The Journal of Open Source Education

A course on Geographic Data Science

Dani Arribas-Bel[1, 2]

1 Geographic Data Science Lab **2** The University of Liverpool

DOI: 10.21105/jose.00042

Software

- Review
- Repository
- Archive

Submitted: 05 January 2019
Published: 26 April 2019

Summary

This paper presents a computational learning module on Geographic Data Science (GDS). This resource is part of a larger set that also includes a series of lecture slides, that has been used for four consecutive academic years to teach the course "Geographic Data Science" at the University of Liverpool.

Statement of need

Data Science (Donoho, 2017) has become one of the most demanded skills thanks to an explosion in the availability of data (Kitchin, 2014). Most of these new sources are, directly or indirecly, geographic in that they can be related to a particular location on a map. However, the vast majority of data science resources available currently ignore the spatial dimension of data, particularly when it comes to the more analytic set of methods covered. At the same time, traditional resources for teaching the handling, visualisation, and analysis of geographic data are based on a paradigm that emphasises graphical interfaces and "point-and-click" software packages. This approach, although valid, limits the flexiblity with which the analyst can effectively move from data to insights, and is more difficult to connect with and benefit from modern advances in both data tools and workflows. This paper presents a pedagogical bridge between the "spatially unaware" set of practices emerging from Data Science, and more traditional resources designed to teach spatial analysis within a Geographic Information Systems (GIS) environment.

Learning objectives

Upon completion, students are able to:

- Demonstrate advanced GDS concepts and be able to use the tools programmatically to import, manipulate and analyse spatial data stored in a variety of formats.
- Understand the motivation and inner workings of the main methodological approcahes of GDS, both analytical and visual.
- Critically evaluate the suitability of a specific technique, what it can offer and how it can help answer questions of interest.
- Apply a number of spatial analysis techniques in Python and explain how to interpret the results, in a process of turning data into information.
- When faced with a new data-set, work independently using GDS tools programmatically to extract valuable insight.

Content

The module represents the computational element of a larger resource used in the delivery of the course "Geographic Data Science" at the University of Liverpool. Materials are organised and made available to the students through the course website, which can be found at:

http://darribas.org/gds18/

The content is organised in three main blocks, each of them with a similar amount of material, and designed to take similar length in its delivery:

- The **first part** (notebooks `lab_00` to `lab_02`) introduces the student to the main computational tools that will become the building blocks of the course. This includes the *Jupyter Notebook*, as well as the basics of non-spatial data manipulation and visualisation using `pandas` and `matplotlib`.
- The **second part** (notebooks `lab_03` to `lab_05`) builds on the previous one and shows how several of the programming patterns learnt for non-spatial data apply directly, or very similarly, when the data provided have a spatial signature. In addition, this part uses Python programming, `geopandas` and `pysal` to introduce the student to building blocks of spatial analysis such as (choropleth) mapping and spatial weights matrices.
- The **third part** (notebooks `lab_06` to `lab_09`) relies on the previous two to show the intuition and application of more advanced, explicitly spatial computational techniques. In particular, this set of notebooks covers exploratory spatial data analysis (ESDA, Anselin, 1999), point patterns (Boots & Getis, 1988), and unsupervised learning, including both geodemographic analysis (Harris, Sleight, & Webber, 2005) as well as regionalisation algorithms (Duque, Ramos, & Suriñach, 2007) that impose an additional spatial constraint.

Taken altogether, the computational module can be seen as a collection of independent learning objects (Norman & Porter, 2007) that can be used individually or repurposed for different contexts, but that also form a coherent learning program that allows the student to progress from basic to more advanced concepts in both Geographic Data Science and Python programming.

Instructional design

There are ten Jupyter notebooks, each covering materials that would usually be presented to the student at a rate of one per week. Since the course expects no previous knowledge on the core theoretical concepts, students first attend a 1h lecture every week where the main ideas behind each notebook are presented, and they are encouraged to examine the notebook before the computer lab (2h/week), following a semi-flipped classroom approach (Lage, Platt, & Treglia, 2000). In the lab, following an enquiry-based learning approach (Hutchings, 2006), they are expected to work individually and at their own pace through each notebook. The student-centered character of the course is complemented through two channels of support. First, within the lab, at least one instructor for every 15 students answers questions one-on-one or in small groups. Second, the class has access to an online discussion forum monitored by the course leader. Within this environment, students are encouraged to post not only questions but also responses both to other students and to their original enquiry, in case they solve it themselves.

Each notebook is conceived as a self-contained computational narrative where theoretical concepts are threaded with programming illustrations. The aim of this strategy is twofold. On the one hand, the programming element presents a vehicle for the student to experiment with more abstract ideas, as well as to illustrate their relevance in practical applications using real world data. On the other hand, the geographic and data science topics covered in class act as a specific case in which several programming techniques such as loops, variables and functions, are used. By embedding these more general notions into a particular context, students are able to experience their utility as tools to solve problems they may face, rather than as abstract computational ideas. In addition, the self-contained nature of each notebook facilitates their use in a variety of contexts, from a full-fledged semester course such as "Geographic Data Science", to shorter intensive bootcamps, or even as repurposed show'n'tell independent master classes.

Experience of use

The materials in this computational resource have been used, updated and maintained for over three years, since the first iteration of "Geographic Data Science" (http://darribas.org/gds15/) was delivered. In this time, the materials have been repurposed for a multiplicity of contexts, with different goals and for different audiences. Most of these experiences fall under the following two scenarios:

- **Semester-long course**: for four consecutive years (2015, 2016, 2017, and 2018), these materials have formed the computational backbone of the "Geographic Data Science" course taught at the University of Liverpool. The course is offered to Year 3 undergraduate students in Geography and Planning, and to MSc students across campus. Every year, a varying cohort from 80 to 120 students with diverse degrees of prior programming experience take the course.
- **Workshops**: the module has also provided building blocks to deliver intensive workshops of length between one half and three days. The audiences for these shorter events range from social scientists with experience in programming but not in geographic analysis, to practitioners with experience in GIS but not in programming. In all these cases, one or more of the notebooks have formed the basis of the materials used to deliver computer sessions. Modifications relate mostly to adjusting the length to fit into the required time, or swaping the data used to provide more relevant examples.

References

Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis. *Geographical Information Systems: principles, techniques, management and applications*, *1*, 251–264.

Boots, B. N., & Getis, A. (1988). *Point pattern analysis* (Vol. 8). Sage Publications, Inc.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. doi:10.1080/10618600.2017.1384734

Duque, J. C., Ramos, R., & Suriñach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review*, *30*(3), 195–220. doi:10.1177/0160017607301605

Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting* (Vol. 7). John Wiley; Sons.

Hutchings, B. (2006). Principles of enquiry-based learning. *Centre for Excellence in Enquiry-Based Learning Resources, University of Manchester, England*. Retrieved from http://www.ceebl.manchester.ac.uk/resources/papers/ceeblgr002.pdf

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education*, *31*(1), 30–43. doi:10.2307/1183338

Norman, S., & Porter, D. (2007). Designing learning objects for online learning. *Commonwealth of Learning*. Retrieved from http://hdl.handle.net/11599/45