MATHÉMATIQUES ET SCIENCES HUMAINES

ROBERT HAINING

Spatial modelling and the statistical analysis of spatial data in human geography

Mathématiques et sciences humaines, tome 99 (1987), p. 5-25

<http://www.numdam.org/item?id=MSH_1987__99__5_0>

Math. Sci. hum. (25e année, n°99, *1987*, p.5-25)

SPATIAL MODELLING AND THE STATISTICAL ANALYSIS OF SPATIAL DATA IN HUMAN GEOGRAPHY

Robert HAINING*

1. INTRODUCTION

Geographical data analysis involves the analysis of one or more variables (i) referenced by location (j) and or time (t). A geographical observation $(z_{i,j,t})$ is sometimes, therefore, represented as an element of a data "cube" the three axes of which reference the variable set, the location set and the time set (Haggett (1981)). Any particular analysis may either be on a sub-block of the cube (for example multi-variate space-time data analysis) or, to use a geological analogy, a "thin section" of the cube parallel to one of the faces (for example multi-variate time series analysis, univariate space-time data analysis or multi-variate spatial analysis), or a "transect" of the cube again parallel to one of the faces (for example uni-variate time series analysis or uni-variate spatial analysis).

Geographical data analysis is much concerned with the analysis of data in space and time and since events often possess spatial and temporal continuity, or persistence, the application of statistical methods in human geography has to recognise that geographical data are unlikely to satisfy the classical assumption of independence which under-pins much standard statistical theory. It is the problems created when this assumption no longer holds which are the focus of this paper.

Spatial data analysis and time series data analysis have several fea-

* Department of Geography, The University, Sheffield.
Paper prepared for the International Congress on "Mathematics for the Social Sciences", Marseille, France. June 22nd - 27th, 1987.

tures in common. Space, like time, imposes an *ordering* on the data set which must be recognised and retained in any analysis. It is not sufficient simply to record the level of unemployment in a region over time -10% for six of the months, 10,5% for three of the months and so on. The temporal order of values can tell us much about whether unemployment is showing signs of increasing or decreasing and when taken with other information may help to indicate why the changes are taking place. Similarly the arrangement properties of values on a map may provide important information on how the map pattern has arisen. Spatial data like temporal data also show different *scales of variation* which need to be separated out if any explanation is to be made of total variation. Finally, spatial data like temporal data may consist of different *components of variation* and a distinction is often drawn between deterministic or functional elements of a series and stochastic elements since these may be associated with different types of generating processes.

But there are also differences between spatial and temporal data analysis. Time has *direction* - the past may have influenced the present, but the present cannot have influenced the past (except in the sense that events may be influenced by expectations of the future). Space has no such natural direction and an event at one point can diffuse in all directions to influence events elsewhere on the map. There are special exceptions to this (the directionality implied by the hierarchical ordering of towns in a central place system) but they are not so commonly encountered. Time is *one-dimensional* whereas space is two-dimensional (again with the exception of studies of linear features such as lines of communication) so that the dependency structures in spatial data are likely to be more complex with the possibility of different dependency patterns in different directions. Temporal data are usually collected in terms of a *regular partitioning* (each month, each year, ...) whereas spatial data are usually recorded in terms of an irregular areal partitioning which may confound and mask the contributions of different scales of variation. Finally, whereas time appears as a *homogeneous medium* within which events occur, space appears as a heterogeneous medium and inter point and inter area distance relationships may be measured in many different ways.

This paper will focus on uni-variate and multi-variate spatial data analysis. There has been a considerable growth in interest in this area amongst statisticians and others in the last few years reflected in the recent publication of books by Ripley (1981), Cliff and Ord (1981) and Upton

and Fingleton (1985) on spatial statistical methods. It is my impression however that these methods are still much less well known than those of time series analysis. They are also much less commonly used, perhaps in part because of the lack of appropriate computer software. The area of spatial statistics shares common ground with some of the methods of geostatistics of which the books by Matheron (1971) and Journel & Huijbregts (1978) are particularly important. Whilst the methods of geostatistics have penetrated areas of physical geography and remote sensing they are not widely used in human geography. In this paper therefore I shall focus on the first set of methods for the analysis of spatial data, the problems that are encountered and provide specific examples of each of the main areas. In some of the examples the data is of a space-time form. However, it is often appropriate to analyse such data as a series of spatial "slices". This is because the time period between successive observations is sufficiently long that observations can be treated as independent in time.

In the following sections four classes of problems will be considered : (1) how to make comparisons between two sets of regional data on a single variable (such as income, mortality rates, etc.) recorded either for two different areas or for the same area at two different points in time; (2) how to establish the existence of a statistical relationship between data on two or more variables collected in the same area at the same time; (3) how to estimate missing values in a spatial record (often encountered in census data); (4) the specification and estimation of "spatial process" models.

In each of these sections an underlying issue is the problem of statistical inference for data where independence of observations does not hold and where the dependency structure in the data has to be estimated and built into the estimation and inference procedures. In the discussion that follows there are certain implicit assumptions and we now specify what these are. Variation in spatial data is assumed to arise from three components : a *deterministic* structured element, a stochastic *structured* element and a local random element or *noise*. The first two elements are often represented in terms of the first two moment properties of some probability distribution, the deterministic element being equated with the mean (not necessarily constant) of that probability distribution. In addition an implicit and untestable assumption is that the variation in geographical data arises from one or more of these three elements and represents a single realisation of some probability model.

This implies that each observation is a specific value of a random variable associated with the given location. Matheron (1971) suggested the term "regionalized variable" for the random variable itself. This "conceptualisation" of reality in terms of probability models has many problems but is justified according to Matheron (1971) if it allows us "to solve effectively practical problems which would otherwise be unsolvable" (p.6). We take the view that this is the case in the problems to be tackled in this paper. In the next section we describe two approaches to the measurement of spatial variation as an essential preliminary to considering the more specific problems identified above.

2. REPRESENTATIONS OF SPATIAL DEPENDENCY

Spatial statistical analysis in geography deals with finite regions (D) . There are two approaches to the modelling of spatial processes on finite spaces : (1) consider the process as the restriction to D of a stationary process defined on a larger region; (2) consider the process as defined on D with border or boundary values set (for example) to the mean of the process. In the second case (unlike the first) the process is not stationary. So, suppose $\underline{V} = \{v_{ij}\}$ denotes the matrix of autocovariances. In the case of processes defined according to (1) , and also isotropic

$$v_{ij} = \begin{cases} \sigma^2 & i = j \\ \sigma^2 \, C(d_{ij}) & i \neq j \end{cases}$$

where σ^2 is a scalar constant and d_{ij} is the distance between regions i and j . In the case of processes defined according to (2)

$$v_{ij} = \begin{cases} \sigma_j^2 & i = j \\ \sigma_i \sigma_j \, C(i,j) & i \neq j \end{cases}$$

where $C(i,j)$ signifies that the autocovariance depends on the locations of i and j .

If the first approach is adopted (usually when the study area contains a large number of observations or is a subset of a larger area as often happens with remotely sensed data) then it is usual to describe the dependency properties of a set of data in terms of the correlations or covariances of the probability distribution. If the second approach is adopted (usually when there are few observations and the study area is a clearly delimited areal unit such as an island sub-divided into counties or a city sub-divided into wards or census tracks) then it will not be possible to estimate co-

variances so that dependency properties are usually described in terms of the relationships between the random variables. We consider each of these two descriptive approaches now.

Suppose the underlying probability model is second order stationary (for definitions see for example Journel and Huijbregts (1978)). If $Z(\underline{x})$ denotes the random variable at location $\underline{x} = (x_1, x_2)$ and E denotes mathematical expectation then

$$E\,(Z(\underline{x})) = m \quad \text{for all} \quad \underline{x}$$

where m is a constant. Furthermore for each pair of variables $\{Z(\underline{x}), Z(\underline{x}+h)\}$ the spatial covariance exists and depends only on the separation distance h, that is

$$C(h) = E[(Z(\underline{x}) - m)\,(Z(\underline{x}+h) - m)] \quad \text{for all} \quad \underline{x}\,.$$

Of course the mean may not be constant in which case m is replaced by a function $m(\underline{x})$ which in human geography is usually some order of trend surface model (see Chorley and Haggett (1965)). When $h = o$ then $C(o)$ is the variance of $Z(\underline{x})$.

$C(h)$ is usually estimated for various distances ($h \geq 0$) or distance bands by :

$$\hat{C}(h) = \sum_i (z(\underline{x}_1)-m)\,(z(\underline{x}_i+h)-m) \,/\, n(h)$$

where z denotes an observation on Z and the sum is taken over all the n(h) pairs that are separated by distance h. (Ripley (1981) p.80) notes that the estimator is unbiased for small distances but that the sampling variance depends on C(h) and that neighbouring values of the correlogram are "substatially correlated"). Since, in general the mean is neither a constant nor known it must be estimated from the data (usually a polynomial equation where the order must also be determined) and an iterative procedure is usually recommended in which the mean is estimated then the set of covariances which are in turn used to provide an improved estimate of the mean and so on until convergence. Further complications arise if directional covariances are required (Ripley 1981). Finally, the covariance function is often standardized to a spatial correlation function by dividing through by C(O).

Usually spatial correlation or covariance functions are used to provide formal descriptions of the patterns of association between observations separated by varying distances. Sibert (1975) computed directional spatial corre-

lations to describe spatial variations in assessed land values in Ann Arbor, Michigan as a preliminary step to developing a predictive model of land values. In other instances the correlation function has been used to try to identify characteristics of the underlying generating process as in a study by Cliff et al (1975) on the diffusion of measles epidemics in S.W. England. Peaks and troughs in an averaged spatial correlation function were interpreted as suggesting a "central place plus wave diffusion" pattern of spread. Haining (1981) using rural population density data for the American Mid West estimated correlations and fitted specific models to these functions in order to test a centre-satellite model of population dispersal.

Spectral analysis, the fourier transform of the spatial covariance function, has also been used to describe spatial variation. Tobler (1969) analysed population density along U.S. highway 40 using spectral methods in order to test whether central place competitiveness was important in structuring the spatial organization of population. By and large however spectral methods have not been widely used in human geography mainly because of the need for large sample sizes and the irregular spatial distribution of much two dimensional geographical data, although if these two conditions are not a problem then this form of analysis has more satisfactory asymptotic sampling theory than that based on covariances (Ripley 1981, p. 80).

We now turn to the second approach, that is the representation of dependency in terms of variate relationships. In this approach attention focuses directly on the random variables themselves. As Ripley (1981) shows there are two distinct approaches to the formulation of variate interaction schemes - the joint (or simultaneous) approach and the conditional approach. In the former the dependency structure is modelled as a set of simultaneous equations in which $Z(\underline{x}_i)$ is expressed as a linear function of some subset of the other (usually neighbouring) random variables. For example in the case of a rectangular lattice system where the sites are referenced by (i,j) on the two orthogonal axes :

$$Z(i,j) = \rho(Z(i+1,j)+Z(i-1,j)+Z(i,j+1)+Z(i,j-1)) + e(i,j) \qquad (2.1)$$

where ρ is a parameter that provides a measure of the strength of variate interaction and $e(i,j)$ is some random or noise process. In practice models of this type are specified by first imposing a graph structure on the system of sites. This specifies which random variables interact with one another. In the case of irregular site distributions this is usually expressed by a con-

tiguity or weighting matrix $\underline{W} = \{ w_{k,\ell} \}$ where

$$w_{k,\ell} = \begin{cases} = 0 & \text{if } k = \ell \\ > 0 & \text{if } \ell \text{ is a neighbour of } k \\ = 0 & \text{otherwise} \end{cases}$$

Here k and ℓ denote two sites and non zero values in $\underline{W}$ may be binary (to denote interaction / no interaction) or weighted in order to reflect, for example, how close together the two sites are. The choice of weighting scheme is clearly an important one and there is much discussion in the literature on how to specify $\underline{W}$. In most applications the weighting matrix $\underline{W}$ is specified in advance to reflect relational properties between the sites. This is particularly important in the case of the non lattice data typical of geographical data sets. Sometimes alternative weighting schemes are tried to see how sensitive results are to the choice of $\underline{W}$. Recommendations to try to estimate $\underline{W}$ as if it were a set of parameters have generally not been followed up in part because of the computational difficulty of such a procedure. Discussion of these issues can be found in Besag (1975) and Cliff and Ord (1981) and Upton and Fingleton (1985).

There are subtle and important differences between the conditional and joint approaches (see Besag 1974) and for the most part it is simultaneous schemes of the sort described by (2.1) that have been most often used in human geography. It is important to realise that all variate schemes generate specific theoretical spatial correlations so that the set of empirical correlations are sometimes used for model specification that is for choosing a variable interaction scheme. However, for any finite lattice, models such as (2.1) will not be stationary if sites on the edge of the lattice have only 2 (corner) and 3 (edge) neighbours. This can easily be shown by evaluating $[(\underline{I} - \rho\underline{W})^T(\underline{I} - \rho\underline{W})]^{-1}$ which is the autocovariance matrix for (2.1) if the $e(i,j)$ process is $N(0,1)$. Therefore these models tie in with the second approach to defining spatial processes on finite domains although for sufficiently large lattices these models are often treated as stationary (Whittle 1954).

Variate spatial interaction models have been used to describe spatial patterns in human geography (Sibert 1975). Some models in human geography are expressed in terms of univariate spatial relationships such as the Bechmann-McPherson model of central place population sizes and are therefore appropriately estimated using the estimation theory for these models. Haining (1980) provides an example in this area. These schemes will figure more prominently in section 5 below.

So far in describing the use of correlation functions and variate interaction models we have assumed that spatial dependency is present in any given set of data. To assume that such dependency may be present is a sensible precaution in spatial analysis but as will be evident in later sections, before adopting the sorts of procedures and remedial action this condition necessitates, it is also sensible to test at some stage whether spatial correlation is present in the data (statistically significant). In fact this was one of the first problems to be tackled and two books by Cliff and Ord (1973, 1981) describe the principal procedures that are available. The most commonly used procedure - the Moran test - is closely related to an estimator for first order spatial correlation. Upton and Fingleton (1985 Ch. 3) have reviewed the range of test statistics and have attempted to evaluate their effectiveness (Fig. 3.15). Tests are also available in the context of the variate interaction model approach. For example a test of hypothesis on ρ in (2.1) is equivalent to a test of spatial correlation since if ρ is not found to be statistically significant this reduces the Z process to a random or white noise process. In general however these sorts of tests are more difficult to apply than the simpler correlation - based tests. There are important distributional differences between tests applied to raw data and those applied to regression residuals and these have been discussed at length in Cliff and Ord (1981).

This concludes the survey of methods for describing spatial dependency. Although measures such as the correlation series can be used to yield substantive insights the next two sections are concerned with the application of statistical methods where interest focuses on other data attributes rather than on the correlation properties of the data per se. The problem is to develop statistically valid procedures given the existence of spatial dependency. We return to the question of modelling spatial dependency in the final section.

3. UNI-VARIATE DATA ANALYSIS WITH SPATIALLY CORRELATED DATA

In this section we consider the problem of making comparisons between two sets of regional data with respect to a single variable. A specific problem might be to identify whether there is a significant difference in the population mean value of a given variable between the two sets of data. We assume initially a constant mean but consider in detail trend surface analysis where the mean is not spatially constant. Throughout we assume $C(h) \geq 0$ for all h which is the most commonly encountered situation in human geography.

If n observations $(z(1),\ldots,z(n))$ are drawn from a set of independent and identically distributed random variables with mean μ and variance σ_z^2 then $\bar{z} = (1/n) \sum_{i=1}^{n} z(i)$, for sufficiently large n, is distributed $N(\mu,\sigma_z^2/n)$. This fundamental results underpins standard statistical inference on means and comparisons between means.

On the other hand if the random variables are spatially correlated the following problems arise :

(i) The true variance of $\bar{z}$ is greater than σ_z^2/n even in those situations where σ_z^2 is known a priori.

(ii) In those situations where σ_z^2 must be estimated from the data the classical estimator

$$S^2 = 1/(n-1) \sum_{i=1}^{n} (z(i)-\bar{z})^2$$

is biased downwards.

A fuller discussion of these and other problems are given in Haining (1987 (c) and (d)) however the main problem is that by underestimating the true sampling variance of $\bar{z}$ statistically significant differences may be inferred which are not valid at the chosen level of significance.

One approach to this problem is to select a model both for the population mean (μ) and the spatial correlation properties of the data. We illustrate the method by reference to a specific problem where the mean was not constant but was, instead, described by a trend surface.

Haining (1978) undertook a study of spatial crop yield variation from an area of the American High Plains using data for over 40 counties from 1879 to 1969. Data were available every 5 or 10 years. The area chosen was marginal in that precipitation declined significantly from east to west and temperature increased from north to south. There were also soil gradients which would be expected to lead to lower wheat and corn yields in the west. The analysis sought to show whether these gradients were reflected in crop yield gradients and whether the introduction of new farming methods (such as dry farming in the early 1900's) and improved strains (especially after 1945) and other human impacts would reflect in a gradual weakening of geographical crop yield trends over time.

The model specified was :

$$
\begin{aligned}
Z(i,j) &= \beta_0+\beta_1 X_1(i) + \beta_2 X_2(j) + e(i,j) \\
e(i,j) &= \rho \sum_{k\in N(i,j)} f(e(k)) + u(i,j)
\end{aligned}
\qquad (3.1)
$$

where $Z(i,j)$ denotes the crop yield (for wheat or corn) in county (i,j) and X_1 is the North-South axis and X_2 is the East-West axis. The terms β_0 , β_1 , and β_2 are the trend surface coefficients. The $e(i,j)$ expression is a model for spatial correlation with $u(i,j)$ a random or white noise term. The term $N(i,j)$ refers to the set of neighbours of the (i,j) county, and f denotes a weighting of $e(k)$. In this case the inclusion of the spatial correlation expression was justified on empirical and theoretical grounds. The method of estimation is that described by Ord (1975) which is a modified version of the Cochrane-Orcutt procedure used in time series analysis.

The correlation element of the model was not always significant but the results, particularly for the corn yield data showed strong decreases in yield both from east to west and north to south in the early years (1879-1924) thereafter trends were no longer significant except for certain isolated years (notably the severe drought year of 1934).

The central problem in applying standard statistical theory to data which is not independent is that the amount of "information" (in the statistical sense) available for parameter estimation is less that the sample size. Spatial dependence in data implies that some of the information carried by an observation is (partially at least) duplicated in other observations (in particular the neighbouring observations). Thus, the effective sample size (call this n') is often substantially less than the actual sample size (n) by an amount that depends on the level of spatial dependence present in the data. It is this problem, the need to accomodate this dependency properly and allow for it in constructing sampling variances which lies at the heart of the inferential problem described in this section. Further discussion of these issues together with other references is given in Haining (1978a).

There are certain problem areas however where the issue of "information loss" can be turned to advantage. If each observation carries information about other observations then if a census record for example is incomplete then the known data may be used to construct estimates of the missing data. (The same principal underlies approaches to image reconstruction with remotely sensed data (see Besag 1986) and map interpolation in geostatistics

(Matheron 1971)). Bennett et al (1987) have suggested a maximum likehood procedure for interpolating missing records in census data which is based on finding a model for the observed data on the given variable (usually some order of trend surface model plus correlated error model as in (3.1)) and then using this to provide estimates of the missing values. The procedure is iterative in that parameter estimates are re-estimated on the basis of estimates of the missing values at the previous round. There are benefits to estimating missing values using only the variable itself as opposed to other variables with which the specified variable is correlated since the latter approach may compromise other sorts of (multi-variate) statistical analyses the researcher may wish to carry out on the census data. In some cases it is an estimate of the missing value (together with some measure of the likely error) that is required. In other cases the aim is to analyse relationships between variables and discarding census tracts where one or more variables have missing values would lead to a considerable waste of data. In the second case missing value estimates are only required in as much as they facilitate the primary research objective. The procedure that has been developed is appropriate to both types of research problems.

4. MULTI VARIATE DATA ANALYSIS WITH SPATIALLY CORRELATED DATA

Correlation and regression are the most commonly utilised statistical techniques for identifying relationships between variables in human geography. Both techniques must be modified in their application if the geographical data are spatially correlated.

The Pearson product moment correlation coefficient, r, is a measure of association for two gaussian variables (Y , Z) where

$$\hat{r} = \sum_{i=1}^{n} (y(i)-\bar{y})(z(i)-\bar{z}) \; / \; \left(\sum_{i=1}^{n} (y(i)-\bar{y})^2 \sum_{i=1}^{n} (z(i)-\bar{z})^2 \right)^{1/2}$$

In the case of uncorrelated ($r = 0$) gaussian variables where observations are independent

$$(n-2)^{1/2} \; \hat{r}(1-\hat{r}^2)^{-1/2} \qquad (4.1)$$

has Student's t distribution with $n-2$ degrees of freedom. Bivand (1980) showed that if Y and Z are spatially correlated the effects on the sampling distribution could be severe with serious underestimation of the real type I errors. Clifford and Richardson (1985) have suggested adjusting the statistic (4.1) by replacing n with n' (the effective sample size)

where n' (which is generally less than n if the data are correlated) is dependent on the spatial correlation in the two sets of data. If nearest neighbour correlation varies from 0.2 to 0.8 in both variables, failure to make the suggested adjustment results in type I errors for a 5% test ranging from 8.2% to 52% . When the adjustment is made type I errors are restricted to the interval 3.2% to 7% . (See also Richardson and Hemon (1981, 1982)).

The correlation coefficient measures the relationship between two variables without taking explicit account of the actual positions of the observations (and the Clifford and Richardson adjustment is designed only to adjust critical values for r for the information loss that spatial dependency introduces). Tjostheim (1978) on the other hand developed an index for ranked data that measures the degree of *spatial* association between two variables. The index computes the distance between each pair of identically ranked observations on the two variables and thus specifically takes account of the physical position of the reported data. In Tjostheims original paper his moment and distribution theory for the statistic assumed that neither of the two variables was autocorrelated which, given the argument underlying this paper, suggests that the statistic as originally developed is rather limited in applicability. There have, however, been studies of the behaviour of the statistic in the case of autocorrelated data (Glick 1982). Note that this statistic does provide additional information to that provided by a correlation coefficient and Hubert and Golledge (1982) provide a nice illustration.

In regression modelling classical Gauss-Markov theory requires that the errors (e(i)) are such that

$$E(e(i)) = 0 \quad \text{for all } i$$

$$E(e(i)\, e(j)) = \sigma^2 \quad \text{if} \quad i=j \qquad (4.2)$$

$$= 0 \quad \text{if} \quad i \neq j \qquad (4.3)$$

If the errors are spatially correlated (as evidenced, for example by applying one of the standard tests for residual autocorrelation - see Cliff & Ord (1981)) then assumption (4.3) does not hold and the principal consequences are that goodness of fit (r^2) measures are inflated (as shown above) and the true standard errors of the slope coefficients are also inflated so that there is the risk that variables will be retained in the final model which are not in fact statistically significant at the chosen level.

Hepple (1976) gives an example of this problem arising. His study related to the effects of sales taxes and transport charges on new cars on the average second hand value of cars in the 49 states of the U.S.A. The independent variable was new car price differentials (between states) attribuable to sales tax and transport cost differences. The parameter estimate for the independent variable was 0.686 with a t value of 3.52 which was highly significant. However autocorrelation was identified in the residuals and a model for the errors as specified by (3.1) was introduced into the regression equation. The estimate of the parameter ρ was 0.816 (which is significant) and the t value on the slope ocefficient of the independent variable fell to 0.099 which is no longer significant. Hepple (1979) has since given a Bayesian analysis of the same regression problem. What is evident from this problem and also stressed by Miron (1984) is that the detection of autocorrelation is often indicative of a misspecified model. Statistical models may be used to allow for this misspecification but sometimes at the price (as in the case of the Hepple example) of having no explanation at all of the observed variation. In many instances it may therefore be preferable to visually inspect the pattern of residuals in order to try to identify new variables that might be responsible for the observed residual correlation as in the case of a study by Haining (1981) of urban population variation in S.W. Wisconsin in which variations in non-service sector employment were suggested as possible causes of the residual correlation pattern. Sometimes residuals of one sign may cluster strongly in one area of the map again indicating possible variables to include in an augmented regression model as in a study of income variation by Haining (1987b).

Recently Mardia and Marshall (1985) have placed the statistical analysis of these sorts of spatial problems on a more rigorous foundation deriving asymptotic results for the sampling distribution of maximum likelihood estimators in the case of gaussian variables. They consider numerical algorithms and discuss procedures for specifying and estimating the parameters of models for the correlated error structure. An earlier paper by Ord (1975) is also of importance in this context, however relatively little is still known about the small sample properties of these estimators and inferential procedures.

5. SPATIAL STATISTICS AND SPATIAL PROCESSES

The term "spatial process" is an ambiguous one but is used here to refer to any process in which the spatial distribution of objects (such as towns, farms, shops, etc.) has an effect on the behaviour of variables (such as income levels,

innovation adoption, price level etc) defined on those objects. So space itself is not a process but spatial relationships are one element in the specification of the process. The significance of this may become clearer with examples.

In this section we concentrate on two types of processes in which spatial considerations enter into model specification - these are spatial competition processes and spatial transfer or interaction processes. We examine a *competition* process first.

Consider two retailers located on the same stretch of road in a city selling (more or less) an identical product. Their potential customers are regularly passing up and down the street and they may be, from time to time at least, noting the prices charged for the goods at either or both of the two stores. The decision of *where* to purchase by customers may, all other things being equal, be influenced by the price charged at the two outlets. If the retailers are sensitive to this state of affairs, a situation may well develop in which prices charged at one retail outlet are responsive to prices charged at the other retail outlet (and vice versa). We have described a simple spatial competition process and if the problem is generalised to a larger number of retailers a price "surface" may develop wihch in part reflects competition between retail sites.

The above arguments underlay two statistical models for spatial price variation in petrol retailing developed by Haining (1984). We consider only one of them here (developed from a very simple, neo-classical equilibrium model) and for further discussion of the problem see Haining (1986). Let $\underline{p}_t$ be an $n \times 1$ vector and denote the prices charged at n retail outlets in a city or area. Further, let $\underline{D}_t$ and $\underline{S}_t$ denote n dimensional demand and supply vectors at time t. A simple model for such spatial price competition might be specified as follows :

$$\underline{D}_t = \underline{A}\,\underline{p}_t + \underline{c} + \underline{u}$$

$$\underline{S}_t = \underline{B}\,\underline{p}_{t-1} + \underline{e}$$

where $\underline{c}$ and $\underline{e}$ are n dimensional vectors of constants and $\underline{A}$ and $\underline{B}$ are $n \times n$ ordered matrices whose rows and columns correspond to the labelling of the n sites. The demand vector is stochastic in that $\underline{u}$ is a random vector with mean $\underline{0}$ and diagonal variance covariance matrix $\sigma^2\underline{I}$.

We assume that $\underline{B} = \{b_{i,j}\}$ is such that

$$b_{ij} \begin{cases} > o & \text{if } i = j \\ = o & \text{if } i \neq j \end{cases}$$

and writing $\underline{A} = \{a_{i,j}\}$ then we assume

$$a_{ij} \begin{cases} < o & \text{if } i = j \\ \geq o & \text{if } i \neq j \end{cases}$$

If market clearance takes place (clearly a very strong assumption) $\underline{D}_t = \underline{S}_t$ and the equilibrium price vector $(\underline{p}_e)$ is given by

$$(\underline{A} - \underline{B})^{-1} (\underline{e} - \underline{c} - \underline{u}) .$$

It follows that the equilibrium price at the i^{th} retail outlet $(p_e(i))$ can be expressed as a function of prices elsewhere, that is

$$p_e(i) \propto \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\, p_e(j) + (c(i)-e(i)) + u(i) \qquad (5.1)$$

Equation (5.1) describe an autoregressive spatial scheme with non negative coefficients $(a_{ij} \geq 0)$ and hence positive spatial autocorrelation at all lags. Since $(c(i)-e(i)) \neq 0$ the model implies the presence of site effects.

The model is simplified by setting some of the a_{ij} to 0 . This can be done to reflect the probable structure of intersite competition and the distances over which such effects are expected to operate. The model defined by (5.1) was hypothesized as a description of the variation in petrol prices in S.W. Sheffield during two periods of intensifying price competition in 1982. The collection $\{a_{ij}\}$ were defined on the basis of bilateral nearest neighbour proximity along the principal radial routeways leading to Sheffield. The statistical model fitted was :

$$p_e(i) = \sum_{g=1}^{k} \beta_g\, x_g(i) + \rho \sum_{j=1}^{n} w_{ij}\, p_e(j) + u(i) \qquad i = 1,\ldots,n$$

where $x_g(i)$ is the value for the g^{th} regressor variable $(g = 1,\ldots,k)$ at site i and $\{w_{ij}\}$ is a binary contiguity matric reflecting the geography of bilateral nearest neighbour inter-site interaction. Data collected on a single day were used to estimate $\{\beta_g\}$ and ρ . However, data were collected at several different times over a 7 month period and separate estimates were obtained on each occasion. (Here then is another example of

a space-time data set analysed as a set of independent spatial data sets. This was justified on the grounds that response times in price adjustment would be far more rapid than the time intervals between successive observations). The regressor variables reflected attributes of the retail sites including whether they were on a main route or not and whether they combined petrol retailing with other activites. Interaction effects were examining by testing for the statistical significance of ρ . (For a similar example relating to land price variation see Ancot and Paelinck (1981). House price variation within an urban area may also show similar regularity arising from the activities of estate agents, since the asking price at quite local scales often reflects realised price levels attained in neighbouring "equivalent" housing units).

We now consider an example of a spatial process in which dependency originates from *transfer* mechanisms. Spatial income variation has attracted a certain amount of theoretical and applied research effort over the last decade or so (see for example Paelinck and Klaassen (1979)). The following simple model has been used to describe geographical income variation :

$$\begin{aligned} \underline{Y} &= \underline{X} + \underline{C} \\ \underline{C} &= c\ \underline{Y} \end{aligned} \qquad (5.2)$$

where $\underline{Y}$ is and $n \times 1$ vector of (n) area income levels, $\underline{X}$ is an $n \times 1$ vector of exogenous expenditures including exports, investment and government outlays and $\underline{C}$ is an $n \times 1$ vector of endogenous expenditures (local consumption by community residents). The parameter c $(0 < c < 1)$ is the income creating local propensity to consume, (see Tiebout (1960)). This model (5.2) is often referred to as a "spatialized" Keynesian income model. It follows that

$$\underline{Y} = (1-c)^{-1}\ \underline{X} \ .$$

Now in the context of, for example, a set of towns, the export income of a town includes income accruing to the town as a result of purchases made by non-residents. Haining (1987b) disaggregated the elements of $\underline{X}$ into "long distance" export income $(\underline{X}_1)$ arising from extra-regional trading and "short distance" export income $(\underline{X}_2)$ linked to the transfer of income between towns arising from non-local consumer spending. Given the definition of $\underline{C}$ in (5.2) then consistent with that assumption is the assumption that

$$\underline{X}_2 = \underline{\Omega}\ \underline{Y}$$

where $\Omega = \{\omega_{ij}\}$ and

$$\omega_{ij} \begin{cases} = 0 & \text{if } i = j \\ \geq 0 & \text{if } i \neq j \end{cases}$$

The term $\{\omega_{ij}\}$ is the propensity for income in j to create income in i where

$$0 < \sum_{i=1}^{n} \omega_{ij} + c < 1$$

The income model now has the form

$$\underline{Y} = (1-c)^{-1} [\underline{X}_1 + \underline{\Omega}\ \underline{Y}]$$

which is analogous to a spatial autoregressive model. The structure of $\underline{\Omega}$ can be simplified by invoking central place assumptions about which towns are likely to receive significant levels of export income from other towns. The system of relationships will tend to be hierarchical with income flowing from the smaller urban places to the larger.

In both the studies outlined above it was important to specify the spatial mechanisms that "bound together" the set of sites and hence imposed an ordering on the spatial system. Different orderings can be compared by fitting the models with $\underline{W}$ matrices chosen to reflect different interaction hypotheses (Anselin 1986). In terms of the variables included most models will be "hybrid" that is, will contain both a traditional regression structure and spatialized variables. In some cases these spatial effects will arise from a small number of variables and can be explicitly stated within the model (as exemplified in this section and where the independent variables are spatially lagged). In other cases where the sources of spatial dependence may be large in number and difficult to specify with any precision they may be subsumed within a statistical model for the error structure (as exemplified in section 4). Doreian (1980, 1981) provides examples of both types of approaches applied to the same problem.

The statistical theory for estimating the parameters of these models and the large sample theory for inferential testing is extensively reviewed in Upton and Fingleton (1985).

6. CONCLUSION

This paper has surveyed some of the main areas of application of spatial

statistics in human geography. Not only is this area of relevance to general problems in univariate and multivariate data analysis it is also of importance in that a number of geographical models involve a spatial specification so that methods are needed to estimate and test for spatial relationships in a set of data. We have only touched the surface of the latter area for in addition to the competition and transfer mechanisms described in section 5 there is the wider class of spatial diffusion processes. These processes underly many observed geographic distributions involving information transfer (in the case of adoption patterns) or contagion processes in the case of the spread of diseases (Cliff et al 1975).

We have also not explored the relevance of spatial statistical theory in the setting up of computer based data storage, data retrieval, data manipulation and display systems (Geographic Information Systems) or in the analysis of remotely sensed data. The need to recognise the importance of spatial statistics in both areas has recently been emphasized in the publication of the Chorley Report (1987). These and other issues have been reviewed elsewhere (Haining 1987c). In conclusion however, it is important to emphasize that these methods are of general relevance in all areas of the social sciences concerned with the analysis of spatially referenced data.

BIBLIOGRAPHIE

ANCOT J.P., PAELINCK J.H.P., "Quelques éléments de l'économétrie spatial des prix du sol" in *Analyse spatiale et utilisation du sol*, J.M. Juriot ed., Université de Dijon, l'Economie du Centre Est, n°1, 1981.

ANSELIN L., "Non nested tests on the weight structure in spatial autoregressive models; some Monte Carlo results", *Journal of Regional Science*, 1986, 267-284.

BENNETT R.J., GRIFFITH D.A., HAINING R.P., "Statistical analysis of spatial data in the presence of missing observations : an application to census analysis", 1987 (forthcoming).

BESAG J., "Spatial interaction and the statistical analysis of lattice systems", *Jo. Royal Statistical Society*, 36, Series B, 1974, 192-225.

BESAG J., "The statistical analysis of non-lattice data", *The Statistician*, 24, 3, 1975, 179-195.

BESAG J., "On the statistical analysis of dirty pictures", *Jo. Royal Statistical Society*, 48, Series B, 1986, 259-302.

BIVAND R., "A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations", *Quaestiones Geographicae*, 6, 1980,

5-10.

Lord CHORLEY, *Handling Geographic Information : Report of the Committee of enquiry chaired by Lord Chorley*, Dept. of Environment HMSO (208pp.), 1987.

CHORLEY R.J., HAGGETT P., "Trend surface mapping in Geographical Research", *Transactions of the Institute of British Geographers*, 37, 1965, 47-67.

CLIFF A.D., HAGGETT P., ORD J.K., BASSETT K., DAVIES R., *Elements of Spatial Structure*, Cambridge, CUP, 1975 (258 pp.)

CLIFF A.D., ORD J.K., *Spatial Autocorrelation*, London, Pion, 1973 (178 pp.)

CLIFF A.D., ORD J.K., *Spatial Processes*, London, Pion, 1981.

CLIFFORD P., RICHARDSON S., "Testing the association between two spatial processes", *Statistics and Decisions*, Supplement Issue n° 2, 1985, 155-160.

DOREIAN, "Linear models with spatially distributed data", *Sociological Methods & Research*, 9, 1980, 29-60.

DOREIAN, "Estimating linear models with spatially distributed data", in *Sociological Methodology 1980*, Leinhardts ed. , San Francisco, Jossey Bass, 1981, 359-388.

GLICK B.J., "A spatial rank order correlation measure", *Geographical Analysis*, 14, 1982, 177-181.

HAGGETT P., "The edges of space", in *European Progress in Spatial Analysis*, R.J. Bennett ed., London, Pion, 1981, 51-70.

HAINING R.P., "A spatial model for High Plains agriculture", *Annals, Association American Geographers*, 68, 1978, 493-504.

HAINING R.P., "Intra-regional estimation of central place population parameters", *Journal of Regional Science*, 20, 3, 1980, 365-375.

HAINING R.P., "Spatial Interdependencies in population distributions : a study in univariate map analysis. 1. Rural Population densities", *Environment and planning*, A, 13, 1981, 65-84.

HAINING R.P., "Testing a spatial interacting markets hypothesis", *Review of Economics and Statistics*, 66, 1984, 576-583.

HAINING R.P., "Intraurban retail price competition : corporate & neighbourhood aspects of spatial price variation", in *Spatial pricing and differentiated markets*, G. Norman ed., London, Pion, 1986, 144-164.

HAINING R.P., "Trend surface models with regional and local scales of variation with an application to aerial survey data", *Technometrics*, 1987 (forthcoming).

HAINING R.P., "Small area aggregate income models : theory and methods with an application to urban and rural income data for Pennsylvania", *Regional Studies*, 1987 (forthcoming).

HAINING R.P., "Geography and spatial statistics : current positions, future developments", in *Re-modelling Geography*, W. MacMillan ed., 1987(c), (forthcoming).

HAINING R.P., "Estimating spatial means with an application to remotely sensed data", 1987(d). Communication in *Statistics : Theory and Methods* (forthcoming).

HEPPLE L., "A maximum likelihood model for econometric estimation with spatial series", in *Theory and Practice in Regional Science*, I. Masser ed., London, Pion, 1976, 90-104.

HEPPLE L., "Bayesian analysis of the linear model with spatial dependence", in *Exploratory and Explanatory Statistical Analysis of Spatial Data*, Bartels C.P.A. & Ketellapper R.H. ed., The Hague, Martinus Nijhoff, 1979, 179-199.

HUBERT L.J., GOLLEDGE R.G., "Comparing rectangular data matrices", *Environment & Planning*, A, 14, 1982, 1087-1095.

JOURNEL A.G., HUIJBREGTS Ch.J., *Mining Geostatistic*, London, 1978, 600 pp.

MARDIA K.V., MARSHALL R.J., "Maximum likelihood estimation of models for residual covariance in spatial regression", *Biometrika*, 71, 1984, 135-146.

MATHERON G., *The theory of Regionalized Variables and its Applications*, *Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau*, n° 5, 1971.

MIRON J., "Spatial autocorrelation in regression analysis : a beginners guide", in *Spatial Statistics and Models*, G.L. Gaile and C.J. Willmott eds., Dordrecht, Reidel, 1984.

ORD J.K., "Estimation methods for models of spatial interaction", *Jo. American Statistical Association*, 70, 1975, 120-126.

PAELINCK J.H.P., KLAASSEN L.H., *Spatial Econometrics*, Farnborough, Saxon House, 1979.

RICHARDSON S., HEMON D., "On the variance of the sample correlation between two independent lattice processes", *Journal of Applied Probability*, 18, 1981, 943-948.

RICHARDSON S., HEMON D., "Autocorrelation spatiale : ses conséquences sur la corrélation empirique de deux processus spatiaux", *Revue de Statistique Appliquée*, 30, 1982, 41-51.

RIPLEY B., *Spatial Statistics*, New York, Wiley, 1981.

SIBERT J.L., *Spatial autocorrelation and the optimal prediction of assessed values*, Ann Arbor, Michigan Geographical Publications n° 14, 1975.

TIEBOUT C.M., "Community income multipliers : a population growth model", *Journal of Regional Science*, 2, 1960, 75-84.

TJOSTHEIM D., "A measure of association for spatial variables", *Biometrika*, 65, 1978, 109-114.

TOBLER W., "The spectrum of U.S. 40 ", *Papers & Proceedings of the Regional Science Association*, 23, 1969, 45-52.

UPTON G., FINGLETON B., *Spatial Data analysis by example Vol. 1*, London, Wiley, 1985, 410 pp.

WHITTLE P., "On stationary processes in the plane", *Biometrika*, 41, 1954, 434-449.