ENVS3/563 - Module Handbook

Dani Arribas-Bel

Contents

ENVS363/563

Geographic Data Science

Welcome to Geographic Data Science, a course taught by Dr. Dani Arribas-Bel in the Autumn of 2016 at the University of Liverpool.

The timetable for the course is:

- **Lectures**: *Tuesdays* 9:00am-10:00am, `REN-LT8`
- **Computer Labs**: *Fridays* 3:00pm-5:00pm, `ENG-HHTC`

Locations

- `REN-LT8`: Rendall Building, Lecture Theatre 8 [Map]
- `ENG-HHTC`: Harrison Hughes Building (Engineering), Computer lab (top floor) [Map]

Contact

Dani Arribas-Bel - D.Arribas-Bel [at] liverpool.ac.uk
Lecturer in Geographic Data Science
Office 508, Roxby Building,
University of Liverpool - 74 Bedford St S,
Liverpool, L69 7ZT,
United Kingdom.

Overview

Aims

The module provides students with core competences in Geographic Data Science (GDS). This includes the following:

- Advancing their statistical and numerical literacy.
- Introducing basic principles of programming and state-of-the-art computational tools for GDS.
- Presenting a comprehensive overview of the main methodologies available to the Geographic Data Scientist, as well as their intuition as to how and when they can be applied.
- Focusing on real world applications of these techniques in a geographical and applied context.

Learning outcomes

By the end of the course, students will be able to:

- Demonstrate advanced GIS/GDS concepts and be able to use the tools programmatically to import, manipulate and analyse spatial data in different 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 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 GIS/GDS tools programmatically to extract valuable insight.

Feedback strategy

The student will receive feedback through the following channels:

- Formal assessment of two summative assignments. This will be on the form of reasoning of

the mark assigned as well as comments specifying how the mark could be improved. This will be provided no later than three working weeks after the deadline of the assignment submission.
- Direct interaction with Module Leader and demonstrators in the computer labs. This will take place in each of the scheduled lab sessions of the course.
- Online forum maintained by the Module Leader where students can contribute by asking and answering questions related to the module.

Key texts and learning resources

Access to materials, including lecture slides and lab notebooks, is centralized through the use of a course website available in the following url: `http://darribas.org/gds16`

Specific readings, videos, and/or podcasts, as well as academic references will be provided for each lecture and lab, and can be accessed through the course website.

Syllabus

Week 1: Introduction

- Lecture: Geographic Data Science.
- Tutorial: Manipulating data in Python - Tidy Data.

Week 2: Spatial Data

- Lecture: Geo-Data.
- Tutorial: Manipulating data in Python - Advanced Tricks.

Week 3: (Geo)Visualization

- Lecture: (Geo)Visualization.
- Tutorial: Manipulating geospatial data in Python.

Week 4: Choropleth mapping

- Lecture: Choropleth mapping.
- Tutorial: Mapping deprivation.

Week 5: Spatial Weights

- Lecture: Spatial Weights.
- Tutorial: Spatial Weights with PySAL.

Week 6: ESDA

- Lecture: Exploratory Spatial Data Analysis (ESDA).

- Tutorial: ESDA in Python.

Week 7: Clustering

ASSIGNMENT 1 due on Wednesday, November 9th-2016.

- Lecture: Clustering.
- Tutorial: Geodemographic analysis.

Week 8: Point Data

- Lecture: Point Data.
- Tutorial: Exploring Twitter patterns.

Week 9: (Spatial) causal inference

- Lecture: Spatial causal inference.
- Tutorial: Assignment Clinic.

Week 10

- Lecture: Assignment 1 recap.
- Tutorial: Assignment Clinic.

Week 11: Geographic Data Science in Action

- Lecture: Geographic Data Science *in the wild*.
- Tutorial: Assignment Clinic.

Week 12

ASSIGNMENT 2 due on Monday, December 12th-2016.

Assessment

The final mark for the course is composed of the following three components:

- Assignment 1 (47.5%)
- Assignment 2 (47.5%)
- Contribution to module's discussion forum (5%)

Assignments 1 and 2 are described below. Students should keep in mind the following information regarding the submission of assignments:

- **Submission is electronic only** and will be managed through Turnitin.
- Assignments will be prepared in the Jupyter Notebook file format and then converted into a self-contained HTML file that will then be submitted on Turnitin.

The contribution to the discussion forum is an all-or-nothing 5% of the mark that can be obtained by contributing *meaninfully* to the online discussion board setup for the course. *Meaningful* contributions include both questions and answers that demonstrate the student is committed to make the forum a more useful resource for the rest of the group.

Marking Criteria

This course follows the standard marking criteria (the general ones and those relating to GIS assignments in particular) set by the School of Environmental Sciences. In addition to these generic criteria, the following specific criteria relating to the code provided will be used:

- **0-15**: the code does not run and there is no documentation to follow it.
- **16-39**: the code does not run, or runs but it does not produce the expected outcome. There is some documentation explaining its logic.
- **40-49**: the code runs and produces the expected output. There is some documentation explaining its logic.
- **50-59**: the code runs and produces the expected output. There is extensive documentation explaining its logic.
- **60-69**: the code runs and produces the expected output. There is extensive documentation, properly formatted, explaining its logic.
- **70-79**: all as above, plus the code design includes clear evidence of skills presented in advanced sections of the course (e.g. custom methods, list comprehensions, etc.).
- **80-100**: all as above, plus the code contains novel contributions that extend/improve the functionality the student was provided with (e.g. algorithm optimizations, novel methods to perform the task, etc.).

Assignment 1 - *Raising awareness of multiple deprivation*

- Type: `Coursework`
- [Equivalent to 2,500 words] Maps, code and 500 words.
- Due on **Wednesday, November 9th-2016** (Week 7).
- 47.5% of the final mark
- Chance to be reassessed
- Electronic submission only.

In this assignment, you will take the role of the data editor of a local newspaper that wants to write about the geography of deprivation. In order to raise awareness of the problem among your readers, you will have to create a compelling visualization that is intuitive and attractive but also rigorous. In addition, in order to convince your most skeptical and data-savvy readers, you will have to provide the code used to create the visualization in a way that allows reproducibility.

Using data from the Index of Multiple Deprivation, as well as from the Census, create at least three and no more than five maps to display different angles and interesting patterns related to multiple deprivation in a British town other than Liverpool. Complement the maps with a short

description of what they show, stressing the relevant aspects you would want your readers to focus on. Keep in mind this needs to be short and to the point, as the report will be passed to a journalist who will draft the final article for the newspaper. In addition to the figures and text, provide data and annotated code that allows to replicate the visualization.

Minimum requirements (complete *all*)

- Choose a city/local authority in the UK that is not Liverpool, preferably one you know.
- Obtain the Index of Multiple Deprivation as well as census demographic data.
- Compose a map with different layers.
- Include a "zoom" of the global map by subsetting the original data.
- At least three and no more than five maps.
- Up to 500 words describing the patterns in the maps.

Optional suggestions (include *at least one*)

- Join deprivation indices from areas to building data to create a more aesthetic visualization.
- Discuss deprivation at different geographical scales.
- Compare the effect of different choropleth classification algorithms on visualizing deprivation.
- Explore the composition of the multiple deprivation as it relates to Census more basic variables (e.g. income, building age, etc.).
- Try to characterize the overall pattern found in the maps (is it concentrated, dispersed, agglomerated into different hotspots or something totally different?).
- Cross-check empirical findings with "common wisdom" about the areas where you have local knowledge.
- Begin to explore the underlying social processes for the empirical findings of your analysis.
- Exchange some of the maps for non-spatial graphics (scatter plots, bar charts, etc.).

Data

- CDRC Census Geodata pack.
- CDRC Census Data pack.
- 2015 Index of Multiple Deprivation.

Assignment 2 - *Targetting areas*

- Type: `Coursework`
- [Equivalent to 2,500 words] Three maps/tables, code and 750 words.
- 47.5% of the final mark
- Chance to be reassessed
- Final Assessment
- Due on **Monday, December 12th-2016** (Week 12).
- Electronic submission only.

In this assignment, you will take the role of a real-world data scientist tasked to identify areas to direct investments. You are consulting for the City of Liverpool on a program to target investments towards particularly disadvantaged areas that are nevertheless popular or have the

potential to become so. The Economic Development division knows that only five local super output areas (LSOAs) will be funded but would like to know which ones.

Choose one of the given questions, develop a data strategy, deploy it, and present the results in a rigorous but intuitive fashion, together with the code.

Minimum requirements (complete *all*)

- Combine *at least two* datasets, potentially among those used in the course.
- Employ *at least two* techniques from the set of analytics covered in the course.
- Justify why you have chosen the methods you use and how they help you answer the question at hand. Critically discuss their limitations too.
- Provide a list of the top five areas that you recommend be funded for improvement.
- Explain clearly how you have arrived at the list and how the decision has been informed by the data analysis and the methodologies employed.
- Include documented code and data that allow the replication of the analysis presented.

Suggestive paths (optional)

- Combine a LISA analysis of deprivation with kernel density maps of Twitter activity to identify areas of high values at both.
- Combine several relevant variables into a geodemographic analysis to obtain candidate areas and display the results in an aesthetically pleasant choropleth.

Data

This assignment can use any of the datasets employed in the course, and/or any other datasets you consider useful. If you are thinking of including additional datasets, or have ideas in this respect, please get in touch with the module lead (Dani Arribas-Bel).