GEOG0114: Principles of Spatial Analysis

Justin van Dijk[1] and Joanna Wilkin[2]

2020-10-23

[1]Department of Geography, https://www.mappingdutchman.com/
[2]Department of Geography, https://www.geog.ucl.ac.uk/people/academic-staff/joanna-wilkin

Contents

Principles of Spatial Analysis

Welcome

Welcome to **Principles of Spatial Analysis**, one of the four core modules for the MSc in Geographic and Social Data Science here at UCL Geography.

This module provides an introduction into the theory, methods and tools of spatial analysis essential for your career as a Data Scientist. This module has been designed in conjunction with your module on Geographic Information Systems (GIS) and Science (GIScience) from CASA to provide you with an extensive introduction into GIScience, spatial analysis and their associated tools. It has a specific focus on the principles, properties and parameters that are part of spatial analysis and how to understand and apply these effectively within geographic and data science-oriented research.

The first half of the module provides detailed introductions into spatial concepts such as scale and geography, spatial dependency and autocorrelation, and spatial heterogeneity and spatial regression models. The second half then focuses on the applications of spatial analysis within current data science research, including cluster analysis and Bayesian modelling.

Get in touch

Dr Jo Wilkin will be taking **Week 1-4**, **9**, and **10** of the module. You can contact her at **j.wilkin [at] ucl.ac.uk** or, for online office hours, you can book a half hour slot with Jo using MS Bookings.

Dr Justin van Dijk will be taking **Week 5-8** of the module. You can contact him at **j.t.vandijk [at] ucl.ac.uk** or, for online office hours, you can book a half hour slot with Justin using MS Bookings.

The module is further supported by two Postgraduate Teaching Assistants: Alfie Long and Jakub Wyszomierski.

Noticed a mistake in this resource?

Please let us know through the GitHub issues tab, send us a message over MS Teams, or contact us by e-mail.

Course information

Module structure

This module consists of ten self-led workshops, ten interactive seminar discussions and ten help sessions. Each week, we'll provide an online workshop, provided as a worksheet with videos and instructions to complete the practical component of the workshop. All online classes will be held on the Principles of Spatial Analysis 'team'.

In addition, each week will have its own reading list or additional 'recommended' (optional, not required!) online tutorials we know of that you might want to also complete.

Weekly topics

Week	Date	Topic
1	05/10/2020	Spatial analysis for data science
2	12/10/2020	Representation, scale and geography in spatial analysis
3	19/10/2020	Spatial properties, relationships and operations
4	26/10/2020	Spatial dependence, spatial autocorrelation and defining neighbours
5	02/11/2020	Exploratory Spatial Data Analysis and regression
reading week	*reading week*	*reading week*
6	16/11/2020	Mapping clusters with point pattern analysis
7	23/11/2020	Geostatistics, interpolation and raster analysis
8	30/11/2020	DBScan and cluster analysis for urban applications

Week	Date	Topic
9	09/12/2020	*To be confirmed*
10	04/12/2020	Creating reproducible research and module recap

Learning objectives

By the end of the module, you should:

- have a good understanding of the principles underlying the analysis of spatial data in general and spatial statistics in particular;
- be able to use GIS software and tools for generating and visualising summary statistics;
- be able to examine, analyse and simulate a range of spatial patterns and processes;
- be able to use geostatistical tools to analyze and interpolate spatial patterns;
- appreciate the many different sources of uncertainty in spatial data and spatial processing and how to address such issues in analysis and research;
- be able to master the key concepts in network analysis with a focus on social and spatial networks (now in Intro to Data Science and Advanced Data Science modules);
- be able to explain several novel applications of spatial analysis techniques within geographic and social data science applications.

Reading list

We link to books and resources throughout each practical. The full reading list for the course is provided on the UCL library reading list page for the course. Alternatively, you can always easily find the link to the Reading List in the top right of any Moodle page for our module, under "Library Resources".

This Reading List will be updated on a weekly basis, in preparation for the week to come, so you may see some weeks without reading for now. But please check back at the **start of each week** as the lecture, seminar and/or workshop material is released for that week to check for new readings. All reading for that week will be provided by the time your learning materials are released - so you will not need to check the reading list for updates as the week progresses.

Module assessment details

The assessment for Principles of Spatial Analysis is set across two pieces of separate coursework, weighted at 50% each.

1) The first piece of coursework will involve the completion of a spatial analysis project, based on the theory, concepts and application learnt during the module. A 1,500 word report will be submitted, alongside the code used within the project, which describes the analysis undertaken and the results of the analysis. Further guidance will be provided in Week 5 of the module.

2) The second piece of coursework will be a written (1,500 word) review on a current data science application that uses spatial analysis as its core methodology. The application can be drawn from the lecture material, particularly during Weeks 8 and 9, or one of your own choice. Further guidance will be provided in Week 9 of the module.

Guidance for both pieces will be uploaded to this section of the Moodle, once provided.

Useful additional resources

Besides the mandatory and recommended reading for this course, there are some additional resources that are worth checking out:

1. MIT's introduction course on mastering the command line: The Missing Semester of Your CS Education
2. A useful tool to unpack command line instructions: explainshell.com
3. Online resource to develop and check your regular expressions: regexr.com
4. Selecting colour palettes for your map making and data visualisation: colorbrewer 2.0

Chapter 1

Spatial analysis for data science

1.1 Session outline

The first week of PSA will introduce how geography and Geographical Information Science (GIScience) fits within the wider data science discipline and why there is a need to specialise in spatial and social analysis for data science.

To provide this understanding, you will first work your way through a short document written specifically for this week to provide an extensive overview to why a geographical understanding to data science is critical to accurate and valid data analysis. Next, through a recorded lecture, you'll be given a short introduction to Geographical Information Systems (GIS) tools for spatial data science and an explanation to how these tools have changed over the last decade, including a shift from traditional GIS software towards programming-based analysis in research applications.

We'll then show you examples of different types of GIS software through a recorded tutorial. To gain a practical basic understanding of the differences across these software, including their ease of use, the recorded tutorial will show you the steps and processing to create a simple choropleth map of population and population density in London. This week plays a formative role in providing everyone with baseline from which to not only pursue this module, but the other technical modules on the MSc.

This week's content is available on Moodle.

Chapter 2

Geographical representation

2.1 Session outline

The first part of this week will look at spatial representation data models, i.e. how we transform geographic features and phenomena into geographic data for use within GIS. We will then explore the role of scale and geography within spatial analysis and provide you with a critical understanding of how both can impact and effect the analysis of data, particularly when looking at 'event' type data, i.e. the occurrence of a specific phenomenon over space.

We will then introduce you to the role and usage of administrative geographies and discuss how they are subject to the Modifiable Area Unit Problem as well as its the consequences, including ecological fallacies. We will then discuss methods to account for these issues, including population standardization, as well as highlight alternative methods for representing data beyond traditional choropleth maps. The interactive lecture will also introduce the role of projections and what considerations you should make when choosing the projection for your analysis; projections are further discussed in Week 3 of CASA0005.

The practical component of the week puts these issues into practice with an analysis of crime data from the UK and its various administrative geographies, as well as voting patterns in the USA. The practical component also introduces the two types of data joining primarily used in spatial analysis: attribute and spatial joins.

This week's content is available on Moodle.

Chapter 3

Spatial properties and relationships

3.1 Session outline

Understanding spatial properties, relationships and how they are used within spatial operations are the building blocks to spatial data processing and analysis. This week, we look to provide you with a thorough introduction into using spatial operations (and the properties and relationships associated with them) through an introductory lecture, a research-based analysis (with demo and practical) and then a research task which we will look at during this week's seminar.

Within the lecture, we will highlight the different ways of conceptualizing key spatial properties, such as distance, and the impact this may have on measurement. We then focus on their application within spatial operations, and how they can be used for the selection, subset and validation of data. We then look at the core terminology used to define spatial relationships and how they can be used to process datasets, using the operations previously mentioned.

The practical utilises these concepts to investigate the accessibility of greenspace for schools across London. Recent research (Bijnens et al, 2020) has shown that children brought up in proximity to greenspace have a higher IQ and fewer behavioral problems, irrespective of socio-economic background. Here we will look to understand whether there are geographical patterns to schools that have high versus low access of greenspace and where a lack of greenspace needs to be addressed.

For the practical, we provide an introduction to the research problem and outline how we devise a research methodology to be able to investigate our research questions. We then look at the required processing steps to create the final

dataset that can be used in our analysis. This is followed by a short demo in which Jo will demonstrate the analysis visually in QGIS. We then ask you to recreate the analysis by creating a script in R-Studio (code provided) - this will allow you to replicate the analysis for other cities within the U.K, or even further afield if you can extract the same data. Finally, in preparation for this week's seminar, we ask you to watch a five minute video from a local news channel in Jo's hometown - ready to discuss as a possible research task in Friday's seminar.

This week's content is available on Moodle.

Chapter 4

Spatial autocorrelation

4.1 Session outline

This week, we focus on the first of two key properties of spatial data: spatial dependence. Spatial dependence is the idea, as introduced in the first week via Tobler's Law (1970), that the observed value of a variable in one location is often dependent (to some degree) on the observed value of the same value in a nearby location. For spatial analysis, this dependence can be assessed and measured statistically by considering the level of spatial autocorrelation between values of a specific variable, observed in either different locations or between pairs of variables observed at the same location. Spatial autocorrelation occurs when these values are not independent of one another and instead cluster together across geographic space. A critical first step of spatial autocorrelation is to define the criteria under which a spatial unit (e.g. an areal or point unit) can be understood as a "neighbor" to another unit. As highlighted in the previous week, spatial properties can often take on several meanings, and as a result, have an impact on the validity and accuracy of spatial analysis. This multiplicity also can be applied to the concept of spatial neighbours which can be defined through adjacency, contiguity or distance-based measures. As the specification of these criteria can impact the results, the definition followed therefore need to be grounded in particular theory that aims to represent the process and variable investigated.

This week looks at spatial dependence and autocorrelation in detail, focusing on the different methods of assessment. As part of this, we look at the multiple methods to defining spatial neighbours and their suitability of use across different spatial phenomena – and how this approach is used to generate spatial weights for use within these spatial autocorrelation methods as well as their potential to generate spatially-explicit variables.

We put these learnings into practice through an analysis of spatial dependence

of our areal crime data from Week 2, experimenting with the deployment of different neighbours and the impact of their analyses.

This week's content will be made available on 26/10.

Chapter 5

Exploratory data analysis

5.1 Session outline

This week, we focus on the second of the two key properties of spatial data: spatial heterogeneity. With the underlying process (or processes) that govern a spatial variable likely to vary across space, a single global relationship for an entire region of study may not adequately model the process that governs outcomes in any given location of the study region. As a result, multiple methods have been developed to incorporate 'space' into traditional regression models, including spatial lag models, spatial error models, and Geographical Weighted Regression.

This week provides the building blocks to conducting a statistical and spatial investigation into the relationships between spatial variables, first looking at the concept of Exploratory Spatial Data Analysis (ESDA) and how to understand data distributions, descriptive statistics, transformations as well as check for outliers and trends. We then look at the three types of spatial regression models in turn to understand their methodology and potential advantages and limitations in accounting for space when modelling relationships.

In this week's session we will explore potential factors that may contribute to our crime rate in London, using previous literature to identify variables for inclusion in our analysis. We conduct an ESDA of these variables, followed by statistical and spatial regression analysis to determine whether our variables adhere to previous literature and do contribute to crime in our area of study.

This week's content will be made available on 02/11.

5.2 Feedback

Please take a moment to give us some feedback on this week's content.

Chapter 6

Point pattern analysis

6.1 Session outline

In our previous practicals, we have aggregated our event data into areal units, primarily using administrative geographies, to enable its easy comparison with other datasets provided at the same spatial scale, such as the census data used in the previous week, as well as to conduct spatial autocorrelation tests. However, when locations are precisely known, spatial point data can be used with a variety of spatial analytic techniques that go beyond the methods typically applied to areal data. The set of methods unique to point data are often referred to as point pattern analysis and geostatistics.

This week, we focus on point pattern analysis, whilst next week we will look at geostatistics. Within point pattern analysis, we look to detect patterns across a set of locations, including measuring density, dispersion and homogeneity in our point structures. We will look first at basic forms of point pattern analysis, including mean centers and the standard deviational ellipse, and then more powerful analysis methods, including kernel density estimation and Ripley's K function. These latter functions help determine and/or show whether points have a random, dispersed or clustered distribution pattern at a certain scale.

To put these measures and methods into action, our practical component looks to test the hypothesis that "the majority of bicycle theft in London occurs with walking distance of a train or tube station", by assessing clusters and hotspots of bike theft in relation to a transportation locations.

This week's content will be made available on 16/10.

6.2 Feedback

Please take a moment to give us some feedback on this week's content.

Chapter 7

Geostatistics

7.1 Session outline

Following on from Week 3 and in conjunction with this week in your CASA005 module, we will also be focusing on rasters, exploring further the creation and application of rasters from vector datasets using geostatistics. During the lecture, we will provide a more in-depth and detailed explanation behind the geostatistics methods of interpolation, looking at various deterministic and geostatistical techniques. We then introduce methods to using vector datasets with rasters, utilizing a GIS method known as zonal statistics.

Following on from Week 3's practical, we look to study further the role of greenspace quality in relation to schools and how this varies across different geographies. This week, we look at greenspace quality from a personal health perspective, analyzing air quality across our greenspaces. By analyzing the air quality of greenspace, we can provide a more detailed understanding to how "accessible" the greenspaces are to the surrounding schools.

This week's content will be made available on 23/11.

7.2 Feedback

Please take a moment to give us some feedback on this week's content.

Chapter 8

Cluster analysis

8.1 Session outline

DBScan is a density-based clustering algorithm that is commonly used in data mining and machine learning. Across a set of points, DBSCAN will group together points that are close to each other based on a distance measurement and a minimum number of points. It also marks as outliers the points that are in low-density regions. The algorithm can be used to find associations and structures in data that are hard to find through visual observation alone, but that can be relevant and useful to find patterns and predict trends. Whilst the algorithm is nearly 25 years old, it still is incredibly relevant to many applications within data science today, including within spatial analysis. Novel use of the algorithm has seen researchers detect and delineate urban areas from building footprint data. This use presents significant opportunities for those working to provide spatial data within countries where until recently, significant reference data did not exist.

To understand this potential application, we look at the issues of data scarcity and sparsity within LICs and MICs, which do not currently have extensive mapping data or national mapping agencies to provide this data. We look at current initiatives to address this data dearth, with a specific focus on those working on the automated extraction of building footprints from satellite imagery using machine learning.

Within the practical component, we look to deploy a simple version of DBScan on building data subset from the Tanzanian building dataset created by Microsoft. We attempt to detect and delineate our own urban areas within the dataset and discuss the advantages of novel changes made by those working with the algorithm for this specific purpose.

This week's content will be made available on 30/10.

8.2 Feedback

Please take a moment to give us some feedback on this week's content.

Chapter 9

Bayesian modelling

9.1 Session outline

Bayesian methodology is an approach to statistical inference that has existed for a long time. However, its applications have been limited until recent advancements in computation and simulation methods has made its deployment and use within big datasets viable. The concept underlying Bayesian spatial modeling is Bayes' theorem, which considers both the distributions of the data and the unknown coefficient estimates (LeSage and Pace, 2009). Bayesian spatial modeling embraces most, if not all, spatial models in the literature, such as the spatial lag model, the spatial error model, and geographically weighted regression discussed in Week 5. As a result, recent decades have experienced a rapid growth in the application of Bayesian spatial modeling to epidemiology, demography, and environmental health research, in addition to other geography-related disciplines. Bayesian modelling has been used to estimate population distributions, disease spread and air pollution.

This week provides you with a general introduction to Bayesian modelling for geographical applications, with a focus on understanding the overall methodology as well as selection of variables for use within the model.

The practical component looks at R-INLA package and how it can be used within health modelling.

This week's content will be made available on Moodle on 09/12.

Chapter 10

Reproducible research

10.1 Session outline

In our final week of PSA, we will recap the main principles of spatial analysis that you have learnt over the last nine weeks. We will then look to tidy one of our projects from the previous week to use it within CodeOcean, an online platform that hosts code and data to create "reproducible runs" and "capsules" of research projects. The platform (alongside others!) enables you to create an anonymized version of your project, ready for its submission to journals as part of their increasing requirement to pass reproducibility and openness tests, e.g. the International Journal of Geographic Information Science. The live session of this week will be extended to present the two pieces of coursework you will assessed on as part of this module, as well as to answer any questions you may have on these assessments.

This week's content will be made available on Moodle on 09/12.