A review of spatial statistical techniques for location studies*

Roger Bivand
Department of Geography
Norwegian School of Economics and Business Administration

August 1998

Abstract

While the new economic geography of trade and location has, understandably enough, concentrated on developing models of stylised relationships, it now seems that a review of some techniques which may be applied in empirical testing could prove useful. It is this task that will be approached here, conditioned by the advances taking place in new economic geography on the one hand, and in spatial data analysis on the other.

Spatial data analysis ranges from the visualization and exploration of spatial data, through spatial statistics to spatial econometrics. The techniques involved are intended to explore for and demonstrate the presence of dependence between observations in space. Typically, observations are classified into three broad types: fields or surfaces with values at least theoretically observable over the whole study area, as in geostatistics, point patterns representing the occurrence of an observation, such as reported cases in epidemiology, and finally lattice observations, where attribute values adhere to a tesselation of the study area. This last form has much in common with time series studies, and shares a number of key testing techniques with econometrics.

The paper reviews chosen techniques which can be applied in new economic geography. Point patterns, for instance, can be readily used to attempt to detect clustering. Lattice observations are used in the study of dynamic externalities, and consequently the effects of testing hypotheses based on spatial series should be examined. Finally, attention will be drawn to problems arising from spatial non-stationarity, when causal relationships may vary across space, and from the modifiable areal unit problem, when test results are influenced by the choice of spatial aggregation employed.

1 Introduction

The relationships between knowledge, especially conceptual knowledge, scientific disciplines, and empirical observations are often far from simple. Insight and intuition play an important role in leading to new conclusions. A striking example of the significance of location in such intuition is the mapping of the locations of cholera deaths in London in 1854 by Dr John Snow, indicating that proximity to the Broad Street water pump could be important. By disabling the pump, Dr Snow ended the epidemic (cf. Tufte, 1997). Few of us will be able to make such dramatic interventions just by mapping our data, but where data are located at geographical coordinates, it seems unfortunate not to examine the possibility that local dependence may be part of the story.

Spatial statistics span many disciplines, with methods varying in relation to the specific research questions being addressed, whether predicting ore quality in mining, examining suspiciously high frequencies of disease events, or handling the vast data volumes being generated by GPS (global positioning system) and satellite remote sensing. A unique feature of spatial data is that geographical

*This paper has been prepared in connection with the CEPR symposium on New Issues in Trade and Location (2277), Lund, Sweden, 28–30 August, 1998.

location provides a key shared either exactly or approximately between data sets of different origins. Census data can be overlayed over patient or customer data; environmental data can be integrated with disease frequencies; problems which hitherto did not admit ready empirical testing are becoming approachable. Geographical information systems are contributing to the development and spread of spatial statistical methods, which have, largely since their inception, remained within narrow research confines, at least partly because they were seen as being computationally burdensome.

In this review, I will concentrate on indicating the kinds of research problems to which spatial statistical methods can be applied, with particular reference to trade and location where possible. It should be admitted that the number of such applications is as yet very limited, but this does not appear to be because there are no opportunities — rather it seems that Krugman's argument about lack of mutual acquaintance also applies here (1995). Further, few of the econometric tools economists are furnished with provide suitable estimation methods. Haining (1990) gives a broad general introduction to the field, supplemented by Hepple (1996) and Getis and Ord (1996). Three recent surveys, including available software, are Levine (1996), Gatrell and Bailey (1996), and Bivand (forthcoming).

Having examined research traditions in trade and location in relation to the testing of models against empirical observations, basic issues in spatial statistics will be discussed, focusing on how the relationships between locations are expressed. We move next to the analysis of point patterns and fields. Most of the paper deals with lattice data typical of social science research problems. Starting from the exploratory analysis of spatial data — also a vital stage in point pattern analysis and geostatistics, global measures of spatial association are presented before the most recent work on local indicators is reviewed. Attention is also drawn to the modifiable areal unit problem often present in the analysis of lattice data. Finally, we turn to spatial econometrics, firstly the detection of spatial dependency in estimation results based on the assumption that the mutual location of observations is without importance, and secondly the explicit modelling of this dependence. This section is concluded by a discussion of a method of geographical weighting, providing a way of revealing non-stationarity in spatial data under analysis.

1.1 Research traditions

Many of the points taken by Krugman (1995) in his account of the development of economic geography and regional science are well taken, and deserve to be received with more grace than has appeared in replies so far. Economics does not however share the institutional contexts and research traditions of geography, which, like economics, has both a disciplinary core and "flavours" extending in many directions. Several of these are manifestly present within this review, particularly medical and physical geography. At present, the clear focus of many quantitative and applied geographers is on geographical information systems and collaboration with disciplines like computer science and surveying. In spatial statistics, the key breakthrough occurred in 1973, with the publication of Cliff and Ord's *Spatial Autocorrelation.* Cliff has deepened his concern with epidemiological modelling in geography (Cliff and Haggett, 1996), while Ord, a statistician of note[1], works from time to time with geographers such as Getis (Getis and Ord, 1996).

An intelligent description of the current setting and research traditions of quantitative geography has been written by Hepple (1998) in reply to an aggressive social-theoretic attack, in which all contact with correlation and regression is condemned for the links between Galton and Pearson and late nineteenth century eugenics. I feel that it is worth noting Hepple's approach. He is careful to express understanding for the criticism advanced, and proceeds to use the same methods as the social theorists in order to demonstrate that, with a more contextual reading admitting additional information from the period in question, one would have found that opinions were also divided. In the case of regression,

[1] J. K. Ord joined in the work of updating *Kendall's advanced theory of statistics* in the late 1970's, and has been involved in both the 4th and 5th editions.

Hepple advances a persuasive case for arguing that Yule, studying what we would now term social exclusion, made a more important contribution, and that consequently it would be premature to condemn quantitative methods as such on the basis of just some of their associations.

Indeed, it will be useful for our present discussion to cite Yule (after Hepple, 1998, p. 279[2]):

> The investigation of causal relationships between economic phenomena presents many problems of peculiar difficulty, and offers many opportunities for fallacious conclusions. Since the statistician can seldom or never make experiments for himself, he has to accept the data of daily experience, and discuss as best he can the relations of a whole group of changes; he cannot, like the physicist, narrow down the issue to the effect of one variation at a time. The problem of statistics are in this sense far more complex than the problems of physics.

Before we proceed to take up key issues raised in using the spatial 'data of daily experience', a few words on a very few selected examples of empirical work in trade and location. The geography of innovation in the context of knowledge spillovers is a research area with substantial interest, but where opportunities for interaction with spatial statistics do not yet seem to have been exploited sufficiently (Jaffe, Trajtenberg and Henderson, 1993, Audretsch and Feldman, 1996). In work on dynamic externalities and growth in cities, Henderson (1997, p. 455) does admit that the residuals "may be correlated for all counties within a metropolitan area", and uses a simple ad hoc diagnostic. The study of the determinants of economic growth using cross-sectional regressions (Sala-i-Martin, 1994, Barro, 1997), despite technical sophistication, does not seem to have opened for the testing of hypotheses concerning residual or structural neighbourhood effects. In conclusion, attention can be fruitfully drawn to the work of Francophobe economists; Thisse (1997) sums up lucidly the indeterminacy of regional bounding, showing how processes like spillover render the construction of entities for empirical purposes problematic — we will return to this issue again as the modifiable areal unit problem.

2 Basic issues in spatial statistics

Since observations of spatial data are as unlikely to be independent as observations on time series, it is perhaps surprising that not more use has been made of this source of information. With an adequate choice of explanatory variables, this spatial dependence may be readily drawn into a model, and cease to be a nuisance. However, spatial dependence is not necessarily just a nuisance, but may help us to capture important facets of the realities of economic processes (cf. Hendry and Mizon, 1978). The literature on spatial statistics is substantial (see Cliff and Ord, 1973, 1981, Ripley, 1981, Upton and Fingleton, 1985, Griffith, 1988, Anselin, 1988, Haining, 1990, and more recently Cressie, 1993, and Bailey and Gatrell, 1995). We will here give a brief introduction to some of the key issues.

In their now classic survey of problems in analysing spatial data, Duncan, Cuzzort, and Duncan express the focus of this study in the following way:

> Interest in areal distributions merges more or less imperceptibly into a concern with the 'spatial structure' of communities, economies, and societies. At the present time it is difficult to appreciate the magnitude of effort which was required to establish the concept of an economy or a society as a territorially organised system (1961, p. 16).

They continue to identify four perspectives on spatial differentiation, which they describe as: (a) chorographic interest in areal differentiation; (b) interest in areal distribution; (c) interest in spatial

[2] from Yule, G., 1897, "On the theory of correlation", *Journal of the Royal Statistical Society*, 60, p. 812.

structure; and (d) concern with the explanation of areal variation (page 19). They deserve credit for taking up the problems which spatial data pose for analysts of society, and of change in society. Since spatial data are neither the outcome of controlled experiments, nor do they result from random samples, it is clear that beyond mapping and informal inference from patterns, specific spatial statistical methods are required.

Data from which statistical inferences are to be drawn ought to fulfill a number of criteria, key among which is that they are independent of each other. The founders of statistics were keenly aware of the difficulties of making inferences from spatial and time series data. Student describes the problem in detail in a paper published over ninety years ago (1914), while the way in which Galton posed the problem is discussed below. In time series, we know that later observations may depend on earlier ones; this dependence is termed autocorrelation. First order autocorrelation is between the current observation and its immediate predecessor. The ordering of the data is clear, although the choice of temporal units does make a difference, for example hourly, daily, weekly or monthly data may display different forms. In the time series case, it is usual to speak of a temporal data generation process. This can be thought of as an unobservable curve, generated both in relation to its own previous values and in relation to the current and possibly previous values of other variables. If we observe it at discrete and regularly spaced intervals, we get time series data, from which we can try to estimate the underlying, unobservable curve.

Spatial data may be viewed as observations taken at discrete points on a surface, rather than a curve, since we are in two dimensions, not just the single dimension of time series. It is in this sense that we can speak of underlying, unobservable spatial data generation processes, about which we would like to infer. The inferences which we would like to be able to make are about these processes, which for a variety of reasons may not be directly observable. Using political behaviour as an example, we could seek to establish the identities of voters, hoping to link their ballot papers to their other characteristics, such as place of residence or birth, sex, age, occupation, etc. An exit poll could be used to achieve this, but then the focus would be on the individual level, rather than on the local, territorial, or ecological links. While we have to accept that we can not make inferences about individual behaviour from ecological data (Langbein and Lichtman 1978), it is often both necessary and relevant to study spatial data generating processes at the aggregate level. Aggregation in itself should not be avoided, not least because it often returns in one form or other as classifications used as explanatory variables related to cleavages, be it socio-occupational class, organisation, or some other structuring variable above the level of the individual.

Spatial aggregation brings with it a number of specific problems. The boundary effects at the edges of the study area are often impossible to control for. If we are concerned with reconstructing unobservable surfaces, then we are faced with the hypothetical question of whether the surface extends outside the study area, even though we have no observations (Haggett 1981). If we had possessed data from beyond the study area, would it alter our inferences about the shape of the surface at the edges of our study area?

In addition, the often arbitrary nature of the assignment of observation units to aggregates, known as the modifiable areal unit problem, has to be recognised. It has been demonstrated that there is a relationship between the coherence or simplicity of the process generating the surface we are trying to make inferences about, and the way in which the observations are aggregated (Openshaw and Taylor 1979). They separate the scale problem, where results change from less aggregated to more aggregated spatial units, from the aggregation problem caused by arbitrary choices made in zoning, that is assigning basic spatial units to contiguous zones. Zones in turn imply the contiguity of member units, while groups require no contiguity. Openshaw and Taylor were able to demonstrate that the interaction between spatial autocorrelation and the zoning procedure directly affected resulting statistics (1979, p. 142). It is quite clear that the results of analysis are dependent on the particular lattice of areal unit boundaries chosen, and that different results may be yielded by analyses using different boundaries. For this reason, units of observation may be termed zones, to show that they have been subject to a process of aggregation from basic spatial units for which data may often not be available.

Finally, the non-stationarity of variance across the study area is a problem analogous to that faced in many studies using statistical inference. We recall that regression, for example, assumes that the variance of the error term should be constant, and not vary with the independent variable. In the time series case, we can say that the series is stationary if it has a constant mean, and fluctuates about that mean with a constant variance. The mean may of course be a residual after the removal of estimated structural features of the curve underlying the observations. In order to make inferences about the curve, it is important that the variance about the estimate should not vary in time. In the same way, with spatial data we should be aware of problems that arise in inference if variance about the estimated surface is not constant over the whole study area.

Haining (1990, p. 22-26) provides a useful discussion of many of the issues involved in inferring from spatial data. If it is possible that observations being treated as independent in fact derive from a shared ancestor, then they will not contribute separate degrees of freedom to the formal test used for inference, or to the judgement involved in the drawing of informal conclusions. Further light is thrown on the difficulties involved in reaching substantive conclusions by Haining (1991), in a discussion of the Clifford-Richardson adjustment of the "effective" sample size for bivariate correlation.There is a clear link between the method suggested by Haining (1991, page 215), for the calculation of the relevant adjustment, and the family of distance statistics summarised by Getis and Ord (1996). Both the adjustment method and distance statistics rely on the explication of the correlation structure at varying distances.

Summing up, we are often faced by non-experimental data for sites or zones, which we would like to analyse. Abstracting from zones to simplify the argument, we are in an inherently multivariate situation, where each site stands in a relationship to every other one. We are faced with a set of probably non-independent random variables $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{R}^2\}$, commonly referred to as a spatial stochastic process, and where $\mathbf{s}$ are the point location coordinates. A typical data set then consists of observed $y(\mathbf{s}_i)$, and is referred to as a realization of the spatial process. It is only a single observation from the joint probability distribution of the random variables $\{Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots\}$, from which little can be gleaned about the relationships between these sites, even given that we accept that they are reasonably representative in some sense (Bailey and Gatrell, 1995, p. 24–28).

On this basis, we will now proceed to review the component areas of spatial statistics, dealing in turn with point pattern analysis, geostatistics, and the analysis of lattice data.

3 The analysis of point patterns

Point pattern analysis is concerned with the location of events, and with answering questions about the distribution of those locations, specifically whether they are clustered, randomly or regularly distributed. Point pattern analysis is very sensitive to the definition of the study area, since a regularly distributed pattern can be made to seem clustered by including large margins within the study area. Measures are also subject to boundary corrections, and most often study area boundaries have to be defined as convex polygons over the study area, or in the simplest form as rectangles bounding the points under analysis. It is of course always important to plot the events to detect outliers visually, together with the boundaries being applied (Bailey and Gatrell, 1995, Cressie, 1993).

The simplest way of exploring point pattern data is by examining a two-dimensional frequency distribution of counts within equal-area units imposed on the study area, giving an impression of how the intensity of the point process varies; this can be extended to kernel estimation. Nearest neighbour distances are also used to analyse intensity. Intensity in this sense is a first order property, the mean number of events per unit area at point $\mathbf{s}$. Spatial dependence is captured by the second order properties of a spatial point process, which involve the relationship between numbers of events in pairs of arbitrary areas within the chosen study area: $\gamma(\mathbf{s}_i, \mathbf{s}_j) = \gamma(\mathbf{s}_i - \mathbf{s}_j) = \gamma(\mathbf{h})$. For a stationary process, this

relationship depends on the distance and direction between the pair of areas; when the relationship depends on distance alone, the process is termed isotropic.

Having an empirical data set is not sufficient to test for divergences from randomness. In general, tests are conducted against a standard model for complete spatial randomness following a homogeneous Poisson process over the study area. This implies that any of the events could have occurred anywhere in the study area, and that the locations of the events are mutually independent. This is enough for a start, but quickly encounters difficulties, when the underlying control distribution is not homogeneous across the study area. Further, one may wish to test hypotheses that the incidence of events is raised at or near given locations. Both of these issues have attracted substantial contributions in the past decade, and methods are now available for testing point patterns against hypotheses of non-randomness in relation to a second control variable with a varying spatial distribution (Cuzick and Edwards, 1990, Diggle, 1990, Diggle and Chetwynd, 1991, Diggle and Rowlingson, 1994, Kingham, Gatrell and Rowlingson, 1995, Gatrell et al., 1996).

These developments have led to empirical work using point patterns for cases — observed events — and controls — for the underlying non-homogeneous distribution. In this framework, K functions are defined for a labelled stationary isotropic point process for case-case, control-control, and case-control pairs for distances up to an arbitrary maximum, and the difference is calculated between the case-case and control-control pairs for the chosen distance steps. An confidence interval envelope can be constructed around the null of no difference, permitting the analyst to detect at which distances significant differences occur between the distances between cases and between controls. These methods have been employed by Jones, Langford and Bentham (1996) to explore the outcomes of road accidents, and within the field of location by Sweeney and Feser (1998) to examine small manufacturing business location patterns in North Carolina. They find conclusive evidence of plants from 8–49 employees, with 8–17 employee plants displaying clustered locations at ranges up to 15 kilometres, while the larger 18–49 employee plants clustered at all spatial scales within the bound calculated. Large plants with over 205 employees were found to seek dispersed locations significantly.

4 Geostatistics

Geostatistical methods most often start from observations at points of single or multiple attributes, and are concerned with their statistical interpolation to a field or continuous surface assumed to extend across the whole study area. It is of course possible to interpolate in a deterministic way, or to use polynomial regression on the site coordinate values to predict a trend surface, but these methods do not give the degree of statistical control to be had from variogram analysis and subsequently modelling by kriging. Geostatistical methods are also subject to a variant of the modifiable areal unit problem, known as the change of support problem (Cressie, 1996); although a surface is assumed to exist throughout the study area, it is not feasible to gather data at all of the $\mathbf{s}$ in the study area, or to know on the basis of the sample points how they represent the study area. Geologists are also vitally interested in finding anomalies, perhaps similar to clusters; the same applies to environmental scientists examining the distribution of radioactive isotopes, who are concerned to locate “hot-spots”.

In practice a sample data set may be treated for systematic variation in the first two moments before geostatistical analysis begins. The next step is to use variograms for exploring spatial variability between all pairs of points a specified distance apart. Measures are taken across the whole map, and can be taken assuming isotropy, or in a chosen direction. The chief sources for exploratory variography and variogram modelling are Cressie (1993), Isaaks and Srivastava (1989), and Deutsch and Journel (1992). Semivariogram analysis and modelling has been attracting growing attention in the spatial analysis of data from others than the earth sciences over recent years. Among other examples, geostatistical methods have been employed in medical as well as physical geography (Oliver and Webster 1986, Webster, Oliver, Muir and Mann 1994).

The distance measure **h** is a vector expressing distance and direction, within specified tolerances, and thus has a natural head and tail. The head and tail variables can be the same, but can differ; in such bivariate cases causal effect is manifested in the direction and at the distance specified. It is assumed that the same dependency relationships between locations will be manifest irrespective of placing in the study area, although the relations may be anisotropic. The classical semivariance measure is:

$$\gamma(\mathbf{h}) = \frac{1}{2N\mathbf{h}} \sum_{i=1}^{N(\mathbf{h})} (x_i - y_i)^2$$

where $N(\mathbf{h})$ is the number of pairs fulfilling relationship **h**, x_i is the tail value, and y_i is the head value. The covariogram is similarly defined:

$$C(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} x_i y_i - \left(\frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} x_i \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} y_i \right)$$

The semivariance is thus the sum of squared differences between pairs of values at distance **h**, divided by twice the number of such pairs. This is analogous to the Geary statistic, while the covariance corresponds to a distance-banded Moran's *I* statistic (described in sections on the analysis of lattice data below). Many further semivariance estimators are available, providing robustness to outliers, and perhaps a better separation of the structured aspect related to the overall distribution of the phenomenon from the often erratic local behaviour of the phenomena.

Modelling is derived from the fitting of one or more of a family of functions to the observed curve, adjusted with respect to a number of parameters. The principle advantage of using geostatistical methods is yielded when the resultant models are used for prediction to other locations within the study area, using both results of trend analyses, and of local dependencies. These result in surfaces of fitted values, perhaps plotted over a regular grid and contoured, and more importantly surfaces of variances, permitting confidence intervals to be constructed around model predictions over the study area. For environmental scientists in general, and mining geologists in particular, attempting to squeeze the most information possible out of each sample core, these methods have proved to be of considerable value. Social science applications are limited chiefly because there are relatively few phenomena which can reasonably be supposed to exist as surfaces of this nature, although by the use of analogy, one might relax this limitation. There are some parallels between work in geostatistics and the treatment of non-stationarity using geographically weighted regression discussed below.

5 Exploratory spatial data analysis and lattice data

In this section, we will find that applications of spatial statistics to trade and location become more realistic, not least because the methods used and the underlying dependency structures appear more like econometrics in the time series domain. While the methods discussed above are related to those for more typical social science lattice data, they are perhaps more similar to the application of time series methods in engineering or the physical sciences — the kinds of processes economists and human geographers are involved in studying are only seldom events of the point pattern kind, or surfaces analysed in geostatistics. The understanding, however, of stationarity and isotropy that they bring with them does however carry over into studies of lattice data, including attempts to detect the spatial range at which neighbourhood effects, spillovers, make themselves felt. Two recent surveys covering the area of exploratory spatial data analysis explicitly are by Bivand (forthcoming) and D. Unwin (1996).

5.1 Visualization

D. Unwin (1996) specifically focuses on visualization as a necessary first step in all spatial data analysis, simply because the position of particular attribute values on a map induces associative processes in the analyst, drawing upon analogies, possible prior information, or memory (for instance of possible sources of data error). In geostatistical analysis, Haslett et al. (1991) and Cook et al. (1996, 1997) have introduced linked variogram cloud plots, displaying the values of the squared differences of the pair of head and tail observations $(x_i - y_i)^2$ in one window, and the specific tail to head line on a map in a second window. By moving a pointing device about the variogram cloud plot, the analyst is able to see where on the map display the chosen pairs are located. In general, linked plot technology for dynamic data visualization is becoming an important part of the modern statistical toolbox, perhaps exemplified by XLispStat (Tierney, 1990) and XGobi (Buja, Cook and Swayne, 1996), neither of which is specifically designed for spatial data, but where both have been successfully utilised (Cook et al., 1996, 1997, Brunsdon and Charlton, 1996). Further examples of visualization techniques for socio-economic data are given by A. Unwin (1996).

5.2 Global measures of spatial association

At this point, a number of definitions and explanations of standard spatial statistical notations are required. A measure of spatial dependence is bound to make some assumptions about the underlying data generation process or processes. Among the assumptions that have been used in studies of autocorrelation, the one implying least about our prior knowledge of relationships between observations for spatial units, say point sites, or bounded zones exhaustively dividing up the study area, is based on contiguity. It is not usual to be able to estimate these relationships from data, involving as they do $N^2 - N$ interactions, omitting those within zones; they are not the same as zonal fixed effects either, although the elimination of such fixed effects in panel studies can alter the ways in which interaction may appear.

Cliff and Ord (1973, p. 11–13) provide the initial formalization of the relationships as a generalized weighting matrix, most usually termed **W**. The most recent systematization, reviewing the Markovian properties of some weighting matrices is given by Bavaud (1998). In a recent study reviewing the use of different forms of weighting matrices, Griffith (1995) has demonstrated that a parsimonious specification of the relationships between observations is to be preferred to one making assumptions about say distance decay. Brett and Pinkse (1997) also note differences in inference which can occur in using distance bands and contiguities, which they call “Hotelling neighbours” for obvious reasons.

It is usual in the literature to define the contiguity relation in terms of sets $N_{(i)}$ of neighbours of zone or site i. These are coded in the form of a weights matrix **W**, with a zero diagonal, and the off-diagonal non-zero elements often scaled to sum to unity in each row (a.k.a. standardized weights matrices), with typical elements:

$$w_{ij} = \frac{c_{ij}}{\sum_{j=1}^{N} c_{ij}}$$

where $c_{ij} = 1$ if i is linked to j and $c_{ij} = 0$ otherwise. This implies no use of other information than that of neighbourhood set membership. Set membership may be defined on the basis of shared boundaries, of centroids lying within distance bands, or other a priori grounds.

Figure 1A shows the way in which the sets of contiguous neighbours of each zone are constructed; in Figure 1B, neighbours are defined within a fixed distance from the zone in question. In table form, the sets of neighbours for selected zones are shown in Table 1.

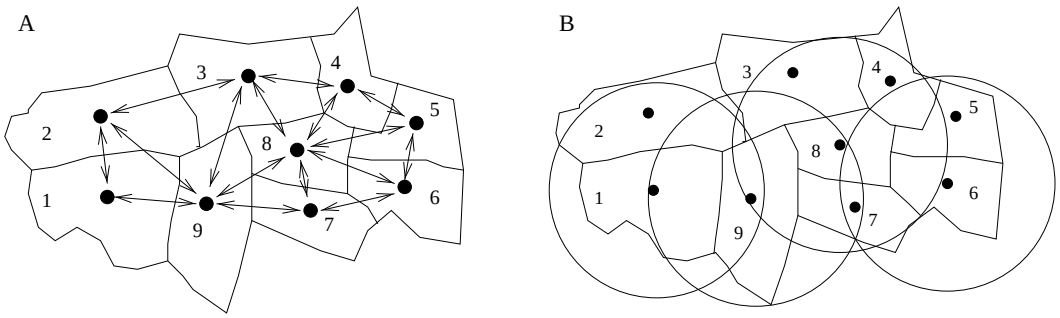

Figure 1: Lattices of irregular polygon zones and point sites.

Table 1: Neighbourhood sets for lattices shown in Figure 1.

Zone	A: contiguity		B: distance	
	number	neighbours	number	neighbours
1	2	(2, 9)	2	(2, 9)
...				
6	3	(5, 7, 8)	2	(5, 7)
...				
8	6	(3, 4, 5, 6, 7, 9)	4	(3, 4, 7, 9)
9	5	(1, 2, 3, 7, 8)	3	(1, 7, 8)

As Getis and Ord point out (1992, p. 190), there are good reasons for examining patterns of spatial dependence at a more local scale. If we do not have good reason to suppose that the process in question is spatially stationary, it seems natural to apply distance-based tests to the observed spatial series. For use with distance statistics, one defines a symmetric one/zero spatial weighting matrix using the distance between the coordinates of a point associated with the observations. The choice of point for non-site series is not arbitrary, nor is the choice of the distance metric. Here the administrative centres of the observation units have been taken as adequately representing the location of the observation. Distance has been assumed to be the simple Euclidean distance between points, ignoring barriers and other factors. Distance has further been banded on the basis of the frequencies of interpoint distances, and the furthest nearest neighbour distance as shown in Figure 2. A typical element of the non-standardized spatial weight matrix $\mathbf{C}(d)$ for distance d is defined as:

$$c_{ij}(d) = \begin{cases} 1 & \text{if hypot}(i,j) \leq d, i \neq j \\ 0 & \text{otherwise} \end{cases}$$

and $\text{hypot}(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

The extent to which results are affected by the choice of points representing zones, and the choice of a simple representation of distance is unknown. Distance banded spatial weight matrices may be stored in the same fashion as contiguity matrices, and may also be represented as sliced increments, again reducing storage requirements.

In Figure 2A, the nearest neighbours of each zone are shown. It is zone 9 that has the furthest nearest neighbour distance, at 50 km from zone 7, while zone 3 is 39 km from zone 8. Figure 2B illustrates the use of distance bands, at 30, 60, 90, and 120 km. Table 2 shows the incremental neighbourhood sets for zone 8 for these bands. If zones were permitted to be their own neighbours, then zone 8 would belong to the set of neighbours for band 1.

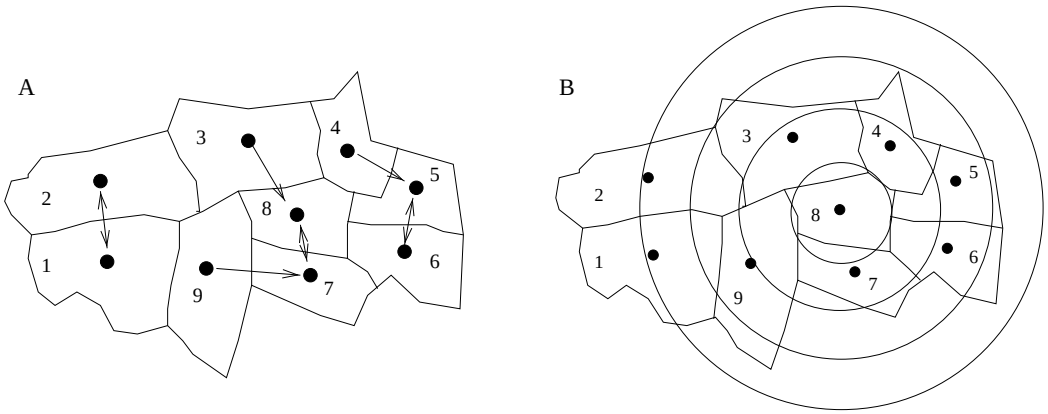

Figure 2: Nearest neighbours and distance bands.

Table 2: The incremental neighbourhood sets of zone 8 (Figure 2B).

Band	Distance	Number	Neighbours
1	< 30	0	
2	30 - 60	3	(3, 4, 7)
3	60 - 90	3	(5, 6, 9)
4	90 - 120	2	(1, 2)

Table 3: Spatial lag values for zone 8.

Zone 8	Number	Neighbours	Lagged value
	6	(3, 4, 5, 6, 7, 9)	
Sum	6	(15.0, 17.0, 19.0, 18.0, 17.0, 14.0)	100.00
Average	6	Each contributes 1/6	16.67

We can exemplify the spatial lag using the neighbourhood set for zone 8 from Figure 1 and Table 1. If the set of observations from all the nine zones is (10.0, 12.0, 15.0, 17.0, 19.0, 18.0, 17.0, 16.0, 14.0), then we can see from Table 3 how the spatially lagged value is calculated as a sum or an average of the values of six neighbours of zone 8, in this case. The average lagged value of 16.67 corresponds closely to the observed value of 16.0.

Using these constructions, we can define two commonly used global measures of spatial autocorrelation (Cliff and Ord, 1973, p. 12, Haining, 1990, p. 230), Moran's I:

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} z_i z_j}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \sum_{i=1}^{N} z_i^2}$$

taking differences from the mean: $z_i = x_i - \bar{x}$, and the Geary coefficient:

$$C = \frac{(N-1)}{2(\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij})} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (x_i - x_j)^2}{\sum_{i=1}^{N} z_i^2}.$$

In addition, mention should be made of the general class of cross product statistics due to Mantel (1967), and developed by Hubert et al. (1981):

$$\Gamma = \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} \xi_{ij}.$$

If we set $\omega_{ij} = w_i j$, we can express the Moran coefficient as $\xi_{ij} = (x_i - \bar{x})(x_j - \bar{x})$, while the Geary measure takes the form: $\xi_{ij} = (x_i - x_j)^2$. Γ yields a general framework for the development of additional measures, including space-time interaction and multivariate tests (Haining, 1990, p. 230–231).

The Moran and Geary coefficients may be tested using analytical expectations and variances (Cliff and Ord, 1973) based largely on the neighbourhood structure assumed in the spatial weighting matrix, and are asymptotically normally distributed. In addition to tests for interval scaled variables, there are also join-count statistics for nominal variables, based as the name suggests on counting the numbers of same-colour and different-colour joins between neighbours defined by the weighting scheme adopted. Lowell (1997) provides a review of these measures in the light of more recent developments. A study adapting Moran's I to heteroscedasticity has been conducted by Waldhör (1996), who is concerned with situations when testing the observed estimate of the statistic against a null in which any permutation of values to zones is not equally likely, an assumption underlying the analytical expectation and variance of the measure. Finally, new measures have been introduced by Sherman and Carlstein (1994) and Sherman (1996) using a method-of-moments solution using only the data at hand, and by Brett and Pinkse (1997) for spatial independence based on characteristic functions. The Moran statistic has been used in studies of prices in international trade by Aten (1996, 1997).

5.3 Local indicators of spatial association

While global measures permit us to test for spatial patterning over the whole study area, it may be the case that there is significant autocorrelation in only a smaller section, which is swamped in the context

of the whole. Both distance statistics (Getis and Ord, 1992, 1996, Ord and Getis, 1995), and the local indicators of spatial association derived by Anselin (1995, see also Getis and Ord, 1996), resemble passing a moving window across the data, and examining dependence within the chosen region for the site on which the window is centred. The specifications for the window can vary, using perhaps contiguity or distance at some spatial lag from the considered zone or point.

There are clear connections here both to the study of point patterns — although methods for boundary correction have not been specifically added to weighting matrix definitions yet — and to geostatistics, since these statistics have application to the exploration of non-homogeneities in relationships between locations across the study area. They are however subject to a correlation problem, that estimated values of the local indicator for neighbouring zones or sites will be correlated with each other because they are necessarily calculated from many of the same values, recalling that neighbouring placements of the moving window will most likely overlap. Ord and Getis (1995) provide suitable adjustments to critical values of the G_i and G_i^* statistics.

By extension from the global measure Γ presented above, Getis (1991, see also Getis and Ord, 1996, Anselin, 1995) defines:

$$\Gamma_i = \sum_{j=1}^{N} \omega_{ij}\xi_{ij},$$

where Γ_i is the measure for location i defined in terms of the weighting matrix with elements ω_{ij}, and ξ_{ij} captures the interaction between the attribute values at locations i and j. Getis and Ord (1996) define six different measures, the local Moran I_i: $\xi_{ij} = (x_i - \bar{x})(x_j - \bar{x})$, three local Geary-type statistics (C_i, K_{1i}, and K_{2i}) with $\xi_{ij} = (x_i - x_j)^2$, and the G_i and G_i^* statistics with $\xi_{ij} = (x_j)$ and $\xi_{ij} = (x_i + x_j)$ respectively (G_i and G_i^* differ in that G_i^* includes the attribute value at location i as well as those at $j \in N_{(i)}$). G_i and G_i^* have been shown to be asymptotically normally distributed as the number of neighbours of location i, $j \in N_{(i)}$, increases, for instance by increasing the radius d around i used to define the weighting matrix.

The uses to which local statistics have been put are to identify "hot-spots", to assess stationarity prior to the use of methods assuming that the data do conform to this assumption, and other checks for heterogeneity in the data series. A typical application is to plot the estimates values of a local statistic with increasing distance from a selected location i, perhaps also controlling for direction (Getis and Ord, 1996, Bivand, 1997). In addition, Anselin (1996) has suggested that a plot of x_i against its spatial lag $\sum_j w_{ij}x_j$, termed a Moran scatterplot, particularly used with dynamic linked visualization, may assist in revealing local patterning.

Examples of the application of local statistics in relation to topics in economic geography are O'Loughlin and Anselin (1996), examining trade bloc formation — challenging assertions made by Krugman, and by Barkley et al. (1995) and Bao and Henry (1996) in exploring the use of local indicators in assessing the appropriateness of definitions of functional economic areas. Talen and Anselin (1998) have also used these methods to evaluate the measures used to define accessibility to public playgrounds, a study in the equity of urban service delivery.

5.4 The Modifiable Areal Unit Problem (MAUP)

Having outlined the MAUP above, it remains here to indicate progress in addressing and in part resolving the issues involved. Arbia (1989) made a major contribution by studying in depth a range of links between the presence of spatial autocorrelation and the MAUP; until that time most analysts had chosen to sidestep Openshaw and Taylor's (1979) potentially devastating finding that the results of statistical analysis of data for spatial zones could be varied at will by changing the zonal boundaries. The problem includes two parts, the problem of scale, involving the aggregation of smaller units into

larger ones, and the problem of alternative allocations of component spatial units to zones, also known as gerrymandering.

A further positive contribution was made by Fotheringham and Wong (1991), followed up by Amrhein (1995), Fisher and Langford (1995), Amrhein and Reynolds (1996), and Morphet (1997). Openshaw (1996) summaries many of the technologies now available for choosing zoning systems to optimize results. Perhaps the most active group of recent publications has resulted from collaboration between social statisticians experienced in complex survey design and geographers, including Holt, Steel, Tranmer, and Wrigley (1996) and Holt, Steel, and Tranmer (1996), and Wrigley et al. (1996). Focusing closely on the scale and zoning effects, they conclude that the use of well chosen grouping variables to adjust the area-level results may yield reliable estimates of underlying individual-level relationships, thus providing at least a partial solution to the MAUP with respect to the "ecological fallacy", the drawing of individual-level inferences based on area-level analyses.

6 Spatial econometrics and lattice data

Estimation methods for models using lattice data and taking spatial dependence into account are as mature as global statistics for spatial autocorrelation (Ord, 1975, Hepple, 1976); the form of model most commonly used is known as the simultaneous autoregression (SAR). Ten years have now passed since Anselin and Griffith (1988) surveyed the regional science and economic geography literature to see how far these methods were being applied to data sets for which they should have been suited. The low penetration they reported seemed related to the lack of access to these tools in standard statistical packages, addressed subsequently by Anselin and Hudak (1992), Griffith (1993), Bivand (1992), and others. The most substantial effect has been achieved by Anselin's "SpaceStat" program, permitting the estimation of most of the specification tests and models described in the literature (1995b).

Examples of the application of these methods by economists are Dubin's estimation of a hedonic regression with cross-section data (1988), an analysis of spatial patterns in household demand by Case (1991), and two detailed studies of fiscal policy interdependence between U.S. states (Case, Rosen and Hines, 1993, Besley and Case, 1995). In addition, mention can be made of some recent studies taking up location problems: Anselin, Varga and Acs (1997) challenge and refine Jaffe's conceptual framework for the analysis of local geographical spillovers between university research and high technology innovations, modifying previous conclusions. Bernat (1996) evaluates manufacturing and regional economic growth across U.S. states in relation to hypotheses based on Kaldor's laws. Bivand and Szymanski (1997) have investigated the attenuation of neighbourhood effects, suggested to stem from local yardstick competition, following the introduction of compulsory competitive tendering for refuse disposal services in English local authorities. A classic study on price autocorrelation in space is reported in Haining (1983, 1984). In all of these examples, the inclusion of information about the mutual location of the observations makes a difference to the conclusions drawn.

We will now present briefly the basic models of spatial econometrics. Assuming that the variance of the disturbance term is constant, we start from the standard linear regression model:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

where $\mathbf{y}$ is an $(N \times 1)$ vector of observations on a dependent variable taken at each of N locations, $\mathbf{X}$ is an $(N \times k)$ matrix of exogenous variables, β is an $(k \times 1)$ vector of parameters, and ε is an $(N \times 1)$ vector of disturbances. The two alternative forms of spatial dependence models are the spatial lag model:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\beta + \varepsilon.$$

and the spatial error model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \qquad \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \varepsilon,$$

where λ is a scalar spatial error parameter, and $\mathbf{u}$ is a spatially autocorrelated disturbance vector. These two models can also be related through the Common Factor model (see Burridge, 1981, Bivand, 1984). The use of the non-spatial linear model with spatial data is equivalent to assuming, in the above parameterisation, that $\rho = \lambda = 0$. The spatial lag and spatial error models can only be combined for estimation if the neighbourhood specifications, here the $\mathbf{W}$ matrices, of the lag and error components differ; for testing, however, the same matrix may be employed.

Dependence between observations in econometrics can stem both from a hypothesised data generation process, such as the kind presented above, and from omitted variable biases, possible even both simultaneously. The spatial lag model is clearly related to a distributed lag interpretation, in that the lagged dependent variable, $\mathbf{Wy}$, can be seen as equivalent to the sum of a power series of lagged independent variables stepping out across the map, with the impact of spillovers declining with successively higher powers of ρ. This may be termed a structural autoregressive relationship, and one would expect it to be based on economic processes. The alternate model presupposes a shared spatial process affecting all of the variables, and is perhaps more often to be interpreted as indicating missing variables.

6.1 Specification testing

A recent line of research in analysing spatial data, mainly associated with Luc Anselin, has focused on how to establish the characteristics of the dependence between observations, whether dependence can be demonstrated and how it ought to be represented. (see for instance Anselin, 1988b, 1990, Anselin and Rey, 1991, Anselin and Florax, 1995, Anselin et al. 1996). Burridge (1980, 1981) made the first attempt to extend the tests for regression misspecification given by Cliff and Ord (1973), using Lagrange multiplier techniques to derive simpler procedures. These have been followed up by Anselin and collaborators, and are now at a stage at which their use in all cases in which geographical cross-sectional data are being analysed should be expected.

A problem solved in Anselin et al. (1996) is that of tests for spatial lag and spatial error specifications being mutually contaminated by each other, that is the original LM test for non-zero ρ also responds to non-zero λ and vice-versa. The new tests take into account the possible non-zero value of the nuisance parameter, and appear to discriminate well between the two alternative forms. Results obtained by Bivand and Szymanski (1998) indicate that these refined LM tests are of considerable use in model specification, and that test results, drawn from OLS residuals from the initial model, are confirmed by likelihood ratio test results from maximum likelihood estimates of ρ and λ for the spatial lag and spatial error models respectively.

Work on global tests for mis-specification is continuing, with Tiefelsdorf and Boots (1995) and Hepple (1998b) arriving independently at exact distributions of Moran's I as a test statistic for regression residuals, using results on ratios of quadratic forms in normal variables. Tiefelsdorf and Boots have also extended their results to the local Moran's I_i statistic (1997).

6.2 Modelling spatially dependent data

Ord (1975) gives the Maximum Likelihood methods for estimating the spatial lag and spatial error SAR models; no satisfactory alternatives have been found subsequently, chiefly because of the important role of the Jacobian expressing the spatial transformation of either the dependent variable in the spatial lag model, or the disturbance in the spatial error model. Unlike the time series case, the logarithm of the

determinant of the $(N \times N)$ asymmetric matrix $(\mathbf{I} - \lambda\mathbf{W})$ or $(\mathbf{I} - \rho\mathbf{W})$ does not tend to zero with increasing sample size; it constrains the parameter values to their feasible range between the inverses of the smallest and largest eigenvalues of $\mathbf{W}$, since for positive autocorrelation, as $\rho \rightarrow 1$, $\ln|\mathbf{I} - \rho\mathbf{W}| \rightarrow -\infty$, and analogously for λ. The log-likelihood function for the spatial lag model is:

$$\ell(\beta, \rho, \sigma^2) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 + \ln|\mathbf{I} - \rho\mathbf{W}|$$
$$-\frac{1}{2\sigma^2}[\mathbf{y}'(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{I} - \rho\mathbf{W})\mathbf{y}]$$

and $\beta = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{y}$, where $\hat{\rho}$ is the ML estimate, and for the spatial error model:

$$\ell(\beta, \lambda, \sigma^2) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 + \ln|\mathbf{I} - \lambda\mathbf{W}|$$
$$-\frac{1}{2\sigma^2}[(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{I} - \lambda\mathbf{W})'(\mathbf{I} - \lambda\mathbf{W})(\mathbf{y} - \mathbf{X}\beta)]$$

To complete the model, the variance-covariance matrix of the parameters needs to be estimated. In many cases it is approximated numerically following non-linear optimization of the likelihood function, but SpaceStat derives its estimates of the asymptotic standard errors analytically (Anselin, 1995b, Anselin and Hudak, 1992). For larger N, this can take considerable time, requiring the inversion of an $N \times N$ matrix.

As Pace and Barry (1997, 1997b, 1997c) have conclusively demonstrated, a feasible solution to modelling situations with large N is to exploit the sparse nature of the spatial weighting matrix, both saving memory and making computation practical in reasonable time without supercomputer resources. They were able to compute results for a model of the median price of dwellings over all the 20,640 block groups in California from census data, improving the fit of the model over OLS results, halving the median absolute residual, finding a highly significant spatial lag coefficient estimate, and recording several significant sign changes among the independent variables (1997b). They also provide a profile likelihood solution to the calculation of coefficient estimate standard errors, avoiding the computation of the information matrix.

Hepple (1995, 1995b), LeSage and Pan (1995), and LeSage (1997) propose the widening of spatial econometrics to include Bayesian techniques, not least because of the information that this yields around the specific point estimates reached in standard modelling. Pinkse and Slade (1996) and Dubin (1997) have begun work on the application of spatial econometric techniques to discrete-choice models, noting that non-spherical disturbances are extremely difficult to handle in the limited dependent variable context. Pinkse and Slade are concerned to be able to detect spatial clustering or dispersion of in retail gasoline contract types across branded service stations in Vancouver, while Dubin models the behaviour of automobile dealers.

Simply in order to give a flavour of the kinds of issues involved, I will briefly run through one of the standard examples, first analysed in this context by Hepple (1976).

Hanna (1966) proposed that the 1960 value of 1955-9 used cars would be higher in states that had higher sales taxes and/or higher transport charges added to the price of new vehicles, a hypothesis confirmed by his ordinary least squares results (Table 4). Hepple (1976) used Hanna's study to illustrate the effects of error dependence in regression modelling, and demonstrated that this finding was spurious. The price variable is significantly autocorrelated (the standard variate of Moran's I is 8.07 under randomisation, prob. < 0.001), as is the least squares error term (Moran's I = 4.25, prob. < 0.001). Hepple drew the conclusion that the problem was in the error term, not least because at that time other tests were not available.

Table 4: Modelling used car prices in 1960, 49 U.S. states, (t-values in parentheses).

	OLS	Spatial lag	Spatial error	Autoregression
Constant	1435.97	332.70	1526.24	282.55
	(52.8)	(2.97)	(56.39)	(2.81)
Sales tax/other charges	0.69	0.18	0.11	—
	(3.96)	(1.61)	(1.05)	—
ρ	—	0.77	—	0.82
	—	(9.94)	—	(12.52)
λ	—	—	0.80	—
	—	—	(11.58)	—
R^2	0.25	0.73	0.73	0.73
σ^2	3181.96	1080.39	1088.18	1093.91

Testing the OLS model using the standard Lagrange multiplier tests gives highly significant results for both of the alternative specifications, but using the new LM tests accommodating the alternative non-zero nuisance parameters yields values of 8.42 for the test for an underlying spatial lag model (χ^2 with 1 d.f., prob. = 0.004), and of 0.035 for an underlying spatial error model (prob. = 0.851). A likelihood ratio test between the estimated spatial lag and spatial error models just fails to find in favour of the spatial lag model (LR = 1.80, prob. = 0.121).

The consequences of taking spatial dependence into account are quite clear. The error variance of the two spatial models is much smaller than that of the least squares regression estimates, and the proportion of the variance in used car prices explained has risen from a quarter to three quarters. The coefficient of the cost variable is no longer significant at the $\alpha = 0.05$ level. Perhaps unsurprisingly, the spatial lag ρ and error λ coefficient estimates are highly significant. Were we to prefer the spatial lag model, we could interpret the results to indicate that ρ represents the influence of the average price in contiguous states, indicating that price setting involves the comparison of prices across state lines. From the final column in Table 4, we see that the residual variance of the autoregressive model, dropping the tax/charges variable altogether, does very nearly as well as the spatial error model, and indeed the LR test to differentiate between the autoregression and the spatial lag model does not come down strongly for the latter (LR = 2.60, prob. = 0.067).

6.3 Geographically weighted regression

As global measures of spatial association have been supplemented by local indicators, Fotheringham, Charlton, and Brunsdon (1996, 1997) and Brunsdon, Fotheringham, and Charlton (1996) have been developing weighting schemes to allow possible differences in local parameter estimates for regression models to be revealed. Moving from the global to local settings, one would perhaps expect the local parameter estimates to vary, but within the bounds of their global standard error based confidence intervals, that is with divergences of more than ± 2 less than five times in a hundred. The weighting scheme used so far is distance based, weighting zone *i* with unity, and with weights declining with increasing distance from *i*. There are similarities with kernel regression techniques, although these use weighting in attribute space, rather than across the observations. Currently, cross-validation is used to select an appropriate global bandwidth parameter, which then determines the form of the distance decay function used to define the weights for each observation. There are clearly substantial difficulties involved in making statistical inferences from results of this kind of procedure, although it has proved very useful in showing up missing variables.

7 Final comments

The motivation for this contribution has undoubtedly been rather missionary, because there are two possible reasons why spatial econometric methods have not been more widely adopted in economics, and particularly in the very relevant areas of trade and location, supposing that researchers are concerned to test their hypotheses on empirical data. The first, which is not improbable, is that the methods are not yet adequate, but here one can see economists, like Pace, Pinkse, Dubin, or Case, contributing with new variants or increments to the existing body of work. Indeed, economists tend to be very welcome to publish in the key journals in the field, such as *Environment and Planning A*, *Geographical Analysis*, or *Regional Science and Urban Economics*. The second is that we have not done a very good marketing job, and although this paper is not going to impress my colleagues specialising in that "black art", I hope that it may increase curiosity, and that the lengthy list of references can give that curiosity something to feed on. Nonetheless, *caveat emptor*.

References

Amrhein, C. G. 1995 Searching for the elusive aggregation effect: evidence from statistical simulation. *Environment and Planning A*, 27, 105–120.

Amrhein, C. G. and Reynolds, H. 1996 Using spatial statistics to assess aggregation effects. *Geographical Systems*, 3, 143–158.

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

Anselin, L. 1988b Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, 20, 1–17.

Anselin, L. 1990 Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics*, 20, 141–163.

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

Anselin, L. 1995b *SpaceStat version 1.80 user's guide.* (Morgantown, WV: Regional Research Institute, West Virginia University).

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

Anselin, L., Bera, A. K., Florax, R. and Yoon, M. J. 1996 Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26, 77–104.

Anselin, L. and Florax, R. 1995 Small sample properties of tests for spatial dependence in regression models: some further results In L. Anselin and R. Florax (eds) *New directions in spatial econometrics*, (Berlin: Springer), 21-74.

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

Anselin, L. and Hudak, S. 1992 Spatial econometrics in practice: a review of software options. *Regional Science and Urban Economics*, 22, 509–536.

Anselin, L. and Rey, S. 1991 Properties of tests for spatial dependence in linear regression models. *Geographical Analysis*, 23, 112–131.

Anselin, L., Varga, A. and Acs, Z. 1997 Local geographic spillovers between university research and high technology innovations. *Journal of Urban Economics*, 42, 422–448.

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

Aten, B. 1996 Evidence of spatial autocorrelation in international prices. *Review of Income and Wealth*, 42, 149–163.

Aten, B. 1997 Does space matter? International comparison of the prices of tradables and nontradables. *International Regional Science Review*, 20, 35–52.

Audretsch, D. and Feldman, M. 1996 R&D spillovers and the geography of innovation and production. *American Economic Review*, 86, 630–640.

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

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

Barkley, D. L., Henry, M., Bao, S. and Brooks, K. 1995 How functional are economic areas? Tests for intra-regional spatial association using spatial data analysis. *Papers in Regional Science*, 74, 297–316.

Barro, R. J. 1997 *Determinants of economic growth.* (Cambridge MA: MIT Press).

Bavaud, F. 1998 Models for spatial weights: a systematic look. *Geographical Analysis*, 30, 152–171.

Bernat, G. A. 1996 Does manufacturing matter? A spatial econometric view of Kaldor's laws. *Journal of Regional Science*, 36, 463–477.

Besley T. and Case, A. 1995 Incumbent behavior: vote-seeking, tax-setting and yardstick competition. *American Economic Review*, 85, 25–45.

Bivand, R. S. 1984 Regression modelling with spatial dependence: an application of some class selection and estimation methods. *Geographical Analysis*, 16, 25–37.

Bivand, R. S. 1992 SYSTAT-compatible software for modelling spatial dependence among observations. *Computers and Geosciences*, 18, 951–963.

Bivand, R. (1997) Scripting and tool integration in spatial analysis: prototyping local indicators and distance statistics. In Z. Kemp (ed) *Innovations in GIS 4*, (London: Taylor & Francis), 127–138.

Bivand, R (forthcoming) Software and software design issues in the exploration of local dependence. *The Statistician.*

Bivand R. and Szymanski S. 1997 Spatial dependence through local yardstick competition: theory and testing. *Economics Letters*, 55, 257–265.

Bivand R. and Szymanski S. 1998 Modelling the impact of the introduction of compulsory competitive tendering. Mimeo, Department of Geography, Norwegian School of Economics and Business Administration, Bergen, Norway.

Brett, C. and Pinkse, J. 1997 Those taxes are all over the map!: a test for spatial independence of municipal tax rates in British Columbia. *International Regional Science Review*, 20, 131–151.

Brunsdon, C. and Charlton, M. (1996) Developing an exploratory spatial analysis system in XLisp-Stat. In D. Parker (ed) *Innovations in GIS 3*, (London: Taylor & Francis), 135–145.

Brunsdon, C., Fotheringham, A. S. and Charlton, M. 1996 Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28, 281–298.

Buja, A., Cook, D. and Swayne, D. 1996 Interactive high dimensional data visualization. *Journal of Computational and Graphical Statistics*, 5, 78–99.

Burridge, P. 1980 On the Cliff-Ord test for spatial autocorrelation. *Journal of the Royal Statistical Society B*, 42, 107–8.

Burridge, P. 1981 Testing for a Common Factor in a spatial autoregressive model. *Environment and Planning A*, 13, 795–800.

Case, A. C. 1991 Spatial patterns in household demand. *Econometrica*, 59, 953–965.

Case, A. C., Rosen, H. S. and Hines, J. R. 1993 Budget spillovers and fiscal policy interdependence: evidence from the states. *Journal of Public Economics*, 52, 285–307.

Cliff, A. D. and Haggett, P. 1996 The impact of GIS on epidemiological mapping and modelling. In P. Longley and M. Batty (eds) *Spatial analysis: modelling in a GIS environment* (Cambridge: Geoinformation International), 321–343.

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

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

Cook, D., Majure, J. J., Symanzik, J. and Cressie, N. 1996 Dynamic graphics in a GIS: exploring and analysing multivariate spatial data using linked software. *Computational Statistics*, 11, 467–480.

Cook, D., Symanzik, J., Majure, J. J. and Cressie, N. 1997 Dynamic graphics in a GIS: more examples using linked software. *Computers and Geosciences*, 23, 371–385.

Cressie, N. A. C. 1993 *Statistics for spatial data.* (New York: Wiley).

Cressie, N. A. C. 1996 Change of support and the modifiable areal unit problem.*Geographical Systems*, 3, 159–180.

Cuzick, J. and Edwards, R. 1990 Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society B*, 52, 73–104.

Deutch, C. V. and Journel, A. G. 1992 *GSLIB: geostatistical software library and user's guide.* (Oxford: Oxford University Press).

Diggle, P. J. 1990 A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society A*, 153, 349–362.

Diggle, P. J. and Chetwynd, A. G. 1991 Second order analysis of spatial clustering foe inhomogeneous populations. *Biometrics*, 47, 1155–1163.

Diggle, P. J. and Rowlingson, B. 1994 A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society A*, 157, 433–440.

Dubin, R. A. 1988 Estimation of regression coefficients in the presence of spatially autocorrelated error terms. *Review of Economics and Statistics*, 70, 466-474.

Dubin, R. A. 1997 A note on the estimation of spatial logit models. *Geographical Systems*, 4, 181-194.

Duncan, O. D., Cuzzort, R. P., Duncan, B. 1961 *Statistical geography: problems in analysing areal data.* (Glencoe, Illinois: Free Press).

Fisher, P. F. and Langford, M. 1995 Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, 27, 211–224.

Fotheringham, A. S., Charlton, M. and Brunsdon, C. 1996 The geography of parameter space: an investigation of spatial non-stationarity. *International Journal of Geographical Information Systems*, 10, 605–627.

Fotheringham, A. S., Charlton, M. and Brunsdon, C. 1997 Two techniques for exploring non-stationarity in geographical data. *Geographical Systems*, 4, 59–82.

Fotheringham, A. S. and Wong, D. W. S. 1991 The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23, 1025–1044.

Gatrell, A. and Bailey, T. 1996 Interactive spatial data analysis in medical geography. *Social Science & Medicine*, 42, 843–855.

Gatrell, A., Bailey, T., Diggle, P. and Rowlingson, B. 1996 Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 21, 256–274.

Getis, A. 1991 Spatial interaction and spatial autocorrelation: a cross products approach. *Environment and Planning A*, 23, 1269–1277.

Getis, A. and Ord, J. K. 1992 The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24, 189-206 (cf. also *Geographical Analysis*, 25, 276).

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

Griffith, D. A. 1988 *Advanced spatial statistics.* (Dordrecht: Kluwer).

Griffith, D. A. 1993 *Spatial regression analysis on the PC: spatial statistics using SAS.* (Washington DC: Association of American Geographers).

Griffith, D. A. 1995 Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In S. L. Arlinghaus and D. A. Griffith (eds) *Practical handbook of spatial statistics* (Boca Raton: CRC Press), 65–82.

Haggett, P. 1981 The edges of space. In Bennett, R. J. (ed) *European progress in spatial analysis* (London: Pion), 51–70.

Haining, R. P. 1983 Modelling intra-urban price competition: an example of gasoline pricing. *Journal of Regional Science*, 23, 517–528.

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

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

Haining, R. P. 1991 Bivariate correlation with spatial data. *Geographical Analysis*, 23, 210–227.

Hanna, F. A. 1966 Effects of regional differences in taxes and transportation charges on automobile consumption. In S. Ostry, and T. K. Rhymes, (eds) *Papers on regional statistical studies* (Toronto: Toronto University Press), 199-223.

Haslett, J., Bradley, R., Craig, P., Unwin, A. and Wills, G. 1991 Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *American Statistician*, 45, 234–242.

Henderson, J. V. 1997 Externalities and industrial development. *Journal of Urban Economics*, 42, 449–470.

Hendry, D. F. and Mizon, G. E. 1978 Serial correlation as a convenient simplification, not a nuisance: a comment on a study of the demand for money by the Bank of England. *Economic journal*, 88, 549–563.

Hepple, L. W. 1976 A Maximum Likelihood model for econometric estimation with spatial series. In I. Masser (ed) *Theory and practice in regional science*, (London: Pion), 90–104.

Hepple, L. W. 1995 Bayesian techniques in spatial and network econometrics: 1 model comparison and posterior odds. *Environment and Planning A*, 27, 447–469.

Hepple, L. W. 1995b Bayesian techniques in spatial and network econometrics: 2 computational methods and algorithms. *Environment and Planning A*, 27, 615–644.

Hepple, L. 1996 Directions and opportunities in spatial econometrics. In P. Longley and M. Batty (eds) *Spatial analysis: modelling in a GIS environment* (Cambridge: Geoinformation International), 231–246.

Hepple, L. 1998 Context, social construction, and statistics: regression, social science, and human geography. *Environment and Planning A*, 30, 225–234.

Hepple, L. 1998b Exact testing for spatial autocorrelation among regression residuals. *Environment and Planning A*, 30, 85–108.

Holt, D., Steel, D., Tranmer, M. and Wrigley, N. 1996 Aggregation and ecological effects in geographically based data. *Geographical Analysis*, 28, 244–261.

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

Hubert, L. J., Golledge, R. G. and Constanzo, C. M. 1981 Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, 13, 224–233.

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

Jaffe, A., Trajtenberg, M. and Henderson, R. 1993 Geographic localization of knowledge spillovers as evidenced by patent citation. *Quarterly Journal of Economics*, 63, 577–598.

Jones, A., Langford, I. and Bentham, G. 1996 The application of K-function analysis to the geographical distribution of road traffic accident outcomes in Norfolk, England. *Social Science & Medicine*, 42, 879–885.

Kingham, S., Gatrell, A. C. and Rowlingson, B. 1995 Testing for clusters of health events within a geographical information system framework. *Environment and Planning A*, 27, 809–821.

Krugman, P. 1995 *Development, geography and economic theory.* (Cambridge MA: MIT Press).

Langbein, L. I. and Lichtman, A. J. 1978 *Ecological inference.* (Beverly Hills: Sage).

Lesage, J. P. 1997 Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, 20, 113–130.

Lesage, J. P. and Pan Z. 1995 Using spatial contiguity as Bayesian prior information in regional forecasting models. *International Regional Science Review*, 18, 33–53.

Levine, N. 1996 Spatial statistics and GIS. *Journal of the American Planning Association*, 62, 381–391.

Lowell, K. 1997 Effect(s) of the "no-same-color-touching" constraint on join-count statistics: a simulation study. *Geographical Analysis*, 29, 339–353.

Mantel, N. 1967 The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209–220.

Morphet, C. S. 1997 A statistical method for the identification of spatial clusters. *Environment and Planning A*, 29, 1039–1055.

O'Loughlin, J. and Anselin, L. 1996 Geo-economic competition and trade bloc formation: United States, German, and Japanese exports, 1968–1992. *Economic Geography*, 72, 131–160.

Oliver, M. and Webster, R. 1986 Combining nested and linear sampling for determining the scale and form of spatial variation of regionalised variables. *Geographical Analysis*, 18, 227–242.

Openshaw, S. and Taylor, P. J. 1979 A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In Wrigley, N. (ed) *Statistical applications in the spatial sciences* (London: Pion), 127–144.

Ord, J. K. 1975 Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70, 120–126.

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

Pace, R. K. and Barry, R. 1997 Quick computation of spatial autoregressive estimators. *Geographical Analysis*, 29, 232–247.

Pace, R. K. and Barry, R. 1997b Sparse spatial autoregressions. *Statistics and Probability Letters*, 33, 291–297.

Pace, R. K. and Barry, R. 1997c Performing large-scale spatial autoregressions. *Economics Letters*, 54, 283–291.

Pinkse, J. and Slade, M. 1996 Contracting in space: an application of spatial statistics to discrete-choice models. Mimeo, Department of Economics, University of British Columbia, Vancouver, Canada.

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

Sala-i-Martin, X. 1994 Cross-sectional regressions and the empirics of economic growth. *European Economic Review*, 38, 739–747.

Sherman, M. 1996 Variance estimates for statistics computed from spatial lattice data. *Journal of the Royal Statistical Society B*, 58, 509–523.

Sherman, M. and Carlstein E. 1994 Nonparametric estimation of the moments of a general statistic computed from spatial data. *Journal of the American Statistical Association*, 89, 496–500.

Student, 1914 The elimination of spurious correlation due to position in time or space. *Biometrika*, 10, 179-180.

Sweeney, S. H. and Feser, E. J. 1998 Plant size and clustering of manufacturing activity. *Geographical Analysis*, 30, 45–64.

Talen, E. and Anselin, L. 1998 Assessing spatial equity: an evaluation of measures of accessibility to public playgrounds. *Environment and Planning A*, 30, 595–613.

Thisse, J.-F. 1997 De l'indétermination des régions et de quelques inconvénients qui en résultent. *L'Espace géographique*, 26, 135–148.

Tierney, L. 1990 *LISP-STAT: an object-oriented environment for statistical computing and dynamic graphics.* (New York: Wiley).

Tiefelsdorf, M. and Boots, B. 1995 The exact distribution of Moran's *I*. *Environment and Planning A*, 27, 985–999.

Tiefelsdorf, M. and Boots, B. 1997 A note on the extremities of local Moran's I_is and their impact on global Moran's *I*. Geographical Analysis, 29, 248–257.

Tufte, E. R. 1997 *Visual explanations.* (Cheshire, Conn: Graphics Press).

Unwin, A. 1996 Geary's contiguity ratio. *Economic and Social Review*, 27, 145–159.

Unwin, D. J. 1996 GIS, spatial analysis and spatial statistics. *Progress in Human Geography*, 20, 540–551.

Upton, G. J. G. and Fingleton, B. 1985 *Spatial data analysis by example: point pattern and quantitative data.* (London: Wiley).

Waldhör, T. 1996 The spatial autocorrelation coefficient Moran's *I* under heteroscedasticity. *Statistics in Medicine*, 15, 887–892.

Webster, R., Oliver, M. A., Muir, K. R. and Mann, J. R. 1994 Kriging the local risk of a rare disease from a register of diagnoses. *Geographical Analysis*, 26, 168–185.

Wrigley, N., Holt, T., Steel, D. and Tranmer, M. 1996 Analysing, modelling and resolving the ecological fallacy. In P. Longley and M. Batty (eds) *Spatial analysis: modelling in a GIS environment* (Cambridge: Geoinformation International), 23–40.