University of Chicago
Department of Sociology
Autumn 2016

SOCI 20253/GEOG 20500, SOCI 30253

Introduction to Spatial Data Science

Luc Anselin

Meet:	Mon, Wed 1:30-2:50pm (place TBA)
Office:	407 Social Science Research Building
E-Mail:	anselin@uchicago.edu
Office Hours:	Mon, Wed 3:00-4:00pm and by appointment
Prerequisite:	multivariate statistics, 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 and computer science. 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. The types of data considered include observations at the point level (e.g., locations of crimes, commercial establishments, traffic accidents), data gathered for aggregate units, such as census tracts or counties (e.g., unemployment rates, disease rates by area, crime rates), and data measured at spatially located sampling points (such as air quality monitoring stations and urban sensors). Specific topics covered include the implementation of formal spatial data structures, geovisualization and visual analytics, spatial autocorrelation analysis, variogram analysis, cluster detection, regionalization, point pattern analysis and spatial data mining. An important aspect of the course is to learn and apply open source geospatial software tools, such as R and GeoDa.

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 in a lecture/lab format. The lab sessions are on Mondays in the Computer Science Instructional Lab (CSIL), on the first floor of the Crerar Library. The lecture is on Wednesdays in a location TBA.

The computer lab will have all required software installed. However, you are strongly encouraged to use your own laptop and install the software yourself (everything is cross-platform, free and open source).

The course will use Chalk as the main communication mechanism. All required readings, software guides and data will be available from the course site.

All assignments, papers etc. must be submitted as a pdf digital file: NO PAPER 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. Over the course of the semester, there will be three intermediate deadlines, with deliverables that are part of the grading of the final paper:

- A fully spelled-out research question (including a brief description of the data to be used) – 1p., due Oct. 12
- A descriptive analysis of the spatial characteristics of the main variables in the study – max 2pp. (not including graphs and figures), due Oct. 26
- A detailed outline of the final paper – 2pp., due Nov. 9
- Final paper is due Dec. 5

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

In addition, there will be eight weekly assignments, each consisting of short computational exercises that use specific spatial analytical software (GeoDa and R). The assignments are graded pass/fail and you must pass all eight (and complete on time) in order to be able to receive an A grade in the course. If you don't pass at least

six, you will fail the course. You can re-do an assignment as many times as necessary in order to reach the required six, but only on-time assignments can result in an A grade.

Software

The class uses only open source software (free and cross-platform). The software will be installed in the CSIL computer lab, but you are strongly encouraged to install it on your own machine. Everything can be readily downloaded from the web.

- **GeoDa**, available from http://geodacenter.github.io/download.html
- **R** and its associated spatial data analysis packages, especially `spdep`, `gstat` and `spatstat`, everything available from http://cran.r-project.org
- Recommended: **jupyter** notebooks, available from http://jupyter.readthedocs.io/en/latest/install.html (requires a Python installation, preferably through the Anaconda distribution – see instructions on the jupyter web site)

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 lecture slides provide the formal background. Specific readings will be assigned each week and made available on the course Chalk web site.

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
  - slides with an overview of the functionality in Version 1.8
- 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, parts of which may be made available during the course of the semester

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)

Quick introduction to R

- W.N. Venables, D.M. Smith and the R Core Team (2016). *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics* Version 3.3.1 (June 2016)
  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
- 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)
- inside-R, a community site for R – sponsored by Revolution Analytics
  - http://inside-r.org
- RStudio blog
  - for the latest on R for data science
  - https://blog.rstudio.org
- 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
- Garrett Grolemund and Hadley Wickham (2016). *R for Data Science*. O'Reilly
  - http://r4ds.had.co.nz (draft in progress)
  - a collection of data science "skills" using R, with an emphasis on "data munging" using specialized R packages
- 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

- 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 (2015). *An Introduction to R for Spatial Analysis and Mapping*. Sage.
    - a more in-depth intermediate overview of GIS operations and some spatial analysis in R with lots of illustrations

Grading

- Class participation: 10%
- Assignments: 40%
- Project Paper: 50%
    - research question 5%
    - data descriptive analysis 5%
    - outline 10%
    - final paper 30%

Tentative Course Outline

Introduction and overview (lab 9/26)
- Overview of the class
- Overview of software
- Introductory demonstration of GeoDa and R (RStudio, Jupyter notebooks)

Spatial Data Science
(9/28)
- Spatial data science
- Spatial effects
- Spatial data structures
- Spatial data and spatial analysis

- Lab (10/3): getting data into GeoDa, Table manipulations, joins, subselection, spatial data structures in R (sp)

- Assignment 1: Data acquisition

Visual Analytics
(10/5)
- Principles of visual analytics, EDA, ESDA
- Linking and brushing
- Conditioning
- Map types, cartogram

- Box plot, PCP, conditional plots, scatterplot matrix
- Brushing scatterplots/maps to assess heterogeneity

- Lab (10/10): visual analytics in GeoDa

- Assignment 2: Assessing spatial heterogeneity

Spatial Autocorrelation
(10/12)
- Spatial autocorrelation principles
    - Spatial randomness
    - Positive and negative spatial autocorrelation
- Spatial weights

- Lab (10/17): spatial weights in GeoDa and R spdep

- Assignment 3: Properties of spatial weights

(10/19)
- Spatial autocorrelation statistics (global)
    - Join count, Moran's I, Geary's c
    - Moran scatter plot
    - Nonparametric spatial correlogram

- Lab (10/24): global spatial autocorrelation in GeoDa and R spdep

- Assignment 4: Sensitivity analysis of spatial autocorrelation

(10/26)
- Variogram
    - principles of geostatistics
    - exploratory variography
    - directional effects

- Lab (10/31): exploratory variography with R gstat

- Assignment 5: Exploring the variogram

Spatial Clusters
(11/2)
- Local spatial autocorrelation
    - Local Moran, local Gi
- Scan statistics

- Lab (11/7): cluster detection with LISA (GeoDa) and scan statistics (R SpatialEpi)

- Assignment 6: Cluster detection

(11/9)
- Spatial cluster detection principles
    - K-means, hierarchical clustering, contiguity constrained clustering
- Regionalization
    - Zoning procedures, max-p, neural networks

- Lab (11/14): cluster detection in R (spdep), Geogrouper (clusterpy)

- Assignment 7: Sensitivity analysis of cluster detection techniques

Events in Space
(11/16)
- Principles of point pattern analysis

- Lab (11/21): classical point pattern analysis in R spatstat

- Lab (11/28): network point pattern analysis in R spatstat

- Assignment 8: Point pattern analysis

Future of Spatial Data Science
(11/30)
- Machine learning, spatial data mining