16
Spatial statistics

A GETIS

The field of spatial statistics is based on the assumption that nearby georeferenced units are associated in some way. More and more, the GIS community needs to draw upon the work of the spatial statistician to help find meaning in spatial data. The precursors of current spatial statistical researchers include those who sought to describe areal distributions, the nature of spatial interactions, and the complexities of spatial correlation. The spatial statistical methods in current use, and upon which research is continuing, include: spatial association, pattern analysis, scale and zoning, geostatistics, classification, spatial sampling, and spatial econometrics. In a time-space setting, scale, spatial weights, and spatial boundaries are especially difficult problem areas for further research. Those working in GIS welcome comprehensive packages of spatial statistical methods integrated into their software.

1 INTRODUCTION

It is common for statisticians to confine their attention to *data description*, including exploratory analysis, and *induction*, the development of generalisations about a defined population on the basis of a sample drawn from that population. Map-oriented researchers have long been interested in data description and induction, usually searching the statistics literature for ideas on how to extract as much information as possible from georeferenced data. The search is often directed toward biometry, psychometry, geology, econometrics, and statistics (see also Fischer, Chapter 19). A relatively small area of original research that cuts across these fields can be attributed to the work of spatial statisticians, who can be distinguished by their attention to georeferenced or spatial data. In recent years, some of this work has been spurred by the development of GIS. In this chapter, spatial data analysis with particular emphasis on the uses and applications of spatial statistics in a GIS environment is discussed.

Spatial statistics can be considered a distinct area of research. Traditional statistical theory bases its models on assumed independent observations. Although common sense tells us that in most real-world situations independence among observations on a single variable is more the exception than the rule, independence is still a suitable benchmark from which to identify statistically significant non-independent phenomena. The field of spatial statistics is based on the non-independence of observations; that is, the research is based on the assumption that nearby units are in some way associated (Tobler 1979). Sometimes this association is because of a spatial spillover effect, such as the obvious economic relationship between city and suburb. Sometimes the association is a distance decline effect; that is, as distance increases from a particular observation, the degree of association between observations lessens. An example is the influence of an earthquake; its effect declines with distance from the epicentre.

Statistics, in general, and spatial statistics with its emphasis on location, are the glue that holds much of our research efforts together. In the search for a high degree of certainty, we look to statistics. As the GIS community matures, it will draw more and more

upon the work of the statistician to help to find meaning in spatial data and in the development of GIS functionality.

The approach in this chapter is to trace briefly and selectively, in section 2, the pre-GIS contributions of map-oriented researchers to spatial statistics. In section 3, brief synopses of statistical analytical devices that spatial analysts use are provided. The distinction is made between the work of the spatial statisticians (those statisticians, biologists, econometricians, atmospheric/oceanic scientists, and geographers who seek to identify the statistical parameters of spatially distributed variables) and the geo-statisticians (those statisticians, geologists, and mining engineers who analyse their data at a number of scales in terms of spatial trend and correlation). Each subsection contains examples of, and key references to, current research. Finally, in section 4, some of the problems and challenges that face spatial researchers are outlined, with a reflection on the nature of statistical work in a GIS environment.

2 PRE-GIS USES OF SPATIAL STATISTICS

Geography has a long history of development of clever cartographic devices that allow for particularly insightful views of spatial data. From Lalanne's (1863) hexagonal railway patterns to the map transformations of Tobler (1963), the pre-GIS literature is filled with interesting ideas designed to enable spatial data to 'speak for themselves'. The desire to make maps a useful part of analysis led pre-computer geographical writers to try to find ways to depict spatial distributions of data in imaginative ways. It was just a short step from interesting depictions on maps to statistical tests on their significance relative to some supposition about the meaning of the maps. Now we have powerful computers and extensive software that guide us toward the production of new and unusual maps. Concomitantly, new statistical devices have been developed, albeit at a slower rate of growth, to answer questions about mapped patterns. Knowing that the spatial perspective is an important aspect of knowledge, analysts seek better ways to depict data on maps and to test hypotheses based on some expected pattern form or structure. Four themes can be considered antecedents of what has become the modern statistical analysis of spatial data.

2.1 Statistical analysis of areal distributions

Although the roots of his work go back to the nineteenth century, Neft (1966), working under the direction of Warntz, was the first to produce a comprehensive, mathematically consistent system for describing areal distributions. Drawing on the work of Carey (1858), Mendeleev (1906), the location theorists – Thünen (1826), Weber (1909), Christaller (1935), and Lösch (1954) – and the ideas of the social physicists – particularly those of Stewart (1950) and Warntz and Neft (1960) – Neft described the statistical moments of areal distributions. For point distributions, he produced statistical measures of skewness and kurtosis of average position (various centroids), spatial dispersion, and surfaces. In addition, he addressed one of the challenging tasks still very much on the agenda of current spatial statistical researchers: producing valid measures of statistical association of spatial variables.

2.2 Spatial interaction

There is no more important topic for the spatial analyst who deals with human issues than the study of the interaction of activities in one place with those in another. Research in this area has a long and distinguished history, dating back to Carey (1858), Ravenstein (1885), Reilly (1929), Zipf (1949), and Stewart and Warntz (1958). The famous Newtonian formula, $m_1 m_2 / d^2$, where m_1 and m_2 are measures of mass at sites 1 and 2, and d is the distance separating the masses at those sites, was the foundation stone. Modified by spatial theorists, this physical law has been used to great advantage to study and to predict a wide variety of human spatial interactions, such as transportation movements, the spread of information, and the potential for economic growth. Modern expositions of the theory and statistical estimation procedures make up a significant portion of modern transportation and marketing literature (Birkin et al, Chapter 51; Gatrell and Senior, Chapter 66). The work of Wilson (1967) must be singled out as a relatively recent attempt to derive practical spatial interaction theory. Rather than depend on physical science analogies, Wilson devised probabilistic laws that described possible human movement. Nowadays an important use of GIS is to allow for the manipulation of data so that parameters that describe movement can be calibrated and evaluated.

2.3 Spatial correlation

Before the 1960s, only a modest literature had developed in geography on perhaps the most challenging spatial question: in an unbiased way, how is one to account for the correlation in spatially distributed variables? The fundamental ideas concerning the measurement of, and testing for, spatial autocorrelation were spawned in geography by Robinson (1956), and Thomas (1960) saw the difficulties in dealing with dependent unequally sized units. Through their work and that of others, the modifiable areal unit problem was addressed and spatial residuals from regression were evaluated. It was during this period that the statisticians Moran (1948) and Geary (1954) developed their measures of spatial autocorrelation. Building on the work of Moran (1948) and Krishna Iyer (1949), Dacey (1965) addressed the issue of the possible association among contiguous spatial units. These join count statistics led to the work of Cliff and Ord, whose monograph 'Spatial Autocorrelation' (1973) opened the door to a new era in spatial analysis. In section 3, we outline the link between the Cliff–Ord work and modern approaches to spatial statistical analysis.

2.4 Hypotheses about settlement patterns

Much of the excitement in the University of Washington's Department of Geography during the late 1950s and early 1960s centred on understanding and testing the theories of the economic geographer, Walter Christaller, and the economist, August Lösch. From the standpoint of spatial statistics, of note is the work of Dacey (1963), who by taking the lead from the plant ecologists such as Clark and Evans (1954), tested various statistical distributional theories using sets of georeferenced data that represented the location of towns in a settlement system. From this work, a point pattern 'industry' developed that featured the work of King (1962), Getis (1964), Harvey (1966), Clark (1969), and Rogers (1969).

3 SPATIAL STATISTICS IN CURRENT USE

The types of statistical methods popular today are a function of both the nature of the problems being studied and the availability of computers. Seven areas of research are listed that are particularly favoured. Each is described in terms of the kinds of problems being solved, their general formulation (if not discussed in detail elsewhere in this book), and their usefulness to the GIS community of analysts. In addition, current research themes are noted together with key references. Such areas of inquiry as spatial neural nets, spatial fuzzy sets, and simulated annealing are just now being developed and are discussed by Fischer (Chapter 19).

3.1 Spatial association

The Cliff–Ord monograph enabled researchers to assess statistically the degree of spatial dependence in their data, and, in so doing, to search for additional or more appropriate variables, and to avoid many of the pitfalls that arise from autocorrelated data. Many GIS contain the Cliff–Ord routines that allow for the calculation of spatial autocorrelation. Much of present-day interest in spatial analysis derives directly from the 1973 Cliff-Ord monograph and the authors' subsequent (1981) more complete discussion. These shed light on the problem of model mis-specification owing to autocorrelation and demonstrated statistically how one can test residuals of a regression analysis for spatial randomness. They explicated the nature of the spatial weight matrix and provided step-by-step procedures for applying statistical tests on Moran's I and Geary's c, the two major autocorrelation statistics.

Finding the degree of spatial association (autocorrelation) among data representing related locations is fundamental to the statistical analysis of dependence and heterogeneity in spatial patterns. Like Pearson's product–moment correlation coefficient, Moran's statistic is based on the covariance among designated associated locations, while Geary's takes into account numerical differences between associated locations. The tests are particularly useful on the mapped residuals of an ordinary least squares regression analysis. Statistically significant spatial autocorrelation implies that the regression model is not properly specified and that one or more new variables should be entered into the regression model.

Mantel (1967), Hubert (1979), and Getis (1991) have shown that statistics of this nature are special cases of a general formulation, gamma, that is defined by a matrix representing possible locational associations (the spatial weights matrix) among all points, multiplied by a matrix representing some specified non-spatial association among the points. The

non-spatial association may be an economic, social, or other relationship. When the elements of these matrices are similar, high positive autocorrelation ensues. Gamma describes spatial association based on covariances (Moran's statistic, *I*), or subtraction (Geary's statistic, *c*), or addition (the *G* statistic of Getis and Ord 1992). These statistics are global insofar as all measurements between locations are taken into account simultaneously. Aspinall (Chapter 69) provides examples in the realm of landscape conservation.

When the spatial weights matrix is a column vector, gamma becomes local; that is, association is sought between a single point and all other points (I_i, c_i, G_i). Research on local statistics has been especially active recently because they lend themselves to kernel-type analyses in a GIS where datasets are large (Anselin 1995; Getis and Ord 1992; Ord and Getis 1995). Local statistics have been used to classify remotely-sensed data (Getis 1994), and to show associations between neighbourhoods' crime rates (Anselin 1993) and countries' conflict propensities (O'Loughlin and Anselin 1991).

Some current research themes in this area are:

- the identification of spatial spillover or nuisance autocorrelation (Anselin and Griffith 1988; Anselin 1990a; Anselin and Rey 1991);
- characteristics of the structure of spatial weight matrices (Griffith 1988; Anselin 1986; Boots and Kanaroglou 1988);
- heterogeneity issues in local measurements of spatial association (Bao and Henry 1996);
- determining the exact distribution of spatial autocorrelation statistics (Tiefelsdorf and Boots 1994);
- alternatives to the Cliff–Ord approach (Kelejian and Robinson 1995);
- multivariate spatial association (Wartenberg 1985).

3.2 Pattern analysis

Popular in the 1960s was point pattern analysis based on the spatial homogeneous Poisson process (see also Fischer, Chapter 19). It was common to find a researcher working at a light table making measurements from numbered points to the first nearest neighbour of each point. Now, with the use of digitised georeferenced data, we are easily able to take measurements from all points to all other points. In addition, measurements of line segments, distances between line intersections, areas, and characteristics of areas such as perimeter length, neighbouring areas, and so on, are basic within most GIS.

Pattern analysis in the spatial sciences grew out of an hypothesis-testing tradition, not out of the extensive pattern recognition literature. Nearest neighbour work continues today, but the work of Clark and Evans (1954) has now been modified for the sake of unbiasedness to take into account the length of the perimeters of study areas (Donnelly 1978) and the distance to study area boundaries (refined nearest neighbour analysis: Diggle 1979; Boots and Getis 1988).

In recent years, point pattern analysis has regained its vigour as an area of study as a result of the ability of computers to handle large numbers of objects. Statistical approaches are usually based on hypotheses of complete spatial randomness (CSR), that is, the theoretical pattern is assumed to be representative of: (a) objects that are located independently of each other; and (b) a study area where each location has an equal chance of receiving an object. The pattern analyst tests hypotheses about the spatial characteristics of point, line, or area patterns. These geometric forms represent everything from the location of individuals suffering from an infectious disease to the shape of hardened basalt flows (Boots and Getis 1988).

Related to pattern analysis is the continuing interest that ecologists have in studying plant and animal distributions. Surprisingly, only in recent years have plant ecologists become aware that because of dependence among nearby observations, a particular pattern of plants may not represent a suitable sample for model testing (Franklin 1995). A set of key references in this area may be found in Potvin and Travis (1993).

Perhaps the most important developments in recent years are the applications of *K*-function analysis to the study of point patterns, and the use of Voronoi polygons to study spatial tessellations (Boots, Chapter 36; Okabe et al 1992). In addition, fractals study is a promising area for pattern analysis (Batty and Longley 1994).

3.2.1 The idea behind the use of *K*-function analysis

The *K*-function is the ratio of the sum of all pairs of points within a pre-specified distance, *d*, of all points to the sum of all pairs of points regardless of distance. The function is adjusted to take into account distances that are closer to the boundary of the study area than to *d*. The original *K*-function by

Ripley (1977) was modified by Besag (1977) to take into account the need to stabilise variance, and Getis (1984) generalised the formula to include the weighting of points, such that the sum of pairs of points became the sum of the multiples of the weights associated with each member of a pair of points. Diggle (1983) has done much to exploit this formulation to show many new features of patterns. For example, not only can one easily show the difference between an existing pattern and a random pattern but one can also develop theoretical expectations for other than random patterns. In addition, patterns divided into different point types (marked patterns) can be studied easily. For testing purposes, an envelope of possible outcomes under the hypothesis of say, randomness, is usually constructed by means of a Monte Carlo simulation. Studies of the spatial distribution of vegetation dominate the empirical literature of *K*-function analysis (Diggle 1983), but the method has been used for the study of human population distribution (Getis 1983) and disease distribution (Morrison et al 1996). Recently, Gatrell et al (1996) showed that the *K*-function can be used as an indicator of time-space clustering; that is, one simultaneously finds pairs of points separated by designated units of time and distances in space. This approach is particularly useful for identifying disease clustering over time.

3.2.2 Successful applications to spatial phenomena

Some themes of current interest in pattern analysis are:

- the development and testing of time-space pattern models (Griffith 1996; Gatrell et al 1996; Jacquez 1995);
- search for pockets of extreme values in large data-sets (Ord and Getis 1995; Haslett et al 1991);
- development of pattern models based on differences, absolute differences, and similarities between nearby observations (Getis and Ord 1996).

3.3 Scale and zoning (the modifiable areal unit problem)

The problem of scale effects was made particularly clear by the results of Openshaw and Taylor's (1979) study of voting behaviour in Iowa. They showed that the level of spatial aggregation and arrangement of spatial units (zoning) has a marked effect on the correlation of variables. Fotheringham and Wong (1991) identified the extent of the spatial bias in a multivariate regression analysis and Fotheringham et al (1995) carried out similar research in a *p*-median problem context. The most comprehensive treatment to date is that of Arbia (1989), who identified the relationship between levels of autocorrelation and spatial unit aggregation. A recent study by Holt et al (1996) shows the scale problem to be an area selection problem. Some themes being pursued include:

- spatial aggregation biases (Okabe and Tagashira 1996; Tobler 1989);
- the relationship of spatial autocorrelation to scale differences (Arbia et al 1996);
- the effect of different zoning on results of various types of analyses (Openshaw 1996; Green and Flowerdew 1996);
- identifying scale effects by use of principal axis factor analysis (Hunt and Boots 1996);
- scale effects on parameters of spatial models (Amrhein and Reynolds 1996; Wrigley et al 1996).

This theme is developed by Openshaw and Alvanides (Chapter 18).

3.4 Geostatistics

The *variogram* (or semivariogram) (Cressie 1991) plays a useful role as the function that describes spatial dependence for a regional (georeferenced) variable. The term 'intrinsic stationarity' is used to indicate the natural increase in variance between observations of a regional variable as distance increases from each observation. The semivariance – a measure of the variance as distance increases from all points or areas (blocks) – eventually reaches a value equal to the variance for the entire array of data locations, regardless of distance. Clearly, at zero distance from a point, the semivariance is also zero, but the semivariance increases until, at a distance called the range and a semivariance value called the sill, it is equal to the variance. The function describing the semivariance is usually spherical, exponential, or Gaussian.

The variogram is essential for *Kriging*, which is a technique for estimating the value of a regional variable from adjacent values while considering the dependence expressed in the variogram. There are many kinds of kriging, each designed to give the highest possible confidence to the estimation of a variable at non-data locations. If there is no bias in the variogram, and all required assumptions are met, the kriged values, as opposed to trend surface, triangulated irregular network (TIN), or other estimation devices, will be optimal.

A large amount of literature has developed in geostatistics. The definitive text by Cressie (1991) details many instances where the geostatistical approach has proved helpful. These include studies of soil-water tension, wheat yields, acid deposition, and sudden infant death syndrome. Aspinall (Chapter 69) and Wilson (Chapter 70) discuss applications in landscape conservation and agriculture, respectively. The variogram has now been introduced into several GIS, and programs that can be interfaced with GIS are available to help construct variograms and to apply the kriging process (GS+ 1995; GEO-EAS 1988; S+SpatialStats 1996). The geographical literature on practical applications is building rapidly. Of particular interest is the work of Oliver and Webster (1990).

3.5 Classification

Interest in this problem rises or falls depending on the challenges presented by the subject matter and the type of data used. As part of any image analysis of remotely-sensed data, grouping algorithms are needed. Supervised and unsupervised classification schemes have been developed that allow for pixel values to be identified with a particular category of, say, land cover. Spectral, regression tree, autocorrelation, neural network, and fuzzy logic schemes have been adapted to deal with the problems of aggregation. Themes being pursued include:

- evaluation of neural pattern classifiers (Fischer, Chapter 19; Fischer et al 1997);
- the degree of supervision needed in finding statistically significant groupings (Gong and Howarth 1990);
- effects of resolution and sensitivity on various classification schemes (Marceau et al 1994);
- incorporation of non-remotely-sensed data in decision tree algorithms (Michaelsen et al 1996);
- application of classification routines to the results of spectral-unmixing (Mertes et al 1995);
- classification routines applied to hyperspectral and high spatial resolution data (Barnsley, Chapter 32; Barnsley and Barr 1996).

3.6 Sampling issues

Just as the jury selection process affects the outcome of a trial, so does the sampling scheme influence research results. Spatial sampling is a particularly difficult problem to deal with, since the idea (unlike many jury selection processes) is to select an unbiased sample, but finding independent observations is impossible. Spatial sampling requires that the researcher recognise the degree of dependence in the data. Very often, the surfaces from which samples are taken are complex and oddly shaped, presenting difficult problems to overcome in the statistical analysis. For many years, considerable effort was given to making sense from small samples. The challenge now is to make sense of large datasets (Fischer, Chapter 19; Openshaw and Alvanides, Chapter 18), and one means of so doing is to sample from them. Research in this area includes:

- line transects and variable circular plots, including kernel sampling (Quang 1992);
- network sampling (Faulkenberry and Garoui 1991);
- cluster and systematic sampling (Thompson 1992);
- spatial sample size (Ripley 1981; Goodchild and Gopal 1989; Haining 1990);
- strip and stratified adaptive cluster sampling (Thompson 1992);
- heterogeneous data sampling (Griffith et al 1994).

3.7 Spatial econometrics

The fundamental work in this area can be traced to Paelinck (1967; see also Paelinck and Klaassen 1979). Anselin has made spatial econometrics accessible to a wide audience with his text (1988) and software (Chapter 17; 1992). In addition, texts by Haining (1990), Griffith (1988), and Upton and Fingleton (1985) have helped to widen the appeal of these methods in geography. As Anselin says, the approach is 'model driven'; that is, the focus is on regression parameter estimation, model specification, and testing when spatial effects are present. Regression models constitute the leading approach for the study of economic and social phenomena. The assumptions required for the basic linear regression model, however, do not satisfy the needs of spatial regression models, which must take into account spatial dependence and/or spatial heterogeneity. Spatial dependence occurs when there is a relationship between observations of one or more variables at one point in space with those at another point in space, while spatial heterogeneity results from data that are not homogeneous – for example, population by areas which vary considerably by size and shape.

A number of spatial autoregressive models have been developed that include one or more spatial weight matrices that describe the many spatial

associations in the data. The models include either a single general stochastic autocorrelation parameter, a series of autocorrelation parameters, one for each independent variable conditioned by spatial effects (dependency or heterogeneity), an error term autocorrelation parameter, or some combination of these. Parameter estimation procedures can be complex. The usual approach is to use diagnostic statistics to test for dependence and/or heteroscedasticity among the spatially weighted variables or the error term. Fortunately, SpaceStat, designed for the exploration and testing of spatial autoregressive models, is sufficiently user friendly to allow for the development of final autoregressive models (see Anselin, Chapter 17, for a general discussion). In addition, the package has been linked explicitly to several GIS, including ArcView (1995) and Idrisi (Eastman 1993).

Several other approaches have been taken to specify the influence of spatial effects in a regression model environment. Casetti's (1972) expansion method is designed to increase the number of variables in a regression model to take into account secondary, but influential, spatial variables, such as the x, y coordinates of georeferenced variables. This approach uses the parameters of the expansion variables as the indicators of the spatial effects.

In another development, Getis (1990, 1995) suggests transforming the spatially autocorrelated model into one without spatial autocorrelation embedded within it. By filtering out the spatial autocorrelation, the ordinary least squares model can be estimated and evaluated using R^2. By use of the Getis–Ord statistics mentioned earlier, variables are transformed to become relatively free of dependency effects. The filtered spatial components are re-entered into the regression equation as separate spatial variables.

The list of recent research themes, many of which can be found in the volume edited by Anselin and Florax (1995), can be divided into two parts: spatial modelling and estimation. The spatial modelling themes are:

- robust approaches to testing spatial models (Anselin 1990b);
- mis-specification effects in spatial models (Florax and Rey 1995; Hepple 1996);
- data problems in spatial econometric modelling (Haining 1995);
- the general linear model and spatial autoregressive models (Griffith 1995);
- multiprocess mixture (space-time) models (LeSage 1995);
- adaptive filtering and dependence filtering for spatial models (Foster and Gorr 1986; Getis 1995).

Parameter estimation is a subject central to model development. For models having spatial parameters or variables, a number of issues have arisen. The robustness, consistency, and reliability of parameters is a function of underlying theoretical distributions. The assumption of asymptotic normality has been called into question in some cases, and in others, sample sizes must be large before normality assumptions can be invoked. Bayesian approaches have been introduced in order to bring more information to bear on parameter estimation. Maximum likelihood procedures are fundamental to spatial model estimation, but data screening and filtering have been suggested as ways to simplify estimation. Current research includes:

- estimation of regression parameters in spatially constructed regression equations (Florax and Folmer 1992; Kelejian and Robinson 1993);
- estimating space-time probit models (McMillen 1992);
- estimating logit models with spatial dependence (Dubin 1995);
- spatial parametric instability (Casetti and Poon 1995);
- small sample properties of tests for spatial dependence (Anselin and Florax 1995a).

4 PROBLEMS, CHALLENGES, AND FUTURE DIRECTIONS

At the heart of spatial science are the statistical and mathematical techniques that allow for confirmatory statements to be made about the relationship between variables in a spatial setting. The thrust in recent years has been to develop more and better ways to describe data (see Anselin, Chapter 17). The exploratory data analysis movement has given researchers a bevy of fast ways to view data. Much of this work has been created as a response to the large and detailed datasets that are becoming available. At the same time, relatively few new methods have been offered to allow for the confirmation of hypotheses, which is well behind in the race for new understanding of spatial phenomena (Anselin and Getis 1992).

A number of barriers hamper the modellers and others seeking verification of their suppositions.

Many of these obstacles derive from flawed data. Problems include poor data quality, inadequate data coverage, incompatible datasets, inappropriate data, and the inability to handle large datasets (see the contributions to Section 1(b) of this book). While the main goal of spatial statistical analysis is to assist in data interpretion, it cannot improve on flawed data (Goodchild 1992). In addition, at least three further obstacles stand in the way of confirmatory analysis; these are described below.

4.1 Scale

Although much has been written about the nature of the scale problem, there are few useful suggestions for dealing with it. Perhaps the most often suggested solution is to attempt to solve the scale problem at a number of scales in the hope that a certain robustness to the process will allow results to be generalised to a number of spatial scales. In order to understand this issue, however, more research is needed on the nature of distributional parameters when data are aggregated. A wider review of scale issues is provided by Weibel and Dutton (Chapter 10).

4.2 Spatial weights effects

Identifying and describing spatial association is the goal of much research. The intrinsic stationarity of the variogram represents an empirically derived theory of spatial effects. In essence, it is the spatial weights matrix of the autoregressive models and the spatial association statistics. Thus, the spatial weights matrix is the manifestation of our understanding of spatial association. Too often, the contiguity spatial weights matrix is chosen simply because no further understanding of distance, or interaction, or association is assumed. The spatial association statistics, such as I, c, and G, could be put to good use as indicators of the appropriateness of particular spatial weights matrices. To understand better issues such as dependence and spillover, generalisation between global and local scales, and data heterogeneity and homogeneity, appropriate mathematical constructs – such as eigenvectors – must be related to the form and structure of our data. In addition, types of variable – economic, social, physical – must be related to the geometry of their spatial representation.

4.3 Boundary effects

Related to the above two problem areas is the issue of boundary effects. In spatial studies, the delineation of boundaries bear heavily upon results. Although many truncated probability distributions have been derived, they have not been used to good effect to account for spatial boundary conditions. A number of statistical procedures, such as refined nearest neighbour analysis and K-function analysis, take into consideration the effect of boundaries, but spatial scientists have yet to consider boundary effects systematically. Stochastic approaches to modelling take into account impervious, reflecting, and other types of boundary conditions, but this work has not yet entered the mainstream of spatial science.

The problems discussed above describe at most half of the challenge. Increasingly, the temporal dimension is becoming a part of formerly static models of spatial human and physical processes. Deeper understanding usually comes from the study of differences in space as well as time. Bringing these two fundamental dimensions into a modelling framework where parameters can be estimated is a considerable challenge.

5 GIS AND SPATIAL STATISTICS

With regard to GIS, this volume makes clear how well suited these systems are for the exploration and manipulation of spatial data. Initially, the contributions to GIS were in the form of commands that allowed for the rectification of inconsistencies between a number of coverages (spatial variables) of the same geographical region. Much was made of the fundamental data model, that is, raster or vector. The main purpose was to link georeferenced datasets that are either in pixel or polygonal spatial form so that various combinations of variables could be mapped. As sophistication increased, functions were developed that allowed for new data to be derived from the various coverages, and for back and forth movement between data models.

For the most part, however, testing of hypotheses using statistical methodology was left for non-GIS statistical packages. It was quite enough to develop the technology and the functions that allow for data manipulation. Naturally, exploratory analytical functions were developed. A great deal of progress has been made in this regard, mainly from the standpoint of graphical summaries of data distributions together with simple summary measures like means and standard deviations. The need for more sophisticated analyses, voiced by many academics, is now getting a hearing in GIS literature

(Longley and Batty 1996a). Analysts are now beginning to take advantage of the data processing and data manipulation qualities of GIS to help create and test models using statistical methodology. A number of packages have been developed that enable researchers to interface with GIS-formatted datasets. Some of these are: S+Gislink links S+SpatialStats with ARC/INFO (1996), SpaceStat links with ArcView (Anselin 1997), Bailey and Gatrell's *Interactive Spatial Data Analysis* (1995), and Regard (Haslett et al 1990). Other packages, such as GS+, can be adapted to GIS requirements.

The effect of the new technology on spatial statistical analysis has led to a broadening of the process of hypothesis testing (Getis 1993). Heretofore, the hypothesis-testing process was straightforward, with little opportunity to recast hypotheses while in the testing process. Now, the approach is much more flexible. Note that in Figure 1 a step has been added to the traditional approach of hypothesis guided inquiry, and most steps have been expanded to include more opportunities to assess data from different vantage points. The added step, *data manipulation*, presents researchers with opportunities to use larger samples, view data over a series of map scales, and generally be in a stronger position to carry out statistical tests by means of simulations, sensitivity analyses, and bootstrap methods. Note, however, that each of these aproaches broadly adheres to what Goodchild and Longley (Chapter 40) term the 'linear project design'.

The flurry of activity in recent years has led to the publication of a number of edited volumes and special journal issues that provide examples of the various themes that are designed to wed spatial statistical analysis with GIS. Included among these are books edited by Fotheringham and Rogerson (1994), Frank and Campari (1993), Fischer and Nijkamp (1993), Fischer et al (1996), Longley and Batty (1996b), and Fischer and Getis (1997). Given the attention paid to this subject, in the next years we might expect a full-fledged statistical package, in the SPSS sense, integrated with the most comprehensive GIS.

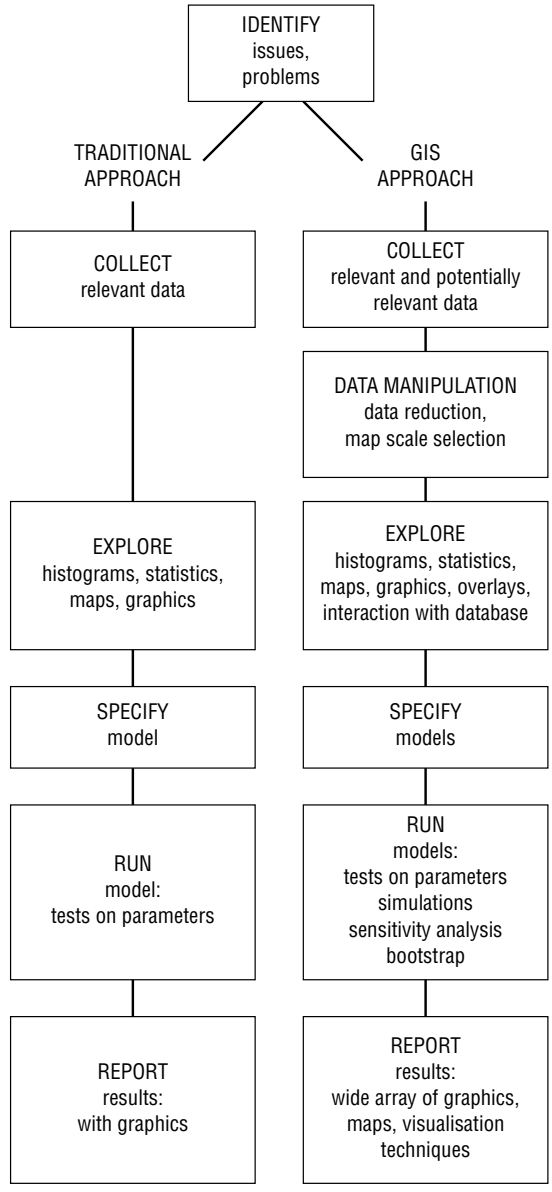

Fig 1. Traditional and GIS approaches to spatial statistics analysis.

References

Amrhein C G, Reynolds H 1996 Using spatial statistics to assess aggregation effects. *Geographical Systems* 3: 143–58

Anselin L 1986 Non-nested tests on the weight structure in spatial autoregressive models: some Monte Carlo results. *Journal of Regional Science* 26: 267–84

Anselin L 1988 *Spatial econometrics: methods and models*. Dordrecht, Kluwer

Anselin L 1990a What is special about spatial data? Alternative perspectives on spatial data analysis. In Griffith D A (ed.) *Statistics, past, present and future*. Ann Arbor, Institute of Mathematical Geography: 63–77.

Anselin L 1990b Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics* 20: 141–63

Anselin L 1992 *SpaceStat: a program for the analysis of spatial data*. NCGIA, Santa Barbara, University of California

Anselin L 1993 *Exploratory spatial data analysis and geographic information sysytems*. West Virginia University, Regional Research Institute, Research Paper 9329

Anselin L 1995 Local indicators of spatial association – LISA. *Geographical Analysis* 27: 93–115

Anselin L, Bao S 1997 Exploratory spatial data analysis linking SpaceStat and ArcView. In Fischer M M, Getis A (eds) *Recent developments in spatial analysis: spatial statistics, behavioural modelling, and computational intelligence*. Berlin, Springer

Anselin L, Florax R J G M (eds) 1995 *New directions in spatial econometrics*. Berlin, Springer

Anselin L, Getis A 1992 Spatial statistical analysis and geographic information systems. *Annals of Regional Science* 26: 19–33

Anselin L, Griffith D A 1988 Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association* 65: 11–34

Anselin L, Rey S 1991 Properties of tests for spatial dependence in linear regression models. *Geographical Analysis* 23: 112–31

Arbia G 1989 *Spatial data configuration in statistical analysis of regional economic and related problems*. Dordrecht, Kluwer

Arbia G, Benedetti R, Espa G 1996 Effects of the MAUP on image classification. *Geographical Systems* 3: 159–80

ArcView v 2.1 1996 *The geographic information system for everyone*. Redlands, ESRI

Bailey T C, Gatrell A C 1995 *Interactive Spatial Data Analysis*. Harlow, Longman/New York, John Wiley & Sons Inc.

Bao S, Henry M 1996 Heterogeneity issues in local measurements of spatial association. *Geographical Systems* 3: 1–13

Barnsley M J, Barr S L 1996 Inferring urban land-use from satellite sensor images using kernel-based spatial reclassification. *Photogrammetric Engineering and Remote Sensing* 62: 949–58

Batty M, Longley P 1994 *Fractal cities: a geometry of form and function*. London/San Diego, Academic Press

Besag J 1977 Discussion following Ripley. *Journal of the Royal Statistical Society* B 39: 193–5

Boots B N, Getis A 1988 *Point pattern analysis.* Newbury Park, Sage

Boots B N, Kanaroglou P S 1988 Incorporating the effect of spatial structure in discrete choice models of migration. *Journal of Regional Science* 28: 495–507

Carey H C 1858 *Principles of social science*. Philadelphia, Lippincott

Casetti E 1972 Generating models by the expansion method: applications to geographic research. *Geographical Analysis* 4: 81–91

Casetti E, Poon J 1995 Econometric models and spatial parametric instability: relevant concepts and an instability index. In Anselin L, Florax R J G M (eds) *New directions in spatial econometrics*. Berlin, Springer: 301–21

Christaller W 1935 *Die zentralen Orte in Süddeutschland*. Jena, G Fischer

Clark P J, Evans F C 1954 Distances to nearest neighbor as a measure of spatial relationships in populations. *Science* 121: 397–8

Clark W A V 1969 Applications of spacing models in intra-city studies. *Geographical Analysis* 1:391–9

Cliff A D, Ord J K 1973 *Spatial autocorrelation*. London, Pion

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

Cressie N 1991 *Statistics for spatial data*. Chichester, John Wiley & Sons

Dacey M F 1963 Order neighbor statistics for a class of random patterns in multidimensional space. *Annals of the Association of American Geographers* 53: 505–15

Dacey M F 1965 A review of measures of contiguity for two and *k*-color maps. In Berry B J L, Marble D F (eds) *Spatial analyses: a reader in statistical geography*. Englewood Cliffs, Prentice-Hall: 479–95

Diggle P J 1979 Statistical methods for spatial point patterns in ecology. In Cormack R M, Ord J K *Spatial and temporal analysis in ecology*. Fairland, International Cooperative Publishing House: 95–150

Diggle P J 1983 *Statistical analysis of spatial point patterns*. London, Academic Press

Donnelly K P 1978 Simulations to determine the variance and edge effect of total nearest neighbour distance. In Hodder I (ed.) *Simulation methods in archaeology*. Cambridge (UK), Cambridge University Press: 91–5

Dubin R 1995 Estimating logit models with spatial dependence. In Anselin L, Florax R J G M (eds) *New directions in spatial econometrics*. Berlin, Springer: 229–42

Eastman J R 1993 *Idrisi: a geographical information system*. Worcester (USA), Clark University

Faulkenberry G D, Garoui A 1991 Estimating a population total using an area frame. *Journal of the American Statistical Association* 86: 445–9

Fischer M M, Getis A (eds) 1997 *Recent developments in spatial analysis: spatial statistics, behavioural modelling, and computational intelligence*. Berlin, Springer.

Fischer M M, Gopal S, Staufer P, Steinnocher K 1997 Evaluation of neural pattern classifiers for a remote sensing application. *Geographical Systems* 4: 195–223

Fischer M M, Nijkamp P (eds) 1993 *Geographic information systems, spatial modelling, and policy evaluation*. Berlin, Springer

Fischer M M, Scholten H, Unwin D (eds) 1996 *Spatial analytical perspectives on GIS in environmental and socio-economic sciences*. London, Taylor and Francis

Florax R J G M, Folmer H 1992 Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. *Regional Science and Urban Economics* 22: 405–32

Florax R J G M, Rey S 1995 The impacts of mis-specified spatial interaction in linear regression models. In Anselin L, Florax R J G M (eds) *New directions in spatial econometrics*. Berlin, Springer: 111–35

Foster S A, Gorr W L 1986 An adaptive filter for estimating spatially varying parameters: application to modeling police hours in response to calls for service. *Management Science* 32: 878–89

Fotheringham A S, Densham P J, Curtis A 1995 The zone definition problem in location–allocation modelling. *Geographical Analysis* 27: 60–77

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

Fotheringham A S, Wong D W S 1991 The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23: 1025–44

Frank A U, Campari I (eds) 1993 *Spatial information theory: a theoretical basis for GIS*. Berlin, Springer

Franklin J 1995 Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography* 19: 474–99

Gatrell A C, Bailey T C, Diggle P J, Rowlingson B S 1996 Spatial point pattern analysis and its application in geographical epidemiology. *Transactions, Institute of British Geographers* 21: 256–74

Geary R 1954 The contiguity ratio and statistical mapping. *The Incorporated Statistician* 5: 115–45

GEO-EAS 1988, Las Vegas, United States Environmental Protection Agency, Environmental Monitoring Systems Laboratory

Getis A 1964 Temporal land-use pattern analysis with the use of nearest neighbor and quadrat methods. *Annals of the Association of American Geographers* 54: 391–9

Getis A 1983 Second-order analysis of point patterns: the case of Chicago as a multi-center urban region. *Professional Geographer* 35: 73–80

Getis A 1984 Interaction modelling using second-order analysis. *Environment and Planning A* 16: 173–83

Getis A 1990 Screening for spatial dependence in regression analysis. *Papers of the Regional Science Association* 69: 69–81

Getis A 1991 Spatial interaction and spatial autocorrelation: a cross-product approach. *Environment and Planning A* 23: 1269–77

Getis A 1993 GIS and modelling prerequisites. In Frank A U, Campari I (eds) *Spatial information theory: a theoretical basis for GIS*. Berlin, Springer: 322–40

Getis A 1994 Spatial dependence and heterogeneity and proximal databases. In Fotheringham A S, Rogerson P (eds) *Spatial analysis and GIS*. London, Taylor and Francis: 105–20

Getis A 1995 Spatial filtering in a regression framework: experiments on regional inequality, government expenditures, and urban crime. In Anselin L, Florax R J G M (eds) *New directions in spatial econometrics*. Berlin, Springer: 172–88

Getis A, Ord J K 1992 The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24: 189–206

Getis A, Ord J K 1996 Local spatial statistics: an overview. In Longley P, Batty M (eds) *Spatial analysis: modelling in a GIS environment*. Cambridge (UK), GeoInformation International: 269–85

Gong P, Howarth P J 1990 An assessment of some factors influencing multispectral land cover classification. *Photogrammetric Engineering and Remote Sensing* 56: 597–603

Goodchild M F 1992 Geographical information science. *International Journal of Geographical information Systems* 6: 31–45

Goodchild M F, Gopal S 1989 *Accuracy of spatial databases*. London, Taylor and Francis

Green M, Flowerdew R 1996 New evidence on the modifiable areal unit problem. In Longley P, Batty M (eds) *Spatial analysis: modelling in a GIS environment*. Cambridge (UK), GeoInformation International

Griffith D A 1988 *Advanced spatial statistics: special topics in the exploration of quantitative spatial data series*. Dordrecht, Kluwer

Griffith D A 1995 The general linear model and spatial autoregressive models. In Anselin L, Florax R J G M (eds) *New directions in spatial econometrics*. Berlin, Springer: 273–300

Griffith D A 1996a Computational simplifications for space-time forecasting within GIS: the neighbourhood spatial forecasting model. In Longley P, Batty M (eds) *Spatial analysis: modelling in a GIS environment*. Cambridge (UK), GeoInformation International

Griffith D A, Haining R, Arbia G 1994 Heterogeneity of attribute sampling error in spatial datasets. *Geographical Analysis* 26: 300–20

GS+ Geostatistics for the environmental sciences 1995 Plainwell, Gamma Design Software

Haining R P 1990 *Spatial data analysis in the social and environmental sciences*. Cambridge (UK), Cambridge University Press

Haining R P 1995 Data problems in spatial econometric modelling. In Anselin L, Florax R J G M (eds) *New directions in spatial econometrics*. Berlin, Springer: 156–71

Harvey D 1966 Geographic processes and the analysis of point patterns. *Transactions, Institute of British Geographers* 40: 81–95

Haslett J, Bradley R, Craig P, Unwin A, Wills G 1991 Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician* 45: 234–42

Haslett J, Wills G, Unwin A 1990 SPIDER [Regard] – an interactive statistical tool for the analysis of spatially distributed data. *International Journal of Geographical Information Systems* 4: 285–96

Hepple L W 1996 Directions and opportunities in spatial econometrics. In Longley P, Batty M (eds) *Spatial analysis: modelling in a GIS environment*. Cambridge (UK), GeoInformation International

Holt D, Steel D G, Tranmer M 1996 Area homogeneity and the modifiable areal unit problem. *Geographical Systems* 3: 181–200

Hubert L J 1979 Matching models in the analysis of cross-classifications. *Psychometrika* 44: 21–41

Hunt L, Boots B 1996 MAUP effects in the principal axis factoring technique. *Geographical Systems* 3: 101–22

Jacquez G M 1996 Disease cluster statistics for imprecise space-time locations. *Statistics in Medicine* 15: 873–85

Kelejian H H, Robinson D P 1993 A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. *Papers in Regional Science* 72: 297–312

Kelejian H H, Robinson D P 1995 Spatial correlation: a suggested alternative to the autoregressive model. In Anselin L, Florax R J G M (eds) *New directions in spatial econometrics*. Berlin, Springer: 75–95

King L J 1962 A quantitative expression of the pattern of urban settlement in selected areas of the United States. *Tijdschrift voor Economische en Sociale Geografie* 53: 1–7

Krishna Iyer P V A 1949 The first and second moments of some probability distributions arising from points on a lattice, and their applications. *Biometrika* 36: 135–41

Lalanne L 1863 Untitled. *Comptes Rendus des Séances de l'Academie des Sciences*, 57(July–Dec): 206–10

LeSage J P 1995 A multiprocess mixture model to estimate space-time dimensions of weekly pricing of certificates of deposit. In Anselin L, Florax R J G M (eds) *New directions in spatial econometrics.* Berlin, Springer: 359–97

Longley P, Batty M 1996a Analysis, modelling, forecasting, and GIS technology. In Longley P, Batty M (eds) *Spatial analysis: modelling in a GIS environment*. Cambridge (UK), GeoInformation International: 1–15

Longley P, Batty M (eds) 1996b *Spatial analysis: modelling in a GIS environment*. Cambridge (UK), GeoInformation International

Lösch A 1954 *The economics of location*. New Haven, Yale University Press

Mantel N 1967 The detection of disease clustering and a generalised regression approach. *Cancer Research* 27: 209–20

Marceau D J, Gratton D J, Fournier R A, Fortin J P 1994 Remote sensing and the measurement of geographical entities in a forest environment: 2. The optimal spatial resolution. *Remote Sensing of the Environment* 49: 105–17

McMillen D P 1992 Probit with spatial autocorrelation. *Journal of Regional Science* 32: 335–48

Mendeleev D I 1906 *K poznaniyu rossii* (Russian information). St Petersburg, A S Suvorina

Mertes L K, Daniel D L, Melack J M, Nelson B, Martinelli L A, Forsberg B R 1995 Spatial patterns of hydrology, geomorphology, and vegetation on the floodplain of the Amazon River in Brazil from a remote sensing perspective. *Geomorphology* 13: 215–32

Michaelsen J, Schimel D S, Friedl M A, Davis F W, Dubayah R C 1996 Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. *Journal of Vegetation Science* 5: 673–86

Moran P A P 1948 The interpretation of statistical maps. *Journal of the Royal Statistical Society B* 10: 243–51

Morrison A C, Getis A, Santiago M, Rigau-Peres J G, Reiter P 1996 Exploratory space-time analysis of reported dengue cases during an outbreak in Florida, Puerto Rico, 1991–92. *American Journal of Tropical Medicine*

Neft D S 1966 *Statistical analysis for areal distributions*, Monograph Series, 2, Philadelphia, Regional Science Research Institute

Okabe A, Boots B, Sugihara K 1992 *Spatial tessellations: concepts and applications of Voronoi diagrams*. New York, John Wiley & Sons Inc.

Okabe A, Tagashira N 1996 Spatial aggregation bias in a regression model containing a distance variable. *Geographical Systems* 3: 77–100

Oliver M A, Webster R 1990 Kriging: a method of interpolation for geographical information systems. *International Journal of Geographic Information Systems* 4: 313–32

O'Loughlin J, Anselin L 1991 Bringing geography back to the study of international relations: dependence and regional context in Africa, 1966–78. International Interactions 17: 29–61

Openshaw S 1996 Developing GIS-relevant zone-based spatial analysis methods. In Longley P, Batty M (eds) *Spatial analysis: modelling in a GIS environment*. Cambridge (UK), GeoInformation International

Openshaw S, Taylor P 1979 A million or so correlation coefficients: three experiments on the modifiable a real unit problem. In Bennett R J, Thrift N J, Wrigley N (eds) *Statistical applications in the spatial sciences*. London, Pion

Ord J K, Getis A 1995 Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* 27: 286–306

Paelinck J 1967 *L'efficacité de la politique économique regionale*. Namur, Faculté des Sciences Economiques: 58

Paelinck J, Klaassen L 1979 *Spatial econometrics*. Farnborough, Saxon House

Potvin C, Travis J 1993 Concluding remarks: a drop in the ocean... . *Ecology* 74: 1674–6

Quang P X 1993 Nonparametric estimators for variable circular plot surveys. *Biometrics* 49: 837–52

Ravenstein E G 1885 The laws of migration. *Journal of the Royal Statistical Society* 48: 52

Reilly W J 1929 Methods for the study of retail relationships. *University of Texas, Bulletin*, 2944

Ripley B D 1977 Modelling spatial patterns. *Journal of the Royal Statistical Society B* 39: 172–94

Ripley B D 1981 *Spatial statistics*. New York, John Wiley & Sons Inc.

Robinson A H 1956 The necessity of weighting values in correlation of areal data. *Annals of the Association of American Geographers* 46: 233–6

Rogers A 1969 Quadrat analysis of urban dispersion: 1. Theoretical techniques. *Environment and Planning* 1: 47–80

S+SpatialStats 1996 Seattle, MathSoft, Inc.

Stewart J Q 1950 The development of social physics. *American Journal of Physics* 18: 239–53

Stewart J Q, Warntz W 1958 Macrogeography and social science. *Geographical Review* 48: 167–84

Thomas E N 1960 Maps of residuals from regression: their characteristics and uses in geographic research. *State University of Iowa, Department of Geography, Report*, 2

Thompson S K 1992 *Sampling*. New York, John Wiley & Sons Inc.

Thünen J H von 1826 *Der isolierte Staat in beziehung auf Landwirtschaft und Nationalokonomie*. Jena, G Fischer

Tiefelsdorf M, Boots B 1994 The exact distribution of Moran's *I*. *Environment and Planning A* 27: 985–99

Tobler W R 1963 Geographic area and map projections. *Geographical Review* 53: 59–78

Tobler W R 1979 Cellular geography. In Gale S, Olsson G (eds) *Philosophy in geography*. Dordrecht, Reidel: 379–86

Tobler W R 1989 Frame independent spatial analysis. In Goodchild M G, Gopal S (eds) *The accuracy of spatial databases*. London, Taylor and Francis: 115–22

Upton G J, Fingleton B 1985 *Spatial statistics by example,* Vol. 1. Chichester, John Wiley & Sons

Warntz W, Neft D S 1960 Contributions to a statistical methodology for areal distributions. *Journal of Regional Science* 2: 47–66

Wartenberg D 1985 Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis* 17: 263–83

Weber A 1909 *Über den Standort der Industrien*. Tubingen

Wilson A G 1967 A statistical theory of spatial distribution models. *Transportation Research* 1: 253–69

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

Zipf G K 1949 *Human behavior and the principle of least effort*. Reading (USA), Addison-Wesley.