University of Chicago
Department of Sociology
Autumn 2019

SOCI 20253/GEOG 20500/ENST 20510
SOCI 30253, MACS 54000

Introduction to Spatial Data Science

Luc Anselin and Marynia Kolak

Meet:	Mon, Wed 1:30-2:50pm, Stuart Hall 101
Offices:	CSDS, 2nd Floor Searle Chemistry 224 (Anselin), 232 (Kolak)
E-Mail:	anselin@uchicago.edu and mkolak@uchicago.edu
TAs:	Hanock Spitzer haspitzer@uchicago.edu and TBN
Office Hours:	Mon, Wed 3:00-4:00pm and by appointment
Prerequisite:	STAT 22000 (or equivalent), familiarity with GIS is helpful, but not necessary

SYLLABUS

Course Description

Spatial data science is an evolving field that can be thought of as a collection of concepts and methods drawn from both statistics/spatial statistics and computer science/geocomputation. These techniques deal with accessing, transforming, manipulating, visualizing, exploring and reasoning about data where the locational component is important (spatial data). The course introduces the types of spatial data relevant in social science inquiry and reviews a range of methods to explore these data.

We will primarily focus on data gathered for aggregate units, such as census tracts or counties (e.g., unemployment rates, disease rates by area, crime rates), and will only briefly consider data measured at spatially located sampling points (such as air quality monitoring stations and urban sensors) and observations at the point level (e.g., locations of crimes, commercial establishments, traffic accidents).

Specific topics covered include the special nature of spatial data, geovisualization and visual analytics, spatial autocorrelation analysis, cluster detection and regionalization. An important aspect of the course is to learn and apply open source geospatial software tools, primarily GeoDa, but also R.

Objectives

1. Learn principles of spatial data science and its application to social science research questions
2. Learn to distinguish which methods are appropriate for a given research question
3. Gain an appreciation for the assumptions and limitations associated with each technique
4. Learn how to interpret and present the results of a spatial data analysis in a coherent fashion
5. Learn how to use appropriate open source software tools to carry out spatial data analytical applications

Organization

The class will meet twice a week, alternating between a lecture and a lab. The lecture will typically be on Mondays, the lab on Wednesdays.

All software used in the class is free and open source. You should install everything on your laptop. If you do not have a laptop, arrangements can be made through the library to make one available to you. The software is also installed on the machines in various computer labs on campus.

The course will use Canvas as the main communication mechanism. All materials, including software guides and data will be available from the course site. Note that the Canvas course number used is SOCI 20253.

All assignments, papers etc. must be submitted as a pdf digital file to Canvas: NO PAPER copy and no Word docs, no exceptions.

Requirements

The main goal for the course is for you to complete a final project/paper that carries out an in-depth spatial data analysis of a research problem of your choice. You will apply a subset of the methods covered in class and use your own data or data provided by the instructor (your own data is preferred). This paper will be due at the end of the semester, December 9. In addition to the final paper, there are two intermediate deliverables: an outline of the research question and data; and a summary of the methods used with initial results. Specific deadlines will be posted.

More details will be provided in class and on the Canvas course web site as the quarter progresses.

In addition, there will be four assignments, each consisting of a short computational exercise that uses specific spatial analytical methods. The assignments are graded pass/fail and you must pass all four (and complete on time) in order to be able to receive an A grade in the course. If you don't pass at least three, you will fail the course. You can re-do an assignment as many times as necessary in order to reach the required three, but only on-time assignments can result in an A grade.

Software

The class uses only open source software (free and cross-platform). Everything can be readily downloaded from the web.

- **GeoDa**, available from http://geodacenter.github.io/download.html. Make sure to have the latest version 1.14 (August 2019)
- **R** (3.6.1. or later) and its associated spatial data analysis packages, everything available from http://cran.r-project.org
- Recommended: **RStudio**, a graphical user interface to R, available from https://www.rstudio.com/products/rstudio/download3/

Readings

There is no text for the course. There are many excellent books on data science, but to date the treatment of spatial aspects is still in its infancy. The GeoDa Workbook notes at http://geodacenter.github.io/documentation.html and the lecture slides provide the formal background. There is also a set of lab notes that mimic the GeoDa operations in R, available at https://spatialanalysis.github.io/tutorials/.

Specific readings will be assigned each week and made available on the course Canvas web site. These readings are complementary and provide additional background on the material. They are not required, but if you are serious about spatial data science, you may want to take a look at them.

General background can be found in the following annotated bibliography (these are *not* required reading, but provided as a guide to the literature)

GeoDa

- https://spatial.uchicago.edu/geoda
  - a brief description of the GeoDa functionality
- http://geodacenter.github.io/documentation.html
  - GeoDa Workbook (in progress)
- Luc Anselin (2005). *Exploring Spatial Data with GeoDa, A Workbook*. (available on the course web site)
  - a bit dated in terms of the interface, but the substance hasn't changed; new version in the works, see above

Data Science

- Cathy O'Neil and Rachel Schutt (2013). *Doing Data Science, Straight Talk from the Frontline*. O'Reilly.
  - very readable introduction to data science (among many others)
- Garrett Grolemund and Hadley Wickham (2017). *R for Data Science*. O'Reilly.
  - also available online from the book's web site http://r4ds.had.co.nz
  - a collection of data science "skills" using R, with an emphasis on "data munging" using specialized R packages – highly recommended if you are serious about data science using R
- Benjamin Baumer, Daniel Kaplan and Nicholas Horton (2017). *Modern Data Science with R*. CRC Press.
  - an in-depth treatment of data science operations, with an emphasis on data handling and data base queries
- David Donoho (2015). *50 Years of Data Science*.
  - available from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf
  - a critical look at the data science "hype" from a statistician's perspective; excellent perspective on historical precedents in the work of Tukey (EDA), Chambers (computing with data), Cleveland (graphics), etc.

Quick introduction to R

- W.N. Venables, D.M. Smith and the R Core Team (2019). *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics* Version 3.6.1 (July 2019) https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
  - the *classic* introduction and overview of the R language and its use for statistical analysis
- Paul Torfs and Claudia Brauer (2014). *A (very) short introduction to R*. http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf
  - an introductory overview and quick start (only 12pp.)

General references for R (admittedly biased, my favorites)

- Robert Kabacoff (2015). *R in Action (2nd Edition)*. Manning Publications.
  - a very readable introductory overview of R functionality and R programming
  - associated web site for Quick R http://www.statmethods.net
- Michael J. Crawley (2013). *The R Book (2nd Edition)*. Wiley
  - *the* comprehensive guide to R (must have if you are going to do any serious work in R) – a bit dated now, but still a classic
- R Bloggers
  - compilation of *daily* blogs pertaining to the use of R in statistics and data science
  - https://www.r-bloggers.com

- RStudio blog
    - for the latest on R for data science
    - https://blog.rstudio.org
- Microsoft R Application Network (Microsoft's release of R and supporting materials)
    - https://mran.microsoft.com
- guides to R books, tutorials, etc.
    - https://cran.r-project.org/other-docs.html
    - https://www.r-project.org/other-docs.html
- for hard-core R programmers, Hadley Wickham's Advanced R web site (as well as his books on ggplot2, packages and advanced R)
    - http://adv-r.had.co.nz
- Deborah Nolan and Duncan Temple Lang (2015). *Data Science in R. A Case Studies Approach to Computational Reasoning and Problem Solving*. CRC Press.
    - worked real-live data science case studies (advanced)

Spatial data analysis in R

- The R ecosystem for spatial data analysis
    - http://cran.r-project.org/web/views/Spatial.html
- Robin Lovelace, James Cheshire, Rachel Oldroyd and others (2015). Introduction to visualizing spatial data in R
    - http://github.com/Robinlovelace/Creating-maps-in-R
    - a good introductory overview to GIS operations in R
- Guy Lansley and James Cheshire (2016). *An Introduction to Spatial Data Analysis and Visualization in R*
    - a more extensive introduction to mapping, spatial data manipulation and visualization in R
    - http://www.spatialanalysisonline.com/An%20Introduction%20to%20Spatial%20Data%20Analysis%20in%20R.pdf
- Roger Bivand, Edzer Pebesma and Virgilio Gomez-Rubio (2013). *Applied Spatial Data Analysis with R* (2nd Edition). Springer, New York, NY.
    - more advanced, in-depth coverage of spatial statistical packages in R (assumes quite a bit of R expertise)
- Chris Brunsdon and Lex Comber (2019). *An Introduction to R for Spatial Analysis and Mapping (2nd Edition)*. Sage.
    - a more in-depth intermediate overview of GIS operations and some spatial analysis in R with lots of illustrations
- Robin Lovelace, Jakub Nowosad and Jannes Muenchow (2019). *Geocomputation with R.* CRC Press.
    - GIS and spatial analytical operations using a range of R packages
    - https://geocompr.robinlovelace.net

Grading

Assignments:	40%
Project Paper:	
Research question and data sources	10%
Methods outline/initial results	10%
Final report	40%

Tentative Course Outline (subject to change)

Introduction and overview (W10/02)

- Introduction and logistics
- Overview of the software

Spatial Data Science

- *Spatial data science* (M10/07)
  - Important concepts
  - Spatial data and spatial analysis

- Lab (W10/09): *Spatial Data Handling*
  - Data wrangling, aggregation, spatial join

Visual Analytics

- *Visual Analytics (1)* (M10/14)
  - Principles of visual analytics, EDA, ESDA
  - Linking and brushing
  - EDA basics
  - Scatter plot (smoothing, brushing)
  - Scatter plot matrix
  - Parallel Coordinate Plot (PCP)
  - Conditional plots

- Lab (W10/16): *Classic EDA*

- *Visual Analytics (2)* (M10/21)
  - Map types (choropleth, outlier maps)
  - Principles of map design
  - Cartogram
  - Co-location maps
  - Conditional maps
  - Mapping rates

- Lab (W10/23): *Geovisualization*

- Assignment 1: Data acquisition (due 10/25)

Spatial Autocorrelation

- *Spatial Autocorrelation Principles* (M10/28)
    - Spatial randomness
    - Positive and negative spatial autocorrelation
    - Spatial weights

- Lab (W10/30): *Spatial Weights*

- *Spatial Autocorrelation Statistics* (M11/04)
    - Join count, Moran’s I, Geary’s c
    - Moran scatter plot
    - Variogram
    - Nonparametric spatial correlogram

- Lab (W11/06): *Global Spatial Autocorrelation*

- Assignment 2: Visualizing relationships among multiple variables (due 11/08)

- *Local Spatial Autocorrelation* (M11/11)
    - Local Moran cluster maps
    - Local Geary, Local join counts
    - Multivariate extensions

- Lab (W11/13): *Local Spatial Autocorrelation Statistics*

Cluster Detection

- *Unsupervised Learning* (M11/18)
    - Curse of dimensionality
    - PCA and MDS maps
    - K-means clustering
    - Hierarchical clustering

- Lab (W11/20): *Cluster Detection*

- Assignment 3: Identifying local clusters and spatial outliers (due 11/22)

- *Spatially Constrained Clustering* (M11/25)
    - Spatial constraints
    - Weighted clustering
    - Skater, Redcap
    - Max-p

- (W11/27) No Class – Thanksgiving

- Lab (M12/02): *Spatially Constrained Clustering*

- *Review* (W12/04)

- Assignment 4: Spatial clusters (due 12/06)

- Final project due 12/09