CHAPTER THREE

Spatial Statistics and the Analysis of Health Data

ROBERT HAINING

Haining, R. (2014). Spatial statistics and the analysis of health data. In GIS and Health (pp. 47-66). CRC Press.

3.1 Introduction

Spatial epidemiology is the analysis of spatial and space-time distributions of disease data. Such analysis enables the identification of populations with high relative risks for particular diseases and may help to isolate possible causal factors for subsequent analysis by individual level study designs. Area studies may also be important in their own right in the case of certain diseases, such as respiratory, water-borne or those linked to exposure to radiation, where individual level epidemiological studies would be incapable of establishing accurate, individual exposure levels to critical risk factors. Area-based studies, particularly where similar results are found at different times and in different places may give aetiological clues.

Health services research is concerned with health promotion and disease prevention and focuses on questions to do with the need for, provision and use of, particular services as well as the effects of changes in service provision. As with spatial epidemiology, health services research may be directed at specific geographically defined populations seeking to identify and address questions arising from health inequalities in a population. For example, it addresses questions about how service provision impacts on different groups living in different parts of a service catchment thereby contributing to an understanding of the extent to which the health needs of specific populations are being met.

There is growing awareness of the role GIS can play in handling the large volumes of spatially referenced data routinely collected at small spatial scales in the field of public and environmental health (de Lepper *et al.*, 1994). At one level this can mean providing a facility that will help with the mapping and display of such data—although if goals remain as limited as this a mapping package is probably adequate. At a higher level the availability of GIS opens up the possibility for more detailed handling and interrogation of spatially referenced data including the types of questions cited above. However, as has been frequently mentioned (for example, Haining, 1994b) the analysis capability of GIS is quite limited and this in turn limits the extent to which GIS can become a general purpose tool for spatial analysis—in any field of research. The

results of an earlier project developing GIS in the area of health needs assessment, and from which the work described in Section 3.5 grew, is reported in Haining (1996).

In this chapter we shall discuss some of the areas of spatial statistics that are important for spatial epidemiology and health services research. Following a description of some important spatial statistical models and techniques in Section 3.2 there will be a brief discussion in Section 3.3 of the problem of constructing an appropriate areal framework for analysis. Section 3.4 will discuss data quality and Section 3.5 will report on a project that is extending GIS capability to include spatial statistical analysis. Spatial data analysis requires the user to approach data in often quite novel ways, something that is not made any easier by a lack of approriate software for implementing specialist techniques.

3.2 Spatial statistics

This section is divided into two main parts: methods for the exploratory analysis of spatial data and methods for analysing relationships including fitting models and testing hypotheses. Only methods for describing pattern and analysing relationships in area data, what Cressie (1991) calls 'lattice data', will be discussed. Formally this means that we assume that the study region *(R)* has been divided into a set of zones or areas *(D)*—for example a British city divided into enumeration districts or wards or a nation divided into subregions such as counties or health authority regions. Attached to each area $(i \in D)$ is a random vector $\mathbf{Z}_i$ that describes the set of attributes attached to the area. The attributes may include not only medical attributes of the population but demographic, socioeconomic and environmental attributes.

3.2.1 Conceptual models of spatial variation: pattern detection and exploratory analysis

Exploratory spatial data analysis (ESDA) is a collection of statistically robust techniques for identifying different forms of spatial variation in spatial data. It represents the extension of exploratory data analysis (EDA) into the domain of spatial data (Hoaglin *et al.*, 1983). An underlying data model for EDA on a data set for a single variable Z assumes two main components to the data:

$$\text{Data } (Z) = \text{smooth} + \text{rough} \tag{3.1}$$

and EDA techniques, which utilise visual and graphical tools as well as numerical measures, are designed to assist in the identification of these components. The 'smooth', sometimes called 'fit', comprises large scale regular features of the data on Z. The 'rough', sometimes called 'residuals', comprises small scale features of the data on Z that relate to individual data cases or small subsets of cases within the full data set. In the case of spatial data, where the analyst is seeking to describe spatial characteristics of a single variable in R, we shall define the 'smooth' component as regional scale or 'global' patterns in the data. One type of regional scale pattern is a trend the presence of which reflects the existence of an overall gradient in some disease such as, for

example, the often commented on, south-east to north-west increase in heart disease mortality in England. In addition to this large scale gradient other 'smooth' properties may be present in the form of variation around the trend. Superimposed on the trend surface or gradient might be spatial covariation—adjacent areas tending to have similar levels of Z. Such a feature is also termed 'spatial autocorrelation'. As an illustration of these two 'smooth' components, asthma rates may decrease with increasing distance from the centre of a large conurbation, reflecting perhaps an overall decrease in air pollution away from the centre, and this might be represented as a trend surface. Superimposed on this general trend, however, might be spatial autocorrelation in asthma rates reflecting spatial covariation (spatial similarity) in air pollution levels or the general mobility of the urban population in their work and other travel patterns that bring them into regular contact with air pollution levels in areas adjacent to their residential area within the conurbation. In addition to these two global pattern properties representing 'smooth' properties of the surface there may be 'local' elements of the spatial pattern, for example individual or small clusters of areas with particularly high (hot spot) or low (cold spot) rates. These would represent 'rough' properties of the surface. Such 'local' elements of pattern in asthma rates might reflect pockets of gentrification near the city centre enjoying lower rates than the global elements of the surface because these households have the resources to enjoy a more healthy lifestyle, or pockets of deprivation in suburban council housing areas suffering higher rates than the global elements of the surface because these households have poorer housing conditions than the average suburban dweller. The underlying spatial data model thus comprises 'global' and 'local' scale patterns corresponding to the terms 'smooth' and 'rough' in the exploratory data model. In summary:

$$\begin{aligned}\text{Data } (Z) = (&\text{trend} + \text{spatial covariation})\\ &\underbrace{\qquad\qquad}_{\text{global or smooth}}\\ &+ (\text{individual or groups of hot (and cold) spot areas})\\ &\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{local or rough}}\end{aligned}$$

The partition into different forms of spatial variation described above is only one possible model for spatial variation in an ESDA framework. Getis and Ord (1992), for example, distinguish between global and local scales of variation and, in the context of a positive valued attribute with a natural origin, suggest a model that classifies a map on the basis of the degree of spatial concentration in either large or small attribute values. (In a later paper, Ord and Getis (1995), the requirement of a natural origin is removed.) In the terminology of (3.1) the smooth element is the existence of a general propensity for large (or small) values to be found together but within such a global picture there may exist smaller groups of areas of particularly high or low concentration and these would seem to equate with the rough element of (3.1). In a disease map, for example, there might be a general tendency for all areas with high death rates to concentrate together in one or more parts of the map (mainly 'smooth', little or no 'rough'), but at the other extreme there might be no evident concentration of similar sized rates except in a relatively small number of adjacent areas in one or two parts of the map where high rates are found together (little or no 'smooth', mainly 'rough'). The distinction between 'spatial covariation' in the first conceptual model and 'spatial concentration' in this does not appear to be sharp and

it may be possible to reconcile these two conceptual (ESDA) models within a single more general model of spatial variation. However, whatever the model the description of spatial variation is dependent on the scale and nature of the spatial partition *(D)*. The presence of spatial concentration or local clusters superimposed on larger scale gradients or trends will depend on the size of the areal units relative to the scale of the background variation responsible for the local scales of variation.

A number of techniques have been proposed to assist in identifying the 'smooth' and 'rough' elements of a spatial data set. A distinction is drawn between 'general' or 'global' tests, also sometimes called 'whole map' statistics, concerned with identifying overall regional patterns (the 'smooth') and 'focused' or 'local' tests which concentrate on individual areas, $i \in D$, (the 'rough'). This focus on the 'local' might be a sweep through all $i \in D$ to look for evidence of 'rough' wherever it might be found, or it might be concerned with only one or two areas *(i)* in D perhaps because they possess a special attribute which the analyst feels might have implications for health such as a nuclear plant or waste incinerator. See also the chapter by Kulldorff in this volume.

Cressie (1994) proposed median polish with row and column effects to identify gradients in a spatial dataset. His original application was to gridded data but Cressie and Read (1989) carried out a median polish on sudden infant death syndrome data for a set of counties in North Carolina. Unfortunately the method when applied to non-gridded data requires some *ad hoc* decisions to be taken to transform the irregular county data to a grid. Ord and Getis (1995) applied the G_i *(*$i \in D$*)* statistic to county level AIDS data in California to test for trends in the incidence of AIDS away from San Francisco. The G_i statistic measures spatial concentration and is an example of a focused or local statistic, it was computed here only for i=San Francisco but taking successive distance bands at increasing distance from San Francisco. The plot of the resultant G_i values against distance declined with increasing distance from San Francisco providing evidence of a general decline in AIDS with distance from that city. Haining (1990) used a sequence of box plots computed over increasing distance bands from the centre of Glasgow to show the presence of trend in standardised mortality rates. The presence of spatial covariation can be explored in a variety of ways—rigorously by using an appropriate spatial autocorrelation statistic (Cliff and Ord, 1981), more informally, for example using a scatter plot and plotting each value against the mean of the neighbouring areas (Haining, 1990). In both cases the analyst may wish to first remove any spatial gradient in the data set. Ord and Getis (1995) construct the global G statistic which in the context of their model of spatial variation measures the general propensity for attribute values of similar size to concentrate together.

Testing for the presence of 'rough' elements of the data might be undertaken by looking for spatial outliers or even clusters of outliers after removing the trend and spatial covariation elements in the data. Suppose trend has been extracted from a spatial data set and the resultant data set plotted by taking each value and plotting it against the mean of its neighbouring areas. A simple way of signalling outliers is to run a least squares regression through the scatter. Outliers from the regression are indicative of spatial outliers from these 'smooth' properties of the data (Haining, 1990, pp. 197–227). Anselin (1995) has recently drawn attention to USA's (Local Indicators of Spatial Association) which are statistics for picking up local spatial properties. These properties can include local concentrations of events detected by the G_i and G_i^* statistics of Getis and Ord (1992), or local patterns of covariation detected by variogram clouds, pocket plots (Cressie, 1994; 1991) and local Moran plots (Anselin, 1995). There is the potential to generate enormous numbers of

Table 3.1 A summary of components of spatial variation and corresponding techniques

Global/'smooth'		Local/'rough'	
Model component	Technique	Model component	Technique
Trend	Median polish G_i test Box plots	Hot spots and cold spots	Residuals outlier tests
Spatial covariation	Moran test Geary test Scatter plots	Local spatial covariation	Cloud plots Pocket plots Local Moran plots
Concentration	G test	Local concentrations	G_i, G_i^* tests

diagnostic statistics for any set of N areas and it is perhaps appropriate to add a warning note about generating a large number of statistical estimates based on very small subsets of the data (see, for example, the warning in Cressie, 1994). The comparable statistical methodology seems to be that concerned with using the evidence to assess data influences on estimating statistical models or identifing outliers but this is based on deleting individual cases from the full data set rather than the other way round. Also if the user has no explicitly articulated model for the spatial variation it may not be clear what the measures signify about the spatial distribution of values. Table 3.1 provides a summary of techniques in relation to these conceptual models.

3.2.2 Mathematical models of spatial variation: model specification

Rigorous methods based on explicit models of the data are also available. Large scale gradients can be represented by trend surface models. These are linear or higher order functions in the spatial coordinates of the area. The location of each area might for example be represented by the X, Y coordinates of its (population or geometric) centroid. Such models can be augmented by the addition of terms that capture the spatial autocorrelation in the residuals from the trend surface. Spatial autocorrelation models typically represent the value of an attribute *(Z)* at location *i* *(Z_i)* as some function of the values of the same attribute at the neighbours of *i (N(i))*. There are many such models and a discussion of these is beyond the scope of this review but the interested reader can obtain an extended discussion in Haining (1990, pp. 65–117, 249–282) which also discusses the technical problems associated with their fitting. However by way of example a linear trend surface model with autocorrelated errors capturing the two components mentioned above for a single attribute Z might be represented as:

$$Z_i = \beta_0 + \beta_1 X_i + \beta_2 Y_i + \mathrm{u}_i$$
$$\mathrm{u}_i = \rho \Sigma_{j \in N(i)} \mathrm{u}_j + \mathrm{e}_i \qquad (3.2)$$

To revert briefly to the earlier terminology, in this model the terms in X and Y represent the large scale linear gradient (with parameters β_0, β_1 and β_2) and the term in u represents the spatial autocovariance around the trend (with parameter ρ). Together these form the 'smooth'. The term in e is the residual or 'rough' and is an

independent and identically distributed noise process. There is a brief introduction to these models in Richardson (1992).

3.2.3 Analysing relationships in geographical populations

Testing for relationships between, for example, disease events or uptake levels of a service and population and environmental characteristics involves the use of correlation and regression. While at first sight these might seem familiar statistical tools, their application to geographical data poses a number of problems. These problems are both interpretative, stemming from the fact that relationships refer to aggregates of individuals, and technical, stemming from the properties of the underlying random process generating Z.

Problems of interpretation fall into several categories. The *ecological fallacy* (or *ecological bias*) is the difference between estimates of relationships at the aggregate level from those at the individual level. Awareness of this is important if we are not to overstate the strength of any relationship in the individuals of a population that has been observed in the spatially aggregated data. Further, the particular form of the spatial aggregation can also affect the estimate of the relationship and is referred to as the *modifiable areal units problem,* the term stemming from the fact that areal units are not 'natural' but usually arbitrary constructs. This latter problem contains two effects: one effect derives from holding the scale of the aggregation constant but grouping different individuals together—effectively selecting different areal boundaries while holding the overall size and number of areal units constant; the other effect derives from reducing the number but increasing the size of the areal units. The latter might more properly be called a 'scale effect'. It has been noted that as the size of areal units increases, and hence their number decreases, the measure of association tends to increase. The effect of increasing areal unit size is bound up with loss of variability or smoothing in the data induced by the aggregation process. In the case of bivariate correlation this variance appears in the denominator of the statistic which is why this measure of association increases. For example if the population attribute is the Townsend index of deprivation then as larger and larger areal units are used, then the area will tend to contain a population with a greater and greater mixture of deprivation levels. The deprivation score for the area becomes less and less representative of the population in the same way that the mean is a very limited descriptor of a frequency distribution if that frequency distribution possesses a large spread. However, if the analyst is tempted to draw the conclusion from this that therefore it is better to work with small regions to try to ensure homogeneity of the population variables this runs up against a counter problem. Rate estimation, such as computing the standardised incidence rate, is more reliable when computed for large populations than small. Where population counts are small, the addition or subtraction of a few cases will have a far greater effect on computed rates than in the case where population counts are much larger. This is true for any type of rate estimation but is particularly true in the case of relatively rare diseases where the occurrence of a small number of cases can have a large impact on computed rates. There is clearly something of a trade off required here—areas with large enough populations to generate reliable rates but homogeneous with respect to those factors the analyst wishes to explore as possibly helping to explain those rates.

There is a further problem in the case of testing for relationships in spatial data which is the problem of *spurious correlation*. A particular form of this problem arises when 'potential confounder variables…show the same regular spatial pattern' (Richardson, 1992, p. 183). For example, in trying to disentangle the contribution of an environmental influence such as air pollution from deprivation associated with poor housing conditions on respiratory disease rates it tends to be the case that the most deprived groups often live in the centres of cities where air pollution may also be greatest. In such circumstances it may be useful to investigate any relationship both before and after the removal of a gradient particularly if it is suspected that the gradient may reflect the influence of confounding variables. An example is described in Haining (1991a).

Technical problems, particularly in analysing relationships between, for example, disease rates and deprivation levels stem from two properties of such data. As described in the previous section values for either or both disease and deprivation may be *spatially autocorrelated*. This may arise, for example, because the continuity of attribute characteristics is at a scale larger than the areal units. It may also arise because the underlying random process is responsible for generating similar levels in adjacent areas as in the case of an infectious disease for example. In the presence of spatial autocorrelation, while the construction of the Pearson correlation coefficient remains unchanged the inference theory is greatly altered (Clifford and Richardson, 1985; Clifford *et al.*, 1989). Allowance has to be made for the presence of spatial autocorrelation in the two variables and this is done by computing the 'effective sample size' based on estimates of that spatial autocorrelation. Haining (1991b) discusses implementation of the method with a medical application and extends the findings to the Spearman rank correlation coefficient.

The presence of spatial autocorrelation in the residuals (strictly speaking the errors) of a regression model violates one of the statistical assumptions underlying least squares regression and may result in invalid inferences. The estimates of the sampling errors of the parameters of the regression model are underestimated when residuals are positively autocorrelated. As a result independent variables may be retained in the model as significant when they are not (type 1 error). The coefficient of multiple determination, measuring the goodness of fit of the model, is inflated. The underlying cause of residual autocorrelation is often the omission of independent variables from the model that are spatially autocorrelated and which have a significant influence on the variation in the disease rate under investigation. The Moran test is often used for testing for the presence of residual autocorrelation and if it is found to be present either the analyst must try to identify the missing independent variables or fit a regression with correlated errors model or some similar 'spatial regression' as described below. Unfortunately fitting procedures for these models are not routinely available in standard statistical software packages—nor indeed is the Moran test. There is an extended discussion of the regression model with autocorrelated errors in Haining (1990, pp. 123–129) and Haining (1994a).

Another commonly occurring technical problem is non-constant variance, heteroscedasticity, of regression residuals (again strictly speaking the errors). As before this undermines inference. The presence of heteroscedasticity arises from variation in the number of observed cases of the disease between regions. The underlying process generating the observed disease count in any area *(i)* is binomial. If the binomial parameter p_i, which is the probability that an individual catches the disease in area i goes to zero and n_i, which is the number of individuals in area i, goes to infinity under

the condition that their product converges to a finite constant (λ_i) then the observed count in area i is Poisson with parameter λ_i. Poisson regression can therefore be used to model the variation in observed counts which is then conceptualised as the outcome of sampling variation, explained variation associated with specified independent variables and error. Alternatively the counts can be converted to standardised rates, logarithm (to the base e) transformed and the resulting variable is then normally distributed with mean p_i and standard deviation which is a function of the reciprocal of the observed count in each area. The argument, described in detail in Pocock *et al.* (1981) and summarised in Haining (1991a), demonstrates that regression modelling may either be undertaken using Poisson regression (for which no test for residual spatial autocorrelation appears to exist nor a fitting procedure where spatial autocorrelation needs to be allowed for in the model) or by normal regression modelling but for which there is the additional problem of heteroscedastic error.

3.2.4 Special issues in analysing relationships in areal data

The previous section has illustrated that while standard techniques of correlation and regression can be called upon to analyse relationships in the case of area disease data there are problems of interpretation but also problems of a technical nature which make the use of these statistical methods difficult. However, there is a further issue which originates from an earlier comment about the fact that geographical areas are artificial constructs. Area boundaries, typically those used for analysis such as health authority regions, wards or enumeration districts, have no relevance to the process generating the disease counts. Put slightly differently the occurrence or non-occurrence of a disease is not influenced by the distribution of ward boundaries. For certain kinds of disease the open and permeable nature of areal unit boundaries is important. In the case of an infectious disease such as an influenza outbreak the rate of the disease in an area *(i)* may in part be a function of the rate of the disease in neighbouring areas *(N(i))* stemming from the interpersonal communication between residents of nearby, even adjacent areas. This might lead to a model specification of the form:

$$Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \phi \Sigma_{j \in N(i)} Z_j + \mathrm{e}_i \tag{3.3}$$

where Z is the standardised incidence rate of the infectious disease, X_1 and X_2 are independent variables, perhaps measuring socioeconomic characteristics, and the final term in the regression model is modelling the effect of rates in the set of neighbouring areas on the rate in i. The parameters of the model are β_0, β_1, β_2 and ϕ, and e_i is the independent error term. This model is called a spatial regression model with spatially lagged dependent variable. The model cannot be fit by ordinary least squares, unlike the equivalent time series version of this model, and the maximum likelihood procedure yields an estimator with better statistical properties (Haining, 1990).

In the case of certain kinds of disease induced by environmental factors a further variant of the usual regression model might be appropriate. Rates of respiratory disease in area i might be a function of levels of air pollution. Individuals resident in area i, as a result of their general patterns of mobility within the city, might be exposed to levels of the risk factor in areas other than just area i. Thus if Z is the standardised incidence rate of the environmentally induced condition (respiratory

disease) and X_2 is the relevant environmental variable (air pollution) then this might lead to a model specification of the following form:

$$Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma \Sigma_{j \in N(i)} X_{2i} + e_i \tag{3.4}$$

where X_1 is an independent variable where there are no spatial 'spillover' effects in the effect it has on Z, and the parameters of the model are β_0, β_1, β_2 and γ, and e_i is the independent error term. This is called a spatial regression model with spatially lagged independent variables. The model can be fit by least squares regression but does raise the problem of multicollinearity particularly if the variable that is lagged is itself spatially autocorrelated which in the case of many environmental variables is likely to be the case. It should be noted that in all the regression models ((3.2), (3.3) and (3.4)) both the spatial and non-spatial effects are 'whole map' effects in that the relationships are assumed to hold for all $i \in D$ and parameters are spatially invariant. Diagnostics exist to assess the extent to which individual areas do not appear to fit with this assumption, or affect parameter estimates in the case of (3.2) and (3.4) but not (3.3). (See, for example, Haining (1994a).)

The problem of measuring the association between two variables where data refer to areas was considered by Tjostheim (1978) who developed a statistic to measure the degree to which ranked values on two variables occupy positions that are close together in space. The statistic was later generalised by Hubert and Golledge (1982). The statistic is defined:

$$\Lambda = \Sigma_i d(l_F(i), l_G(i))$$

where $l_F(i)$ is the location of rank i on variable F and $l_G(i)$ is the location of rank i on variable G and d (.,.) is a measure of spatial separation. The statistic does not appear to have attracted many applications but measuring the association between the incidence of high levels of respiratory illness and the geographical distribution of air pollution is one area for the reasons discussed above. The statistic is of particular interest because measuring association between variables is complicated in those situations where, because of the nature of the spatial units, exposure levels can be a function of conditions in many units.

This section has reviewed some of the statistical methods for exploring geographically referenced health data and measuring and modelling relationships between health data and socioeconomic and environmental data.

3.3 Constructing an areal system for analysis

The aggregation of health event data to an areal framework involves the loss of some of the original detail although the benefits are that it may then be possible to look at associations between health events and socioeconomic and environmental variables that cannot be studied at the level of individual cases. This may be because of confidentiality reasons or because such data are not available at the level of individuals or because individual level estimates of exposure to risk factors are impossible to compute reliably.

However, one of the consequences of following this route is that the choice of areal aggregation becomes critical. If the aim is to measure the association between

subpopulations and socioeconomic and (or) environmental variables then the areas should satisfy a number of important criteria:

- Availability of other information. Socioeconomic data are available in the UK through the census and the smallest areal unit is the enumeration district (ED) containing on average between 125–220 households. The next unit up is the ward which contains on average about 20 EDs. Health event data, which are postcoded, have to be converted to the ED framework (rather than the other way around). This is because postcodes are not census units and because unit postcodes are very small which would result in much suppression of data for confidentiality reasons were the census data to be purchased from the Office of Population Censuses and Surveys (OPCS) in this form.
- Size. Health event data are converted to rates (typically directly or indirectly standardised rates controlling for the population numbers, age and sex composition of the areas). If areas are small, rates are unlikely to be robust in the sense that with small populations the addition or subtraction of small numbers of disease cases in an area can have a large effect on the size of the computed rate. Not only should all areas be of sufficient size it is also arguable that all areas should be of similar size so that rates *between* areas are of equivalent robustness.
- Homogeneity. Linking disease rates to socioeconomic measures (for example, deprivation) will be unsatisfactory even within the terms of ecological analysis if the area contains considerable variability (heterogeneity) with respect to the socioeconomic characteristic. A spatial framework based on wards will suffer from this problem in the case of a variable like material deprivation (Ubido and Ashton, 1993). EDs also contain heterogeneous populations but the heterogeneity is less pronounced than in the case of wards.

Haining *et al.* (1994) constructed a spatial framework for analysing the association between the incidence of colorectal cancer and material deprivation by obtaining the Townsend index of material deprivation for each of the 1159 EDs in the Sheffield Health Authority Metropolitan District from which were produced 48 Townsend Deprivation Regions.

A requirement for this type of analysis is a good regionalisation algorithm. Openshaw and Liang Rao (1994) review some of the algorithms that exist for constructing optimal zoning systems (regionalisations) based on census EDs. Wise *et al.* (1997) review region building algorithms and describe a procedure available in SAGE (see Section 3.5) that is appropriate for meeting the homogeneity and size criteria important in the case of medical geography and which by emphasising speed of operation may be suitable as part of a programme of ESDA.

3.4 Data quality

Quantitative analysis depends critically on data quality. There are several data quality issues that arise in the sorts of analyses identified in the earlier sections.

3.4.1 Health data

Perhaps the main concern here (setting aside issues of clinical diagnosis which is outside the scope of this discussion) is the accuracy with which health events can be assigned to

EDs for the purpose of exploring data properties and associations. Address listings can be assigned an exact (to within one metre) National Grid reference using the Ordinance Survey ADDRESS-POINT data but at approximately 12 pence per address this can quickly become expensive. Alternatively, and indeed if only the postcode rather than the full address is available, the Postcode Address File (PAF) can be used which is considerably cheaper but only accurate to within 100 metres. A new directory linking postcodes to EDs has been developed for the 1991 census which is more accurate than PAF. However, both of the cheaper routes run the risk of assigning health events to the wrong ED. In the case of the PAF a 100 metre level of accuracy is clearly a problem in urban areas where EDs are not geographically very large. The 1991 directory has an assignment problem when dealing with postcodes that overlap two or more EDs (Collis *et al.,* 1998).

The implication here is that the locational accuracy of health event data is compromised by any assignment process other than the (expensive) OS ADDRESSPOINT data. When dealing with relatively rare events such misassignment can have a serious impact on estimated rates. Where EDs are grouped into large contiguous areas the risks of inaccurate assignment are reduced.

3.4.2 Census data

Census data are collected every ten years which immediately sets a limit on the accuracy of the recorded counts for periods other than the date of the census itself. (And, of course, there are inherent errors in the census data at the time of collection associated with undercounting for example.) The application of methods described here to non-census years is subject to a declining level of accuracy in the data with inaccuracy being greatest as the time approaches for the next census. One way to reduce the effect is to focus analyses on the period around the date of the census although, of course, this may not always be appropriate. Areas subject to considerable migration or urban redevelopment for example suffer particularly from these effects.

The accuracy of UK census data is also affected by the process of Barnardisation by which counts are randomly altered by 0, ± 1 for reasons of confidentiality. Confidentiality is also the reason for the suppression of data in the case of spatial units with very small populations. In the case of the larger spatial units such as wards this is less of a problem than in the case of small spatial units like EDs. Such inaccuracy in the census data undermines the computation of several important measures described above including the Townsend index of material deprivation, which is based on standardised values of four census variables, and the expected counts on which standardised rate estimates are based.

There is further discussion on the quality of cancer data in terms of diagnosis, statistical coverage and the linkage of the cancer data with residence data in the chapter by Teppo in this volume. An important concern is that when dealing with small regions, small data inaccuracies can have a serious effect on computed rates, including standardised incidence rates, especially in the case of rare diseases. Bayes adjustment is often applied to standardised incidence rates which helps to ensure comparability of rates computed across areas with different observed numbers of cases (Clayton and Kaldor, 1987). Areas with small observed counts have their rates driven towards the regional mean. There is further discussion of the use of Bayes smoothing in Cressie (1992).

3.5 GIS and spatial statistics

There have been a number of purpose written packages built to permit various forms of spatial statistical analysis. Such packages include GAM (Openshaw *et al.*, 1987), INFOMAP (Bailey, 1990), REGARD (Haslett *et al.*, 1991), MANET (Unwin, 1996) and SpaceStat (Anselin 1990). The software has been largely written from scratch, a problem sometimes eased by drawing on toolkits (such as Tcl and Tk) but still necessitating writing software to do things that GIS is already good at.

Approaches that have sought to attach statistical analysis to GIS can be classified into one of two types: loose coupling and close coupling. In the first case linkage is via data files as in the case of linking ARC/INFO and GLIM (Kehris, 1990). In the second case the system is built round one package modified as necessary to allow other features to be incorporated and calling other packages. Examples of this type of coupling include the work of Ding and Fotheringham (1992), Batty and Yichun (1994) who used ARC/INFO as the starting point and Brunsdon and Charlton (1995) and Gatrell and Rowlingson (1994) who used statistical packages (XLisp-Stat and S-Plus respectively) as the starting point. For reviews of work in this area see Goodchild *et al.* (1992), Haining and Wise (1991) and Haining *et al.* (1996).

Geographic information systems are potentially of considerable value in implementing the analyses identified in the previous sections. They provide database and mapping facilities and in addition they provide some useful specialist facilities. For example, they facilitate the merging and mapping of data recorded on different spatial frameworks. They also have capabilities that assist region building as described in Section 3.3 where clusters of EDs are merged to form larger areas (Wise *et al.*, 1997). GIS is also potentially of considerable value where the analyst wishes to explore the effects of accessibility. Consider, for example, the uptake of a screening programme like breast cancer which in the case of Sheffield is offered at a single site. To what extent may geographical patterns of uptake reflect problems of public transport access to the screening site from different parts of the city? Suppose there is a proposal to close one or more of several service sites (such as an Accident and Emergency unit). What is the geographical pattern of current usage at each of the existing sites and what might the effects be on travel times for different geographical populations arising from closing any one of the current sites?

Notwithstanding these examples of the use of GIS in this area of research, the current statistical analytical capability of ARC/INFO is still modest. Most of the statistical analysis capabilities described in the preceding sections cannot as yet be implemented within any GIS as far as this author is aware. Currently the author is involved in an ESRC funded project jointly with Stephen Wise and Jingsheng Ma at the Department of Geography, the University of Sheffield, to incorporate spatial statistical analysis capability in ARC/INFO of specific relevance to area-based analyses of health data. This involves drawing on ARC/INFO's powerful database management and map drawing facilities while adding additional modules for other necessary activities that ARC/INFO is either unable to perform or is unsuited to performing. These other activities include graphs and spreadsheets as well as a number of spatial statistical analysis techniques as described above. The prototype version of SAGE (Spatial Analysis in a GIS Environment) has been developed in a client-server architecture with ARC/INFO acting as the server while a program consisting of a number of other visual and non-visual functions complementing those

in ARC/INFO acts as a client. The system operates a series of linked windows through which the user can interact rapidly with the data, including highlighting cases in one window (such as a statistical outlier or a set of extreme cases in a box plot or frequency plot) to have them highlighted in other windows including the map window (Haining *et al.*, 1996).

Figures 3.1 to 3.3 illustrate some SAGE sessions. One facility in SAGE allows the user to build regions from small spatial units (for example, enumeration districts) by merging them in ways that allow the user to control for intra regional heterogeneity and inter regional equality of given variables.

Figure 3.1 shows a regionalisation of Sheffield based on 1200 enumeration districts aggregated into about 200 Townsend deprivation regions. The other windows show a histogram of the interquartile range of the enumeration district Townsend scores that comprise each region (homogeneity) and a second histogram of population at risk counts for the set of regions (equality). The regions have been constructed to try to achieve a set of regions with uniform Townsend scores (for the enumeration districts that comprise them) and similar population counts. The histograms show the extent to which these objectives have been realised by the non-hierarchical regionalisation routine that allows the user to specify the number of regions that is required. (For more details of the algorithm see Wise *et al.*, 1997.)

Figure 3.2 shows a session in which a box plot of the 200 incidence rates for colorectal cancer in Sheffield (1979–1983) has been displayed and large (outlier) values from the plot highlighted. These cases are then automatically highlighted in the map and spreadsheet windows. An associated histogram of the Getis-Ord, G_i^*, statistic computed up to and including lag-two adjacency is shown simultaneously and the regions that have been highlighted on the box plot are also highlighted on the histogram. It appears that regions with high incidence rates are not by and large embedded in areas (groups of regions) that have concentrations of high rates since most of the corresponding G_i^* values are not on the right-hand tail of the distribution. Figure 3.3 shows the tail of the histogram of the G_i^* values highlighted. This reveals regions at the centres of concentrations of relatively high incidence rates. The regions are then highlighted on the map, the spreadsheet and the box plot. The response times of the system are good enabling the user to implement interactive ESDA virtually instantaneously as well as confirmatory procedures some of which (such as spatial statistical model fitting) require longer response times (>5 minutes) when the areal system exceeds 500 spatial units.

3.6 Conclusion

The developments reported in this chapter are primarily limited to the linkage of health data with socioeconomic data collected by census units. As noted in the discussion, however, other data may also be relevant to the explanation of the spatio-temporal incidence of disease including environmental data. Environmental data may constitute surfaces rather than areal aggregates so that the spatial index for the observed event is a sample point in two-dimensional space. Examples of this include ground levels of radiation or atmospheric pollution. In order to add environmental data it may be necesary to convert point samples (for example, of air quality) into a surface of air quality so that estimates can be attached to different parts of the city (Collins, 1996). This line of argument suggests that the extension of these methods to

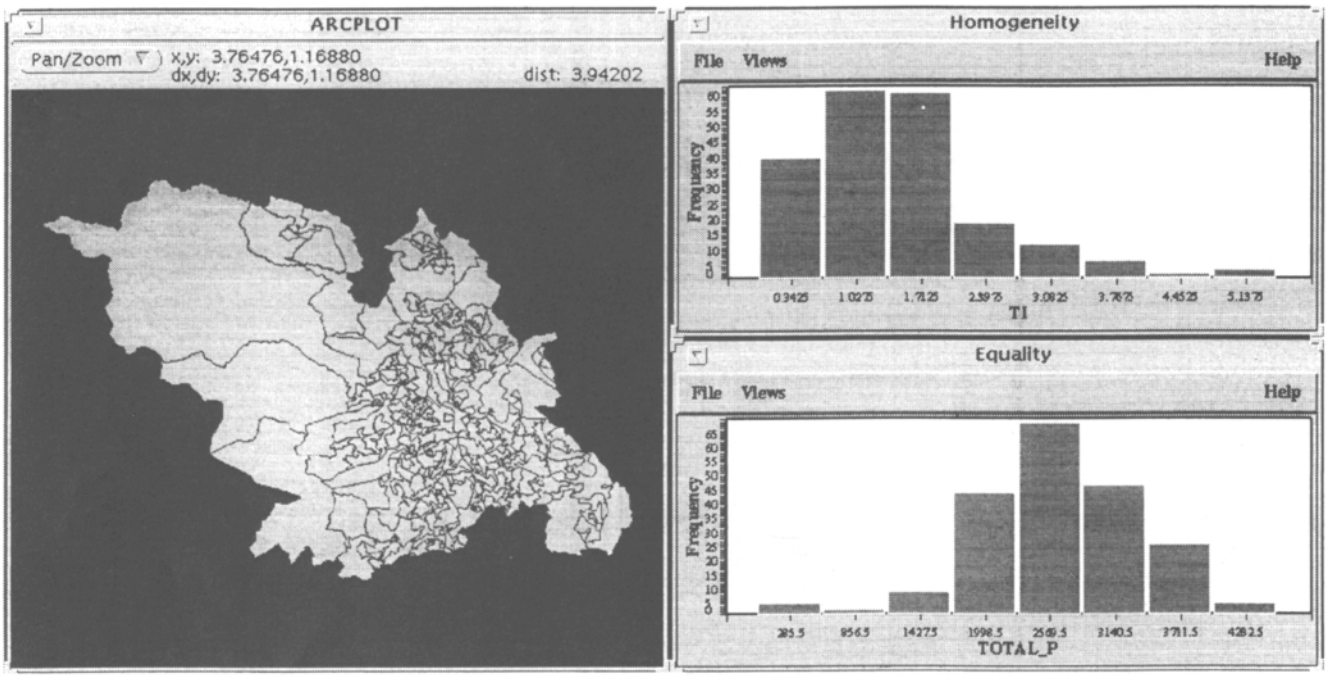

Figure 3.1 Deprivation regions in Sheffield: homogeneity and equal population objectives.

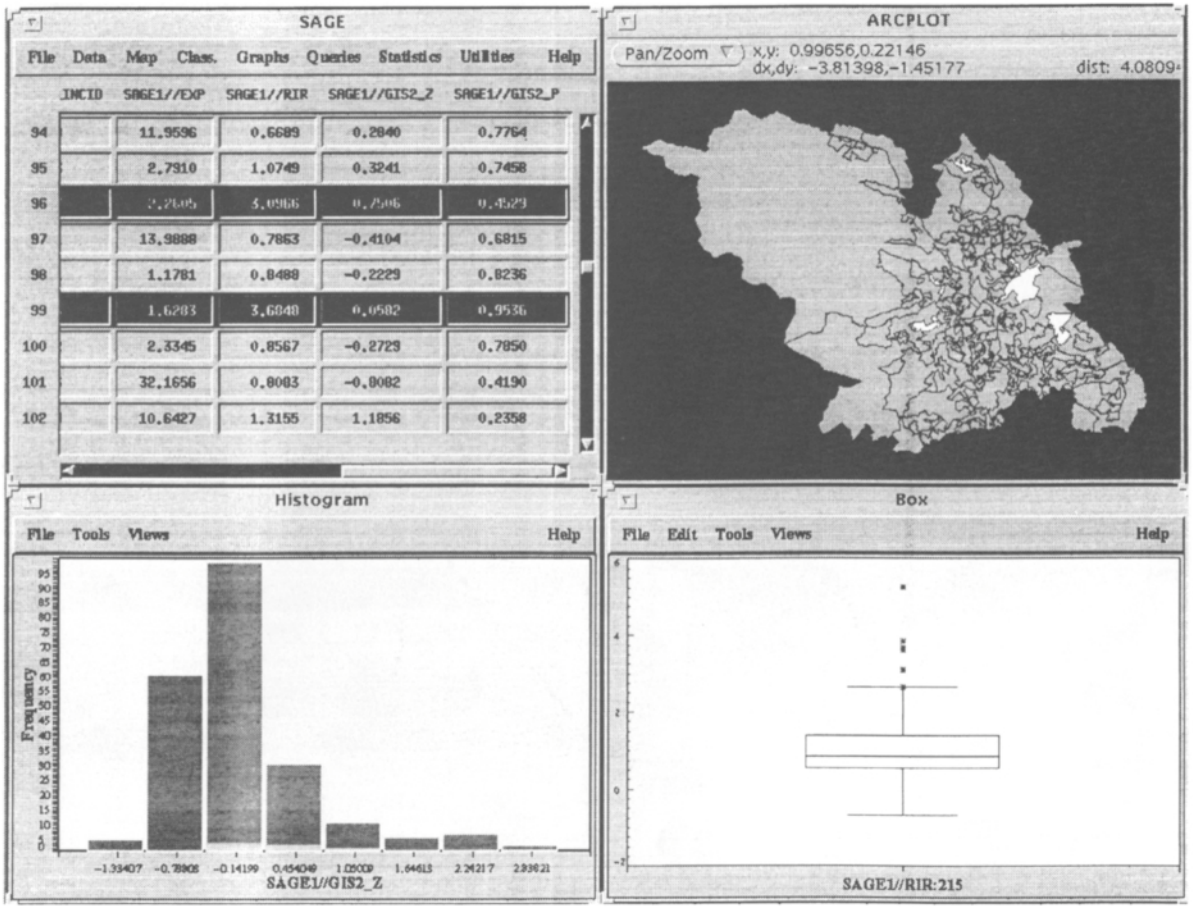

Figure 3.2 Analysis of colorectal cancer incidence rates: identification of regions with large rates.

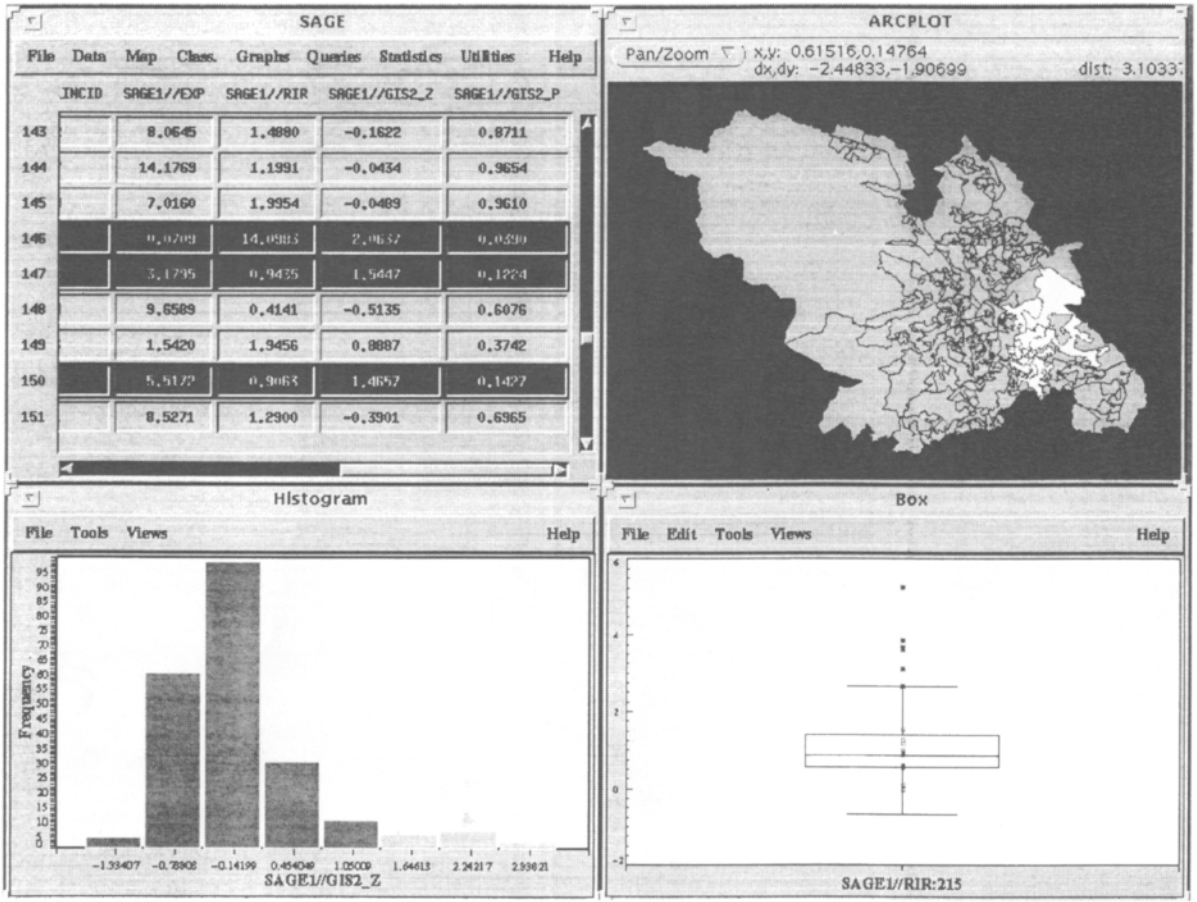

Figure 3.3 Analysis of colorectal cancer incidence rates: identification of areas with concentrations of high rates according to the G_i^* statistic.

this form of health data analysis may require the incorporation of spatial interpolation methods, such as kriging, in which point sample readings are interpolated to a larger area together with estimates of sampling error. A number of interpolation methods have been developed (Ripley, 1981). This in turn extends the requirements for GIS if it is to be able to offer the range of analytical capabilities necessary to undertake spatial analysis of health data.

Acknowledgments

The author wishes to acknowledge receipt of ESRC research grant R000234470 which has made the research reported here on SAGE possible. The author also wishes to thank Steve Wise and Jingsheng Ma, who have collaborated on this ESRC project, for many helpful discussions on the subject matter of this chapter.

References

ANSELIN L. (1990) *Space Stat: A Program for the Statistical Analysis of Spatial Data.* Department of Geography, University of California, Santa Barbara.

ANSELIN L. (1995) Local indicators of spatial association—LISA. *Geographical Analysis,* **27** (2), 93–115.

BAILEY T.C. (1990) GIS and simple systems for visual interactive spatial analysis. *The Cartographic Journal,* **27,** 79–84.

BATTY M. and YICHUN, X. (1994) Urban analysis in a GIS environment: population density modelling using ARC/INFO. Pages 189–220 in Fotheringham S. and Rogerson P. (eds) *Spatial Analysis and GIS,* Taylor and Francis, London.

BRUNSDON C. and CHARLTON M. (1995) A spatial analysis development system using LISP. *Proc. GISRUK '95,* pp. 155–160.

CLAYTON D. and KALDOR J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics,* **43,** 671–681.

CLIFF A.D. and ORD J.K. (1981) *Spatial Processes: Models and Applications.* Pion, London.

CLIFFORD P. and RICHARDSON S. (1985) Testing the association between two spatial processes. *Statistics and Decisions,* Suppl. 2, 155–160.

CLIFFORD P., RICHARDSON S. and HEMON D. (1989) Assessing the significance of the correlation between two spatial processes. *Biometrics,* **45,** 123–134.

COLLINS S.E. (1996) *A GIS approach to Modelling Small Area Variations in Air Quality.* PhD Thesis, University of Huddersfield.

COLLINS S.E., HAINING R.P., BOWNS I.R., CROFTS D.J., WILLIAM T.S., RIGBY A. and HALL D. (1998) Errors in postcode to enumeration district mapping and their effect on small area analysis of health data. *Journal Public Health Medicine* (forthcoming).

CRESSIE N.A.C. (1991) *Statistics for Spatial Analysis.* Wiley, New York.

CRESSIE N. (1992) Smoothing regional maps using empirical Bayes predictors. *Geographical Analysis,* **24,** 75–95.

CRESSIE N. (1994) Towards resistant geostatistics. Pages 21–44 in Verly G. *et al.* (eds) *Geostatistics for Natural Resources Characterization,* Reidel, Dordrecht.

CRESSIE N. and READ T.R.C. (1989) Spatial data analysis of regional counts. *Biometrical Journal,* **6** 699–719.

DE LEPPER M.J., SCHOLTEN H. and STERN R. (1995) *The Added Value of Geographical Information Systems in Public and Environmental Health.* Kluwer, Dordrecht.

DING Y. and FOTHERINGHAM, S. (1992) The integration of spatial analysis and GIS. *Computers, Environment and Urban Systems,* **16,** 3–19.

GATRELL A.C. and ROWLINGSON, B. (1994) Spatial point process modelling in a GIS environment. Pages 147–164 in Fotheringham, S. and Rogerson P. (eds) *Spatial Analysis and GIS,* Taylor and Francis, London.

GETIS A. and ORD J.K. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis,* **24,** 189–206.

GOODCHILD M.G., HAINING R.P. and WISE S.M. (1992) Integrating geographic information systems and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Systems,* **16,** 407–424.

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

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

HAINING R.P. (1991b) Bivariate correlation with spatial data. *Geographical Analysis,* **23** (3), 210–227.

HAINING R.P. (1994a) Diagnostics for regression modeling in spatial econometrics. *Journal of Regional Science,* **34,** 325–341.

HAINING R.P. (1994b) Designing spatial data analysis modules for geographical information systems. Pages 45–64 in Fotheringham S. and Rogerson P. (eds) *Spatial Analysis and GIS* , Taylor and Francis, London.

HAINING R.P. (1996) Designing a health needs GIS with spatial analysis capability. Pages 53–65 in Fischer M., Scholten H. and Unwin D. (eds) *Spatial Analytical Perspectives in GIS,* Taylor and Francis, London.

HAINING R.P. and WISE S.M. (eds) (1991) *GIS and Spatial Data Analysis: Report on the Sheffield Workshop*. ESRC Regional Research Laboratory Initiative. Discussion Paper No. 11.

HAINING R.P., WISE S.M. and BLAKE M. (1994) Constructing regions for small area analysis: material deprivation and colorectal cancer. *Journal of Public Health Medicine,* **16,** 429–438.

HAINING R.P., WISE S.M. and MA J. (1996) Design of a software system for interactive spatial statistical analysis linked to a GIS. *Computational Statistics,* **11,** 449–466.

HASLETT J., BRADLEY R., CRAIG P.S., WILLS G. and UNWIN A.R. (1991) Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *American Statistician,* **45,** 234–242.

HOAGLIN D.C., MOSTELLER F. and TUKEY J.W. (1983) *Understanding Robust and Exploratory Data Analysis*. Wiley, New York.

HUBERT L.J. and GOLLEDGE R.G. (1982) Measuring association between spatially defined variables: Tjostheim's index and some extensions. *Geographical Analysis,* **14,** 273–278.

KEHRIS E. (1990) *A Geographical Modelling Environment Built Around ARC/INFO*. North West Regional Research Laboratory Report 13.

OPENSHAW S., CHARLTON M., WYMER C. and CRAFT A.W. (1987) A Mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems,* **1,** 335–358.

OPENSHAW S. and LIANG RAO (1994) Re-engineering 1991 census geography: serial and parallel algorithms for unconstrained zone design. Paper presented to the *Dublin Meeting of the Regional Science Association,* Dublin, 1994.

ORD J.K. and GETIS A. (1995) Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis,* **27,** 286–306.

POCOCK S.J., COOK D.G. and BERESFORD S.A. (1981) Regression of area mortality rates on explanatory variables: what weighting is appropriate? *Applied Statistics,* **30,** 286–296.

RICHARDSON S. (1992) Statistical methods for geographical correlation studies. Pages 181–204 in Elliott P., Cuzick J., English D. and Stern R. (eds) *Geographical and Environmental Epidemiology: Methods for small area studies*. Oxford University Press, Oxford.

RIPLEY B.D. (1981) *Spatial Statistics*. Wiley, Chichester.

TJOSTHEIM D. (1978) A measure of association for spatial variables. *Biometrika,* **65,** 109–114.

UBIDO J. and ASHTON J. (1993) Small area analysis. *Journal of Public Health Medicine*, **15**, 137–143.

UNWIN A., HAWKINS G., HOFMAN H. and SIEGL G. (1996) Interactive graphics for data sets with missing values—MANET. *Journal Computational and Graphical Statistics*, **5**, 113–122 (other information at http://wwwl.math.uni-augsburg.de/Manet)

WISE S., MA J. and HAINING R.P. (1997) Regionalization tools for the exploratory spatial analysis of health data. Pages 83–100 in Fischer M. and Getis A. (eds) *Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling and Computational Intelligence*. Springer, Berlin.