GENERAL ASSEMBLY

SPATIAL ANALYSIS

Matt Brems, Data Science Immersive

AGENDA

- Introduction
- Spatial Statistics
- Mapping Codealong

INTRODUCTION

- Today, we're continuing our discussion of correlated data by jumping into spatial data. We'll delve into how to integrate this with temporal data tomorrow.

- Recall: How are spatial and temporal data different than the data with which we're usually working?

- Recall: Why is that a problem?

WHY WORK WITH SPATIAL DATA?

- Analysis – to find a statistical model that adequately explains the dependence observed in spatial data, often toward the goal of understanding how spatial location affects the variable of interest.

- Prediction – to predict or forecast values of the spatial process.

- Adjustment – removing the “spatial component” so that we can answer another question. (i.e. decomposing results into trend and spatial components.)

- Simulation – investigating how the stochastic process behaves through repeated simulations.

TYPES OF SPATIAL DATA

- Spatial data is interpreted as a realization of a stochastic process. (Stochastic processes are sets of random variables which allow our modeling processes to work more nicely by relying on the properties of randomness.)

- We would formally write this as $\{Y(s)|s \in D\}$, where $Y(s)$ are our random variables, s is our spatial input, and D is the "spatial domain."

- You can think of s as the different locations in space, $Y(s)$ as the value of interest at those locations in space (i.e. temperature, amount of snow, etc.), and D is the list of all possible locations s.

- The spatial domain D determines what resources we have at our disposal.

TYPES OF SPATIAL DATA

- If our spatial domain D is a set of non-overlapping regions, then we are working with an areal process. (i.e. states, ZIP codes)

- If our spatial domain D is continuous (likely two-dimensional or three-dimensional space), then we are working with a geostatistical process. (i.e. rainfall)
- "I want to understand what is happening."

- If our spatial domain D is a collection of random points, then we are working with a point pattern process. (i.e. locations of terror attacks)
- "I want to understand where it is happening."

AREAL DATA

- Areal processes (D is comprised of non-overlapping regions) are particularly difficult to work with because geography is messy.

- Whereas geostatistical and point pattern processes usually involve “distances” in the normal sense, areal processes are a bit more vague about how “distance” or “relationships” between different regions are defined.

WEIGHT/PROXIMITY/DISTANCE MATRIX

- We often use a weight matrix W to assess how closely related two regions are. This describes, for region i, how much weight to assign to region j.

- $w_{ii} = 0$ in all cases. (Why would this be the case?)

- While the choice is otherwise arbitrary, there are a few standard choices:
- $w_{ij} = 1$ if region i borders region j (or is within a certain distance of j), otherwise 0.
- $w_{ij} = \rho$ if region i borders region j, ρ^2 if i and j are one neighbor apart, and so on. (In this case, $0 < \rho < 1$ makes the most sense.)
- $w_{ij} = \text{dist}\{centroid\ i, centroid\ j\}$.

ACTIVITY

- For each of the three examples on the previous slide, generate a weight matrix. The different “regions” are the people sitting in your row (so most people will have five or six different regions). Note that your matrix will be $n \times n$, where n is the number of people in your row!

- You should work with the people closest to you, but note that you will all get slightly different weight matrices given that you’re all sitting in slightly different spots!

AREAL EDA

- Statistical EDA (aside from plotting) is difficult with areal data.

- The most commonly used statistical test is the permutation test using Moran's I statistic. (https://pysal.readthedocs.io/en/latest/library/esda/moran.html)

- H_0: no spatial correlation versus H_A: spatial correlation exists.

- Note: Small p-values allow you to conclude that spatial correlation exists, but doesn't provide information about the direction of the association!

GEOSTATISTICAL & POINT PATTERN EDA

- Plotting is your best bet here.

- Given the natural distance, you can also include geographic features in your EDA and use "standard" hypothesis tests/confidence intervals!

- If you're working with "irregularly placed data gatherers," be careful to check assumptions of any hypothesis test you'd use. When in doubt, defer to nonparametric hypothesis tests.

A MODELING STRATEGY

- If you believe that an additive model makes sense, then:

- 1. Try to estimate the trend by building a model $\hat{f}(s)$. (i.e. linear regression)
- 2. For every observation $\mathrm{Y}(s)$, calculate $\hat{\varepsilon}(s) = \hat{f}(s) - Y(s)$.
- 3. Examine the residuals and generate the covariance of the residuals.

REFERENCES

- This lecture draws heavily from Peter Craigmile's lectures. Peter is a professor of statistics at The Ohio State University and his lecture notes on spatial statistics can be found here: http://www.stat.osu.edu/~pfc/teaching/5012_spatial_statistics/