Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

KEYNOTE ADDRESS: GEOGRAPHY AND STATISTICIANS

M.F. Goodchild[1]

ABSTRACT

Spatial data present unique problems to statistics because of their inherent properties, particularly of spatial heterogeneity and spatial dependence. The paper reviews the practical implications of these properties in such areas as modifiable reporting zones and ecological fallacies. It looks at the development of the spatial tradition in statistics, and the statistical tradition in geography. Current interest in geographic information systems offers an opportunity to bring the two groups closer together, and to address longstanding issues in a systematic and practical manner. New developments in exploratory spatial analysis and visualization add to the impact of digital technology and enhance the value of the spatial perspective.

KEY WORDS: Spatial data; Spatial statistics; Thematic mapping; GIS.

1. 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 exceptions, see Griffith and Amrhein 1991).

---

[1] M.F. Goodchild, Director, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA 93106, U.S.A.

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.

2. 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.

3. 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.

3.1 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).

3.2 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).

3.3 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.

4. 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.

5. 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.

6. 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.

ACKNOWLEDGMENT

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? Report 89-4, Santa Barbara, CA: *National Center for Geographic Information and Analysis*.

Anselin, L. (1990). SPACESTAT: A program for the statistical analysis of spatial data, Santa Barbara, CA: Department of Geography, University of California.

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. Report 91-5, Santa Barbara, CA: National Center for Geographic Information and Analysis.

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, N.J.: Prentice Hall.

Haslett, J., Bradley, R., Craig, P., Unwin, A., and others (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *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: Oxford University Press.

Maguire, D.J., Goodchild, M.F., and Rhind, D.W. (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 N. Wrigley (ed.) *Statistical Applications in the Spatial Sciences*, London: Pion, 127-144.

Richardson, S. (1990). Some remarks on the testing of association between spatial processes, *Spatial statistics: Past, Present and Future*, (ed.) 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.