# WHAT PROBLEM? SPATIAL AUTOCORRELATION AND GEOGRAPHIC INFORMATION SCIENCE

Michael F. Goodchild[1]

With the benefit of 40 years of hindsight, it is the second word of the title that strikes me as most remarkable about the original Cliff and Ord paper (Cliff and Ord, 1969). Why was spatial autocorrelation perceived in 1969 as a *problem*, and has that perspective changed over the past 40 years, particularly given developments in geographic information science? To examine this question it is necessary to go back further, to the origins of statistics, and to the issues involved in applying statistical methods in a geographic context. Today our understanding of the nature of geographic information has advanced remarkably, and with it has come a much more nuanced view of the value of statistical methods, and more specifically of statistical inference. In this essay I explore these issues in detail, building on advances in geographic information science, and ending with some conclusions regarding the appropriate training of spatial analysts.

## CONTROLLED AND NATURAL EXPERIMENTS

A psychologist looking for patterns and effects in the operation of the human brain clearly makes the most valuable contribution when those effects can be demonstrated to apply to the entire human population. Similarly a geneticist examining the expression of genes, or a physicist measuring the properties of atoms, would like the results of those

[1] Center for Spatial Studies, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA. Phone +1 805 893 8049, FAX +1 805 893 3146, Email good@geog.ucsb.edu

studies to be as general as possible. General, otherwise known as nomothetic knowledge is always valued more highly by science than specific or idiographic knowledge, but in all three of these examples it would be impossible for the researcher to demonstrate generality with complete confidence. The psychologist would have to study every human brain, including past and future brains, and instead is likely to rely on samples, taking the results of a few examinations of current brains and attempting to generalize beyond the sample to the larger population.

The apparatus of inferential statistics was developed to provide a sound logical and mathematical basis for this process, of course. The sample must be representative of the population, because if it is not there is the possibility of bias in the results, and for the sample to be representative it is necessary that every member of the population have an equal and independent chance of being chosen. Seeds sown in an agricultural experiment, for example, must be representative of the larger population of seeds that will eventually be used in practice; subjects chosen for a psychological experiment must be representative of the broader population whose inferred properties will eventually be discussed by the researcher; and the errors present in the sample of measurements acquired by a measuring instrument must be representative of the errors generally produced by the instrument. In all of these cases the researcher controls the choice of samples, conceptualizes the population, and reviews the assumptions inherent in the process of inference.

The use of statistical inference spread rapidly across the sciences in the 20th century, and departments of statistics were established in many universities. Generations of students were trained in these methods, and it became standard practice to submit any numerical result to inferential testing. In some cases this was done using confidence limits, by computing the uncertainty that would occur when a numerical result was generalized to a population. In other cases the emphasis was placed on the probability associated with a counter or null hypothesis. In the case of relationships between variables, for example, the common null hypothesis is the lack of relationship, or coefficients equal to zero in the population. Three measures result: the strength of the relationship in the sample; confidence limits on that strength when the sample is generalized to the population; and 1 minus the probability that such a strength could arise in the sample despite a lack of relationship in the population, commonly termed the relationship's significance.

These methods have become so pervasive that it is common to apply them to all data, however appropriate the underlying assumptions. For example, it is easy to find in the literature instances of the application of inferential methods to samples when no concept of a population has been articulated, or when samples were clearly not drawn randomly and independently from such a population. These problems are particularly common in so-called *natural* experiments, or analyses of data that arise in the real world without any control by the researcher. Rather than select a random sample of US counties, for example, a researcher might examine relationships among census variables for all US counties, since the data are readily available and there is no need to sample in the interests of time or expense. But what population have these samples been drawn from,

and was the process of drawing representative? If not, the assumptions on which the logic of statistical inference is based are clearly invalid.

Here of course is explanation for the use of the word "problem" in the Cliff and Ord paper. Counties are largely arbitrarily defined reporting zones, with boundaries that cut across the grain of landscape organization, and it is impossible to argue convincingly that two counties have been drawn independently from some larger population, especially when the counties are adjacent. Three options might be considered at this point. First, one might limit the analysis to a truly random sample of the roughly 3100 US counties, limiting the number and spacing them far enough apart so that the assumption of independence is tenable. But this means rejecting the vast majority of the data, widening the confidence limits of the conclusions, and possibly missing some important patterns and effects. Second, one might build spatial effects directly into the model, changing its algebraic form so that spatial dependencies are explicitly accounted. This option has been explored extensively (Anselin, 1988; Arbia, 2006; Cressie, 1993; Griffith, 1988; Haining, 2003), and researchers now have access to a rich resource of alternative formulations. Third, one might reject inference altogether, arguing that what US counties reveal is true only of US counties, and that generalization to some hypothetical and undefined population has no merit. This last option is the most problematic, since it flies in the face of the inferential tradition, despite the straightforward and compelling arguments that support it.

GEOGRAPHIC INFORMATION SCIENCE

The rise of GIS (geographic information systems) beginning in the 1960s has brought with it a new set of perspectives on spatial autocorrelation. To develop digital representations of the phenomena distributed over the Earth's surface, and to design the systems that can store, process, and edit those representations requires a degree of understanding of the inherent nature of those phenomena. Much of this knowledge has its roots in map-making, and the practices of cartography that developed over centuries, but the digital environment of GIS, with its formal coding systems, has brought a new level of focus and rigor.

The first GIS, developed by the Government of Canada and IBM in the mid 1960s (Foresman, 1998), used a single, uniform representation that partitioned geographic areas into irregular zones of approximately homogeneous type. Land use, for example, provided one basis for partitioning, as did land capability for agriculture. These so-called *area-class maps* (Mark and Csillag, 1989) are in effect mappings from location **x** to class $c(\mathbf{x})$, and rely on the fact that classes assigned to nearby points tend to be positively correlated, thus allowing zones to emerge of useful size and approximate homogeneity. Geostatistics, or the theory of regionalized variables (Goovaerts, 1997), provides one formal framework for such data, since it focuses on spatial autocorrelation as a monotonically decreasing function of distance. In this instance spatial autocorrelation is not a problem, but a fortunate characteristic of a wide range of spatially distributed phenomena.

Over the ensuing years GIS developed into the comprehensive solution to the representation of many types of phenomena that we see today. Surfaces, more formally known as interval/ratio fields in contrast to the nominal and ordinal fields of area-class maps, are mappings from location **x** to a value $z(\mathbf{x})$, and are typified by the surface of topographic elevation. They can be represented in numerous ways, including digital models of isolines, and again the inherent spatial autocorrelation of such surfaces provides the essential economies that allow complex surfaces to be represented in manageable volumes. Spatial interpolation, the widespread practice of inferring the values of fields from sample points, relies entirely on the almost universal existence of positive spatial autocorrelation in such phenomena. Again, spatial autocorrelation is not a problem but the basis of a solution.

While geostatistics and spatial statistics provide the theoretical framework, the observation that spatial autocorrelation is endemic in geographic information is more likely to be discussed loosely as *Tobler's First Law of Geography*. In a 1970 paper (Tobler, 1970) Waldo Tobler noted that the proposition "nearby things are more similar than distant things" might provide a simple, universal basis for estimating missing data values in geographic information. All methods of spatial interpolation use this principle (abbreviated as TFL) in some way, whether formally through Kriging or procedurally in methods such as inverse-distance weighting. In effect it amounts to an empirical law (Goodchild, 2004), and its generality makes the occasional exceptions particularly interesting.

Anselin (1989) was perhaps the first to generalize away from this case, and to ask the broader question "What is special about spatial?" Besides TFL, he argued that *spatial heterogeneity* was also pervasive, in other words that like many time series the Earth's surface exhibited uncontrolled variance or non-stationarity. Like the weather, geographic phenomena do not oscillate around a mean, but drift from one locally average condition to another. If this were not so, European explorers would not have found "new" worlds, but new samples of the old. Because of spatial heterogeneity it is impossible to conceive of an average location on the Earth's surface, and one cannot understand its full range simply by extrapolating from local conditions.

Like spatial autocorrelation, spatial heterogeneity poses a substantial "problem" for traditional statistical inference, in this case because a sample taken in a limited area cannot be used to generalize about conditions everywhere. Similarly it implies that the results of any study of a limited area depend explicitly on the bounds of the area: shift the bounds, and the conclusions will change. Generalization in the geographic context clearly is not the same as generalization from controlled experiments.

In recent years a series of papers have described various *place-based* analytic techniques that are specifically designed to explore spatially heterogeneous phenomena (Fotheringham and Brunsdon, 1999). They include Anselin's LISA (Anselin, 1995) and the various forms of bivariate and multivariate analysis developed by Fotheringham and colleagues, building on the original Geographically Weighted Regression (Fotheringham, Brunsdon, and Charlton, 2002). Whether or not one believes in the value of exploring

spatial heterogeneity, these techniques seem to reflect a fundamental problem in the environmental and social sciences. Although it is traditional to see perfect explanation as the goal of modeling, in reality such a goal is clearly unattainable with respect to the complex phenomena dealt with in the environmental and social disciplines. Models in these fields will always be underspecified, and it seems likely that the residual variation will be at least partially spatial. Thus GWR and related techniques seem a perfectly reasonable response to the reality of complex sciences.

I have suggested that several other empirical "laws" of geographic information might be formulated, besides TFL and spatial heterogeneity (Goodchild, 2004), and that the search for such principles is a suitable goal of geographic information science. They could have practical value in the design of geographic information systems and databases, allowing designers to take advantage of systematic properties to create efficient storage structures and algorithms.

## BROADER IMPLICATIONS

I have argued, in effect, that spatial autocorrelation is a problem only because the methods of statistical inference developed for the analysis of controlled experiments do not transfer well to the world of natural experiments, and thus to the normal context of geographic research. We insist that our students learn such methods, and then point out to them that the context in which they will be applied is exceptional and problematic. Yet from that perspective the world of controlled experiments is itself the exception, being limited to those situations in which the researcher is able to identify a population, select

cases from it randomly and independently, and then make inferences about the population. This ideal situation occurs so rarely in geographic research that one wonders why the associated assumptions have persisted, and indeed dominated, for so long. Why are we not content simply to describe specific parts of the heterogenous world that we see around us, using the rigorous methods of science, and why do we insist instead that inferences be made about some poorly conceived and non-existent hypothetical world?

Scale economies might provide one explanation. It makes good sense to develop a one-size-fits-all approach to scientific inference, based on a single data model (the rectangular table of cases and variables) and a single set of assumptions, and to offer a single sequence of courses in statistics to all comers. The discipline of geography has rarely had access to the kinds of resources needed to mount special courses of its own, and has often relied instead on courses offered by others.

Another explanation might lie in the sheer complexity of spatial methods. It is not easy to develop successful models of spatial effects, or to adapt statistical methods to the problem of spatial autocorrelation. But the approach I have suggested above, that of ignoring inference on the grounds that the world is fundamentally heterogeneous, leads to a set of techniques that are far simpler, and avoids the complexities of spatial effects almost entirely.

However one explains the past, the growth of spatial methods and technologies over the past few years, and the widespread availability of massive spatial databases, suggest that

the time has come for a more aggressive approach. We devote large amounts of time as a society to the development of mathematical and verbal intelligence, but almost none to the kinds of critical thinking necessary to work effectively with spatial concepts and with phenomena embedded in space and time (NRC, 2006). It would make excellent sense to reorganize our curricula, to ensure that students learn the realities of geographic research at the outset, rather than proceeding first to adopt unrealistic assumptions, then to encounter the "problem" or problems that make these assumptions untenable, and then finally to achieve spatial understanding, once the cohort has been reduced to a small minority of survivors.

REFERENCES

Anselin, L., 1988. *Spatial Econometrics: Methods and Models*. Boston: Kluwer.

Anselin, L., 1989. *What Is Special about Spatial Data? Alternative Perspectives on Spatial Data Analysis*. Technical Report. Santa Barbara: National Center for Geographic Information and Analysis.

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

Arbia, G., 2006. *Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence*. New York: Springer.

Cliff, A.D. and J.K. Ord, 1969. The problem of spatial autocorrelation. In A.J. Scott, editor, *London Papers in Regional Science*, pp. 25-55. London: Pion.

Cressie, N.A.C., 1993. *Statistics for Spatial Data*. New York: Wiley.

Foresman, T.W., 1998. *The History of Geographic Information Systems: Perspectives from the Pioneers*. Upper Saddle River, NJ: Prentice Hall.

Fotheringham, A.S. and C. Brunsdon, 1999. Local forms of spatial analysis. *Geographical Analysis* 31: 340-358.

Fotheringham, A.S., C. Brunsdon, and M.E. Charlton, 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Hoboken, NJ: Wiley.

Goodchild, M.F., 2004. The validity and usefulness of laws in geographic information science and geography. *Annals of the Association of American Geographers* 94(2) 300-303.

Goovaerts, P., 1997. *Geostatistics for Natural Resources Evaluation*. New York: Oxford.

Griffith, D.A., 1988. *Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series*. Boston: Kluwer.

Haining, R.P., 2003. *Spatial Data Analysis: Theory and Practice*. New York: Cambridge University Press.

Mark, D.M. and F. Csillag, 1989. The nature of boundaries on 'area-class' maps. *Cartographica* 26: 65-77.

National Research Council, 2006. *Learning to Think Spatially: GIS as a Support System in the K-12 Curriculum*. Washington, DC: National Academies Press.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit Region. *Economic Geography* 46: 234-240.