Interactive spatial data analysis

Trevor C. Bailey
Department of Mathematical Statistics
and Operational Research
University of Exeter

Anthony C. Gatrell
Department of Geography
Lancaster University

LONGMAN

1

Spatial data analysis

Here we give an introduction to spatial data analysis and distinguish it from other forms of data analysis. We identify different classes of spatial problem and types of spatial data, and draw a loose distinction between techniques for visualising, exploring and modelling of such data. We then explain the scope and organisation of the book in terms of these distinctions and introduce some fundamental issues related to spatial analysis and modelling. We conclude with a short discussion of some practical problems which confront the spatial analyst and which we urge our reader to bear in mind throughout subsequent chapters.

1.1 Introduction

Imagine you work for the research department of a major political party and have been asked to look at the relationship between support for that party and levels of personal income. You have survey data from a sample of electoral districts which includes the percentage support for your political party and corresponding average per capita personal income. Since you are the kind of person that we would expect to read this book—that is, you have at some time taken an introductory course in statistical methods, even though you have not necessarily specialised in the subject—we hope that you would feel fairly confident about assessing the nature and strength of any relationship between these two variables. You remember something about relevant basic techniques that might be of use. For example, perhaps you might produce a scatter plot of the two variables against each other and, if there appears to be a reasonable degree of association, you might go on to calculate a correlation coefficient between them and formally test that its value is statistically significant. You

may wish to model more precisely the quantitative nature of the relationship, maybe by fitting a linear regression between the two variables, and so on.

These kinds of general purpose data analyses will undoubtedly give some insight into the strength of the relationship in which you are interested; but you will probably be rightly wary of placing too much confidence in the results. You realise that there will be a whole host of other factors affecting voting preference, which you have not accounted for, and which may be confounding any relationship that you have established. Some of these you may be able to address by straightforward extensions to your earlier analysis, for example by including additional 'explanatory' variables into your regression model. However, one factor which may not immediately strike you as a problem, but in which we are very interested in this book, is that the observational units you have been analysing (electoral districts) have a spatial configuration—some may neighbour others, some may be a long way apart—and this may be relevant to your analysis.

It is very likely, for example, that both party support and per capita income in one district are related to those in neighbouring districts, in addition to any relationship that they may have with each other. The reasons for this may be too complex to disentangle and account for separately in your model except by reference to the geographical proximity of the districts as a surrogate measure. Even though you are not interested in a spatial question, one may have entered through the 'back door'. Most elementary general purpose data analysis makes the fundamental assumption that the observational units analysed represent *independent* pieces of evidence about the relationship under study. In your case, depending on your particular sample of electoral districts, the values of the variables which you are analysing may well be *spatially correlated* across observational units. In essence you may have less independent evidence about the relationship than you think and somehow you need to allow for this in your analysis.

In fact, you are in a similar, although possibly more complex, situation to a colleague of yours who has been asked to explore the way in which overall support for the party is related to changes in national economic growth. Her problem is that she is, by necessity, forced to examine any possible relationship in these variables by reference to data for successive time periods. However, it is very likely that party support in any time period will be partly a function of what happened in preceding time periods, in addition to any relationship with economic growth in the current period—data for her observational units (time periods) are serially or *temporally correlated*. If she were to analyse such data using standard methods which are intended for independent observations (such as the views of different, randomly selected individuals in an opinion poll), then she may well come to misleading conclusions—similarly, for you and your analysis. In her case the possibilities for the structure of such temporal correlation are relatively straightforward—time is one dimensional and the direction of correlation will be backwards in time, the only question being how far. Your case is possibly more complex—any spatial correlation in your data occurs in two dimensions and no particular direction is ruled out *a priori*.

She needs to consult a book on *time series analysis*. You, we presume, have already purchased (or borrowed!) this one on *spatial data analysis* and commenced reading it. We hope you will persevere, because this book should not only prove useful to you in modifying your analysis of non-spatial hypotheses to account for the use of data which are spatially or geographically referenced, but will also help you to address explicitly spatial questions which are not discussed at all in many elementary texts on data analysis—such as forecasting the effect that forthcoming changes in the boundaries of electoral districts might have on the party's fortunes in future elections.

Furthermore, if your area of work or interest is outside party politics, you will be pleased to know that we have not written this book solely to enhance the promotional prospects of political party researchers! We will be concerned generally with ways of analysing all varieties of data in a spatial context. Here are some other examples of the kinds of problem we are interested in, just to set the scene. We hope that you will agree that they are all important, practical, problems.

- Seismologists collect data on the regional distribution of earthquakes. Does this distribution show any pattern or predictability over space?
- Public health specialists (epidemiologists) collect data on the occurrence of diseases. Does the distribution of cases of a disease form a pattern in space? Is there some association with possible sources of environmental pollution? Is there any evidence that a particular disease is passed on from one individual to another?
- Police wish to investigate if there is any spatial pattern to the distribution of burglaries. Do the rates of burglaries in particular areas correlate with socio-economic characteristics of those areas?
- Environmental scientists (but also others concerned with 'image' data) collect data obtained from satellites. Such data may be very 'noisy'. Can we filter out the noise to reveal underlying pattern?
- Geologists wish to estimate the extent of a mineral deposit over a particular region, given data on borehole samples taken from locations scattered across the area. How can we make sensible estimates?
- A groundwater hydrologist collects data on the concentration of a toxic chemical in samples collected from a series of wells. Can we use these samples to construct a regional map of likely contamination?
- Retailers wish to use socio-economic data, available for small areas from the population Census, to assess the likely demand for their products if they open, or expand, an outlet. How are we to classify such areas?
- The same retailers collect information on movements of shoppers from residential 'zones' to stores. Can we build models of such flows? Can we predict changes in such flows if we expand an outlet or open a new one?

These examples illustrate the breadth of application of spatial data analysis, and suggest that the subject is of relevance in many different fields. Whether we are geographers, statisticians, economists, sociologists, epidemiologists, planners, biologists or environmental scientists, we are often faced with

problems of a spatial nature, or with problems that involve spatial data. Our aim in this book is to show how the spatial nature of the data can be recognised explicitly and, if necessary, properly incorporated into our analyses.

Clearly, some of these disciplines will find a spatial dimension to their analyses more important than others. For example, geologists interested in searching for mineral deposits and predicting the volume of ore in rock have seen a major branch of their subject (and of spatial data analysis), 'geostatistics', emerge around this problem. Similarly, geographers have an obvious interest in spatial analysis techniques. For other researchers, such as sociologists and botanists, the spatial dimension, although often a necessary consideration in analysis, may be of secondary rather than primary importance in the sorts of problems they study.

This wide and varying interest in spatial data analysis has meant some disciplines have tended, independently of others, to develop their own distinctive styles of analysis. As a result, terminology and notation differ from discipline to discipline and inevitably a fair amount of 're-inventing the wheel' has gone on. This means that the subject of spatial data analysis is somewhat scattered throughout the literature of various disciplines and can appear, particularly to those coming to the field for the first time, as a collection of *ad hoc* techniques, with little sense of coherence or underlying theory, and can consequently seem rather difficult to get to grips with. This is not really so, as is well demonstrated by a number of recent unifying texts on the subject. However, such texts tend to require a fairly advanced statistical background. We hope in this book to follow their example and emphasise the coherence of the subject by structuring our discussion accordingly, but to do so at a more introductory level, addressing an audience which may not be so statistically specialised.

In the remainder of this chapter, we commence this task by first defining more precisely what we mean by spatial data analysis and what distinguishes it from its non-spatial counterpart. We try to convince our reader that 'space can make a difference'; in other words, that the recognition of a spatial dimension in analysis may yield different, and more meaningful, results, than an analysis which ignores it. From this we are able to move on to consider a set of illustrative case studies, which set the scene for the different classes of spatial problem addressed in the various subsequent parts of the book; each of which is related to one of these particular types of spatial problem. We follow this by a general discussion of the different types of spatial phenomena and spatial relationships that may arise in analysing any of these classes of problem. We then introduce some basic general concepts which we will apply in each part of the book to further structure the spatial analysis methods presented. In particular, we draw a distinction, which is useful, although not cast in tablets of stone, between methods which allow us to *visualise* spatial data, those which allow us to *explore* any structure and suggest possible hypotheses and finally, those that involve more formal statistical *models* of such data. Discussion of the latter gives us an opportunity to introduce some important ideas concerning the statistical modelling of spatial phenomena, which arise in

different contexts throughout the book. We conclude the chapter with a short discussion of a number of important practical problems which may confront the spatial analyst, acknowledging that many of these will never be resolved by analytical techniques alone, but will ultimately involve the insight, experience and more subjective judgement of the analyst in the particular case in question. We encourage the reader to bear these in mind in relation to any of the subsequent methods presented in the book.

1.2 Spatial versus non-spatial data analysis

In broad terms one might define spatial analysis as the quantitative study of phenomena that are located in space. However, it would indeed be ambitious to attempt to cover such a broad field in one book! We have therefore decided to limit our discussion in the chapters which follow, to that important subset of the subject which we shall refer to as *spatial data analysis*.

The particular distinction which we draw between *spatial analysis* and *spatial data analysis* needs some clarification. We are concerned in this book with the situation where observational data are available on some process operating in space and methods are sought to describe or explain the behaviour of this process and its possible relationship to other spatial phenomena. The object of analysis is to increase our basic understanding of the process, assess the evidence in favour of various hypotheses concerning it, or possibly to predict values in areas where observations have not been made. The data with which we are concerned constitute a sample of observations on the process from which we attempt to infer its overall behaviour.

By defining spatial data analysis in this way we place ourselves firmly in the area of statistical description and modelling of spatial data and accordingly restrict attention in the book to a particular set of methods. In doing so we have to exclude from our coverage some important quantitative methods which would be included under the more general heading of spatial analysis. For example, although we discuss very briefly various forms of network analysis, such as those concerned with routing problems, minimisation of transportation costs or the optimal siting of facilities, we do not go into great detail. These classes of problem do of course involve spatial data, but not primarily observational data of the type discussed above, and understanding, explaining, or predicting the data is not the objective of the analysis; these problems instead involve mathematical optimisation techniques. There are clearly areas where such methods do interact with those with which we are concerned and have defined as spatial data analysis. For example, network analysis may be useful to compute travelling times, which are then used as measures of spatial proximity in the modelling of some observed spatial process of interest. Alternatively, models of observational data on movement of people or goods between locations, may provide necessary input to methods concerned with the optimal siting of facilities. We shall endeavour to point out

such links and, where appropriate, will refer the reader to references on relevant spatial analysis methods beyond the scope of our definition of spatial data analysis.

Having clarified what we mean by spatial data analysis as opposed to spatial analysis more broadly conceived, we need to make explicit a further distinction at the outset—that between *spatial* data analysis and *non-spatial* data analysis. We do not intend in this book to discuss all forms of statistical analysis that may be useful in relation to data that happen simply to be located in space. This would require us to present (yet again!) many of the standard techniques found in numerous general purpose texts on statistical data analysis. We have no wish to do this. In fact, we assume that our reader already has some familiarity with these sorts of techniques. Rather, we wish to focus on modifications, extensions and additions to such techniques which consider explicitly the importance of the locations, or the spatial arrangement of the objects being analysed; in other words, on *spatial* as opposed to *non-spatial* data analysis.

It is unnecessary for us to become too pedantic about precisely where this dividing line between spatial and non-spatial data analysis actually lies. We have no wish to become involved in theoretical 'hair splitting' between what is, and what is not, a spatial statistical method. For our purposes it will suffice to say that spatial data analysis is involved when data are spatially located *and* explicit consideration is given to the possible importance of their spatial arrangement in the analysis or in the interpretation of results.

A good example of the distinction we are trying to convey comes from the field of research known as island biogeography. Consider the relationship between number of plant species and geographical area for a set of small islands. There is a wealth of empirical evidence to suggest that the logarithm of the number of species is related to the logarithm of the area of the island. One explanation of this kind of empirical relationship is simply that as area increases there is a greater probability of a range of available habitats. Spatial data analysis is not necessarily involved at this stage; the mere fact that the units of observation (islands) are locational, or that one of the variables involved is geographical area, does not itself make the analysis a spatial one. However, other theories are more explicitly 'spatial'. We might expect, *a priori*, that the isolation of an island is an important factor, in terms of its distance from other islands or from a continental area. Indeed, some authors have considered modifications to estimation of the relationship between species number and area which allow for such factors. Work on the distribution of bird species in Pacific islands, for example, shows that isolation, in terms of distance from New Guinea, does indeed reduce the number of species. Such analysis is more clearly spatial, since the relative locations of the spatial units (their proximity to the mainland) are being exploited in the analysis.

The reason for our intended focus in this book on spatial, as opposed to non-spatial, forms of data analysis, is that when spatial data are involved the former often yield different and more meaningful results than analyses which ignore the spatial dimension. We tried to convince our imaginary political

researcher of this in the introduction to this chapter. Let us now consider in more detail a further example to consolidate the point. We do not set this up as an exemplar of 'ideal' spatial data analysis; indeed, there are shortcomings. But it does illustrate well the importance of a spatial perspective when dealing with spatial data.

Consider the very geographical problem of trying to model spatial variation in precipitation in California. Suppose we take a set of 30 monitoring stations, distributed across the state as in Figure 1.1. For each of these we have recordings of: average annual precipitation (the 'response' variable of interest); and altitude, latitude, and distance from the coast, each of which is a possible covariate that might explain the variation in precipitation.

We include the data, along with a simple map, on the disk we supply with this book, as we do with other data sets we will be using in later chapters. Of course, to 'look' at this data you will need some software. You may already be familiar with simple mapping packages, or even more advanced software for handling geographical data. If so, you can 'import' the data we supply into such software. If not, then you can use the software that we have provided with this book to 'view' the data sets. In Chapter 2 we shall be looking in detail at using computers to view, manipulate and analyse spatial data, and at that time

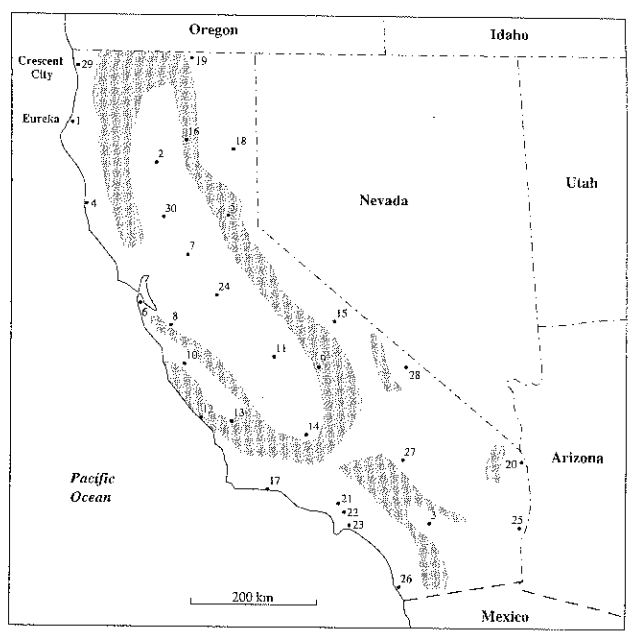

Fig. 1.1 Locations of rainfall measurement sites in California (shaded areas are mountains)

you can place the software package that we have provided into the wider context of what is currently possible and potentially desirable in terms of computing developments related to our subject. It will become apparent there that this is an area of rapid technological change and in no way do we wish to tie our text to any specific computer package. Therefore, throughout the book we avoid specific reference to use of our software, except in the series of computer exercises that accompany chapters. This is a book about spatial data analysis, and not about any particular software package. However, in the computer exercises we do encourage the reader to try out some of the ideas that we have described in the text, using our software and the data sets and maps that we have provided with it. In the course of discussing these exercises, we do become involved in specific details of using our package. At points in the text where there is a relevant computer exercise for the reader to refer to, we simply place a symbol in the margin to indicate this, such as the one which you see here. If you wish to follow our scheme you may like to refer to our first exercise now. If you are the kind of reader who prefers not to be distracted with our practical computer exercises, or wishes to use alternative computer software for spatial analysis and mapping, you can simply ignore the symbols in the margin, and rest assured that we will not worry you with details of the exercises in the main text.

Returning to our main discussion, in the paper that originally reported the Californian rainfall data, Peter Taylor fits a standard multiple regression model to the data. This involves trying to find the best linear combination of the covariates that explains variation in precipitation. Taylor finds that all three covariates are significant predictors of rainfall and about 60 per cent of the variation in precipitation is explained by them. Taylor then proceeds to map the residuals (the differences between the observed values of precipitation at the stations and those predicted by his three-variable model). He does this in order to see if there is any spatial pattern to these differences. The mountain ranges in the area are also indicated in Figure 1.1 and he finds a 'clustering' of negative residuals on the leeward side of the mountains. In other words, the model over-predicts precipitation at these locations. He therefore introduces a fourth covariate, where stations are coded 1 if they lie in the lee of mountains, 0 otherwise. When this variable is added to the model the explained variation in precipitation rises to 74 per cent, the rain shadow effect being highly significant. Mapping the residuals from this new model reveals no obvious spatial patterning, though two sites are still poorly predicted.

In an extension to the analysis of these data, Kelvyn Jones has shown how further, graphical, exploratory analysis of the data yields benefits. One modification is to examine the nature of relationships between the response variable and each covariate, in order to check for non-linearity; for example, although precipitation increases with altitude this may not be a simple linear increase with height. But, interestingly, the incorporation of an interaction effect, between distance from the coast and the rain shadow variable, generates a particularly good fit. The final model suggests that precipitation will be lowest when a station is both well inland *and* within the rain shadow.

Although there appears to be no obvious spatial relationship in the final residuals, missing from the analysis reported so far is any formal check that no *spatial correlation* or 'spatial persistence', remains in these unexplained variations in precipitation. We mentioned this consideration in our earlier example relating to support for a particular political party and levels of per capita income. In this case it may be that the amount of rainfall observed at one site will be similar to that measured close by, for reasons other than those explicitly incorporated into the model presented so far. If so then we may be able to improve our model by taking this into account in some way. We shall see in later chapters exactly how to employ methods which seek to allow for such residual spatial dependence. At this stage the point we wish to make is simply that there are possibly yet further adjustments that we might have to make to our precipitation model to allow for residual rainfall levels at one site being similar to those nearby.

This seems to us to be a useful example of the importance of an 'awareness of space' when analysing spatial data. There is an emphasis on spatial exploration of both data and results, in the form of both mapping and graphical plots. This then informs the more formal model-fitting. Disparities between model and reality are mapped and interpreted in terms of the spatial arrangement of the observations and their geographical context. We could then go on to use the model in a spatially predictive sense, by generating predictions of precipitation at locations which are not themselves monitoring sites. In this way, there is a close interaction between the statistical modelling and spatial interpolation.

As our earlier illustrations have indicated there are several important classes of problem for which such an explicitly spatial perspective is useful in data analysis. In the next section we shall look in more detail at a set of four empirical studies, and relate each of these to the problem areas dealt with in the four subsequent parts of the book.

1.3 Classes of problem in spatial data analysis

We hope that the discussion so far has given our reader a clear idea of the kind of methods that we will be concerned with in subsequent chapters; in other words, how we have defined the subject of spatial data analysis for the purposes of this book. The next obvious question is how we intend to structure our discussion of the various techniques which fall under this heading. We want now to explain this by examining a range of particular case studies, relating each of these examples to particular parts of the book. As we said in the Preface, this is an applied rather than a theoretical book and so we will start as we mean to go on, letting the structure of the book be dictated by practical applications.

In each case we believe that the applications considered deal with quite important contemporary problems. We hope that this will have the additional

spatial data analysis. In this paper we describe how to implement exploratory spatial. data analysis tools into a commercial desktop GIS product.Â Interactive statistical graphics are reviewed in the contexts of spatial data and geographical information systems (GIS). GIS provide the user with an active geographical view of the dataâ€”a map that can be used as an entry point to the data base. Prototype softwareâ€”SPIDERâ€”illustrates the possibilities of using statistical graphics as further views of the data, which can be made active and thus provide alternative means of querying the data. These views can be cross-referenced by 'linking'. It is argued that such a system can provide a very rich environment for pursuing exploratory st