Spatial Modelling for Data Scientists

Francisco Rowe, Dani Arribas-Bel

November 6, 2022

Table of contents

Welcome

This is the new website for the course *Spatial Modeling for Data Scientists* taught by Dr. Francisco Rowe at the University of Liverpool, United Kingdom. The website is *work in progress* and will be fully updated by February 1st 2023. The course provides an intuitive understanding of models and analytical approaches to manipulate, visualise and interrogate spatial data.

 A compilation of this web course is hosted as a GitHub repository that you can access:

- As a download of a `.zip` file that contains all the materials.
- As an html website.
- As a pdf document
- As a GitHub repository.

Contact

Francisco Rowe - `F.Rowe-Gonzalez [at] liverpool.ac.uk`
Senior Lecturer in Quantitative Human Geography
Office 507, Roxby Building,
University of Liverpool - 74 Bedford St S,
Liverpool, L69 7ZT,
United Kingdom.

1 Overview

Access to all materials, including lecture notes, computational notebooks and datasets, is centralised through the use of the course website available in the following url:

https://gdsl-ul.github.io/san/

The module handbook, including the assessment description, criteria and module programme, and videos for each teaching week can be accessed via the module Canvas site:

ENS453 Spatial Modelling for Data Scientists

1.1 Aims

This module aims to provides students with a range of techniques for analysing and modelling spatial data:

- build upon the more general research training delivered via companion modules on *Data Collection and Data Analysis*, both of which have an aspatial focus;
- highlight a number of key social issues that have a spatial dimension;
- explain the specific challenges faced when attempting to analyse spatial data;
- introduce a range of analytical techniques and approaches suitable for the analysis of spatial data; and,
- enhance practical skills in using *R* software packages to implement a wide range of spatial analytical tools.

1.2 Learning Outcomes

By the end of the module, students should be able to:

- identify some key sources of spatial data and resources of spatial analysis and modelling tools;
- explain the advantages of taking spatial structure into account when analysing spatial data;

- apply a range of computer-based techniques for the analysis of spatial data, including mapping, correlation, kernel density estimation, regression, multi-level models, geographically-weighted regression, spatial interaction models and spatial econometrics;
- apply appropriate analytical strategies to tackle the key methodological challenges facing spatial analysis – spatial autocorrelation, heterogeneity, and ecological fallacy; and,
- select appropriate analytical tools for analysing specific spatial data sets to address emerging social issues facing the society.

1.3 Feedback

- *Formal assessment of two computational essays.* Written assignment-specific feedback will be provided within three working weeks of the submission deadline. Comments will offer an understanding of the mark awarded and identify areas which can be considered for improvement in future assignments.
- *Verbal face-to-face feedback.* Immediate face-to-face feedback will be provided during lecture, discussion and clinic sessions in interaction with staff. This will take place in all live sessions during the semester.
- *Online forum.* Asynchronous written feedback will be provided via an online forum maintained by the module lead. Students are encouraged to contribute by asking and answering questions relating to the module content. Staff will monitor the forum Monday to Friday 9am-5pm, but it will be open to students to make contributions at all times.

1.4 Computational Environment

To reproduce the code in the book, you need the most recent version of R and packages. These can be installed following the instructions provided in our R installation guide.

1.4.1 Dependency list

The list of libraries used in this book is provided below. If you have followed the instructions provided in our R installation guide and will be using Docker, you can relax and these libraries have already been installed for you.

If you have `natively` installed R and RStudio, you need to ensure you have installed the list of libraries used in this book following the steps provided here.

- `arm`
- `car`
- `corrplot`

- FRK
- gghighlight
- ggplot2
- ggmap
- GISTools
- gridExtra
- gstat
- jtools
- kableExtra
- knitr
- lme4
- lmtest
- lubridate
- MASS
- merTools
- plyr
- RColorBrewer
- rgdal
- sf
- sjPlot
- sp
- spgwr
- spatialreg
- spacetime
- stargazer
- tidyverse
- tmap
- viridis

1.5 Assessment

The final module mark is composed of the *two computational essays*. Together they are designed to cover the materials introduced in the entirety of content covered during the semester. A computational essay is an essay whose narrative is supported by code and computational results that are included in the essay itself. Each teaching week, you will be required to address a set of questions relating to the module content covered in that week, and to use the material that you will produce for this purpose to build your computational essay.

Assignment 1 (50%) refer to the set of questions at the end of Chapters @ref(points), @ref(flows) and @ref(spatialecon). You are required to use your responses to build your computational essay. Each chapter provides more specific guidance of the tasks and discussion that you are required to consider in your assignment.

Assignment 2 (50%) refer to the set of questions at the end of Chapters @ref(mlm1), @ref(mlm2), @ref(gwr) and @ref(sta). You are required to use your responses to build your computational essay. Each chapter provides more specific guidance of the tasks and discussion that you are required to consider in your assignment.

1.5.1 Format Requirements

Both assignments will have the same requirements:

- Maximum word count: 2,000 words, excluding figures and references.
- Up to three maps, plot or figures (a figure may include more than one map and/or plot and will only count as one but needs to be integrated in the figure)
- Up to two tables.

Assignments need to be prepared in *R Notebook* format and then converted into a self-contained *HTML* file that will then be submitted via Turnitin. The notebook should only display content that will be assessed. Intermediate steps do not need to be displayed. Messages resulting from loading packages, attaching data frames, or similar messages do not need to be included as output code. Useful resources to customise your *R notebook* can be found on the R Markdown website from RStudio:
* A Guide
* R Markdown: The Definitive Guide by Xie, Allaire, and Grolemund (2018)
* R Markdown Reference Guide

Two R Notebook templates will be available via the *module Canvas site*.

Submission is electronic only via Turnitin on *Canvas*.

1.5.2 Marking criteria

The Standard Environmental Sciences School marking criteria apply, with a stronger emphasis on evidencing the use of regression models, critical analysis of results and presentation standards. In addition to these general criteria, the code and outputs (i.e. tables, maps and plots) contained within the notebook submitted for assessment will be assessed according to the extent of documentation and evidence of expertise in changing and extending the code options illustrated in each chapter. Specifically, the following criteria will be applied:

- **0-15**: no documentation and use of default options.
- **16-39**: little documentation and use of default options.
- **40-49**: some documentation, and use of default options.
- **50-59**: extensive documentation, and edit of some of the options provided in the notebook (e.g. change north arrow location).

- **60-69**: extensive well organised and easy to read documentation, and evidence of understanding of options provided in the code (e.g. tweaking existing options).
- **70-79**: all above, plus clear evidence of code design skills (e.g. customising graphics, combining plots (or tables) into a single output, adding clear axis labels and variable names on graphic outputs, etc.).
- **80-100**: all as above, plus code containing novel contributions that extend/improve the functionality the code was provided with (e.g. comparative model assessments, novel methods to perform the task, etc.).

2 Spatial Data

This Chapter seeks to present and describe distinctive attributes of spatial data, and discuss some of the main challenges in analysing and modelling these data. Spatial data is a term used to describe any data associating a given variable attribute to a specific location on the Earth's surface.

2.1 Spatial Data types

Different classifications of spatial data types exist. Knowing the structure of the data at hand is important as specific analytical methods would be more appropriate for particular data types. We will use a particular classification involving four data types: lattice/areal data, point data, flow data and trajectory data (Fig. 1). This is not a exhaustive list but it is helpful to motivate the analytical and modelling methods that we cover in this book.

Lattice/Areal Data. These data correspond to records of attribute values (such as population counts) for a fixed geographical area. They may comprise regular shapes (such as grids or pixels) or irregular shapes (such as states, counties or travel-to-work areas). Raster data are a common source of regular lattice/areal area, while censuses are probably the most common form of irregular lattice/areal area. Point data within an area can be aggregated to produce lattice/areal data.

Point Data. These data refer to records of the geographic location of an discrete event, or the number of occurrences of geographical process at a given location. As displayed in Fig. 1, examples include the geographic location of bus stops in a city, or the number of boarding passengers at each bus stop.

Flow Data. These data refer to records of measurements for a pair of geographic point locations. or pair of areas. These data capture the linkage or spatial interaction between two locations. Migration flows between a place of origin and a place of destination is an example of this type of data.

Trajectory Data. These data record geographic locations of moving objects at various points in time. A trajectory is composed of a single string of data recording the geographic location of an object at various points in time and each record in the string contains a time stamp. These data are complex and can be classified into explicit trajectory data and implicit trajectory data. The former refer to well-structured data and record positions of objects continuously

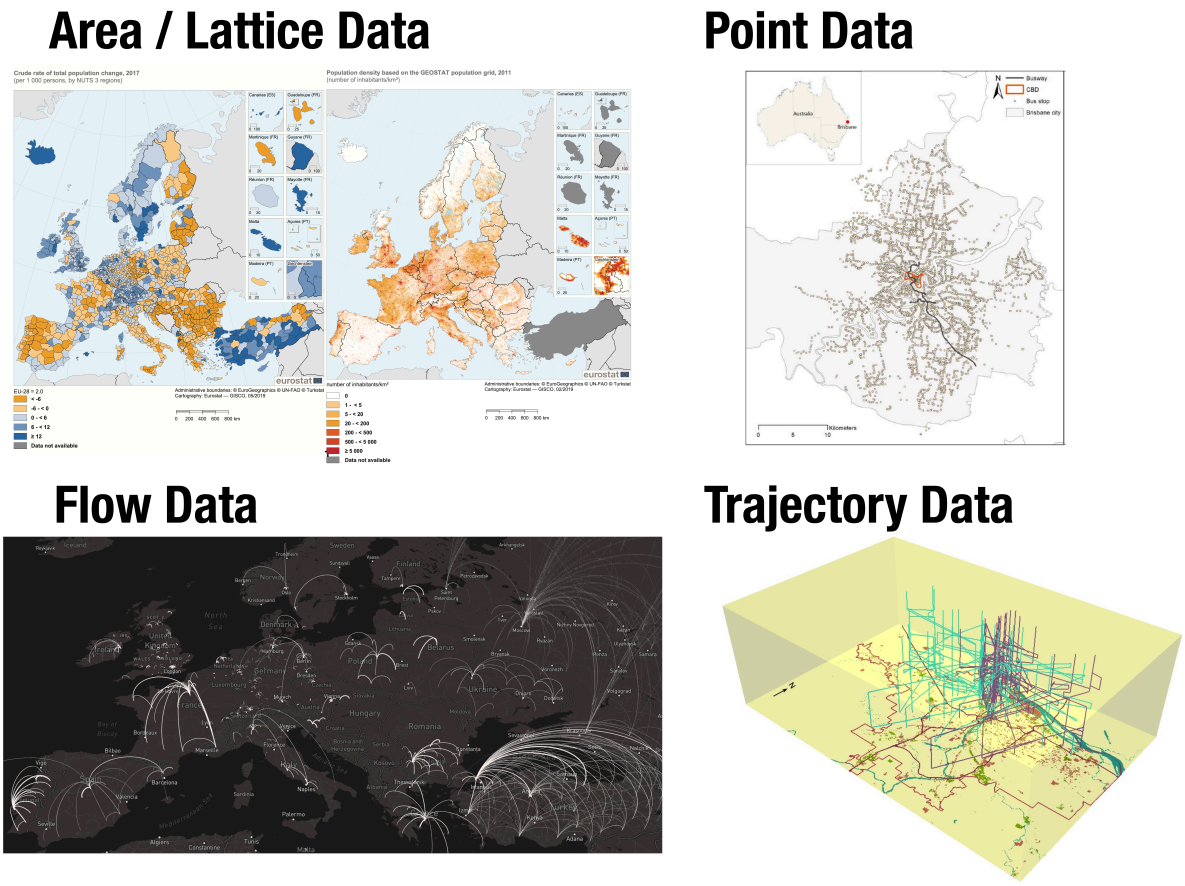

Figure 2.1: Fig. 1. Data Types. Area / Lattice data source: Önnerfors et al. (2019). Point data source: Tao et al. (2018). Flow data source: Rowe and Patias (2020). Trajectory data source: Kwan and Lee (2004).

and intensively at uniform time intervals, such as GPS data. The latter is less structured and record data in relatively time point intervals, including sensor-based, network-based and signal-based data (Kong et al. (2018)).

In this course, we cover analytical and modelling approaches for point, lattice/areal and flow data. While we do not explicitly analyse trajectory data, various of the analytical approaches described in this book can be extended to incorporate time, and can be applied to model these types of data. In Chapter @ref(sta), we describe approaches to analyse and model spatio-temporal data. These same methods can be applied to trajectory data.

2.2 Hierarchical Structure of Data

The hierarchical organisation is a key feature of spatial data. Smaller geographical units are organised within larger geographical units. You can find the hierarchical representation of UK Statistical Geographies on the Office for National Statistics website. In the bottom part of the output below, we can observe a spatial data frame for Liverpool displaying the hierarchical structure of census data (from the smallest to the largest): Output Areas (OAs), Lower Super Output Areas (LSOAs), Middle Super Output Areas (MSOAs) and Local Authority Districts (LADs). This hierarchical structure entails that units in smaller geographies are nested within units in larger geographies, and that smaller units can be aggregated to produce large units.

```
Simple feature collection with 6 features and 4 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 335071.6 ymin: 389876.7 xmax: 339426.9 ymax: 394479
Projected CRS: Transverse_Mercator
      OA_CD   LSOA_CD   MSOA_CD    LAD_CD                       geometry
1 E00176737 E01033761 E02006932 E08000012 MULTIPOLYGON (((335106.3 38...
2 E00033515 E01006614 E02001358 E08000012 MULTIPOLYGON (((335810.5 39...
3 E00033141 E01006546 E02001365 E08000012 MULTIPOLYGON (((336738 3931...
4 E00176757 E01006646 E02001369 E08000012 MULTIPOLYGON (((335914.5 39...
5 E00034050 E01006712 E02001375 E08000012 MULTIPOLYGON (((339325 3914...
6 E00034280 E01006761 E02001366 E08000012 MULTIPOLYGON (((338198.1 39...
```

Next we quickly go through the components of the output above. The first line indicates the type of feature and the number of rows (features) and columns (fields) in the data frame, except for the geometry. The second and third lines identify the type of geometry and dimension. The fourth line `bbox` stands for bounding box and display the min and max coordinates containing the Liverpool area in the data frame. The fifth line `projected CRS` indicates the coordinate reference system projection. If you would like to learn more about the various components of spatial data frames, please see the *R* `sf` package vignette on Simple Features.

2.3 Key Challenges

Major challenges exist when working with spatial data. Below we explore some of the key longstanding problems data scientists often face when working with geographical data.

2.3.1 Modifible Area Unit Problem (MAUP)

The Modifible Area Unit Problem (MAUP) represents a challenge that has troubled geographers for decades (Openshaw 1981). Two aspects of the MAUP are normally recognised in empirical analysis relating to *scale* and *zonation*. Fig. 2 illustrates these issues

- *Scale* refers to the idea that a geographical area can be divided into geographies with differing numbers of spatial units.
- *Zonation* refers to the idea that a geographical area can be divided into the same number of units in a variety of ways.

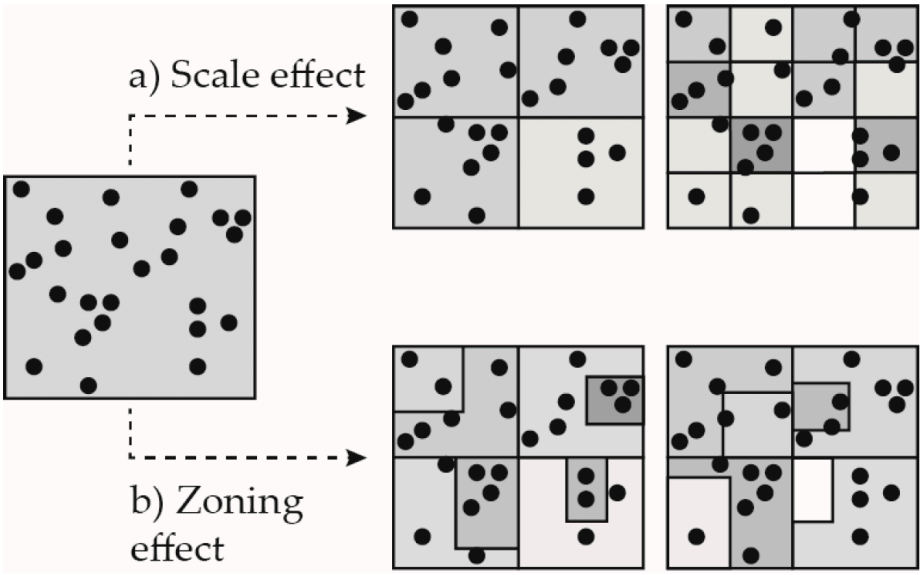

Figure 2.2: Fig. 2. MAUP effect. (a) scale effect; and, (b) zonation effect. Source: loidl2016mapping.

The MAUP is a critical issue as it can impact our analysis and thus any conclusions we can infer from our results (e.g. Fotheringham and Wong 1991). There is no agreed systematic approach on how to handle the effects of the MAUP. Some have suggested to perform analyses based on different existing geographical scales, and assess the consistency of the results and identify potential sources of change. The issue with such approach is that results from analysis

at different scales are likely to differ because distinct dimensions of a geographic process may be captured at different scales. For example, in migration studies, smaller geographies may be more suitable to capture residential mobility over short distances, while large geographies may be more suitable to capture long-distance migration. And it is well documented that these types of moves are driven by different factors. While residential mobility tends to be driven by housing related reasons, long-distance migration is more closely related to employment-related motives (Niedomysl 2011).

An alternative approach is to use the smallest geographical system available and create random aggregations at various geographical scales, to directly quantify the extent of scale and zonation. This approach has shown promising results in applications to study internal migration flows (Stillwell, Daras, and Bell 2018). Another approach involves the production of "meaningful" or functional geographies that can more appropriately capture the process of interest. There is an active area of work defining functional labour markets (Casado-Díaz, Martínez-Bernabéu, and Rowe 2017), urban areas (Arribas-Bel, Garcia-López, and Viladecans-Marsal 2019) and various forms of geodemographic classifications (Singleton and Spielman 2014; Patias, Rowe, and Cavazzi 2019). However there is the recognition that none of the existing approaches resolve the effects of the MAUP and recently it has been suggested that the most plausible 'solution' would be to ignore the MAUP (Wolf et al. 2020).

2.3.2 Ecological Fallacy

Ecological fallacy is an error in the interpretation of statistical data based on aggregate information. Specifically it refers to inferences made about the nature of specific individuals based solely on statistics aggregated for a given group. It is about thinking that relationships observed for groups necessarily hold for individuals. A key example is Robinson (1950) who illustrates this problem exploring the difference between ecological correlations and individual correlations. He looked at the relationship between country of birth and literacy. Robinson (1950) used the percent of foreign-born population and percent of literate population for the 48 states in the United States in 1930. The ecological correlation based on these data was `0.53`. This suggests a positive association between foreign birth and literacy, and could be interpreted as foreign born individuals being more likely to be literate than native-born individuals. Yet, the correlation based on individual data was negative `-0.11` which indicates the opposite. The main point emerging from this example is to carefully interpret analysis based on spatial data and avoid making inferences about individuals from these data.

2.3.3 Spatial Dependence

Spatial dependence refers to the spatial relationship of a variable's values for a pair of locations at a certain distance apart, so that these values are more similar (or less similar) than expected for randomly associated pairs of observations (Anselin 1988). For example, we could think of observed patterns of ethnic segregation in an area are a result of spillover effects of pre-existing

patterns of ethnic segregation in neighbouring areas. Chapter @ref(spatialecon) will illustrate approach to explicitly incorporate spatial dependence in regression analysis.

2.3.4 Spatial Heterogeneity

Spatial heterogeneity refers to the uneven distribution of a variable's values across space. Concentration of deprivation or unemployment across an area are good examples of spatial heterogeneity. We illustrate various ways to visualise, explore and measure the spatial distribution of data in multiple chapters. We also discuss on potential modelling approaches to capture spatial heterogeneity in Chapters @ref(spatialecon), @ref(mlm1) and @ref(sta).

2.3.5 Spatial nonstationarity

Spatial nonstationarity refers to variations in the relationship between an outcome variable and a set of predictor variables across space. In a modelling context, it relates to a situation in which a simple "global" model is inappropriate to explain the relationships between a set of variables. The geographical nature of the model must be modified to reflect local structural relationships within the data. For example, ethnic segregation has been positively associated with employment outcomes in some countries pointing to networks in pre-existing communities facilitating access to the local labour market. Inversely ethnic segregation has been negatively associated with employment outcomes pointing to lack of integration into the broader local community. We illustrate various modelling approaches to capture spatial nonstationarity in Chapters @ref(mlm2) and @ref(gwr).

3 Data Wrangling

References

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Vol. 4. Springer Science & Business Media.

Arribas-Bel, Daniel, M-À Garcia-López, and Elisabet Viladecans-Marsal. 2019. "Building (s and) Cities: Delineating Urban Areas with a Machine Learning Algorithm." *Journal of Urban Economics*, 103217.

Casado-Díaz, José Manuel, Lucas Martínez-Bernabéu, and Francisco Rowe. 2017. "An Evolutionary Approach to the Delimitation of Labour Market Areas: An Empirical Application for Chile." *Spatial Economic Analysis* 12 (4): 379–403.

Fotheringham, A Stewart, and David WS Wong. 1991. "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis." *Environment and Planning A* 23 (7): 1025–44.

Kong, Xiangjie, Menglin Li, Kai Ma, Kaiqi Tian, Mengyuan Wang, Zhaolong Ning, and Feng Xia. 2018. "Big Trajectory Data: A Survey of Applications and Services." *IEEE Access* 6: 58295–306.

Kwan, Mei-Po, and Jiyeong Lee. 2004. "Geovisualization of Human Activity Patterns Using 3d GIS: A Time-Geographic Approach." *Spatially Integrated Social Science* 27: 721–44.

Niedomysl, Thomas. 2011. "How Migration Motives Change over Migration Distance: Evidence on Variation Across Socio-Economic and Demographic Groups." *Regional Studies* 45 (6): 843–55.

Önnerfors, Åsa, Mariana Kotzeva, Teodóra Brandmüller, et al. 2019. "Eurostat Regional Yearbook 2019 Edition."

Openshaw, Stan. 1981. "The Modifiable Areal Unit Problem." *Quantitative Geography: A British View*, 60–69.

Patias, Nikos, Francisco Rowe, and Stefano Cavazzi. 2019. "A Scalable Analytical Framework for Spatio-Temporal Analysis of Neighborhood Change: A Sequence Analysis Approach." In *International Conference on Geographic Information Science*, 223–41. Springer.

Robinson, WS. 1950. "Ecological Correlations and Individual Behavior." *American Sociological Review* 15 (195): 351–57.

Rowe, Francisco, and Nikos Patias. 2020. "Mapping the Spatial Patterns of Internal Migration in Europe." *Regional Studies, Regional Science* 7 (1): 390–93.

Singleton, Alexander D, and Seth E Spielman. 2014. "The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom." *The Professional Geographer* 66 (4): 558–67.

Stillwell, John, Konstantinos Daras, and Martin Bell. 2018. "Spatial Aggregation Methods for Investigating the MAUP Effects in Migration Analysis." *Applied Spatial Analysis and Policy* 11 (4): 693–711.

Tao, Sui, Jonathan Corcoran, Francisco Rowe, and Mark Hickman. 2018. "To Travel or Not to Travel:'weather'is the Question. Modelling the Effect of Local Weather Conditions on Bus Ridership." *Transportation Research Part C: Emerging Technologies* 86: 147–67.

Wolf, Levi John, Sean Fox, Rich Harris, Ron Johnston, Kelvyn Jones, David Manley, Emmanouil Tranos, and Wenfei Winnie Wang. 2020. "Quantitative Geography III: Future Challenges and Challenging Futures." *Progress in Human Geography*, 0309132520924722.

Xie, Yihui, Joseph J Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. CRC Press.