PLAN 672: Urban Data Analytics in R

Jan 10, 2022 ·

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 in R. The objective of these analytical techniques is to inform both short term operational decisions and long term planning in cities. 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.

Unlike other courses of similar vein, I put inordinate emphasis on data visualisation and communication. The point of data analysis is to tell a compelling story, not to use latest analytical techniques.

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.

Course Details

- **Instructor:** Nikhil Kaza
- **Teaching Assistant:** Xijing Li
- **Classroom:** New East 101
- **Hours:** MW 4:40PM - 5:55PM
- **Office Hours:** go.unc.edu/kaza
- **TA Office Hours:** M 12-1 PM, Th 2-3 PM on Zoom
- **Web:** nkaza.github.io & sakai.unc.edu

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 to learn about 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 a 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 have to 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.

It may be helpful to go through the materials from STOR 320: Introduction to Data Science

Textbooks & Readings

The following books are used implicitly in the class. You are not required to buy any of them but they are very useful to have on your bookshelf.

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. ISBN: 978-1-938377-05-1.

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

Wickham, Hadley (2016). Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

All the above books are about principles of information display and design rather than about data analysis techniques. Information visualisation is very important and much more so than analytical techniques though enough attention is not devoted to them. While we may not 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.

For general purpose statistics, I have always enjoyed Tim Harford's podcast called More or Less. He has a recent book out that succinctly details the attitudes you want to take towards data analysis and telling stories with data. I highly recommend his new book.

Harford, Tim (2021). The Data Detective: Ten Easy Rules to Make Sense of Statistics. New York: Riverhead Books. ISBN: 978-0-593-08459-5.

The following books will get you started on some analytical techniques and can serve as a reference.

Bivand, Roger S., Edzer Pebesma, and Virgilio GÃfÂ³mez-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. first. Sebastapol, CA: O' Reilly. URL: http://r4ds.had.co.nz/ (visited on May. 25, 2018).

The following book is excellent for covering the latest techniques for Geospatial data in R

Lovelace, Robin, Jakub Nowosad, and Jannes Muenchow (2019). Geocomputation with R. 1 edition. Boca Raton: Chapman and Hall/CRC. ISBN: 978-1-138-30451-2. URL: https://geocompr.robinlovelace.net/ (visited on Dec. 01, 2019).

Course Policies

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

Logistics

- Your health and well-being is of paramount importance. You may also be primary care givers and might have substantial and increased demands on your time. You may not be able meet the requirements of the course, for any number of other reasons. Reach out to me, if you need any help, including if you need extensions or want to take an Incomplete or deal with it differently. I will deal with these on ad-hoc basis.
- I don't need any advance notification for intermittent absences. You should make appropriate judgments based on your health and your peers. However, you are responsible for keeping up with the material. Because the materials are posted on-line and in advance, you should be able to work through the code. If you have issues, please use Microsoft Teams, Office Hours and other resources available to you.
- Microsoft Teams will be used for troubleshooting and announcements. Sakai will continue to be used for HW, lab and assignment submissions. In addition, please create an account on rstudio.cloud (see below). You can sign in using your Github login.
- You should expect that the datasets you will need to download may not be available, because of server outages and lack of personnel to attend to these outages. We will cross those bridges, when we get to them.
- All bets are off, if I fall ill. Hopefully, there is enough course infrastructure already set up for you to achieve the learning objectives on your own. The department will figure out how to assign grades, in that eventuality.

Deadlines & Extension Requests

Completed lab session materials are due by the end of lab (6 PM) in Sakai. You only need to submit one lab work for each topic.

Homework assigned for the week is due on the deadline specified 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 labs and homework needs to be submitted as two files 1) a R markdown file (*.Rmd) and 2) html output (*.html) of the Markdown file.

Readings/Resources

The weekly readings are provided as resources and references. You are not required to read all the materials in detail. But the readings are useful to learn the material in depth and troubleshoot some issues. In some cases, the software and techniques in the Resources may be dated. Please use the web to adapt and update them.

Tutorials

Often labs are accompanied by tutorials. The tutorials are usually self-contained and self-explanatory. In R, there are multiple ways to achieve the results, each with their own advantages and disadvantages. The tutorials may include different ways of data munging and analysis to expose you to different techniques. It is not implied that one is better than the others, though we all have our own preferences. If in doubt, rely on benchmarking.

Equipment

We will conduct the class in the New East lab. However, it would be helpful if every student has 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

While all assignments are posted on this website, they are to be submitted exclusively on Sakai and on time. Please refrain from emailing your submissions to the instructor or the TA.

I am going to use a 'Specification Grading' in this course.

- Lab reports to be submitted at the end of the class (6 PM). (Individual/Collaborative)
- (Mostly) weekly homework (HW) programming assignments (Individual/Collaborative)
- Critique of a data visualisation. (Assignment 1) (Group)
- Final term project. (Assignment 2) (Individual)

The assignments will be graded on an Satisfactory/Unsatisfactory scale. Satisfactory grade is equivalent to a B+ letter grade. The focus of these assignments is on learning outcomes such as mastery of the material, making innovative connections in the material and on-time submission.

You will need to achieve Satisfactory grades on at least 5 HW and 5 labs and at least one of the major assignments to achieve a L (equivalent to a C grade). Fewer than 4 Satisfactory grades in labs and HW will result in a failing grade. In addition to the Satisfactory grades in 7 labs and 7 HW, Satisfactory grades should be achieved in the both the major assignments to achieve a P.

Exceptional performance in the final term project, in addition to Satisfactory grade in 9 labs and 9 HW the other requirements, will result in H/A grade.

This grading scale will be adjusted if the deliverables change depending course progress.

Consent to Share

I am reserving the right to post your ungraded submissions (including HW) in Sakai (restricted to the class) for current and future students, as examples, without comment or recommendation. If you wish to decline to consent, please send me a note. No explanation is necessary and opting out does not affect your grades.

Attendance and Participation

If you don't attend classes, but submit the requirements on time, there is no penalty. Group assignments will get a single grade for the group. Continuous absences that affect the progress in the course should be discussed with the instructor to figure out remedial action.

E-mail

Sakai messaging system should be the preferred way to communicate with the instructor. Before you email either of us about homework or lab sessions, you should use resources on the web and on Sakai. Google, Stack Overflow and

Microsoft Teams are your friends.

Asynchronous Communication & Troubleshooting

We will use Microsoft Teams for asynchronous communication and troubleshooting. We can follow guidelines like these that allow you to get to answers quickly:

- https://stackoverflow.com/help/how-to-ask
- https://codereview.stackexchange.com/help/how-to-ask

We will use RStudio Cloud for troubleshooting in this course. Think of RStudio Cloud as an instance of RStudio in the cloud where you can share not only your script but also the whole environment. This increases the likelihood that others can replicate your results or troubles. Instructions are located here.

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. **All help including fragments of borrowed code**, however, should be explicitly acknowledged. Severe penalties are imposed for non-attribution. In particular, please pay attention to the copyright restrictions and attribution requirements associated with the R-code that you might find elsewhere.

Additional Help

Please set up a time on my calendar to discuss any additional help you may require.

I strongly recommend that you avail yourself of the R open labs run by Lorin Bruckner and Matt Jensen.

You can also request additional help from Lorin and Matt by setting up appointments with them from their website. In addition, Odum Institute has walk-in consultations and some of them have expertise in R.

Phil McDaniel and Amanda Henley are excellent resources for tracking down geospatial datasets and troubleshooting issues with them.

You will probably extensively use Stack Overflow to troubleshoot and debug your code. Please be mindful of how you should ask questions on Stack Overflow including providing minimum reproducible code and datasets.

There are organisations that are devoted to ensuring diversity in the R community. See for example, R-ladies meetup groups and Slack channels. Local groups may or may not be active.

Schedule (Tentative)

Introductory materials

Jan 10 (Mon), Jan 12 (Wed) Introduction. R & QGIS

Tutorials/Slides

- Data & Cities
- Create R markdown files and knit to HTML
- Create a map in QGIS
- Introduction to R - UCLA slides

Homework

HW1

- Download the building violations data from Chicago. Using QGIS, map and style the building violations by for at least 4 periods since 2006. Export these maps as PNG/JPG files to embed them in a R markdown file.
- In the same R markdown file, using the pictures and some (any) basic R code (such as to produce historgrams) tell a short story about the evolution of building violations in Chicago.

Resources/Readings

- Chapters 1, 4, 8 & 21 of (Grolemund and Wickham 2017)
- Modules 1, 2 & 3 of QGIS training manual
- (Resch and Szell 2019)
- (Kitchin 2016)
- (Singleton, Spielman, and Folch 2018)

Jan 19 (Wed) Using Git and Github

This week is coordinated with PLAN 372.

Jan 24 (Mon), Jan 26 (Wed), Jan 31 (Mon) Exploratory Data Analysis & Visualisation

Tutorials/Slides

- Starting with tidyverse & ggplot

Homework

HW 2

In the tutorial, we used data for O_3, a Criteria Air Polluatant (CAP). In this homework perform similar analysis for NO_2, another CAP. For this you will need to download the data from EPA and clean it.

You will need to tell a story about this pollutant and its distribution (in various senses). You may consider the following questions to structure your anlayses. In no way, these are exhaustive. Feel free to tell the story that interests you.

- How many sensors are measuring NO_2?
- Of what percentage of the year are the NO_2 sensors active? Where are the sensors that are not active? Map them.
- How does the geographical distribution of NO_2 AQI differ from that of O_3
- Which cities (CBSA, not counties) are worst affected by NO_2?
- Does the spatial extent of the CBSA matter? (i.e. if you use the 2000 definitions of MSA vs 2018 definitions of CBSA does your conclusions change? Hint: Modifiable Areal Unit Problem)
- Is there a correlation between NO_2 and O_3? Do different correlations matter? (i.e. correlations among AQIs of NO_2 and O_3 at site level, vs correlation of days AQI>100 at CBSA level etc.)
- What does the scatterplot look like? What does facetting the scatter plot by state tell us (pick 5 or so states)?
- Link this with other data (temperature, population etc.)? Where to acquires these datasets? How to link them?

Other Deliverables

- Assignment 1 proposal due on Wiki in Sakai by Friday 6 PM. Groups finalised. The proposal consists of 2-3 paragraphs and consists of the problem statement, concepts you will bring to bear and why it is important. It is posted so that everyone else has access and knowledge of the projects that your peers are working on. The proposal is posted on the wiki, on individual pages.

Resources/Readings

- Chapters 7 & 9-16 of (Grolemund and Wickham 2017)
- (Wickham 2010)
- (Tufte 2001)

Feb 2 (Wed), Feb 7 (Mon), Feb 9 (Wed) Maps & Flows

Tutorials/Slides

- Geospatial Data in R

Homework

HW 3

Download the motor vehicle traffic collisions data from NYC Open data portal.

Answer the following questions

- Which locations have high incidences of traffic collisions?
- How are these high traffic collisions locations different at different times of the day?
- What are the most frequent casues of collisions and how do they differ by location, time of the day and day of the week?
- Visualise the correlation between home values in a block group and traffic collisions and tell a story.

Resources/Readings

- Chapters 1-8 (Lovelace, Nowosad, and Muenchow 2019)

Feb 14 (Mon), Feb 16 (Wed) Scraping Web for (Un)Structured Data

Tutorials/Slides

- Using Census APIs
- Scraping Structured Data from Google | Alternate tutorial with OSM

- Unstructured Data

Homework

HW 4

Scrape Craigslist for houses available for rent in the Triangle and the Triad area. Using Google places, for each of those houses, extract different amenities (coffee shops, hospitals etc.) and construct a Walk Score like index for each listing. Tell a story about these indices (e.g. choice of amenities conditioning the results or spatial variation etc.)

Resources/Readings

- (Munzert et al. 2014)
- (Schweitzer 2014)
- (Boeing and Waddell 2017)

Feb 21 (Mon) Assignment 1 Presentations

Analytical Techniques

Feb 23 (Wed), Feb 28 (Mon) Analysing Text

Tutorials/Slides

- Matching Messy Texts
- Sentiment Analysis of Emails

Homework

HW 5

In this week's homework, I want you to analyse tweets in/referencing a city of your choice (e.g. Chicago). You are welcome to use a package that makes this easy for you such as `twitteR` or `rtweet`. However, you will need to set up a developer account and acquire credentials. (Hint: You may follow this tutorial). The key tasks are think through the following issues and write a story.

- Analyse the sentiments of the tweets
- Think through how the sentiments are changing
- Do official accounts (such as @cta and @cta311) differ in the tweets compared to general public?
- How prevalent is the location information in the tweets? i.e. Are the tweets useful to planners to pin point issues and track them?
- Is twitter a useful mechanism to identify prevalence of sexual assault in public spaces such as transit?

As usual, you are welcome to choose the scope and the above are not meant to be exhaustive.

Note that you will need to use your own credentials and submitting them using a Rmd file might reveal them to me and to your classmates. To prevent this, I am asking you to only submit an html file.

Resources/Readings

- (Boeing 2019)
- (Chen, Silva, and Reis In press)

Mar 2 (Wed), Mar 7 (Mon) Networks

Tutorials/Slides

- Network Analysis of Bikeshare systems
- Spatial Relationships as Networks

Homework

HW 6

An input output table is a square matrix with rows representing originating industry sector and columns representing destination industry sector. The value in each cell represents the flow from one sector to another. The higher the number, the more tightly connected the two industries are.

Download the Multi-Regional Input Output Matrix for China from (Mi et al. 2018)

In this example, each node would be a combination of region-industry and the weight on the link is the interaction (flows) between the two nodes. There are 30 regions and 30 industries, so the full graph could potentially have 900 nodes and 810000 links (lot of them have zero or near zero weights). Note that this matrix is not symmetric and the diagonal elements are non-zero

Do some data cleaning and explicitly convert the matrix to graph/network. In particular, be explicit about thresholds

that you will use to include the links in your network.

Tell a story about this network, by looking at (but not limited to) the following questions.

1. Identify the central region-industries in the entire network by constructing a betweenness centrality score. What can you conclude about the spatio-economic structure of China from these.
2. What other centrality measures are appropriate for this network and why?
3. Partition this network into communities using at least two different community detection algorithms. Do these results make intuitive sense?
4. Pick a subset of industries in your favourite region and create a subgraph of all the region-industries they are linked to. Identify the potential disrupters of this subgraph. Are these different from considering the entire flow pattern?
5. What is the relationship with this industry network with other network (labor markets, migration flows, road infrastructure etc.). Are there any connections you can draw? For this, you may have to scour the web for different datasets. Some of your classmates who are Mandarin speakers are probably better suited for this task. Again none of these are mandatory.

Other Deliverables

- Assignment 2 proposal due on Wiki in Sakai by Friday 6 PM. Groups finalised. The proposal consists of 2-3 paragraphs and consists of the problem statement, concepts you will bring to bear and why it is important. It is posted so that everyone else has access and knowledge of the projects that your peers are working on. The proposal is posted on the wiki, on individual pages.

Resources/Readings

- (Boeing 2019)
- (Nelson and Rae 2016)
- (Kaza and Nesse 2021)

Mar 9 (Wed), Mar 21 (Mon) Analysing Raster Datasets

Tutorials/Slides

- Basic Raster Analysis in R
- Urban Landscape Metrics
- Land Suitability Analysis

Homework/Deliverables

HW 7

Improve the land suitability analysis for locating a landfill in Durham County. For example, you can pick a non-trivial subset of the following

1. By explicitly accounting for Environmental Justice criteria (this is mandatory for everyone)
2. By having a hierarchy of criteria and using the AHP process. e.g. Environmental, Social, Institutional at upper level and within each of those have multiple sub criteria (for e.g. poverty rates, minority population etc. are part of Social). Please make sure to ensure consistency using the AHP process
3. By including soil characteristics from USGS (https://nrcs.app.box.com/v/soils)
4. By accounting for size of the site.
5. By accounting for cost of parcel assembly.
6. By accounting for the fact that some highways are limited access highways and can only be acccessed at the interchanges.

Other Deliverables

- Final term paper proposal due on Wiki in Sakai by Friday 6 PM. The proposal is no more than 2 pages and should include succinctly, the research questions, datasets that are being used and the appropriateness of analytical techniques to the research question and the datasets. You should have acquired the datasets and performed exploratory analyses before submitting the proposal. The proposal should demonstrate both suitability and feasibility.

Resources/Readings

- (McCarty and Kaza 2015)
- (Watson and Hudson 2015)
- (Rincón, Khan, and Armenakis 2018)

Mar 21 (Mon) Dimensionality Reduction

Tutorials/Slides

Homework

- [HW7 Posted]

Resources/Readings

- (Frenkel and Ashkenazi 2008)
- (Clifton et al. 2008)
- (Golan et al. 2019)

Mar 28 (Mon), Mar 30 (Wed) Supervised Classification with Trees and Forests

Tutorials/Slides

- Classifying Remote Sensing Images

Homework

HW 8

Efforts to protect development along the coast through the placement of coastal protection infrastructure have fundamentally changed the composition of shorelines along the U.S. coast (Gittman et al. 2015). Coastal protection infrastructure, also called shoreline armoring, are physical structures typically made of rock or concrete that are placed along ocean-facing and inland shorelines in order to either (a) offer protection from storm surges and flooding, or (b) stabilize coastal land and halt erosion. Examples of these types of structures include seawalls, riprap, rock revetments, and bulkheads. These protections rarely accomplish their goals. Despite this, owners (including govts.) continue to invest in this infrastructure.

In this homework, I want you to test the efficacy of different predictive (non-linear) model of whether or not a parcel is armored or not. The data set is coastal parcels for two counties in FL (posted in Sakai), which contains the variable `type` that refers to whether th parcel is armored or not. TYou have a wide range of options with regards to the predictors. You are welcome to construct additional predictors by taking advantage of the parcel boundaries and location (such as distance to highway interchange or size of the parcel etc.). Make sure that report on your preferred predictive model based on different accuracy statistics and validation. What are the most important/useful predictors?

Readings/Resources

- (Reades, De Souza, and Hubbard 2019)
- (Stevens et al. 2015)
- (Tribby et al. 2017)

Apr 4 (Mon), Apr 6 (Wed) Clustering & Unsupervised Classification

Tutorials/Slides

- Mapping Crime Clusters in Manchester
- Unsupervised Classification of Non-spatial Data

Homework/Deliverables

HW 9

Frank Baumgartner's most recent book is *Suspect Citizens* (Cambridge, 2018), focusing on racial differences in the outcomes of routine traffic stops. In this homework, I want you to analyse a small portion of the summary dataset (by police agency), I acquired from him, to identify clusters of patterns. You are welcome to choose a subset of the variables, but you should pick at least some "Stop" and "Search" variables such as "Total stops" and "Total number of searches of Hispanic drivers." Noting that the data is from different years, construct a reasonable 'cross-sectional' dataset and tell a compelling story about the nature of traffic stops in the United States. The dataset, along with the code book is posted on Sakai.

Resources/Readings

- (Bates 2006)
- (Clapp and Wang 2006)
- Chapters 7 & 9 (Bivand, Pebesma, and Gómez-Rubio 2013)

Apr 11 (Mon), Apr 13 (Wed) Neural Networks & Deep Learning

Tutorials/Slides

Homework

- [HW 10 Posted]

Readings/Resources

- Chapters 1-5 (Chollet and Allaire 2018)
- (Gebru et al. 2017)
- (Law, Brooks, and Russell 2019)

Apr 18 (Mon), Apr 20 (Wed) Individual Work on Assignment 3

Apr 25 (Mon), Apr 27 (Wed) Assignment 3 Presentations (subject to change based on exam day)

References

Bates, Lisa K. 2006. "Does Neighborhood Really Matter?: Comparing Historically Defined Neighborhood Boundaries with Housing Submarkets." *Journal of Planning Education and Research* 26 (1): 5–17. https://doi.org/10.1177/0739456X05283254.

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.

Boeing, Geoff. 2019. "Urban Spatial Order: Street Network Orientation, Configuration, and Entropy." *Applied Network Science* 4 (1): 1–19. https://doi.org/10.1007/s41109-019-0189-1.

Boeing, Geoff, and Paul Waddell. 2017. "New Insights into Rental Housing Markets Across the United States: Web Scraping and Analyzing Craigslist Rental Listings." *Journal of Planning Education and Research* 37 (4): 457–76. https://doi.org/10.1177/0739456X16664789.

Chen, Yiqiao, Elisabete A. Silva, and José P. Reis. In press. "Measuring Policy Debate in a Regrowing City by Sentiment Analysis Using Online Media Data: A Case Study of Leipzig 2030." *Regional Science Policy & Practice*, In press. https://doi.org/10.1111/rsp3.12292.

Chollet, Francois, and J. J. Allaire. 2018. *Deep Learning with R*. 1 edition. Shelter Island, NY: Manning Publications.

Clapp, John M., and Yazhen Wang. 2006. "Defining Neighborhood Boundaries: Are Census Tracts Obsolete?" *Journal of Urban Economics* 59 (2): 259–84. https://doi.org/http://dx.doi.org/10.1016/j.jue.2005.10.003.

Clifton, Kelly, Reid Ewing, Gerrit-Jan Knaap, and Yan Song. 2008. "Quantitative Analysis of Urban Form: A Multidisciplinary Review." *Journal of Urbanism: International Research on Placemaking and Urban Sustainability* 1 (1): 17–45. https://doi.org/10.1080/17549170801903496.

Frenkel, Amnon, and Maya Ashkenazi. 2008. "Measuring Urban Sprawl: How Can We Deal with It?" *Environment and Planning B: Planning and Design* 35 (1): 56–79. https://doi.org/10.1068/b32155.

Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. "Using Deep Learning and Google Street View to Estimate the Demographic Makeup of Neighborhoods Across the United States." *Proceedings of the National Academy of Sciences* 114 (50): 13108–13. https://doi.org/10.1073/pnas.1700035114.

Golan, Yael, Nancy Wilkinson, Jason M. Henderson, and Aiko Weverka. 2019. "Gendered Walkability: Building a Daytime Walkability Index for Women." *Journal of Transport and Land Use* 12 (1). https://doi.org/10.5198/jtlu.2019.1472.

Grolemund, Garrett, and Hadley Wickham. 2017. *R for Data Science*. First. Sebastapol, CA: O' Reilly. http://r4ds.had.co.nz/.

Kaza, Nikhil, and Katherine Nesse. 2021. "Characterizing the Regional Structure in the United States: A County-based Analysis of Labor Market Centrality." *International Regional Science Review* 44 (5): 560–81. https://doi.org/10.1177/0160017620946082.

Kitchin, Rob. 2016. "The Ethics of Smart Cities and Urban Science." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 374 (2083). https://doi.org/10.1098/rsta.2016.0115.

Law, Stephen, Piage Brooks, and Chris Russell. 2019. "Take a Look Around: Using Street View and Satellite Images to Estimate House Prices." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10 (5). https://dl.acm.org/doi/abs/10.1145/3342240.

Lovelace, Robin, Jakub Nowosad, and Jannes Muenchow. 2019. *Geocomputation with R*. 1 edition. Boca Raton: Chapman and Hall/CRC. https://geocompr.robinlovelace.net/.

McCarty, J., and N. Kaza. 2015. "Urban Form and Air Quality in the United States." *Landscape and Urban Planning* 139: 168–79. https://doi.org/10.1016/j.landurbplan.2015.03.008.

Mi, Zhifu, Jing Meng, Heran Zheng, Yuli Shan, Yi-Ming Wei, and Dabo Guan. 2018. "A Multi-Regional Input-Output Table Mapping China's Economic Outputs and Interdependencies in 2012." *Scientific Data* 5 (1): 1–12. https://doi.org/10.1038/sdata.2018.155.

Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. 1 edition. Chichester, West Sussex, United Kingdom: Wiley.

Nelson, Garrett Dash, and Alasdair Rae. 2016. "An Economic Geography of the United States: From Commutes to Megaregions." *PLOS ONE* 11 (11): e0166083. https://doi.org/10.1371/journal.pone.0166083.

Reades, Jonathan, Jordan De Souza, and Phil Hubbard. 2019. "Understanding Urban Gentrification Through Machine Learning." *Urban Studies* 56 (5): 922–42. https://doi.org/10.1177/0042098018789054.

Resch, Bernd, and Michael Szell. 2019. "Human-Centric Data Science for Urban Studies." *ISPRS International Journal*

of Geo-Information 8 (12): 584. https://doi.org/10.3390/ijgi8120584.

Rincón, Daniela, Usman T. Khan, and Costas Armenakis. 2018. "Flood Risk Mapping Using GIS and Multi-Criteria Analysis: A Greater Toronto Area Case Study." *Geosciences* 8 (8): 275. https://doi.org/10.3390/geosciences8080275.

Schweitzer, Lisa. 2014. "Planning and Social Media: A Case Study of Public Transit and Stigma on Twitter." *Journal of the American Planning Association* 80 (3): 218–38. https://doi.org/10.1080/01944363.2014.980439.

Singleton, Alex David, Seth Spielman, and David Folch. 2018. *Urban Analytics*. First edition. Los Angeles: SAGE Publications Ltd.

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. 2015. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data." *PLOS ONE* 10 (2): e0107042. https://doi.org/10.1371/journal.pone.0107042.

Tribby, Calvin P., Harvey J. Miller, Barbara B. Brown, Carol M. Werner, and Ken R. Smith. 2017. "Analyzing Walking Route Choice Through Built Environments Using Random Forests and Discrete Choice Techniques." *Environment & Planning B : Urban Analytics & City Science* 44 (6): 1145–67. https://doi.org/10.1177/0265813516659286.

Tufte, E. R. 2001. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Watson, Joss J. W., and Malcolm D. Hudson. 2015. "Regional Scale Wind Farm and Solar Farm Suitability Assessment Using GIS-assisted Multi-Criteria Evaluation." *Landscape and Urban Planning* 138 (June): 20–31. https://doi.org/10.1016/j.landurbplan.2015.02.001.

Wickham, Hadley. 2010. "A Layered Grammar of Graphics." *Journal of Computational and Graphical Statistics* 19 (1): 3–28. https://doi.org/10.1198/jcgs.2009.07098.