources such as mortgage loans from outside people's neighborhoods may contribute to lower crime [40].

Drawing on this emerging evidence, in this paper we aim to understand the factors that affect crime rates focusing on interdependencies between neighborhoods based on geographic as well as social proximity. We use spatial statistical models to identify the significant community characteristics that may affect domestic and sexual violence by using incident-level crime data as well as local and federal data sources to incorporate socioeconomic characteristics of the neighborhoods. Socioeconomic variables such as unemployment rate and median income are collected from the American Community Survey (ACS), and commuting data from

KDD'18, August 2018, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACS are used to capture social proximity between neighborhoods defined by block groups from Census.

The first location that we choose for our study is Detroit, Michigan because crime is known to be a problem in Detroit, so there is potential for effective interventions. Detroit's "Crime Index" is listed as 2 (out of 100) on NeighborhoodScout.com. They say that Detroit is safer than just 2% of cities [2]. The second community we focus on is Arlington County, Virginia due to our engagement with county and the local government agencies including the police department. This allows us to identify issues that are of particular interest to them. Hence, the findings of our analysis could be used to impact their decision-making. Arlington was rated as safer than 40% of US Cities by NeighborhoodScout.com. The two communities we study in this manuscript are diverse enough to test the robustness and scalability of our approach.

In this paper, we rely on the Conditional Autoregressive (CAR) model for analyzing spatially aggregated or "areal" data. CAR models are popular in sociology, political science, epidemiology, and many other disciplines. We consider three classes of statistical models, all three of which fall under the category of spatial generalized linear mixed models (SGLMMs): the Besag-York Mollie (BYM) model, the Leroux model, and the sparse SGLMM. Each of the three models can capture geographical and social proximity, but do so in different ways. For each model, we consider two versions – one that allows for geographical proximity alone, and another that combines both geographic and social proximity. We find in all three cases that the combined model improves the fit and the performance of the statistical models, for both Arlington and Detroit.

We have studied, for the first time in this context, three different spatial modeling approaches that allow us to examine the impact of including the new social proximity measures in addition to more standard geographic proximity measures. Our study pushes the criminology and sociology fields to rethink the social processes and mechanisms that affect crime beyond a predominant focus on internal or spatially proximate processes and highlights the significance of inter-connectivity of ecological units that have previously been treated as if they were isolated islands. Moreover, our study makes key methodological contributions by integrating a novel measure of social proximity using commuting data provided by US Census Bureau.

The paper is organized as follows. In the next section, we summarize the relevant studies that focus on crime. Section 3 provides the modeling framework and presents the statistical models used in this study. Section 4 summarizes the data sources, and in Section 5 we present the results describing the metrics used to compare the models as well as to evaluate the fitness. Section 6 concludes.

2 RELATED WORK

Recent literature on crime has focused on modeling the diffusion of crime between neighborhoods or communities [41]. Neighborhoods have been traditionally treated as closed systems where there is no transfer of ideas or behaviors inside of a city once you leave some geographic bound. We know this is not the case, as there is a national and international transfer of people, attitudes, and customs due to the modern ease of travel and communication. Some people may become victimized outside their home area as a result of routine travel and activities, an idea that gained attention under the routine activities theory [14]. Similarly, crime pattern theory suggests that the places where people live, work, and play constitute activity nodes and paths that increases one's odds to become involved in crime as a victim or offender [9, 10]. Evidence shows that a majority of violent crimes (e.g., about 70 percent of homicides in Pittsburgh) occur outside of the neighborhood of residence of the involved victim or offender [18, 39]. Residents' travel patterns to work, shopping, and recreation, were found to be positively correlated with violent and property crime [15, 38].

We are therefore 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" [27], where certain neighborhoods are predominantly targeted, mainly based on race and other demographic information. In our modeling framework, we use the Herfindahl Index as a metric for diversity in order to avoid these controversial policing strategies.

We are interested in investigating what other ties, such as social, might exist between neighborhoods, in order to create more accurate estimation of crime and more effective policing interventions. Social proximity is a relatively new topic in the criminology literature and it is a new technique to combine social proximity and geographic proximity. Wang et al.[41] use taxi data to establish social proximity between neighborhoods for crime rate inference in Chicago. They establish links between areal units where people get in and out of a taxi. We will use commuting data, available through the Census, to establish social proximity. We believe that commuting data may provide a more reliable way to estimate social proximity. Commuting data is almost always symmetric- if someone goes to work from block group A to B, they will come home from B to A, which reflects a symmetric transfer of information. With taxi trips, travelers may be going from work to a restaurant or from bar to bar which may not illustrate as strong or as consistent of a social tie as commuting.

We compare models that incorporate purely geographic proximity with models that incorporate both social and geographic proximity, to evaluate whether commuting data adds information about ties between communities that may affect crime. Wang et al.[41] also use Point-of-Interest (POI) data (i.e., relevant venues such as schools, hospitals, churches that provide information about neighborhood functions), and use linear regression models for crime rate inference. Our model also combines social and geographic proximity but through the Bayesian CAR model structure.

3 PROPOSED METHODS

Our goal is to develop a model that combines geographic and social proximity to provide a more flexible way to analyze how crime may be related across communities. Typical models rely on geographic proximity but do not incorporate other ways that communities may be alike rather than just sharing a border.

We consider three statistical models, described below, and compare their results when including geographic proximity with their results when simultaneously account for both geographic proximity and social proximity. These are the BYM, Leroux, and sparse spatial generalized linear mixed model (SGLMM); the sparse SGLMM controls for potential spatial confounding present in the BYM and Leroux models. By considering these different models we also demonstrate the robustness of our approach to a host of different assumptions.

3.1 Modeling Framework

In this framework, we assume that our study region $\mathcal{S}$ is partitioned into K non-overlapping areal units. In our study, the areal units correspond to the 1,822 block groups in Detroit and 181 block groups in Arlington. The aggregated crime counts for each block group are represented by the set of responses $\mathbf{Y} = (Y_1, ..., Y_K)$. Spatial variation in the response is modeled using a matrix of covariates $\mathbf{X}_{p \times K} = (x_1, ..., x_K)$ and a spatial structure component ψ = $(\psi_1, ..., \psi_K)$ that accounts for any spatial autocorrelation that is not captured by the covariate effects.

We adopt the Conditional Autoregressive model (CAR) model [22] for analyzing areal unit data. We will use particular specifications of the CAR model, BYM, Leroux, and sparse SGLMM, which are described in detail below. These models are a special case of a Gaussian Markov Random Field (GMRF) [22]. We use the implementations of the BYM and Leroux models in the CARBayes package in R and the implementation of the sparse SGLMM model in the ngspatial package in R [19, 22].

We use a Generalized Linear Mixed Model (GLMM) to fit our spatial areal unit data, given in Equation 1 below.

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

The parameters Y_k, x_k, and ψ_k represent the response, covariates, and spatial structure component, respectively, as defined above. The expected value of Y_k, $E(Y_k)$ is denoted in the model as μ_k. The vector of regression parameters, denoted by $\beta = (\beta_{\mathbf{1}}, ...\beta_{\mathbf{p}})$, is assumed to have a multivariate Gaussian prior distribution. The parameters of this covariance matrix, μ_β and Σ_β are specified in the CARBayes package. We use the default for μ_β as a zero-mean vector and Σ_β as a diagonal matrix, where the diagonal elements are 100,000. As our response is count data over the areal units, we use a Poisson form of the GLMM. We assume that $Y_k \sim Poisson(\mu_k)$ and $ln(\mu_k) = \mathbf{x_k^T} \beta + \psi_{\mathbf{k}}$.

In order to model spatial correlation, we define a neighborhood matrix, $\mathbf{W}$, which is a non-negative, symmetric, K×K matrix, where K is the number of areal units. In each of the model specifications below (BYM, Leroux, and sparse SGLMM), the distribution of ψ parameter depends on the form of the neighborhood matrix. We denote the (k,j)th element of the neighborhood matrix by w_{kj}, which represents spatial closeness between areas $(\mathcal{S}_k, \mathcal{S}_j)$. The diagonal elements of this matrix, w_{kk}, are always 0. Positive values of w_{kj} indicate geographical closeness, and $w_{kj} = 0$ indicates non-closeness. The most commonly used structure to indicate closeness is $w_{kj} = 1$. In fact, the current framework of many of the statistical models, such as the sparse SGLMM in the R package *ngspatial*, only allow for binary elements in the neighborhood matrix.

Detroit

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

Arlington

Variable	Geographic	Social	Geo. & Soc.
Number of regions	173	142	173
Number of nonzero links	982	524	1436
Percentage nonzero weights	3.28	2.60	4.80
Average # of weights	5.68	3.69	8.30

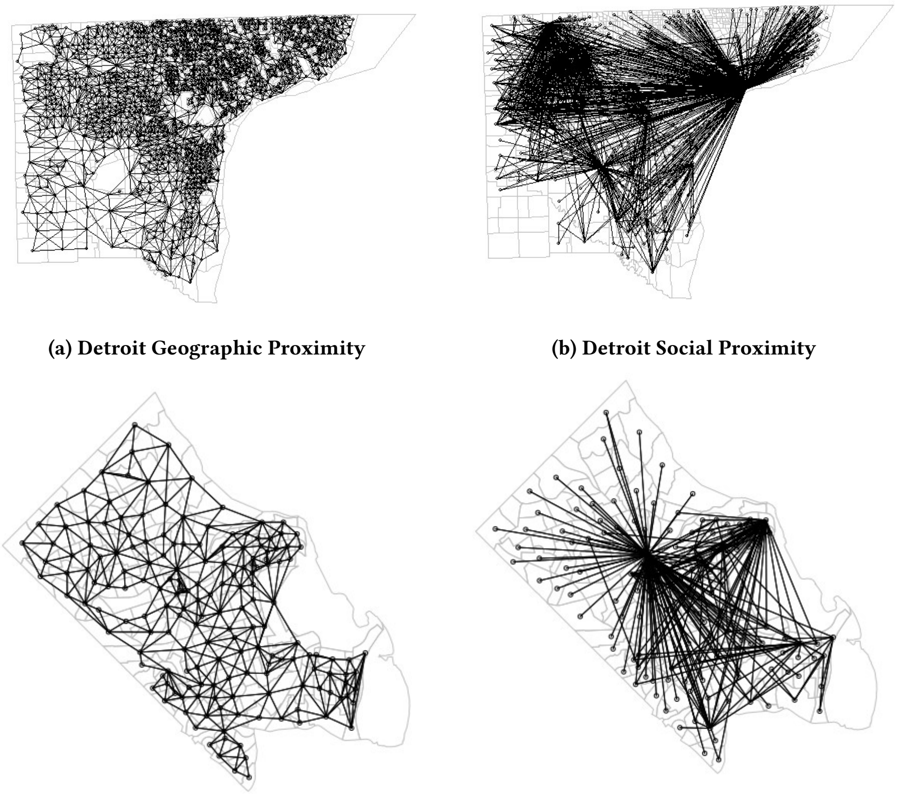
Table 1: Descriptive Statistics of Neighborhood Structures

For our purposes, we let $w_{kj} \in \{0, 1\}$, where $w_{kj} = 1$ indicates that two areal units, k and j, are neighbors and $w_{kj} = 0$ otherwise. In our future work, we will consider other values of w_{kj} that will define proximity between areas that are not adjacent.

In order to combine social and geographic proximity, we create a neighborhood matrix, $\mathbf{W}$, for both the geographic and social proximity. For geographic proximity, we define a neighbor as any block group that shares a border. For social proximity, we define a neighbor as any block group where there is a certain level of commuting between two block groups. We recognize that commuting is technically a directed activity but we treat it as undirected. We believe that regardless if it is the origin or the destination, it is still establishing a tie between the two block groups, as people will return home after work. Therefore, this creates a symmetric neighborhood matrix. For the purposes of our model, we define a cutoff for the number of commuters in order to establish meaningful social proximity ties between two block groups. In our case, we define the cutoff to be 15 commuters between two block groups in order for a tie to be defined for Detroit. This includes about 0.89% of the commuting ties. In Arlington County, we define the cutoff to be 10 commuters, which makes up about 3.54% of the commuting ties.

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

In Figure 1, we see the neighborhood structures depicted on the map of the block groups of Detroit and Arlington County. 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, we can see that some folks are commuting across the counties, as there are clearly social hubs in some areas of the county where many people commute to/from.



(a) Detroit Geographic Proximity

(b) Detroit Social Proximity

(c) Arlington Geographic Proximity

(d) Arlington Social Proximity

Figure 1: Plot of Neighborhood Structure

3.2 BYM Model

The first model we consider for ψ_k, the spatial random effects that capture the spatial dependence among the observations, is based on the CAR model proposed by Besag, York, and Mollié (BYM) [5]. It is also called the intrinsic CAR model and has been very widely used, especially in disease mapping. In this model, there are two sets of random effects, spatially autocorrelated and independent. The full model specification in the Bayesian framework is given in Equation 2 below. This model is equivalent to the multivariate specification $\phi \sim N(0, \tau^2 \mathbf{Q}(\mathbf{W})^{-1}),$ where $\mathbf{Q}(\mathbf{W}) = \text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}$.

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

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

3.3 Leroux Model

We adopt the model presented by Brian Leroux et al. [11] as our second model for spatial random effects (ψ_ks). This model provides improved parameter interpretability, particularly of the variance parameters of the spatial random effects. The Bayesian model specification is given in Equation 3.

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

The Leroux model uses only a single random effect, ϕ_k, and incorporates another spatial autocorrelation parameter, ρ, and 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 $\mathbf{Q}(\mathbf{W}, \rho) = \rho[\text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}] + (\mathbf{1} - \rho)\mathbf{I}$. This version of the CAR model has been widely accepted to be the most appealing CAR model, from both theoretical and practical standpoints [21].

3.4 Sparse SGLMM

Finally, we consider the sparse SGLMM model [19]. This model addresses potential spatial confounding issues – spatial confounding is the phenomenon by which the spatial random effects act as if they are multicollinear with the covariates ("fixed effects", in our case the demographic variables). As in standard regression, this multicollinearity can impact our ability to interpret the regression coefficients (β's). This phenomenon was described in Reich et al.. For instance, in the model parameterization given in Equation 1, we observe that $g(\mu_k) = x_k^T \beta + \psi_k$. Reich et al. [32] note that if we **P** is the orthogonal projection onto the regression manifold C(**X**), and we construct the eigendecomposition of **P** and **I**-**P** to obtain orthonormal bases, such as $\mathbf{K}_{n \times p} and L_{n \times (n-p)}$, for C(**X**) and C(**X**)$^{\perp}$, then Equation 1 can be rewritten as $g(\mu_k) = x_k^T \beta + k_i' \gamma + l_i' \delta$ where $\gamma_{p \times 1}$ and $\delta_{(n-p) \times 1}$ are random coefficients. This shows how K is the source of the confounding as it is multicollinear with the covariates. Hughes and Haran resolve this issue by removing K from the model while also reducing the dimensionality of the spatial random effects for computing efficiency. An implementation of this model is available in the R package ngspatial [19].

4 DATA

In this section, we describe the data sources used in our statistical models. We use crime data from the Police Data Initiative for Detroit, and data provided by Arlington County Police Department. In order to incorporate socioeconomic characteristics of the neighborhoods, as well as the commuting behavior, we use data provided by the US Census Bureau. We describe the data sources in detail below.

4.1 Crime Data

To be able to use the models described above, with a Poisson distribution, we use crime count data for both Arlington and Detroit. With the ACS demographic information, we incorporate total population into our models to ensure that crime frequency is not confounded with total population.

4.1.1 Detroit: Police Data Initiative. The crime data for Detroit, Michigan was provided by the Police Data Initiative (PDI) [1]. This is a recently popularized data source, used as a dataset for a Kaggle-like challenge by the American Statistical Association. It was created by the Police Foundation to support academic research. They encourage folks in the community to use their data in order to create more effective relationships between law enforcement and local citizens. There are several datasets hosted by the PDI, including data on accidents/crashes, community engagement, officer-involved shootings, and complaints. In this study, we focus on the Calls for Service (CFS) which include the individual 911 calls.

For Detroit, the dataset includes 566,553 crimes from September, 2016 to November, 2017. There is a location (lat/lon) associated with each call, as well as some basic information on crime type and the priority of the call. After taking a subset of the data for domestic/sexual violence related calls, we have 5,679 incidents for this time interval.

For Detroit, we use call description codes such as "RAPE IP or JH" (rape in progress or just happened) or "ASSAULT OR SEX ASSAULT DELTA", where delta indicates a high-severity crime (on the range alpha to echo). Figure 2 (top) illustrates the number of domestic violence crimes in Detroit by block group. We observe that there is a large peak in crimes in the northeast part of Detroit. Most of the city is pretty evenly distributed other than this peak, and many block groups, 912 out of 1822, have no domestic/sexual violence crimes reported during this time.

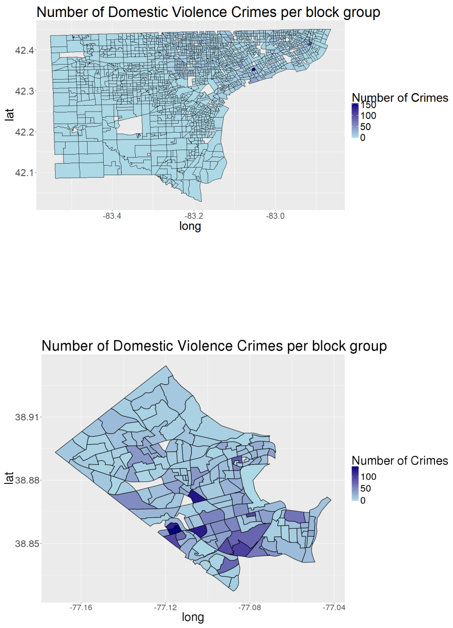


Figure 2: Number of domestic violence incidents by block group in (top) Detroit, and (bottom) Arlington County

4.1.2 Arlington: Arlington County Police Department (ACPD). For Arlington County, we use CFS data provided by ACPD for 2005-2015. The dataset includes a total of 1,064,099 Calls for Service (CFS) and 261,944 incidents, and detailed information such as the time and location (lat/lon) of the call, type of the call and the incident (could be different), responding (dispatched) units, individuals involved. We focus again on the domestic/sexual violence incident types, including "ASSAULT FAMILY, DOMESTIC FAMILY DISTURBANCE" or "RAPE-GUN". For our analysis, we use the full range of dates available, from 2005 to 2015. The total number of domestic/sexual violence crimes in this time interval is 3,856. There is just 1 block group, out of 181, with no crimes present in this time period. The volume of domestic violence incidents in Arlington County by block group is illustrated in the bottom panel of Figure 2.

4.2 Socioeconomic Characteristics: ACS

We use the American Communities Survey (ACS) to obtain socioeconomic information for all of the block groups in Detroit (Michigan) and Arlington County (Virginia). ACS is a national survey that is released every few years through the US Census Bureau and includes variables such as income information to migration information. The block groups for Detroit and Arlington County can be seen in Figure 1. In Detroit, there are 1,822 block groups, of which only 1,706 have complete publicly available ACS data. For Arlington County, 173 of the block groups (out of 181) have publicly available data. The data of the remaining block groups are not released due to concerns of identifiability, a common practice with ACS. The crime rates in these block groups are not high and represent a small portion of the data, so this is not a large concern for our analysis, though this is something we plan to address in future work. We use the 2015 ACS data in order for our data to be representative of the demographics present with the crime data.

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

We calculate the *Herfindahl Index (HI)* [33] using the set of variables on race. HI 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. 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 use HI, shown in equation 4, as a measure of diversity in the block group.

$$HI = 1 - \left(\left(\frac{\text{\# white}}{\text{total pop}} \right)^2 + \left(\frac{\text{\# black}}{\text{total pop}} \right)^2 + \left(\frac{\text{\# Asian}}{\text{total pop}} \right)^2 + \ldots \right) \quad (4)$$

where $HI = 0$ indicates that the block group is composed of one racial category (i.e., no diversity). The higher the index, the more diverse the block group is.

4.3 Social Proximity: LODES

Finally, we incorporate data on commuting to capture social proximity using LODES (Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics) provided by US Census Bureau. This data is also used as the source for the Census app OnTheMap. It shows how many people are commuting into and out of a given geographical area, and how many stay in the area for work. Using this data source, we obtain commuting data for all block groups in Detroit and Arlington County, as well as for the respective states. In this study, we focus on commuting behavior within the boundaries of these counties and exclude the commuters that are leaving the counties from the analysis. We plan to incorporate inter-county commuting in future work.

5 RESULTS

In this section, we explain our metrics for model comparison, methods for assessing for spatial autocorrelation, and the results of our models. We give details about how adding social proximity improved the model fit and provide evidence on the lack of spatial confounding.

	Detroit		Arlington	
Proximity Type	Moran's I	p-value	Moran's I	p-value
Geographic	0.22656	0.000999	0.035268	0.1738
Geo. & Soc.	0.25188	0.000999	0.053927	0.04496

Table 2: Moran's I and p-value

5.1 Metrics

In order to compare our models, we adopt the *Deviance Information Criterion **(DIC)***, and the *Percentage Deviance Explained **(PDE)*** that are commonly used for model comparison in a Bayesian framework.

The Deviance Information Criteria (DIC) is a measure that combines the "goodness of fit" of a model and its "complexity" [37]. We measure the fit via the deviance, where $D(\theta) = -2logL(data|\theta)$. Complexity is measured by the estimate for effective number 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 5 below.

$$\text{DIC} = D(\bar{\theta}) + 2p_D = \bar{D} + p_D \quad (5)$$

A model with a smaller DIC is better supported by the data than a model with a larger DIC, similar to AIC that is commonly used in model comparison.

5.2 Spatial Autocorrelation Assessment

Before fitting our models, we want to provide evidence of spatial autocorrelation in the data, which necessitates spatial modeling, based on our two neighborhood matrices. We use *Moran's I*, a common method for assessing spatial processes, which is calculated as:

$$\text{I} = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{ij} w_{i \neq j})(Y_i - \bar{Y})^2} \quad (6)$$

When spatial association is large, the ratio (Moran's I) is large. 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 [16]. Therefore, for spatial modeling to be appropriate, we would like the p-value to be less than an α of 0.05.

Table 2 summarizes Moran's I as well as the p-values for this test. For Detroit, we obtain p-values are less than 0.05, indicating that spatial autocorrelation is present, using the neighborhood matrices for both geographic and combined proximity models. However, the Moran's I statistic for Arlington is statistically significant only for the combined model. Moran's I depends on the residuals of the linear model, which accounts for total population. Therefore, there is no concern for confounding of crime count and population.

5.3 Model Comparison Results

In this section, we compare the six models – the BYM model, Leroux Model and sparse SGLMM model, all for both geographic and combined (geographic and social) proximity – using the model comparison criteria outlined in Section 5.1.

Table 3 summarizes the comparison of the models, using DIC and Percentage Deviance Explained for both Detroit and Arlington. For Detroit, we observe that when we combine social and geographic

Detroit

Criterion	Geo. BYM	Combined BYM	Geo. Leroux	Combined Leroux	Geo. Sparse SGLMM	Combined Sparse SGLMM
DIC	8923.92	8845.05	8029.72	7005.55	8522.32	8251.78
PDE	56.93	61.01	63.91	68.12	38.96	58.63

Arlington

Criterion	Geo. BYM	Combined BYM	Geo. Leroux	Combined Leroux	Geo. Sparse SGLMM	Combined Sparse SGLMM
DIC	1149.64	1147.77	1144.86	1138.02	1725.10	1614.36
PDE	81.64	81.68	81.52	81.56	53.94	52.84

Table 3: Model Comparison

proximity and compare it to the results of just the geographic proximity, there is an improvement in every model in terms of percentage deviance explained and DIC. All of the models that incorporate both geographic proximity and social proximity have higher percentage deviance explained and lower DIC, indicating a better model fit.

For Arlington, we see that the fits are relatively similar between geographic proximity alone and geographic proximity with social proximity. There is a consistent decrease in DIC across all models. The percent deviance explained does not change much over all three model types, and there is a mix of small increases and decreases in the percent deviance explained. We believe that adding geographic proximity provides a slightly better model fit for Arlington County.

Therefore, based on the model fit results, we conclude that adding meaningful commuting ties between block groups improves the model fit as compared to only using geographic proximity.

We also compare model performance between model types. For Detroit, in the case of just geographic proximity, the Leroux model outperforms both the BYM model and the sparse SGLMM model. For the sparse SGLMM model in the geographic case, there is actually a quite low percentage deviance explained. In the case where we combine social and geographic proximity, we see once again that the Leroux model outperforms the BYM and sparse SGLMM models in terms of DIC and percentage deviance explained. For Arlington County, in both cases of geographic proximity alone and with combined proximities, we see that there are only minor differences between the BYM and Leroux models. The BYM models have slightly higher percent deviance explained and slightly higher DIC's. The improvement of the DIC from BYM to Leroux is greater than the improvement of the percent deviance explained from BYM to Leroux, so we conclude that the Leroux is a better fit for Arlington as well.

5.4 Fitted Values

Next, we compare the models to see the fitted values compared to the actual observed counts per block group. Figure 3 provides the fitted values for all six models. For Detroit, 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 sparse SGLMM models are predicting much lower values compared to the actual number of crimes, shown in Figure 2. Both of the Leroux models are picking up on the couple of block groups that have high rates of domestic/sexual violence, where these are not as accurately captured in the BYM models.

For Arlington, the fitted values for the BYM and Leroux models look quite similar to the actual number of crimes in the block groups, shown in Figure 2. The sparse SGLMM model did not pick up on the peaks as well as the BYM and Leroux models. Similar to the model comparison metrics used earlier (DIC and PDE), we do not notice a large difference in the fitted values between geographic alone and the combination of geographic and social proximity. However, due to the improvement in DIC, we conclude that adding social proximity provides a better model fit.

5.5 Coefficients

In addition to the accuracy of the estimation of crime and the fit of the model, we are also interested in the estimates of the coefficients of the covariates, the demographic information we have used in our model. Specifically, we are interested to see if there are any problems with spatial confounding in our case. To assess this, we will see if the estimates for the coefficients and their confidence intervals are similar between the BYM/Leroux Model and the sparse SGLMM. We will focus on the comparison between the Leroux model and the sparse SGLMM, the latter of which controls for spatial confounding. The presented coefficients are for the Arlington data analysis and we find a similar trend for Detroit.

The estimated values of the posterior median as well as the 95% credible intervals for the model parameters are in Table 4 for the combined Leroux Model and in Table 5 for the combined sparse SGLMM model. Our covariates are all of the demographic variables that we collected from Census through the American Communities Survey.

Based on Tables 4 and 5, we see that there is not a large difference between the combined Leroux Model and the combined sparse SGLMM model coefficients. Therefore, spatial confounding does not appear to impact our estimates. The models agree that median income has a very small coefficient- too small for an exact estimate to be provided for the Leroux model. Both models show an insignificant estimate for the coefficient of the unemployment rate and percentage male, which we use to show gender diversity. They both show there is a positive estimate for the Total Population, that is quite similar in fact between the two models. This positive relationship makes intuitive sense, as we may expect crime to increase with the total population. There is also a positive estimate for the coefficient of the Herfindahl Index. In the Combined Leroux model, the

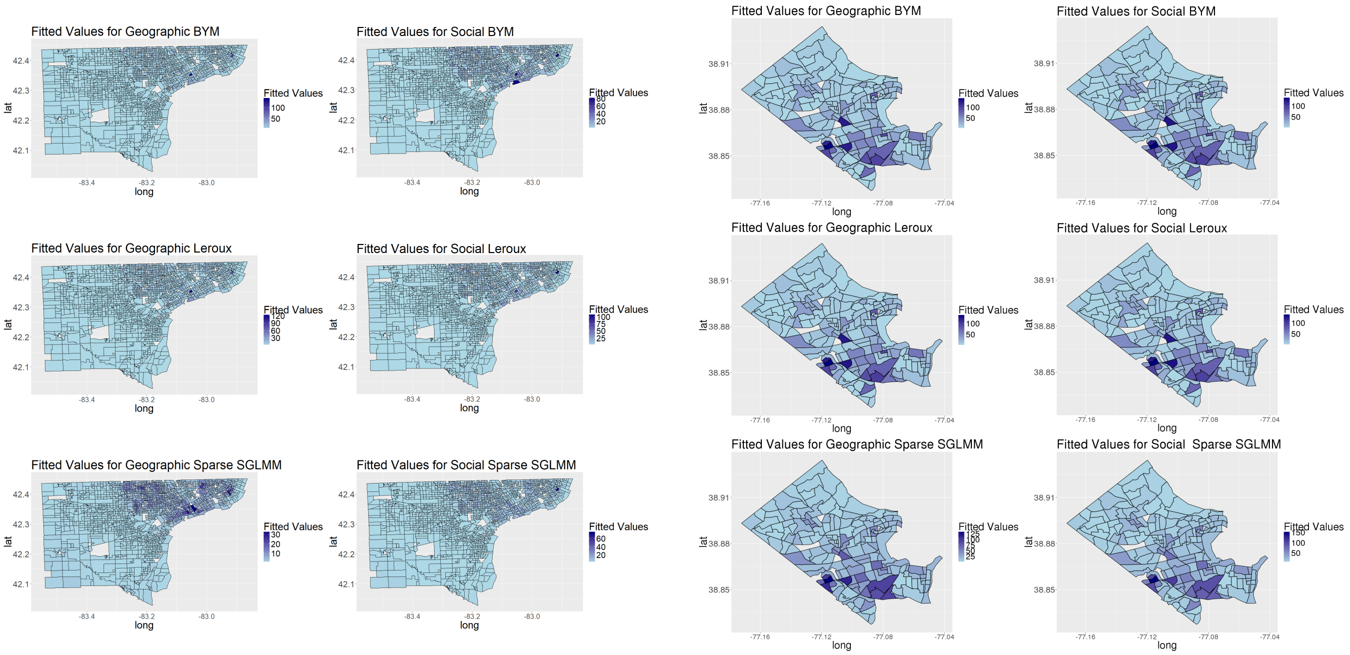

Fitted Values for Geographic BYM
Fitted Values for Social BYM
Fitted Values for Geographic Leroux
Fitted Values for Social Leroux
Fitted Values for Geographic Sparse SGLMM
Fitted Values for Social Sparse SGLMM
Fitted Values for Geographic BYM
Fitted Values for Social BYM
Fitted Values for Geographic Leroux
Fitted Values for Social Leroux
Fitted Values for Geographic Sparse SGLMM
Fitted Values for Social Sparse SGLMM
Fitted Values
lat
long
42.4
42.3
42.2
42.1
-83.4
-83.2
-83.0
38.91
38.88
38.85
-77.16
-77.12
-77.08
-77.04

Figure 3: Fitted Values, All Models

credible interval includes 0 whereas in the sparse SGLMM model, the credible interval does not include 0 but is quite close to 0. In summary, we believe that for our data, spatial confounding is not a major concern.

	median	credible interval	
Term	0.5	0.025	0.975
Intercept	1.6395	-0.2839	3.5849
Median Income	0.0000	0.0000	0.0000
Unemployment Rate	-0.1647	-1.3830	0.9723
Total Population	0.0005	0.0002	0.0007
Percentage Male	0.1976	-1.5813	2.0374
Median Age	-0.0018	-0.0266	0.0205
Herfindahl Index	2.2455	1.1624	3.5081

Table 4: Combined Leroux

	median	credible interval	
Term	0.5	0.025	0.975
Intercept	2.9514	2.2451	3.7022
Median Income	-3.037e-06	-4.542e-06	-1.577e-06
Unemployment Rate	-0.4852	-0.9795	0.02126
Total Population	0.0003	0.0002	0.0004
Percentage Male	-0.0111	-0.0708	0.4846
Median Age	-0.0094	-0.0182	-0.0003
Herfindahl Index	1.4943	1.1172	1.8672

Table 5: Combined Sparse SGLMM

6 CONCLUSION

The social science literature has typically modeled neighborhoods and crime focusing on social forces that are predominantly internal or spatially proximate. We demonstrate the value of predicting crime using a new feature that measures external social forces, based on social rather than spatial proximity and operationalized though commuting. The findings contribute conceptually and methodologically to the literature by highlighting the significance of inter-connectivity among ecological units. Using the social proximity feature is important because the employment-based mobility of urban residents may diffuse norms or resources that affect local crime risk. This study makes key methodological contributions by integrating for the first time, publicly available data on commuting and a new measure of social proximity, with methods from spatial statistics to improve crime prediction. Specifically, by adding information on social proximity to spatial models, we create more accurate estimation of crime in communities. By using multiple modeling frameworks, we have demonstrated the robustness of our conclusions to a variety of different assumptions about the underlying spatial dependence, including to potential confounding between the spatial random effects (representing spatial dependence due to geographic or social proximities) and fixed effects (demographic covariates). Still, the usual caveats are necessary here, namely, that the data sets we use are observational and there is always the possibility that unmeasured covariates could impact conclusions drawn from such a study. Despite limitations, the methods we used here, which combine geographic and social links, will be useful beyond modeling crime, to examine more generally the spread of influence through social and ecological networks.

7 ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under IGERT award DGE-1144860, "Big Data Social Science" and the Data Science for Public Good (DSPG) Program at the Social and Decision Analytics Laboratory at the Biocomplexity Institute of Virginia Tech. The authors also acknowledge the support of the Arlington County Police Department. Graif also acknowledges grant support from the National Institute of Health (NIH 1K01-HD093863-01) and from the Penn State University's Population Research Institute (NICHD R24-HD041025).

REFERENCES

[1] [n. d.]. Datasets. ([n. d.]). https://www.policedatainitiative.org/datasets/

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

[3] Luc Anselin. 2002. Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural economics* 27, 3 (2002), 247–267.

[4] Luc Anselin, Jacqueline Cohen, David Cook, Wilpen Gorr, and George Tita. 2000. Spatial analyses of crime. *Criminal justice* 4, 2 (2000), 213–262.

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

[6] Michele C Black, Kathleen C Basile, Matthew J Breiding, Sharon G Smith, Mikel L Walters, Melissa T Merrick, Mark R Stevens, et al. 2011. The national intimate partner and sexual violence survey: 2010 summary report. *Atlanta, GA: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention* 19 (2011), 39–40.

[7] Rémi Boivin and Maurizio D'Elia. 2017. A Network of Neighborhoods: Predicting Crime Trips in a Large Canadian City. *Journal of research in crime and delinquency* 54, 6 (2017), 824–846.

[8] Rémi Boivin and Marcus Felson. 2017. Crimes by Visitors Versus Crimes by Residents: The Influence of Visitor Inflows. *Journal of Quantitative Criminology* (2017), 1–16.

[9] Patricia Brantingham and Paul Brantingham. 1995. Criminality of place. *European journal on criminal policy and research* 3, 3 (1995), 5–26.

[10] Patricia L Brantingham and Paul J Brantingham. 1993. Nodes, paths and edges: Considerations on the complexity of crime and the physical environment. *Journal of Environmental Psychology* 13, 1 (1993), 3–28.

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

[12] F Stephen Bridges, Kimberly M Tatum, and Julie C Kunselman. 2008. Domestic violence statutes and rates of intimate partner and family homicide: A research note. *Criminal Justice Policy Review* 19, 1 (2008), 117–130.

[13] Christopher R Browning, Catherine A Calder, Bethany Boettner, and Anna Smith. 2017. Ecological Networks and Urban Crime: The Structure of Shared Routine Activity Locations and Neighborhood-Level Informal Control Capacity. *Criminology* 55, 4 (2017), 754–778.

[14] Lawrence E Cohen and Marcus Felson. 1979. Social change and crime rate trends: A routine activity approach. *American sociological review* (1979), 588–608.

[15] Marcus Felson and Rémi Boivin. 2015. Daily crime flows within a city. *Crime Science* 4, 1 (2015), 31.

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

[17] Corina Graif, Alina Lungeanu, and Alyssa M Yetter. 2017. Neighborhood isolation in Chicago: Violent crime effects on structural isolation and homophily in inter-neighborhood commuting networks. *Social networks* 51 (2017), 40–59.

[18] Elizabeth R Groff and Tom McEwen. 2007. Integrating distance into mobility triangle typologies. *Social Science Computer Review* 25, 2 (2007), 210–238.

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

[20] John Hughes and Murali Haran. 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75, 1 (2013), 139–159.

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

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

[23] Stephen A Matthews and Tse-Chuan Yang. 2013. Spatial Polygamy and Contextual Exposures (SPACEs) Promoting Activity Space Approaches in Research on Place And Health. *American Behavioral Scientist* 57, 8 (2013), 1057–1081.

[24] Stephen A Matthews, Tse-Chuan Yang, Karen L Hayslett, and R Barry Ruback. 2010. Built environment and property crime in Seattle, 1998–2000: A Bayesian analysis. *Environment and Planning A* 42, 6 (2010), 1403–1420.

[25] Daniel P Mears and Avinash S Bhati. 2006. No community is an island: The effects of resource deprivation on urban violence in spatially and socially proximate communities. *Criminology* 44, 3 (2006), 509–548.

[26] RE Morgan and JL Truman. 2014. Nonfatal domestic violence 2003–2012. *Washington, DC: Bureau of Justice Statistics, US Department of Justice* (2014).

[27] Cathy O'Neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books.

[28] Paul M Ong and Douglas Houston. 2002. Transit, Employment and Women on Welfare1. *Urban Geography* 23, 4 (2002), 344–364.

[29] World Health Organization et al. 2013. *Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence.* World Health Organization.

[30] Andrew V Papachristos, David M Hureau, and Anthony A Braga. 2013. The corner and the crew: the influence of geography and social networks on gang violence. *American sociological review* 78, 3 (2013), 417–447.

[31] Ruth D Peterson and Lauren J Krivo. 2010. *Divergent social worlds: Neighborhood crime and the racial-spatial divide.* Russell Sage Foundation.

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

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

[34] Robert J Sampson. 2012. *Great American city: Chicago and the enduring neighborhood effect.* University of Chicago Press.

[35] David R Schaefer. 2012. Youth co-offending networks: An investigation of social and spatial effects. *Social networks* 34, 1 (2012), 141–149.

[36] Clifford R Shaw and Henry D McKay. 1942. Juvenile delinquency and urban areas. *Chicago, Ill* (1942).

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

[38] Brian J Stults and Matthew Hasbrouck. 2015. The effect of commuting on city-level crime rates. *Journal of Quantitative Criminology* 31, 2 (2015), 331–350.

[39] George Tita and Elizabeth Griffiths. 2005. Traveling to violence: The case for a mobility-based spatial typology of homicide. *Journal of Research in Crime and Delinquency* 42, 3 (2005), 275–308.

[40] Maria B Velez, Christopher J Lyons, and Blake Boursaw. 2012. Neighborhood housing investments and violent crime in Seattle, 1981–2007. *Criminology* 50, 4 (2012), 1025–1056.

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

[42] Per-Olof H Wikström, Vania Ceccato, Beth Hardie, and Kyle Treiber. 2010. Activity fields and the dynamics of crime. *Journal of Quantitative Criminology* 26, 1 (2010), 55–87.

[43] David A Wolfe, Claire V Crooks, Vivien Lee, Alexandra McIntyre-Smith, and Peter G Jaffe. 2003. The effects of children's exposure to domestic violence: A meta-analysis and critique. *Clinical child and family psychology review* 6, 3 (2003), 171–187.