Chapter 7: Spatial Analyses of Statistical Data in GIS

7.1 Introduction

Visualisation helps us to understand spatial patterns but often further investigation is needed to describe or explain these patterns. With quantitative data this frequently means that researchers will want to perform some form of statistical analysis. When using GIS, an analysis should use not just use conventional statistical techniques, which only focus on the attribute data, but should also incorporate the spatial component of the data. Bringing the two together is known as *spatial analysis*, or *geographical data analysis* (GDA).

Caution is required when using GIS-based analysis of spatial data, and lessons can be learned from the experience of GIS research in human geography. Early enthusiasm for a GIS-based approach led protagonists to claim that GIS provided a cohesive and scientific framework that could re-unite geography as a discipline. This was presented within a highly quantitative framework (see, for example, Openshaw 1991). Not surprisingly, this led to a backlash that particularly focused on the overtly quantitative and scientific approaches that GIS research was taking at this time. This, it was argued, marked a return to 'the very worst sort of positivism' (Taylor 1990, 211). It was also argued that as an academic sub-discipline, GIS lacked a strong epistemology and any treatment of ethical, economic or political issues (see, in particular, Pickles 1995).

As time has gone on, some of the more extreme claims about what GIS may achieve have been moderated, while some of the criticisms of it are being addressed. In particular, it has become clear that GIS has more to offer than simply being a quantitative, positivist, number-crunching operation and, as will be discussed in Chapter 8, it has a potentially significant role to play in qualitative research as well as quantitative. In addition, through *Geographic Information Science* (GISc, see Chapter 1), the role of quantitative GIS-based work as an academic sub-discipline is being increasingly well defined.

While there has been much enthusiasm for including statistical functionality as part of the GIS toolbox among researchers, this has not been mirrored among the GIS software vendors. Many software packages claim to have analytic functionality but this usually refers to the inclusion of overlay, buffering and the other forms of spatial manipulation described in Chapter 4, rather than the more statistical techniques dealt with in this chapter. This has advantages as well as disadvantages. In particular, it encourages researchers to determine for themselves what types of techniques are relevant to their

data, and to devise methods of performing the analysis either within the GIS software or outside it. Effectively this means that at present we are forced to think about what techniques are appropriate to our data rather than simply being spoon-fed with vendor-provided 'solutions' that may or may not be appropriate. While this may lead to cumbersome computing, it is hoped that it should lead to more appropriate analyses.

7.2 What makes spatially referenced data special?

Spatially referenced data have special characteristics that offer both advantages and disadvantages to the researcher wanting to perform statistical analysis. The advantages are basically that they allow us to ask questions such as 'where does this occur?', 'how does this pattern vary across the study area?', 'how does an event at this location affect surrounding locations?', and 'do areas with high rates of one variable also have high rates of another?'. Conventional statistical techniques such as *correlation* and *regression* tend to produce a summary statistic that quantifies the strength of a relationship within a dataset or between two of more sets of variables, for example, a correlation coefficient of 0.8 between *X* and *Y*. This approach is termed *global* or *whole-map analysis* and is undesirable from a GIS perspective because it ignores the impact of space by simply providing a summary of the average relationship over the whole study area. In reality any relationship is likely to vary over space and when performing analysis on spatially referenced data we should attempt to use techniques that allow us to examine not simply what the average relationship is over the whole map, but how the relationship varies across the map. This is termed *local analysis* (Fotheringham 1997). Some examples of how this can be done are given in section 7.3. For now it is sufficient to make the point that by using techniques that explicitly incorporate the spatial component of the data we are able to use the data to develop a more sophisticated understanding. Ignoring the effects of space, as many traditional techniques do, means that we are limiting the knowledge that can be gained from the data.

There are four major disadvantages associated with spatially referenced data. These are: data quality and *error propagation*; *spatial autocorrelation*; the *modifiable areal unit problem* (MAUP); and *ecological fallacy*. Many of the issues regarding data quality with spatially referenced data were discussed in Chapter 4. Here it is sufficient to say that there is always *error*, inaccuracy, and uncertainty in the spatial component of data. When data are combined through overlay or similar operations the error and uncertainty become cumulative, a process known as error propagation. GIS software packages do not handle

this well, as their data model only allows a single concisely defined representation for each spatial feature, so the possible impact of error must always be considered.

Many statistical techniques make the assumption that the observations used in the study are independent pieces of evidence. This is usually referred to as being 'independently random'. If we are studying the locations of the data to try to find a geographical pattern then we are working on the principle that the locations are being influenced by some underlying cause that varies over space. The data are referred to as being spatially autocorrelated, and this invalidates the assumption of independent randomness. Spatial autocorrelation is similar to temporal autocorrelation but is more complicated, as it operates in many directions simultaneously. The degree of spatial autocorrelation in a dataset may be quantified, as is described later in this chapter.

Many socio-economic datasets, for example the census, are published as totals for administrative units. The boundaries of these units are, to all intents and purposes, arbitrary and random, defined as a result of politics and inertia rather than because they say anything meaningful about the population that they are sub-dividing. The question this raises is: if the arrangement of the administrative units were changed, would the results of any analyses based on them also change? Openshaw and Taylor (1979) vividly demonstrated that it could. They compared data on the percentage of the population voting Republican in the 1968 Congressional election with the percentage of the population aged over 60 for 99 counties in Iowa. By aggregating these data to six regions using different arrangements they could produce correlations from -0.99 to 0.99 and almost any result in between. In other words, the results of the analysis were totally dependent on the way in which the data were aggregated. Similar phenomena have also been demonstrated using regression analysis (Fotheringham and Wong 1991), and they occur as a result of two effects combining. The first, the scale effect, simply means that as data are aggregated they become increasingly averaged or smoothed as extreme values are merged with more normal areas. The second is connected to the actual arrangements of the boundaries, which can lead to results being 'gerrymandered' in a similar way to election results. Taken together these effects are referred to as the modifiable areal unit problem (MAUP).

At one level the MAUP is highly worrying, as it means that the results of any analysis using spatially aggregate data are highly suspect and could simply be the result of the administrative units used to analyse the data. Some statistical and GIS-based approaches to the MAUP have been suggested (see Openshaw and Rao 1995; Fotheringham *et al*. 2000), but these are either statistically or geographically complex and are not entirely

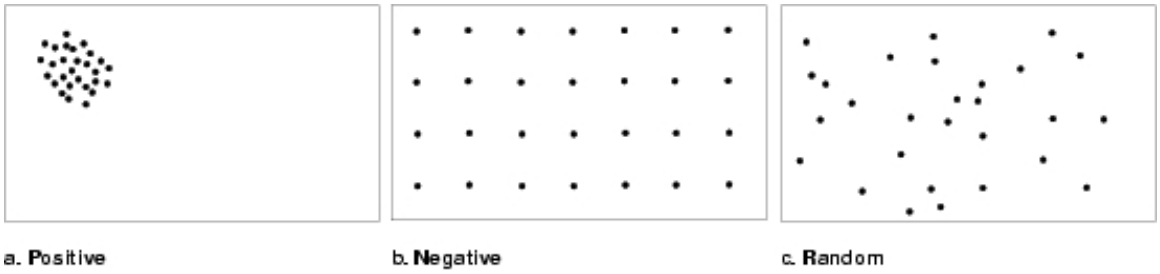
satisfactory. A more pragmatic approach is to avoid additional aggregation by using the data in as close to their raw form as possible while interpreting the results of any analysis bearing in mind the possible impact that modifiable areal units may be having.

Ecological fallacy is closely related to the modifiable areal unit problem. It deals with the fact relationships found in aggregate data may not apply at the household or individual level. For example, if areas with high rates of unemployment also have high crime rates it is a mistake to think that a person is more likely to commit crime if he or she is unemployed. This has been known since the 1950s, yet the increasing potential to analyse spatially aggregate data in GIS means that there is a temptation to ignore it. There have been attempts to find mathematical or statistical solutions to ecological fallacy but, as with the MAUP, these provide computationally and statistically complex ways into the problem without actually resolving it.

7.3 Spatial analysis techniques

Having introduced the advantages and disadvantages associated with the statistical analysis of spatially referenced data, a few types of techniques that can be used are described here. It is not the intention to describe them in great detail as they are well described elsewhere (see the bibliography), but to give a flavour of the types of approaches that can be used.

The type of spatial data often determines what type of spatial analysis technique is most appropriate as different techniques must be used to take advantage of the characteristics of point, line and polygon datasets. It is sometimes sensible to analyse polygon data as points based on their centroids or points as polygons using *Thiessen polygons* (see Chapter 4).



[Types of spatial autocorrelation in point patt]

Point pattern analysis is concerned with attempting to determine whether the distribution of points is random or whether it either clusters (positive spatial autocorrelation) or is evenly distributed (negative spatial autocorrelation) as shown in

Figure 7.1. The easiest way of doing this is termed *quadrat analysis* whereby the study area is sub-divided into regular grid squares and the number of events in each square is counted. This is not particularly satisfactory, as the results tend to be heavily dependent on the size and arrangement of the grid squares. Better techniques focus on the distance between each point and its nearest neighbour or on kernel estimations, which effectively produces a moving average of the density of points at each location on the map. Global summary statistics from either of these approaches will simply say that the data clusters, is evenly distributed or is randomly distributed. It is better to use local techniques that allow summaries such as 'the data cluster in this location but are randomly distributed here'.

Pure point pattern analysis is simply concerned with analysing the spatial component of the data and ignores attribute. This is not possible with polygon data where the arrangement of the polygons will usually be arbitrary, such as with census data based on administrative units. For these it is necessary to devise *proximity measures* that allow us to quantify the influence that each polygon has on its neighbours. These can either be based on the distance between the centroids or based on some measure on the relationship between polygon boundaries. They may also be binary where polygon *i* either has an influence on polygon *j* or it does not, or there may be some degree of quantification on the strength of the relationship. Simple binary proximity measures include whether two polygons share a boundary or whether their centroids lie within a set distance of each other. More complex measures include ratios based on the length of shared boundary between two polygons compared to their total perimeters, or distance decay models whereby the influence one polygon has on another declines as the distance between the two centroids increases.

Once a proximity measure has been calculated, this may be used to analyse the polygon or point data set. A simple form of analysis is to test for spatial autocorrelation using techniques such as *Geary's coefficient* or *Moran's coefficient* that analyse proximity and attribute together. Traditionally these provide global summaries but techniques such as *Geary's* G_i allow us to perform local analysis of these questions (Getis & Ord, 1996). More sophisticated forms of spatial analysis that combine the analysis of spatial and attribute data include *kriging* and *geographically weighted regression* (GWR). Kriging is a sophisticated interpolation technique that attempts to estimate a continuous trend surface from a set of known sample points (Isaaks and Srivastava 1989). GWR explicitly incorporates space into regression analysis. Rather than simply producing a single, global regression equation GWR produces an equation for each point or polygon

in the dataset based on the location's relationship with its neighbours. This can be used to produce maps that show how relationships vary over space (Fotheringham *et al.* 2002).

Analysis of line data is often different from analysis of other forms of data, as it tends to concentrate on flows. This is termed *network analysis*. This can involve problems such as the way in which the shortest route between a set of points on the network is calculated, termed the *travelling salesman problem*. This can be made more sophisticated by classing lines on the network in different ways, for example, different types of road or railways may have different journey times or journey costs associated with them. Networks can also be used in *location-allocation models* that attempt to find the most efficient location on a network. An example of this might be calculating the most efficient location for an industrial complex based on certain assumptions about the rail network. These can then be compared with actual locations.

7.4 Spatial analysis in historical GIS

A variety of authors have made use of spatial analysis techniques to analyse historical data. Bartley and Campbell (1997) wanted to produce a multi-variate land-use classification for medieval England from a single source, the Inquisitions Post Mortem (IPM), which gave a detailed breakdown of landowners' estates on their death. To give as comprehensive coverage of the country as possible they used 6,000 IPMs covering a 50-year period. These were given place names that allowed them to be allocated to points on the map. These points represent large areas with a variety of land-uses, and they wanted to create a more realistic and flexible representation of medieval land-use. To do this they reallocated the point data onto a raster surface, arguing that this offered a more valid representation of continuous data than either points or polygons. This was done by interpolating from the points on a raster surface such that the value of each pixel was calculated using all eligible IPM points within a set distance of the pixel, with nearer IPMs being given more weight than further ones. They then used cluster analysis to allocate each cell to one of six land-use classes. To take into account the uncertainty in their model, the technique they used allowed cells that were hard to classify to be given second choice alternatives. Through this imaginative and sophisticated use of spatial statistics they were able to create a detailed classification of pre-Black Death England that in turn will allow other studies of localities to be interpreted.

Bartley and Campbell use sophisticated handling of spatial data and statistical techniques to create a complex surface from a point layer. Cliff and Haggett (1996) use basic classifications, overlay and exploration through bar charts and basic statistics to

integrate data from a variety of sources to explore the cholera epidemic in London in 1849. They believed that there were two main factors affecting the variation in severity of the epidemic in different parts of London: drainage, and the source of the water supply. To model poor drainage they created a polygon layer that distinguished high and low risk areas based on height above the level of the River Thames. London's water supply came either from reservoirs or wells, or directly from the Thames itself. Areas supplied by water extracted directly from the Thames were believed to carry a higher risk factor than those supplied from wells and reservoirs. Sub-dividing London in this way provides a second polygon layer. *Overlaying* these two layers sub-divides London into four types of area: those at risk as a result of both poor drainage and a polluted water supply, those at risk only as a result of poor drainage, those at risk only as a result of a polluted water supply, and those with neither risk factor. They then calculated the death rate from cholera in each type of area. Even simply graphing the data shows a clear pattern: areas with both risk factors had a death rate that was nearly twice as high as the metropolitan average. Those with defective drainage alone were also above average, those with polluted water were around average, and those with neither factor were well below. Conventional statistical analysis using analysis of variance (ANOVA) confirms this pattern.

Gregory (2000) performs an analysis of census and vital registration data to show how manipulating the spatial and temporal component of the data can increase information about attribute. He does this by working on net migration over a 50-year period from 1881 to 1931. The census provides population figures split into five-year age bands and by sex. The *Registrar General's Decennial Supplements* provides the number of deaths sub-divided by age and by sex. In theory this gives enough information to calculate the net migration rate in ten-year age bands and by sex. Subtracting the number of, for example, women aged 5 to 14 at the start of a decade from the number aged 15 to 24 at its end and then subtracting the number of deaths in this cohort should give the net migration rate. Boundary changes make this calculation highly error prone, as any population change caused by a boundary change will appear to be net migration. By *areal interpolation* of all the data onto a single set of districts, age and sex, specific net migration rates can be calculated at district-level over the long term. This analysis, therefore, takes very basic demographic data and uses it to provide a long time-series of new data that have far more spatial and attribute detail than it was possible to create from manual methods.

Martin (1996) compares the 1981 and 1991 censuses. He interpolated data from the two censuses he was interested in onto a raster grid consisting of square pixels with 200m sides. In this way he was able to compare data between the two censuses with far more spatial detail than Gregory, but was unable to include earlier censuses as they did not provide sufficient spatial detail to allow accurate interpolation.

7.5 Conclusions

Although a significant amount of progress has been made in determining the research agenda for quantitative analysis of statistical data, there is still some way to go before this functionality becomes a standard part of the GIS toolkit, and even further before it becomes a common part of historical analyses. There is a large amount of potential for the statistical analysis of spatially referenced data, but the limitations of the data must be remembered. This means that researchers must select techniques that do not rely on the assumption of independent randomness, and must bear in mind the limitations of their data, in particular: its quality; potential problems caused by error propagation; the possible impact of modifiable areal units; and the potential for incorrect inference due to ecological fallacy.

As the use of GIS among historians becomes increasingly sophisticated, it is to be expected that increasing use will be made of spatial statistical techniques. The fact that these are not well integrated into GIS software as yet provides opportunities for researchers to devise the techniques that are most suitable for their own data, although these must be devised within the limitations of the source data. The impact of the spatial nature of the data must be borne in mind here and, in particular, unnecessary aggregation should be avoided not only through space, but also through attribute and time. Such techniques should not just produce a single summary statistic for the entire study area but should show how the impact of the phenomenon varies across the map. The results of the technique may be map-based rather than using traditional numeric summaries.

Further reading from chapter 7:

References giving in **bold** are key references.

Spatial analysis and GIS

Anselin, L., 1999. Interactive techniques and exploratory spatial data analysis. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, eds. *Geographical Information Systems: principals, techniques, management and applications.* 2nd edition. Chichester: John Wiley, 1999, 239-251

Fischer, M.M., 1999. Spatial analysis: retrospect and prospect. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, eds. *Geographical Information Systems: principals, techniques, management and applications.* 2nd edition. Chichester: John Wiley, 1999, 283-292

Fischer, M.M., Scholten, H.J. and Unwin, D., 1996, eds. *Spatial analytical perspectives on GIS*. London: Taylor & Francis.

Fotheringham, A.S., 1997. Trends in quantitative methods I: stressing the local. *Progress in Human Geography*, 21, 88-96

Fotheringham, A.S., Brunsdon, C. and Charlton, M., 2000. *Quantitative geography: perspectives on spatial data analysis.* London: Sage.

Fotheringham, A.S. and Rogerson, P.A., 1994, eds. *Spatial analysis and GIS*. London: Taylor & Francis.

Longley, P. and Batty, M., 1996, eds. *Spatial Analysis: modelling in a GIS environment*. Cambridge: GeoInformation International, 1996, 41-54

Longley, P.A., Brooks, S.M., McDonnell, R., and MacMillan, B., 1998, eds. *Geocomputation: a primer*. Chichester: John Wiley & Sons.

Openshaw, S., 2000, eds. *GeoComputation*. London: Taylor & Francis.

Openshaw, S. and Alvanides, S., 1999. Applying geocomputation to the analysis of spatial distributions. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, eds. *Geographical Information Systems: principals, techniques, management and applications.* 2nd edition. Chichester: John Wiley, 1999, 267-282

Openshaw, S. and Clarke, G., 1996. Developing spatial analysis functions relevant to GIS environments. In: M.M. Fischer, H.J. Scholten and D. Unwin, eds. *Spatial analytical perspectives on GIS*. London: Taylor & Francis, 1996, 21-38

Critiques of quantitative analysis in GIS

Flowerdew, R., 1998. Reacting to Ground Truth. *Environment and planning A*, 30, 289-301

Openshaw, S., 1991. A view on the GIS crisis in geography or using GIS to put Humpty-Dumpty back together again. *Environment and planning A*, 23, 621-628

Pickles, J.J., 1995, ed. *Ground Truth: The Social Implications of Geographic Information Systems* New York: Guildford Press.

Pickles, J.J., 1999. Arguments, debates and dialogues: the GIS-social theory debate and the concern for alternatives. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, eds. *Geographical Information Systems: principals, techniques, management and applications.* 2nd edition. Chichester: John Wiley, 1999, 49-60

Taylor, P.J., 1990. Editorial comment: GIS. *Political Geography Quarterly*, 9, 211-212

Data quality and error propagation

See Chapter 4.

The modifiable areal unit problem

Fotheringham, A.S. and Wong, D., 1991. The modifiable areal unit problem in multi-variant statistical analysis. *Environment and Planning A*, 23, 1025-1044

Green, M. and Flowerdew, R., 1996, New evidence on the modifiable areal unit problem. In: P. Longley and M. Batty, eds. *Spatial Analysis: Modelling in a GIS Environment*. Cambridge: GeoInformation International, 1996, 41-54

Openshaw, S., 1984. *The Modifiable Areal Unit Problem*. Concepts and Techniques in Modern Geography, 38. Norwich, Geobooks.

Openshaw, S. and Rao, L., 1995 Algorithms for re-aggregating 1991 Census geography. *Environment and Planning A*, 27, 425-446

Openshaw, S. and Taylor, P.J., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: N. Wrigley, ed. *Statistical Applications in the Spatial Sciences*. London: Pion, 1979, 127-144

Ecological fallacy

Jones, K. and Duncan, C., 1996. People and places: the multilevel model as a general framework for the quantitative analysis of geographical data. In: P. Longley and M. Batty, eds. *Spatial analysis: modelling in a GIS environment.* Cambridge: GeoInformation International, 1996, 79-104

Robinson, W.S., 1950. Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, 351-357

Robson, B.T., 1969. *Urban Analysis*. Cambridge: Cambridge University Press.

Wrigley, N., Holt, T., Steel, D. and Tranmer, M. 1996. Analysing, modelling, and resolving the ecological fallacy. In: P. Longley and M. Batty, eds. *Spatial analysis: modelling in a GIS environment.* Cambridge: GeoInformation International, 1996, 25-40

Spatial statistics

Bailey, T.C. and Gatrell, A.C., 1995. *Interactive spatial data analysis.* Harlow: Longman.

Cliff, A.D. and Ord, J.K., 1981. *Spatial processes: models and applications*. London: Pion.

Fotheringham, A.S., Brunsdon, C. and Charlton, M., 2002. *Geographically Weighted Regression: the analysis of spatially varying relationships*. Chichester: John Wiley.

Getis, A., 1999. Spatial statistics. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, eds. *Geographical Information Systems: principals, techniques, management and applications.* 2nd edition. Chichester: John Wiley, 1999, 239-251

Getis, A. and Ord, J.K., 1999. Local spatial statistics: an overview. In: P.A. Longley and M. Batty, eds. *Spatial analysis: modelling in a GIS environment*. Cambridge: GeoInformation International, 1996, 261-277

Goodchild, M.F., 1987. *Introduction to spatial autocorrelation*. Concepts and Techniques in Modern Geography, 47. Norwich: GeoAbstracts.

Isaaks, E.H. and Srivastava, R.M., (1989) *An introduction to applied geostatistics*. Oxford: Oxford University Press.

Odland, J., 1988. *Spatial autocorrelation*. Sage: London.

Web links

The Spatial Analysis Laboratory, Department of Agriculture and Consumer Economics: http://sal.agecon.uiuc.edu/default.php

The Center for Advanced Spatial Analysis, UCL, London: http://www.casa.ucl.ac.uk/

The GeoDa Exploratory Spatial Data Analysis package: http://sal.agecon.uiuc.edu/geoda_main.php

The CommonGIS package: http://www.ais.fhg.de/SPADE/products/index_eng.html or http://www.esds.ac.uk/International/support/commonGIS.asp

Case studies

Bartley, K. and Campbell, B.M.S., 1997 Inquisitiones Post Mortem, GIS, and the creation of a land-use map of medieval England, *Transactions in GIS*, 2, 333-346

Cliff, A.D. and Haggett, P., 1996. The impact of GIS on epidemiological mapping and modelling. In: P. Longley and M. Batty, eds. *Spatial analysis: modelling in a GIS environment*. GeoInformation International: Cambridge, 1996, 321-344

Diamond, E. and Bodenhamer, D.J., 2001, Race and decline of mainline Protestantism in American cities: a GIS analysis of Indianapolis in the 1950s. *History and Computing*, 13, 25-44

Gregory, I.N., 2000. Longitudinal analysis of age- and gender- specific migration patterns in England and Wales: a GIS-based approach. *Social science history*, 24, 471-503

Martin, D., 1996. Depicting changing distributions through surface estimations. In: P. Longley and M. Batty, eds. *Spatial analysis: modelling in a GIS environment*. GeoInformation International: Cambridge, 1996, 105-122

Pearson, A. and Collier, P., 1998. The integration and analysis of historical and environmental data using a Geographical Information System: landownership and agricultural productivity in Pembrokeshire c. 1850. *Agricultural history review*, 46, 162-176. See also Chapter 2.