1 Think Spatially

Basic Concepts of Spatial Analysis and Space Conceptualization

THEORY

Learning Objectives

This chapter

- Presents the basic concepts, terms and definitions pertaining to spatial analysis
- Introduces a spatial analysis workflow that follows a describe–explore–explain structure
- Presents in detail the reasons that spatial data are special – namely spatial autocorrelation, scale, the modifiable area unit problem, spatial heterogeneity, the edge effects and the ecological fallacy
- Explains why conceptualization of spatial relationships is extremely important in spatial analysis
- Presents the approaches used to conceptualize spatial relationships
- Explains how distance, contiguity/adjacency, neighborhood, proximity polygons and space–time window are used in space conceptualization
- Defines the spatial weights matrix, which is essential to almost every spatial statistic/technique
- Introduces the real-world project along with the related dataset to be worked throughout the book

After a thorough study of the theory and lab sections, you will be able to

- Implement a comprehensive workflow when you conduct spatial analysis
- Distinguish spatial from nonspatial data
- Understand why spatial data should be treated with new methods (e.g., spatial statistics)
- Understand the importance of applying conceptualization methods according to the problem at hand
- Understand essential concepts for conducting spatial analysis such as distance, contiguity/adjacency, neighborhood, proximity polygons and space–time
- Describe the spatial analysis process to be adopted for solving the real-world project of this book
- Presents the project's data with ArcGIS and GeoDa

1.1 Introduction: Spatial Analysis

"In God we trust. All others must bring data," said W. Edwards Deming (American statistician and professor, 1900–1993), as without *data*, there is little to be done. Counting objects or individuals and measuring their characteristics is the basis for almost every study. With the advent of *geographic information systems* (GIS), it is simple to link nonspatial data (e.g., income, unemployment, grades, sex) to spatial data (e.g., countries, cities, neighborhoods, houses) and create large *geodatabases*. In fact, when data are linked to *location*, then analysis becomes more intriguing, and *spatial analysis* and the science of geography take over, as raw data are of a little value. Analyzing data through spatial analysis methods and techniques allows us to add value by creating information and then knowledge. Within this context, spatial analysis can be defined in various ways:

- **Spatial analysis** is a collection of methods, statistics and techniques that integrates concepts such as location, area, distance and interaction to analyze, investigate and explain in a geographic context patterns, actions, or behaviors among spatially referenced observations that arise as a result of a process operating in space.
- **Spatial analysis** is the quantitative study of phenomena that manifest themselves in space (Anselin 1989 p. 2).
- **Spatial analysis** studies "how the physical environment and human activities vary across space – in other words, how these activities change with distance from reference locations or objects of interest" (Wang 2014 p. 27).
- **Spatial analysis** is "the process by which we turn raw data into useful information, in pursuit of scientific discovery, or more effective decision making" (Longley et al. 2011).
- **Spatial (data) analysis** is "a set of techniques designed to find pattern, detect anomalies, or test hypotheses and theories based on spatial data" (Goodchild 2008 p. 200).
- **Spatial analysis** is a broad term that includes (a) spatial data manipulation through geographical information systems (GIS), (b) spatial data analysis in a descriptive and exploratory way, (c) spatial statistics that employ statistical procedures to investigate if inferences can be made and (d) spatial modeling which involves the construction of models to identify relationships and predict outcomes in a spatial context (O' Sullivan & Unwin 2010 p. 2).

Why Conduct Spatial Analysis?

Spatial analysis concepts, methods, and theories make a valuable contribution to analysis and understanding of

- **Social Systems:** Spatial analysis methods can be used to study how people interact in social, economic and political contexts, as space is the underlying layer of all actions and *interconnections* among people.

- **Environment:** Spatial analysis methods can be applied in studies related to natural phenomena and climate change hazards, natural resources management, environmental protection and *sustainable* development.
- **Economy:** Spatial analysis methods can be used to analyze, map and model interrelations among humans and various economic dimensions of economic life.

The main advantage of spatial analysis is the ability to reveal patterns in data that had not previously been defined or even observed. For example, using spatial analysis techniques, one might identify the *clustering* of a disease occurrence and then develop mechanisms for preventing expansion or even eliminating it (Bivand et al. 2008). In this respect, spatial analysis leads to better *decision making* and *spatial planning* (Grekousis 2019).

In a broad sense, there are four types of spatial analysis:

- **Spatial point pattern analysis:** A set of data points is analyzed to trace if it exhibits one of three states: *clustered, dispersed*, *random*. Consider, for example, a spatial arrangement of stroke events in a study area. Are they clustered to a specific region or are strokes randomly distributed across space? Spatial analysis proceeds with a further investigation, such as to determine the driving factors that lead to this clustering (potentially the existence of nearby industrial zones and related pollution). Point patter analysis also includes centrographics, a set of spatial statistics utilized to measure the center, the spread and the directional trend of point patterns. In this type of analysis, data typically refer to the entire population and not to a sample.
- **Spatial analysis for areal data:** Data are aggregated into predefined zones (e.g., census tracts, postcodes, etc.), and analysis is based on how neighboring zones behave and whether relations and interactions exist among them (i.e., clusters of nonspatial data also form clusters in space). For example, do people with high or low income cluster around specific regions, or are they randomly allocated? *Spatial dependence*, *spatial heterogeneity*, *spatial autocorrelation*, *space conceptualization* (through spatial weights matrix) and *regionalization* (*spatial clustering*) are central notions in this type of analysis.
- **Geostatistical data analysis (continuous data):** Geostatistical analysis is the branch of statistics analyzing and modeling *continuous field* variables (O'Sullivan & Unwin 2010 p. 115). In this respect, geostatistical data comprise a collection of sample observations of a continuous phenomenon. Using various geostatistical approaches (e.g., interpolation), values can be calculated for the entire surface. Pollution, for instance, is monitored by a limited network of observation locations. To estimate pollution for every single point, we may apply interpolation techniques through geostatistical analysis. Geostatistical analysis is not covered in this book.

- **Spatial modeling:** Spatial modeling deals mainly with how spatial dependence, spatial autocorrelation and spatial heterogeneity can be modeled in order to produce reliable, predicted spatial outcomes. Spatial modeling can be used, for example, to model how the value of a house is related to its location. Spatial regression and spatial econometrics are key methods in spatial modeling.

Spatial Analysis Workflow

As spatial analysis is a wide discipline with a large variety of methods, approaches and techniques, guidance on how to conduct such analysis is necessary. This book introduces a new spatial analysis workflow that follows a describe–explore–explain structure in order to address what, where and why questions respectively (see Figure 1.1).

Step A: Describe (What). This is the first step in the spatial analysis process. It describes the dataset through *descriptive statistics*. Descriptive statistics are used to summarize data characteristics and provide a useful understanding about the distribution of the values, their range and the presence of outliers.

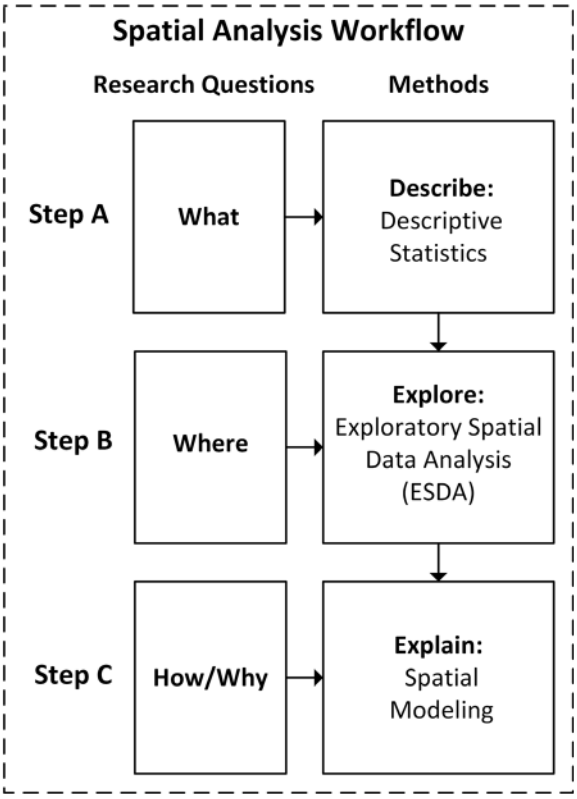
Spatial Analysis Workflow
Research Questions
Methods
Step A
What
Describe: Descriptive Statistics
Step B
Where
Explore: Exploratory Spatial Data Analysis (ESDA)
Step C
How/Why
Explain: Spatial Modeling

Figure 1.1 Spatial analysis workflow.

Cambridge University Press
978-1-108-49898-2 — Spatial Analysis Methods and Practice
George Grekousis
Excerpt
More Information

This step typically answers "what?" questions, such as what is the mean income of a neighborhood, or what is the population proportion living under the poverty level? This step offers an initial understanding of the dataset and its specific characteristics. Still, if the data have not been collected appropriately, then no analysis can lead to accurate and useful results. For this reason, any dataset should be checked for consistency and accuracy before any deeper analysis takes place. Datasets without detailed reports explaining the methods used and accuracies achieved should be avoided (always cite in your studies the link to the database used and the report that describes the methods used to collect the data along with the associated quality controls).

Step B: Explore (Where). In the second step, *exploratory spatial data analysis* (ESDA) is applied to explore and map data, locate outliers, test underlying assumptions or identify trends and associations among them, such as spatial autocorrelation presence or spatial clustering. In this step, we mostly answer "where?" questions, such as where are the areas with low/high values in income, is there any spatial clustering in the distribution of income per capita, where is it located, and where are the crime hot spots in a city?

Step C: Explain (Why/How). In the last step, explanatory statistical analysis through a spatial lens is applied to explain and understand causes and effects through models. In this step, we attempt to answer "why?/how?" questions. These methods do not just identify associations but also attempt to (a) unveil relations that explain why something happens and (b) trace the drivers behind a change. Typical questions in a geographic context include why do crime events cluster in a specific area? Is there any link to the specific socioeconomic characteristics of this area? Why is income per capita linked to location, how is income related to the size of a house? Which are the driving forces behind sea-level rise, and how does population increase drive urban land cover changes? In this type of analysis and in the context of this book, we treat variables as either independent or dependent. The dependent variable (effect) is the phenomenon/state/variable we attempt to explain. For example, if an analysis concludes that population increase (driver-independent variable) accounts for x% of urban land cover change (effect-dependent variable), then there is a linkage (a relation) established that explains the degree that the driver influences the effect. We have now built a model that explains why something happens which additionally can be used for predictions. This is a step beyond steps A and B, which mostly address "what happens" or "where something happens." Spatial regression and spatial econometrics will be described in this book concerning this stage of analysis. From the spatial analysis perspective, several additional questions could be also addressed at this stage: Can we learn something from this dataset and the applied methodology? Has new knowledge been created? What is the next step? How should future research proceed? When spatial analysis is completed, the knowledge created enhances decision making and spatial planning.

1.2 Basic Definitions

Box 1.1 More than 20 terms (in italics) related to spatial analysis, spatial statistics and *spatial thinking* (one more) have been mentioned in the preceding section. Some terms might be comprehensive, others entirely new and others quite vague. Let us start by building a common vocabulary and presenting some key definitions in this section. Definitions, terms and formulas typically vary among books, which confuses not only nonspecialists but scientists as well, causing much misunderstanding. This confusion has also hampered statistical, GIS and spatial analysis software, especially when referred to equations and formulas. This book presents the most commonly used names and symbols for terms and statistics.

Definitions

Spatial statistics employ statistical methods to analyze spatial data, quantify a spatial process, discover hidden patterns or unexpected trends and model these data in a geographic context. Spatial statistics can be considered part of ESDA, spatial econometrics and remote sensing analysis (Fischer & Getis 2010 p. 4). They are largely based on inferential statistics and hypothesis testing to analyze geographical patterns so that spatially varying phenomena can be better modeled (Fischer & Getis 2010 p. 4). Spatial statistics quantify and map what the human eye and mind intuitively see when reading a map depicting spatial arrangements, distributions or trends (Scott & Janikas 2010 p. 27; see also Chapter 2).

Spatial modeling deals with the creation of models that explain or predict spatial outcomes (O'Sullivan & Unwin 2010 p. 3).

Geospatial analysis is the collection of spatial analysis methods, techniques, and models that are integrated in geographic information systems (GIS; de Smith et al. 2018). Geospatial analysis is enriched with GIS capabilities and is used to design new models or integrate existing ones in a GIS environment. It is also used as an alternative term for "spatial analysis" (de Smith et al. 2018); strictly speaking, however, spatial analysis is part of geospatial analysis (see Box 1.2).

Box 1.2 It is not always easy to distinguish between the terms "geographic," "spatial" and "geospatial." These terms have been defined by many experts within different scientific contexts. The definitions provided here are not exhaustive; they serve as a basis for a common terminology. Even within the science of geography, terms can overlap, and distinctions can be vague. The term "geographic" refers to a location relative to the earth's surface combined with some type of representation. On the other

Box 1.2 (*cont.*)

hand, the term "spatial" does not refer solely to the earth's surface; its meaning is extended to a location combined with additional attribute data. The term "geospatial" is more computer oriented and refers to information based on both spatial data and models, combining geographical analysis with spatial analysis and modeling.

Spatial data refers to spatial entities with geometric parameters and spatial reference (coordinates and coordinate system) that also have other nonspatial attributes (see Figure 1.2; Bivand et al. 2008 p. 7). For example, we can describe a city by its population, unemployment rate, income per capita or the average monthly temperature. When these data are linked to a location through spatial objects (e.g., city postcodes), then we get spatial data. The range of attributes to be joined to the spatial objects depends on the problem being studied and the availability of datasets (e.g., census). Images that are georeferenced are also considered spatial data.

Conceptually, there are two ways to represent the geographic entities and as such to represent the world digitally: the object view of the world and the field view of the world (Haining 2010 p. 199).

Object view is a representation that describes the world with distinctive spatial objects that are georeferenced to a specific location using coordinates. In the object view, spatial objects are modeled as points, lines or polygons (also called features). This data model is called vector data model (O'Sullivan & Unwin 2010 p. 6). The object view of the world and the vector data model can be used to map, for example, demographic or socioeconomic data. Spatial objects might be represented differently when the scale of analysis changes.

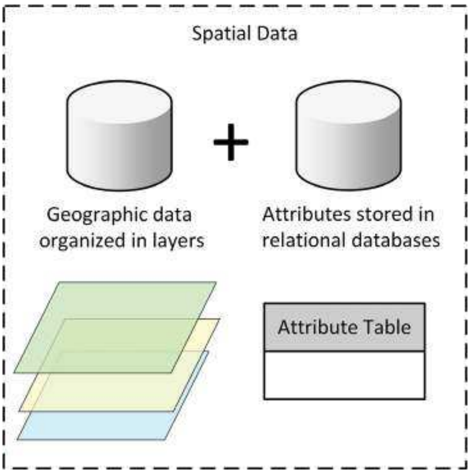
Spatial Data
+
Geographic data organized in layers
Attributes stored in relational databases
Attribute Table

Figure 1.2 Spatial entities linked to attributes create spatial data.

For instance, a city might be represented as a point feature when examined on a national scale and as a polygon feature at a more local level.

Field view is a representation that describes the world as a surface of continuously varying properties (O'Sullivan & Unwin 2010 p. 7). The field view of the world is more appropriate for depicting a continuous phenomenon/ property (e.g., temperature, pollution, land cover type, height). A way to record a field is through the raster data model. In this model, rectangular cells (called pixels) organized in a regular lattice, depict the geographic variation of the property studied. Another way to model fields is by using triangulated irregular networks (TINs). We can convert spatial data from one model to the other (i.e., vector to raster conversion or vice versa) according to the study's needs.

Variable is any characteristic an object or individual might have. For example, age, height and weight are all variables characterizing humans, animals, or objects.

Attributes are information carried by spatial data. They are stored as columns in a GIS table. An attribute field is equivalent to a variable in classic statistics and has become the preferred term in GIS analysis, but these terms can be used interchangeably. An attribute might be the population of a postcode or the per-capita annual income in a census tract.

Data are produced when we measure the characteristics of objects or individuals.

Value is the result of a measurement (or response) of a characteristic. In statistics, the term **score** is also used to describe a variable's value.

Outlier is an unusual, very small or very large value compared to the rest of the values.

Dataset is a collection of variables of any kind of objects. Typically, a spatial dataset has a tabular format, whereby columns contain the attributes and rows contain the spatial entities.

Population is the entire collection of observations/objects /measurements about which information is sought.

Sample is a part of the entire population.

Level of measurement of a variable describes how its values are arranged in relation to each other (de Vaus 2002 p. 40). Variables/attributes are grouped at three levels of measurement: nominal, ordinal and interval or ratio (Haining 2010 p. 201; see Table 1.1).

- **Nominal** variables are variables with values that cannot be ordered. For example, race may be set as White = 1, Asian = 2, Hispanic = 3. This is a nominal variable, as the values "1, 2, 3" do not to reveal rank but are used as labels for the various categories. We cannot add the values of two different objects like, let's say, "1+3 = 4," as "4" does not reflect any meaningful value. Another example of nominal variables is the "Name of city" (e.g., Athens, Beijing, New York) or the "Land Cover Name" (e.g., Forest, Urban, Water). This type of attribute provides descriptive

Table 1.1 Level of measurement per data structure model (vector/raster) and examples per data type. Applicable logical and arithmetic operations are mentioned in parentheses. Many statistical procedures and techniques cannot be used at all levels of measurements, as different logical and arithmetic operations apply to different levels. For example, binary logistic regression is designed for dichotomous dependent variables and cannot be used for ratio variables. The level of measurement defines the pool of the statistical procedures to be used later in the analysis. From the statistical perspective, more techniques can be used to analyze ratio variables than can be used for nominal and ordinal variables; thus, ratio variables are preferred (de Vaus 2002 p. 43).

Level of measurement	Vector data model (object view)			Raster data model (field view)
	Point	Line	Polygon	Pixel
Nominal $(=,\neq)$	City name	Road name	Postcode ID	Land cover type
Ordinal $(=,\neq,>,<)$	City most desirable to live (ranked)	Road classification type: (Avenue, Highway)	Postcode classification according to education attainment	Forest land cover subclasses
Interval $(=,\neq,>,<,+,-)$	Poverty level	Width of road	Poverty level for a postcode	Ground temperature
Ratio $(=,\neq,>,<,+,-,x,/)$	Population	Road freight	Postcode data: population, income per capita	Pollution PM2.5

information and can be used to label polygons on a map. The applicable operators are “equal” or “not equal” $(=, \neq)$.

- **Ordinal** variables are variables whose categories can be ordered but whose numerical differences are not meaningful and cannot be calculated. For example, the variable “Student” might get the following values: “Exceptional” = 1, “Good” = 2, “Need to study harder” = 3. We can order categories from top to bottom (or vice versa), but there is no meaning in subtracting (“Exceptional” – “Good” = -1). We can apply the operators “equal,” “not equal,” “larger than” and “smaller than” $(=, \neq, >, <)$. Spatial entity’s attributes measured at nominal or ordinal levels are also called “categorical.”
- **Interval and ratio variables** (also called “numerical”) are variables for which each observation can be expressed in a numerically meaningful way. Numbers are not used only as labels but may be used to calculate statistics (e.g., the average). If the values of a numerical variable are limited to specific categories, then the variable is a discrete numerical, also called interval. The interval level is a class of ratio level. In interval-level measurement, categories are defined by fixed distances. Interval data allow for the operation of addition and subtraction (Haining 2010 p. 201). Still, interval variables do not preserve ratios (O’Sullivan & Unwin 2003 p. 13). Dichotomous variables (e.g., for the variable “sex,” an individual might be Male = 1, Female = 0, or the inverse) can also be regarded as discrete

interval-level variables. In this case, zero stands for the absence of something. If the set of possible values is not limited to some categories between low and high values, then the variable is a continuous numerical (also called "ratio"). Ratio variables have a meaningful zero. In ratio variables, we can use all operators (=, ≠, >, <, +, −, x, /).

To analyze variables/attributes, statistical methods can be employed. There are three major branches of classic statistics: **descriptive statistics, inferential statistics** and **explanatory statistics** (Linneman 2011 p. 20). We will deal with all of them in this book.

Descriptive statistics is a set of statistical procedures that summarize the basic characteristics of a given distribution. Descriptive statistics usually summarize a specific sample and are not appropriate for making inferences regarding the total population (unless we have the entire population at hand). As a result, they are not developed on the basis of probability theory as inferential statistics are. In this sense, the results of descriptive statistics apply only to the specific dataset they have been calculated for. Descriptive statistics make use of tables, graphs and simple statistical procedures (Linneman 2011 p. 21).

Inferential statistics is the branch of statistics that analyzes samples to draw conclusions for the entire population. Typical approaches for dealing with inferential statistics include tests of significance (hypothesis testing), confidence interval and Bayesian inference.

Explanatory statistics is the branch of statistics that uses methods and techniques to identify relations among variables and potentially "explain" causalities. In this type of statistics, variables are treated as dependent or independent (Linneman 2011 p. 21). The dependent variable is what we attempt to explain through a set of independent variables. Regression analysis is typically used in explanatory statistics.

1.3 Spatial Data: What Makes Them Special?

Consider that a realtor stores the contact numbers of his clientele in his cell phone. These contacts are data stored in his phone's memory in a casual type of database. If these data are linked to location (in terms of coordinates or addresses through geocoding), then they are transformed into spatial data. Each contact is now attached to a single spatial object (e.g., a point denoting the home address of each client that carries additional information – the attributes – such as phone number, name, date of birth or e-mail) that can be mapped. Transforming data to spatial data offers a lot more than a glossy visualization. It allows for in-depth geoprocessing analysis and advanced spatial querying. For example, which is the closest client to a specific point, where do the majority of clients live, where are clients who spend the most located, what is the best route to their homes, what percent of clients live within a zone of