GEOGRAPHIC INFORMATION ANALYSIS
David O'Sullivan and David J. Unwin

Chapter 2

The Pitfalls and Potential of Spatial Data

CHAPTER OBJECTIVES

In this chapter, we attempt to:

- Justify the view that spatial data are in some sense special
- Identify a number of problems in the statistical analysis of spatial data associated with *spatial autocorrelation*, *modifiable areal units*, the *ecological fallacy*, and *scale*, what we call the "bad news"
- Outline the ideas of *distance*, *adjacency*, *interaction,* and *neighborhood* —the "good news" central to much spatial analysis
- Show how *proximity polygons* can be derived for a set of point objects
- Introduce the idea that these relations can be summarized using matrices and encourage you to spend some time getting used to this way of organizing geographic data

After reading this chapter, you should be able to:

- List four major problems in the analysis of geographic information
- Outline the geographic concepts of distance, adjacency, interaction, and neighborhood and show how these can be recorded using matrix representations
- Explain how proximity polygons and the related Delaunay triangulation can be developed for point objects

2.1. INTRODUCTION

It may not be obvious why spatial data require special analytic techniques, distinct from standard statistical analysis that might be applied to ordinary data. As the spatial aspect of more and more data has been recognized due to

the diffusion of GIS technologies and the consequent enthusiasm for mapping, this is an important issue to understand clearly. The academic literature is replete with essays on why space either is or is not important, or on why we should take the letter *G* out of GIS. In this chapter, we consider some of the reasons why adding a spatial location to some attribute data changes them in fundamental ways.

There is bad news and good news here. Some of the most important reasons why spatial data must be treated differently appear as problems or pitfalls for the unwary. Many of the standard techniques and methods documented in statistics textbooks are found to have significant problems when we try to apply them to the analysis of spatial distributions. This is the bad news, which we deal with first in Section 2.2. The good news is presented in Section 2.3. This boils down to the fact that geospatial referencing inherently provides us with a number of new ways of looking at data and the relations among them. The concepts of *distance*, *adjacency*, *interaction,* and *neighborhood* are used extensively throughout this book to assist in the analysis of spatial data. Because of their all-encompassing importance, it is useful to introduce these concepts early on in a rather abstract and formal way so that you begin to get a feel for them immediately. Our discussion of these fundamentals also includes *proximity polygons*, which appear repeatedly in this book and are increasingly being used as an interesting way of looking at many geographical problems. We hope that by introducing these concepts and ideas early on, you will begin to understand what it means to *think spatially* about the analysis of geographic information.

2.2. THE BAD NEWS: THE PITFALLS OF SPATIAL DATA

Conventional statistical analysis frequently imposes a number of conditions or assumptions on the data it uses. Foremost among these is the requirement that samples be random. The most fundamental reason that spatial data are special is that they almost always violate this requirement. The technical term describing this problem is *spatial autocorrelation*, which must therefore come first on any list of the pitfalls of spatial data. Other closely related tricky problems frequently arise, including the *modifiable areal unit problem*, associated issues of *scale* and *edge effects,* and the *ecological fallacy*.

Spatial Autocorrelation

Spatial autocorrelation is a complicated name for the obvious fact that *data from locations near one another in space are more likely to be similar than data from locations remote from one another*. If you know that the elevation

at point X is 250 m, then you have a good idea that the elevation at point Y, 10 m from X, is probably in the range 240 to 260 m. Of course, there could be a huge cliff between the two locations. Location Y *might* be 500 m above sea level, although it is highly unlikely. It will almost certainly not be 1000 m above sea level. On the other hand, location Z, 1000 m from X, certainly could be 500 m above sea level. It could even be 1000 m above sea level or even 100 m *below* sea level. We are much more uncertain about the likely elevation of Z because it is farther away from X. If Z is instead 100 km away from X, almost anything is possible, because knowing the elevation at X tells us very little about the elevation 100 km away.

If spatial autocorrelation were not commonplace, then geographic analysis would be of little interest and geography would be irrelevant. Again, think of spot heights: we know that high values are likely to be close to one another and in different places from low values—in fact, these are called *mountains* and *valleys*. Many geographic phenomena can be characterized in these terms as local similarities in some spatially varying phenomenon. Cities are local concentrations of population—and much else besides: economic activity and social diversity, for example. Storms are local foci of particular atmospheric conditions. Climate consists of the repeated occurrence of similar spatial patterns of weather in particular places. *If geography is worth studying at all, it must be because phenomena do not vary randomly through space*. The existence of spatial autocorrelation is therefore a given in geography. Unfortunately, it is also an impediment to the application of conventional statistics.

The nonrandom distribution of phenomena in space has various consequences for conventional statistical analysis. For example, the usual parameter estimates based on samples that are not randomly distributed in space will be biased toward values prevalent in the regions favored in the sampling scheme. As a result, many of the assumptions we are required to make about data before applying statistical tests become invalid. Another way of looking at this is that spatial autocorrelation introduces *redundancy* into data, so that each additional item of data provides less new information than is indicated by a simple assessment based on n, the sample size. This affects the calculation of confidence intervals and so forth. Such effects mean that there is a strong case for assessing the degree of autocorrelation in a spatial data set before doing any conventional statistics at all. Diagnostic measures for the autocorrelation present in data are available, such as *Moran's I,* and *Geary's C*, and these will be described in Chapter 7. Later, in Chapter 10, we introduce the *variogram cloud*, a plot that also helps us understand the autocorrelation pattern in a spatial data set.

These techniques help us to describe how useful knowing the location of an observation is if we wish to determine the likely value of an attribute

measured at that location. There are three general possibilities: *positive autocorrelation*, *negative autocorrelation,* and *noncorrelation* or *zero autocorrelation*. Positive autocorrelation is the most commonly observed case and refers to situations where nearby observations are likely to be similar to one another. Negative autocorrelation is much less common and occurs when observations from nearby locations are likely to be different from one another. Zero autocorrelation is the case where no spatial effect is discernible and observations seem to vary randomly through space. It is important to be clear about the difference between negative and zero autocorrelation, as students frequently confuse the two.

Describing and modeling patterns of variation across a study region, effectively *describing the autocorrelation structure*, is of primary importance in spatial analysis. Again, in general terms, spatial variation is of two kinds: first- and second-order. *First-order* spatial variation occurs when observations across a study region vary from place to place due to changes in the underlying properties of the local environment. For example, the rates of incidence of crime might vary spatially simply because of variations in the population density, such that they increase near the center of a large city. In contrast, *second-order* variation is due to interaction effects between observations, such as the occurrence of crime in an area making it more likely that there will be crimes surrounding that area, perhaps in the shape of local "hotspots" in the vicinity of bars and clubs or near local street drug markets. In practice, it is difficult to distinguish between first- and second-order effects, but it is often necessary to model both when developing statistical methods for handling spatial data. We discuss the distinction between first- and second-order spatial variation in more detail in Chapter 4 when we introduce the idea of a spatial process.

Although autocorrelation presents a considerable challenge to conventional statistical methods and remains problematic, quantitative geographers have made a virtue of it by developing a number of autocorrelation measures into powerful descriptive tools. It would be wrong to claim that the problem of spatial autocorrelation has been solved, but considerable progress has been made in developing techniques that account for its effects and in taking advantage of the opportunity it provides for useful geographic descriptions.

The Modifiable Areal Unit Problem

Another major difficulty with spatial data is that they are often aggregates of data originally compiled at a more detailed level. The best example is a national census, which is collected at the household level but reported for

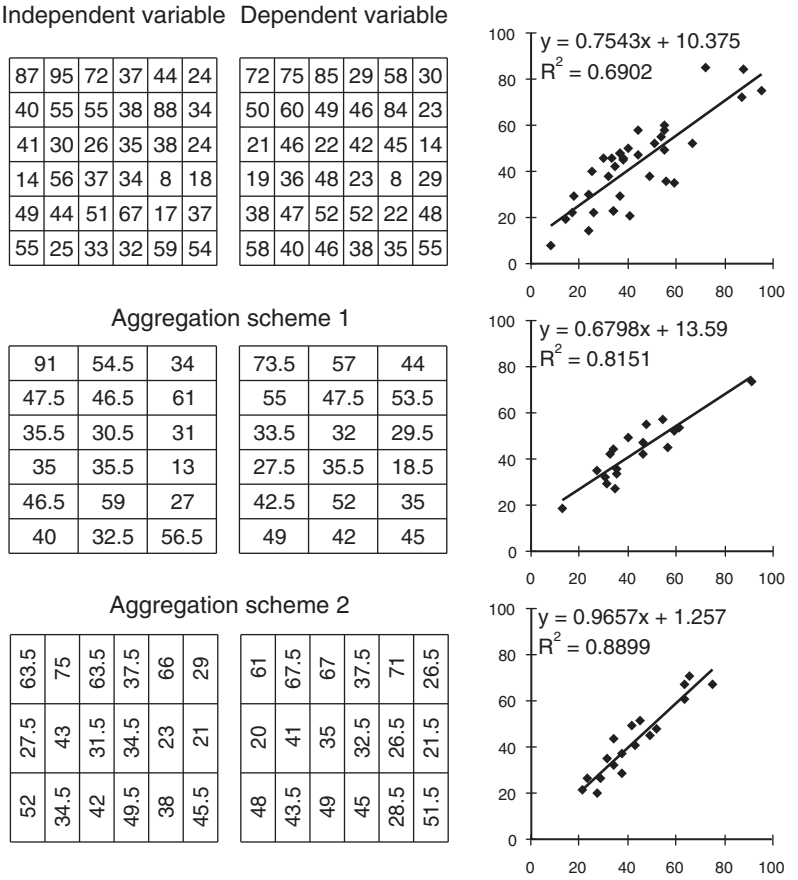

Figure 2.1 An illustration of MAUP.

practical and privacy reasons at various levels of aggregation such as city districts, counties, and states. The problem is that the aggregation units used are *arbitrary* with respect to the phenomena under investigation, yet the units used will affect statistics determined on the basis of data reported in this way. This difficulty is referred to as the *modifiable areal unit problem* (MAUP). If the spatial units in a particular study were specified differently, we might observe very different patterns and relationships. The problem is illustrated in Figure 2.1, where two different aggregation schemes applied to a spatial data set result in two different regression results. There is a clear impact on the regression equation and the coefficient of determination, R^2. This is an artificial example, but the effect is general and is not widely understood, even though it has been known for a long time (see Gehlke and Biehl, 1934). Usually, as shown here, regression relationships are strengthened by aggregation. In fact, using a simulation

approach, Openshaw and Taylor (1979) showed that with the same underlying data, it is possible to aggregate units together in ways that can produce correlations *anywhere* between −1.0 and +1.0!

The effect is not altogether mysterious, and two things are happening. The first relates to the scale of analysis and to aggregation effects such that combining any pair of observations will produce an outcome that is closer to the mean of the overall data, so that after aggregation, the new data are likely to be more tightly clustered around a regression line and thus to have a stronger coefficient of determination. This effect is shown in our example by both the aggregation schemes used producing better fits than the original disaggregated data. Usually this problem persists as we aggregate up to larger units. A second effect is the substantial differences between the results obtained under different aggregation schemes. The complications are usually referred to separately as the *aggregation* effect and the *zoning* effect.

MAUP is of more than academic or theoretical interest. Its effects have been well known for many years to politicians concerned with ensuring that the boundaries of electoral districts are defined in the most advantageous way for them. It provides one explanation for why, in the 2000 U.S. presidential election, Al Gore, with more of the popular vote than George Bush, still failed to become president. A different aggregation of U.S. counties into states could have produced a different outcome (in fact, it is likely that in this very close election, switching just one or two northern Florida counties to Georgia or Alabama would have produced a different outcome.)

The practical implications of MAUP are immense for almost all decision-making processes involving GIS technology, since with the now ready availability of detailed but still aggregated maps, policy could easily focus on issues and problems which might look very different if the aggregation scheme used were changed. The implication is that our choice of spatial reference frame is itself a significant determinant of the statistical and other patterns we observe. Openshaw (1983) suggests that a lack of understanding of MAUP has led many to choose to pretend that the problem does not exist in order to allow *some* analysis to be performed so that we can "just get on with it." This is a little unfair, but the problem is a serious one that has had less attention than it deserves. Openshaw's suggestion is that the problem be turned into an exploratory and descriptive tool, as has been done with spatial autocorrelation. In this view, we might postulate a relationship between, say, income and crime rates. We would then search for an aggregation scheme that maximizes the strength of this relationship. The output from such an analysis would be a spatial partition into areal units. The interesting geographic question then becomes "Why do these zones produce the strongest relationship?" Perhaps because of the computational complexities and

the implicit requirement for very detailed individual-level data, this idea has not been widely taken up.

The Ecological Fallacy

MAUP is closely related to a more general statistical problem: the *ecological fallacy*. This arises when a statistical relationship observed at one level of aggregation is assumed to hold because the same relationship holds at a more detailed level. For example, we might observe a strong relationship between income and crime at the county level, with lower-income counties being associated with higher crime rates. If from this we conclude that lower-income individuals are more likely to commit a crime, then we are falling for the ecological fallacy. In fact, it is only valid to say exactly what the data say: that lower-income counties tend to experience higher crime rates. What causes the observed effect may be something entirely different—perhaps lower-income families have less effective home security systems and are more prone to burglary (a relatively direct link); or lower-income areas are home to more chronic drug users who commit crimes irrespective of income (an indirect link); or the cause may have nothing at all to do with income.

It is important to acknowledge that a relationship at a high level of aggregation *may* be explained by the same relationship operating at lower levels. For example, one of the earliest pieces of evidence to make the connection between smoking and lung cancer was presented by Doll (1955; cited by Freedman et al., 1998) in the form of a scatterplot showing per capita national rates of cigarette smoking and the rate of death from lung cancer for 11 countries. A strong correlation is evident in the plot. However, we would be wrong to conclude, based on this evidence alone, that smoking is a cause of lung cancer. It turns out that it is, but this conclusion is based on many other studies conducted at the individual level. Data on smoking and cancer at the country level can still only support the conclusion that countries with larger numbers of smokers tend to have higher death rates from lung cancer.

Having now been made aware of the problem, if you pay closer attention to the news, you will find that the ecological fallacy is common in everyday and media discourse. Crime rates and (variously) the death penalty, gun control or imprisonment rates, and road fatalities and speed limits, seat belts, or cycle helmet laws are classic examples. Unfortunately, the fallacy is almost as common in academic discourse! It often seems to arise from a desire for simple explanations, but in human geography things are rarely so simple. The common thread tying the ecological fallacy to MAUP is that statistical relationships may change at different levels of aggregation.

Scale

This brings us neatly to the next point. The *geographic scale* at which we examine a phenomenon can affect the observations we make, and this must always be considered prior to spatial analysis. We have already encountered one way that scale can dramatically affect spatial analysis since the object type that is appropriate for representing a particular entity is scale dependent. For example, at the continental scale, a city is conveniently represented by a point. At the regional scale, it becomes an area object. At the local scale, the city becomes a complex collection of point, line, area, and network objects. The scale we work at affects the representations we use, and this in turn is likely to have effects on spatial analysis; yet, in general, the correct or appropriate geographic scale for a study is impossible to determine beforehand, and due attention should be paid to this issue.

Nonuniformity of Space and Edge Effects

A final significant issue distinguishing spatial analysis from conventional statistics is that *space is not uniform*. For example, we might have data on crime locations gathered for a single police precinct. It is very easy to see patterns in such data, hence the *pin maps* (see Chapter 3) seen in any self-respecting movie police chief's office. Patterns may appear particularly strong if crime locations are mapped simply as points without reference to the underlying geography. There will almost certainly be clusters simply as a result of where people live and work, and apparent gaps in (for example) parks or at major road intersections. These gaps and clusters are not unexpected but arise as a result of the nonuniformity of the urban space with respect to the phenomenon being mapped. Similar problems are encountered in examining the incidence of disease, where the location of the at-risk population must be considered. Such problems also occur in point patterns of different plant types, where underlying patterns in soil types, or simply the presence of other plant types, might lead us to expect variation in the spatial density of the plants we're interested in.

A particular type of nonuniformity problem, which is almost invariably encountered, is due to *edge effects*. These arise where an artificial boundary is imposed on a study, often just to keep it manageable. The problem is that sites in the center of the study area can have nearby observations in all directions, whereas sites at the edges of the study area only have neighbors toward the center of the study area. Unless the study area has been very carefully defined, it is unlikely that this reflects reality, and the artificially produced asymmetry in the data must be accounted for. In some specialized

areas of spatial analysis, techniques for dealing with edge effects are well developed, but the problem remains poorly understood in many cases.

2.3. THE GOOD NEWS: THE POTENTIAL OF SPATIAL DATA

It should not surprise you to find out that many of the problems outlined in Section 2.2 have not been solved satisfactorily. Indeed, in the early enthusiasm for the quantitative revolution in geography (in the late 1950s and the 1960s), many of these problems were glossed over. Unfortunately, dealing with these problems is not simple, so that only more mathematically oriented geographers and relatively small numbers of statisticians have paid much attention to the issues. More recently, with the advent of GIS and a much broader awareness of the significance of spatial data, there has been considerable progress, with new techniques appearing all the time. Regardless of the sophistication and complexity of the techniques adopted, the fundamental characteristics of spatial data are critical to unlocking their potential. To give you a sense of this, we continue our overview of what's special about spatial data, not focusing on the problems that consideration of spatial aspects introduces, but instead examining some of the potential for additional insight provided by an examination of the locational attributes of data.

The important spatial concepts that appear throughout this book are *distance*, *adjacency*, and *interaction*, together with the closely related notion of *neighborhood*. These appear in a variety of guises in most applications of statistical methods to spatial data. Here we point to their importance, outline some of their uses, and indicate some of the contexts where they will appear. The reason for the importance of these ideas is clear. In spatial analysis, while we are still interested in the distribution of values in observational data (classical descriptive statistical measures like the mean, variance, and so on), we are now *also* interested in the distribution of the associated entities *in space*. This *spatial distribution* can only be described in terms of the relationships between spatial entities, and spatial relationships are usually conceived in terms of one or more of the relationships we call distance, adjacency, interaction, and neighborhood.

Distance

Distance is usually (but not always) described by the simple crow's flight distance between the spatial entities of interest. In small study regions, where the Earth's curvature effects can be ignored, simple *Euclidean*