COMBINING GEOGRAPHIC AND SOCIAL PROXIMITY TO MODEL URBAN DOMESTIC AND SEXUAL VIOLENCE

CLAIRE KELLING, DEPARTMENT OF STATISTICS
GIZEM KORKMAZ, BIOCOMPLEXITY INSTITUTE OF VIRGINIA TECH
CORINA GRAIF, DEPARTMENT OF SOCIOLOGY AND CRIMINOLOGY
MURALI HARAN, DEPARTMENT OF STATISTICS

MOTIVATION

- Understand crime in urban areas
- Investigate the socio-demographic attributes of the communities as well as the interactions between neighborhoods
- Incorporate social proximity as strong social ties can lead to the transfer of ideas, customs, and behaviors

DATA SOURCES

- Police Data Initiative
- Arlington County Police Department
- U.S. Census Bureau
  - American Community Survey
  - LODES (Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics)

PROPOSED METHODS

Neighborhood Matrix Structure, W

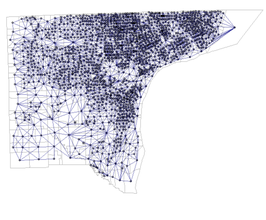

Figure 1: Detroit, Geog

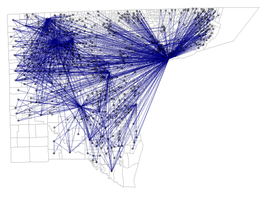

Figure 3: Detroit, Social

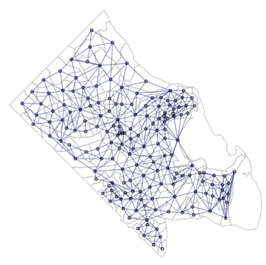

Figure 2: Arlington, Geog

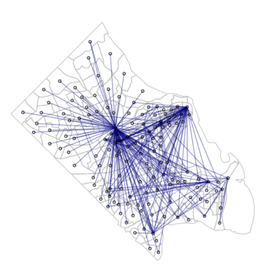

Figure 4: Arlington, Social

Spatial Generalized Linear Mixed Model (SGLMM)

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

- Y_k, x_k, and ψ_k represent the response, covariates, and spatial structure component
- $\mu_k = E(Y_k)$
- $\beta = (\beta_{\mathbf{1}}, ...\beta_{\mathbf{p}})$, is assumed to have a multivariate Gaussian prior distribution
- Assume $Y_k \sim$ Poisson (μ_k) and $ln(\mu_k) = \mathbf{x_k^T}\beta + \psi_{\mathbf{k}}$

Demographics

- income
- unemployment rate
- population
- gender
- Herfindahl Index
- age

MODELING FRAMEWORK

We use three models for ψ_k, the spatial random effects for modeling spatial dependence, in order to demonstrate the robustness of our technique.

- **BYM (Besag, York, and Mollié) [1]:** $\psi_k = \phi_k + \theta_k$
  - Two sets of random effects, spatially autocorrelated and independent
  - Only their sum is identifiable
  - Multivariate specification: $\phi \sim N(0, \tau^2\mathbf{Q}(\mathbf{W})^{-\mathbf{1}})$, where $\mathbf{Q}(\mathbf{W}) = \text{diag}(\mathbf{W1}) - \mathbf{W}$
- **Leroux [2]:** $\psi_k = \phi_k$
  - Provides improved parameter interpretability
  - Widely accepted, theoretically and practically [3]
  - Multivariate specification: $\phi \sim N(0, \tau^2\mathbf{Q}(\mathbf{W}, \rho)^{-\mathbf{1}})$, where $\mathbf{Q}(\mathbf{W}, \rho) = \rho[\text{diag}(\mathbf{W1}) - \mathbf{W}] + (\mathbf{1} - \rho)\mathbf{I}$
- **Sparse SGLMM [4]:**
  - Addresses potential spatial confounding issues
  - **Spatial confounding**: the phenomenon by which spatial random effects act as if they are multicollinear with the covariates ("fixed effects", in our case the demographic variables)
  - Can impact our ability to interpret the regression coefficients [5]

RESULTS

Model Comparison:

- *Deviance Information Criteria* **(DIC)**
  - Measure that combines the "goodness of fit" and "complexity" [6]
  - Deviance, $D(\theta) = -2\text{logL}(\text{data}|\theta)$
  - Complexity is measured by the estimate for effective number of parameters, $p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta})$
    $$\text{DIC} = D(\bar{\theta}) + 2p_D = \bar{D} + p_D \quad (2)$$
  - Smaller DIC indicates that the model is better supported by the data
- *Percentage Deviance Explained* **(PDE)**
  - Larger PDE means the model is better supported by the data

Detroit

BYM:	Geog	Comb
DIC	8923.9	8845.0
PDE	56.9	61.0
Leroux:	**Geog**	**Comb**
DIC	8029.7	7005.6
PDE	63.9	68.1
Sparse SGLMM:	**Geog**	**Comb**
DIC	8522.3	8251.8
PDE	39.0	58.6

Arlington

BYM:	Geog	Comb
DIC	1149.6	1147.8
PDE	81.6	81.7
Leroux:	**Geog**	**Comb**
DIC	1144.9	1138.0
PDE	81.5	81.6
Sparse SGLMM:	**Geog**	**Comb**
DIC	1725.1	1614.4
PDE	53.9	52.8

Spatial Confounding:

- We analyzed if there are any problems of spatial confounding and compared the estimated values of the posterior median as well as the 95% credible intervals for the model parameters for the combined Leroux Model and for the combined sparse SGLMM model.
- We found there is no large difference between the combined Leroux Model and the combined sparse SGLMM model coefficients.
- Therefore, spatial confounding does not appear to impact our estimates.

Coefficient Analysis:

- Both models show that median income has a very small coefficient, and both show an insignificant estimate for the coefficients of the unemployment rate and percentage male, which we use to show gender diversity.
- Also, both models show there is a positive estimate for the Total Population, and it is quite similar between the two models.
- There is also a positive estimate for the coefficient of the Herfindahl Index; in the Combined Leroux model, the credible interval includes 0 whereas in the sparse SGLMM model, the credible interval does not include 0 but is quite close to 0.

CRIME ESTIMATION

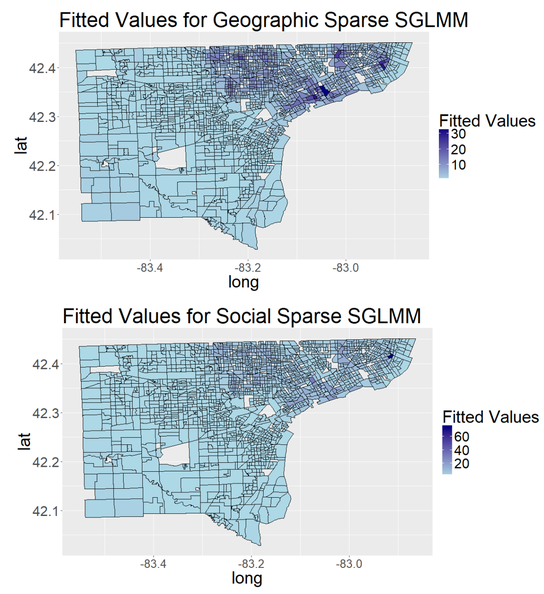

Figure 5: Fitted Values (Detroit)

CONCLUSIONS

- Employment-based mobility may diffuse norms or resources.
- Incorporating social proximity results in more accurate estimates of crime.
- We demonstrate the robustness of our conclusions to a variety of different assumptions about the underlying spatial dependence:
  - spatial random effects and
  - fixed effects (demographics).
- Our approach could be useful beyond modeling crime, to examine more generally the spread of influence through social and ecological networks.

FUTURE WORK

- Extend to other communities
- Study other types of crime
- Include crime estimates outside of the given County (to include complete commuting proximity)
- Include other measures of social proximity, such as taxi/Uber data

ACKNOWLEDGEMENTS

This material is based on work supported by the National Science Foundation under IGERT Grant DGE-1144860, Big Data Social Science.
This work was also supported by

- Data Science for Public Good Program, Social and Decision Analytics Laboratory, Biocomplexity Institute of VT
- Arlington County Police Department
- NIH, Grant 1K01-HD093863-01
- Penn State University's Population Research Institute (NICHD R24-HD041025)

REFERENCES

[1] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.
[2] G Leroux Brian, Xingye Lei, Norman Breslow, M Halloran, and Berry Donald Elizabeth. Estimation of disease rates in small areas: a new mixed model for spatial dependence. *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191, 2000.
[3] Duncan Lee. A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2(2):79–89, 2011.
[4] John Hughes. ngspatial: A Package for Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data. *The R Journal*, 6(2):81–95, 2014.
[5] Brian J Reich, James S Hodges, and Vesna Zadnik. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206, 2006.
[6] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.