Spatial Data and Statistical Methods: A Chronological Overview

65

Robert Haining

Contents

Abstract

We review some of the special properties of spatial data and the ways in which these have influenced developments in spatial data analysis. We adopt a historical perspective beginning in the early twentieth century before moving to the development of spatial autocorrelation statistics in geography's Quantitative Revolution. Phases of development after the Quantitative Revolution are divided into emergence of spatial econometrics, the development of exploratory methods for spatial data analysis, and local statistics for handling heterogeneity. We then consider more recent advances in the areas of spatial data mining, the "new" geostatistics, and Bayesian hierarchical statistical modeling of spatial data.

R. Haining
Department of Geography, University of Cambridge, Downing Place, Cambridge, UK
e-mail: rph26@cam.ac.uk

M.M. Fischer, P. Nijkamp (eds.), *Handbook of Regional Science*,
DOI 10.1007/978-3-642-23430-9_71,

65.1 Introduction

Spatial statistics is used for the *analysis* of spatial data, that is, "the reduction of spatial patterns to a few clear and useful summaries" (Ripley 1981, p. 1), and for *comparing* such summaries "with what might be expected from theories of how the pattern might have originated and developed" (Ripley 1981, p. 1). In order to test theoretical expectations against data collected through observation, statistical models are often necessary. There are, of course, many different types of models that are critical to the progress of science, but statistical models are formalizations of theory in terms of random variables and their associated probability distributions. We might compare several different models to see which one best fits the data; we might take a single model and see how well it fits the data.

There are several different types of spatial data encountered in geography and regional science. *Point pattern* or *point process* data arise where each data value refers to the location of a discrete object the size of which is sufficiently small relative to the study area that it can be treated as a point (e.g., the location of factories in a region). Interest may focus on how the points are distributed within the region (e.g., are the factories in a particular economic sector spatially clustered or random?). If interest focuses on say the distribution of an attribute attached to each point (e.g., are the factories that have been closed in the last 12 months clustered given the distribution of factories in that sector?), we refer instead to a *marked point process*. Some variables that originate as point data are reported as discrete-valued *regional counts* (e.g., the number of residential burglaries recorded by UK Census Output Area (COA)) or their attributes as continuous-valued *regional averages* (e.g., average household disposable income) or *rates* (e.g., the number of burglaries per 100 households) or *ratios* (e.g., area-standardized disease-specific mortality ratios obtained by dividing the observed number of deaths due to a particular disease in each area by the expected number of cases given the area's population size and its age and sex composition). The reporting areas may be irregular in shape as in the above examples or form a regular grid. When *regional data* refer to small areas (such as COAs or a fine grid), then there may be interest in constructing a density map which is a smoothed representation of the data (e.g., a population density map, a burglary risk map). Some spatial data are *point samples from a continuous surface*, where the point *sampling* has been performed according to some design (e.g., random, stratified random, systematic). The data may refer to levels of surface soil contamination or ground level atmospheric pollution. There may be interest, for example in *geostatistics*, in constructing a map of the attributes spatial variability or interpolating to points or areas on the map where no data have been collected. Spatial data may also take the form of *objects* that have both location and extent and which may or may not fill the space. Interest may focus on modeling the distribution of the objects such as vegetation or land use patches. As a final example, data may refer to the nodes or vertices of a *network* (e.g., a rail, road, or airline network). On the vertices of the network are recorded origin–destination or (directed) flow data such as numbers of people or tonnage of goods moving between two nodes or line segments on the network in a period of time.

Analysis may be concerned to understand population migration or trade *flow data*, for example, using origin- and destination-specific factors and the distance between the origins and destinations (Fischer and Wang 2011).

In any analysis, it may be necessary to combine data of different types: regular areas, irregular areas, and point data, as illustrated, for example, in Elliott et al. (2009). Different data types raise different problems for statistical analysis and modeling, and for a more formal definition of a number of different forms of spatial data, see Cressie (1991, pp. 8–9)

Obtaining useful summaries of spatial data is complicated because single numbers (such as the mean and standard deviation for summarizing aspects of the distribution of data values) are not sufficient for describing the spatial variation in data values. Maps and graphs, perhaps several of both, are required in order to describe both the distributional and spatial variation in the data. Spatial data may show different patterns of variability on different parts of the map and at different scales.

Statistical analysis and modeling of spatial data introduces other considerations. The classical theory of statistical inference assumes data are obtained by randomly *sampling* from the population (so that the data set can be considered representative of the population), and observations are independent and identically distributed (i.i.d.). For this reason, the underlying probability models of classical statistics are i.i.d. But in the case of spatial data, even if sample observations have been collected by a process of random *sampling*, if sample values are sufficiently close together in geographical space, they will not be independent because the population from which the data have been drawn is said to be *spatially autocorrelated*. Data values are not independent; the structure of that dependence may not be the same everywhere on the map (nonstationarity), and there may also be small pockets of high (or low) data values (*disease clusters*, crime, or unemployment hot spots). The mean value of an attribute may not be the same everywhere on the map. In addition to nonindependence, data values across a map may not be uniform in the sense of coming from some common underlying distribution. They may display what is termed *spatial heterogeneity*. These two often-encountered characteristics of spatial data introduce special considerations when undertaking statistical modeling including the need to construct valid (or "permissible") models to describe spatial variability which can then be used for inference with spatial data.

When data are collected in the form of counts or rates by area, there may be further issues to consider. If area data values are averages, rates, or ratios and if in addition areas possess large populations, then such summaries may obscure or conceal *within-area heterogeneity* – subpopulations within areas with markedly different averages (or rates or ratios). On the other hand, if the areas possess small populations, while such a framework might preserve more of the underlying spatial variability in the data (and each small area might be homogeneous), individual area estimates will typically have much larger standard errors. As a further consequence of this, if the map is partitioned into areas, some of which have large while others have small populations, then the data may be *heteroscedastic* – that is, each observation has been drawn from a different probability distribution with

a different variance. If area data values refer to an ecological covariate (i.e., expressing a property of an area that is not reducible to measures at the individual level, e.g., social capital, area deprivation, social cohesion), then the values of this variable will depend on the scale of the partition and the configuration of the boundaries (the zoning). This is one example of the "*modifiable areal unit problem*" which states that the results of spatial analyses and modeling are conditional on the particular partition through which we observe spatial variation. These and other challenges that confront the analysis of spatial data have been reviewed at some length elsewhere (see Haining (2009)).

This chapter provides an overview of some of the major developments in spatial statistics with particular relevance to geography and regional science. It is divided into four sections. The first section briefly reviews the statistical origins of what is now called spatial statistics and describes how and why one area of spatial statistics came to be part of geography's *Quantitative Revolution* in the 1950s, 1960s, and early 1970s. The second section discusses the emergence of *spatial econometrics* and its links (and overlaps) with spatial statistics. What characterizes *spatial econometrics* and in what sense does it stand apart from, indeed distinct from, the field of spatial statistics? The third section considers the emergence of *exploratory spatial data analysis* in the 1980s and follows this with the development of "local statistics" for analyzing spatial *heterogeneity*. The fourth section is an overview of some recent developments in the field: *spatial data mining*, what I shall term the "new" *geostatistics* and Bayesian hierarchical spatial statistical modeling. The purpose of this chapter is to show the reader the development path of spatial statistics and how this can be seen as a response to the distinctive properties and challenges presented by spatial data.

65.2 Where Did It All Start?

65.2.1 The Statistical Origins

The roots of spatial statistics can be traced back to at least the early twentieth century and the involvement of classically trained statisticians in analyzing the data from agricultural uniformity trials carried out at Rothamsted in England. In such analyses, a component of yield variation is due to processes operating at a scale greater than the size of the spatial units used to report the data so that crop yields in adjacent plots tend to be similar. Classically trained statisticians became interested in the problem of how to carry out field trials (which might be testing different management methods and crop varieties in relation to soil type) using experimental designs that would control for such attributes and hence yield stronger inference.

The analysis of agricultural yield data was also motivation for the seminal paper by Whittle (1954) in which he made explicit the link between the problem of analyzing such data and two-dimensional stochastic models. In that paper, he defined the *simultaneous spatial autoregressive (SAR) model* on

a regular square lattice in which if X(i,j) denotes the random variable at location (i,j) of a regular square lattice then

$$X(i,j) = a[X(i-1,j) + X(i+1,j)] + b[X(i,j-1) + X(i,j+1)] + e(i,j) \quad (65.1)$$

where a and b are parameters and e(i,j) is a normally distributed white noise (i.i.d.) process. Whittle's paper notes that unlike time series modeling, spatial process modeling needs to allow for dependence to extend in all directions (not just from past to future as in time series modeling) and that in two dimensions, the dependence structure on the north–south axis might differ from that on the east–west axis. Twenty years later, Besag (1974) in an equally seminal paper presented the theory of conditionally specified spatial models (for continuous- and discrete-valued random variables) and included in his set of models the *conditional spatial autoregressive (CAR) model* for normally distributed random variables. For readers interested in the differences between these models, see, for example, Haining (2003 pp. 297–304).

A slightly earlier strand of development, predating Whittle's work, saw the publication of papers addressing the problem of how to test for what is now referred to as *spatial autocorrelation*. In each case, the null hypothesis of no *spatial autocorrelation* is tested against a nonspecific alternative that the observations are autocorrelated. The tests, *Geary's c test*, based on squared differences between observations in adjacent areas; *Moran's I test*, based on the cross product between observations in adjacent areas; and Krishna Iyer's *join count test* for nominal level data, based on counting the number of adjacent areas in the same class or in different classes, are reviewed in Cliff and Ord (1973). None of these early tests made any allowance for the topological and/or geometrical structure of the areas making up the areal system other than whether pairs of areas were adjacent (shared a common border) or not.

These developments were applicable to the case of area data. *Geostatistics* was developed for spatial data collected as point or block samples from a continuous surface and with a quite different aim in mind. Matheron (1963) developed a comprehensive theory of optimal *interpolation* in geographical space on the basis of sample data. In purely geographical terms, we might think of this as a theory for drawing maps of continuous phenomena on the basis of a scattered sample of observations. Spatial variation in any particular set of data was described by estimating the semi-*variogram* (a squared difference statistic) and then finding a best fit model for this empirical semi-*variogram* from the set of "permissible" models. A rich array of models is available for describing spatial variation. For further discussion of this area of spatial statistics and its antecedents, see Haining et al. (2010) which also includes comparative comments with the literature cited above.

65.2.2 From Statistics into Geography and Regional Science

It was during the *Quantitative Revolution* in the 1950s and 1960s that some aspects of this statistical theory began to filter into geography. Researchers in sociology had

already recognized (in some cases long before geography's *Quantitative Revolution*) that the theory and tools of classical statistics could not be applied uncritically to the analysis of geographical data (see, e.g., Neprash (1934) as well as other papers in the same journal supplement). It was not until the 1960s that the "problem" of *spatial autocorrelation* began to be examined carefully by geographers and a key deficiency of the earlier tests devised by Moran, Geary, and Krishna Iyer, their topological invariance, confronted.

Geography's *Quantitative Revolution* was not merely a methodological revolution – the aligning of geography's methods with those used by the quantitative sciences – it was also a revolution in terms of how the subject matter of geography should be addressed. It was a revolution in the sense that geographers became interested in the development of theory. But it would be theory that would only survive so long as it withstood rigorous attempts to refute it through empirically grounded research. Models represented the translation of theory into a form that would enable empirical testing to be performed, and although not the only form of model building that entered the geographical literature, statistical modeling was a key element in this agenda. So herein lay the nub of the problem. Because of the importance of statistical modeling to theoretical geography, these issues could not be set to one side. And there were of course precedents to believing that none of these problems were insuperable, least of all coping with the effects of spatial dependence. Time series analysis had undergone a transformation in the first half of the twentieth century and now underpinned the practice of econometrics which in turn supported the testing of economic theory. Geographical statistics was in need of a similar transformation.

It was the work of Cliff and Ord that reported important breakthroughs in constructing statistics for testing for *spatial autocorrelation* on the sorts of irregular areal frameworks social scientists most frequently worked with. In Cliff and Ord (1973), they developed the inference theory for modified versions of *Geary's c* and *Moran's I* statistics introducing a "weighting" term into the formulation of the statistic that allowed adjacency to be specified much more generally than had hitherto been possible. The reader interested in this aspect of geography's history should refer to the special issue of the journal *Geographical Analysis* (2009(4)).

Two areas of geography benefitted most directly from this innovative work. The first was in the application of the regression model. *Regression models* enable data analysts to empirically test the relationship between a dependent variable and a set of explanatory or independent variables. This statistical model has in the past and indeed continues to play a very important role in developing and testing theory in many areas of quantitative science. In the case of classical least squares *regression modeling*, population inference (hypothesis testing, parameter estimation) is based on the assumption that model errors are i.i.d., and failure to satisfy this assumption results in underestimation of type I errors in hypothesis testing. Regression residuals are estimates of model errors, and Cliff and Ord (1973) provide an inference theory for testing for nonindependence of model errors using the least squares residuals.

The second area to benefit was the testing of specific types of spatial theory. Cliff and Ord (1973) considered the area of spatial *diffusion modeling* where different theoretical processes were simulated and then compared with observed outcomes. But, among many other complications associated with this approach, evaluating the correspondence between simulated output and empirical observation must compare not only the frequency distribution of numbers of adopters by area but also the spatial arrangement of the counts.

Another example was in the area of economic geography and the analysis of those economic processes that by definition are embedded into geographical space (e.g., urban and regional development, location theory, land use change, regional and international trade, spatial price competition). During the *Quantitative Revolution*, those economic geographers who went in search of stronger theoretical perspectives began to take a close interest in the work of location theorists. They also began to engage with the newly emerging field of regional science. Early books on the tools and methods of regional science paid little attention to statistical modeling, but in the 1970s, the field of *spatial econometrics* began to emerge becoming to regional scientists what econometrics had become to economists. The purpose behind its development was to provide the statistical tool kit to enable regional scientists to test spatial economic theory. We now turn to discuss this development.

65.3 Spatial Econometrics

Anselin (1988), and again most recently in his extended review of the field in 2010, credits Jean Paelinck with first use of the term *spatial econometrics* in an address to the Dutch Statistical Association in 1974. The term was used to "designate a growing body of the regional science literature that dealt primarily with estimation and testing problems encountered in the implementation of multiregional econometric models" (Anselin 1988, p. 7). Another significant date in the development of the field is 1979 when Paelinck and Klaassen's (1979) "*Spatial Econometrics*" was published. Anselin (2010) chooses that year as "the historical starting point for *spatial econometrics*" (p. 3).

Anselin (2010), as others had done before him, argues that *spatial econometrics* and spatial statistics should be seen as distinct. The distinction is defined by the types of problems that are tackled. Whereas spatial statistics is fundamentally data driven, *spatial econometrics* (like econometrics), is fundamentally theory driven. *Spatial econometrics* has been developed explicitly to fit *spatial regression models* to test spatial economic theory – in this sense moving away from Paelinck's original definition of the field.

Providing spatial econometricians do not cut themselves off from the rich vein of statistical theory and models generated by spatial statisticians, then there may be advantage to be gained from distinguishing between *spatial econometrics* and spatial statistics. But the justification for the distinction is not entirely convincing.

In contrast to earlier days in geography's *Quantitative Revolution*, statisticians today see many opportunities for fruitful interaction on broad classes of spatial problems and would not accept the view that their model building is purely data driven, by implication "atheoretical." As Cressie and Wikle (2011, p. 14) have recently observed, "...Statistics has become more a Science than a branch of Mathematics. . .."

Spatial econometrics today is principally concerned with how to specify, fit, and then carry out diagnostic checks on *regression models* when working with locationally (or spatially) referenced data. The data can be cross-sectional (purely spatial) or *spatiotemporal* with measurements on one or several variables. Underlying these models are usually theories about how distance (to a particular location such as a city center), spatial configuration (the spatial distribution of objects within a space, such as whether areas of poverty are scattered or ghettoized within an urban area), or spatial gradients (between neighboring areas in terms of socioeconomic characteristics) help to explain variation in a dependent variable. What is observed (e.g., area crime rates, regional economic performance) is not necessarily an outcome purely of circumstances *within the places themselves* because what is observed, and the variation we want to explain, may be the outcome of processes that operate across geographical space (e.g., different forms of interaction).

Methodologically *spatial econometrics* focuses on two properties commonly encountered when handling geographical data: spatial dependence (autocorrelation) and spatial *heterogeneity*. We shall consider approaches to the handling of spatial *heterogeneity* in the next section and focus here on just the handling of spatial dependence in *spatial econometrics*.

Typically, spatial dependence is handled by specifying lagged variables in the *regression model*. In the case of lagging the dependent variable, a model might be specified of the form

$$\begin{aligned} Y(i) &= b_0 + b_1 X_1(i) + \ldots + b_k X_k(i) + \rho \sum_{j \in N(i)} w(i,j)\, Y(j) + e(i) \\ \sum_{j=1}^{n} w(i,j) &= 1; \quad i = 1, \ldots, n \end{aligned} \tag{65.2}$$

where the $\{e(i)\}$ are i.i.d $N(0, \sigma^2)$, b_0 is the intercept coefficient, and $b_1,\ldots,b_k$ are the regression coefficients on the independent variables X_1 to X_k. The parameter ρ is the spatial interaction parameter for the weighted average of the dependent variable (Y(i)). For any given site i, this weighted average takes in the values at sites neighboring i but excluding site i (N(i)). Thus, $w(i,j) > 0$ if $j \in N(i)$ and $w(i,i) = 0$ for all i. Models specified in this way and in which the influence of neighboring sites is usually stronger the closer they are to i ($w(i,j) > w(i,k)$ if j is closer to i than k is to i) have a long history in the statistical modeling of certain types of economic interaction processes including price competition effects. Clearly, other forms of weighting could be constructed to reflect the structure of economic interactions across space (Haining 1990).

Lagging may also be specified on one or more of the independent variables as in, for example, the model

$$\begin{aligned} Y(i) &= b_0 + b_1 X_1(i) + \ldots + b_k X_k(i) + b_{r;lag} \sum_j c\,(i,j)\, X_r(j) + e(i) \\ \sum_{j=1}^{n} c(i,j) &= 1; \quad i = 1, \ldots, n \end{aligned} \tag{65.3}$$

where in this case Y(i) is modeled to have an association with the independent variable X_r as a function not only of X_r's value at i but also its value at neighboring locations. (For this reason, we use the notation c(i,j) (rather than w(i,j)) to distinguish the spatial averaging in Eq. (65.3) from that in Eq. (65.4).) This type of model is sometimes encountered in house price modeling where characteristics of the neighborhood where the house is located and nearby neighborhoods may impact on price. This type of spatial averaging or smoothing may also be encountered in environmental epidemiology where the air pollution level in neighboring areas is treated as a risk factor because people move about in their day-to-day lives and are thus exposed to levels of air pollution in areas other than where they reside.

In the absence of well-defined explanatory variables to include in the model, the spatial lagging may be applied to the errors

$$\begin{aligned} Y(i) &= b_0 + b_1 X_1(i) + \ldots + b_k X_k(i) + u(i) \\ u(i) &= \theta \sum_{j \in N(i)} w(i,j)\, u(j) + e(i) \\ \sum_{j=1}^{n} w(i,j) &= 1; \qquad i = 1, \ldots, n \end{aligned} \tag{65.4}$$

where the terms in Eq. (65.4) are as defined above and θ is now the spatial interaction parameter associated with the errors. Forms of this model and Eq. (65.2) have been used in the modeling of origin–destination *flows* (Fischer and Wang 2011, pp. 64–67).

As Anselin (2010) points out, the methodology associated with the fitting of this class of models has continued to evolve. He reviews in some detail the notable strides that have been made both in the rigor with which these and other models can be fitted and in the availability of software to implement the fitting. One problematic aspect in this evolutionary development is the specification of the *weights matrix* {w(i,j)}. Adjacency is the default option for many analysts in specifying the *weights matrix* with two areas being defined as *neighbors* if they share a common border. Most software makes this an easy option to implement. But adjacency may not always be appropriate depending on the model to be fitted and whether, for example, there is a need to capture other forms of spatial relationship including

hierarchical dependency structures and complex patterns of spatial competition (Haining 1990). Another approach is to define the elements of the *weights matrix* based on the similarity of the areas in terms of one or more covariates when borrowing data spatially to strengthen small area inference, for example, social and other interpersonal networks may be used to underpin spatial relationships based on the presence or absence of social relationships. Lu et al. (2007) define, as part of a Bayesian *hierarchical model*, an intrinsic conditional autoregressive model where the weights, w(i,j), are Bernoulli distributed with parameter p(i,j) and where logit(p(i,j)) is a linear function of a set of covariates (z(i,j)) based on known features of the pair of areas i and j.

The links between methodological advance and the evolution of spatial economic theory are only touched upon in Anselin (2010) – in that sense, his review is concerned with theoretical *spatial econometrics* (statistical methods) rather than applied *spatial econometrics* (economic models). Over time, applied *spatial econometrics* has tended to become synonymous with *regression modeling* applied to spatial data where *spatial autocorrelation* and spatial *heterogeneity* in particular are present and need to be accommodated. Its treatment of spatial effects reflects the growing "legitimization of space and geography" (Anselin 2010, p. 8) in the quantitative social sciences more generally. But the subfield perhaps needs to be more than that if it is to justify its separate identity from spatial statistics and fully justify its "econometric" label. A close link with mainstream economic theory would seem essential in order to provide economic legitimacy to models (systems of equations) within which geography and spatial relationships have been, in economic terms, rigorously embedded (Fingleton (2000)).

65.4 New Kinds of Geographical Exploration

65.4.1 Exploratory Spatial Data Analysis

Exploratory data analysis (EDA) is a collection of techniques for summarizing data properties, detecting patterns in data, identifying unusual or interesting features in data, detecting errors, distinguishing accidental from important features in a data set, and formulating hypotheses from data. *EDA* might also be used in later phases of analysis, for example, in assessing model fit. Techniques are typically visual (charts, graphs, and figures) and/or numerical (resistant statistics, i.e., statistics not greatly affected by a small number of extreme values). *Exploratory spatial data analysis* (ESDA) extends the definition of *EDA* to spatial data, extending the set of visual tools to include the map and the set of numerical tools to include, for example, spatial *cluster detection* statistics (Haining 1990, 2003).

GIS or GIS-like software, for example, GeoDa, have provided excellent platforms for these tools (see www.geodacenter.asu.edu). Advances in computer technology have had a particularly big impact on *ESDA* with the development of new visualization techniques such as *brushing* (highlighting cases in one graph such as a segment of a boxplot and seeing them highlighted in another graph or on a map),

dynamic brushing (brushing using a moving window), and various forms of *dynamic interactivity* (allowing the user to modify the graphics themselves to better explore data properties, e.g., rescaling and rotating three-dimensional plots).

A significant challenge in undertaking these forms of analysis with area data (where data values refer to areal aggregates such as census tracts) is the problem of comparability especially when dealing with small numbers of events (e.g., numbers of cases of a disease by small area). In many areas of *scientific visualization*, different data values are directly comparable (e.g., the results of experiments, data values taken from time windows of the same length). But area data across a region often refer to polygons of different physical sizes and with different baseline populations. This raises two distinct problems for *ESDA*. Comparing rates and ratios across such a map is potentially misleading. A rate computed for an area with a small denominator population has a larger error variance than a rate computed for an area with a large denominator population. This may necessitate using different sized symbols or other visual devices to distinguish between high- and low-precision data values. Extreme rates (and ratios, such as standardized ratios where an observed count is divided by an expected count) are most often found when the denominator population is small, but statistically significant rates (relative to some baseline) are most often found when the denominator population is large (see, e.g., Haining 2003, pp. 194–199). Sometimes, the areas with the largest populations are physically the smallest (e.g., the census tracts in urban as compared to rural areas) and may be hard to see depending on the scale of the map. One solution to this is the cartogram, where, for example, each area is physically transformed so that its size is proportional to its population. As computer technology has advanced, it has become possible to develop many different forms of cartogram, some that more closely reflect the area as the viewer is used to seeing it which may help him or her to better navigate and hence read the map. For numerous examples of *cartograms* see, for example, www.worldmapper.org and www.sasi.group.shef.ac.uk/maps.

65.4.2 The Local Revolution

We noted above that spatial *heterogeneity* is a property often present particularly when analyzing spatial data over a large geographic area. *Heterogeneity* may be illustrated in the following terms. Assume we are dealing with the counts associated with the number of new cases of a disease in each of n areas during an interval of time. Suppose the generating process for these counts is dependent on an underlying set of risk parameters $(\lambda_1,\ldots,\lambda_n)$. If all the λ_i are identical, then the risk surface is said to be homogeneous. If for at least some areas $(i \neq k)$, $\lambda_i \neq \lambda_k$, then the risk surface is said to be spatially heterogeneous.

Clusters of cases might arise from a process in which events occur independently of each other but where there is spatial variation in the levels of the risk factors across the map. So, an unusually large number of observed cases in area k may be the product of a high value of λ_k which may be due, in turn, to high levels of the relevant

risk factors in area k that determine the value of λ_k. If adjustment is made for these factors, then the existence of the cluster may be accounted for. Clusters of cases might also be due to a contagion process where although the underlying risk map may be uniform, when one case occurs, it triggers others giving rise to a spatially clustered pattern. Examples of this type of process include repeat offending in the same neighborhood for cases of burglary and the occurrence of cases of an infectious disease. This is referred to as *global heterogeneity*, extra variation in the data that can be analyzed using a global model. This *heterogeneity* may be spatially uncorrelated, but it may be spatially correlated if, for example, the spatial scale of the process exceeds the size of the observational units used to collect data.

However, there is often interest in identifying the specific locations of local clusters (hot spots) of cases in an area referred to as *local heterogeneity*. Kulldorff's scan test, a likelihood-based test statistic, has been widely adopted to test for the presence of spatial clusters in point as well as area data (Kulldorff 1997). This test uses moving windows (circles) of varying size. Each of these many circles represents a possible cluster. The test measures the unusualness of each potential cluster using a local likelihood ratio statistic which compares a null hypothesis that cases occur in the population at risk with equal probability whether individuals are inside or outside the circle against an alternative hypothesis that cases inside the circle have a higher probability of occurrence than those outside the circle. The circle with the highest local likelihood ratio statistic is considered the most likely cluster. The question posed is "how unusual is this most unusual collection of events?" (Waller 2009, p. 312). By using Monte Carlo hypothesis testing, the scan test is able to answer this question avoiding the *multiple testing problem*. In addition to the scan test that looks for clusters wherever they might be on the map, another class of techniques tests for whether there is an unusually large number of cases around a specific location such as a point source of pollution or other source of possible contamination. This is referred to as a *focused test* (see, e.g., Haining 2003, pp. 263–5).

That *heterogeneity* may be present in the relationship between a dependent variable and the set of independent variables that explain its spatial variation underlies another important set of local spatial statistical techniques. Such a *regression model* might take the form

$$Y(i) = b_0 + b_1(i)X_1(i) + \ldots + b_k(i)X_k(i) + e(i) \qquad i = 1, \ldots n \tag{65.5}$$

where the terms are as defined for Eq. (65.2), but now the regression coefficients depend on i so that parameter values differ for each observation. Additional modeling assumptions have to be introduced in order to fit a model of this type as otherwise there is insufficient data to estimate the parameters.

Consider the hedonic *regression modeling* of house prices in an area large enough to encompass different climatic regimes so that house buyers attach different values to housing attributes depending on location with respect to these different regimes. If the geographic area can be partitioned into different areas, then a *spatial regimes* model may be used in which model parameters are allowed to differ

from one area to another (but not within areas) using dummy variables in the regression that distinguish between the different areas (Anselin 1988). Spatial *heterogeneity* however might show a form of variation where the parameters vary continuously across the study area, rather than discretely, preventing any prior partitioning. In this case, other methods might be implemented, and the interested reader is referred to Jones and Casetti (1992) for the *expansion method* and to Fotheringham et al. (2000) for the method of *geographically weighted regression* (*GWR*). In the case of the *expansion method*, the parameters are expressed as a function of a finite number of other variables called expansion variables (z). To take a simple example of the *spatial expansion method* where parameter variation is treated as a function of spatial location, then we might assume

$$b_1(i) = \varphi_0 + \varphi_1 z_1(i) + \varphi_2 z_2(i) \;\; i = 1, \ldots n \tag{65.6}$$

where φ_1 and φ_2 are parameters and the variables z_1 and z_2 are the coordinates defining the centroid of each area i. In this case, the spatial expansion is a linear or first-order trend surface which is the additional modeling assumption about how the regression parameters vary spatially. Higher-order trend surfaces could be used or indeed other types of variables. By contrast, *GWR* is based on obtaining local estimates of each parameter where a separate model is fitted to each area. For any $b_j(i)$, for example, data at i are used as well as data from areas close to i but giving most weight to those data values nearest to i. Many possible weighting functions (spatial kernels) can be specified. The additional modeling assumption is that the data in nearby areas to any i carry information about the value of the parameter in i (a form of *spatial autocorrelation*). For a comparative overview of these and other methods for allowing local variation in *regression model* parameters, including *spatially varying coefficients model*s, see Lloyd (2011, pp. 109–143).

Heterogeneity may be associated with other properties of the attribute such as its spatial dependency structure. The spatial dependency structure might be different on different parts of the map. Again, depending on the nature of the spatial correlation, there may be a single, global, model of spatial variation that can accommodate the apparent *heterogeneity*. However, there may be circumstances where a global model of spatial variation will not provide a useful model for the data, for example, where there is either theoretical or empirical evidence (or both) that data from particular parts of a map reflect the outcome of special and distinctive local processes. In *geostatistics*, for example, different *variograms* may be needed for different map segments in order to implement kriging.

65.5 Into the Twenty-First Century

In this section, we reflect on some of the areas of spatial statistics which are shaping and will probably continue to form a significant part of the research agenda in spatial statistics in the coming years. We look at the following areas: *spatial data mining*, the "new" *geostatistics*, and *Bayesian spatial hierarchical modeling*.

65.5.1 Spatial Data Mining

Spatial data mining is the process of discovering interesting but potentially useful patterns in spatial databases. It therefore shares at least some of the objectives of ESDA described in an earlier section. But *spatial data mining* is concerned with the development of automated methods that can be applied to large (and very large) spatial databases. Extracting patterns from large databases underpins decision making in many organizations including those concerned with public health, crime and disorder, land use and transportation, and environmental management.

In common with the relationship of ESDA to EDA, *spatial data mining* when compared to other forms of (nonspatial) *data mining* has the special challenge of recognizing spatial relationships and spatial *neighbors* and taking into account the special properties of spatial data. The location and spatial extension of objects need to be embedded into algorithms. "Neighbor relations" need to be examined for many objects within the same analysis and the term "neighbor" interpreted in many different ways for a thorough interrogation. Moreover, given the size of databases and hence the time taken to process data, it has to be possible to achieve efficient implementation for the purposes of, among others, detecting spatial clusters, spatial outliers and co-location, and relationship patterns among different classes of point, line, and polygon objects such as the distribution of an animal species and wildlife habitats. This is one aspect of the "process of stimulus and convergence" between Geographic Information Systems (*GIS*) and spatial data analysis which began in the 1960s and discussed by Goodchild and Haining (2004): "it is more difficult to analyse the vast amounts of (spatial) data available. . ., and to test new theories and hypotheses without computational infrastructure; and the existence of such infrastructure opens possibilities for entirely new kinds of theories and models, and new kinds of data" (p. 382). *GIS* has an important role to play in providing the necessary computational infrastructure for spatial *data mining*. For further discussion as well as numerous examples of *spatial data mining*, see Miller and Han (2009).

65.5.2 The "New" Geostatistics

Traditionally, *geostatistics* has been viewed as a tool to enable physical and environmental scientists to analyze sample data obtained from a continuous surface. But more recently, the methods of *geostatistics* have been adapted to predict and map *regional data* in the form of small area counts. Oliver et al. (1998) use *binomial co-kriging* to analyze the risk of childhood cancer in the English West Midlands. Population size variation across the areal units is taken into account with pairs of areas with larger populations (and hence more reliable rates) given more weight in the estimation of the *variogram*. If the population at risk is large and the probability of having the disease is small so that the *small number problem* arises, *Poisson kriging* can be used.

Geostatistical change of support methods have been used to create maps that help to reduce the visual bias that can arise when mapping data where the subareas

vary in physical size. Areas that are physically large can visually dominate a map. The methodology involves deconvolution of the *variogram* obtained from areal data in order to construct a point support *variogram*. Area-to-point kriging is used to provide point support predictions. Population size variation is allowed in estimating the deconvoluted *variogram*. See Haining et al. (2010) for many references on the methodology including area to area kriging for irregularly shaped areas which can be used to tackle other change of support problems.

Geostatistics is also being used to model spatial variation in a dependent variable in terms of a set of independent variables where the data refer to irregular areas. Kerry et al. (2010) use the spatial components from area to area factorial *Poisson kriging* to identify the most important spatial scales at which crime rates vary and to identify which explanatory variables are statistically significant at those different scales. This represents another important extension of *geostatistical* theory, one that offers insights into the scale-dependent nature of relationships.

65.5.3 Bayesian Hierarchical Modeling

We conclude this section with comments on some new approaches to modeling spatial data. In the last 10–15 years, *Bayesian model*s have emerged as important tools in geography and regional science research, made possible by important breakthroughs in computational methods and the availability of inexpensive high-speed computers and of software for fitting spatial models (e.g., WinBUGS and facilities in R and MATLAB). In earlier years, spatial modeling was overwhelmingly *frequentist* or likelihood based: data values, $\mathbf{x}$, are assumed a random sample from $\mathbf{X}$, a random variable with a specified probability distribution depending on a set of fixed parameters, $\boldsymbol{\phi}$. The likelihood function for $\boldsymbol{\phi}$ given the data $\mathbf{x}$ is then defined, $L(\boldsymbol{\phi}|\mathbf{x})$, and parameter estimates are based on maximizing the likelihood function. Hypothesis testing is based on likelihood ratios for different values of $\boldsymbol{\phi}$. Inference is based on repeated *sampling*.

With *Bayesian* inference, however, the parameters are also random variables with their own distribution. This means that in addition to specifying the distribution of $\mathbf{X}$ for the observed data $\mathbf{x}$, it is necessary to also specify the distribution of $\boldsymbol{\phi}$, called the prior distribution, which depends on a further set of parameters. These parameters in turn can be modeled by prior distributions (hyper-priors). The combination of these conditional distributions produces the posterior distribution, and by *sampling* the posterior distribution, inference summaries can be obtained such as the posterior mean, credible intervals (the *Bayesian* version of the *frequentist*'s confidence interval), and probabilities of interest (such as the probability of a risk parameter exceeding a critical threshold). In *Bayesian* analysis, instead of handling spatial dependency effects in the data model for $\mathbf{X}$, which complicates the likelihood and often makes model fitting by maximum likelihood difficult (for an early discussion of this, see Whittle 1954), these effects can be handled in the prior distribution instead and fitted using the software referred to above. There are now many examples of this type of modeling (see, e.g., Le Sage 2000; Lu et al. 2007).

Specifying probability models in terms of a sequence of linked conditional models offers a means of modeling complex systems in ways that quantify the inherent uncertainties within scientific research relating to the data (level 1), the specification of the model (level 2), and model parameters (level 3). "Hierarchical statistical modelling represents a way to express uncertainties through well defined levels of conditional probabilities" (Cressie and Wikle 2011, p. 15). Cressie et al. (2009) provide a discussion and application of *hierarchical models* in *ecological analysis*.

Spatial effects are typically handled through a spatially structured random effects term as the following example illustrates. Suppose the researcher is modeling small area disease counts where x(i) is the number of cases in area i. The data model (level 1) specifies x(i) as the realization of a Poisson random variable (X(i)) with intensity parameter $\lambda(i) = E(i)\theta(i)$ where E(i) is the number of cases expected in area i given its population composition and $\theta(i)$ is the area-specific relative risk in area i. This level of the *hierarchical model* expresses the uncertainty in the data given the model specification including its parameters. At level 2, we define the model that reflects our understanding of what determines area level relative risk. For example, we may set

$$\mathrm{Log}[\lambda(i)] = \mathrm{Log}[E(i)] + b_0 + b_1 Z_1(i) + \ldots + b_k Z_k(i) + u(i) + s(i) \quad i = 1, \ldots, n \qquad (65.7)$$

where $Z_1, \ldots, Z_k$ define a set of k area-specific covariates with parameters b_1 to b_k that explain variation in relative risk and {u(i)} and{s(i)} are random effects. The {u(i)} are i.i.d. normal random effects, and the {s(i)} are given an *intrinsic conditional spatial autoregressive* (ICAR) specification (Haining 2003). These two terms model the scientific uncertainty in the model specification (e.g., competing theoretical understandings of the determinants of relative risk, our understanding of exposure to risk factors) as well as the effects of *overdispersion* and *spatial autocorrelation* in the spatial variation in relative risk and hence in the spatial distribution of the counts. At level 3, for a fully *Bayesian* analysis, the parameters at level 2 are treated as random variables and given probability distributions. As noted in an earlier section of this chapter, this could be extended to include the weights that define which areas are treated as *neighbors* in the *ICAR* specification. Choices about probability distributions could be informed by scientific understanding, but they might also be a way of allowing for uncertainty in our knowledge (Cressie et al. 2009). For an extension of these models to the multivariate case including multivariate spatial effects, see, for example, Gelfand and Vounatsou (2003).

65.6 Conclusions

One of the earliest items on the agenda of the USA's National Center for Geographic Information and Analysis (NCGIA) was spatial *data quality* emphasizing its fundamental importance to the development of good science. Understanding data uncertainty, arising from all the stages by which a complex geographical

reality is translated into spatial data, remains at the heart of good spatial science. In the light of the preceding comments, it should also link closely with modeling. At about the same time, attention was also being drawn to the importance of software development for spatial data analysis. In addition to progress in these two areas, the field of spatial data analysis has grown in many other ways. But static, cross-sectional in time, spatial data analysis is restricted to analyzing and modeling the "here and now" of some wider process. A series of spatial analyses over time can shed light on change but in other respects remains limited. The understanding that has been gained by the progress made in spatial statistics forms an essential element in the emergence of *spatiotemporal* data analysis. With the huge growth in space-time data sets and the potential they offer to advance scientific understanding, this represents one of the key areas for future growth.

References

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

Anselin L (2010) Thirty years of spatial econometrics. Pap Reg Sci 89:3–25

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc, B 36:192–225

Cliff AD, Ord JK (1973) Spatial Autocorrelation. Pion, London

Cressie N (1991) Statistics for spatial data. Wiley, New York

Cressie N, Wikle C (2011) Statistics for spatio-temporal data. Wiley, New York

Cressie N, Calder CA, Clark TS, Ver Hoef JM, Wikle CK (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical modelling. Ecol Appl 19:553–570

Elliott P, Richardson S, Abellan JJ, Thomson A, de Hoogh C, Jarup L, Briggs DJ (2009) Geographic density of landfill sites and risk of congenital abnormalities in England. Occup Environ Med 66:81–89

Fingleton B (2000) Spatial econometrics, economic geography, dynamics and equilibrium: a 'third way'? Environ Plan A 32:1481–1498

Fischer M, Wang J (2011) Spatial data analysis: models, methods and techniques. Springer, Heidelberg

Fotheringham S, Brunsdon C, Charlton M (2000) Quantitative geography: perspectives on spatial data analysis. SAGE, London

Gelfand A, Vounatsou P (2003) Proper multivariate conditional autoregressive models for spatial data. Biostatistics 4:11–25

Goodchild MG, Haining RP (2004) GIS and spatial data analysis: converging perspectives. Pap Reg Sci 83:363–385

Haining RP (1990) Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge

Haining RP (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge

Haining RP (2009) The special nature of spatial data. In: Fotheringham AS, Rogerson PA (eds) The SAGE handbook of spatial analysis. SAGE, Los Angeles, pp 5–24

Haining RP, Kerry R, Oliver M (2010) Geography, spatial data analysis and Geostatistics: an overview. Geogr Anal 42:7–31

Jones JP III, Casetti E (1992) Applications of the expansion method. Routledge, London

Kerry R, Goovaerts P, Haining RP, Ceccato V (2010) Applying geostatistical analysis to crime data: car-related thefts in the Baltic States. Geogr Anal 42:53–77

Kulldorff M (1997) A spatial scan statistic. Commun stat: theory methods 26:1481–1496
Le Sage J (2000) Bayesian estimation of limited dependent variable spatial autoregressive models. Geogr Anal 32:19–35
Lloyd CD (2011) Local models for spatial analysis. CRC Press, Boca Raton
Lu H, Reilly CS, Banerjee S, Carlin B (2007) Bayesian areal wombling via adjacency modelling. Environ Ecol Stat 14:433–452
Matheron G (1963) Principles of geostatistics. Econ Geol 58:1246–1266
Miller H, Han J (2009) Geographic data mining and knowledge discovery. CRC Press, Boca Raton
Neprash JA (1934) Some problems in the correlation of spatially distributed variables. J Am Stat Assoc 29(suppl):167–168
Oliver MA, Webster R, Lajaunie C, Mann JR, Muir KR, Parkes SE, Cameron AH, Stevens MCG (1998) Binomial cokriging for estimating and mapping the risk of childhood cancer. Math Med Biol 15:279–297
Paelinck J, Klaassen L (1979) Spatial econometrics. Saxon House, Farnborough
Ripley BD (1981) Spatial statistics. Wiley, New York
Waller LA (2009) Detection of clustering in spatial data. In: Fotheringham AS, Rogerson PA (eds) The SAGE handbook of spatial analysis. SAGE, Los Angeles, pp 299–320
Whittle P (1954) On stationary processes in the plane. Biometrika 41:434–449