Grouping Data

⚠ Changes before 22/23

- □ Point to clustering code and show (A/H)DBSCAN, k-Means, etc. as applied to the Airbnb data.
- □ Check can run SOM code. Are there any other algorithms we should cover?
- □ Add 'Module Dialogue' form, e.g. (to be tweaked):
  - Do you understand what is being taught on the module?
  - Do you know how you will be assessed?
  - Can you access the learning resources?

Connects to both CASA0001 *and* CASA0007 and CASA0005.

Overview

In this final week we will be focussing in the live session on planning for the Data+Policy Briefing and Individual Reflection, but you should *also* be looking at how this module connects with ideas covered in CASA0001 (UST), CASA0005 (GIS), and CASA0007 (QM). We will also be looking more widely to the future of quantitative methods, the potential of a *geographic* data science, and the ways in which we can move between spatial and non-spatial paradigms of analysis within the same piece of work.

! Important

This week's Learning Outcomes are:

1. An appreciation of how clustering as *part of an analytical pipeline* differs from the material covered in CASA0007 and so enhances our understanding of 'paradigms' in CASA0001.
2. A general appreciation of how different clustering algorithms work and how this differs from classifcation.
3.

Preparation

Lectures

You are *strongly* advised to watch these videos on classification and clustering; *however*, you will not be asked to present any of these because our attention has shifted towards the final assessments. You should, by now, be familiar with the concept of how to cluster data from the QM module (CASA0007), so this week is actually focussed on how to move beyond *k*-means. The point is to contextualise these two approaches as part of a data science 'pipeline' and to contrast to them with the more theoretical aspects covered elsewhere. We are less interested in the *mathematical• and* technical* aspects, and more interested in how one might go about selecting the *appropriate* algorithm for a particular problem.

Session	Video	Presentation	Notes
Classification	Video	Slides	Notes
Clustering	Video	Slides	Notes

Readings

Come to class prepared to briefly present:

- (Shapiro and Yavuz 2017) URL
- (Singleton and Arribas-Bel 2021) DOI

You may also want to look at the following reports / profiles with a view to thinking about employability and how the skills acquired in this module can be applied beyond the end of your MSc:

- Geospatial Skills Report
- AAG Profile of Nicolas Saravia
- (Wolf et al. 2021) DOI

Activities

- Padlet: [Collaborative Agenda]
- Complete the short Moodle quiz associated with this week's activities.

Practical

In the practical we will run through a variety of clustering algorithms with a view to comparing their performance on real-world data, and we will see why k-means should almost never be your *default* choice when encountering a new data set.

Connections

The practical focusses on:

- ☐ The use of dimensionality reduction as a precursor to clustering.
- ☐ The choice of k in k-means and extract of representative centroids.
- ☐ The use of alternative clustering algorithms (DBSCAN, OPTICS, Self-Organising Maps, and ADBSCAN).

The practical can be downloaded from GitHub.

References

Shapiro, W., and M. Yavuz. 2017. "Rethinking 'distance' in New York City." Medium. https://medium.com/topos-ai/rethinking-distance-in-new-york-city-d17212d24919.

Singleton, Alex, and Daniel Arribas-Bel. 2021. "Geographic Data Science." *Geographical Analysis* 53 (1):61–75. https://doi.org/10.1111/gean.12194.

Wolf, Levi John, Sean Fox, Rich Harris, Ron Johnston, Kelvyn Jones, David Manley, Emmanouil Tranos, and Wenfei Winnie Wang. 2021. "Quantitative Geography III: Future Challenges and Challenging Futures." *Progress in Human Geography* 45 (3). SAGE Publications Sage UK: London, England:596–608.