Introduction

0.1 About the book

This book is about methods for analysing quantitative spatial data. ‘Spatial’ means each item of data has a geographical reference so we know where each case occurs on a map. This spatial indexing is important because it carries information that is relevant to the analysis of the data. The book is aimed at those studying or researching in the social, economic and environmental sciences. It details important elements of the methodology of spatial data analysis, emphasizes the ideas underlying this methodology and discusses applications. The purpose is to provide the reader with a coherent overview of the field as well as a critical appreciation of it.

There are many different types of spatial data and different forms of spatial data analysis so it is necessary to identify what is, and what is not, covered here. We do so by example:

1 *Data from a surface*. The data that are recorded have been taken from a set of *fixed* (or given) locations on a continuous surface. The continuous surface might refer to soil characteristics, air pollution, snow depth or precipitation levels. The attribute being measured is typically continuous valued. Note that for some of these variables (e.g. snow depth) a point observation is sufficient whilst for others (e.g. air pollution) an areal support or block is necessary in order to provide a measure for the attribute value.

The attribute need not be continuous valued and could be categorical. Land use constitutes a continuous surface. The surface might be divided into small parcels or blocks and land-use type recorded for each land parcel.

The data may originate from a sample of points (or small blocks) on the surface. The data may originate from exhaustively dividing up the surface into tracts and recording a representative value for the attribute for each tract. Once the data have been collected the location of each observation is treated as fixed.

2 *Data from objects*. In this case, data refer to point or area objects that are located in geographic space. An individual may be given a location according to their place of residence. At some scale of analysis the set of retail outlets in a town can be represented by points; even towns scattered across a region may be thought of as a set of points. Attributes may be continuous valued or discrete valued, quantitative or qualitative. Objects may be aggregated into larger groupings – for example populations aggregated by census tracts. Now attribute values are representative of the aggregated population. Again, once the data have been collected, the locations of the points or areas or aggregate zonings are treated as fixed.

The purpose of analysis may be to describe the spatial variation in attribute values across the study area. The next step might be to explain the spatial pattern of variation in terms of other attributes. Description might involve identifying interesting features in the data, including detecting clusters or concentrations of high (or low) values and the next step might be to try to understand why certain areas of the map have a concentration of high (or low) values. In some areas of spatial analysis such as geostatistics the aim may be to provide estimates or predictions of attribute values at unsampled locations or to make a map of the attribute on the basis of the sample data.

That the locations of attribute values are treated as fixed is in contrast to classes of problems, not treated here, where it is the location of the points or areas that are the outcome of some process and where the analyst is concerned to describe and explain the location patterns. These are referred to as *point pattern* data and *object* data (Cressie, 1991, p. 8). To give an example: suppose data were available on the location of all retail outlets of a particular type across an urban area. In addition, for each site, attribute data have been recorded including the price for a particular commodity. The methods of this book would be appropriate for describing and explaining the variation in prices across the set of retail outlets treating their locations as fixed. If we are interested in describing and explaining the location pattern of the individual retail outlets within the urban area, as the outcome of a point location process, then this falls outside the domain of the methods here. Note however that if we are willing to view the location problem through a partitioning of the urban area into a set of *fixed* areas that have been defined independently of

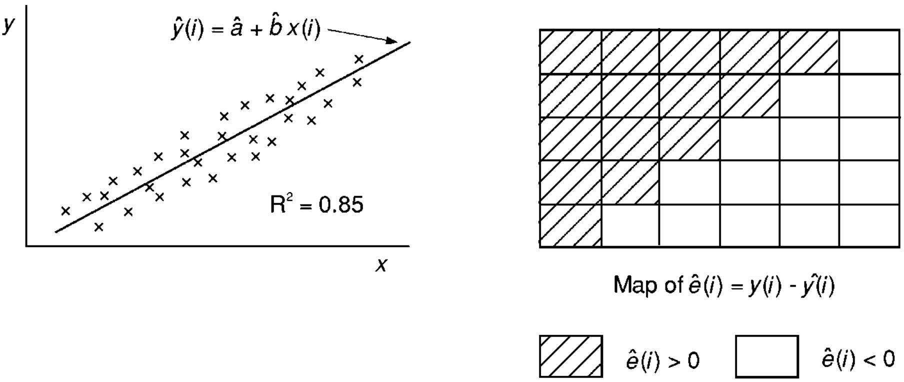

Figure 0.1 Evidence in assessing the adequacy of fit of a regression model

the distribution of retail sites, then the number of retail sites (0, 1, 2, . . .) in each area becomes an attribute and the methods of data analysis in this book are relevant.

The methods of this book are appropriate for analysing the variation in, for example, disease, crime and socio-economic data across a set of areal units such as census tracts or fixed-point sites. They are also appropriate where a sensor has recorded data across a study area in terms of a rectangular grid of small areas (pixels) from which land use or other environmental data have been obtained.

There are two aspects to variation in a spatial data set. The first is *variation* in the data values disregarding the information provided by the locational index. The second is *spatial* variation – the variation in the data values across the map. *Describing* these two aspects of variation calls for two different terminologies and involves different strategies. *Explaining* variation – that is finding a model that will account for the variation in an attribute – could, as an outcome, also provide a good explanation of its *spatial* variation. It is also possible that a model that apparently does well in describing attribute variation leaves important aspects of its spatial variation unexplained. For example all the cases that are very poorly fitted by the model might be in one part of the map. This would arise in regression analysis if all the cases that have the largest positive residuals are in one part of the map and all the cases that have the largest negative residuals are in another. In figure 0.1 the goodness-of-fit statistic ($R^2 \times 100$) equals 85%. X has accounted for 85% of the variation in the response variable (Y). This suggests an adequate model. But there is a strong spatial structure to the pattern of positive and negative residuals ($\hat{e}(i)$). The analyst will conclude that the model is in need of further development if parameters of interest are to be properly estimated or hypotheses properly tested and will need to consider strategies for achieving this.

0.2 What is spatial data analysis?

The term 'spatial analysis' has a pedigree in geography that can be traced back to at least the 1950s and for an overview of historical developments at that time see Berry and Marble (1968, pp. 1–9). Spatial analysis is a term widely used in the Geographical Information *Systems* (GIS) and Geographical Information *Science* (GISc) literatures. A definition of spatial analysis is that it represents a collection of techniques and models that explicitly use the spatial referencing associated with each data value or object that is specified within the system under study. Spatial analysis methods need to make assumptions about or draw on data describing the spatial relationships or spatial interactions between cases. The results of any spatial analysis are not the same under re-arrangements of the spatial distribution of values or reconfiguration of the spatial structure of the system under investigation (Chorley, 1972; Haining, 1994, p. 45).

Spatial analysis has three main elements. First it includes *cartographic modelling*. Each data set is represented as a map and map-based operations (or implementing map algebras) generate new maps. For example buffering is the operation of identifying all areas on a map within a given distance of some spatial object such as a hospital clinic, a well, or a linear feature such as a road. Overlaying includes logical operations (.AND.; .OR.; .XOR.) and arithmetic ($+$; $-$; $\times$; /) operations. The logical overlay denoted by .AND. identifies the areas on a map that simultaneously satisfy a set of conditions on two or more variables (Arbia et al., 1998). The arithmetic overlay operation of addition sums the values of two or more variables area by area (Arbia et al., 1999).

Second, spatial analysis includes forms of *mathematical modelling* where model outcomes are dependent on the form of spatial interaction between objects in the model, or spatial relationships or the geographical positioning of objects within the model. For example, the configuration of streams and the geography of their intersections in a hydrological model will have an effect on the movement of water through different areas of a catchment. The geographical distribution of different population groups and the distribution of their density in a region may have an influence on the spread of an infectious disease whilst the location of topographical barriers may have an influence on the colonization of a region by a new species. Finally, spatial analysis includes the development and application of statistical techniques for the proper analysis of spatial data and which, as a consequence, make use of the spatial referencing in the data. This is the area of spatial analysis that we refer to here as *spatial data analysis*.

There are many features to spatial data that call for careful consideration when undertaking statistical analysis. Although the analysis of spatial dependence is a critical element in spatial data analysis and central for example in specifying sampling designs or undertaking spatial prediction, an excessive attention to just that aspect of spatial data can lead the analyst to ignore other issues. For example: the effect of an areal partition on the precision of an estimator or the wider set of assumptions and data effects that determine whether a model can be considered adequate for the purpose intended. In this sense spatial data analysis is a *subfield* of the more general field of data analysis. In defining the skills and concepts necessary for undertaking a proper analysis of spatial data there is, then, an important role for areas of statistical theory developed to handle other types of, non-spatial, data. In adopting this rather broader definition of spatial data analysis a link is maintained to the wider body of statistical theory and method.

0.3 Motivation for the book

This book is a descendant of *Spatial Data Analysis in the Social and Environmental Sciences* (1990) which dealt with the same types of spatial data. The earlier book reviewed models for describing and explaining spatial variation and discussed the role of robust methods of fitting partly in response to the perceived nature and quality of spatial data. That book described both exploratory spatial data analysis and spatial modelling. Exploratory spatial data analysis (ESDA) includes amongst other activities the identification of data properties and formulating hypotheses from data (Good, 1983). It provides a methodology for drawing out useful information from data. The findings from exploratory analysis provide input into spatial modelling. Modelling involves specification, parameter estimation and inference (testing hypotheses, computing confidence intervals, assessing goodness of fit) through which the analyst hopes to estimate parameters of interest and test hypotheses. In assessing a model, the tools and methods of ESDA may again play a useful role, leading to further iterations of model specification, estimation and inference.

Over the last decade or so there have been a number of developments, theoretical and practical, which have had important implications for the conduct of spatial data analysis (Haining, 1996). We briefly sketch some of these developments by way of illustration which also provides some motivation for the timing of this book.

One of the first research agendas set by the United States National Center for Geographic Information and Analysis (NCGIA) after its founding in the second half of the 1980s was in the area of spatial data accuracy (Goodchild and Gopal, 1989). Research in this area particularly into the nature of error in spatial data and how such error may propagate as a result of performing different types of map operations like overlaying or buffering has important implications for the conduct of data analysis. It helps to define the limits to what may be concluded from the analysis of spatial data (Arbia et al., 1999). This remains an area of special importance at a time when there are ever-growing volumes of spatially referenced data produced both by government agencies and the private sector. This research focus on data accuracy and data quality in turn led to a focus on issues of spatial representation (when geographic reality is translated into digital form) and what terms such as 'quality', 'accuracy' and 'error' mean in the case of spatial data.

Exploratory spatial data analysis (ESDA) methods were not widely used in the late 1980s, although Cressie (1984) had written about methods for exploring geostatistical data and Openshaw and colleagues had developed a 'geographical analysis machine' for looking for clusters of events in inhomogeneous spatial populations (Openshaw et al., 1987). There has been considerable interest in this area since that time together with new research into visualization tools and associated software to support ESDA. Notable in this respect has been the pioneering work of statisticians including Haslett and colleagues (e.g. Haslett et al., 1990, 1991; Unwin et al., 1996) and geographers including Monmonier and MacEachren (e.g. Monmonier and MacEachren, 1992). One aspect of ESDA that has attracted interest is the development of local statistics (based on using spatial subsets of the data) in order to detect local properties and describe spatial heterogeneity. The complex nature of spatial variation has long been a subject of comment. The development of local statistics to compliment the array of familiar 'whole map' or global statistics that provide descriptions of the average properties of a map is in part a response to the recognition of the heterogeneous nature of spatial variation (e.g. Getis and Ord, 1992, 1996; Anselin, 1995; Fotheringham et al., 2000). New techniques for the detection of spatial clusters of events (such as clusters of a particular disease or crime) have been developed which represent additions to the spatial analysis toolkit (Besag and Newell, 1991; Kulldorff and Nagarwalla, 1995).

In the area of statistical modelling of spatial data, Bayesian approaches attracted attention in the 1990s, in part because of the availability of numerical methods within new software for fitting a wide range of models (Gilks et al., 1996). Prior to the early 1990s much spatial modelling was based on spatial modifications to the linear regression model in which, for example, spatial dependence was modelled through the response variable. There were few applications of Bayesian methods (for an exception see Hepple, 1979). Bayesian methods have introduced other ways for modelling the effects of spatial dependence.

Over the last decade there have been important advances in software. The requirement to write ones own software with the attendant anxiety of making subtle (and not so subtle) programming errors, always acted as a brake on the utilization of spatial analysis methods particularly outside statistics. Bailey and Gatrell (1995) provided software as part of their book although this software was largely for teaching purposes. There have been considerable advances in making software available including much that can be downloaded free off the web. This book is not linked to any one piece of spatial analysis software but the appendix gives a list of software that can be used to implement many of the methods discussed in this book.

Geographic Information Systems (GIS) are software systems for capturing, storing, managing and displaying spatial data. The ability to directly capture events (such as the location of the offence by an officer attending the crime scene) on to a GIS has important implications for the rapid and timely accumulation of spatial data and the conduct of analysis. One of their most important capabilities for the purpose of spatial data analysis is that they provide a platform for integrating different data sets that may not necessarily be referenced to the same spatial framework. The problem of *how* to link data sets that derive from incompatible spatial frameworks (for example linking pixel-based environmental data and enumeration district-based population data) has attracted considerable interest. Less attention however seems to have been paid to the *consequences* of such linkage on the conduct of analysis and the interpretation of results, given the errors and uncertainties that such linkage necessarily induces in the database. Commercial GIS (e.g. ArcGIS Geostatistical Analyst) now provides some spatial analysis capability including statistical analysis. This opens up at least in some areas of research the possibility for a seamless environment for the storage, management, display and also *analysis* of spatially referenced statistical data.

The breadth of disciplinary interest in spatial data analysis is evident from earlier books in the field (Ripley, 1981; Upton and Fingleton, 1985; Anselin, 1988; Haining, 1990; Cressie, 1991; Bailey and Gatrell, 1995) and edited volumes such as Fotheringham and Rogerson (1994), Fischer et al. (1996) and Longley and Batty (1996). The continued vitality of the field over the last decade is illustrated by the growing number of applications and the increasing number of journals that have carried theme issues (see, e.g., special issues of *Papers in Regional Science*, 1991 (3); *Computational Statistics*, 1996 (4); *International*

Regional Science Review, 1997 (1&2); *The Statistician*, 1998 (3); *Journal of Real Estate Finance and Economics*, 1998 (1); *Statistics in Medicine*, 2000 (19, parts 17 and 18)).

The emergence of an area of quantitative research is due to the availability of good-quality data, the emergence of well-formulated hypotheses that can be expressed in mathematical terms, the availability of appropriate mathematical and statistical tools and techniques and the availability of technology for facilitating analysis. This diagnosis seems to apply to geographical and environmental epidemiology and to health services research (Cuzick and Elliott, 1992) where the availability of geo-coded health data and the growth of geographical databases and the development of new statistical techniques has generated considerable research activity. The expanding use of spatial analysis methods reflects the significance of place and space in theorizing disciplinary subfields such as the interest in area contextual effects in explaining health behaviours.

Developments in criminology seem to reflect a similar pattern: the collection of geo-coded offence, offender and victim data and the development of local statistics and their availability through spatial analysis software such as provided by the United States' National Institute of Justice. Such work is given further impetus through theorizing the role of spatial relationships and the context of place in shaping offence, offender and victim geographies (Bottoms and Wiles, 1997).

0.4 Organization

Chapter 1, discusses the relevance of spatial data analysis in selected areas of scientific and policy-related research and provides motivation for the rest of the book. Chapter 2, discusses the nature of spatial data and the relationship between the spatial data matrix, the foundation of all the analysis in this book, and the geographical reality it seeks to capture. This chapter draws heavily on the geographical information science literature. This leads to a discussion of data quality issues and methods of quantifying spatial dependence. This book is not just about the issues for the analysis of spatial data raised by spatial dependence but, as already noted, it is an important property that influences many stages and aspects of data analysis.

Chapter 3 discusses sources of spatial data and concentrates in particular on spatial sampling. There is a section on obtaining simulated data from spatial models. Chapter 4 looks at the implications of different aspects of data quality for the conduct of spatial data analysis. The emphasis is on techniques that address some of the problems often encountered with spatial data such as missing values, data on incompatible areal systems and inference problems associated with ecological (spatially aggregated) data.

Chapters 5 to 7 deal with exploratory spatial data analysis. Chapter 5 is a short chapter describing different conceptual models of spatial variation that might be used to underpin a programme of exploratory data analysis in the sense of specifying in a quite informal way what spatial structures might be looked for in a data set. Chapter 6 deals with visual methods for exploring spatial data whilst chapter 7 describes numerical methods for identifying data properties concentrating on spatial smoothing, clustering methods and map comparison methods. Splitting in this way makes the discussion of ESDA more manageable, in my view, but it is important to remember that visual and numerical methods are complementary.

Chapter 8 describes some of the implications for carrying out statistical tests when data are not independent. Tests of differences of means, bivariate correlation tests and chi-square tests on spatial data are discussed. These topics are approached through the concept of the 'effective sample size' which means calculating the amount of information about the process contained in the dependent set of data. The 'nuisance' aspect of spatial dependence (in the statistical sense) is that positive spatial dependence reduces the amount of information the analyst has for making inferences about the population. This reduction is relative to the amount of information the analyst would have if the observations were independent.

Chapters 9 to 11 discuss the modelling of spatial data. Chapter 9 describes statistical models for spatial variation when the data are from a continuous surface and when data refer to areal aggregates or point or area objects. Just as chapter 5 describes models that underpin ESDA, chapter 9 describes the range of descriptive and explanatory models that underpin formal data analysis where the aim is to estimate parameters of interest and test hypotheses. Models for representing spatial variation are mentioned and occasionally used in earlier chapters. The reader is encouraged, after completing chapter 2, to dip into section 9.1 especially for material on spatial covariance and spatial autoregressive models.

Chapter 10 discusses and provides examples of descriptive spatial modelling where the aim is to find a model to represent the variation in a response variable. The coverage includes trend surface models with independent errors and trend surface models with spatially correlated errors. Models for describing spatial variation in discrete-valued regional variables are also treated. Bayesian methods for disease mapping are applied. Chapter 11 discusses and provides examples of explanatory modelling using regression where the spatial variation in a response is modelled in terms of covariates. This

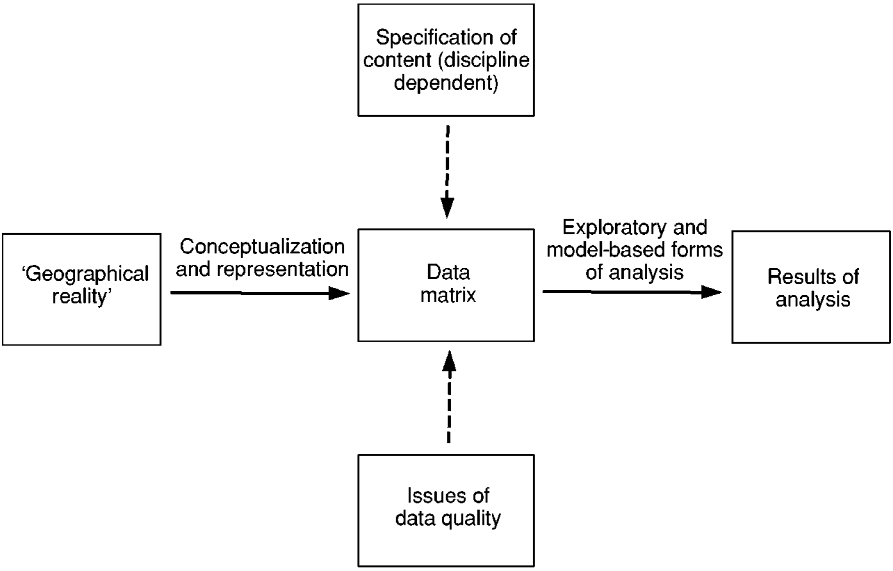
Specification of content (discipline dependent)
'Geographical reality'
Conceptualization and representation
Data matrix
Exploratory and model-based forms of analysis
Results of analysis
Issues of data quality

Figure 0.2 Overall structure of the book

chapter includes a review of different methodologies for modelling in non-experimental sciences. There is an appendix on available software.

Figure 0.2 represents the overall structure of the book.

0.5 The spatial data matrix

Underlying all the analyses here is a data matrix. We now indicate the content of that matrix as well as introducing some notation.

Let $Z_1, Z_2, \ldots, Z_k$ refer to k variables or attributes and $\mathbf{S}$ to location. The type of spatial data set to be considered in this book can be represented as:

$$
\begin{array}{cl}
\begin{array}{cc} \text{Data on the } k \text{ variables} & \quad\text{Location} \end{array} & \\
\left[\begin{array}{ccccc}
z_1(1) & z_2(1) & \cdots & z_k(1) & \mathbf{s}(1) \\
z_1(2) & z_2(2) & \cdots & z_k(2) & \mathbf{s}(2) \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
z_1(n) & z_2(n) & \ldots & z_k(n) & \mathbf{s}(n)
\end{array}\right] &
\begin{array}{l}
\text{Case } 1 \\
\text{Case } 2 \\
\vdots \\
\text{Case } n
\end{array}
\end{array}
\tag{0.1}
$$

The use of the lower case symbol on Z and $\mathbf{S}$ denotes an actual data value whilst the number inside the brackets, 1, 2, . . . etc, references the particular case. Attached to every case (i) is a location $\mathbf{s}(i)$. In a later chapter we shall be more specific about how the location of a case is referenced and what other information on $\mathbf{s}(1), \ldots, \mathbf{s}(n)$ may need to be recorded in order to undertake analysis – such as which other sites (cases) represent spatial neighbours of any given site (case). At this stage, however, and since we are only interested in two-dimensional space, it is sufficient to note that there will be occasions when the referencing will involve two co-ordinates. Together these fix the location of the case with respect to two axes that are at right angles to one another (orthogonal). So, the bold font for **s** signals that this is a vector and may contain more than one number for the purpose of identifying the spatial location of the case: for example $\mathbf{s}(i) = (s_1(i), s_2(i))$. In this book we only look at methods that treat the locations as fixed – we will not be looking at problems where there is randomness associated with the locations of the cases.

The structure (0.1) can be shortened to the form:

$$\{z_1(i), z_2(i), \ldots, z_k(i) \mid \mathbf{s}(i)\}_{i=1,\ldots,n} \tag{0.2}$$

and when no confusion arises the notation outside the curly brackets will be dropped.

In addition to possessing a spatial reference, data also have, at least implicitly, a temporal reference. The type of data set specified by 0.2 might be re-expressed in the form:

$$\{z_1(i,t), z_2(i,t), \ldots, z_k(i,t) \mid \mathbf{s}(i), t\}_{i=1,\ldots,n} \tag{0.3}$$

where t denotes time. However, all data values are meant to refer to the same point in time which is why t will be suppressed in the notation. The implications of this assumption will need careful consideration in any particular analysis because it is not always possible to have data on different attributes referring to the same time period. Population censuses for example are only taken every 10 years in the UK. At this stage we simply note that this is not a book about analysing space–time variation, except inasmuch as we might compare results arising from separate analyses of two or more time periods.

On various occasions throughout the book, depending on the context, the variables or attributes $Z_1, Z_2, \ldots, Z_k$ will be divided into groups and labelled differently. In the case of data modelling, the variable whose variation is to be modelled will be denoted Y. In regression Y is called the *response* or *dependent* or *endogenous* variable. The variables used to explain the variation in the response are called *explanatory* or *independent*, or *exogenous* or *predictor* variables and are usually labelled differently such as $X_1, X_2, \ldots, X_k$.