Sociospatial Data Science

Christopher Prener, Ph.D.

2017-09-22

Contents

Preface

This text is a companion text for both of my research methods courses at Saint Louis University:

- SOC 4650/5650 - Introduction to Geographic Information Science
- SOC 4930/5050 - Quantitative Analysis: Applied Inferential Statistics

The goal of the text is to create a reference for the intangible, subtle or disparate skills and ideas that contribute to being a successful *computational* social scientist. In writing this text, I draw inspiration from the work of Donald Knuth.[1] Knuth has discussed his experiences in designing new software languages, nothing that the developer of a new language

> …must not only be the implementer and the first large-scale user; the designer should also write the first user manual… If I had not participated fully in all these activities, literally hundreds of improvements would never have been made, because I would never have thought of them or perceived why they were important…

While there is nothing particularly new about what I am writing here, and I am certainly not developing a new language for computing, the goal of this text remains similar to Knuth's experience. By distilling some of key elements for making a successful transition to being a *professional developer* of knowledge rather than a *casual consumer*, I hope to both improve the course experience itself and also create an environment that fosters a successful learning experience for you.

In both classes, the course names are deceptive. We are not only concerned with statistical work or mapping. Rather, we are more fundamentally concerned with research methods. In particular, we are concerned with *high quality* research methods and the *process* of conducting research. We therefore focus on a combination of mental habits and technical practices that make you a successful researcher. Some of the skills and techniques

[1] Donald Knuth is the developer of TeX, a computer typesetting system that is widely used today for scientific publishing in the form of LaTeX. He also established the concept of literate programming, which forms the basis of some of the practices we follow with `R`.

that we will discuss this semester are not taught as often in graduate programs, let along undergraduate programs. Instead, they are often the products of "learning the hard way". These "habits of mind and habits of method" are broadly applicable across methodologies and disciplines.

License

Chapter 1

Introduction

The first part of this text is designed to help get you oriented to coursework in computational social science. Both Introduction to Geographic Information Science (SOC 4650/5650) and Quantitative Analysis: Applied Inferential Statistics (SOC 4930/5050) are focused on building students' capacities to address social science research questions using tools that have been the traditional domain of computer and information scientists. The growth and application of these tools in a variety of disciplines both inside and outside of the social sciences has come to be known as data science. Both courses are taught from this perspective, so that while we focus on social science data, the tools and techniques are broadly applicable across disciplines.

1.1 What is data science?

Given data science's new emergence, its definition remains both contested and often unclear. For me, there are four key aspects to focus on when considering what constitutes data science:

1. Statistics
2. Programming
3. Visualization and Communication
4. Substantive Knowledge

I think of this as a "full stack" approach to computational research. You want to be well versed not only the techniques for generating and analyzing data, but also substantively in the academic literature about your area of interest as well as ways to communicate your findings with other researchers and the wider public.

1.1.1 Statistics

Statistics covers the mathematical techniques that we use to draw inference from our data. It is the main subject of one my courses, Quantitative Analysis. In Introduction to GIS, we do not explicitly cover much in the way of inferential statistics. However, the course is desinged to prepare you for a next level, Intermediate GIS course that covers spatial statistics.

1.1.2 Programming

Computer programming is an essential part of data science broadly and computational social science more specifically. Using a programming language means that our work can be easily reproduced. This emphasis on **reproducibility** is a response to a growing fear in many disciplines that results are replicable from study

to study, which raises questions about the validity of much of the research work that we do. Our goal in both courses is to produce research that is as reproducible as possible.

This book introduces two programming languages - `R` and Python. In Quantitative Analysis, we will be focused exclusively on learning `R` and using it to produce statistical analyses. In Introduction to GIS, we will spend a lot of time in `R`, but we will also use some Python to help pass data from `R` to ArcGIS, the mapping application we will use. I will also provide some additional Python lessons to folks who are interested, but these will not be required for the course.

1.1.3 Visualization & Communication

Visualization is a fundamental aspect of data science work. It is how we make our results easily digestible and accessible to a wider audience, many of whom may not be able to interpret statistical output but can learn from a well-designed scatter plot. In Quantitative Analysis, we will focus on building plots to communicate information about statistical distributions and the relationships between our variables. In Introduction to GIS, we will use the same fundamental skills to build simple maps in `R`. We will extend our emphasis on visualization to ArcGIS, where we will focus on producing cartographically rich depictions of our data.

Separetely, we will discuss the presentation of statistical data in tables using LaTex in Quantitative Analysis as well as the producing to conference-style presentations to communicate research findings. In Introduction to GIS, we will focus on producing conference-style posters instead. These are different mediums, but they rely on the same design fundamentals that are covered in this text.

1.1.4 Substantive Knowledge

Substantive knowledge covers two tangentially related topics: the ability to work well in groups and the ability to digest and integrate an academic literature into your own research. Each of these topics receive some focus in both Quantitative Analysis and Introduction to GIS. Each class has some group work associated with the completion with weekly lab assignments. Introduction to GIS also has a group work component associated with the final project. In each class, students' final projects are focused on a specific content area that requires at least some background research and knowledge. Synthesizing this knowledge and integrating it into your final projects is a key piece of addressing this facet of data science.

Part I

First Steps

Chapter 2

Getting Started

Before you begin the semester, there are a number of things that I recommend that you do to help set yourself up for success. Before you do *anything* else, you should read through the **Syllabus** and the **Reading List**. Make sure you have a good sense of what is *required* for the course. If you have questions, bring them to the first day of class!

2.1 Account Signups

2.1.1 Get Started with Slack

We'll be using the messaging platform Slack as a space for "virtual office hours". Slack is a messaging system used by teams of all kinds. If you can text, you can use Slack. You will need to sign-up for the SOC 5050 Slack organization here. You will need to complete the signup process even if you use Slack for other purposes. Consider installing either the desktop or the mobile apps for Slack to keep in touch and receive push alerts!

2.1.2 Get Started with GitHub

The website that is hosting this wiki is called GitHub. GitHub is used by programmers, data scientists, and researchers for hosting computer code, data, and project materials. We will be using GitHub extensively this semester. You will need a free account, which you can sign up for one from GitHub's homepage. If you already have a GitHub account, you do not need a new one. *Once you have a GitHub user name, send Chris a Direct Message via Slack with it so that you can be added to the SOC 5050 organization.*

2.1.3 Get Started with LaTeX

We'll be doing a little bit of writing using LaTeX, which is a markup language that makes technical writing easier. We'll be using ShareLaTeX this semester for this purpose. ShareLaTeX is a bit like Google Docs, but for LaTeX. It is a "freemium" service - please don't pay for any additional features - you won't need them! You can sign-up for ShareLaTeX on their website.

2.2 Get Started with Software

If you will be using your own computer in class, you'll want to install a number of applications. If you aren't using your own computer, you can skip this section! All of these applications are available in our classroom, and - lucky you - you get 24-hour access to Morrissey Hall for the semester.

2.2.1 Computer Prep

Before you install your software, you should do the following:

1. Make sure your operating system is up-to-date. If you are able, I would also recommend upgrading your computer to the most recent release of its operating system that the computer can run.
2. We'll be sharing computer files throughout the semester, so you should ensure that you have functioning anti-virus software and that it is up-to-date. You can get anti-virus software for free from SLU. Go to `ITS Software Downloads` under `Tools` on mySLU.
3. You'll also need to download files, so you'll need to make sure you have some free space on your hard drive. If you have less than 10GB of free space, you should de-clutter!
4. Make sure you know how to access your computer's file management system.

- On macOS, this means being comfortable with Finder.app.
- On Windows, this means being comfortable with Windows Explorer.

2.2.2 Software Installation

Now that your computer is up-to-date

1. The computing language `R` needs to be downloaded and installed. You can download it from the University of California-Berkeley. Choose "Download R for (Mac) OS X" or "Download R for Windows".
2. RStudio is a graphical user interface for `R` that will make learning the language and using it much, much easier. You should download the *free* version of RStudio from their website. Choose the installer for your platform, and ping me on Slack if you have any questions.
3. GitHub Desktop is a client for interacting with GitHub that makes downloading and uploading files a breeze. You can download it from the developer's website.
4. Atom is a text editor that is produced by the same folks who operate GitHub. Download Atom from the developer's website.

2.3 Get Access to Books and Readings

2.3.1 Books

There are three books required for this course. Each book has been selected to correspond with one or more of the course objectives. The books are:

1. Freedman, David, Robert Pisani, and Roger Purves. 2007. *Statistics*. 4th edition. New York, NY: W.W. Norton and Company.
2. Wheelan, Charles. 2014. *Naked Statistics: Stripping the Dread from the Data*. New York, NY: W.W. Norton and Company.

3. Wickham, Hadley and Garrett Grolemund. 2016. *R for data science*. Sebastopol, CA: O'Reilly. Webbook Available.

All of the books are available in the bookstore. They can also be ordered online. If you would rather use ebooks, those are acceptable for this course as well.

2.3.2 Check Out the Readings for Week 01

All but one of the Week 01 readings are available on our course's electronic reserves site, and the password is posted in Slack on the `#helpdesk-coursework` channel. The initial section of Wickham and Grolemund can be found via the webbook.

2.4 Administrative Tasks

There are two forms that all students must fill out by Tuesday, September 5th:

1. the Student Information Sheet, which gives me some info about you and gives you the chance to let me know about any initial concerns you might have.
2. the un-graded Diagnostic Assessment, which is designed to get a sense of where each student's math skills are currently. Please don't consult outside materials as you do this - if you are not sure how to answer, make the most educated guess you can and move on. If you look answers up it defeats the purpose of this exercise!

Bibliography