2

The Special Nature of Spatial Data

Robert Haining

This chapter describes some of the special or distinguishing features of spatial data opening the way to methodological issues that will be treated in more depth in later chapters. The use of the term 'special' should not be taken to imply that no other types of data possess these features. Spatial data analysis is a sub-branch of the more general field of quantitative data analysis and has sometimes suffered from not paying sufficient attention to that fact. Many of the data properties that will be encountered are found in other types of (non-spatial) data but when found in spatial data, may possess a particular structure or properties may arise in particular combinations.

The chapter will first define what is meant by spatial data and then identify properties. It will be helpful, in order to put structure on this discussion, to distinguish 'fundamental' properties of spatial data from properties that are due to the chosen representation of geographical space and from properties that are a consequence of measurement processes by which data are collected for the purpose of storage in the spatial data matrix (SDM). The SDM is what the analyst works with. We conclude by considering the implications of these properties for the methodology of spatial data analysis.

Geographic Information Science (GISc) is the generic label that is frequently used, particularly by geographers, to define the area of science that involves the analysis of spatially referenced data – that is data where each case has some form of locational co-ordinate attached to it. Data is the lynch pin in the process of "doing science" and it is essential that methodologies for spatial data analysis are tuned to the properties of spatial data.

The science undertaken with spatial data is usually 'observational' rather than 'experimental'. This is important. Much spatial data are not collected under controlled situations. We often cannot choose the values of independent variables in order to generate a satisfactory experimental design. There is no replication (in order, for example, to assess the effects of measurement error) and the analyst must take the world as he or she finds it. There may be further problems in specifying what the appropriate locational co-ordinate is when studying certain types of processes and outcomes. All this has implications for the quality of spatial data and for the methodologies that can be employed. We worry not only about the quality of our data but exactly what it is we are observing in any given situation. A consequence of this is that much of the data collected may be used to build a model of the situation under study which can then be used to estimate parameters and test hypotheses. We shall see that some of the fundamental properties of spatial data raise major problems in this regard.

2.1. SPATIAL DATA AND THEIR PROPERTIES

A spatial datum comprises a triple of measurements. One or more *attributes* (X) are measured at a set of *locations* (i) at *time t*, where t may be a point or interval of time. So, if k attributes are measured at n locations at time t, we can present the spatial data in the form:

$$\{x_j\,(i;t)\,;\ j = 1, \ldots, k;\ \ i = 1, \ldots, n\}. \quad (2.1)$$

Equation (2.1) expresses in shorthand much of the content of the SDM. The record of when the observation was taken (t) may be suppressed if analysis is concerned with only a single time period but may be retained if there are to be a series of comparative studies through time or if different attributes were recorded at different times and the analyst needs to be aware of this. Such data may come from a variety of different sources including national censuses; public or private agency records (e.g., national health service, police force areas, consumer surveys); and satellite imagery; environmental surveys; and primary surveys. The data may be collected from a census or from a sampling process. For the purposes of analysis data from different sources may be required. Studies in environmental epidemiology utilise health, demographic, socio-economic and environmental data. These data may come with differing degrees of quality and may not all be collected on the same areal framework (Brindley *et al.*, 2005).

To understand the properties of spatial data we need to understand the relationship between equation (2.1) and the ‘real world’ from which the data are taken. In order to undertake data analysis the complexity of the real world must be captured in finite form through the processes of conceptualization and representation (Goodchild, 1989; Guptill and Morrison, 1995; Longley *et al.*, 2001). We shall focus here only on the issues associated with capturing spatial variation, but the reader should note that there are conceptualization and representation issues associated with the way attributes and time are captured as well.

The first step in this process, which ultimately leads to the construction of the SDM, involves conceptualizing the geography of the real world. There are two views of the geographical world in GISc – the field and the object views. The field view conceptualizes space as covered by surfaces with the attribute varying continuously across the space. This is particularly appropriate for many types of environmental and physical attributes. The object view conceptualizes space as populated by well-defined indivisible objects, a view that is particularly appropriate for many types of social, economic and other types of data that refer to populations. Objects are conceptualized as points, lines or polygons.

These two views constitute models of the real world. In order to reduce a field to a finite number of bits of data then the surface may be represented using a finite number of sample points at which the attribute is recorded or it may be represented using a raster grid. Pixels are laid down independently of the underlying field and its surface variation. Alternatively, the surface may be represented by polygons that partition the space into areas with uniform characteristics (e.g., vegetation zones). How well any field is captured by these different representations will depend on the density of the points or the size of the raster in relation to surface variability. There is a large theoretical and empirical literature on the efficiencies of different spatial sampling designs – for example the properties of random, systematic and stratified random sampling given the nature of variation in the surface to be sampled (see, e.g., Cressie, 1991; Ripley, 1981). The process of discretizing in this way involves a loss of information on surface variability.

This loss of detail on variability also arises when selecting a representation based on the object view. A city may comprise many households (points) but for confidentiality reasons information about households is aggregated into spatially defined groups (polygons) – output areas in the case of the 2001 UK census, enumeration districts prior to 2001 (Martin, 1998). Again aggregation into polygons involves a loss of information. There may be a further loss of information in capturing the polygon itself in the database. It may be captured using a representative point (such as its centroid) and its spatial relationship to other polygons captured using a neighbourhood weights matrix.

The conceptualization of a geographic space as a field or as an object is largely dictated by the attribute. However, representation – the process by which information about the geography of the real world is made finite using geometric constructs – involves making choices (Martin, 1999). These choices include the size and configuration of polygons, the location and density of sample points.

2.1.1. Fundamental properties

Fundamental properties are inherent to the nature of attributes as they are distributed across the earth's surface. There is a fundamental continuity (structure) to attributes in space that derives from the underlying processes that shape the human and physical geographical world. We shall discuss examples of these processes in section 2.2.2. The geographical world would be a strange place if levels of attributes changed suddenly and randomly as we moved from one point in space to another close by. Continuity is also a fundamental property of attributes observed in time. If we know the level of an attribute at one position in space (time) we can make an informed estimate of its level at adjacent locations (points in time). The information that is carried in a piece of data about an attribute at a given location provides information on what the level of the attribute is at nearby locations. However as distance increases then the similarity of attribute values weakens and in the GISc literature this is often referred to as Tobler's First Law of Geography ('…near things are more related than distant things'). Although Tobler's First Law is clearly an oversimplification, and in relation to some types of spatial variation just plain wrong, it is nonetheless a useful aphorism.

Testing for spatial autocorrelation was one of the high-profile research agendas in geography during the quantitative revolution. Geographers adapted spatial autocorrelation statistics based on the join-count statistic, the cross product statistic and the squared difference statistic that had been developed for quantifying spatial structure on regular areal frameworks (grids). These statistics were developed to test for statistically significant spatial autocorrelation on irregular areal frameworks (Cliff and Ord, 1973). The null hypothesis (no spatial autocorrelation) was assessed against a non-specific alternative hypothesis (spatial autocorrelation is present). We shall see how this argument was developed in later years with the introduction and use by geographers of models for spatial variation.

In the earth sciences, dealing principally with point data from surfaces, the quantification of structure was based on the use of the empirical semi-variogram which uses a squared difference statistic (Isaaks and Srivastava, 1989). The advantage of the latter route was that it led naturally to model specification and model fitting using theoretical semi-variograms. Of course these quantitative measures and tests of hypothesis depend on the scale of analysis. That is, they depend on the size of the polygons in terms of which data are reported, the inter-point distance between samples on a continuous surface. Thus the chosen representation has an important influence on the quantification of this fundamental property and hence its presence within any spatial dataset. If samples are taken at sufficient distances apart the level of spatial autocorrelation is likely to be much reduced relative to the case where samples are taken close together.

Autocorrelation statistics are also used to capture temporal structure in attribute values but there are important differences with the spatial situation. Time has a natural uni-directional flow (from past to present) whereas space has no such order. The two dimensional nature of space means that dependency structures might vary not just with distance but direction too giving rise to anisotropic dependency structures with structure along the north–south axis differing from the east–west axis. The presence of spatial autocorrelation, that attribute values are not statistically independent, has fundamental implications for the conduct of spatial analysis.

Spatial autocorrelation, in statistical terms, is a second order property of an attribute distributed in geographic space. In addition there may be a mean or first-order component of variation represented by a linear, quadratic, cubic (etc.) trend. We can think of these as two different scales of spatial variation although the distinction may be hard to make and quantify in practice. As Cressie (1991) remarks: 'What is one person's (spatial) covariance may be another persons mean structure' (p. 25). It has often been remarked that spatial variation is heterogeneous. This type of decomposition (plus a white noise element to capture highly localized heterogeneity) is one way of formally capturing that heterogeneity using what are termed 'global' models. Another approach is to only analyze spatial subsets, that is allow model structure to vary locally.

2.1.2. Properties due to the chosen representation

We have already noted that the extent to which our data retains fundamental properties depends on the chosen representation. We now turn to look at other properties that stem directly or indirectly from the chosen representation.

Representing spatial variation using polygons is employed in many branches of science that handle spatially referenced data. Two of the generic consequences of working with data aggregates are: intra-areal unit heterogeneity and inter-areal unit heteroscedasticity.

Whether the data refer to a continuously varying phenomenon (field view) or aggregations of individuals like households (object view) the effect of bundling data into spatial aggregates has the effect of smoothing variation. In the case of environmental data and the use of pixels then the degree of smoothing will clearly depend on the size of the pixels. The larger the pixels the greater the degree of smoothing. A non-intrinsic partition, where the polygons are defined in terms of attribute variability with the aim of maximizing within unit homogeneity and maximizing between-unit heterogeneity will not produce this effect to the same extent. This second process shares common ground with the process of regionalization – to which it is sometimes compared.

Intra-unit heterogeneity is a particular problem for many types of social science data particularly in those cases where area boundaries are chosen arbitrarily as was the case with the UK census for example prior to 2001. Attributes reported for an area may represent percentages or means of attribute values associated with the individuals (people or households) that have been aggregated and the analyst may have no information on the variability around the mean. If an ecological or contextual attribute is calculated for an area (social capital say, or area deprivation) again the calculation is conditional on the chosen representation and the scale of the partition.

One of the conclusions that might be drawn from this is that it is better to have small areal aggregates rather than large ones. Assuming spatial structure, a reasonable supposition given the discussion in section 2.1.1, then smaller areas should be more homogeneous than larger areas and their mean values should be more representative of their area's population. But such spatial precision comes at the cost of statistical precision. Data errors or small random fluctuations in numbers of events (household burglaries; disease outcomes) will have a big effect on the calculation of rates when populations are small. Take the case of a standardized mortality ratio. If the expected count is small, for example 2.0, then the ratio itself (observed count divided by the expected count) rises or falls by 0.5 with each addition or subtraction of a single case. This will have implications for determining the statistical significance of counts – whether there are significantly more cases than would be expected on the basis of chance alone. It will also have implications for determining the statistical significance of differences in counts between areas which in turn raises problems for the detection of significant crime hotspots or disease clusters.

In summary, there is a trade-off that is linked to the number of individual elements in a polygon. A polygon containing few individuals will tend to be more homogeneous but statistical quantities, such as rates and ratios, tend to be unreliable in the sense that small errors and random fluctuations can impact severely on the calculated values. Polygons containing many individuals will generate robust rates and ratios but often conceal much higher levels of internal heterogeneity.

In practice an area is sometimes partitioned into polygons of varying size and this can yield a secondary effect on data properties. A rate calculated for a polygon where the denominator attribute is small has a larger variance than a rate computed for a polygon where the denominator attribute is large. Moreover there is a mean-variance dependence in the rate statistics. Take the case where the denominator is the number of households ($n(i)$). Rates are observed counts of some attribute (number of burglaries) in polygon $i(O(i))$ divided by the number of households. It follows from the binomial model for $O(i)$ that:

$$E\left[O(i)/n(i)\right] = (1/n\,(i))\,E\left[O(i)\right] = p\,(i)\,;$$

$$\begin{aligned}\mathrm{Var}\left[O(i)/n\,(i)\right] &= (1/n\,(i))^2\ \mathrm{Var}\left[O(i)\right]\\ &= p(i)(1-p(i))/n(i)\end{aligned} \tag{2.2}$$

where $E[\ldots]$ and Var[…] denote mean and variance and $p(i)$ is the probability that any individual in area i (e.g., number of households) has the characteristic (e.g., been burgled) that is being counted. The mean and the variance in equation (2.2) are clearly not independent. It also follows from equation (2.2) that the standard error of the estimate of the rate $p(i)$ which is:

$$\left[\,p(i)\,(1-p(i))\,/n(i)\right]^{1/2}$$

is inversely related to the number of households. It follows that any real spatial variation in rates could be confounded by variation in $n(i)$ (the number of households) or alternatively spatial variation in rates could be an artifact of any spatial structure in

$n(i)$ (see Gelman and Price, 1999 who give examples from disease mapping in the USA).

Standardized ratios provide an estimate of the true but unknown area-specific relative risk of the selected disease under the assumption of an independent Poisson model for the observed counts. It follows from the properties of the Poisson distribution that the standard error of the standardized ratio is $O(i)^{1/2}/E(i)$. Using a normal approximation for the sampling distribution of the standardized ratio, $SR(i)$, approximate 95% confidence intervals can be computed:

$$SR\,(i) \pm 1.96 \left[O(i)^{1/2}/E(i) \right].$$

However there are problems here when making comparisons. The standard error tends to be large for areas with small populations and small for areas with large populations because of the effect of population size on $E(i)$. So extreme ratios tend to be associated with small populations but ratios that are significantly different from 1.0 tend to be associated with areas with large populations (Mollie, 1996).

These examples are intended to illustrate the way in which data properties can be induced by the chosen representation. In certain circumstances the geographical structure of the representation (for example the geography of which areas have large and which have small denominator values) could induce a geographical structure on the statistics which when mapped could then give rise to a misleading impression about trends or patterns in the data.

2.1.3. Properties due to measurement processes

The final step in the creation of the SDM involves obtaining measurements on the attributes of interest given the chosen representation.

Data quality can be assessed in terms of four characteristics: accuracy, completeness, consistency and resolution. As noted above, a spatial datum comprises a triple of measurements: the attributes, location and time. Thus the quality of each of these three measurements needs to be assessed against the four characteristics. What is of interest here, however, is how measurement problems might introduce certain properties into the data (Guptill and Morrison, 1995).

A common assumption in error analysis is that attribute errors are independent. This is likely to hold less often in the case of spatial data. Location error may lead to overcounts in one area and undercounts in adjacent areas because the source of the overcount is the set of nearby areas that have lost cases as a result of the location error. So, count errors in adjacent areas may be negatively correlated (Haining, 2003, pp. 67–70). Location error can be introduced into a spatial data set as a result of having to put data, collected on different spatial frameworks, onto a common spatial framework. Areal interpolation methods are used but these are based on assumptions about how attributes are distributed within areal units and these assumptions often cannot be tested. The consequence is that further levels and patterns of error are introduced into the database (Cockings *et al.*, 1997).

In the case of remotely sensed data, the values recorded for any pixel are not in one-to-one relationship with an area of land on the ground because of the effects of light scattering. The form of this error depends on the type and age of the hardware and natural conditions such as sun angle, geographic location and season. The point spread function quantifies how adjacent pixel values record overlapping segments of the ground so that the errors in adjacent pixel values will be positively correlated (Forster, 1980). The form of the error is analogous to a weak spatial filter passed over the surface so that the structure of surface

AQ : Short 1999 not listed in refernce. Please check.

variation, in relation to the size of the pixel unit, will influence the spatial structure of error correlation. Linear error structures also arise in remotely sensed data (Short, 1999). Finally, we note that the effects of error propagation may further complicate error properties when arithmetic or cartographic operations are carried out on the data and source errors are compounded and transformed via these operations (Haining and Arbia, 1993).

Data incompleteness may induce false patterns in spatial data. Data incompleteness refers to the situation where there are missing data points or values or where there are under or overcounts arising from the reporting process. 'Spatially uniform' data incompleteness raises problems for analysis but spatial variation in the level of data incompleteness with, for example, undercounting 6 more serious in some parts of the study area than others can seriously affect comparative work and the interpretation of spatial variation. Missing or inaccurately located cases in a point pattern of events may result in failure to detect a local cluster of cases (Kulldorff, 1998).

Incompleteness in cancer data leads to forms of under or overcounting which give rise to spatial variation that is an artifact of how the data were collected. In the case of official crime statistics geographical differences between large counties in England may be due to differences in police investigative and reporting practices. On the intra-urban scale, burglaries in suburban areas will, on the whole, be well reported for insurance purposes, but in some inner city areas there may be under reporting either because there is no 'incentive' or because of fear of reprisals. The Census provides essential denominator data for computing small area rates. However refusals to cooperate can lead to undercounting and the 1991 Census in the UK was thought to have undercounted the population by as much as 2% because of fears that its data would be used to enforce the new local 'poll tax'. Inner city areas show higher levels of undercounting than suburban areas where populations are easier to track. Finally, since there are 10 year gaps between successive censuses, population in- and out-flows in many areas may be such as to preserve the essential socio-economic and demographic characteristics of the areas. On the other hand some areas of a city, especially inner-city areas, may experience population mobility and redevelopment which result in marked shifts that have implications for the reliability of the data in the years following the Census.

Finally, in the case of some imagery, some areas of the image may be obscured because of cloud cover. A distinction should be drawn between data that are 'missing at random' from data that are missing because of some reason linked to the nature of the population or the area. Weather stations temporarily out of action because of equipment failure produce data missing at random. On the other hand mountainous areas will tend to suffer from cloud cover more than adjacent plains and there will be systematic differences in land use between such areas. This distinction has implications for how successfully missing values can be estimated and whether the results of data analysis will be biased because some component of spatial variation is unobservable.

Figure 2.1 provides a summary of the points raised in this section.

2.2. IMPLICATIONS OF DATA PROPERTIES FOR THE ANALYSIS OF SPATIAL DATA

In this section we turn to a consideration of the implications of the properties of spatial data for the conduct of spatial analysis. Again we shall simply introduce ideas which will be taken up in more detail in later chapters. We divide this section into situations where spatial properties can be exploited to help solve problems and situations where spatial properties introduce complications for the conduct of data analysis.

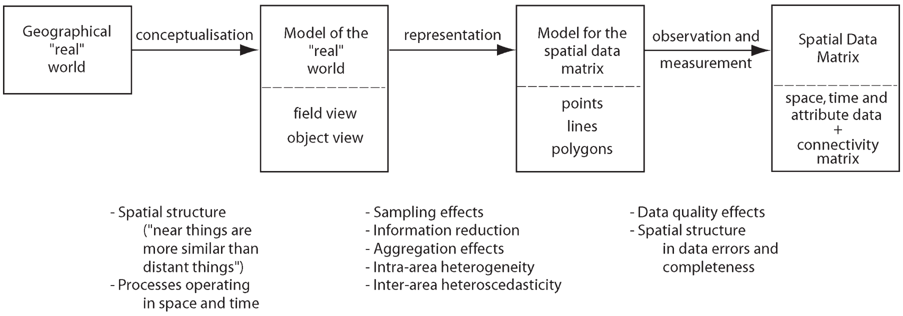
Geographical "real" world
conceptualisation
Model of the "real" world
field view
object view
representation
Model for the spatial data matrix
points
lines
polygons
observation and measurement
Spatial Data Matrix
space, time and attribute data + connectivity matrix
- Spatial structure ("near things are more similar than distant things")
- Processes operating in space and time
- Sampling effects
- Information reduction
- Aggregation effects
- Intra-area heterogeneity
- Inter-area heteroscedasticity
- Data quality effects
- Spatial structure in data errors and completeness

Figure 2.1 Processes involved in constructing the spatial data matrix and the data properties that are present or introduced at each stage.

2.2.1. Taking advantage of spatial data properties to tackle problems

Consider the following problems:

- Samples of attribute values have been taken across an area. The analyst would like to construct a map to describe surface variation using the information contained in the sample. Perhaps instead the analyst just wishes to estimate the surface at a point, or set of points, where no sample has been taken and estimate the prediction error.
- A spatial database has been assembled but the database contains data that are 'missing at random' in the sense that there are no underlying reasons (such as suppression or confidentiality) why the particular values are missing. The analyst wants to estimate these missing values.

In both these cases we might expect to exploit some formalized version of the notion that data points near together in space carry information about each other. Both of these examples constitute a form of the spatial interpolation problem and solutions such as kriging exploit the spatial structure inherent in the surface as well as the configuration of the sample points to provide an estimate of surface values together with an estimate of the prediction error (Isaak and Srivastava, 1989). It is intuitive that any solution that did not use the information contained in the location co-ordinates of sample data values would be considered an inefficient solution.

Consider another group of problems:

- Aggregated data are obtained on race (black/white) and voting behaviour (did vote/did not vote). Counts in the 2 × 2 table are known but the real interest lies in the voting behaviour at the constituency level.
- Unemployment estimates have been obtained from a survey for each of a number of small areas in a region. The small area estimators are unbiased but, because of small sample sizes have low precision. Conversely the region wide estimator has high precision, but as an estimate for any of the small area levels of unemployment is biased. A similar situation arises when estimating relative risk levels across the small areas of a larger region using the standardized mortality ratio.

In both these cases there is again an opportunity to exploit some formalized version of the notion that data points nearby in space carry information about each other. One solution is to 'borrow information' or 'borrow strength' so that the low precision of small area estimates are raised by using data from nearby areas (Mollie, 1996; King, 1997).

These nearby areas provide additional data (helping to improve precision) and because they are nearby should reflect an underlying situation that is close to the small area in question so will not introduce a serious level of bias.

2.2.2. Where spatial data properties introduce complications for data analysis

Spatial analysis is often called upon to address scientific *questions* relating to outcomes (numbers of cases of a disease, distribution of house prices, regional economic growth rates) that are a consequence of processes that by their nature are spatial. Haining (2003) identifies four generic groups of spatial processes. A *diffusion* process is one where some attribute is taken up by a population so that at any point in time some individuals have the attribute (e.g., an infectious disease) and some do not. If the diffusion process operates in ways that are constrained by distance then there is likely to be spatial structure in the geography of those who do and those who do not have the attribute in question. An *exchange and transfer* or *mixing* process is one where places become similar in attribute values (per capita income; employment) as a result of flows of goods or services that bind their economic fortunes together or where patterns of movement and mixing perhaps at different scales introduce a measure of spatial homogeneity into structures. A third type of spatial process is an *interaction* process in which outcomes at one location (e.g., the price of a commodity) are observed and as a result of the competition effect influence outcomes (prices) at another location. Finally, there is a *dispersal* process in which individuals spread across space (such as the dispersal of seeds around a parent plant) so that counts reflect the geography of the dispersal mechanism.

These generic spatial processes – processes that operate in geographic space – generate data where spatial structure emerges as a fundamental property of the data. Process shapes or at least influences attribute variation and the resulting data that are collected possess dependency structures that reflect the way the process plays out across geographic space.

Not all processes of interest are 'spatial' in the sense described above. Many of the processes of interest to geographers play out across geographic space in response to the place-based characteristics of areas (the particular mix of attributes they possess) and the spatial relationships between those areas. Outcomes in places (whether for example economic, social, epidemiological or criminological) are not necessarily merely the consequence of the properties of those places – as places – but may also be the consequence of relational and contextual influences. The distance between places; the difference between adjacent places in terms of relevant attributes; the overall configuration of places across a region, are all facets of relation and context that may impact on outcomes and modify the role of 'place' in influencing outcomes. Two places may be identical in terms of their place-based characteristics but differ significantly in terms of their relational and contextual attributes with neighbouring areas and these differences may explain why (for example) two similarly affluent neighbourhoods experience quite different levels of assault and robbery; why two similarly deprived neighbourhoods experience quite different levels of health outcomes.

We now examine briefly how these features of how attribute values are generated impact on the choice of methodology for the purpose of data analysis. We distinguish between exploratory spatial data analysis and model based forms of analysis that allow hypothesis testing and parameter estimation.

Exploratory spatial data analysis

Exploratory data analysis (EDA) comprises a collection of visual and numerically resistant techniques for summarizing data properties, detecting patterns in data, identifying unusual or interesting features in data including possible data errors and formulating hypotheses. Exploratory spatial data analysis (ESDA) undertakes these activities with respect to spatial data so that cases can be located on a map and the spatial relationships between cases assumes importance because they carry information that is likely to be relevant to the analysis (Cressie, 1984; Haining *et al.*, 1998; Fotheringham and Charlton, 1994). It is important to be able to answer questions such as: 'where does that subset of cases on the scatterplot or that subset of cases on the boxplot, occur on the map?' 'What are the spatial patterns and spatial associations in this geographically defined subset of the map?' In the case of regression modelling do the large positive residuals, for example, cluster in one area of the map?

ESDA and the software that supports ESDA needs to be able to handle the spatial index and be able to handle the special queries that arise because of the spatial referencing of the data. Thus the map becomes an essential visualization tool (Dorling, 1992). The linkage between a map window and other graphics windows, so that cases can be simultaneously highlighted in more than one window, becomes an essential part of the conduct of ESDA (Andrienko and Andrienko, 1999; Monmonier, 1989).

Visualizing spatial data raises particular problems, in part because of some of the properties discussed in earlier parts of this chapter. We highlight two here. First, it has been noted that data values, particularly rates and ratios, may not be strictly comparable because standard errors are population size dependent. So if areas vary substantially in terms of population counts (used as the denominators for a rate) then extreme values and even patterns detected by visual inspection might be associated with that effect rather than real differences between areas. Second, areas that partition a region might be very different in physical size. This may mean that the viewer of a map has their attention drawn to certain areas of the cartographic display (those areas with physically large spatial units) whilst other areas are ignored. This may be particularly important if in fact it is the small areas that have the larger populations so that it is their rates and ratios (rather than the rates and ratios associated with the physically larger but less densely populated areas) that are the more robust. One solution to this problem is to use cartograms so that areas are transformed in physical extent to reflect some underlying attribute such as population size (Dorling, 1994). This comes at a cost because the individual areas in the resulting cartogram may be hard for the analyst to place. There may be a need for a second, conventional, map linked to the cartogram, so the analyst can highlight areas on the cartogram and see where they are on the conventional map.

Conventional visualization technology is often based on the assumption that all data values are of equal status so that the viewer can extract information from visual displays without worrying about the statistical comparability of the data values that are displayed. This assumption may break down when dealing with spatially aggregated data (Haining, 2003).

Model fitting and hypothesis testing

If n data values are spatially autocorrelated then one of the consequences of this for the application of standard statistical inference procedures is that the information content of the data set is less than would be the case if the n values were independent. This means that the degrees of freedom available for testing hypotheses is not a simple function of n. We shall take the example of testing for significant bivariate correlation between two variables to illustrate this point.

Suppose n pairs of observations, $\{(x(i), y(i))\}_i$ are drawn from a bivariate normal distribution. Pearson's product moment correlation coefficient (r) is the statistic used to measure the association between X and Y. If the observations on the two variables are independent (there is no spatial autocorrelation in either X or Y), then if the null hypothesis is of no association between X and Y then a test statistic is given by:

$$(n-2)^{1/2}\,|r|\,\left(1-r^2\right)^{-1/2} \qquad (2.3)$$

which is t distributed with $(n-2)$ degrees of freedom.

These distributional results do not hold if X and Y are spatially correlated. The problem is that when spatial autocorrelation is present the variance of the sampling distribution of r, which is a function of the number of pairs of observations n, is underestimated by the conventional formula which treats the pairs of observations as if they were independent. The effect of spatial autocorrelation on tests of significance have been extensively studied (for reviews see Haining, 1990, 2003) and shown to be very severe when both X and Y have high levels of spatial autocorrelation.

Clifford and Richardson (1985) obtain an adjusted value for $n(n')$ which they call the 'effective sample size'. This value, n', can be interpreted as measuring the equivalent number of independent observations so that the solution to the problem lies in choosing the conventional null distribution based on n' rather than n. An approximate expression for this quantity is:

$$n' = 1 + n^2\left(\text{trace}\left(\mathbf{R}_x\mathbf{R}_y\right)\right)^{-1} \qquad (2.4)$$

where $\mathbf{R}_x$ and $\mathbf{R}_y$ are the estimated spatial correlation matrices for X and Y respectively. (For a discussion of estimators see Haining, 1990, pp.118–120.) The null hypothesis of no association between X and Y is rejected if:

$$\left(n'-2\right)^{1/2}\,|r|\,\left(1-r^2\right)^{-1/2} \qquad (2.5)$$

exceeds the critical value of the t distribution with $(n'-2)$ degrees of freedom.

This illustrates a general problem. Since the n observations are positively spatially autocorrelated, the information content of the sample is over-estimated if n is used – it needs to be deflated. The sampling variance of statistics are underestimated leading the analyst to reject the null hypothesis when no such conclusion is warranted at the chosen significance level. For the effects of spatial dependency on the analysis of contingency tables see, for example, Upton and Fingleton (1989) and Cerioli (1997).

AQ : Upton and Fingleton 1989 not listed in refernce. Please

To make further progress in understanding the importance of spatial data properties and the complications they introduce we need to introduce models for spatial variation – or data generators for spatial variation. Such models are important. By specifying a model to represent the variation in the data (including the spatial variation), the analyst is able to construct tests of hypothesis with greater statistical power than is possible if testing is against a non-specific alternative. There are a number of possible formal models for spatial variation of which the simultaneous spatial autoregressive (SAR), the conditional spatial autoregressive (CAR) and the moving average (MA) models are probably the best known. We will briefly look at the first two but the interested reader will need to follow up the literature to gain a fuller understanding of these models and their properties (Whittle, 1954; Besag, 1974, 1975, 1978; Ripley, 1981; Cressie, 1991; Haining, 1978, 1990, 2003).

A multivariate normal CAR model which satisfies the first order (spatial) Markov property and thus might be thought of as the simplest departure from spatial independence can be written as follows (Besag, 1974; Cressie, 1991, p. 407):

$$\begin{aligned} &E\left[X(i)=x(i)\,\middle|\,\left\{X(j)=x(j)\right\}_{j\in N(i)}\right] \\ &\quad = \mu + \sum_{j\in N(i)} \tau\, w(i,j)\,[X(j)-\mu]\,, \\ &\quad i=1,\ldots,n \qquad (2.6) \end{aligned}$$

and:

$$\mathrm{Var}\left[X(i) = x(i) \,\middle|\, \{X(j) = x(j)\}_{j \in N(i)}\right] = \sigma^2,$$
$$i = 1, \ldots, n$$

where $E[\ldots|.]$ and $\mathrm{Var}[\ldots|.]$ denote conditional expectation and variance respectively, μ is a first-order parameter and τ is the spatial interaction parameter. The Markov property means observations are *conditionally* independent given the values at neighbouring sites. $\{w(i, j)\}$ denotes the neighbourhood structure of the system of areas and $w(i,j) = 1$ if i and j are neighbours ($j \in N(i)$) and $w(i, i) = 0$ for all i. $\mathbf{W}$ is the $n \times n$ matrix of $\{w(i, j)\}$ and is sometimes called the connectivity matrix. It is a requirement that τ lies between $(1/\omega_{\min})$ and $(1/\omega_{\max})$ where $\omega_{\min}$ and $\omega_{\max}$ are the smallest and largest eigenvalues of $\mathbf{W}$. For a fuller introduction to the Markov property for spatial data including how to construct higher-order spatial Markov models see, for example, Haining (2003, pp. 297–299). This approach allows the construction of a hierarchy of models of increasing complexity. As noted in Haining (2003), however, the Markov property does not have the natural appeal it has in the case of time series, because space has no natural ordering. So the neighbourhood structure can often seem rather arbitrary especially in the case of the non-regular areal frameworks used to report Census and other social and economic data.

If the analyst of regional data does not attach importance to satisfying a Markov property another option is available called the SAR model specification. A form of this model was first introduced into statistics by Whittle (1954). Let $\mathbf{e}$ be independent normal $\mathrm{IN}(\mathbf{0},\ \boldsymbol{\sigma}^2\mathbf{I})$ where $\mathbf{I}$ is the identity matrix and $e(i)$ is the variable associated with site $i(i = 1, \ldots, n)$. Define the expression:

$$X(i) = \mu + \sum_{j \in N(i)} \rho w(i, j)\,[X(j) - \mu] + e(i), \quad i = 1, \ldots, n. \qquad (2.7)$$

where ρ is a parameter. The bounds on ρ are set by the largest and smallest eigenvalues of $\mathbf{W}$ just as in the case of the CAR model. This is the model most often seen in the spatial analysis and regional science literature although the reason for its hegemony is far from clear and seems to be largely based on a combination of historical accident (in the sense that time series modelling preceded spatial data modelling and methods were transferred across) and subsequent 'lock-in'.

These models can be embedded into, for example, regression models either as additional covariates (as in the case of equation (2.7)) or as models for the error structure where the errors (in practice the residuals) are tested and found to show evidence of spatial autocorrelation (Anselin, 1988; Ord, 1975). It is well known that fitting regression models by ordinary least squares when errors are spatially (positively) autocorrelated gives rise to some damaging consequences. First, although we shall obtain consistent estimates of the regression parameters (there may be some small sample bias), the sampling variance of these estimates may be inflated compared with methods that take account of the spatial autocorrelation in the errors. Second, if the usual least squares formula for the sampling variances of these regression estimates is applied, the variances will be seriously underestimated. The formulae are no longer valid and conventional F and t tests of hypothesis are also not valid. We shall take a very simple example to illustrate these points, where the parameter to be estimated and tests of hypothesis relate to a constant mean μ.

AQ :Ord 1975 not listed in refernce. Please check.

Suppose n independent observations $\{x(i)\}$ are drawn from a $N(\mu, \sigma^2)$ distribution. The sample mean, $\bar{x}$, is an unbiased estimator for μ, and the variance of the sample mean is:

$$\mathrm{Var}(\bar{x}) = \sigma^2/n. \qquad (2.8)$$

If σ^2 is unknown then it is estimated by:

$$s^2 = (1/(n-1)) \sum_{i=1,\ldots,n} (x(i) - \bar{x})^2 \qquad (2.9)$$

so that:

$$Var\,(\bar{x}) = (1/n\,(n-1)) \sum_{i=1,\,\ldots,\,n} (x(i) - \bar{x})^2\,. \tag{2.10}$$

If the n observations are not independent then although the sample mean is still unbiased as an estimator of μ, assuming each $x(i)$ has the same variance (σ^2), the variance of the sample mean is (see for example Haining, 1988, p. 575):

$$Var\,(\bar{x}) = \sigma^2/n + \left(2/n^2\right) \times \sum_{i} \sum_{j(i<j)} \mathrm{Cov}\,(x(i), x(j)) \tag{2.11}$$

where $\mathrm{Cov}(x(i), x(j))$ denotes the spatial autocovariance between $x(i)$ and $x(j)$. So, if there is positive spatial dependence and σ^2 is known then σ^2/n underestimates the true sampling variance of the sample mean. If σ^2 is unknown and is estimated by equation (2.9) then if there is positive spatial dependence the expected value of s^2 is (see, for example, Haining, 1988, p. 579):

$$E\left[s^2\right] = \sigma^2 - [(2/n(n-1)) \times \sum_{i} \sum_{j(i<j)} \mathrm{Cov}\,(x(i), x(j))] \tag{2.12}$$

so that equation (2.9) is a downward biased estimate of σ^2. This further compounds the underestimation of the sampling variance.

Modified methods to take account of spatial dependence are often based on the following argument (see, for example, Haining, 1988). Assume the data $\mathbf{x}^T = (x(1), \ldots, x(n))$, where T denotes the transpose, are drawn from a multivariate normal spatial model with mean vector given by $\mu\mathbf{1}$ and n by n variance–covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{V}$ given, say, by one of the models described above. (In the case of the CAR model (2.6), $\mathbf{V} = (\mathbf{I} - \tau\mathbf{W})^{-1}$.) The log likelihood for the data is:

$$-(n/2)\ln 2\pi\sigma^2 - (1/2)\ln|\mathbf{V}| - \left(1/2\sigma^2\right) \times (\mathbf{x} - \mu\mathbf{1})^T\ \mathbf{V}^{-1}\,(\mathbf{x} - \mu\mathbf{1}) \tag{2.13}$$

where $\mathbf{1}$ is a column vector of 1's and $|\mathbf{V}|$ denotes the determinant of $\mathbf{V}$. For simplicity we assume $\mathbf{V}$ is known. The maximum likelihood estimator of μ is:

$$\widetilde{\mu} = \left(\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}\right)^{-1}\left(\mathbf{1}^T\mathbf{V}^{-1}\mathbf{x}\right). \tag{2.14}$$

The estimator (2.14) is the best linear unbiased estimator (BLUE) of μ. Note that in the case of independence $\mathbf{V} = \mathbf{I}$ (the identity matrix with 1's down the diagonal and zeros elsewhere) and equation (2.14) reduces to the sample mean. In the case $\mathbf{V} \neq \mathbf{I}$ two modifications to the sample mean are occurring. First, the denominator for positive spatial dependence will be less than n. Second, the presence of $\mathbf{V}^{-1}$ in the numerator of equation (2.14) downweights the contribution of any attribute $x(i)$ which is highly correlated with other attribute values $\{x(j)\}$ – that is, where $x(i)$ is part of a cluster of observations.

The variance of $\widetilde{\mu}$ is:

$$\mathrm{Var}[\widetilde{\mu}] = \sigma^2(\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1})^{-1} \tag{2.15}$$

which reduces to σ^2/n if $\mathbf{V} = \mathbf{I}$.

Since the sample mean is an unbiased estimator of μ, one modification is to replace equation (2.8) with equation (2.15). The term $(\mathbf{1}^T\ \mathbf{V}^{-1}\ \mathbf{1})$ is proportional to Fisher's information measure (Haining, 1988, p. 586). It identifies the information about μ contained in an observation. Now equation (2.9) is not the maximum likelihood estimator for σ^2. This is given by:

$$\widetilde{\sigma}^2 = n^{-1}(\mathbf{x} - \mu\mathbf{1})^T\, \mathbf{V}^{-1}(\mathbf{x} - \mu\mathbf{1}). \quad (2.16)$$

A further refinement is to replace equation (2.9) with equation (2.16) substituting the sample mean for μ in equation (2.16) where $\mathbf{V}^{-1}$ plays a role equivalent to the second term in the right-hand side of equation (2.11).

The general results given by equations (2.11) and (2.12) are why adjustments to conventional methods are needed. The evidence suggests that it is the effect of the second term on the right-hand side of equation (2.11) that is the more serious, at least in the usual situation of positive spatial dependence, and that one way to deal with this is to adjust n in equation (2.8) thereby increasing the sampling variance of the sample mean. The size of the adjustment to n will be sensitive to the estimates of the spatial autocorrelation in the data or, if a spatial model is fitted to the data, the choice of model. The problem is further complicated if, as is usually the case, $\mathbf{V}$ is not known and so must be estimated from the data.

Before leaving the normal model it is important to note that aggregated spatial data may violate another of the statistical assumptions of least squares regression. It was remarked in section 2.1 how rates and ratios based on areas with very different population counts will have different standard errors. It follows that the assumption of homoscedasticity (or constant error variance) is likely to be violated when developing models to explain how rates or ratios vary over a region. Data transformations or weighted least squares estimators are used to address these problems (Haining, 1990, pp. 49–50) but such adjustments may need to be implemented whilst also addressing the problems created by residual spatial autocorrelation (Haining, 1991). In addition to the problems created by failure to satisfy statistical assumptions, spatial data often create 'data-related' problems in regression modelling (Haining, 1990, pp. 332–333). For example, the fit of a trend surface model can be influenced by the configuration of the sample data points on the surface where, as a result of the particular distribution, certain values have high leverage (Unwin and Wrigley, 1987); the particular shape of the study region may also influence the trend surface model fit (Haining, 1990, p. 372). These and other issues are reviewed in Haining (1990, pp. 40–50).

We conclude this section by remarking on the implications of intra-area and inter-area spatial dependency and intra-area heterogeneity when modelling a discrete valued response variable such as the count of the number of cases of a disease across a region using the Poisson model. Spatial dependency and heterogeneity are important causes of overdispersion. For example consider a local diffusion process in which individuals are more likely to be infected if they are close to someone already infected. The result is that counts of the number of cases will reveal Poisson overdispersion because there will be areas with large counts (due to the local infection process) and areas with zero counts where the process has not yet started. These considerations require the analyst both to carry out tests for overdispersion and where necessary take appropriate action. The effects of overdispersion in generalized linear modelling are rather similar to those described for the normal model when spatial autocorrelation is detected. If overdispersion is present, ignoring it tends to have little impact on point estimates of the regression parameters (the maximum likelihood estimator is consistent, although some small sample bias might be present). However, standard error estimates for regression parameters are underestimated. Type I errors associated with the model are underestimated which is particularly problematic in relation to predictors that are close to the significance threshold. If the objective is to build a parsimonious model, the presence of overdispersion may result in an analyst constructing a model more complicated than necessary, and that overestimates the variance explained.

Ways of tackling this problem may depend on the reasons for the overdispersion. A conventional approach is through the use of a variance inflation factor (Dobson, 1999). Where the cause is inter-area spatial autocorrelation then a discrete valued 'auto-model' may be used which is analogous to equation (2.6) (see Besag, 1974). More recently attention has focused on the use of spatial random effects models using CAR models fitted using WinBUGS (Law *et al.*, 2006). These models allow for overdispersion through the random effects term. This is an area of current research in spatial modelling since the development of good modelling tools for discrete valued response variables has rather taken a back seat whilst attention for many years has focused – perhaps disproportionately – on the normal model (Law and Haining, 2004).

2.3. DRAWING INFERENCES

One of the main purposes of undertaking spatial statistical analysis is to make population inferences on the basis of the data collected. In concluding this chapter we consider some of the inference pitfalls associated with the analysis of spatial data.

What is the population about which inferences are made in an observational science? If data are point samples from a continuous surface then the population might be the surface itself. Of course the realized surface may be thought of as only one of many possible realizations (the rest not having been observed). However, with or without the concept of a 'superpopulation' of surfaces, making inferences from point samples to the (realized) surface population does represent a legitimate target. This argument is less convincing when the data represent a complete census – for example the data refer to areas and a complete (or nearly complete) enumeration has been carried out. What is the population about which inferences are being made now? A frequent answer to this is that the underlying process is stochastic (chance is an inherent part of the process) so that inferences are directed at the process (its parameters and covariates) rather than the map. The problem with this is that we have access to only one realization of the process and in order to give our inferences some broader validity other assumptions need to be invoked such as that this realization is representative of the underlying process. There may be no way to test such an assumption.

The modifiable areal units problem (MAUP) reminds us that results obtained from analyzing aggregate data are dependent on the particular scale of the partition, and, at the given scale, the particular boundaries used. In general statistical relationships between attributes are stronger the larger the spatial aggregates because variances are reduced. Boundary shifts can influence whether or not disease clusters or crime hot spots are detected at any scale because if boundaries happen to cut through the middle of a cluster this may dilute the effect over two or more areas.

The analysis of aggregated data is particularly problematic and not just because of the MAUP. It is important to remember that conclusions drawn from aggregate data can only be transferred to the individual level under certain conditions. The ecological fallacy is the uncritical transfer of findings at the group level to the individual level. As the famous example cites, the suicide rate in Germany in the 17th century may have been larger in areas with higher percentages of Catholics but that does not mean Catholics were more prone to commit suicide than Protestants. Quite the reverse as individual level data revealed. Aggregation bias raises serious problems for epidemiological studies based on aggregate data and is one reason why it is considered the weakest of the different methodologies for assessing dose–response relationships – even though this may be the only realistic way of obtaining reasonably sound measures of exposure to an environmental risk factor. The problem is that it is not difficult to construct examples where there are complete sign reversals when going from the ecological to the individual level study (Richardson, 1992).

The converse of the ecological fallacy is the atomistic (or individualistic) fallacy which assumes relationships identified at the individual level apply at the group level. There may be group level or contextual effects that need to be taken into account – as for example in the study of youth offending, where the risk of becoming an offender may not depend only on personal and household level risk factors but also neighbourhood and peer group effects. This then raises the problem of defining what the 'neighbourhood' is.

Figure 2.2 provides a summary of the points raised in sections 2.2 and 2.3.

2.4. CONCLUSIONS

Spatial data possess a number of distinctive properties that derive from the fundamental nature of geographic space and the way processes unfold in geographic space, the way that spatial variation is represented for the purpose of storage in a finite digital database and the way spatial data are collected and attributes measured. Many of these properties were recognized early in geography's 'quantitative revolution' most notably the lack of independence in data values collected close together in space. Geographers then and since have made important contributions to the development of relevant statistical theory and practice.

Geographers continue to develop new methods for describing spatial variation and new methods for modelling processes that operate across geographical space. At present there are two strong traditions which provide focuses for research. On the one hand there are methodologies based on 'whole map' or global statistics that seek to capture data properties through models that are fitted to all the data. On the other hand there are methodologies based on 'local' statistics that process geographically defined subsets of the data and do not seek to impose a single statistic or model on the whole data set (Anselin, 1995, 1996; Getis and Ord, 1996; Fotheringham and Brunsdon, 2000). They represent different ways of responding to the need to develop methodologies to meet the analytical challenges posed by the special nature of spatial data.

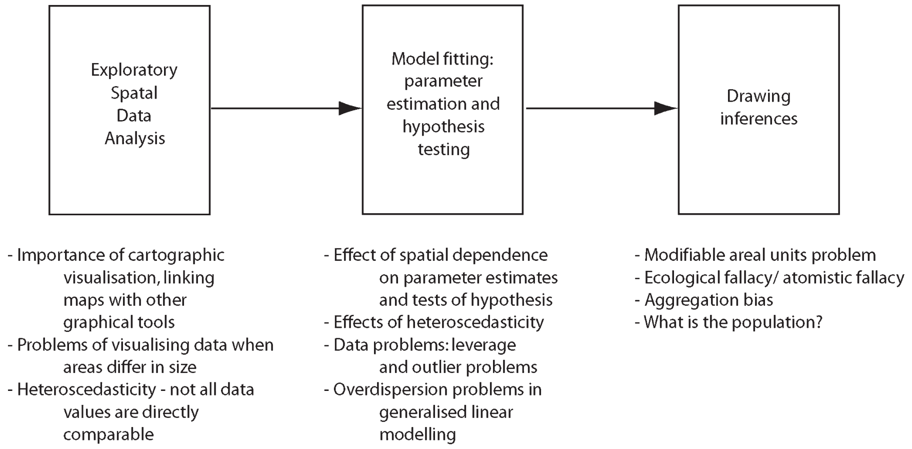
Exploratory Spatial Data Analysis
Model fitting: parameter estimation and hypothesis testing
Drawing inferences
- Importance of cartographic visualisation, linking maps with other graphical tools
- Problems of visualising data when areas differ in size
- Heteroscedasticity - not all data values are directly comparable
- Effect of spatial dependence on parameter estimates and tests of hypothesis
- Effects of heteroscedasticity
- Data problems: leverage and outlier problems
- Overdispersion problems in generalised linear modelling
- Modifiable areal units problem
- Ecological fallacy/ atomistic fallacy
- Aggregation bias
- What is the population?

Figure 2.2 Spatial data properties and how they impact at different stages of analysis.

REFERENCES

Andrienko, G.L. and Andrienko, N.V. (1999). Interactive maps for visual data exploration. *International Journal of Geographical Information Science,* **13**: 355–374.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Dordrecht: Kluwer Academic.

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

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fischer, M, Scholten, H.J. and Unwin, D., (eds), *Spatial Analytical Perspectives on GIS,* pp. 111–125. London: Taylor & Francis.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal, Royal Statistical Society,* B, **36**: 192–225.

Besag, J.E. (1975). Statistical analysis of non-lattice data. *The Statistician,* **24**: 179–195.

Besag, J.E. (1978). Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute,* **47**: 77–92.

Brindley, P., Wise, S.M., Maheswaran, R. and Haining, R.P. (2005) The effect of alternative representations of population location on the areal interpolation of air pollution exposure. *Computers, Environment and Urban Systems,* **29**: 455–469.

Cerioli, A. (1997). Modified tests of independence in 2 × 2 tables with spatial data. *Biometrics,* **53**: 619–628.

Cliff, A.D. and Ord, J.K. (1973). *Spatial Autocorrelation.* London: Pion.

Clifford, P. and Richardson, S. (1985). Testing the association between two spatial processes. *Statistics and Decisions*, **Suppl. No. 2**: 155–160.

Cockings, S., Fisher, P.F. and Langford, M. (1997). Parametrization and visualization of the errors in areal interpolation. *Geographical Analysis,* **29**: 314–328.

AQ : 45 Fisher (1935) not found in text. Please check.

Cressie, N. (1984). Towards resistant geostatistics. In: Verly, G., David, M., Journel, A.G. and Marechal, A., (eds), *Geostatistics for Natural Resources Characterization,* pp. 21–44. Dordrecht: Reidel.

Cressie, N. (1991). *Statistics for Spatial Data.* New York: Wiley.

Dobson, A.J. (1999). *An Introduction to Generalized Linear Models.* Boca Raton: Chapman & Hall.

Dorling, D. (1992). Stretching space and splicing time: from cartographic animation to interactive visualization. *Cartography and Geographic Information Systems,* **19**: 215–227.

Dorling, D. (1994). Cartograms for visualizing human geography. Hearnshaw, H.M. and Unwin, D.J., (eds), *Visualization in Geographic Information Systems,* pp. 85–102. New York: J. Wiley & Sons.

Fisher, R. (1935). *The Design of Experiments.* Edinburgh: Oliver & Boyd.

Forster, B.C. (1980). Urban residential ground cover using LANDSAT digital data. *Photogrammetric Engineering and Remote Sensing,* **46**: 547–558.

Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis.* London: SAGE.

Fotheringham. A.S. and Charlton, M. (1994). GIS and exploratory spatial data analysis: an overview of some research issues. *Geographical Systems,* **1**: 315–327.

Gelman, A. and Price, P.N. (1999). All maps of parameter estimates are misleading. *Statistics in Medicine,* **18**: 3221–3234.

Getis, A. and Ord, J.K. (1996). Local spatial statistics: an overview. In: Longley, P. and Batty, M., (eds), *Spatial Analysis: Modelling in a GIS environment,* pp. 261–277.

AQ : Getis and Ord (1996) publisher required

Goodchild, M.F. (1989). Modelling error in objects and fields. In: Goodchild, M. and Gopal, S. (eds), *Accuracy of Spatial Databases,* pp. 107–113. London: Taylor & Francis.

Guptill, S.C. and Morrison, J.L. (1995). *Elements of Spatial Data Quality.* Oxford: Elsevier Science.

Haining, R.P. (1978). The moving average model for spatial interaction. *Transactions of the Institute for British Geographers*, **NS3**: 202–225.

Haining, R.P. (1988). Estimating spatial means with an application to remotely sensed data. *Communications in Statistics, Theory and Methods,* **17**: 573–597.

Haining, R.P. (1990). *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge: Cambridge University Press.

Haining, R.P. (1991). Estimation with heteroscedastic and correlated errors: a spatial analysis of intra-urban mortality data. *Papers in Regional Science,* **70**: 223–241.

Haining, R.P. (2003) *Spatial Data Analysis: Theory and Practice.* Cambridge: Cambridge University Press.

Haining, R.P. and Arbia, G. (1993). Error propagation through map operations. *Technometrics,* **35**: 293–305.

Haining, R.P., Wise, S.M. and Ma, J. (1998). Exploratory Spatial Data Analysis in a geographic information system environment. *The Statistician,* **47**: 457–469.

Isaaks, E.H. and Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics.* Oxford: Oxford University Press.

King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton, New Jersey: Princeton University Press.

Kulldorff, M. (1998) Statistical methods for spatial epidemiology: tests for randomness. *GIS and Health,* eds Gatrell, A. and Löytönen, M. pp. 49–62. London: Taylor & Francis.

Law, J. and Haining, R.P. (2004) A Bayesian approach to modelling binary data: the case of high intensity crime areas. *Geographical Analysis,* **36**: 197–216.

Law, J., Haining R., Maheswaran, R. and Pearson, T. (2006) Analysing the relationship between smoking and coronary heart disease at the small area level. *Geographical Analysis*, **38**: 140–159.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2001). *Geographical Information Systems and Science.* Chichester: Wiley.

Martin, D.J. (1998) Optimizing Census Geography: the separation of collection and output geographies. *International Journal of Geographical Information Science,* **12**: 673–685.

Martin, D.J. (1999). Spatial representation: the social scientists' perspective. In: Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds), *Geographical Information Systems: Volume 1. Principles and Technical Issues, 2nd edition.* pp. 71–89. New York: Wiley.

Mollie, A. (1996). Bayesian mapping of disease. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics,* pp. 359–379. London: Chapman & Hall.

Monmonier, M.S. (1989). Geographic brushing: exhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis,* **21**: 81–84.

Richardson, S. (1992). Statistical methods for geographical correlation studies. In: Elliot, P., Cuzich, J., English, D. and Stern, R., (eds), *Geographical and Environmental Epidemiology: Methods for Small Area Studies,* pp. 181–204. Oxford: Oxford University Press.

Ripley, B.D. (1981). *Spatial Statistics*. New York: Wiley.

Unwin, D.J. and Wrigley, N. (1987). Towards a general theory of control point distribution effects in trend surface models. *Computers and Geosciences,* **13**: 351–355.

Whittle, P. (1954) On stationary processes in the plane. Biometrika, **41**: 434–449.