Michael F. Goodchild
National Center for Geographic Information and Analysis
University of California
Santa Barbara, California

Geography and Statisticians

Introduction

Geography has always been a discipline of discovery. In Classical times it was a form of mathematics, concerned with measuring and projecting the earth and developing methods of navigation. From the 15th to the 19th Centuries it explored and mapped the Earth, and its practitioners were navigators and natural historians. Now that we know the source of the Nile, and new measurements of the height of Mt. Everest no longer capture the public imagination, geography has turned to the study of the processes which form and modify the physical landscape, to the dynamic and evolving relationships between humanity and the environment, and to the processes which shape the geographical diversity of human culture. Geography is infinitely complex: the more closely one looks at any of its aspects, human or physical, the more detail one sees and the more there is to explain and understand (Mandelbrot 1967).

Geography discovered statistics in its Quantitative Revolution which began in the 1950s and continued through the late 1960s. To most, this meant the application of the standard battery of statistical techniques, from student's t and F through principal components analysis and canonical correlation, to observed cases of some phenomenon. The fact that these cases were embedded in a spatio-temporal continuum was not given any great importance – space and time merely provided the sampling frame.

To a smaller group, it became increasingly apparent that geography was a special case. Some discovered spatial statistics, and used techniques such as point pattern analysis to make inferences about spatial processes (Getis and Boots 1978). Others explored spatial autocorrelation (Cliff and Ord 1981), random fields, regionalized variables or geostatistics (Isaaks and Srivastava 1989), and some made significant contributions. But these

are difficult areas, and have remained largely outside the coverage of standard courses in statistics for geographers. Even today, there is very little spatial statistics in the average undergraduate course text, which focuses on the application of the standard battery of (nonspatial) statistical tests to geographical problems (see for example Barber 1988; Clark and Hosking 1986; for a notable exception, see Griffith and Amrhein 1991).

Continuing developments that began in the 1980s are likely to change this situation, for the better. I will call these Geographic Information Systems (GIS), although the term is used in a much broader sense than the currently available set of GIS software packages (see for example Burrough 1986; Maguire, Goodchild and Rhind 1991). This paper is structured into four sections. The first identifies the areas of statistics that are of most relevance to geographical analysis and GIS. The second looks at some important issues in the application of statistics to geographical data in a GIS context. This is followed by a section that focuses on the use of statistical approaches to the problem of accuracy in spatial data, which is a key issue in the use of GIS technology. Finally, the last section looks at issues of implementation.

What Kinds of Statistics?

Perhaps the most striking property of geographical data from a statistical point of view is spatial dependence. Geographers might express this in the form: all places are related but nearby places are more related than distant places. It is almost impossible to conceive of a spatially independent variable, one whose value is unpredictable over the shortest distances. The variogram used in geostatistics shows the empirical variance between observations as a function of the distance between them, and is observed to rise monotonically to a "sill" at a distance termed the "range" and then fail to rise further, for most geographically distributed variables. This "range" is also known in geomorphology as the "grain" of the landscape, the distance beyond which the landscape fails to provide further surprises.

One way of understanding the effect of spatial dependence is as an inflation in the degrees of freedom of a test. An observation taken within a short distance of another one may not be a "new" observation, if the variable is strongly spatially dependent (Richardson 1990).

A second, again often ignored property is spatial heterogeneity (Anselin 1989). It is common for the coefficients of models to vary spatially, and it may be meaningless to try to estimate an "average" value for an arbitrarily defined study area. Rather, geographers may find more value in techniques such as adaptive spatial filtering, expansion methods (Casetti 1972) or geographical brushing (Monmonier 1989) that explicitly estimate the spatial variability of the model's coefficients. In this context, the conventional view that the study area provides a sample from some larger and normally unspecified population is highly inappropriate.

A third issue in the analysis of geographical data is the existence of hierarchical structures. Simple geographical objects may be members of larger, complex objects, and processes often operate across scales. In this sense, the view of a sample as a set of similar objects with associated attributes can be very limiting. From a database point of view, the "flat file" model which is implicit in many statistical tests places a severe restriction on our ability to understand spatial processes.

Finally, geographical space is inherently continuous and infinitely complex; a complete description would require an infinite number of records or "tuples". To describe geographical space in finite terms, some explicit discretization is required, through sampling, aggregation, generalization, modeling or approximation. Geographical data are thus almost always inaccurate, although the level of uncertainty associated with each item may sometimes be known. The objects that populate a spatial database – contour lines, sample points, patches on land cover maps, census tracts – may be no more than artifacts of this process of discretization, and their relationship to the underlying continuum may be difficult to characterize, or even unknown.

To summarize, geography needs statistics that explicitly recognize spatial dependence, and has great difficulty with the independence assumption made in so many statistical tests. Without them, the process of inference is likely to be misleading. It is interested in processes that are spatially non-stationary, and in techniques that can estimate the spatial variation of coefficients, since it is often unreasonable to assume stationarity within the arbitrary limits of a study area. It deals with processes that operate across scales, and with complex hierarchical relationships. And finally, many geographical observations represent arbitrary samplings or discretizations of continuous variation, rather than attributes of real objects.

These underlying problems often surface in concern over two specific issues: the Modifiable Areal Unit Problem and the Ecological Fallacy. Openshaw and Taylor (1979) provided perhaps the most dramatic illustration of the MAUP, when they showed that it was possible to obtain correlations ranging from –0.99 to +0.99 between two spatial variables simply by reaggregating to different reporting zones. Without spatial dependence or spatial heterogeneity, reporting zones would be analogous to samples of an underlying population, and their effects would be no different from sampling effects. But Openshaw and Taylor showed clearly that reporting zones have a much more significant effect, as gerrymandering politicians have known for a long time.

One commits an Ecological Fallacy (Robinson 1950) when one assumes that all components of a geographical aggregate have the properties of the aggregate as a whole. In contrast to the MAUP, the EF would never occur if reporting zones were always perfectly homogeneous. Yet it is made constantly, and the assumption of homogeneity is what lies behind the use of cluster techniques (Weiss 1988) and a number of other methods in market research. It would be easy to compute and report measures of diversity for census reporting zones, and this could be done without significant infringement on confidentiality; the practice of only reporting means, averages and counts must sometimes have the effect of encouraging census users to commit the Ecological Fallacy.

Application Issues

Until the early 1980s, very little was available in the way of digital technology for supporting spatial statistics. The statistical packages such as SAS, SPSS or S offered only non-spatial methods. Spatial information could be processed by computer mapping packages, but only for the purpose of making a map. Limited mapping capabilities also existed in such packages as SAS, but the spatial data input to make maps could not be used to support spatial analysis.

Geographic Information Systems became widely available beginning in the early 1980s, and have diffused rapidly within universities, government agencies and the private sector. Unlike computer mapping packages, their primary purpose is to support the analysis of spatial data, although their capabilities to do so are often limited to simple geometric operations.

Packages to support spatial statistics also appeared during the 1980s, although they have not had the same degree of impact as GIS. Some, such as the widely distributed GEOEAS package for geostatistics developed by the U.S. Environmental Protection Agency, or Anselin's SPACESTAT (Anselin 1990), are stand alone packages with limited mapping capabilities. Griffith (1988) has developed procedures for spatial statistics using SAS and MINITAB, and others have linked GIS to standard statistical packages or standalone modules (Ding and Fotheringham 1991). All are oriented toward carrying out standard tests on well-defined sets of input data.

GIS is a technology of the 1980s, and its underlying interactive paradigm bears little relationship to the batch processing of the 1960s. As such it is an ideal basis for support of an exploratory approach to statistical analysis. In fact there has been increasing interest recently in the concept of Exploratory Spatial Analysis, by analogy to the Exploratory Data Analysis paradigm that invaded

statistics in the late 1970s. It seems particularly inappropriate to adopt a lockstep confirmatory process given the difficulties imposed by the four characteristics of spatial data discussed earlier. Instead, a researcher working with geographical data should be encouraged to see the technology as a way of exploring a poorly formulated set of ideas. Visualization is important to this process, confidence limits may be more helpful than hypothesis tests, and randomization techniques may be more robust than the more traditional parametric tests.

Spatial Dependence

We assume that spatial dependence is always present in geographical data, and must be recognized explicitly unless the spacing of samples is greater than the range. If a test of spatial dependence (Cliff and Ord 1981) shows it to be absent, this inference will always constitute a Type II statistical error – the null hypothesis of an absence of spatial dependence will have been accepted when in fact it is false. Degrees of freedom will always be inflated.

This position is orthogonal to tradition, and particularly in courses where spatial dependence is taught as an exception or problem, and where absence of spatial dependence among residuals is often used as diagnostic of a fully-specified model. Unfortunately models of spatial dependence are difficult, in part because of the simultaneous nature of spatial dependence, which is in contrast to the one-directional nature of temporal dependence. Visualization may be the only way of communicating information about spatial dependence, since the parameters of models of spatial dependence have so little intuitive meaning.

Two families of techniques are available for dealing with spatial dependence, depending on the representation of space. The term "Spatial Statistics" normally refers to techniques that assume a space populated by well-defined objects that interact as wholes (Arbia 1989), even though they may actually be arbitrarily define objects serving as reporting zones for some underlying continuum. Measures of spatial dependence include the Moran and Geary measures of spatial autocorrelation (Cliff and Ord 1981), and focus on the level of dependence between neighboring objects. Autoregressive and moving average models are available to model processes between objects. Much currently available software makes use of standard packages (Griffith 1988). Finally, for these techniques the representation of space is reduced to a simple matrix of adjacencies, proximities or connectivities between objects, and results are invariant under any spatial transformations that preserve this matrix.

The term "Geostatistics" is used for techniques that model a continuous underlying distribution through point or block samples. Interpolation of this underlying surface is often the major objective. Variables are measured on continuous scales, and spatial dependence is described through the variogram or spectrum. Kriging is perhaps the best known technique of geostatistics (Isaaks and Srivastava 1989).

Other Methods of Spatial Statistics

Besides these, spatial statistics includes a range of tests of various models of pattern. In the simplest case, it is often desirable to test whether a pattern of points could have been generated by a given stochastic process. To identify clusters of cases in epidemiology, for example, one needs tests that can distinguish between spatially constant and spatially varying rates of disease, and in the latter case provide estimates of the geographic distribution of rates. In other cases it may be desirable to distinguish between clusters resulting from spatially varying rates, and those resulting from contagious processes. Openshaw's Geographical Analysis Machine (Openshaw *et al.* 1987) is a stimulating example of the use of GIS technology to test for clusters.

Numerous descriptive indices have been defined for geographical patterns. Perhaps the best known of these is the centroid, defined as the mean of the x and y coordinates of a point set, invariant under rotation of the axes, and the point that minimizes the total of squared distances to the point set. Movement of the centroid is a useful indicator of changes in geographical distributions. Centroids also provide the only available information on geographic distributions of population at the lowest level of each national system of reporting zones – the EA in Canada and the ED or block in the U.S. In the U.K., some recent work by Bracken and Martin (1989) is providing estimates of continuous distributions at very large scales by disaggregating centroid counts using intelligently specified kernel functions. Centroids and other measures of geographic central tendency also have normative value as logical places from which to provide service to a dispersed population. More sophisticated and appropriate techniques for central facilities location are discussed in a practical context by Ghosh and Rushton (1987).

Exploratory Spatial Analysis

Exploratory Data Analysis (EDA) is already well accepted as a paradigm in statistical analysis of nonspatial data. It provides a set of techniques for adding to the scientist's natural intuition, and for developing hypotheses that can later be checked more objectively and deliberately. So it seems reasonable to assume that there is a similar potential role for Exploratory Spatial Analysis, or ESA. A small number of papers have already explored the potential of ESA. In fact preliminary indications are that ESA is significantly different from EDA, and may have substantially more potential.

We define ESA here in comparative terms, as a set of techniques designed to improve on or sharpen the scientist's ability to reach inferences by exploring data in its spatial context. The most valuable kinds of ESA would therefore be ones designed to aid the researcher in areas of spatial reasoning where human intuition is not good. Some of these are obvious, and have been used to justify GIS in the past: the difficulty of combining evidence from different spatial sources, overcome in a GIS through the operation of overlay; or the difficulty of combining evidence of different forms – position (pattern), attributes, sound and text – when limited by the capabilities of a map display. The same argument for overlay can be made with respect to detecting change in a series of images: the eye is not good at coregistering and differencing, and thus a GIS can be a very effective tool at detecting change in a spatiotemporal time series. Nor is human intuition good at measuring spatial objects – a longstanding literature in cartography describes the problem of controlling the human eye's perception of the size of an object.

ESA tools can also improve on the eye's difficulty in integrating under a surface, to make estimates of total population within an arbitrary area from a contour map, for example. The reverse operation is also difficult – estimating a density surface from a map of dots. People are not good at reasoning between scales, or at testing patterns against statistical process models – it is easy to find numerous instances of intuition being misled into thinking that a random pattern of dots shows some systematic tendency. And finally, the eye is not good at separating regional from local trends, or detecting spatial outliers, particularly in multivariate data.

Haslett's SPIDER and similar systems (Haslett *et al.* 1991) make use of multiple screen windows to show data from different perspectives – map, time series, scatter plot, histogram – and then link them logically so that actions performed with the mouse in one window are reflected in the others. Pointing to an area on a map causes the corresponding point in a scatterplot to be highlighted. Animation of time series is also an effective but simple way to aid human intuition in gaining insight into spatial processes.

The Accuracy Problem

We have already noted that almost all spatial data are inherently uncertain and inaccurate. At the same time there is a longstanding tradition in cartography to establish standards for the accuracy of features shown on maps, particularly positional accuracy. Many regulate the distance within which the true location of a point must lie 90% of the time – the Circular Map Accuracy Standard (CMAS) – and set it to

a fraction of a millimeter on the finished map product. But for many types of geographical data it is impossible to measure CMAS because the objects being represented are artifacts of the process of discretization, and are not identifiable on the ground. CMAS is also concerned with point accuracy, and cannot be applied to complex line and area objects that are more than a collection of discrete points.

Substantial research has been devoted to developing error models for spatial databases. We define an error model as a stochastic process capable of replicating the distortions present in spatial data as a result of all sources of uncertainty. The variation between replications might be interpreted as the variation between alternative spatial database representations of the same geographical truth due to variation between observers, interpreters, measuring instruments or digitizers. The simplest such error model would be the map equivalent of the Gaussian distribution for scalar measurements – the distribution one would assume if one knew nothing whatever about the exact nature of the error process, other than that error is the additive combination of a large number of independent contributions.

Error modeling research is currently under way at a number of universities. All take the same basic approach: that uncertainty in the database can be described by an error model and associated parameters; that it is desirable to know the effects of this uncertainty in the form of confidence limits on the products from processing of the database; and that the mathematics of error analysis are complex, and therefore should be largely hidden from the user. At Newcastle, Openshaw and his group (Carver 1991) are working with a simple error band model that provides limits to the spatial deviation of every line in the database. At Utrecht, Burrough and his group (Heuvelink, Burrough and Stein 1989) use Monte Carlo simulation and Taylor series expansions to propagate error in the processing of rasters of continuous data. At Santa Barbara (Goodchild, Sun and Yang 1992), we have been using GRASS to simulate the propagation of error in operations on discrete multinomial land cover maps, based on a spatially autoregressive model of uncertainty.

Implementation Issues

If GIS provides the potential for more effective use of spatial statistics in the analysis of spatial data, then what developments are needed to bring this about? Progress to date has been limited, and there are still very few examples of effective spatial statistical research using GIS in the literature. At the same time we have a vast array of statistical software with very little spatial capabilities.

The GIS industry is large and flourishing – estimates of its size range from several hundred millions to several billions of dollars. But its market is largely in areas that require very little analysis – the management of vast collections of spatially distributed facilities by local governments, utility and natural resource companies and agencies. Thus recent developments in GIS software tend to have added much more to their capabilities as database systems than to their analytic power.

There seem to be three models for integration of GIS and techniques for statistical analysis. The first, termed here full integration, would require the addition of analytic capabilities to the GIS itself, and seems unlikely to occur without a significant reorientation of market priorities. The second, termed here loose coupling, is what we have at present. Data from GIS is communicated to analytic codes, either packages or standalone modules, through flat files of ASCII records. In this mode all of the higher level structures of spatial data – hierarchical relationships, adjacencies and complex geometries – are lost, and it is difficult (and may sometimes be impossible) to pass the results of analysis back to the GIS for mapping or storage. The third mode is termed close coupling, and results when the GIS and statistical package share common data models, thus preserving the full structure of the spatial data. Unfortunately at this time none of the readily available statistical packages recognize structures more complex than the simple flat file of records. On the other hand GIS tend to represent complex spatial structures through various implementations of the relational data model.

Priorities

If the pace of current developments in GIS, spatial statistics and ESA is anything to go by, the future looks far from bleak. We are likely to see the emergence of very exciting packages for the analysis of social and economic data, transportation planning, emergency facility management, epidemiology, and site selection, that implement many of our existing methods of analysis in much more effective ways through GIS technology. Multimedia and hypermedia techniques open up entirely new perspectives on spatial data, and invite new forms of visualization and inference that go far beyond the familiar hard copy map. At the same time intuition can be very misleading, and statistical analysis in a spatial context is far from simple and straightforward.

Given the potential, there is a real problem knowing where and how to begin. Spatial statistics and spatial analysis comprise a wide range of methods and there are few organizing principles available to simplify the set. We can distinguish to some extent between analysis and modeling, and between exploratory and confirmatory paradigms, but the most effective distinction from a GIS perspective lies in the data model – the set of objects that is created from the process of discretizing continuous geographic variation.

Spatial statistics relies on remarkably few data models, fewer in fact than the set currently supported by available GIS. Perhaps the simplest is the undifferentiated set of points, used for example in epidemiology and point pattern analysis. Another is the table of attributes plus matrix of proximities, used in the measurement of spatial dependence and in autoregressive and moving average models. More complex are models based on attributes for two sets of objects, plus a rectangular table of interactions or proximities, used for example in modeling flows and spatial interactions. Another is the set of point samples from a continuous distribution, and another is the raster of cells – both of these are common in the study of physical and environmental processes.

GIS has been successful in part because it has managed to integrate many of these data models into a common framework. But while the diversity of models may be a major benefit of GIS, it is a significant barrier to the development of spatial statistics and their integration with GIS. Most successful efforts at software integration rely on the existence of a simple, common data model, and the statistical packages have been built in this fashion around the simple flat file. GIS's main contribution to spatial statistics may turn out to be its insistence on an explicit recognition of data models as the latter's primary organizing principle.

Acknowledgement

The National Center for Geographic Information and Analysis is supported by the National Science Foundation, Grant SES 88-10917.

References

Anselin, L. 1989 'What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis' Technical Paper 89-4, National Center for Geographic Information and Analysis (University of California, Santa Barbara, CA)

Anselin, L. 1990 'SPACESTAT: A Program for the Statistical Analysis of Spatial Data' (Department of Geography, University of California, Santa Barbara, CA)

Arbia, G. 1989 **Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems** (Dordrecht: Kluwer)

Barber, G.M. 1988 **Elementary Statistics for Geographers** (New York: Guilford)

Bracken, I. and Martin, D. 1989 'The Generation of Spatial Population Distributions from Census Centroid Data' **Environment and Planning A** 21:537-543

Burrough, P.A. 1986 **Principles of Geographical Information Systems for Land Resources Assessment** (Oxford: Clarendon)

Carver, S. 1991 'Adding Error Handling Functionality to the GIS Toolkit' **Proceedings, EGIS 91** (Brussels)1:187-194

Casetti, E. 1972 'Generating Models by the Expansion Method: Applications to Geographical Research' **Geographical Analysis** 4(1):81-91

Clark, W.A.V. and Hosking, P.L. 1986 **Statistical Methods for Geographers** (New York: Wiley)

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

Ding, Y. and Fotheringham, A.S. 1991 'The Integration of Spatial Analysis and GIS: The Development of the Statcas Module for ARC/INFO' Technical Paper 91-5, National Center for Geographic Information and Analysis (University of California, Santa Barbara, CA)

Getis, A. and Boots, B.N. 1978 **Models of Spatial Processes: An Approach to the Study of Point, Line and Area Patterns** (Cambridge: Cambridge University Press)

Ghosh, A. and Rushton, G. 1987 **Spatial Analysis and Location-Allocation Models** (New York: Van Nostrand Reinhold)

Goodchild, M.F., Sun, G., and Yang, S. 1992 'Development and Test of an Error Model for Categorical Data' **International Journal of Geographical Information Systems** 6 (in press)

Griffith, D.A. 1988 'Estimating Spatial Autoregressive Model Parameters with Commercial Statistical Packages' **Geographical Analysis** 20:176-186

Griffith, D.A. and Amrhein, C.G. 1991 **Statistical Analysis for Geographers** (Englewood Cliffs, NJ: Prentice Hall)

Haslett, J., Bradley, R., Craig, P., Unwin, A. *et al.* 1991 'Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies' **The American Statistician** 45(3):234-242

Heuvelink, G.B.M., Burrough, P.A. and Stein, A. 1989 'Propagation of Errors in Spatial Modelling with GIS' **International Journal of Geographical Information Systems** 3:303-322

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

Maguire, D.J., Goodchild, M.F. and Rhind, D.W. (editors) 1991 **Geographical Information Systems: Principles and Applications** (London: Longman Scientific and Technical)

Mandelbrot, B.B. 1967 'How Long is the Coast of Britain? Statistical Self Similarity and Fractional Dimension' **Science** 156:636-638

Monmonier, M. 1989 'Geographic Brushing: Enhancing Exploratory Analysis of the Scatterplot Matrix' **Geographical Analysis** 21(1):81-84

Openshaw, S., Charlton, M., Wymer, C. and Craft, C. 1987 'A Mark I Geographical Analysis Machine for the Automated Analysis for Point Data Sets' **International Journal of Geographical Information Systems** 1:335-358

Openshaw, S. and Taylor, P.J. 1979 'A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem' in **Statistical Applications in the Spatial Sciences** edited by N. Wrigley (London: Pion):127-144

Richardson, S. 1990 'Some Remarks on the Testing of Association Between Spatial Processes' in **Spatial Statistics: Past, Present and Future** edited by D. Griffith (Ann Arbor, Michigan: IMaGe):277-309

Robinson, W.S. 1950 'Ecological Correlations and the Behavior of Individuals' **American Sociological Review** 15:351-357

Weiss, M.J. 1988 **The Clustering of America** (New York: Harper and Row)

Michael F. Goodchild
National Center for Geographic Information and Analysis
University of California
Santa Barbara, California

La géographie et les statisticiens

Introduction

La géographie a toujours été une discipline axée sur les découvertes. À l'époque classique, elle était une forme de mathématiques, où l'on s'intéressait à mesurer et à projeter la terre et à élaborer des méthodes de navigation. Du 15e au 19e siècles, la géographie a servi à explorer et à cartographier la terre et ses praticiens étaient des navigateurs et des chercheurs en histoire naturelle. Maintenant que nous connaissons la source du Nil et que les nouvelles mesures de l'altitude du Mont Everest ne captivent plus l'imagination du public, la géographie s'est tournée vers l'étude des processus qui façonnent et qui modifient le paysage physique, vers les liens dynamiques et évolutifs entre l'humanité et l'environnement ainsi que vers celle des processus qui forment la diversité géographique de la culture humaine. La géographie est infiniment complexe : plus on étudie de près l'un quelconque de ses aspects, humains ou physiques, plus on voit de détails et plus on trouve d'éléments à expliquer et à comprendre (Mandelbrot 1967).

La géographie a découvert les statistiques au cours de sa révolution quantitative qui a commencé dans les années 50 et qui s'est poursuivie jusqu'à la fin des années 60. Pour la majorité des personnes, cela signifie l'application de la batterie courante de techniques statistiques, allant du test t de Student et du test F à l'analyse en composantes principales et à la corrélation canonique, appliqués aux cas observés d'un phénomène. On n'a pas accordé une grande importance au fait que ces cas étaient noyés dans un continuum spatio-temporel – l'espace et le temps ne faisaient que fournir la base de sondage.

Pour un plus petit groupe de personnes, il est devenu de plus en plus apparent que la géographie constituait un cas spécial. Certaines personnes ont découvert la statistique spatiale et ont utilisé des techniques comme l'analyse de modèles de points pour faire des inférences à propos des processus spatiaux (Getis et Boots 1978). D'autres ont étudié l'autocorrélation spatiale (Cliff et Ord 1981), les champs aléatoires, les variables régionalisées ou la géostatistique (Isaaks et Srivastava 1989) et certaines personnes ont apporté des contributions importantes. Mais ces domaines sont difficiles et, généralement, ils n'ont pas été inclus dans les sujets traités dans les cours habituels de statistique destinés aux géographes. Même aujourd'hui, on traite très peu de la statistique spatiale dans le manuel typique utilisé dans les cours destinés aux étudiants du premier cycle, qui se concentre sur l'application de la batterie normale de tests statistiques (non spatiaux) aux problèmes géographiques (voir par exemple, Barber 1988; Clark et Hosking 1986; pour une exception marquée, voir Griffith et Amrhein 1991).

Il est probable que les progrès qui se poursuivent et qui ont commencé au cours des années 80 modifieront cette situation pour le mieux. Je vais désigner ces progrès par l'expression «systèmes d'information géographique» (SIG), bien que j'utilise ce terme dans un sens beaucoup plus étendu que celui qui s'applique à l'ensemble des progiciels de SIG disponibles actuellement (voir par exemple Burrough 1986; Maguire, Goodchild et Rhind 1991). La présente communication est divisée en quatre sections. Dans la première, on détermine les domaines de la statistique qui ont le plus de rapports avec l'analyse géographique et les SIG. Dans la deuxième section, on étudie certaines questions importantes portant sur l'application de la statistique aux données géographiques dans un contexte de SIG. Vient ensuite une section qui

porte sur l'utilisation des méthodes statistiques pour résoudre le problème de l'exactitude dans les données spatiales, ce qui constitue un problème-clé pour l'utilisation de la technologie des SIG. Finalement, la dernière section porte sur des questions de mise en application.

Quels genres de statistiques?

Il se peut que la dépendance spatiale constitue la propriété la plus frappante des données géographiques d'un point de vue statistique. Les géographes pourraient exprimer ce concept de la façon suivante : il existe un lien entre tous les endroits mais le lien est plus étroit entre des endroits rapprochés qu'entre des endroits éloignés. Il est presque impossible de concevoir une variable indépendante du point de vue spatial, une variable dont la valeur ne pourrait être prédite sur les distances les plus courtes. Le variogramme utilisé en géostatistique montre la variance empirique entre des observations comme fonction de la distance qui les sépare et on observe que cette variance augmente de façon monotone jusqu'à un «seuil» à une distance désignée par le mot «portée» puis cesse d'augmenter, pour la majorité des variables réparties géographiquement. Cette «portée» est aussi connue en géomorphologie comme la «trame» du paysage, la distance passée par laquelle le paysage cesse de fournir d'autres surprises.

Une façon de comprendre l'effet de la dépendance spatiale consiste à la considérer comme un gonflement des degrés de liberté d'un test. Il se peut qu'une observation prise à peu de distance d'une autre ne soit pas une «nouvelle» observation, s'il existe une forte dépendance spatiale pour la variable (Richardson 1990).

L'hétérogénéité spatiale (Anselin 1989) constitue une deuxième propriété qui, elle aussi, est souvent négligée. Il n'est pas rare que les coefficients des modèles varient spatialement et le fait d'essayer d'estimer une valeur «moyenne» pour une zone étudiée définie arbitrairement peut être dénué de sens. Il se peut plutôt que les géographes préfèrent utiliser des techniques telles que le filtrage spatial adaptatif, les méthodes de développement (Casetti 1972) ou le «brushing» géographique (Monmonier 1989) qui estiment explicitement la variabilité spatiale des coefficients du modèle. Dans ce contexte, l'opinion conventionnelle voulant que la zone étudiée constitue un échantillon tiré d'une population plus grande et normalement non précisée est très inappropriée.

L'existence de structures hiérarchiques constitue un troisième problème dans l'analyse des données géographiques. Des objets géographique simples peuvent faire partie d'objets complexes plus gros et il arrive souvent que les processus puissent s'appliquer quelle que soit l'échelle. Dans ce sens, il peut être très restrictif de considérer un échantillon comme un ensemble d'objets semblables avec des attributs associés. Du point de vue des bases de données, le modèle des «tables à deux dimensions» qui est implicite dans de nombreux tests statistiques limite considérablement notre aptitude à comprendre les processus spatiaux.

Finalement, l'espace géographique est, en soi, continu et infiniment complexe; sa description complète exigerait un nombre infini d'enregistrements ou de «tuples». Pour décrire l'espace géographique en termes finis, il faut effectuer une discrétisation explicite, à l'aide de l'échantillonnage, du groupement, de la généralisation, de la modélisation ou de l'approximation. Donc, les données géographiques sont presque toujours inexactes, bien qu'il se peut que le niveau d'incertitude associé à chaque article soit parfois connu. Les objets qui peuplent une base de données spatiales – les courbes de niveau, les points échantillons, les zones sur les cartes de la couverture végétale, les secteurs de recensement – peuvent n'être que des artefacts de ce processus de discrétisation et leur rapport avec le continuum sous-jacent peut être difficile à caractériser, ou même inconnu.

Pour résumer, en géographie on a besoin de statistiques qui reconnaissent explicitement la dépendance spatiale et on a beaucoup de difficultés avec l'hypothèse d'indépendance faite dans un si grand nombre de tests statistiques. Sans eux, le processus d'inférence sera vraisemblablement trompeur. La géographie s'intéresse à des processus qui ne sont pas stationnaires du point de vue spatial et à des techniques qui peuvent estimer la variation spatiale des coefficients, puisqu'il est souvent déraisonnable de supposer l'immobilité dans les limites arbitraires de la zone étudiée. En géographie, on utilise des processus qui peuvent être appliqués quelle que soit l'échelle et avec des relations hiérarchiques complexes. Et, finalement, de nombreuses observations géographiques représentent des échantillonnages ou des discrétisations arbitraires d'une variation continue plutôt que des attributs d'objets réels.

Ces problèmes sous-jacents surgissent souvent dans des préoccupations portant sur deux problèmes particuliers : le problème de l'unité spatiale modifiable (PUSM) (Modifiable Areal Unit Problem) et celui de l'erreur écologique (EE) (Ecological Fallacy). Openshaw et Taylor (1979) ont peut-être fourni l'illustration la plus spectaculaire du PUSM, quand ils ont montré qu'il suffisait de regrouper à nouveau, selon des zones de déclaration différentes, deux variables spatiales pour que les corrélations établies entre ces dernières aillent de –0,99 à +0,99. Sans dépendance ou hétérogénéité spatiale, les zones de déclaration seraient analogues à des échantillons d'une population sous-jacente et leurs effets ne différeraient pas des effets de l'échantillonnage. Mais Openshaw et Taylor ont montré clairement que les zones de déclaration ont un effet beaucoup plus important, ce que les politiciens qui ont recours au remaniement arbitraire des circonscriptions savent depuis longtemps.

On commet une erreur écologique (EE) (Robinson 1950) quand on suppose que toutes les composantes d'un agrégat géographique possèdent les propriétés de l'agrégat dans son ensemble. Par opposition au PUSM, l'EE ne se produirait jamais si les zones de déclaration étaient toujours parfaitement homogènes. Pourtant, on la fait constamment et l'hypothèse d'homogénéité est ce qui se trouve à la base de l'utilisation des techniques de groupement (Weiss 1988) et d'un certain nombre d'autres méthodes utilisées en étude de marché. On pourrait facilement calculer et déclarer des mesures de diversité pour les zones de déclaration du recensement et cela pourrait se faire sans contrevenir de façon importante aux règles de confidentialité; l'usage qui veut qu'on ne déclare que les mesures de la tendance centrale et les totaux doit parfois avoir l'effet d'encourager les utilisateurs des données du recensement à commettre l'erreur écologique.

Problèmes relatifs aux applications

Jusqu'au début des années 80, la technologie numérique offrait très peu de moyens pour traiter les statistiques spatiales. Les progiciels statistiques comme SAS, SPSS ou S n'offraient que des méthodes non spatiales. On pouvait traiter des renseignements spatiaux à l'aide de progiciels de cartographie par ordinateur, mais seulement pour produire une carte. Certains progiciels comme SAS possédaient aussi des fonctions limitées de cartographie, mais les données spatiales employées pour produire des cartes ne pouvaient être utilisées pour réaliser une analyse spatiale.

C'est au début des années 80 que les systèmes d'information géographique sont devenus disponibles sur une grande échelle et ils se sont répandus rapidement dans les universités, les organismes gouvernementaux et le secteur privé. Contrairement aux progiciels de cartographie par ordinateur, leur objet principal est d'effectuer l'analyse de données spatiales, bien qu'il arrive souvent que les fonctions dont ils disposent à cette fin soient limitées à des opérations géométriques simples.

C'est aussi au cours des années 80 que les progiciels permettant de traiter les statistiques spatiales sont apparus, bien qu'ils n'aient pas eu autant d'incidence que les SIG. Certains, comme le progiciel GEOEAS, largement répandu, utilisé à des fins de géostatique et élaboré par la Environmental Protection Agency des États-Unis, ou SPACESTAT de Anselin (Anselin 1990), sont des progiciels autonomes avec des fonctions de cartographie limitées. Griffith (1988) a élaboré des procédures pour traiter des statistiques spatiales à l'aide de SAS et

de MINITAB, et d'autres personnes ont élaboré des liens entre les SIG et des progiciels statistiques standard ou des modules autonomes (Ding et Fotheringham 1991). Tous ces progiciels visent à réaliser des tests standard sur des ensembles bien définis de données d'entrée.

Les SIG constituent une technologie des années 80 et leur modèle interactif sous-jacent a peu de rapports avec le traitement par lots des années 60. En soi, ils constituent une base idéale pour appuyer une approche exploratoire à l'analyse statistique. En fait, il y a récemment eu un intérêt croissant pour le concept d'analyse spatiale exploratoire, par analogie au modèle de l'analyse exploratoire des données, qui a envahi la statistique à la fin des années 70. Il semble particulièrement inapproprié d'adopter un processus de confirmation par étapes, compte tenu des difficultés imposées par les quatre caractéristiques des données spatiales déjà mentionnées. On devrait plutôt encourager les chercheurs travaillant avec des données géographiques à considérer la technologie comme une façon d'explorer un ensemble d'idées mal formulées. La visualisation est importante pour ce processus, les limites de confiance peuvent être plus utiles que les tests d'hypothèse et les techniques de randomisation peuvent être plus robustes que les tests paramétriques plus traditionnels.

Dépendance spatiale

Nous supposons que la dépendance spatiale est toujours présente dans les données géographiques et qu'elle doit être reconnue explicitement à moins que les échantillons soient distancés de plus que la portée. Si un test de dépendance spatiale (Cliff et Ord 1981) montre que cette dernière est absente, cette inférence constituera toujours une erreur statistique de deuxième espèce – l'hypothèse nulle d'une absence de dépendance spatiale aura été acceptée quand, en fait, elle est fausse. Les degrés de liberté seront toujours gonflés.

Cette position tranche par rapport à la tradition, en particulier par rapport à l'enseignement classique qui considère la dépendance spatiale comme une exception ou un problème et où l'absence de dépendance spatiale entre les résidus est souvent utilisée comme diagnostic d'un modèle entièrement spécifié. Malheureusement, les modèles de dépendance spatiale sont complexes, en partie à cause de la nature simultanée de la dépendance spatiale qui s'oppose à la nature unidirectionnelle de la dépendance temporelle. Il se peut que la visualisation constitue la seule façon de communiquer des renseignements sur la dépendance spatiale, parce que les paramètres des modèles de dépendance spatiale ont très peu de signification sur le plan intuitif.

On dispose de deux familles de techniques pour traiter la dépendance spatiale, selon la représentation de l'espace. L'expression «statistiques spatiales» se rapporte normalement aux techniques qui supposent un espace peuplé d'objets bien définis qui interagissent comme des touts (Arbia 1989), bien qu'il puisse en fait s'agir d'objets définis arbitrairement qui servent de zone de déclaration pour un continuum sous-jacent. Les mesures de la dépendance spatiale comprennent les mesures de Moran et Geary de l'autocorrélation spatiale (Cliff et Ord 1981) et se concentrent sur le niveau de dépendance entre des objets voisins. Des modèles autorégressifs et à moyenne mobile sont disponibles pour modéliser les processus entre des objets. Une bonne partie du logiciel qu'on peut employer actuellement utilise des progiciels standard (Griffith 1988). Finalement, pour ces techniques, la représentation de l'espace est réduite à une simple matrice de facteurs de contiguïté, de facteurs de proximité ou de facteurs de connectivité entre des objets et les résultats sont invariants pour toute transformation spatiale qui préserve cette matrice.

Le terme «géostatistique» est utilisé pour des techniques qui modélisent une distribution sous-jacente continue par tous les échantillons de points et d'îlots. L'interpolation de cette surface sous-jacente est souvent le principal objectif. Les variables sont mesurées sur des échelles continues et la dépendance spatiale est décrite à l'aide du variogramme ou du spectre. Le krigeage est peut-être la technique de géostatistique la plus connue (Isaaks et Srivastava 1989).

Autres méthodes utilisées en statistique spatiale

En plus de ces techniques, la statistique spatiale comprend une gamme de tests de divers modèles de structures. Dans le cas le plus simple, il est souvent souhaitable de faire un test pour savoir si une structure de points aurait pu être produite par un processus stochastique donné. Pour déterminer des grappes de cas en épidémiologie, par exemple, on doit disposer de tests qui permettent de faire la distinction entre des taux de maladies constants dans l'espace et d'autres qui varient dans l'espace et, dans ce dernier cas, les tests doivent pouvoir fournir des estimations de la répartition géographique des taux. Dans d'autres cas, il peut être souhaitable de faire la distinction entre des grappes qui découlent de taux qui varient spatialement et celles qui découlent de processus contagieux. La machine d'analyse géographique (Geographical Analysis Machine) d'Openshaw (Openshaw *et coll.* 1987) constitue un exemple stimulant d'utilisation de la technologie des SIG pour faire des tests portant sur des grappes.

De nombreux indices descriptifs ont été définis pour des structures géographiques. Il se peut que le plus connu de ces indices soit le centroïde, défini comme la moyenne des abscisses et des ordonnées d'un ensemble de points, invariant par rapport à la rotation des axes, ainsi que le point qui minimise le total des carrés des distances à l'ensemble de points. Le mouvement du centroïde constitue un indicateur utile des changements dans les répartitions géographiques. Les centroïdes fournissent aussi les seuls renseignements disponibles sur les répartitions géographiques de la population au plus bas niveau de chaque système national de zones de déclaration – le SD au Canada et le DD ou îlot aux É.-U. Au R.-U., des travaux récents effectués par Bracken et Martin (1989) fournissent des estimations de répartitions continues à de très grandes échelles en décomposant les chiffres pour des centroïdes à l'aide de fonctions noyaux précisées de façon intelligente. Les centroïdes et autres mesures de tendance centrale en géographie ont aussi une valeur normative comme endroits logiques à partir desquels on peut fournir un service à une population dispersée. Des techniques plus évoluées et plus appropriées pour déterminer l'emplacement d'installations centrales sont traitées, dans un contexte pratique, par Ghosh et Rushton (1987).

Analyse spatiale exploratoire

L'analyse exploratoire des données (AED) est déjà bien acceptée comme un modèle en analyse statistique de données non spatiales. Elle fournit un ensemble de techniques pour ajouter à l'intuition naturelle du chercheur et pour élaborer des hypothèses qui peuvent, par la suite, être vérifiées de façon plus objective et délibérée. Il semble donc raisonnable de supposer que l'analyse spatiale exploratoire ou ASE peut jouer un rôle semblable. Dans un petit nombre d'articles on a déjà étudié les possibilités de l'ASE. En fait, les indices préliminaires laissent supposer que l'ASE diffère considérablement de l'AED et qu'elle peut offrir beaucoup plus de possibilités.

Nous définissons ici l'ASE, en termes comparatifs, comme un ensemble de techniques conçues pour apporter des améliorations ou ajouter à l'aptitude d'un chercheur à tirer des conclusions après l'étude de données dans leur contexte spatial. Les genres les plus valables d'ASE seraient donc ceux conçus afin d'aider le chercheur dans des domaines du raisonnement spatial où l'intuition humaine est peu fiable. Certaines de ces analyses sont évidentes et ont déjà été utilisées pour justifier les SIG dans le passé : la difficulté à combiner des données provenant de différentes sources spatiales, surmontée dans un SIG à l'aide de l'opération de recouvrement; ou la difficulté à combiner des données de formes différentes – position (structure), attributs, son et texte – quand elle est limitée par les possibilités

d'un écran sur lequel est affiché une carte. Les raisons invoquées pour le recouvrement s'appliquent également dans le cas de la détection du changement dans une série d'images : l'oeil n'est pas un bon instrument pour enregistrer plusieurs types de données et pour les distinguer, un SIG peut donc se révéler un outil très efficace pour détecter un changement dans une série chronologique spatio-temporelle. L'intuition humaine non plus n'est pas bonne pour mesurer des objets spatiaux – des documents publiés de longue date en cartographie décrivent le problème qu'impose l'obligation de tenir compte de la perception par l'oeil humain de la taille d'un objet.

Les outils de l'ASE peuvent également diminuer la difficulté qu'a l'oeil à intégrer sous une surface, pour effectuer des estimations de la population totale dans une région arbitraire tirée d'une carte de courbes de niveau, par exemple. L'opération inverse aussi est difficile – l'estimation d'une surface de densité à partir d'une carte de points. Les personnes réussissent mal à faire des raisonnements d'une échelle à l'autre ou à effectuer l'essai de structures par rapport à des modèles de processus statistiques – il est facile de trouver de nombreux cas où l'intuition a amené des personnes à penser, à tort, qu'une structure aléatoire de points possédait une certaine tendance systématique. Et finalement, l'oeil n'est pas un bon outil pour séparer les tendances régionales des tendances locales ou pour détecter des valeurs spatiales aberrantes, particulièrement dans des données à plusieurs variables.

Le système SPIDER de Haslett et autres systèmes semblables (Haslett *et coll.* 1991) utilisent plusieurs fenêtres sur un écran pour montrer des données dans différentes perspectives – carte, série chronologique, nuage de points, histogramme – puis pour les coupler logiquement afin que les opérations effectuées à l'aide de la souris dans une fenêtre soient reflétées dans les autres. Le fait de désigner une région sur une carte entraîne la mise en évidence du point correspondant dans un nuage de points. L'animation de séries chronologiques constitue aussi un moyen simple mais efficace d'aider les êtres humains à obtenir une compréhension intuitive des processus spatiaux.

Le problème de l'exactitude

Nous avons déjà fait remarquer que presque toutes les données spatiales sont fondamentalement incertaines et inexactes. En même temps, il existe une tradition bien établie en cartographie qui vise à établir des normes portant sur l'exactitude des objets qui figurent sur les cartes, particulièrement l'exactitude de la position. De nombreux organismes réglementent la distance en dedans de laquelle l'emplacement véritable d'un point doit se trouver dans 90% des cas – la norme de précision cartographique circulaire (NPCC) – et la fixent à une fraction de millimètre sur le produit cartographique fini. Mais, pour de nombreux genres de données géographiques, il est impossible de mesurer la NPCC parce que les objets représentés sont des artefacts du processus de discrétisation et qu'ils ne peuvent être repérés sur le terrain. La NPCC porte aussi sur l'exactitude des points et elle ne peut être appliquée à des objets complexes, qu'il s'agisse de lignes ou de surfaces, qui sont plus qu'une collection de points discrets.

Beaucoup de recherches ont été consacrées à l'élaboration de modèles des erreurs pour des bases de données spatiales. Nous définissons un modèle des erreurs comme un processus stochastique qui peut reproduire les distorsions présentes dans les données spatiales par suite de toutes les sources d'incertitude. La variation entre les répétitions pourrait être interprétée comme la variation entre diverses représentations, dans des bases de données spatiales, de la même réalité géographique en raison d'une variation entre les observateurs, les interprètes, les instruments de mesure ou les numériseurs. Le plus simple de tels modèles des erreurs serait l'équivalent cartographique de la distribution gaussienne pour les mesures scalaires – la distribution que l'on supposerait si l'on ne connaissait rien à propos de la nature exacte du processus d'erreur, autre que cette erreur est la combinaison additive d'un nombre très grand de contributions indépendantes.

Dans un certain nombre d'universités, on réalise actuellement de la recherche en modélisation des erreurs. Toutes ces universités emploient la même approche fondamentale : l'incertitude dans la base de données peut être décrite par un modèle des erreurs et des paramètres qui y sont associés; il est souhaitable de connaître les effets de cette incertitude sous forme de limites de confiance sur les produits obtenus à la suite du traitement de la base de données et la mathématique de l'analyse des erreurs est complexe et devrait donc, dans une large mesure, être dissimulée à l'utilisateur. À Newcastle, Openshaw et son équipe (Carver 1991) travaillent avec un modèle de bande des erreurs simple qui fournit des limites à la déviation spatiale de chaque ligne dans la base de données. À Utrecht, Burrough et son équipe (Heuvelink, Burrough et Stein 1989) utilisent la simulation de Monte-Carlo et des développements en série de Taylor pour propager l'erreur dans le traitement de trames de données continues. À Santa Barbara (Goodchild, Sun et Yang 1992), nous avons utilisé le système GRASS afin de simuler la propagation des erreurs dans des opérations sur des cartes de la couverture végétale multinomiales discrètes, basées sur un modèle spatialement autorégressif de l'incertitude.

Problèmes relatifs à la mise en application

Si les SIG offrent la possibilité d'utiliser de façon plus efficace les statistiques spatiales dans l'analyse de données spatiales, alors de quels progrès avons-nous besoin pour que cela se produise? Jusqu'ici les progrès ont été limités et on trouve encore très peu de documents décrivant des exemples de recherche efficace en statistique spatiale qui emploient des SIG. En même temps, nous disposons d'une gamme étendue de logiciels statistiques qui possèdent très peu de fonctions spatiales.

L'industrie des SIG est considérable et florissante – les estimations de son importance vont de plusieurs centaines de millions à plusieurs milliards de dollars. Mais son marché se retrouve surtout dans des domaines qui exigent très peu d'analyse – la gestion, par les administrations locales, par les organismes et par les sociétés de services publics et d'exploitation de richesses naturelles, de vastes ensembles d'installations réparties dans des régions géographiques. Ainsi, les progrès récents dans le logiciel des SIG tendent à avoir augmenté beaucoup plus leurs possibilités comme systèmes de bases de données que leur puissance analytique.

Il semble exister trois modèles pour l'intégration des SIG et des techniques d'analyse statistique. Le premier, que l'on appellera ici «intégration complète», exigerait l'addition de fonctions analytiques aux SIG eux-mêmes, il semble peu probable que cela se produise sans une réorientation importante des priorités du marché. Le deuxième, que nous appellerons ici «couplage relâché» est ce dont nous disposons actuellement. Les données provenant des SIG sont communiquées à des programmes analytiques, soit des progiciels, soit des modules autonomes par l'intermédiaire de tables à deux dimensions d'enregistrements en caractères ASCII. Dans ce mode, toutes les structures de niveau plus élevé des données spatiales – les rapports hiérarchiques, les contiguïtés et les géométries complexes – sont perdues et il est difficile (et peut être parfois impossible) de retransmettre au SIG les résultats de l'analyse pour cartographie ou stockage. Le troisième mode est appelé «couplage étroit» et il se produit quand le SIG et le progiciel statistique utilisent des modèles de données communs, préservant ainsi la structure complète des données spatiales. Malheureusement, aucun des progiciels statistiques que l'on peut se procurer facilement à l'heure actuelle ne reconnaît de structures plus complexes que la simple table à deux dimensions d'enregistrements. Par contre, les SIG tendent à représenter des structures spatiales complexes à l'aide de diverses

mises en application du modèle de données relationnelles.

Priorités

Si l'on se fie à l'évolution actuelle des SIG, des statistiques spatiales et de l'ASE, l'avenir est loin d'être sombre. Nous verrons probablement apparaître des progiciels très intéressants, pour l'analyse des données sociales et économiques, pour la planification des transports, pour la gestion des installations de crise, pour l'épidémiologie et pour le choix des sites d'implantation, qui mettent en application un grand nombre des méthodes d'analyse dont nous disposons actuellement, de façon beaucoup plus efficace, par l'intermédiaire de la technologie des SIG. Les techniques des multimédias et des hypermédias ouvrent des perspectives totalement nouvelles sur les données spatiales et invitent de nouvelles formes de visualisation et d'inférence qui dépassent de beaucoup la carte familière sur support papier. En même temps, l'intuition peut être très trompeuse et l'analyse statistique dans un contexte spatial est loin d'être simple et directe.

Compte tenu des possibilités, le fait de savoir où commencer et comment s'y prendre pose un problème réel. Les statistiques et l'analyse spatiales comprennent une gamme étendue de méthodes et on dispose de peu de principes d'organisation pour simplifier cet ensemble. Nous pouvons, dans une certaine mesure, faire la distinction entre l'analyse et la modélisation et entre les modèles exploratoires et corroboratifs, mais la distinction la plus efficace, dans la perspective des SIG, se trouve dans le modèle de données – l'ensemble d'objets qui est créé à partir du processus de discrétisation de la variation géographique continue.

Les statistiques spatiales reposent sur un nombre remarquablement faible de modèles de données, moins, en fait, que l'ensemble qui peut actuellement être utilisé par les SIG dont on dispose. Il se peut que le plus simple de ces modèles de données soit l'ensemble de points non différencié utilisé, par exemple, en épidémiologie et en analyse de modèles de points. Un autre de ces modèles est la table d'attributs plus la matrice de facteurs de proximité utilisés pour mesurer la dépendance spatiale et dans les modèles autorégressifs et à moyenne mobile. Les modèles basés sur des attributs pour deux ensembles d'objets plus un tableau rectangulaire d'interactions ou de facteurs de proximité utilisés, par exemple, pour modéliser les flux et les interactions spatiales constituent des modèles plus complexes. Un autre de ces modèles est l'ensemble d'échantillons de points tirés d'une distribution continue et un autre est la trame de cellules – ces deux ensembles sont utilisés couramment dans l'étude de processus physiques et environnementaux.

Les SIG ont connu du succès en partie parce qu'ils ont réussi à intégrer un bon nombre de ces modèles de données dans un cadre commun. Mais, alors que la diversité des modèles peut être un des principaux avantages des SIG, elle constitue un obstacle important à l'élaboration de statistiques spatiales et à leur intégration avec les SIG. La majorité des efforts d'intégration au niveau du logiciel, qui ont été couronnés de succès, sont basés sur l'existence d'un modèle de données commun simple et les progiciels statistiques ont été établis de cette façon autour de la table simple à deux dimensions. Il se pourrait que la principale contribution des SIG à la statistique spatiale soit le fait que, dans ces systèmes, on insiste sur une reconnaissance explicite des modèles de données comme leur principe d'organisation fondamental.

Remerciements

Le National Center for Geographic Information and Analysis est appuyé par la National Science Foundation, subvention SES 88-10917.

Bibliographie

Anselin, L. 1989 'What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis' Rapport technique 89-4, National Center for Geographic Information and Analysis (University of California, Santa Barbara, CA)

Anselin, L. 1990 'SPACESTAT: A Program for the Statistical Analysis of Spatial Data' (Department of Geography, University of California, Santa Barbara, CA)

Arbia, G. 1989 **Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems** (Dordrecht: Kluwer)

Barber, G.M. 1988 **Elementary Statistics for Geographers** (New York: Guilford)

Bracken, I. et Martin, D. 1989 'The Generation of Spatial Population Distributions from Census Centroid Data' **Environment and Planning A** 21:537-543

Burrough, P.A. 1986 **Principles of Geographical Information Systems for Land Resources Assessment** (Oxford: Clarendon)

Carver, S. 1991 'Adding Error Handling Functionality to the GIS Toolkit' **Proceedings, EGIS 91** (Brussels) 1:187-194

Casetti, E. 1972 'Generating Models by the Expansion Method: Applications to Geographical Research' **Geographical Analysis** 4(1):81-91

Clark, W.A.V. et Hosking, P.L. 1986 **Statistical Methods for Geographers** (New York: Wiley)

Cliff, A.D. et Ord, J.K. 1981 **Spatial Processes: Models and Applications** (London: Pion)

Ding, Y. et Fotheringham, A.S. 1991 'The Integration of Spatial Analysis and GIS: The Development of the Statcas Module for ARC/INFO' Rapport technique 91-5, National Center for Geographic Information and Analysis (University of California, Santa Barbara, CA)

Getis, A. et Boots, B.N. 1978 **Models of Spatial Processes: An Approach to the Study of Point, Line and Area Patterns** (Cambridge: Cambridge University Press)

Ghosh, A. et Rushton, G. 1987 **Spatial Analysis and Location-Allocation Models** (New York: Van Nostrand Reinhold)

Goodchild, M.F., Sun, G. et Yang, S. 1992 'Development and Test of an Error Model for Categorical Data' **International Journal of Geographical Information Systems** 6 (sous presse)

Griffith, D.A. 1988 'Estimating Spatial Autoregressive Model Parameters with Commercial Statistical Packages' **Geographical Analysis** 20:176-186

Griffith, D.A. et Amrhein, C.G. 1991 **Statistical Analysis for Geographers** (Englewood Cliffs, NJ: Prentice Hall)

Haslett, J., Bradley, R., Craig, P., Unwin, A. *et al.* 1991 'Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies' **The American Statistician** 45(3):234-242

Heuvelink, G.B.M., Burrough, P.A. et Stein, A. 1989 'Propagation of Errors in Spatial Modelling with GIS' **International Journal of Geographical Information Systems** 3:303-322

Isaaks, E.H. et Srivastava, R.M. 1989 **Applied Geostatistics** (Oxford University Press)

Maguire, D.J., Goodchild, M.F. et Rhind, D.W. (éditeurs) 1991 **Geographical Information Systems: Principles and Applications** (London: Longman Scientific and Technical)

Mandelbrot, B.B. 1967 'How Long is the Coast of Britain? Statistical Self Similarity and Fractional Dimension' **Science** 156:636-638

Monmonier, M. 1989 'Geographic Brushing: Enhancing Exploratory Analysis of the Scatterplot Matrix' **Geographical Analysis** 21(1):81-84

Openshaw, S., Charlton, M., Wymer, C. et Craft, C. 1987 'A Mark I Geographical Analysis Machine for the Automated Analysis for Point Data Sets' **International Journal of Geographical Information Systems** 1:335-358

Openshaw, S. et Taylor, P.J. 1979 'A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem' dans **Statistical Applications in the Spatial Sciences** édité par N. Wrigley (London: Pion):127-144

Richardson, S. 1990 'Some Remarks on the Testing of Association Between Spatial Processes' dans **Spatial Statistics: Past, Present and Future** édité par D. Griffith (Ann Arbor, Michigan: IMaGE):277-309

Robinson, W.S. 1950 'Ecological Correlations and the Behavior of Individuals' **American Sociological Review** 15:351-357

Weiss, M.J. 1988 **The Clustering of America** (New York: Harper and Row)