INTELLIGENT SYSTEMS FOR GISCIENCE: WHERE NEXT?

A GISCIENCE PERSPECTIVE

The SOM algorithm was originally designed to explore complex multidimensional data spaces describing objects that may or may not be located in geographic space, and whose locations may or may not be known. The subject matter of GIScience is thus only a small subset of the subject matter of SOM, since GIScience starts with the assumption that all objects of interest are georeferenced, and goes on to address fundamental issues underlying such information: its nature, representation, storage, handling, analysis, visualization, and modeling (for a recent review of the definitions and content of GIScience see Mark, 2003).

The chapters of this book have explored many facets of what is clearly a complex relationship between SOM and GIScience. In Chapter 2, Bação, Lobo, and Painho show how the basic SOM algorithm can be modified to recognize location explicity, and used to solve certain longstanding problems in GIScience, including regionalization and service area design, through modifications to the basic algorithm. Other authors see SOM in much the same way as its originator and early proponents, as a method for exploring complex multidimensional data sets, and for operationalizing the fundamental scientific task of classification. If the basic objects of analysis are georeferenced, then classes can be mapped, as Kropp and Schellnhuber do, for example, in Chapter 13.

Other authors see SOM as a means of *spatializing* data, in other words organizing objects in a space defined by their similarities, in effect enlisting SOM as a technique for adding locational references to data that are not inherently spatial, and thus extending the techniques of GIScience to spaces other than the geographic. Such applications represent a remarkable implementation of Tobler's First Law of Geography (Tobler, 1970, the empirical tendency for objects that are nearby in geographic space to be more similar than objects that are distant in geographic space, by insisting that an information space created by spatialization have the properties that we as humans recognize to be true of geographic space. Montello *et al.* (2003) have proposed a First Law of Cognitive Geography, based on the observation that people *think* things are more similar if they are near each other, a finding that provides direct support for the logic of spatialization as a tool for visualizing complex multidimensional data sets.

Many years ago Tobler and Wineberg (1971) provided yet another motivation for techniques like SOM that position objects in a readily visualized space. Their interest lay in the locations of ancient settlements in Cappadocia, some of which were known and some unknown. Measures of interaction between settlements were available, in the form of counts of the numbers of tablets found in one settlement that mentioned another settlement. On the basis that interaction would decline systematically with distance once appropriate normalizations had been applied, Tobler and Wineberg were able to use metric scaling methods to estimate the missing locations.

In the late 1960s and early 1970s social scientists were just beginning to appreciate the power of digital computers to support a vast range of new, more complex, and more powerful methods of analysis. Before that time, methods of analysis were essentially manual, with the aid of printed tables of standard statistical distributions, electric calculators, and mechanical sorting machines. Multivariate methods such as factor analysis had been invented many decades previously, but their methods were fundamentally compromised by the lack of powerful machines to do the necessary matrix inversions. Rigid assumptions, such as the normality of distributions, had to be imposed to make analysis tractable, whether or not they were supported by the data; and metrics such as variance were preferred over possibly more useful alternatives simply because they made the analysis feasible with the computational resources of the time.

Today, of course, we are blessed with an abundance of techniques that exploit cheap, powerful computing capabilities to do things with data that were scarcely conceivable before 1970. We also have geographic information systems that make it easy to acquire, store, analyze, and visualize georeferenced information. More fundamentally, perhaps, these new methods have shifted the paradigms of science significantly, towards larger data sets, and a greater emphasis on the exploratory methods and induction over confirmatory methods and deduction, as the chapters in this volume make clear. All of this is consistent with a widespread belief that the simple problems of science have been solved, and that further progress will require a new kind of science that emphasizes collaboration between disciplines, fueled by a search for elusive patterns in complex, multidimensional data sets.

A similar series of transitions are evident in the evolution of GIS. In the early days of the 1960s and 1970s, the focus of developers was on the data structures needed to represent the contents of maps in computers, and in the simplest kinds of analysis -- measurement of area, for example. Later, the functionality of GIS expanded to include a substantial fraction of the known methods of spatial analysis, such as metrics for the measurement of spatial autocorrelation, tests of the randomness of point patterns, methods of spatial interpolation, and methods of density estimation (Longley *et al.*, 2005). The ideas of exploratory spatial data analysis (ESDA) originated in the late 1980s and early 1990s, and were implemented in specialized packages such as Regard (Unwin, 1994; Anselin, 1999) and more recently GeoDa (Anselin, Syabri, and Kho, 2005; geoda.uiuc.edu). But even today the mainstream GIS products support only a small fraction of these ideas.

Essentially, ESDA seeks to create an intuitive, easy-to-use interface to geographic information that encourages exploration, and makes it possible for users to discover patterns and anomalies in data that would not otherwise be apparent. As such, the tests of its success seem to have much in common with other mainstream software environments, and little with traditional GIS, which is known for its complexity and the long training needed to extract useful results. Indeed, ESDA may resemble the kinds of GIS-derived products that are now available in the general marketplace, such as Google's Keyhole (http://www.keyhole.com/) whose user interfaces have more of the look and feel of a video game than a piece of software designed for serious scientific research, and on

which designers expect users to move from complete ignorance to mastery in a few minutes, rather than years.

With this background, it is possible to see a little further into the future of SOM and GIScience. Tools such as SOM, suitably adapted for regionalization, zone design, or classification, or simply for the exploration and visualization of structure within massive data sets, would be valuable additions to the GIS toolbox, and would help in a more general process of moving to simpler, more intuitive user interfaces. At the same time, they raise issues of fundamental significance that are logically part of the GIScience research agenda: what are the appropriate methods of representation of SOM inputs and results; what are the appropriate designs of user interfaces and visualization methods; how should one deal with time, dynamics, and uncertainty; what is the appropriate approach to effects of scale; and how can SOM results be made available for further analysis as part of the GIS database? A strong thread running through the chapters of this book suggests that SOM provides more than a single addition to the spatial analytic toolbox, but instead reflects an entirely new paradigm for ESDA and spatial data mining. If so, does this suggest that it should be implemented in a stand-alone toolbox rather than as part of GIS functionality, and how does the general trend in the GIS software industry to component ware impact this issue? In short, the advent of SOM, and the thinking that lies behind the chapters of this book, raise important questions for GIScience and add significantly to the GIScience research agenda.

REFERENCES

Anselin, L., 1999. Interactive techniques and exploratory spatial data analysis. In P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, editors, *Geographical Information Systems: Principles, Techniques, Management and Applications*. New York: Wiley, pp. 253-266.

Anselin, L., I. Syabri, and Y. Kho, 2005. GeoDa: an introduction to spatial data analysis, *Geographical Analysis* (forthcoming).

Longley, P.A., M.F. Goodchild, D.J. Maguire, and D.W. Rhind, 2005. *Geographic Information Systems and Science*. Second Edition. New York: Wiley.

Mark, D.M., 2003. Geographic information science: defining the field. In M. Duckham, M.F. Goodchild, and M.F. Worboys, editors, *Foundations of Geographic Information Science*. New York: Taylor and Francis, pp. 3-18.

Montello, D.R., S.I Fabrikant, M. Ruocco, and R.S. Middleton, 2003. Testing the first law of cognitive geography on point-display spatializations. In W. Kuhn, M. Worboys, and S. Timpf, editors, *Spatial Information Theory: Foundations of Geographic Information Science*, pp. 316–331. Lecture Notes in Computer Science 2825. Berlin: Springer.

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

Tobler, W.R. and S. Wineberg, 1971. A Cappadocian speculation. *Nature* 231(5297): 39-42.

Unwin, A, 1994. REGARDing geogrpahic data. In P. Dirschedl and R. Ostermann, editors, *Computational Statistics*. Heidelberg: Physica, pp. 345-354.