PLAN 672: Urban Data Analytics in R

Nikhil Kaza

Spring 2020

Class Room: Philips 367
TA: TBD
Email: nkaza@unc.edu

Time: MW 4:40 PM - 5:55 PM
Office Hours: TBD
Office: 314 New East

Course Description & Objectives

This course is about different techniques used in assembling, managing, analysing and predicting using heterogeneous data sets in urban environments. These data sets that are inherently messy and incomplete. Types of data include, point, polygon, raster, vector, text, image and network data; data sets with high cadence and high spatial resolution. This is a survey course for different techniques and approaches in dealing with these data to make short term operational decisions as well as long term planning. As such, the emphasis is on practical urban data analytics rather than in-depth discussion about the suitability and appropriateness of techniques and their associated theoretical assumptions.

This is a companion course to PLAN 673: The Ethics and Politics of New Urban Analytics (Seminar), which deals with problems, opportunities and hidden agendas with data generation, analysis and visualisation in urban settings. Students are encouraged to take them both.

Prerequisites & Preparation

The course will move quickly, cover a large number of analytical techniques, data sets, use cases and disciplinary domains. It requires significant investment on the part of the students to learn the technical skills as well as learn about the substantive urban and regional analyses.

Much of the work in this course will be done using Open Source Software that is usually free.

While it is not a prerequisite, the course assumes a working knowledge of R. R is a programming language and free software environment for statistical computing and graphics. There are a number of online resources that will help you with getting up to speed with R. You will use extensively the documentation, help and examples that R environment provides; i.e. Do not be afraid to use, for example,

```
?qplot
??randomForest
```

to seek help for specific commands.

One disadvantage with R is that it stores all its objects in memory. This means that your computer should have significant RAM to deal with large data sets.

Another disadvantage with R is that it has a *shallow learning curve*. And it has some quirks. In particular, please pay attention to R-Inferno. However, persistence will have long term benefits.

You should have an aptitude for debugging computer code, thinking through edge cases in data sets, identifying and dealing with missing data and messy data sets.

You should expect that the instructions and help provided may not work on your system due to different configurations, mismatched data types and differences in libraries. You should have an aptitude to troubleshoot the problems and figure out workarounds.

Textbooks & Readings

The following textbooks are used implicitly in the class. You should buy them and keep them as a reference.

Brewer, Cynthia A. (2015). Designing Better Maps: A Guide for GIS Users. 2 edition. Redlands, California: Esri Press. ISBN: 978-1-58948-440-5.

Few, Stephen (2015). Signal: Understanding What Matters in a World of Noise. Burlingame, California: Analytics Press.

Tufte, E. R (2001). The Visual display of Quantitative Information. Cheshire, CT: Graphics Press.

All these books are about principles of information display and design rather than about data analysis techiques. Information visualisation is very important and much more so than analytical techniques though enough attention is not devoted to them. While we won't be using these textbooks explicitly in weekly readings, you are expected to critically engage with the materials and thoughtfully follow the principles laid out in the books throughout the course.

The following books are recommended as a reference that will get you started on some analytical techniques.

Bivand, Roger S, Edzer Pebesma, and Virgilio Gomez-Rubio (2013). Applied Spatial Data Analysis with R. 2nd ed. 2013 edition. New York Heidelberg Dordrecht London: Springer. ISBN: 978-1-4614-7617-7.

Grolemund, Garrett and Hadley Wickham (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Sebastopol, CA: O' Reilly media,. URL: http://r4ds.had.co.nz/ (visited on May. 25, 2018).

When readings are assigned, links are usually provided on Sakai or should be available from the library.

Course Policies

The following set of course policies is not meant as an exhaustive list. If in doubt, ask for permission and clarification.

Deadlines & Extension Requests

Students must read assigned material before class session. Completed lab session materials are due by the end of lab in Sakai. Homework assigned for the week is due the following Monday at 6 PM in Sakai. If there is a reason to extend the deadline for the entire class, please discuss with me at least a week ahead and make a cogent case. All homework needs to be submitted as a R markdown file, as well as the html output of the markdown.

Equipment

Every student should have a working laptop that has R and Rstudio installed. The laptops should have sufficient memory and processing speed to deal with large data sets. If you have access to no such equipment, please see me immediately to discuss options.

Grading

- **20%** lab reports to be submitted at the end of the class. (Individual/Collaborative)
- **30%** (Mostly) weekly homework programming assignments that are (usually) due Tue 5 PM.(Individual/Collaborative)
- **10%** Critique of a data visualisation. (Assignment 1) (Group)
- **10%** Discussion & critique of a smart cities data analytics platform. (Assignment 2) (Group)
- **20%** Final term project. (Assignment 3) (Individual)
- **10%** Class & lab participation

Attendance and Participation

If you don't attend classes, but submit the requirements on time, you will be penalised only on the participation grade. Group assignments will be individually graded based on what your peers suggest your participation and leadership in the group is.

E-mail

Sakai messaging system should be the preferred way to communicate with me or the TA. Before you email either of us about homework or lab sessions, you should use resources on the web and on Sakai. Google, Stack Overflow, Sakai forums are your friends. The class has a group email list. Please be considerate to your colleagues and do not spam their Inbox.

Academic Conduct

I firmly believe in learning from your peers and from others. All homework and lab submissions could benefit from collaborations, however, the submissions are individual. This means that interpreting the data and the results, producing the visualisations, drawing appropriate conclusions from the data is necessarily individual even when the strategies can be discussed and developed with others in class or out of class. **All** help, however, should be explicitly acknowledged. Severe penalties are imposed for non-attribution.

Schedule (Tentative)

Jan 8 (Wed): Lec: Introduction; Urban Datasets & Analytics Platforms

Jan 13 (Mon): Lab: Introduction to R & QGIS. Creating R Markdown files.

Jan 15 (Wed): Lab: Data Manipulation in R

Jan 22 (Wed): Lab: Visualise data in R using ggplot and create a basic interactive visualisation

Jan 27 (Mon): Lec: Mapping Flows

Jan 29 (Wed): Lab: Analyse Bike share, GPS & LODES data sets

Feb 3 (Mon): Pres: Assignment 1 Group Presentations

Feb 5 (Wed): Pres: Assignment 1 Group Presentations

Feb 10 (Mon): Lec: Creating Composite Indices & Dimensionality Reduction

Feb 12 (Wed): Lab: Create a sprawl index from census and economic data

Feb 17 (Mon): Lec: Machine Learning & Cluster Analysis

Feb 19 (Wed): Lab: Clustering cities based on water consumption

Feb 24 (Mon): Lec: Visualising, Interpolating & Analysing Point Patterns

Feb 26 (Wed): Lab: Analysing crime clusters in Manchester

Mar 2 (Mon): Assignment 2 Group Presentations

Mar 4 (Wed): Assignment 2 Group Presentations

Mar 16 (Mon): Lec: Text & Corpus Analysis

Mar 18 (Wed): Lab: Scraping Twitter API and sentiment analysis

Mar 23 (Mon): Lec: Network Analysis

Mar 25 (Wed): Lab: Generating accessibility maps

Mar 30 (Mon): Lec: Outlier Detection in Time Series.

Apr 1 (Wed): Lab: High resolution electricity consumption data of campus buildings

Apr 6 (Mon): Lec: Raster & Image Analysis

Apr 8 (Wed): Lab: Urban landscape metrics

Apr 13 (Mon): Lec: Classification with Trees & Forests, Boosting & Bagging

Apr 15 (Wed): Lab: Classifying Remote Sensing Images

Apr 20 (Mon): Lec: Deep Neural Networks