The University Consortium for Geographic Information Science

Research Priorities

Identification of Spatial Clusters

The Priority

To find ways to identify statistically significant clusters on maps.

Description of Research Challenge

There are a dozen or so techniques available that make it possible to identify pockets of unusual activity on maps. These pockets or "hot spots" may be areas of above average criminal activity, disease incidents, or any other phenomenon of interest that can be carefully mapped using some sort of location or coordinate system. Among these techniques are geographic data mining algorithms and kernel density estimation.

The crucial element that is missing in much of this work is the ability to recognize the clusters that did not occur by chance. Several attempts are being made to rectify this situation. Rogerson (1997), Kulldorff (1997), and Ord and Getis (1995, 2001) are among those attempting to identify statistically significant clusters on maps. In particular, Ord and Getis (2001) have challenged themselves to search for hot spots in large spatial data sets. This is a compelling problem for the geographic information sciences. Rapidly advancing technology emphasizes large data sets and the effects of various map scales on research results (Waller, 1994).

The considerable heterogeneity and autocorrelation present in most large georeferenced data sets present a compelling challenge for the spatial scientist attempting to find statistically significant clusters. The task includes finding suitable solutions to the identification of clusters when there is the need to use simultaneous, dependent testing procedures. The simultaneity problem is a result of the large number of similar tests required when a large data set is employed. The dependent test problem arises from the need to use some of the information from the original data set for more than one test.

Importance of Research Challenge

Nothing is more important to those who depend on maps than the ability to separate fact from fiction. As in most of the visual sciences, there is a tendency to fail to control for fundamental characteristics of the background information when attempting to find patterns that are significant. For example, a series of crimes depicted on a map may be confused for a cluster when a control variable, such

Author:
Arthur Getis
Department of Geography
San Diego State University
San Diego, CA 92182
E-mail: getis@mail.sdsu.edu

University Consortium for GIS
Suzy Jampoler, Director
Arthur Getis, President

UCGIS
43351 Spinks Ferry Road
Leesburg, Virginia 20176-5631
TEL: (888) 850-8533
FAX: (703) 771-1635
Internet: http://www.ucgis.org

The UCGIS is a non-profit organization of universities and other research institutions dedicated to advancing the understanding of geographic processes and spatial relationships through improved theory, methods, technology, and data.

as population density, is not taken into account. A disease may appear to be contagious or clustered when the susceptible population itself is clustered.

Comparisons between patterns can be effective only if suitable statistical properties of the patterns are taken into account. For example, is the spatial pattern of terrorist activity in 2002 different than, say, 2001? Homeland security issues dealing with georeferenced data do and will continue to require sophisticated spatial analytic procedures such as those outlined here. Important questions such as these can be answered by engaging in the analytical dissection of map patterns.

Eminent Research Questions

How is statistically valid clustering detected in large, heterogeneous, highly spatially autocorrelated data sets? What are valid procedures and useful tests for finding significance when one must perform a large number of tests simultaneously and where there is likely dependence between tests? What new and useful techniques can be devised to study clustering both in time and space? Can we create new and useful surveillance techniques for the identification of spatio-temporal evolution of clusters? What structural characteristics of maps must be taken into account for the study of clustering? How can spatial filtering (Getis and Griffith 2002) be used in a regression framework to account for the clustering characteristics of key regression variables?

References

Getis, A. and Griffith, D.A. (2002) Comparative Spatial Filtering in Regression Analysis. *Geographical Analysis* 34: 130-40.

Kulldorff, M. (1997) A Spatial Scan Statistic. *Communications in Statistics – Theory and Methods* 26: 1481-96.

Ord, J.K. and Getis, A. (1995) Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27: 286-306.

Ord, J.K. and Getis, A. (2001) Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation. *Journal of Regional Science* 41: 411-32.

Rogerson, P.A. (1997) Surveillance Systems for Monitoring the Development of Spatial Patterns. *Statistics in Medicine* 16: 2081-93.

Waller, L.A. and B.W. Turnbull, B.W. (1994) The Effect of Scale on Tests of Disease Clustering. *Statistics in Medicine* 12: 1969-84.