PLAN 0000: Urban Data Analytics

Nikhil Kaza

null

E-mail: nkaza@unc.edu
Office Hours: TBD
Office: TBD

Web: sia.planning.unc.edu
Class Hours: TBD
Class Room: TBD

Course Description & Objectives

This course is about different techniques used in assembling, managing, analysing and predicting using heterogeneous data sets in urban environments. These include point, polygon, raster, vector, text, image and network data; data sets with high cadence and high spatial resolution; data sets that are inherently messy and incomplete. This is a survey course for different techniques and approaches. 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 0001XXX: Data & Society (Seminar), which deals with the ethics and politics of data 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 Gómez-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,. 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 Tuesday at

5 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

You are adults and free to choose how to spend your time. I don't care to keep track of your attendance in class and labs. If you don't attend classes, but submit the requirements on time, you will be penalised only on the participation grade.

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)

Aug 22 (Wed): Introduction

- Lab Session: Introduction to R & QGIS. Creating R Markdown files.
- Homework:

Aug 29 (Wed): Urban Datasets & Analytics Platforms

- Lab Session: Data Manipulation in R
- Homework:

Sep 5 (Wed): Representation, Visualisation & Cartography

- Lab Session: Visualise data in R using ggplot and create a basic interactive visualisation
- Homework:

Sep 12 (Wed): Mapping Flows

- Lab Session: Analyse Bike share, GPS & LODES data sets
- Homework: Analyse NYC taxi cab trip data, Airsage cell phone data

Sep 19 (Wed): Assignment 1 Group Presentations

Sep 26 (Wed): Creating Composite Indices & Dimensionality Reduction

- Lab Session: Create a sprawl index from census and economic data
- Homework: Create a composite health of housing market index from multiple datasets including unemployment rate, repeat sales, time on market, sale price at census tract level over time.

Oct 3 (Wed): Unsupervised Learning & Cluster Analysis

- Lab Session: Analyse 311 noise complaint data, clustering cities based on water consumption
- Homework: Analyse production and violations data from Pennsylvania unconventional natural gas wells.

Oct 10 (Wed): Visualising, Interpolating & Analysing Point Patterns

- Lab Session: Analysing crime clusters and air quality sensors
- Homework: Analyse Chicago's Array of Things data set and interpolate for the city.

Oct 17 (Wed): Assignment 2 Group Presentations

Oct 24 (Wed): Text & Corpus Analysis

- Lab Session: Scraping Twitter API, sentiment analysis
- Homework: Sexual harassment on public transit

Oct 31 (Wed): Network Analysis

- Lab Session: Analysing hydrological networks and transportation networks, Google API, Accessibility to day care centers
- Homework: Create intersection density, network connectivity of various parts of different cities and cluster the neighbourhoods in different cities.

Nov 7 (Wed): Time Series Analysis, Predictive Models & Outlier Detection

- Lab Session: Analyse 15 min interval electricity consumption data of campus buildings at UNC and predict energy demand for campus
- Homework:

Nov 14 (Wed): Raster & Image Analysis

- Lab Session: Urban landscape metrics, Remote Sensing classification
- Homework:

Nov 21 (Wed): Classification with Trees & Forests, Boosting & Bagging

- Lab Session:
- Homework:

Nov 28 (Wed): Deep Neural Networks

- Lab Session:
- Homework:

Dec 5 (Wed): Assignment 3 Individual Presentations