GENERAL ASSEMBLY

SPATIOTEMPORAL ANALYSIS

Matt Brems, Data Science Immersive

AGENDA

- Introduction

- Integrating Spatial and Temporal Statistics
- 1. Modeling
- 2. Space & Time Dependence
- 3. Training/Testing
- 4. Data Concerns
- 5. Visualization

INTRODUCTION

- Today, we're going to do a high-level overview of spatiotemporal analysis - what it is, some of its challenges, and what problems it can solve.

- You could take multiple graduate-level courses in spatiotemporal analysis, so we won't be experts by the end of the lesson, but we'll have a decent understanding of the basic ideas involved.

WHAT IS SPATIOTEMPORAL DATA?

- Spatiotemporal 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 $\{Y(s,t)|s \in D, t \in T\}$, where $Y(s,t)$ are our random variables, s is our spatial input, D is the "spatial domain," t is the time input, and T is the set of times under consideration.

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

A MODELING STRATEGY

- We'll decompose the data into mean and noise components: $Y(s,t) = \mu(s,t) + \varepsilon(s,t)$.

- If our goal is to study only how space and time affect $Y(s,t)$, we might only include space and time pieces in $\mu(s,t)$, then study $\mu(s,t)$.

- If our goal is to adjust for or negate the spatio-temporal dependencies, then we might only include space and time pieces in $\mu(s,t)$, then study $Y(s,t) - \mu(s,t) = \varepsilon(s,t)$.

- If our goal is to accurately forecast values of $Y(s,t)$, we might include other variables in $\mu(s,t)$ as well – not just space and time variables.

A MODELING STRATEGY

- We'll decompose the data into mean and noise components: $Y(s,t) = \mu(s,t) + \varepsilon(s,t)$.

- Often, it will be most helpful for $\varepsilon(s,t)$ to be a stationary process with a mean of zero. (It makes modeling nicer.)

- This is similar to the ε values in a linear regression model; we want them to have mean zero and have no discernable pattern with respect to the values of our independent variables.

SIMPLIFICATIONS (OFTEN UNREALISTIC)

- **Stationary**: A stationary spatio-temporal process is one that has:
- 1. A constant mean $\mu(s,t)$ that does not depend on space or time.
- 2. A covariance that depends only on spatial lag h and temporal lag u, not the actual points themselves s and t: $\mathrm{Cov}\big(\varepsilon(s+h,t+u),\varepsilon(s,t)\big) = \mathrm{Cov}\big(\varepsilon(h,u)\big)$

- **Isotropy**: An isotropic spatio-temporal process is one where spatial distance matters, but spatial direction does not.

- **Separability**: A separable spatio-temporal process is one where the covariance in space is independent of the covariance in time.

ARE SPACE AND TIME DEPENDENT?

- There are a series of hypothesis tests one can use to identify whether or not space and time are independent.

- http://pysal.readthedocs.io/en/latest/users/tutorials/dynamics.html#space-time-interaction-tests

- Mantel, Knox, and Jacquez Tests. (I prefer Mantel.)

TRAINING AND TESTING

- Training and testing our model is important to manage bias and variance.

- Why might using a simple random sample of 30% of our data for testing be improper?

TRAINING AND TESTING

- Training and testing our model is important to manage bias and variance.
- Why might using a simple random sample of 30% of our data for testing be improper?
- Use stratified random sampling to ensure a good cross-section of space and time data in both the training and testing sets. (Recommended.)
- Alternatively, you may use certain representative locations as the test locations. (This is known as a cluster sample.)
- If your goal is prediction/forecasting, it might make sense for you to use early time periods in your training and later time periods in your testing.

GATHERING DATA

- Gathering spatio-temporal data can be quite difficult:
- 1. Generally need a significant amount of data.
- 2. Differing reporting periods. (i.e. daily vs. weekly)
- 3. Format of data.

- Munging spatio-temporal data can be quite time intensive and, once finished, you may not have enough data to build an accurate model!

VISUALIZATION

- Spatio-temporal data are notoriously difficult to visualize.
- We want to be intentional about how we visualize the data.
- Be smart about how you include “time” in your visuals: https://fivethirtyeight.com/features/the-52-best-and-weirdest-charts-we-made-in-2016/
- Also be smart about how you represent “space.”

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 spatio-temporal statistics can be found here: http://www.stat.osu.edu/~pfc/teaching/Lyon/notes/5_spatio-temporal.pdf