Causal Inference with Spatial Data

Intensive PhD course at Department of Economics, University of Oslo
August 27-31, 2018

Lecturer: Masayuki Kudamatsu (OSIPP at Osaka University)

The purpose of this course is two-fold. First, we expose ourselves to various pieces of empirical economics research that use spatial datasets. This will help you come up with original research ideas by taking advantage of spatial datasets. Second, we have hands-on experiences of using ArcGIS 10 and of programing in Python by replicating the spatial datasets for the actual pieces of empirical economics research. This way, we will be able to find out how to process the spatial datasets needed for our own research.

This course has been taught since 2010, and the lecture slides from the previous years are available at the lecturer's website.

Grading Policy

This course welcomes those who just want to audit the course. For obtaining credit, however, you have to submit a research proposal. The submission deadline is to be announced, but you will have several weeks after the course ends. The guideline for writing a research proposal is attached at the end of this syllabus.

Course Schedule and Reading List

Lecture 1: Introduction to Spatial Data
August 27, 9:00-11:30

Kudamatsu, Masayuki. 2018. "GIS for Credible Identification Strategies in Economics Research." *CESifo Economic Studies*, 64(2): forthcoming.

Read this article for how spatial data can be used for causal inference. The papers listed below are briefly discussed in this article. In the lecture, we will learn the basics of spatial datasets: polygons, raster, map projection, etc.

Lecture 2: Spatial Join
August 28, 9:00-12:00

Alsan, Marcella. 2015. "The Effect of the TseTse Fly on African Development." *American Economic Review*, 105(1): 382–410.

We will replicate how this paper spatially merges ethnic group level data with the grid of weather data, a great source of exogenous variation.

Lecture 3: Buffer
August 28, 13:00-16:00

Conley, Timothy G., and Christopher R. Udry. 2010. "Learning about a New Technology: Pineapple in Ghana." *American Economic Review*, 100(1): 35–69.

We will replicate how this paper matches each agricultural plot with its neighboring plots within a radius of 1km, to control for common shocks in the peer effect estimation.

Lecture 4: Distance
August 29, 9:00-12:00

Nunn, Nathan. 2008. "The Long-Term Effects of Africa's Slave Trades." *Quarterly Journal of Economics*, 123(1): 139–176.

We will replicate how this paper measures, as an instrument variable for slave exports, the distance from each country's centroid to its nearest point on the coast, and also how the number of slaves exported at the ethnic group level is assigned to each country in Africa.

Lecture 5: Zonal Statistics
August 29, 13:00-16:00

Michalopoulos, Stelios. 2012. "The Origins of Ethnolinguistic Diversity." *The American Economic Review*, 102(4): 1508–1539.

We will replicate how this paper calculates the standard deviation of land suitability for agriculture at the 0.5-degree cell level within each "virtual country" (i.e. a 2.5-dgree cell) whose boundary is, by definition, exogenous.

Lecture 6: Elevation
August 30, 9:00-12:00

Duflo, Esther, and Rohini Pande. 2007. "Dams." *Quarterly Journal of Economics*, 122(2): 601–646.

We will replicate how this paper constructs instrument variables for dam construction: the slope of rivers.

Lecture 7: Spatial Regression Discontinuity Design
August 31, 9:00-12:00

Dell, Melissa. 2010. "The Persistent Effects of Peru's Mining Mita." *Econometrica*, 78(6): 1863–1903.

We will replicate how this paper obtains each observation's distance to the treated area boundary and the nearest segment of the boundary, to conduct regression discontinuity design. We will also replicate how this paper calculates the length of roads by taking elevation into account (Peru is a hilly country).

Research Proposal Guideline

1. Choosing a research question

Any research question is fine as long as it is an empirical question. The use of spatial datasets is encouraged, but not mandatory.

For tips to find a good research question, see the following pieces of writing by leading economists:

Steve Pischke
http://econ.lse.ac.uk/staff/spischke/phds/How to start.pdf
Don Davis
http://www.columbia.edu/~drd28/Thesis%20Research.pdf
Ross Levine
http://www2.warwick.ac.uk/fac/soc/economics/staff/mfmcmahon/macro/better_research.pdf
Avinash Dixit
http://www.princeton.edu/~dixitak/home/dixitwrk.pdf (see pages 4-6 in particular)
Hal Varian
http://people.ischool.berkeley.edu/~hal/Papers/how.pdf (see Sections 1 and 2 in particular)
David Levine
http://faculty.haas.berkeley.edu/LEVINE/cheap_advice.html#dissertation

2. Originality

A literature survey is NOT allowed. Your proposal needs to be for a piece of original research.

3. Structure

You should first clearly state the research question (the specific one, not the general one such as "Does credit market failure impede growth?").

Then explain why this question is important and original. Here you may need to cite the existing literature, but never write a lengthy literature review. The purpose of citing the literature is to show why your research is original and contributes to the literautre in an important way.

Finally, describe the empirical research method to answer this research question.

1. Discuss a theory that guides your empirical analysis. It doesn't have to be a mathematical model. But explain possible mechanisms for why your treatment variable affects the outcomes you are going to look at.
2. Describe the data you will use (the sample period, the number of observations, the list of variables to be used, the sampling method employed if it's a survey data).
3. Write summary statistics tables (or figures) with the content empty. You can at least specify what variables to be reported in these tables and figures. Here you may also want to create a table in which you compare the means between treated and control observations with t-statistics reported.
4. Write down the equation(s) to be estimated. Explain what estimation method you use (OLS, IV, SUR, Probit etc.), including how standard errors will be computed.

5. Describe the identifying assumption, that is, the assumption that ensures the interpretation of estimated coefficients as causal impacts.
6. Write down regression tables. You don't have results yet, but you can still write them down to describe what empirical specification each column estimates. What is the dependent variable? How is the sample restricted? Which regressors are included? What F-tests will be reported? And so forth. Add notes to the table, to explain how standard errors are computed, what F-statistics refers to, etc.
7. Write paragraphs to explain the purpose of each column in these regression tables. This is where you need to think hard about the potential threat to your identifying assumption. To deal with each threat, what kind of robustness checks need to be done?

You don't need to write implications of your finding or the conclusion, because you haven't obtained any finding yet.

4. Full paper (Optional)

If you have already obtained some empirical findings, you can write them up in the form of a full paper, which can be a draft chapter in your PhD thesis.

For the structure of a full paper, follow the papers that you read for this course (or any other paper published in top journals) as closely as possible. Pick the paper that's closest to your paper in terms of empirical research design (e.g. RCT, DID, IV, RD, cross-sectional regression, etc.), which may be from a different field of economics than yours, and imitate its structure.