

# The University Consortium for Geographic Information Science

## Research Priorities



### GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY

#### THE PRIORITY

Development of new methods to facilitate the discovery and extraction of useful patterns and associations from large geospatial and temporal databases.

#### DESCRIPTION OF RESEARCH CHALLENGE

Explosive growth in geospatial data and the emergence of new spatial technologies emphasize the need for the automated discovery of geospatial knowledge. It is estimated that the EROS Data Center alone has archived about 200,000 gigabytes of remote sensing imagery and this amount is expected to grow to 2.4 million gigabytes by the year 2005. Many other forms of geospatial databases show similar trends, increasing both the volume and density of geospatial data that is available electronically (e.g. census data, road networks, thematic and topographic maps, digital elevation maps, soil, climate, ecology). With advances in storage technology, our ability to archive these data products has greatly improved as compared to our ability to analyze and extract useful information (patterns, clusters, classes, etc.) from these data products in an automated fashion. And since the datasets span the boundaries of

many traditional domains of inquiry, it is likely that they may contain much that is currently unknown or unstudied. Geographic data mining is emerging as an important research area within geographic and computing science domains with far reaching applications.

Geographic data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large geospatial databases. The complexity of spatial data and intrinsic spatial relationships limit's the usefulness of conventional data mining techniques for extracting spatial patterns. The differences between classical and spatial data mining are similar to the differences between classical and spatial statistics. First, spatial data is embedded in a continuous space, whereas classical datasets are often discrete. Second, spatial patterns are often local (in space and/ or time) whereas classical data mining techniques often focus on global patterns. When it comes to the analysis of spatial data, however, the assumption about the independence of samples is generally false because spatial data tends to be highly autocorrelated. Ignoring spatial autocorrelation when analyzing data with spatial characteristics may produce hypotheses or models that are inaccurate or inconsistent

#### Authors:

Shashi Shekhar  
Department of Computer Science and Engineering  
University of Minnesota  
Minneapolis, MN 55455  
E-mail: shekhar@cs.umn.edu

Ranga Raju Vatsaval  
Department of Forest Resources  
University of Minnesota

Mark Gahegan  
Department of Geography  
Pennsylvania State University

Harvey J. Miller  
Department of Geography  
University of Utah

Barbara Battenfield  
Department of Geography  
University of Colorado

May Yuan  
Department of Geography  
University of Oklahoma

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.

with the dataset. Thus new methods are needed to analyze spatial data to detect spatial patterns. The roots of spatial data mining lie in spatial statistics, spatial analysis, artificial intelligence and machine learning, pattern recognition and image analysis, and high-performance computing. A summary of various techniques can be found in Battenfield et al. (2000).

Several new techniques and applications have been reported in recent years. Examples are location prediction, spatial outlier detection, co-location pattern discovery, and constraint-based clustering. Location prediction is concerned with the discovery of a model to infer locations of a spatial phenomenon from the maps of other spatial features. Spatial outliers are significantly different from their neighborhood even though they may not be significantly different from the entire population. The co-location pattern discovery process finds frequently co-located subsets of spatial event types given a map of their locations. Constraint-based clustering incorporates user-defined constraints into the clustering process. For example, consider a situation where two schools are located on opposite banks of a river. Simple Euclidean-based clustering methods place these two schools in one cluster; however, in reality they may be quite far off in terms of travel distance. The constraint-based clustering is very useful for grouping geographic objects in the presence of obstacles. The spatial extensions are computationally intensive and there is a great need for developing scalable parallel algorithms and efficient data structures for these techniques.

## IMPORTANCE OF RESEARCH CHALLENGE

Spatial data mining plays an important role in diverse scientific domains dealing with large volumes of geospatial data. Some of the well-known applications are: NASA (studying climatological effects of El Nino, land use classification and global change using satellite imagery), NIH (predicting the spread of a disease), NIJ (finding crime hot spots), transportation agencies (detecting instability in traffic), the Army (predicting global hot spots, inferring enemy tactics from blobology, locating lost ammunition dumps), and M(mobile)-commerce (location-based services).

## EMINENT RESEARCH QUESTIONS

Which types of data mining algorithms are appropriate for geospatial data? What are the limitations and how can they be extended to incorporated spatial concepts? What kind of spatial access structures are needed for efficient mining? How can SDM algorithms be scaled to process large geospatial databases? How can domain knowledge be incorporated to improve query processing and mining? What kind of computational infrastructure is needed (e.g. Distributed, GRID)? How can we deal with uncertainty, missing information, inconsistency, and heterogeneity (typical in large geospatial databases)? What kinds of feature selection techniques are appropriate or required? How can we efficiently detect spatial outliers? How can we represent the knowledge (pattern) that we discover? Where will it go after it is 'discovered'? (i.e. how can it be saved or represented by the system). How can we validate these patterns?

## REFERENCES

- Battenfield, B., Gahegan, M., Miller, H.J., and Yuan, M. (2000) *Geospatial Data Mining and Knowledge Discovery*. Washington, D.C., University Consortium for Geographic Information Science Research White Paper (available at <http://www.ucgis.org/emerging/gkd.pdf>)
- Miller, H.J. and Han, J. (eds) (2001) *Geographic Data Mining and Knowledge Discovery*. London, Taylor and Francis
- Tung, A.K.H., Ng, R.T., Lakshmanan, L.V.S., and Han, J. (2001) Constraint-based Clustering of Large Databases. In *Proceedings of the Eighth International Conference on Database Theory (ICDT'01)*, London: 405-19
- Shekhar, S., Huang, Y., Wu, W., Lu, C.T., and Chawla, S. (2001) *What's Spatial about Spatial Data Mining: Three Case Studies*. In Grossman, R.L., Kamath, C., Kegelmeyer, P., Kumar, V., and Namburu, R.R. (eds) *Data Mining for Scientific and Engineering Applications*. New York, Kluwer
- Shekhar, S., Schrater, P.R., Vatsavai, R.R., Wu, W., and Chawla, S. (2002) Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transactions on Multimedia* 4: 174-88