Spatial Statistics Workshop

Claire Kelling

Pennsylvania State University- Department of Statistics

BIGSSS CSS 2019

References: [Shaby, 2017], [Schutte, 2018], [Diggle, 2013], [Schabenberger and Gotway, 2017]

Introduction

Introduction

What is the field of “spatial statistics"?

- Studies entities/events using their geographic properties
- Spatial data can take many forms
- Spatial data is increasingly common, especially at the point/event level

Outline for today:

1. Introduction/Motivation
2. Areal Unit Models
3. Point Process Models
4. Other Topics
5. Workshop in R

Motivation

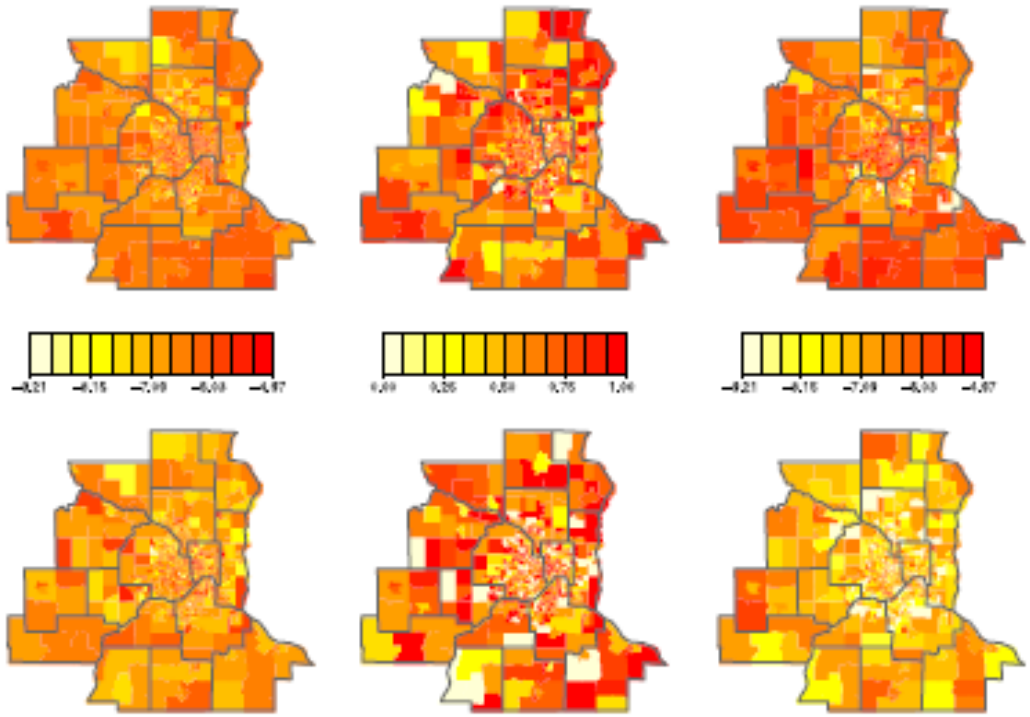


Figure: Covariate information, cancer diagnosis [Liang et al., 2008]

Motivation

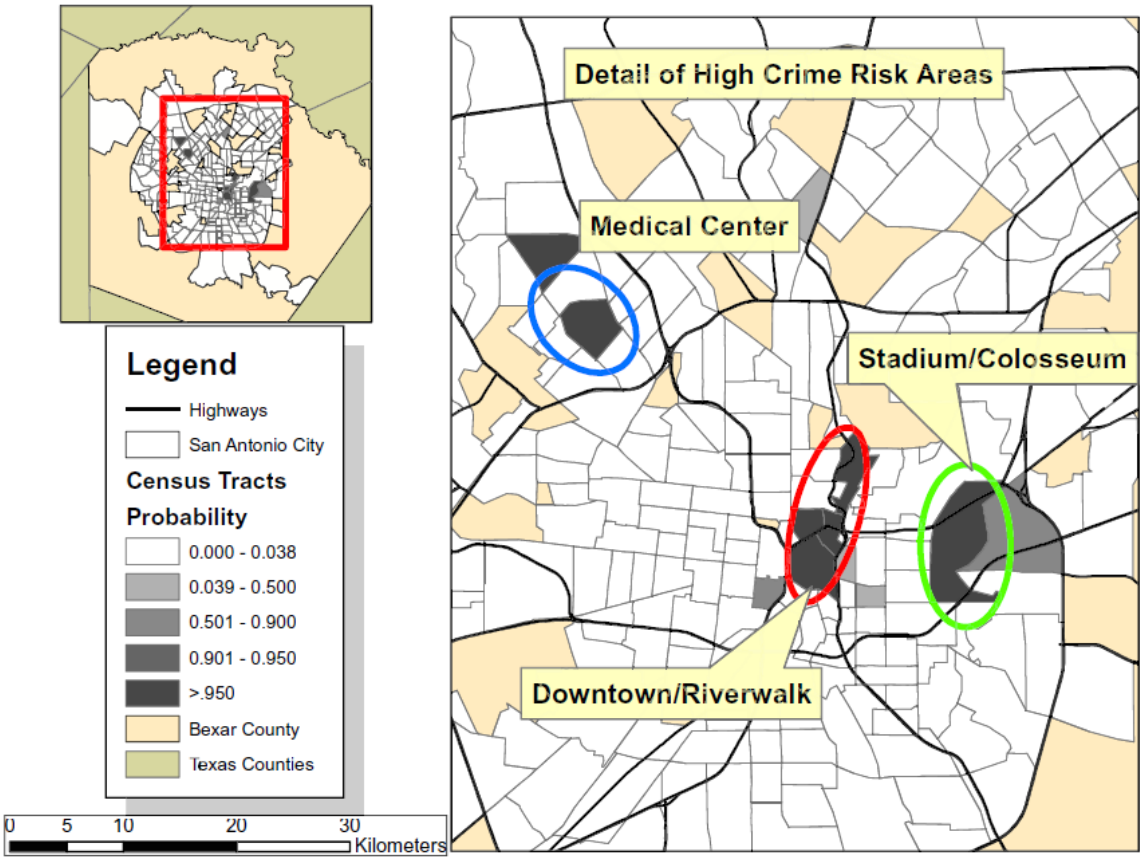


Fig. 2. Map showing areas of San Antonio with very high crime risk and their contexts.

Figure: Violent crime [Sparks, 2011]

Motivation

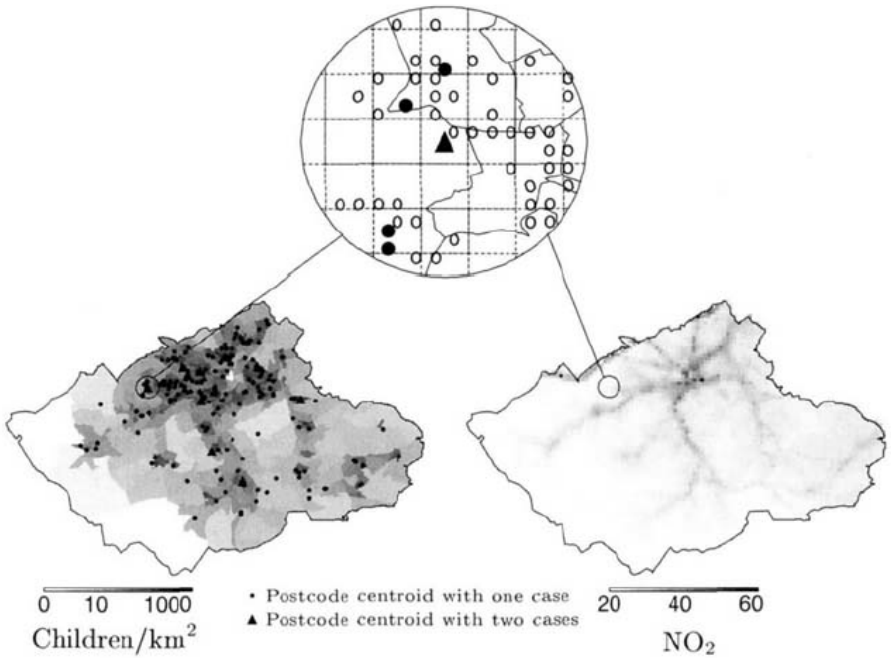


Figure: Wheezing cases [Best et al., 2000]

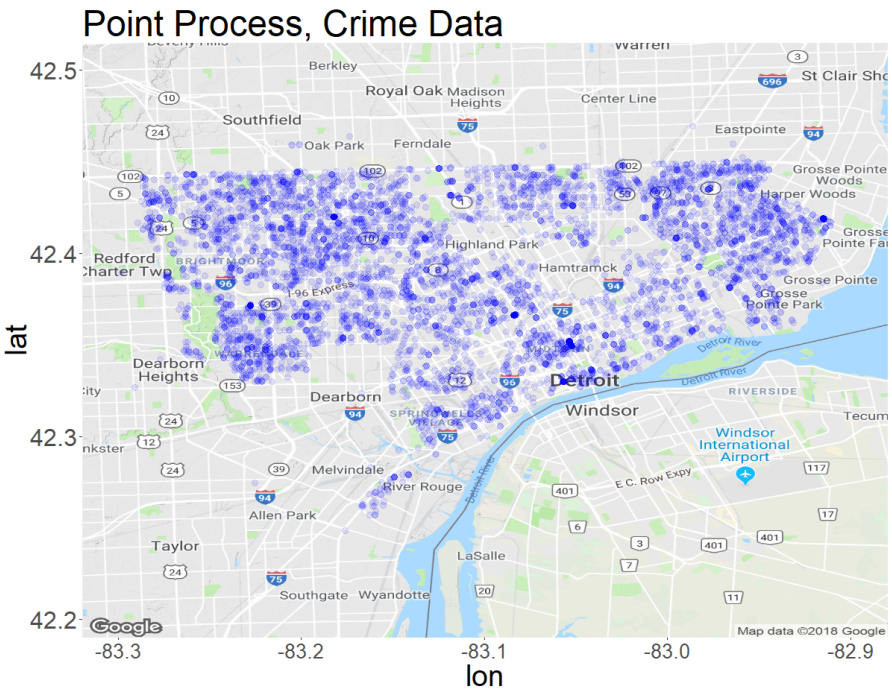
Preliminaries

What are the categories of spatial data?

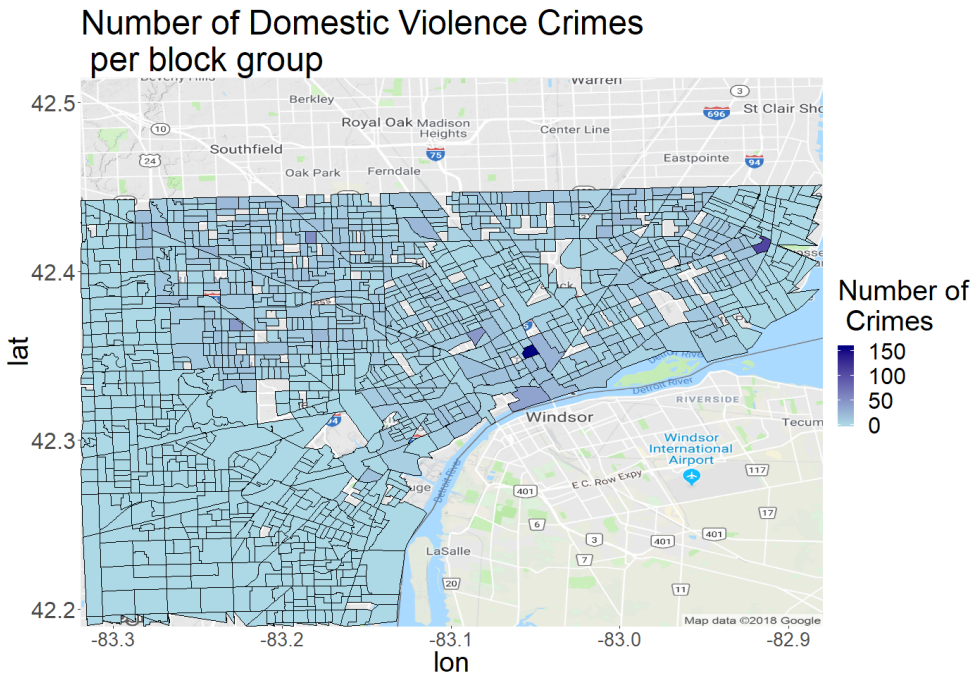
1. Point-referenced (geostatistical)
   - $Y(s); s \in \mathbb{R}^d$ and s varies continuously
   - **Example:** What is the temperature at an un-measured location? (prediction)
2. Areal (lattice)
   - Finite number of areal units, e.g. counties or elements of a grid
   - Observations are typically sums or averages
   - **Example:** Are variations in outcomes related to covariates, such as demographic characteristics? (estimation)
3. Point Process (point patterns)
   - Locations are themselves the data
   - **Example:** Can we distinguish clustering or repulsion?

Spatial Modeling

Point Process Modeling

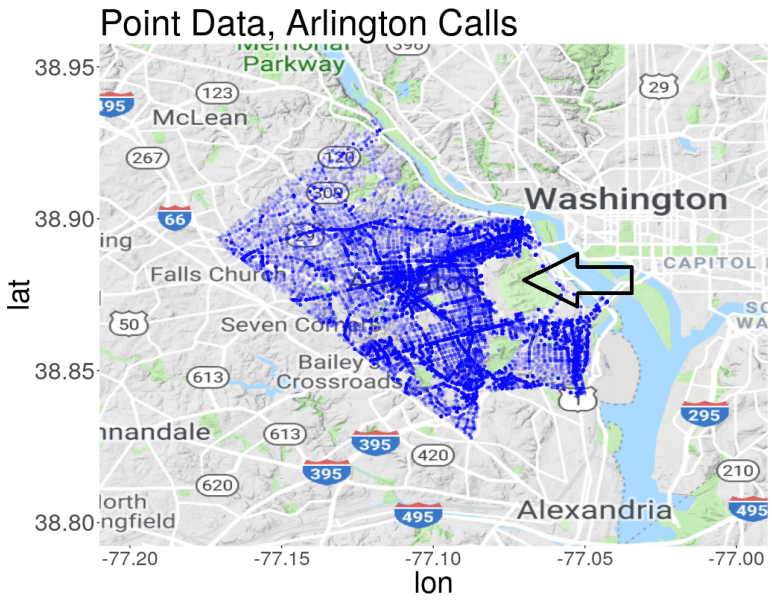


Areal Unit Modeling

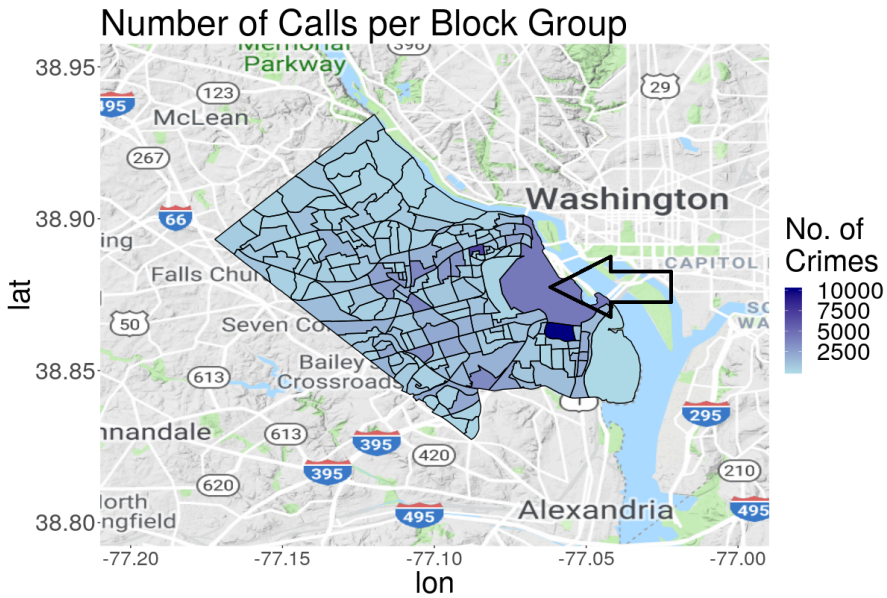


Caution

Crime Point Process



Crime Aggregated by Census Block Groups

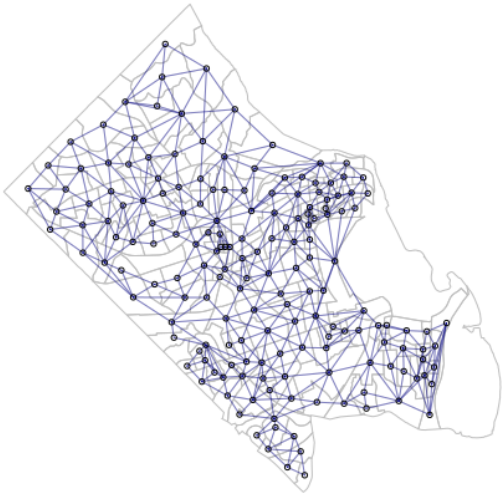


- All of these calls are occurring on the border of Arlington National Cemetery, very close to other communities.
- Also, do not just create maps that are essentially population maps.

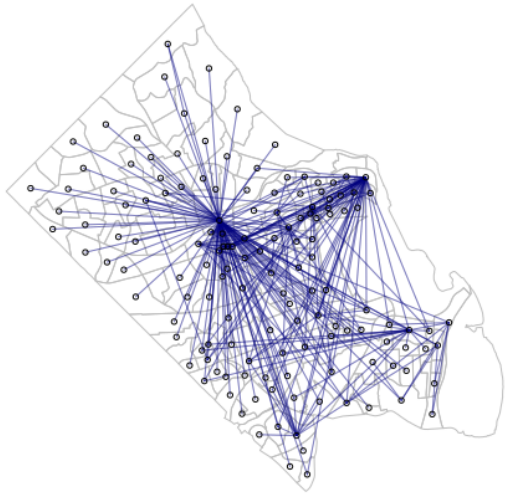
Areal Unit Models

Neighborhood Matrix

Geographic Neighborhood



Social Neighborhood



Goal

Define/capture the dependence between areal units.

Neighborhood Matrix, **W**

Some details on neighborhood matrix, **W**

- non-negative, symmetric, K$\times$K
- (i,j)th element of the neighborhood matrix w_{ij} represents spatial closeness between areas $(\mathcal{S}_i, \mathcal{S}_j)$
- positive values denoting geographical closeness and zero values denoting non-closeness (0-1 is the most common structure)
- $w_{ii} = 0$

Bayes CAR Model Lee [2013] Set Up

CAR = Conditional Autoregressive model for areal data

- study region $\mathcal{S}$ is partitioned into K non-overlapping areal units
- linked to set of responses $\mathbf{Y} = (Y_1, ..., Y_K)$
- spatial variation in the response is modeled by a matrix of covariates $\mathbf{X} = (x_1, ..., x_k)$ and a spatial structure component $\psi = (\psi_1, ..., \psi_k)$
  - $\psi = (\psi_1, ..., \psi_k)$ models any spatial autocorrelation that remains after covariate effects have been accounted for

$$g(\mu_k) = \mathbf{x_k}^\mathbf{T}\beta + \psi_\mathbf{k}$$

BYM Model (Besag, York, and Mollie) [Besag et al., 1991]

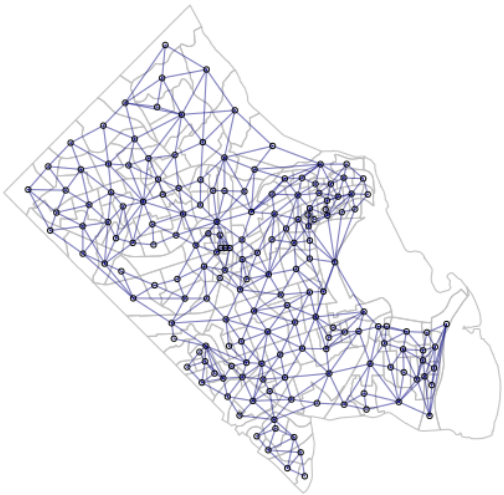
$$\psi_k = \phi_k + \theta_k$$

$$\phi_k|\phi_{-k}, \mathbf{W}, \tau^2 \sim N\left(\frac{\sum_{i=1}^{K} w_{ki}\phi_i}{\sum_{i=1}^{K} w_{ki}}, \frac{\tau^2}{\sum_{i=1}^{K} w_{ki}}\right)$$

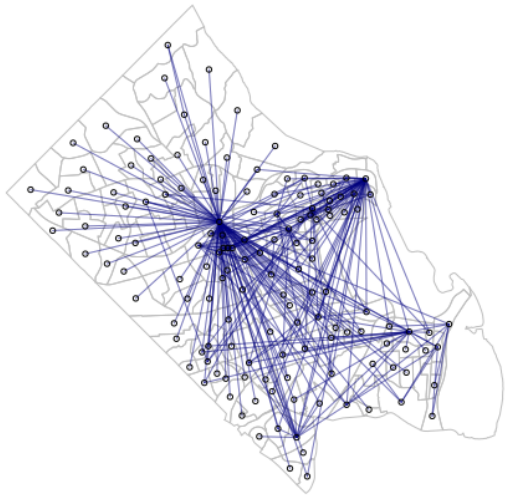
- First CAR model to be proposed.
- Two sets of random effects, spatially autocorrelated and independent
- Requires two random effects to be estimated at each data point, whereas only their sum is identifiable

My Research

Geographic Neighborhood



Social Neighborhood



Findings

- When you modify the neighborhood matrix, W, to incorporate both spatial and social proximity, you create a better model of urban crime.
- This may be applicable to other outcomes of interest that fit into this areal unit modeling framework.

Point Process

Point Process Model Set Up

Goal:

To develop a model that will assess for and characterize spatial structure in the data.

- Can develop a flexible model for the mean surface for events, $\mu(s)$, in order to estimate the overall trend across our area of interest.
- Can use a Gaussian Process (GP) model and estimate the parameters of its covariance function.
- Need to assess for Complete Spatial Randomness

Point Process Basics

Poisson Processes over region A

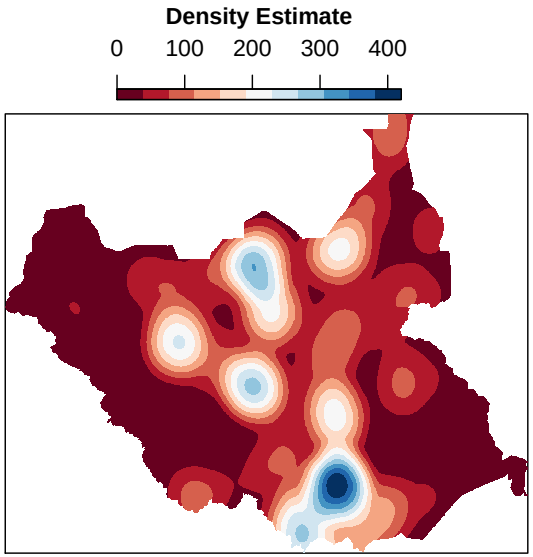
- **Homogeneous Poisson Process**: Static intensity within region A (complete spatial randomness)
- **Inhomogeneous Poisson Process:** Variable intensity within region A
- Depends on *intensity function* $\lambda(s)$

Log Gaussian Cox Process (LGCP)

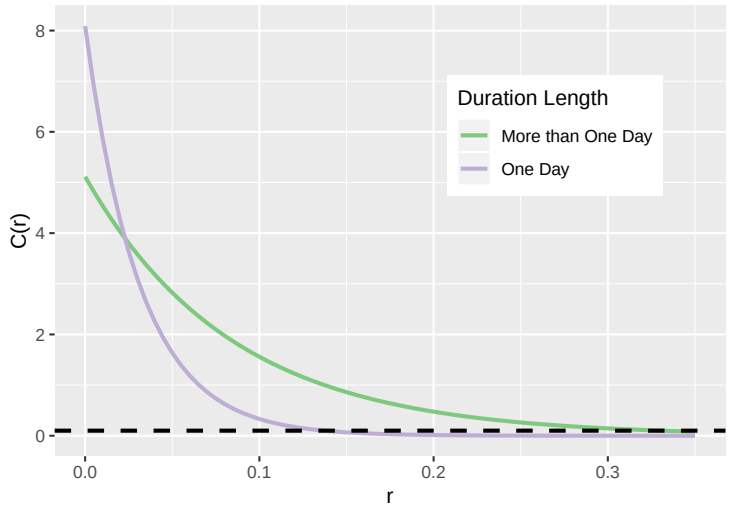
- We build upon the model developed in Liang et al. [2008] for studying disease risk using spatially varying and non-spatially varying covariates.
- This model allows us to combine covariates at different resolutions, whether it be on crime event or the community level, to draw conclusions about the risk of crime in the area of interest as well as to determine the dominating factors leading to that risk.

My Research

Spatial Kernel Fit



Estimate of Covariance Function



Other

Other Topics

Geocoding

- Some of our datasets consist of addresses, not coordinates.
- We have tested various geo-coding techniques to find latitude/longitude coordinates for each of these addresses.
- The results are not always consistent and it can be expensive to get a large number of addresses.

Preliminary Tests

- **Spatial autocorrelation**: Moran's I
- **Complete Spatial Randomness**: Ripley's K

Other Topics

Projections

- It is important to be aware when dealing with spatial data (shape files, point patterns, etc) that projections/coordinate systems are extremely important!
- Projections define what coordinate system you are using (basically)
- What creates this problem?
  - A **map** is a two-dimensional representation of the surface of the earth
  - A **globe** is a three-dimensional model of the earth
  - You lose information when going from 3D to 2D

In R

Packages in R

Popular R packages for spatial statistics

- `maptools`: load and display spatial data
- `rgdal`: reprojections
- `sp`: generate spatial data structures
- `spatstat`: load and maniuplate raster data
- `CARBayes`: fit areal unit CAR models
- `ngspatial`: fit areal unit models while addressing confounding
- `ppm`: fit/explore point process data

Today's Workshop

Find today's workshop presentation and code at:

https://github.com/ckelling/spatial_stats_workshop

References/Appendices

References

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.

Nicola G Best, Katja Ickstadt, and Robert L Wolpert. Spatial poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American statistical association*, 95(452):1076–1088, 2000.

Peter J Diggle. *Statistical analysis of spatial and spatio-temporal point patterns*. Chapman and Hall/CRC, 2013.

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

Shengde Liang, Bradley P Carlin, and Alan E Gelfand. Analysis of minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. *The annals of applied statistics*, 3(3):943, 2008.

Oliver Schabenberger and Carol A Gotway. *Statistical methods for spatial data analysis*. Chapman and Hall/CRC, 2017.

Sebastian Schutte. Spatial event data analysis (in r!). *BIGSSS CSS in Conflict*, 2018.

Ben Shaby. Spatial models. *Introduction to Spatial Statistics*, 2017.

Corey S Sparks. Violent crime in san antonio, texas: An application of spatial epidemiological methods. *Spatial and spatio-temporal epidemiology*, 2(4):301–309, 2011.

Thanks!

Thanks for your attention!
Questions?

Contact information:

Email: cek32@psu.edu

Point Process Model Details

- $z(s) =$ a vector of location-specific, spatial covariates corresponding to crime event (demographic features, housing information, neighborhood characteristics, etc.)
- $\lambda(s) = r(s)\pi(s)$ where $r(s)$ is the population density at location s, or an offset
- $\pi(s) = \exp(z(s)'\beta + \omega(s))$ where $\omega(s)$ is a zero-centered stochastic process, such as Gaussian Process and β is unknown vector of regression coefficients

Adding non-spatial covariates:

- $\pi(s, v) = \exp(\beta_0 + z(s)'\beta + v'\alpha + (v \otimes z(s))'\gamma + \omega(s))$ where s is the location and v is the crime event