

# **DISCOVERING GEOGRAPHIC KNOWLEDGE IN DATA RICH ENVIRONMENTS**

Report of a Specialist Meeting held under the auspices of the Varenus Project

Kirkland, Washington  
March 18-20, 1999

by

Harvey J. Miller  
Department of Geography  
University of Utah  
260 S. Central Campus Dr. Rm 270  
Salt Lake City, Utah 84112-9155  
harvey.miller@geog.utah.edu

Jiawei Han  
School of Computing Science  
Simon Fraser University  
Burnaby BC  
Canada V5A 1S6  
han@cs.sfu.ca

October 8, 1999

## EXECUTIVE SUMMARY

Similar to many scientific and applied research fields, geography has moved from a data-poor and computation-poor to a data-rich and computation-rich environment. The scope, coverage and volume of digital geographic datasets are growing rapidly due to new high-resolution satellite systems, initiatives such as the U.S. National Spatial Data Infrastructure and the automated collection of point-of-sale, logistic and behavioral data. In addition, the types of geographic data collected are expanding from the traditional vector and raster data models to include georeferenced multimedia data. This trend is likely to continue if not accelerate in the foreseeable future.

Traditional spatial statistical and spatial analytical methods were developed in an era when data collection was expensive and computational power was weak. The increasing volume and diverse nature of digital geographic data easily overwhelm mainstream spatial analysis techniques that are oriented towards teasing scarce information from small and homogenous datasets. Traditional statistical methods, particularly spatial statistics, have high computational burdens. These techniques are confirmatory and require the researcher to have *a priori* hypotheses. Therefore, traditional spatial analytical techniques cannot easily discover new and unexpected patterns, trends and relationships that can be hidden deep within very large and diverse geographic datasets.

The National Center for Geographic Information and Analysis (NCGIA) – Project Varenius workshop on “Discovering geographic knowledge in data-rich environments” brought together a diverse group of stakeholders with interests in developing and applying new techniques for exploring large and diverse geographic datasets. This included geographers, geographic information scientists, computer scientists and statisticians. The synergy created by the discussions prior to, during and after the three-day workshop resulted in the identification of research priorities and directions for continued development of “geographic knowledge discovery” (GKD) theory and techniques.

This research report summarizes the activities surrounding the Project Varenius workshop on “Geographic knowledge discovery in data-rich environments.” Section 1 comprises an overview of the workshop. Section 2 lists the workshop participants. Section 3 provides the position papers submitted in response to the open Call for Participation. Section 4 summarizes the workshop presentations and discussion.

Many detailed recommendations for continued research and development in GKD theory and techniques emerged during the three-day workshop. Participants also identified several cross-cutting research issues. These issues are:

- What are the new questions for geographic research? A fundamental question for the geographic and related research communities is “What questions do we want to answer that we could not answer previously?” These communities need to form well-structured (but possibly open-ended questions) in order to guide the computer science and related communities in their tool and algorithm development
- Better spatio-temporal representations in GKD. Current GKD techniques use very simple representations of geographic objects and geographic relationships, for example, point objects and Euclidean distances. Other geographic objects (including lines, polygons and more

complex objects) and geographic relationships (including non-Euclidean distances, direction, connectivity, attributed geographic space such as terrain and constrained interaction structures such as networks) should be recognized by GKD techniques. Time needs to be more completely integrated into geographic representations and relationships. This includes a full range of conceptual, logical and physical models of spatio-temporal objects. Finally, we need to formulate multiple representations (in particular, robust geographic concept hierarchies) and granularities in spatio-temporal representation in order to manage the complexity of GKD.

- GKD using richer geographic data types. Geographic datasets are rapidly moving beyond the well-structured vector and raster formats to include semi-structured and unstructured data, in particular, georeferenced multimedia. GKD techniques should be developed that can handle these heterogeneous datasets.
- User interfaces for GKD. GKD needs to move beyond the technically-oriented researcher to the broader geographic and related research communities. This requires interfaces and tools that can aid these diverse researchers in the GKD process. These interfaces and tools should be based on useful metaphors that can guide the search for geographic knowledge and make sense of discovered geographic knowledge.
- Proof of concepts and benchmarking problems for GKD. There is a strong need for some “examples” or “test cases” to illustrate the usefulness of GKD. This includes a demonstration of GKD techniques leading to new, unexpected knowledge in key geographic research domains. Also important is benchmarking to determine the effects of varying data quality on discovering geographic knowledge. A related issue is research and demonstration projects that illustrate the usefulness of GKD techniques in forecasting and decision support for the public and private sectors.
- Building discovered geographic knowledge into GIS. Current GIS software uses simple representations of geographic knowledge. Discovered geographic knowledge should be integrated into GIS, possibly through inductive geographic databases or online analytical processing (OLAP)-based GIS interfaces.
- Developing and supporting geographic data warehouses. A glaring omission from current research in GKD techniques is the development and supporting infrastructure for *geographic data warehouses* (GDW). To date, a true GDW does not exist. This is alarming since data warehouses are central to the knowledge discovery process. Creating true GDWs requires solving issues in geographic and temporal data compatibility, including differences in semantics, referencing systems, geometry, accuracy and precision. Supporting GDWs may also require restructuring of transaction-oriented databases systems, particularly for flow and interaction data.

## **ACKNOWLEDGEMENTS**

The co-leaders would like to thank the following individuals whose efforts were invaluable in making the “Discovering geographic knowledge in data-rich environments” workshop a success and pleasure:

- The National Center for Geographic Information and Analysis-Project Varenius (Michael Goodchild, Director) for funding the workshop.
- Max Egenhofer (University of Maine) and LaNell Lucius (University of California – Santa Barbara) for their outstanding organization efforts on behalf of the National Center for Geographic Information and Analysis-Project Varenius.
- Steve Smyth (Microsoft), Charles Roche (Mobile GIS LTD) and Raina Smyth for excellent local arrangements (and a special thanks to Steve for use of his beautiful summer home on Lake Washington).
- Phoebe McNeally (University of Utah), Matt Rice (University of California – Santa Barbara), Anthony Tung (Simon Fraser University) and Yi-Hwa Wu (University of Utah) for their meticulous and detailed note taking.
- The members of the Steering Committee for their support in organizing the workshop.
- All of the participants for their stimulating and friendly interaction before, during and after the workshop.

## TABLE OF CONTENTS

EXECUTIVE SUMMARY .....	1
ACKNOWLEDGEMENTS.....	3
TABLE OF CONTENTS .....	4
1 WORKSHOP BACKGROUND.....	5
2 WORKSHOP PARTICIPANTS.....	7
3 RESEARCH STATEMENTS .....	10
3.1 Yvan Bedard.....	10
3.2 David Bennett (and Raja Sengupta) .....	10
3.3 Catherine Dibble.....	12
3.4 Doug Flewelling.....	14
3.5 Mark Gahegan.....	17
3.6 Myke Gluck.....	19
3.7 Kathleen Hornsby.....	21
3.8 Eric D. Kolaczyk .....	25
3.9 Brian Lees .....	27
3.10 Donato Malerba.....	30
3.11 Duane Marble.....	32
3.12 Raymond Ng .....	34
3.13 Jonathan Raper .....	36
3.14 Hanan Samet .....	37
3.15 Shashi Shekhar .....	39
3.16 Monica Wachowicz .....	41
4 SUMMARY OF WORKSHOP PRESENTATIONS AND DISCUSSION.....	43
4.1 Thursday, March 18.....	43
4.1.1 Plenary I - Status and trends in geographic information science.....	43
4.1.2 Plenary II - Methods for geo-spatial data mining .....	43
4.1.3 Panel: Tasks for spatial data mining in large geo-spatial databases .....	45
4.1.4 Breakout groups - Setting the GKD agenda .....	46
4.2 Friday, March 19 .....	48
4.2.1 Plenary III - Geocomputational tools and GKD .....	48
4.2.2 Panel: Research frontiers in geocomputation and GKD .....	49
4.2.3 Panel - GKD and domain-specific research.....	51
4.2.4 Breakout groups - New tools for a geospatial data-rich environment .....	53
4.3 Saturday, March 20.....	54
4.3.1 Plenary IV - Geospatial data and data warehousing .....	54
4.3.2 Panel - Application problems and requirements.....	56
4.3.3 Breakout groups - GKD research and application frontiers .....	57
4.3.4 Synthesis.....	58
5 SEED GRANT PROPOSALS.....	60
5.1 Ontologies for Spatial Data Mining and Geographic Knowledge Discovery in Large Multimedia Spatial Datasets.....	60
5.2 Multi-Scale Tools for Modeling Flows in Geographic Databases.....	61

## 1 WORKSHOP BACKGROUND

Digital geographic datasets are growing rapidly due to activities as the development of the U.S. National Spatial Data Infrastructure, the launching of new satellite systems with higher resolutions, and the day-to-day collection of digital imagery, video, and sound. Society has changed from being data-poor to data-rich, while our techniques for deriving knowledge from the data have remained inferential. The problem has now become not finding the data, but filtering through large volumes of data to finding meaningful geographic knowledge. At the same time, the types of datasets available are changing from the traditional vector and raster formats to include such data types as video and audio, and the location of where these data were collected. We must overcome these limitations and develop new approaches and methods that focus upon separating the relevant from the irrelevant, the meaningful from the background noise.

The goal of the "Discovering geographic knowledge in data-rich environments" workshop and research initiative is to find new automated methods for filtering large amounts of raw geographic data into more user-consumable forms of knowledge. This includes: i) spatial data mining; ii) content-based and knowledge-based retrieval; iii) development of multi-media spatial data types; iv) on-line analytic processing; v) refinement of non-parametric statistics; vi) incorporation of computational intelligence techniques (such as neural networks and AI expert systems) into spatial data analysis.

A key objective of the "geographic knowledge discovery" (GKD) workshop was to bring together the diverse stakeholders in GKD and spatial data mining. In the academic realm, these include:

- i) geographers and other researchers interested in domain-oriented research questions;
- ii) geographic information scientists formulating new digital geographic representations and geocomputational techniques;
- iii) computer scientists formulating computational techniques for knowledge discovery from non-geographic and (increasingly) geographic databases; and,
- iv) other scientists developing analytical techniques for data analysis (e.g., statisticians).

Also critical was bringing together academic researchers and scientists working in industry and other private sector settings.

The Co-leaders and Steering Committee for the GKD workshop that reflected these diverse stakeholders:

### Initiative Co-Leaders

**Jiawei Han**, Simon Fraser University (computer science)

**John R. Herring**, Oracle Corporation

**Harvey J. Miller**, University of Utah (geography)

### Steering Committee

**Larry Band**, University of North Carolina, Chapel Hill (geography)

**Max J. Egenhofer**, University of Maine (geographic information science)

**Suchi Gopal**, Boston University (geography)

**Hans-Peter Kriegel**, University of Munich (computer science)

**Richard R. Muntz**, University of California, Los Angeles (computer science)

**John Roddick**, University of South Australia (computer science)

**Carl Stephen Smyth**, Microsoft Corporation

**Elizabeth A. Wentz**, Arizona State University (geography)

**Aidong Zhang**, State University of New York at Buffalo (computer science)

As can be seen from the participant list (pp. 7 - 9) and the position papers (pp. 10-42) the workshop achieved the desired goal of bringing together a diverse group of GKD stakeholders.

The next section of this report provides the participant list and contact information. Section 3 provides the research statements submitted during the open call for participation. Section 4 provides a summary of the workshop presentations and discussion, including an overall synthesis of research themes. Section 5 summarizes the funded seed grant proposals that resulted from the workshop.

## 2 WORKSHOP PARTICIPANTS

<u>Name</u>	<u>Affiliation</u>	<u>Email</u>
Yvan Bedard	Department of Geomatics Sciences Center for Research in Geomatics Laval University Quebec City, Canada G1K7P4	Yvan.Bedard@scq.ulaval.ca
David Bennett	Department of Geography University of Kansas Lawrence, KS 66045	dbennett@falcon.cc.ukans.edu
Sean Curry	ESRI 50 El Camino Real Berkeley, CA 94705	scurry@esri.com
Catherine Dibble	Department of Geography University of California 3611 Ellison Hall Santa Barbara, CA 93106-4060	cath@econ.ucsb.edu
Max Egenhofer	NCGIA 5711 Boardman Hall University of Maine Orono, ME 04469-5711	max@spatial.maine.edu
Douglas Flewelling	NCGIA 5711 Boardman Hall University of Maine Orono, ME 04469-5711	dougf@spatial.maine.edu
Mark Gahegan	Department of Geography Pennsylvania State University University Park, PA 16802	mark@geog.psu.edu
Myke Gluck	School of Information Studies Florida State University 244 Shores Building Tallahassee, FL 32306-2100	mgluck@lis.fsu.edu
Jiawei Han	Department of Computing Science Simon Fraser University 8888 University Drive Burnaby, B.C. V5A1S6 Canada	han@cs.sfu.ca
Kathleen Hornsby	Department of Spatial Information Science & Engineering 5711 Boardman Hall University of Maine Orono, ME 04469-5711	khornsby@spatial.maine.edu

Eric Kolaczyk	Department of Mathematics & Statistics Boston University 111 Cummington Street Boston, MA 02215	kolaczyk@math.bu.edu
Hans-Peter Kriegel	University of Munich Institute for Computer Science Oettingen Str. 67 D-80538 Muenchen Germany	kriegel@dbs.informatik.uni-muenchen.de
Brian Lees	Department of Geography Australian National University ACT 0200 Australia	Brian.Lees@anu.edu.au
Tim McGrath	Geography Product Unit Microsoft Corporation One Microsoft Way Redmond, WA 98052-6399	tmcgrath@microsoft.com
Heikki Mannila	Microsoft Research One Microsoft Way Redmond, WA 98052	mannila@microsoft.com
Donato Malerba	Dipartimento di Informatica University of Bari Via Orabone 4 70126 Bari Italy	malerba@di.uniba.it
Duane Marble	Center for Mapping The Ohio State University 1216 Kinnear Columbus, OH 43212	marble.1@osu.edu
Phoebe B. McNeally	Department of Geography University of Utah 260 S. Central Campus Dr. Room 270 Salt Lake City, UT 84112-9155	phoebe.mcneally@geog.utah.edu
Harvey J. Miller	Department of Geography University of Utah 260 S. Central Campus Drive Room 270 Salt Lake City, UT 84112-9155	harvey.miller@geog.utah.edu
Richard Muntz	Computer Science Department 4732 Boelter Hall University of California Los Angeles, CA 90095	muntz@cs.ucla.edu
Raymond Ng	Department of Computer Science 2366 Main Mall University of British Columbia Vancouver, B.C. V6T 1Z4 Canada	rng@cs.ubc.ca

Matt Rice	Department of Geography University of California 3611 Ellison Hall Santa Barbara, CA 93106-4060	rice@geog.ucsb.edu
Charles Roche	Mobile GIS LTD Cruachan Sardfields Court Glanmire Co Cork Ireland	charles@mobilegisltd.com
John Roddick	School of Computer and Information Science University of South Australia Mawson Lares 5095 South Australia	roddick@cis.unisa.edu.au
Hanan Samet	Computer Science Department University of Maryland College Park, MD 20742	hjs@cs.umd.edu
Shashi Shekhar	Department of Computer Science University of Minnesota 200 Union Street, SE #4192 Minneapolis, MN 55455	shekhar@cs.umn.edu
Carl Stephen Smyth	Mobile Electronics Group Microsoft Corporation One Microsoft Way Redmond, WA 98052-6399	stevesmy@microsoft.com
Anthony Tung	School of Computing Science Simon Fraser University 8888 University Drive Burnaby, B.C. Canada V5A1S6	khtung@cs.sfu.ca
Monica Wachowicz	WVR Geoinformation Center- DLO PO Box 125 6700 AC Wageningen The Netherlands	wachowicz@sc.dlo.nl
Elizabeth Wentz	Department of Geography Arizona State University Box 870104 Tempe, AZ 85287-0104	wentz@asu.edu
Yi-Hwa Wu	Department of Geography University of Utah 260 S. Central Campus Drive Room 270 Salt Lake City, UT 84112-9155	wu.yi-hwa@geog.utah.edu

### **3 RESEARCH STATEMENTS**

#### **3.1 Yvan Bedard**

Over the last 15 years, we have been deeply rooted in R&D about spatial database modeling, management and implementation at federal and provincial levels in Canada both on academic and professional grounds. We have also been involved in R&D on spatial data warehousing and Spatial OLAP for the last two years, with practical experiences. In particular, we realized a Strategic positioning technical report for the Canadian National Defense on this topic. Dr. Bedard is the scientific leader of two major projects on this topic (and which involve Dr. Jiawei Han) in the context of GEOID, the new Canadian Network of Centres of Excellence in Geomatics.

My Ph.D. student Pierre Marchand and myself are very interested in participating to this meeting because there is a high level of compatibility between our research focus and the issues that will be discussed. In particular, besides the scientific discussions which will take place, it will provide the best opportunity to establish the ground for cooperation and coordination between the NCGIA and GEOID groups on this topic, probably the two largest formal university projects on the topic. Finally, we will be very interested to share with the audience the results of our research, especially the strategic R&D issues identified for the National Defense.

Our approach to knowledge discovery in spatial data is not yet based on automated techniques. We strongly believe that prior to a fully automated spatial data analysis there lies a powerful human space that has not been exploited yet. This space could greatly help spatial data mining in supporting spatial iterative reasoning. We have carried out various research projects that could greatly contribute to help knowledge discovery in data rich environments. Whatever the approach, appropriate integration and availability of spatio-temporal data is the sine qua none condition of effective spatial processing. Moreover, whereas most of the techniques involved in spatial data mining imply an IT approach, we believe that there are other ways to generate knowledge discovery in spatial data.

Data mining techniques have been used in the IT mainstream for quite some time, spatial data mining is a recent phenomenon that is facing multiple barriers. "The reality is that the bulk of our scientific knowledge lacks the spatial specificity in the relationships among variables demanded by these advanced applications." [Berry]. "Managers would rather live with a problem they can't solve than apply a solution they don't understand" [Woosley]

#### **3.2 David Bennett (and Raja Sengupta)**

The goal of our research is to develop interoperable technologies that integrate digital geographical data and analytical tools that have been gleaned from network accessible repositories into systems capable of supporting spatial analysis and problem solving (i.e., on-line analytical processing). Our focus is not necessarily on how to perform data mining operations designed to search through mountains of data in search of a few high quality gems. Rather, we are more interested in how to manage these data once they are returned to the user.

Initial efforts will focus on the development of a system capable of supporting environmental modeling. To address social and environmental problems it is often necessary to understand the impact of human activities on a complex set of geographical processes that interact across space and through time. Spatially explicit computer models can provide insight into spatio-temporal relations and, thus, contribute valuable information to the decision making process. The creation

and management of spatially explicit models, however, is often difficult and time consuming because the real world systems that they represent are large and complex. Furthermore, the use of such models often requires users to construct links among disparate software products, data formats, and computing platforms. Performing these integrative tasks often requires geoprocessing and computer skills that are beyond many potential users. Connections forged among disparate software also require common communication protocols. Such protocols are in development for geographical databases (e.g., SDTS (National Institute of Standards and Technology, 1992), Open-GIS (<http://www.opengis.org>)). However, there is little support for the distribution and sharing of geographical models and analytical tools. Nor is there strong support for the transformation and integration processes that must be performed to make data applicable to a particular project.

Our challenge, then, is to represent, capture, and use the knowledge of geoprocessing experts in digital form in a manner that is flexible, robust, and adaptive (i.e., brittleness must be avoided) and to use these digital experts to expedite the development of spatial models. Conceptually, such a system would be of utility to power users running computationally intensive spatial models as well as the computing novice who simply wants to understand how he will be effected by a proposed transportation project. We have identified a set of four technologies that require further research before such Internet-based GIS technologies reach their full potential. These follow those set for in the "Call for Participation" and are as follows:

- Geoprocessing technologies are needed to bridge multiple vendor formats and heterogeneous computing environments.
- Artificial intelligence technologies are needed to facilitate the acquisition of geographical data and the transformation of these data into a usable form.
- Distributed processing capabilities are needed to harness the collective computational power that resides on networks.
- Advances in the theory and application of public participation GIS are needed to better understand how and to what extent geographical concepts and problem solving techniques can be used by and communicated to the public at large.

If such advances can be made in the theory and application of geoprocessing technologies then decision-makers, analysts, and stakeholders will have access to computing resources that may not otherwise be available and a more level "technological playing field" on which to build consensus and compromise.

We have begun the development of a system capable of supporting some of the tasked outlined above. This system is being built using: distributed problem solving strategies, intelligent agents, existing GIS software, and applicable web-based tools. Collectively, these technologies must model the activities of the geoprocessing expert as he or she goes through the complex task of integrating data and analytical tools collected from various sources across the Internet and to incorporate this behavior into a computer program. The objectives of this first phase of research are to:

- Develop a geographical data model that supports geographical analysis across a heterogeneous computing environment.
- Develop knowledge structures that capture in digital form the human expertise needed to perform specific geoprocessing tasks.
- Develop intelligent agents that use elements derived from objectives 1 and 2 to automate the integration of data and models.

- Evaluate the effectiveness of this work by comparing computer generated solutions to those generated manually by geoprocessing experts.

Future research will expand on these efforts and focus on:

- Developing a geographical model definition language that supports the query, manipulation, and integration of geographical data and analytical tools stored in network accessible repositories.
- Developing distributed computing capabilities that manage environmental models and data on a virtual machine that is distributed across heterogeneous computing environments.
- Testing the ability of novice users to construct spatial analytical tools using this system.
- Testing the ability of novice users to properly interpret the results of spatial analytical tools.

The intended beneficiaries of this research include stakeholders (e.g., individuals impacted by resource management decisions), resource managers, and scientists who do not possess the technical expertise or the time needed to search for and integrate the geographical data and analytical tools needed to make informed decisions or to study spatial processes.

### **3.3 Catherine Dibble**

Grace, resilience, adaptation, and evolution are complementary themes. I have been working with Genetic Algorithms and Holland classifier Genetics Based Machine Learning (GBML) since coming to geography. First, by using an unusual alphabet to discover efficient yet diverse solutions to the P-Median location-allocation facility location problem. Second, by developing a spatial representation for GBML classifiers that supports the implementation and refinement of adaptive expert system rule bases. Spatial classifiers can serve as effective relevance filters or adaptive knowledge bases for large spatial data files such as those generated by remote sensing and environmental monitoring systems, also as adaptive assistants for automated spatial data selection and representation. Most recently, I am a co-founder of UCSB's Computational Laboratories Group and have been developing network and spatial agent-based simulations as computational laboratories for scientists and theorists to model and explore the behavior of complex spatial systems.

Whether the massively detailed data sets come from real-world data or laboratory simulations, in each case the ultimate knowledge-engineering challenge is to design computational tools that provide the greatest possible complementary leverage for human insight. I was the graduate organizer for Barbara Buttenfield's NCGIA Specialist Meeting on Formalizing Cartographic Knowledge, and would very much like to contribute to innovative designs and research strategies for Discovering Geographic Knowledge in Data-Rich Environments.

I've collected the abstracts from my four papers on Discovering Geographic Knowledge in Data Rich Environments. Current research complements these by exploring graph theoretic and pattern theory formalizations for spatial structure in adaptive systems.

*Beyond Data: Handling Spatial and Analytical Contexts with Genetics Based Machine Learning.*

This paper introduces a general class of Genetics Based Machine Learning (GBML) systems for adding spatial context and analytical purpose to the data in Geographic Information Systems (GIS). These flexible, adaptive expert systems inductively learn and apply relevant relationships and rules based on spatial, temporal, and attribute characteristics of geographic data. In particular,

this paper introduces a graceful and efficient new representation for evaluating both absolute and relative spatio-temporal relations among objects or cells in geographic databases. Geographic applications of GBML methodologies have been hindered by the difficulty of comparing spatial and temporal positions in a way that allows efficient evaluation of their relationships across a richly populated spatial database; explicit calculations of distances or buffer zones would be prohibitively expensive considering the many comparisons that are required. The strategy presented here allows GBML systems to efficiently develop and refine relevant rules based on space-time-attribute information. Applications could filter salient data, construct structured yet realistically complex models, or coordinate other GIS processing.

GBML systems do not require tedious and prohibitive knowledge elicitation and program (rule) maintenance in order to be useful (unlike traditional expert systems), nor do they require strong assumptions regarding data precision or functional forms (unlike statistics). Yet the rules and relationships generated by the system are straightforward to edit or to query if desired (unlike neural nets). In addition, the familiar space-time-attribute representations presented here are easy to define for existing databases and thus facilitate the practical development of GBML systems as complements to existing GIS.

#### *Relevance Filters for Spatial Information*

Dealing effectively with overwhelming masses of spatial data in the twenty-first century will require more than improved technologies for faster computers or parallel algorithms. Effectiveness incorporates but should not be confused with efficiency; the more fundamental challenge remains ultimately the selection of relevant data for further attention or processing by humans or computers. This paper defines a relevance filter as a mechanism interposed between an extensive source of input data and a human or computer agent that seeks to make use of some subset of the data for a particular analytical purpose. Consideration of the central principles for the design of such relevance filters raises a number of theoretical and practical research questions related to spatial information, and unifies several surprisingly diverse lines of research.

#### *The Cartographer's Apprentice*

Difficulties in formalizing cartographic knowledge impede the development of automated cartographic systems. The intuitive and complex operations are difficult to capture and describe, and rule-based systems tend to be brittle in the face of changing map purposes and databases. This paper explores the potential contribution of Holland Classifiers and Genetic Algorithms to the development of adaptive software to support and enhance the design and generation of cartographic products from geographic information systems. Along the way, we consider the roles that such adaptive systems may play in helping to discover and formalize cartographic knowledge and in complementing intelligent data base systems that assist and support even expert cartographers.

#### *Generating Interesting Alternatives in GIS and SDSS Using Genetic Algorithms (Catherine Dibble and Paul J. Densham)*

Decisions often are evaluated on the quality of the process that supported them. It is in this context that GIS and SDSS increasingly are being used to generate alternatives to aid decision-makers in their deliberations. Unfortunately, GIS and SDSS typically lack formal mechanisms to help decision-makers explore the solution space of their problem and thereby challenge their assumptions about the number and range of options available.

We describe the use of a genetic algorithm to generate a range of feasible alternatives to location selection problems. The ability of genetic algorithms to search a solution space and selectively focus on promising combinations of criteria makes them ideally suited to such complex spatial decision problems. We also describe the implementation of this algorithm in a microcomputer-based SDSS and present representative results for several location selection problems. Finally, we discuss the inherent parallelism of this algorithm and strategies for its decomposition that will enable it to exploit the efficiency gains of parallel processing computers.

### **3.4 Doug Flewelling**

#### *Evaluating Spatial Character in Subsets of Very Large Datasets*

Very large spatial datasets are becoming commonly available through digital libraries (Smith 1996), data warehouses (Garcia-Molina *et al.* 1995), and research archives (Levy and Marshall 1996). Plans to increase the number and size of these *digital spatial archives* through such efforts as the National Spatial Data Infrastructure (FGDC 1994) are being implemented at the Federal and local level. The availability of data and their suitability for a desired purpose do not necessarily go hand-in-hand. The current state of digital archives is a “buyer beware” market with few controls on what goes into a digital archive or assurances of fitness for use. Efforts such as the Content Standard for Digital Geospatial Metadata (FGDC 1997) and Dublin Core (Weibel *et al.* 1997) are attempting to standardize dataset labeling, but their use is voluntary in many cases. The current euphoria at having and providing access often overlooks the risks and costs associated with using the data. Without knowledge about the suitability or fitness of a dataset for a particular task the user cannot draw meaningful or defensible conclusions. This article addresses the need for a means to evaluate the suitability of very large spatial datasets for a given task.

There are two key elements necessary for scientists to determine that a dataset is the best one for their needs. First they must have some knowledge of the dataset’s contents, through experience with the dataset, through a detailed description or a reliable summary. In the case of detailed description, metadata—data about data—can provide valuable statistics and data history, which can show that the dataset should have the proper contents to address the problem domain. It is possible, however, that the data collected are incomplete or unevenly distributed through space, or may have any number of hidden problems. For instance, even though the metadata show that each station in a weather dataset for 1988 has rainfall as a data field and 86% of the stations reported rainfall, the dataset is of limited use for a scientist studying the 1988 drought if the missing 14% are in the American Midwest. Second, the domain scientist must consider the specific relationships among data elements in a spatial dataset. The relevance of a specific data item in a scientific inquiry is an important factor since there are complex inter-relationships present in spatial data. These specific relationships are, in fact, of greatest concern to spatial scientists and must be preserved in any dataset used.

Scientists who have had experience with a particular dataset or its provider may have developed an understanding of the peculiarities involved. They develop a level of trust—high or low—which effects their future choices of datasets. Inexperienced users of digital archives, however, do not have this advantage in picking their datasets and may have to download several seemingly useful datasets before they find one that meets their needs. In an environment with a charge for network time and data access, finding the right dataset can become expensive. In a traditional market, imagine having to pay a fee to take a shirt off the shelf and try it on, or even having to buy it based on a description like “green plaid, long sleeve, large.” In order to make digital spatial

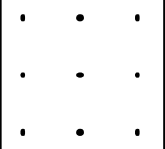
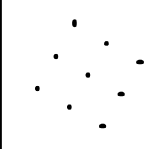
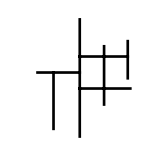



archives useful, users need a mechanism to try out datasets before investing time and money in exploiting the entire dataset. The scientist must be able to assess the fitness of a particular dataset for her tasks and to do so efficiently. The task of *finding* the right data is different from *using* the data in scientific analyses and requires different supporting tools (Flewelling and Egenhofer 1993). Tools to compare and evaluate subsets of large spatial data sets are needed to support effective use of spatial data archives. Assessment of the spatial character of datasets needs to be based on the synoptic or aggregate-level of the data, not the attributes of the individuals in the dataset.

### *Synoptic Attributes of Spatial Datasets*

With regard to this paper, the purpose of creating a subset of a spatial dataset is to identify salient spatial properties of the dataset and to compare those properties to other spatial phenomena. To perform this task meaningfully it is necessary for subsets of spatial datasets to retain those spatial characteristics of the large datasets that are important to making a decision about the larger dataset. For example, a scientist who is looking for a dataset that would help her evaluate temperature change in North America for the last 100 years would need a dataset that is evenly dispersed over the continent. A sample of the dataset should preserve that dispersion, perhaps at the expense of picking the weather stations at large cities or even fully preserving spatial clustering.

There are several spatial properties that might be used to describe individual spatial objects and groups of objects. For our purposes, we need to preserve the properties of sets of spatial objects, such as all populated places in the United States or the streams in a drainage network. Sets of spatial objects have a limited number of spatial properties that are necessary to preserve among which are: density, dispersion, and pattern (Unwin 1981).

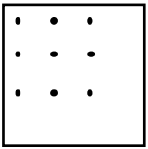
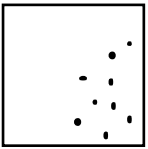
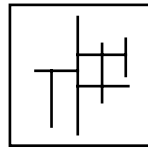
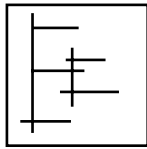


In geography (Goodchild 1992) and geomorphology (Bloom 1978), pattern has traditionally been recorded with descriptive terms such as regular, dendritic, or trellis (Figure 1). While such terms are useful for classification they do not lend themselves to quantitative measures of similarity. Descriptive classifications do not say how dendritic a particular set of lines is or how similar one set is to another. Parametric techniques, such as Horton's (1945) stream ordering, describe elements of pattern depending on the type of spatial objects in the set. However, none of these methods arrive at a one-to-one correspondence between a value and an unique pattern. At best it is possible to partially order the pattern of spatial datasets according to a reference pattern.

Sets of Spatial Objects with Similar Pattern					
Points		Lines		Areas	
A	B	A	B	A	B
					

**Figure 1: Similar pattern in sets of spatial objects of homogeneous type.**

Measuring density in spatial datasets is much less problematic than pattern. Methods for describing density of sets of spatial objects take different forms depending on the type of spatial object (Figure 2). For points, a suitable surrogate for density is simply a count of the number of points within a reference area. Total length of all lines in a reference area can be a measure of

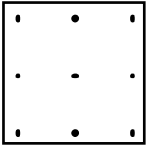
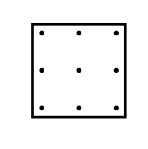
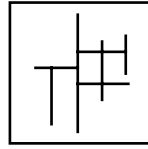
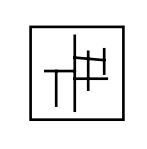


density for line sets. In the case of a set of area objects a more traditional percent of area filled can be used.

Sets of Spatial Objects with Similar Density					
Points		Lines		Areas	
A	B	A	B	A	B
					

**Figure 2: Similar density in sets of spatial objects with respect to an enclosure.**

Dispersion refers to the arrangement of the elements of a spatial dataset within a containing area (Figure 3). For instance, an analysis of population in Maine shows that the largest cities in Maine are in the south and population is clustered around four population centers in Maine. Dispersion is similar to density except that dispersion implies an orientation relative to the containing area. Density measures are invariant under rotation relative to the container, but dispersion measures are not, unless the container is a circle.

While pattern, density, and dispersion are exclusively spatial properties, importance is a fourth property that can play a key role in structuring sets of spatial objects. Importance is an ordering function over the set of objects according to the key concept or concepts being studied (Kadmon 1972). In this manner, similarity is an user-defined concept applicable to the application at hand. For instance, in a dataset on crime in America there would be some debate whether towns like Presque Isle, Maine or Presque Isle, Ohio should be included. However, it is clear New York City must be included. For this particular purpose, New York City is more important than most other cities. Therefore, importance (or rank) according to a particular concept is a property of elements in spatial datasets that may need to be preserved in subsets. We will use the concepts of density, dispersion, and pattern to evaluate the similarity of spatial datasets.

Sets of Spatial Objects with Similar Dispersion					
Points		Lines		Areas	
A	B	A	B	A	B
					

**Figure 3: Similar dispersion of sets of spatial objects within a reference area.**

The automated extraction of geographic knowledge in data-rich environments will require consideration of the synoptic attributes of the dataset as well as the individual attributes of the data elements. If a subset of a spatial dataset retains a large amount of the character of the original dataset then processing times can be reduced when extracting geographic knowledge. General

relationships can be discovered and proven or refined through closer and more detailed examination of the data.

### *References*

- A. L. Bloom (1978) *Geomorphology: A Systematic Analysis of Late Cenezoic Landforms*. Prentice-Hall, Englewood Cliffs.
- FGDC (1994) *The 1994 Plan for the National Spatial Data Infrastructure*. Federal Geographic Data Committee, Technical Report
- FGDC (1997) *Content standard for digital geospatial metadata (revised April, 1997)*. Federal Geographic Data Committee. Washington, D.C., Technical Report
- D. M. Flewelling and M. J. Egenhofer (1993) Formalizing Importance: Parameters for Settlement Selection. in: R. McMaster (Ed.), *11th International Conference on Automated Cartography*, Minneapolis, MN, pp. 167-175.
- H. Garcia-Molina, J. Widom, J. Wiener, W. Labio, B. Lent, and Y. Zhuge (1995) *A Warehousing Approach to Data and Knowledge Integration*. Stanford University, Technical Report
- M. F. Goodchild (1992) Analysis. in: R. F. Abler, M. G. Marcus, and J. M. Olson (Ed.), *Geography's Inner Worlds*. 1, pp. 138-162, Rutgers University Press, New Brunswick, New Jersey.
- R. E. Horton (1945) Erosional development of stream and their drainage basins: hydrophysical approach to quantitative morphology. *Bulletin, Geological Society of America* 56: 275-370.
- N. Kadmon (1972) Automated Selection of Settlements in Map Generalization. *CJ* 9(1): 93-98.
- D. M. Levy and C. C. Marshall (1996) Going Digital: A Look at Assumptions Underlying Digital Libraries. *Communications of the ACM* 38(4): 77-84.
- T. R. Smith (1996) A Digital Library for Geographically Referenced Materials. *Computer* 29(5): 54-60.
- D. Unwin (1981) *Introductory Spatial Analysis*. Methuen & Co., New York.
- S. Weibel, W. Cathro, and R. Iannella (1997) The 4th Dublin Core Metadata Workshop Report. URL <http://www.dlib.org/dlib/june97/metadata/06weibel.html> (Last accessed October 10, 1997).

### **3.5 Mark Gahegan**

#### *Neural Networks and Decision Trees in Geography: Knowledge Discovery through Classification*

My main area of interest is in the development of suitable methods to make proper use of inductive learning tools in a geographic setting, specifically for problems involving complex, high dimensionality datasets. Work to date has focussed on three different application areas: landcover classification (at the floristics level), spatial epidemiology and geological interpretation, and has led to the development of packaged neural network-based classifiers that are self-configuring or adaptive to the structure of a problem (<http://www.cs.curtin.edu.au/gis/donnet/>) and knowledge-based visualisation tools (<http://www.cs.curtin.edu.au/gis/visualisation/>).

The use of decision trees and (artificial) neural networks for data classification in geography and remote sensing has seen a steady rise in popularity. Kamata and Kawaguchi (1993) and Civco (1993) describe neural network classifiers whilst Lees and Ritman (1991), Eklund et al. (1994) and Freidl and Brodley (1997) describe classification approaches based around decision trees.

Initially, the focus of attention was on comparing classifier performance with established methods (eg. Benediktsson, et. al., 1990; Hepner et al., 1990; Paola and Schowengerdt, 1995; Fitzgerald and Lees, 1994). More recent efforts have concentrated on methodologies and customisation that improve performance or reliability; a sign that the technology has reached at least some level of acceptance. For example, Benediktsson, et al., (1993) and German et al. (1997) describe performance improvements and Kanellopoulos and Wilkinson (1997) and Gahegan et al. (1998) address methodological issues from the specific viewpoint of geographic datasets.

The kinds of classification problems that arise in geography or the wider earth sciences are often characterised by their complexity, both in terms of the classes and the datasets used. For example, classes may be difficult to define, may vary with location and over time, and their properties may overlap in attribute space. Datasets increasingly contain many descriptive variables (layers) and often contain a mix of statistical types; for example remotely sensed reflectance values (quantitative data) supplemented with nominal data such as soil type or geology and ordinal data such as slope or aspect. Data 'saturation' seems set to increase with the adoption of sensing devices of greater sophistication, resulting in a higher spatial resolution and many more channels. The complexity of the tasks to which these data are applied is also increasing; for example the classification of deep geological structure from 300 channel airborne electro-magnetics data, or socio-demographic indices from combinations of many indicator variables.

Addressing geographic classification problems successfully with tools based around inductive learning and search requires detailed attention to methodology and often also a good deal of further development and enhancement. To this end a black-box neural classifier has been specifically engineered for application to geographic problems (German and Gahegan, 1996). It is based on a feedforward, multi-layer perceptron, with various enhancements made (German et al., 1997; Gahegan et al., 1998) Importantly, it is designed to be self-configuring, requiring only the same setup as a standard Maximum Likelihood Classifier. The aim is to provide a tool where configuration and control parameters are gleaned from a preliminary analysis of the data. This analysis provides both a start condition and a means to control the adaptive behaviour of the network. The research has led to a range of modifications to the standard Feedforward network, including: scaling of inputs, initial network architecture and weight configuration, activation function, dynamic insertion, deletion and re-assignment of hyperplanes, output cost function and performance evaluation.

Results so far show increased classification accuracy over established techniques. Furthermore, the architecture appears scalable; accuracy and stability are maintained as the complexity of the task is increased (for example by increasing the number of data layers (attributes), the number of target classes or the inseparability of the classes).

### *References*

- Benediktsson, J. A., Swain, P. H. and Ersoy, O. K. (1990). Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE transactions on Geoscience and Remote Sensing*, Vol. 28, No. 4, pp. 540-551.
- Benediktsson, J. A., Swain, P. H. and Ersoy, O. K. (1993). Conjugate gradient neural networks in classification of multisource and very high dimensional remote sensing data. *International Journal of Remote Sensing*, Vol. 14, No. 15, pp. 2883-2903.
- Civco, D. L. (1993). Artificial neural networks for landcover classification and mapping. *International Journal of Geographical Information Systems*, Vol. 7, No. 2, pp. 173-186.
- Eklund, P. W., Kirkby, S. D. and Salim, A. (1994). A framework for incremental knowledge update from additional data coverages. *Proc. 7th Australasian Remote Sensing*

- Conference, Melbourne, Australia, Remote Sensing and Photogrammetry Association of Australia, pp. 367-374.
- Fitzgerald, R. W. and Lees, B. G. (1994). Assessing the classification accuracy of multisource remote sensing data. *Remote Sensing of the Environment*, Vol. 47, pp. 362-368.
- Freidl, M. A. and Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *International Journal of Remote Sensing*, Vol. 61, No. 4, pp. 399-409.
- Gahegan, M., German, G. and West, G. (1998). Some solutions to neural network configuration problems for the classification of complex geographic datasets. To appear in *Geographical Systems*.
- German, G. and Gahegan, M. (1996). Neural network architectures for the classification of temporal image sequences. *Computers and Geosciences*, Vol. 22, No. 9, pp. 969-979.
- German, G., Gahegan, M. and West, G. (1997). Predictive assessment of neural network classifiers for applications in GIS. *Second International Conference on GeoComputation*, Otago, New Zealand, pp. 41-50.
- Hepner, G. F., Logan, T., Ritter, N. and Bryant, N. (1990) Artificial neural network classification using a minimal training set: comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing*, Vol. 56, No. 5, pp. 469-473.
- Kamata, S. and Kawaguchi, E. (1993). A neural network classifier for multi-temporal Landsat images using spatial and spectral information. *Proc. IEEE 1993 International Joint Conference on Neural Networks*, Vol. 3, pp. 2199-2202.
- Kanellopoulos, I. and Wilkinson, G. (1997). Strategies and best practice for neural network image classification. *International Journal of Remote Sensing*, Vol. 61, No. 4.
- Lees, B. G. and Ritman, K. (1991). Decision tree and rule induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. *Environmental Management*, Vol. 15, pp. 823-831.
- Paola, J. D. and Schowengerdt, R. A. (1995). A detailed comparison of backpropagation neural networks and maximum likelihood classifiers for urban landuse classification. *IEEE transactions on Geoscience and Remote Sensing*, Vol. 33, No. 4, pp. 981-996.

### **3.6 Myke Gluck**

Information Retrieval (IR) as opposed to data retrieval (DR) may be described as the seeking of indeterminate information resources to resolve an information need. Determinate information resources are classic 'trivial' exact match or SQL-type retrieval; For example: How many widgets are left at the Chicago warehouse? The response from a system is determinate: 1200, say. However, the desire for some information on the future Asian market for widgets is an indeterminate query. Similarly seeking patterns in data and exploratory data analysis (EDA) are indeterminate searches: there is not just one best response known a priori, if existing at all. Indeterminate information seeking processes explore data rich environments which have historically been text-based or numerically based. Now, however, multimedia resources much of which have a geospatial nature, are exploding: this requires rethinking of the algorithms which match users' queries with the available information collection as well as the criteria for successful searching of large information systems. (For example, we can see glimpses of many of the issues to come in the work of the Alexandria Digital Library project.)

Indeterminate searches often need multiple sources to be retrieved and integrated in effective and efficient ways to address the information need. With the goal of this initiative to find new automated methods for filtering large amounts of raw geographic data into more user-consumable forms of knowledge, I suggest that at least three major components need to come together for improvements to be more than merely incremental. These components involve 1) improved

metadata descriptions and related standards that focus on user's needs, not just content description, 2) more robust and user based measures of information retrieval and data integration effectiveness to guide the retrieval engines, and 3) multimedia interactive tools that allow users to explore FILTERED geographic data more efficiently and effectively to seek patterns in the data or extract meaningful summaries of the information.

#### *Metadata Information Integration*

Metadata are frequently designed by experts in the data and not those with expertise in the usability of data or of deep cognitive understanding of user behavior. The necessary preliminary and tentative modifications of metadata for use by the Alexandria Digital Library Project indicate this gap quite clearly. Solutions to these gaps that would indeed help and not hinder user ability to locate and access the needed information are not trivial and need better understanding of user requirements to resolve information need not just for transfer or interoperability of datasets and software. One aspect of this process is the need for metadata that describes at more than the highest level of a dataset. (Classical book cataloging is not amenable to these processes for this (and many other) reasons: Try to find the sheet with a useful inset map on a larger cartographic sheet in a map library?)

One possible avenue for development is the use of more microlevel metadata within an information bearing object, some of which may be amenable to automated processing. Such higher resolution may permit data integration across information bearing objects so that a single new integrated dataset is delivered to the user which is much more focused on the current information need without having to wade through the extraneous and irrelevant information and data within the larger individual information bearing objects. For example, if information on the hardwood hammocks of the Apalachicola National Forest within the boundaries of Leon County, Florida, automated tools to take the county boundary files and do the necessary relational algebraic operations on the Forest dataset would be of value. Such tools would need to process micro level metadata files and execute the operations transparently to the user, providing a retrieved set that is the integration of the initial raw information bearing objects with quality and accuracy statements of the 'new' integrated or retrieved set. Simultaneously, a third object containing a video clip displaying the hammocks could also be retrieved and integrated into the final delivered information bearing object. The algebra is clear and 'straightforward' when the relational data is selected but tools for discerning the subsets of the data to integrate are not currently available for analyst support. Such tools will require rethinking the role of metadata in retrieval processes, alternative forms of query posing, and the understanding of user needs.

#### *Retrieval Measures*

Traditional information retrieval (IR) used two measures of success: precision and recall. Precision being defined as the ratio of the number of relevant items retrieved to the number retrieved while recall defined as the ratio of the number of relevant items retrieved to the total number of relevant items in the information collection. Automated methods for computing these measures are really not available since they depend upon the assessment of relevance of items and only the user can actually make that judgment. Complicating this is the infeasibility of having users assess the relevance of several million or billion items in a large data store for each posed query. The use of multimedia tools (graphics and sound) to form surrogates and to form a viewable and explorable information space to assist retrieval is encouraging. Multimedia tools allow users to explore clusters of information objects or data subsets and select from a visual/auditory presentation those of interest or seeming to form meaningful patterns. However, the construction of such spaces for multimedia viewing is still dependent upon system assessment of relevance to form clusters. Research into novel measures of retrieval success is needed to avoid

merely restating the relevance question. Historically, researchers have suggested various weighting schemes and similarity measures for assessing the relevance of an information object to a query but these have not yet led consistently to more than minor improvements, if even that, from a user's perspective. Having users determine the weights merely complicates the use process with a steep learning curve for consistent weight assignments to query terms. Few schemes have incorporated the spatial component into retrieval algorithms to attempt to improve precision or recall.

#### *Tools for Exploratory Data Analysis*

Tools to assist users, especially expert users, in the exploration of large filtered or unfiltered data sets is a third component necessary to help automate or support the discovery of geographic knowledge in data-rich environments. Though the initiative seeks automated tools for filtering, automation itself is a continuum. Note a wordprocessor is an automated tool but requires an author to create a work. Thus, it seems reasonable to seek human-aided automated tools as well as machine-aided human tools for improving discovery in such environments.

As an example of a tool that combines automated and manual support for data mining and data exploration, we have been developing a seriation tool. The tool takes a data matrix and re-expresses the raw data into fewer classes replacing the numerical value in the matrix cell with an icon proportional in size to the classed data value. The user then is able to visually explore the data, reclass the data, and to manipulate rows and columns of the classed data to seek meaningful patterns 'manually.' Included in the functions of the tool is an automated seriation algorithm for both dichotomous and multi-valued data. The seriation process is a univariate scaling process much like a visual principal component analysis. Multimedia support for this tool includes simultaneous display of single variable and bivariate choropleth maps, brushing associated with the classed data matrix, and sound functions to 'hear' the classed data. Preliminary user testing indicates that especially for data experts this tool is usable and useful. Research into algorithms for multi-valued seriation and the various visual and auditory functions to coordinate the maps with the matrix for data exploration and data mining are ongoing. This tool as just one in a class of tools that could permit users to easily re-express data (reclass, move median windows, etc.) allowing the analyst to explore the filtered data in conjunction with automated statistical methods seeking meaningful patterns. Such tools could readily incorporate scene analysis and image and sound processing algorithms allowing video and audio data to also be explored along with the numerical and cartographic data.

#### *Position Summary*

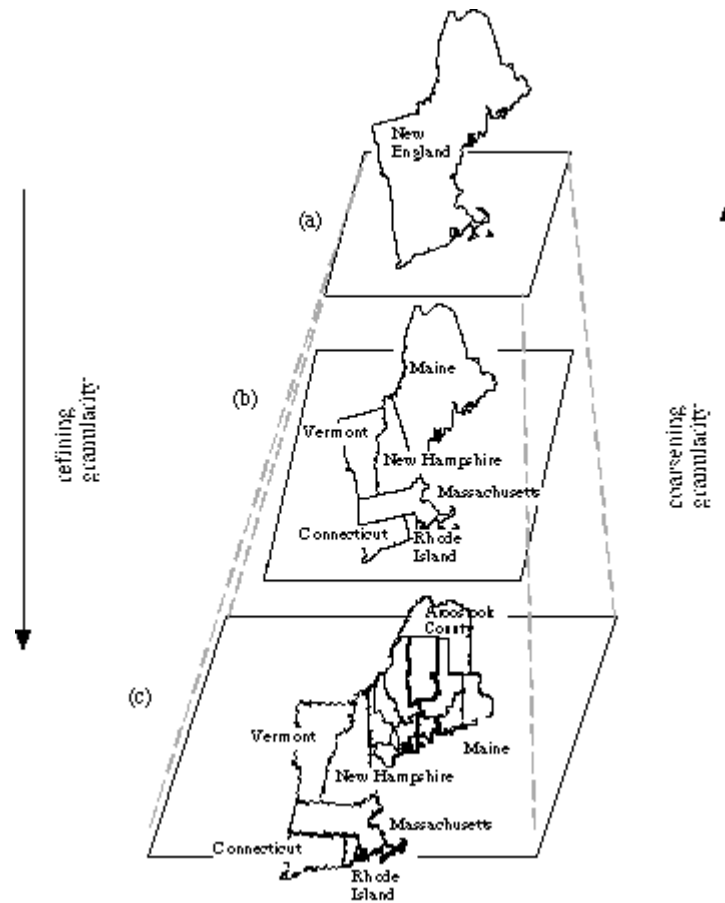
Data mining and knowledge extraction from data rich environments is a complex task. It is a creative process and fully automated tools will still need user intervention to assess the outputs. It seems hopeful to attempt to devise such tools driven by at least three concepts: improved metadata descriptions and related standards that focus on user's needs, not just content description permitting integration of information bearing entities, 2) more robust and user based measures of information retrieval and data integration effectiveness to guide the retrieval engines, and 3) multimedia interactive tools that allow users to explore FILTERED geographic data more efficiently and effectively to seek patterns in the data or extract meaningful summaries of the information.

### **3.7 Kathleen Hornsby**

*Spatio-Temporal Knowledge Representation over Shifting Granularities.*

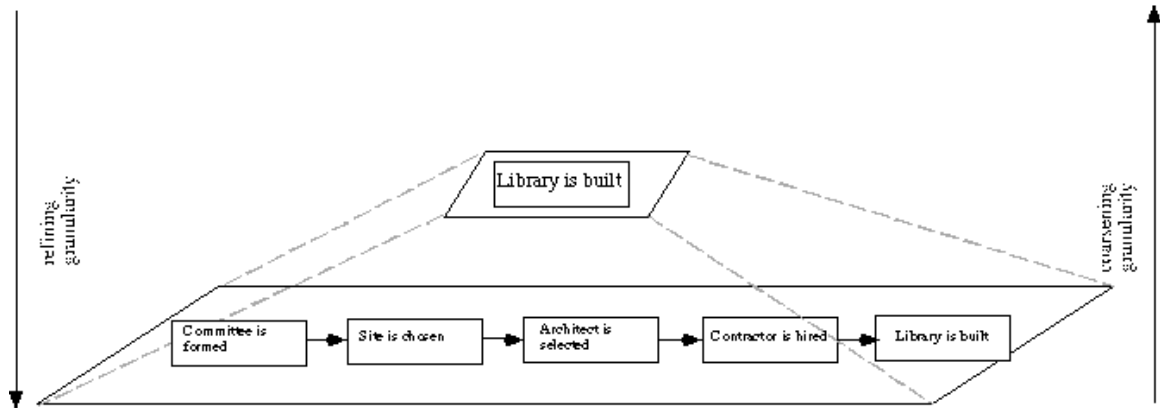
Recent efforts in spatio-temporal knowledge representation to explicitly describe scenarios of change as commonly experienced by geographic phenomena (Claramunt and Thériault 1996; Renolen 1996; Hornsby and Egenhofer 1997; Hornsby and Egenhofer 1998) offer new insights and opportunities for reasoning in data-rich environments. One aspect of developing approaches that convey how phenomena change over space and time is to gain an appreciation that spatio-temporal knowledge representation often requires shifting from one level of detail or granularity to another in order to view change or an entity. Granularity is a term used by the AI community among others to capture the notion that the world is perceived as at different grain sizes or granules (Hobbs 1990; Dyreson and Snodgrass 1995). People view the world at different granularities, abstracting from the world only those things that serve their present interests (Hobbs 1990; Lam and Quattrochi 1992; Goodchild and Proctor 1997). These shifts in granularity enable people to translate the complexities of the real world into simpler representations. The converse, namely shifting to a more detailed view is also a common requirement. In the context of data-rich environments, the need to shift granularities is particularly important. Efforts are underway to define a set of operations that enable users to move between different granularities (Stell and Worboys 1998) but further work is essential. Operations to collapse complex spatio-temporal scenarios into coarser granularities are not currently implemented but are necessary as are operations to support refinements in granularity.

Preliminary work on modeling scenarios of spatio-temporal change through tracking alterations to an object's identity (Al-Taha and Barrera 1994; Hornsby and Egenhofer 1997) have suggested different views of objects that are possible: objects of single identity and objects with parts termed composite objects (Hornsby and Egenhofer 1998). Objects refer to the representation of a real world phenomenon in an information system (Kim 1990) that might exist as a physical entity, such as a building or a mountain, or something conceptual, such as Penobscot County or the United States (Smith 1995). Shifting granularities can be understood from two orthogonal perspectives. First, granularity of objects refers to changes that coarsen a view of an object, for example a particular county, Aroostook County in the State of Maine, to a less detailed conceptualization containing fewer parts, such as the six states that comprise New England, or to the highest level of abstraction, a single object, the New England region, (Figure 1). In certain circumstances, an object may even disappear altogether as objects that are unimportant to the task at hand are abstracted away. Alternatively, a shift in the opposite way moves from a single object to a more refined granularity of objects where composite objects and their parts become relevant.



**Figure 1: Shifting granularities over: (a) New England region, (b) the New England states, and (c) Aroostook County in Maine.**

Second, granularity also applies to transitions between temporally-ordered objects. This type of shift is particularly interesting in the context of data-rich environments. As a simple example, the building of a new public library in a town can be perceived and modeled as one operation on a single object or from a different perspective, multiple operations over multiple objects, where each of the separate tasks and events over time that contributed to the creation of the building might be relevant (Figure 2). This refinement of detail gives us more information about the steps in the planning and building process over time. Adjusting the granularity over transitions allows one to refine or coarsen the temporal resolution as desired for the task at hand. More complex examples can be drawn from analysis of battlefield environments, for instance, where the position and movements of many participants might be tracked over time and reasoning performed at different granularities.



**Figure 2: Shifting granularities over transitions between temporally-ordered objects.**

A set of operations to support shifts in granularities over objects that refine or coarsen people's views and additional operations to support shifts that expand the sequence of transitions or collapse the sequence into fewer transitions will be necessary in data-rich contexts. This research focuses on formalisms to support spatio-temporal knowledge representation under shifting granularities.

#### *References*

- K. Al-Taha and R. Barrera (1994) Identities through time. in: M. Ehlers and D. Steiner (Eds.), International Workshop on Requirements for Integrated Geographic Information Systems, New Orleans, LA, pp. 1-12.
- C. Claramunt and M. Thériault (1996) Toward semantics for modelling spatio-temporal processes within GIS. in: M. Kraak and M. Molenaar (Eds.), 7th International Symposium on Spatial Data Handling, Delft, NL, pp. 47-63.
- C. Dyreson and R. Snodgrass (1995) Temporal granularity. in: R. Snodgrass (Ed.), The TSQL2 Temporal Query Language. pp. 347-394, Kluwer Academic Publishers, Norwell, MA.
- M. Goodchild and J. Proctor (1997) Scale in a digital geographic world. *Geographical and Environmental Modelling* 1(1): 5-23.
- J. Hobbs (1990) Granularity. in: D. S. Weld and J. d. Kleer (Eds.), *Readings in Qualitative Reasoning about Physical Systems*. pp. 542-545, Morgan Kaufmann Publishers, Inc., San Mateo, CA.
- K. Hornsby and M. Egenhofer (1997) Qualitative Representation of Change. in: S. Hirtle and A. Frank (Ed.), *Spatial Information Theory: A Theoretical Basis for GIS*, International Conference COSIT '97, Laurel Highlands, PA. Lecture Notes in Computer Science 1329, pp. 15-33, Springer-Verlag, Berlin.
- K. Hornsby and M. Egenhofer (1998) Identity-based change operations for composite objects. in: T. Poiker and N. Chrisman (Ed.), 8th International Symposium on Spatial Data Handling, Vancouver, Canada, pp. 202-213.
- W. Kim (1990) *Introduction to Object-Oriented Databases*. The MIT Press, Cambridge, MA.
- N. Lam and D. Quattrochi (1992) On the issues of scale, resolution, and fractal analysis in the mapping sciences. *Professional Geographer* 44(1): 88-98.
- A. Renolen (1996) History graphs: Conceptual modelling of spatiotemporal data. in: *GIS Frontiers in Business and Science*, Brno, Czech Republic.

- B. Smith (1995) On drawing lines on a map. in: A. Frank and W. Kuhn (Eds.), COSIT '95, Semmering, Austria, pp. 475-484.
- J. Stell and M. Worboys (1998) Stratified map spaces: a formal basis for multi-resolution spatial databases. in: Eighth International Symposium on Spatial Data Handling, Vancouver, Canada, pp. 180-189.

### **3.8 Eric D. Kolaczyk**

An increasingly important problem in working with spatial data is that of processing and combining information from data at a variety of scales. The quantity of available geospatial data is growing exponentially with developments such as the launching of new satellite systems and the widespread use of geographic information systems (GIS). Both of these technologies facilitate data collection at multiple spatial and temporal scales. The ability to handle scale properly therefore would seem to be an important characteristic for next-generation methods of spatial data mining to possess.

From the standpoint of statistical and/or mathematical analysis, the word scale is by association likely to bring to mind the word wavelet. Since their introduction in the mid-1980's, wavelets arguably have been used as the foundation for the vast majority of multi-scale methods offered for mathematical and statistical analysis in the last decade. At a fundamental level, wavelet-based methods, like their Fourier predecessors, are based on an approach that advocates transforming an object of interest, for example, a vector of time series data or a matrix of image data, by matching it (via a mathematical inner product) against a predetermined collection of shapes -- in this case localized waveforms at a variety of locations and scales. Such an approach has been found to be quite appropriate in, for example, image processing tasks, where the primary interest may be in the visual quality of a final smoothed image or the long-term success in classification of image types. In such contexts, the wavelet transform is used mainly as a tool for more efficiently representing the information within the data, according to a collection of location-scale combinations, in the form of the wavelet coefficients.

With many problems in spatial analysis, however, a wavelet transform seems a less appropriate tool for creating a multi-scale model. For example, an arbitrary wavelet shape may not be easily interpretable within the context generating a given dataset; or, put another way, there may be an important aspect of interpretability possessed by the data in its original form that is lost upon computing a wavelet transform and examining the resulting wavelet coefficients. Instead, often it is desirable to have a framework for creating multi-scale models within which is preserved the natural tendency to envision scale simply as a result of aggregation and disaggregation. Towards this end, an approach based on partitioning (e.g., the operation underlying such algorithms as CART [1]) would seem natural.

Quite recently several authors (including myself) have begun to pursue this line of research in introducing a new class of multi-scale models within the statistics [2,3,4] and digital signal processing [5,6] literatures. Through the use of recursive dyadic partition (RDP) strategies, the simplicity and interpretability of partition-based methods is combined with the ability to create efficient location-scale representations usually attributed only to wavelets. This basic framework initially was developed for standard signal and image processing tasks, yet has the potential to generalize to a variety of tasks arising with the analysis of spatial data in geography. Such generalizations recently have become the focus of a collaborative effort between groups in the Department of Mathematics and Statistics and the Department of Geography at Boston University.

If the data take the form of a time series, the partition-based multi-scale models, in their original formulation, begin by recursively partitioning the data space in half, then quarters, then eighths, and so on and so forth, up to the resolution of the data. For two-dimensional image data, the operations are defined analogously, working in both vertical and horizontal directions. In either case, the result of this partitioning is a hierarchy of aggregations of the data. Beyond the convenient interpretability of these operations, however, lies a fundamentally important and powerful accompanying property: a multi-scale likelihood factorization. Recalling the principle that the information in the data is summarized by the data likelihood, the multi-scale likelihood factorization shows explicitly how to re-express this information in such a way as to isolate at each spatial-scale combination that which is new in moving from a coarser level of aggregation to one level less coarse. In particular, this information is expressed through certain conditional probabilities that possess simple, tractable mathematical expressions for common data-types such as Gaussian or Poisson.

Given the above described structure, specific tasks like estimation and classification may be pursued, typically with the inclusion of additional information in the model through the use of prior distributions on the set of canonical multi-scale parameters accompanying the multi-scale likelihood factorization (e.g., [4,5]). The end result, viewed from a certain level of abstraction, is a type of hierarchical statistical model and, therefore, a type of graphical model [7].

From the vantage of graphical models, it becomes more clear why these multi-scale models have the potential to be generalized to arbitrary nested hierarchies of aggregation, and hence to a variety of contexts in spatial analysis. Accompanying such generalizations are a number of methodological and computational challenges, but many are expected to be tractable due to the modular form possessed by these types of graphical models. Finally, another key feature of this framework to note is that the underlying graph-based structure should allow for user-interpretable graphical displays and tools to accompany the modeling process (e.g., graphical user interfaces (GUIs)). This in turn may provide a natural way for combining our statistical modeling with recent advances in hierarchical spatial databases (e.g., [8]).

### *References*

1. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth International Group : Belmont, California.
2. Engel, J. (1994). A simple wavelet approach to nonparametric regression from recursive partitioning schemes, *Journal of Multivariate Analysis*, 49, 242-254.
3. Donoho, D. L. (1997). CART and best-ortho-basis selection: A Connection, *The Annals of Statistics*, 25, 1870-1911.
4. Kolaczyk, E. D. (1998). Bayesian Multi-Scale Models for Poisson Processes, To be published in the *Journal of the American Statistical Association*.
5. Timmerman, K. E. and Nowak, R. D. (1998). Multiscale modeling and estimation of Poisson processes with applications to photon-limited imaging, To be published in *IEEE Transactions on Information Theory*.
6. Nowak, R. D. and Kolaczyk, E. D. (1998). A Bayesian Multiscale Approach to Poisson Inverse Problems, *Proceedings of the 32nd Asilomar Conference on Signals, Systems, and Computers*.
7. Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon Press.
8. Rigaux, P. and Scholl, M. (1995). Multi-Scale Partitions: Application to Spatial and Statistical Databases. In *Advances in Spatial Databases, Proceedings of the 4th International Symposium*. Egenhofer and Herring (eds.): Springer-Verlag.

### **3.9 Brian Lees**

#### *The Importance of Data and Research Design in Extracting Geographic Knowledge from Large Complex Spatial Data Sets*

During the last 10 years my group at the Australian National University has pursued a continuing strategy concerning the development of methods and software tools for the analysis of very large spatial data sets with diverse characteristics and high levels of redundancy. The main impetus for the focus in this area were the early specifications of the EOS programs and polar-orbiting platforms, as they were then envisaged. A further impetus was given by the 1991 Australian IGBP Workshops on Global Change where a very high reliance was placed on spatial data analysis as an important, future, scientific resource for dealing with the large data flows being discussed. At that stage, techniques for dealing with this sort of data volume were not really well understood.

Our initial work in Decision Trees was targeted at producing a data selection tool to deal with pre-analysis data preparation for the HIRIS instrument. It soon became apparent that we required a large data set, whose characteristics were well understood, to facilitate the development and testing of our techniques. This led to a parallel project, the establishment of the Kioloa Pathfinder Global Land Cover Test Site and its associated spatial databases.

The Kioloa data set consists of high quality GIS, remotely sensed and point data sets. The behaviours of disparate data types, changes in scale and temporal sequence have all be examined using this data set. Over the 10 years we have progressively investigated the refinement of non-parametric statistics and the incorporation of computational intelligence techniques such as decision trees (Moore et al., 1991; Lees and Ritman, 1991)), a wide range of artificial neural net configurations (Fitzgerald and Lees, 1993, 1994, 1996), and genetic algorithms (Payne, 1998)) and compared their performances with a number of more conventional parametric techniques (Lees, 1994, 1996a, 1996b). Our major insights have resulted in the development of a more strategic approach to analysis. In particular, our spatial data mining experiments using Kohonen nets have made us very aware of the importance of data descriptors.

#### *Problem Statement*

The implications of adopting computational intelligence techniques have not yet been fully appreciated by many researchers working in spatial data analysis. These computational solutions comprise several critical components. Firstly, the new hardware configurations without which they could not be implemented. Secondly, the new algorithms themselves. Thirdly, the data, and fourthly, the problem. Perhaps this fourth element could be better described as the problem statement. Success depends on the adequacy of all of these, and on their correct integration. The data, the problem statement and the links between the algorithm and the data, between the data and the problem statement, and the problem statement and the algorithm have all received much less attention than the matching of algorithms and hardware. As we become increasingly engaged in harnessing our new computer power to the extraction of geographic knowledge in data rich environments it is important to consider these other, equally important, facets of quantitative investigation, analysis and prediction using spatial data.

### *Data*

After decades of accepting the questionable proposition that most phenomena in the natural world are normally distributed we are now adopting non-parametric methodologies with enthusiasm. The fact that many of these can be parallelised has, perhaps, added to this enthusiasm. There are some costs in this enthusiasm. To produce results matching the quality of the more conventional, parametric, techniques in the analysis of non-normally distributed data using non-parametric methods, much larger and more carefully structured samples are needed. By definition, non-parametric, supervised inductive learning systems have no information on the distribution of the data other than that can be inferred from the learning sample. Few of the studies which have appeared in the literature indicate that this has been recognised. That this is such a problem is perhaps due to the fact that many of those involved are not data gatherers, but data processors. There is a real temptation to use 'legacy' data sets for experiments and this leads to the use of proportional samples. Given the error minimisation rule on which many of these systems are based, the use of a proportional sample as a learning sample will bias the system towards the largest categories.

### *Data Models*

Much of the published work on data models focuses on the data model as the rationale for organising data in the computer. In computer science it is a means of capturing the semantics of the data through definitions of the operations related to classes, describing which combinations of operations are legal, which combinations of operations are equivalent, and consistency constraints among data. This bias towards the computer science view of data models is quite understandable as it is a necessary tool to deal with the data, but many phenomena have not been carefully scrutinised by domain experts in the same way and I suspect that, when this happens, the whole concept of data model will become considerably more complex and critical.

When we start to consider whether the measure used to code the data is appropriate to the phenomena we wish to examine we need to remember that many disciplines, including geography, routinely classify data as part of their collection protocols. This pre-analysis processing is often not recognised as such but can be a major limitation to accurate prediction based on such sampling. All too often phenomena distributed as a continuum are discretised into gaussians on the assumption that this is an appropriate data model for the phenomenon. That this is unnecessary now that we are no longer bound by the cartographic model of spatial data has not really penetrated the consciousness, and standard procedures, of many disciplines.

All other things being equal, if one can provide a learning system with some indication of how values in an attribute relate one to another, then the system will do a better job. Humans like to simplify these relationships as we are unable to deal very effectively with high frequency variability in data. By coding data to suit human perceptions, we degrade it and remove information a non-human learning system may be able to interpret. Deriving appropriate measures to code data for a machine learning system requires a knowledge of both the attribute, and the interactions of attributes relating to the phenomenon being modelled. This is a vital step in the successful extraction of geographic knowledge in data rich environments.

### *Data Scale & Data Domain*

Both spatial and temporal variability are strongly scale dependent. There is a general trend in most land cover data for spatial autocorrelation to be low at fine scale, to rise to a maximum at an intermediate scale and then to decline. One can see a similar pattern in many forms of temporal data. Spatial and temporal variability also depends on the data space, or domain, in which one views the data. Spatial data can be envisaged as existing in a number of discrete domains (Lees,

1994; Aspinall & Lees, 1995). In each of these there exist topological relationships, but these relationships vary from domain to domain.

This view of data existing in 'domains' has proven useful in dealing with the complex analysis of data sets from disparate sources. It has been discussed before in the context of decision tree analysis of mixed data sets from GIS, remote sensing and ground point observation. In these cases, the analysis tools were envisaged as stepping from 'domain' to 'domain' to partition the data set optimally. Our, later work with artificial neural nets envisaged this process taking place in parallel, rather than in series. The basic concept can even be extended to encompass Harvey's (1973) discussion of methodological problems at the interface of spatial and social analysis.

There is often confusion about what can, reasonably, be used as an analytical 'data domain' or space. Most spatial analysis tools available in GIS are based on Waldo Tobler's 'First Law of Geography'. This states that 'everything is related to everything else, but near things are more related than distant things' (Tobler, 1970). The data domains are constructs within which this relationship is optimal. If Tobler's First Law is truer of our data set in 'Environmental Data Space' than it is in 'Geographic Space', then it suggests that the former 'space' is the most appropriate context for analysis. Domain knowledge is fundamental to constructing the necessary spaces for analysis, and for understanding the relationships between the spaces. In many problems different parts of the analysis need to be carried out in different data spaces. A conceptual framework for Spatial Data Mining which incorporates this concept tends to lead to more successful, and fewer idiosyncratic, results.

### *References*

- Aspinall, R. & Lees, B.G. 1995. 'Sampling and analysis of spatial environmental data.' in Waugh, T.C. & Healey, R.G. (eds) *Advances in GIS Research*, Taylor and Francis, Southampton, 1086-1099.
- Fitzgerald, R.W. & Lees, B.G., 1993. Assessing the classification accuracy of multisource remote sensing data. *REMOTE SENSING OF THE ENVIRONMENT*, 41: 1-25.
- Fitzgerald R.W., & Lees, B.G. 1994. 'Spatial context and scale relationships in raster data for thematic mapping in natural systems.' in Waugh, T.C. & Healey, R.G. (eds) *Advances in GIS Research*, Taylor and Francis, Southampton, 462-475.
- Fitzgerald, R.W., & Lees, B.G. 1996. 'Temporal context in floristic classification.' in Lees, B.G. (ed) *Neural Network Applications in the Geosciences. Special Volume, Computers and Geosciences*, 22:9, 981-994. (C1).
- Harvey, D. (1973) *Social Justice in the City*, London, Edward Arnold.
- Kohonen, T., 1984. *Self-Organisation and Associative Memory*. Springer Verlag, New York.
- Lees, B.G. 1994. 'Decision trees, artificial Neural Networks and Genetic Algorithms for classification of remotely sensed and ancillary data.' *Proceedings 7th Australasian Remote Sensing Conference, Remote Sensing and Photogrammetry Association, Australia Ltd, Floreat, W.A.* v1: 51-60.
- Lees, B.G., 1996. 'Sampling strategies for machine learning using GIS' in *GIS and Environmental Modelling: Progress and Research Issues*, Goodchild, M.F., Steyart, L., Parks, B., Crane, M., Johnston, C., Maidment, D., and Glendinning, S. (eds), GIS World Inc, Fort Collins, Co. ISBN 1-882610-17-2. (B).
- Lees, B.G. 1996a. 'Inductive modelling in the spatial domain.' in Lees, B.G. (ed) *Neural Network Applications in the Geosciences. Special Volume, Computers and Geosciences*, 22:9, 955-957. (C1).
- Lees, B.G. 1996b. 'Improving the spatial extension of point data by changing the data model' in *Integrating GIS and Environmental Modelling* (eds) Goodchild, M. et al., Santa Barbara : National Centre for Geographic Information and Analysis, WWW;CD.

- Lees, B.G. & Ritman, K. 1991. A decision tree and rule induction approach to the integration of remotely sensed and GIS data in the mapping of vegetation in disturbed or hilly environments. ENVIRONMENTAL MANAGEMENT, 15, 823-831.
- Moore, D.M., Lees, B.G. & Davey, S. 1991 'A new method for predicting vegetation distributions using Decision Tree Analyses in a Geographic Information System.' ENVIRONMENTAL MANAGEMENT, 15, 59-71.

### **3.10 Donato Malerba**

Information extraction from maps is not a trivial task since objects in the charts must be interpreted from their spatial characteristics such as shape, scale and contours. In order to recognize a symbol in a map when it is not explicitly reported in the legend, some mechanisms for defining and manipulating the symbols at a semantic level are necessary.

It is possible to define different tasks corresponding to the steps performed by human experts, such as geologists, during their process of recognition and interpretation of morphological concepts on a map. The first task is feature selection, which is useful for the recognition of the major components in the map and is centered on the shape of objects, their size and ground characteristics. Then the expert is ready to single out the major elements in the map and their possible relations. A rather complex description language is required to describe the spatial characteristics of an object in the map, concerning the shape, the orientation, the dimensions etc., as well as the relations intra-object and inter-objects. A suitable logic language in clause form can be used for this purpose.

The problem of describing the map without losing the ontological expressive power of the image, may be solved by superimposing a grid, that is by dividing the map into regular observation elements, called cells, the dimension of which is related to scale factors and to the selected level of detail. Psychological studies (Fischler and Firschein, 1987) have shown that very little can be recognized when a scene is viewed partially through a small window: Correct interpretation must be decided entirely by the context. Thus, the choice of a suitable grid on the map in order to indicate the cells should also take into account the problem of losing the continuity of a spatial concept when it appears over different, although adjacent, cells.

The subsequent steps are based on the use of explicit background knowledge and context. In fact, the individuation of morphologies characterizing the landscape, the selection of the important environmental elements, both natural and artificial, the recognition of forms of territorial organization requires abstraction processes and deep domain knowledge. Although the human process of image interpretation may be cast as a process of pattern recognition, techniques for the induction of intensional descriptions of patterns, that is, techniques of conceptual learning, appears to be more suitable for this task.

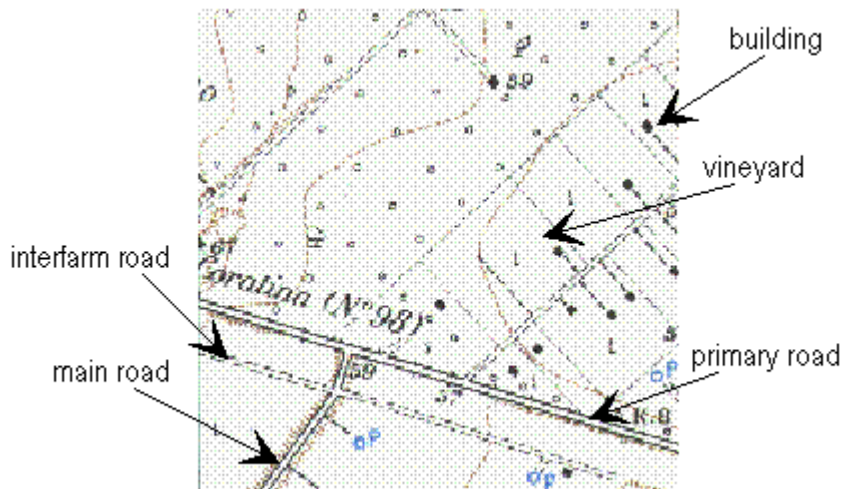
In a study concerning the Apulian region (Italy), an attempt to interpret maps by applying conceptual machine learning techniques has been made. The territory considered for this study covers 246 Km<sup>2</sup> of the landscapes around the Ofanto River, spanning from the zone of Canosa until its mouth. The goal is that of recognizing four morphological concepts deemed relevant for environmental protection: regular grid systems of farms, fluvial landscape, system of cliffs, and royal cattle track. The former consists of partitions of farms, which are arranged in a rectilinear way. The fluvial landscape is individuated by the presence of waterways, fluvial islands and embankments. The system of cliffs is related to the emergence of limestone as a single block.

Finally, the royal cattle track, exclusively present in Southern Italy, is a road for the transhumance.

The examined area is covered by five maps drawn to the scale of 1:25000. All maps are produced by the Italian Military Geographic Institute (IGM). They are segmented into regular square cells so that the problem of recognizing the four morphological concepts can be cast as a problem of labeling each cell with at most one of the following labels: Regular-grid, fluvial-landscape, system-of-cliffs and royal-cattle-track. Unlabelled cells are considered uninteresting with respect to the goal of environmental protection.

The choice of a suitable grid on the map is critical, since a cell too big might not be consistently labeled in one way, while a cell too small may cause loss of important information which is fragmented into several, adjacent cells. In this study, the gridding system superimposed over IGM charts to the scale of 1:25000 has proven itself a suitable resolution for the recognition task. In particular each cell covers an area of 1 km<sup>2</sup>. The choice has been made after having observed that a pool of fifteen geomorphologists and experts in territory planning were able to give a unique interpretation to several cells cut out of the maps.

Each cell is described by means of a first-order logic language (see Figure 1). Some descriptors used to represent the cartographic objects are attributes, while others define relations between objects. They have been defined by studying the behavior of geomorphologists and experts in territory planning during their map interpretation process. Particular attention has been paid to the definition of general descriptors so that they can be appropriate to describe maps even when the scale of representation or the concepts to be learned change.



```
[class(x1)=royal_cattle_track] :-
    contain(x1,x2), contain(x1,x3), contain(x1,x4), contain(x1,x5), contain(x1,x6),
    type_of(x2)=primary_road, type_of(x3)=interfarm_road, type_of(x4)=main_road,
    type_of(x5)=vineyard, type_of(x6)=building, color(x2)=black, color(x3)=black,
    color(x4)=black, color(x6)=black, relation(x2,x3)=almost_parallel,
    relation(x2,x4)=almost_perpendicular, relation(x3,x4)=almost_perpendicular,
    distance(x2,x3)=115, density(x5)=low, density(x6)=low, outside(x2,x3,x5),
    geographic_direction(x1,x2)=north_west, geographic_direction(x1,x3)=north_west
```

**Figure 1: An example of cell and its corresponding description**

Classification rules of the morphological concepts have been automatically generated using a conceptual learning system developed at the University of Bari, namely ATRE (Malerba et al., 1998). Forty-five training examples are taken from the same map. Some of the learned rules are reported below:

class(X)=royal\_cattle\_track :-

contain(X,Y), distance(Z,Y) in [90.0 .. 130.0] , type\_of(Y)=interfarm\_road.

class(X)=fluvial\_landscape :-

contain(X,Y), type\_of(Y)=river\_line, color(Y)=blue.

They can be automatically translated into the following natural language sentences: "if there is an interfarm road at a distance between 90 and 130 m from another object, then the cell contains a royal cattle track," and "if there is a blue line of type 'river' then the cell contains a fluvial landscape."

The interpretation task is performed using the induced rules, and the performance of the whole process is evaluated in terms of predictive accuracy, which is higher than 95% on an independent set of eighty-six cells extracted from the other four maps.

### *References*

M.A. Fischler, & O. Firschein (1987). *Readings in Computer Vision*, Morgan Kaufmann, Los Altos, CA.

D. Malerba, F. Esposito, & F.A. Lisi (1998). Learning recursive theories with ATRE, in H. Prade (Ed.), *Proc. of the 13th European Conference on Artificial Intelligence*, 435-439, John Wiley & Sons, Chichester, England.

## **3.11 Duane Marble**

### *Exploratory Data Analysis of Spatial–Temporal Interactions*

In a recent series of provocative lectures Freeman Dyson (1997) stated that "If we are looking for new directions in science, we must look for scientific revolutions. When no scientific revolution is under way, science continues to move ahead along old directions." Dyson notes that he recognizes two basic kinds of scientific revolutions, those driven by new concepts and those driven by new tools. While concept-driven revolutions that result in explaining old things in new ways have received the majority of methodological attention, they are actually relatively rare in science. Tool-driven revolutions that generally result in the discovery of new things that have to be explained have received considerably less methodological attention but, never the less, are comparatively more commonly encountered.

The development of geographic information systems (GIS) over the last three decades, with the recent melding of GIS and spatial analysis within the last decade of this century to form what is now being called geographic information science (GIScience), is created a new tool-driven revolution that appears highly likely to dominate scientific research in geography, and other disciplines concerned with the spatial/temporal structure of our world, during the early decades of

the coming century. The wide spread adoption of GIS technology in both the public and private sectors has also led to the creation and general distribution of spatial and spatial/temporal databases of increasingly massive size and detail. Currently, the conceptual basis for much of spatial analysis is inadequate to actively assist in understanding the complex phenomena that we are encountering in these databases. This situation results from a long term, myopic focus by much of the spatial analysis community upon conceptually simple, spatially aggregative problems coupled with an exceedingly slow acculturation by many in the same community to the potential, major advances made available as a result of the rapidly development of GIS technology.

As Dyson noted, tool-driven revolutions focus upon the discovery of new things that must be explained and in geography the explosive growth of disaggregative, spatial-temporal databases have provided us with a host of things that are crying out to be explored. Some years ago, Walter Isard (1960) noted that location and interaction are two sides of the same coin. Aside from the seminal works such as that of Ullman (1957) and Berry (1966), it seems puzzling, until the question of tools is considered, that geographers and others have paid far more attention to the former than the latter. Very few tools for the analysis of complex data sets covering human interaction (transportation, migration and communication) are available and this has significantly hampered researcher activities in this area. This interaction between tools and problems is well known (Marble, 1990).

In an effort to correct this imbalance and to improve our knowledge of spatial and spatial/temporal interactions, the exploratory data analysis of spatial-temporal flows, based in scientific visualization and statistics, has been the basis of a structured research effort at The Ohio State University for several years (Marble, et al, 1995 and Marble, et al, 1997). These research activities are directed toward the design and proof-of-concept testing of scientific visualization and dynamic graphics-based tools for the exploratory analysis of spatial-temporal, interregional flow systems. This series of powerful, interactive tools can assist researchers in generating hypotheses relating to the spatial and temporal structure of interregional flows. It is our hope that the availability of such tools will lead to a resurgence of interest in flows, as opposed to locations, among geographic researchers.

The data representation found to be most useful involved the use of multi-dimensional dyadic origin-destination matrices. An important initial step in the research was to examine and significantly improve upon the useful but limited methods of computer-based flow mapping that had been proposed earlier. It was necessary to develop visualization approaches that would permit examination of the total set of flows and the identification within this set of "interesting subsets." Following this, our attention was turned to creating an ability to quickly, easily, and selectively explore the relationships that exist between various components of the flow data set (e.g., inflows vs. outflows) and between the various interregional flows and selected characteristics of the origin and destination regions (e.g., out-migration as related to unemployment in the region of origin). This was accomplished in our prototype system by introducing forward and backward brushing approaches.

Still another approach involved the use of fast algorithms for projection pursuit reduction of dimensionality in the multi-dimensional flow matrices to permit visualization of flows to take place. However one of the most interesting – and certainly the most frustrating – problem encountered was to move from a "interesting" case developed through browsing the database to the automatic identification of a set of similar cases within the database.

Much still remains to be done in the development of tools for the exploration of massive interregional flow data sets and it is my belief that participation in this Workshop will lead to both new thinking and, hopefully, to new patterns of interaction with other researchers.

### *References*

Berry, B. J. L. (ed.), 1966. *Essays on Commodity Flows and the Spatial Structure of the Indian Economy*. Department of Geography Research Paper No. 111, The University of Chicago.

Dyson, Freeman, 1997. *Imagined Worlds: The Jerusalem–Harvard Lectures*. Cambridge: Harvard University Press.

Isard, Walter, 1960. Interregional flow analysis, chapter in Walter Isard, *Methods of Regional Analysis*. New York, NY: John Wiley and Sons, Inc.

Marble, Duane F., 1990. "The potential methodological impact of geographic information systems on the social sciences," in Allen, Zubrow and Green (eds.), *Interpreting Space: GIS and Archaeology*. London: Taylor & Francis, Ltd.

\_\_\_\_\_, Z. Gou and L. Liu, 1995. Visualization and exploratory data analysis of interregional flows, in D. D. Moyer and T. Ries (eds.), *Proceedings, 1995 Geographic Information Systems for Transportation (GIS-T) Symposium*. Washington, D.C.: American Association of State Highway and Transportation Officials (AASHTO).

\_\_\_\_\_, 1997. "Recent advances in the exploratory analysis of interregional flows in space and time," (with Zaiyong Gou, Lin Liu and James Saunders), in Z. Kemp (ed.), *Innovations in Geographic Information Systems*. London: Taylor & Francis. [Keynote address given at the 4th National United Kingdom GIS Research Conference, Canterbury, England, April 1996.]

\_\_\_\_\_, 1999. "Geographic information system technology and decision support systems," *Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS-32)*. Los Alamitos, CA: Institute of Electrical and Electronics Engineers.

Ullman, E. L., 1957. *American Commodity Flows: A Geographical Interpretation of Rail and Water Traffic Based on Principles of Spatial Interchange*. Seattle, WA: University of Washington Press.

### **3.12 Raymond Ng**

There are the two aspects to my research contributions: data mining and content-based analysis. With respect to the first aspect, my first publication on data mining [1] about 5 years ago, in fact, started with spatial data in mind. At that time, because of one of my research grants, I worked with academic and industry people in GIS, forestry and remote sensing. The thesis of that publication is that natural distance functions exist for many forms of spatial data, and clustering can be an effective and efficient way to discover certain kinds of knowledge in spatial data.

My research group and I continued along the lines of analyzing spatial clusters. Focusing on geographic data, we developed a suite of techniques for further analyzing the identified spatial clusters [2]. Those techniques range from finding proximity relationships to shape matching with geographic features. A prototype was developed and demos were given in various international

conferences. To this date, I believe the prototype is still among the state-of-the-art systems for certain kinds of spatial analysis.

Another line of my data mining research that I believe is highly relevant to the workshop is outlier detection. There are many data analysis and mining techniques (e.g., clustering) that designate certain parts of the data as outlying and noisy. But the notions of outliers supported by those techniques are basically second-class to whatever the primary tasks happen to be. Two problems arise. First, those notions of outliers may not be general and may be too specific to the primary tasks, whatever that may be. Second, inefficiencies occur because effort is spent on the primary tasks. Thus, for outlier detection research, we seek to define what an outlier is in general, and develop efficient ways to identify exactly those outliers and nothing else.

My outlier detection research consists of two parts: depth-based methods and distance-based methods. In computational statistics, one can find outliers by defining some notions of depth over the dataset. One well-known notion is called Tukey depth. In a recent paper, my collaborators and I develop an algorithm for computing 2-D Tukey depth contours [3]. Our algorithm is more robust and orders of magnitude faster than state-of-the-art algorithms. Our algorithm has been generalized to 3-D. While this could be sufficient for many kinds of geographic data, the general problem with depth contours is that their computation does not scale up well with dimensionality, i.e., could be exponential with respect to the dimensionality.

Distance-based methods for outlier detection, in contrast, do not suffer from this problem. It is linear with respect to the dimensionality of the dataset. In [4], my collaborator and I show a few properties of distance-based outliers, arguing why they are rather natural semantically, and present a few algorithms for computing distance-based outliers embedded in large and high-dimensional datasets. From an applicability standpoint, a weakness of distance-based outliers is that they require the existence of distance functions that are meaningful. (As such, distance-based methods and depth-based methods are complementary.) I believe this is not a problem for many kinds of spatial and geographic data.

As pointed out earlier, finding outliers is often a very natural form of knowledge discovery for surveillance applications. In those applications, the user may already know what the norm is. What the user does not know, and can get help from knowledge discovery modules, are what forms the abnormalities manifest themselves, and when those abnormalities arise. I believe this is an area that has been largely overlooked, and can be very valuable for analyzing and understanding spatial data.

With respect to content-based analysis from images, my research revolves around three aspects: compression/extraction, indexing and querying. When dealing with multimedia data, compression is essential. Compared with compression, which can largely be considered as "syntactic", feature extraction is more "semantic". However, "a picture is worth a thousand words"; numerous features can be extracted from each image. Thus, there is the critical issue of feature selection. My work described in [5] is concerned with feature selection for indexing. That framework has been extended to deal with sub-image querying, whereby the user only gives specifications to part of the images.

I believe feature extraction for image indexing and querying is very different from feature extraction for knowledge discovery. In the former case, features are selected to maximize (in a loose sense) discriminating power - essentially for differentiation. In the latter case, the situation is less obvious. If we are looking for summaries, feature selection to maximize discriminating power appears to be counter-productive. But we do want summaries that do not necessarily

include every single image in the database; in other words, a certain amount of discriminating power will be needed. Where to draw the line, I believe, is a very hard question. If we are looking for outliers, the outcome also depends on what features are used. How to perform feature selection appropriate for outlier detection is a non-trivial question as well.

### *References*

- [1] Ng, R. and Han, J., "Efficient and Effective Clustering Methods for Spatial Data Mining", Proceedings of 20th International Conference on Very Large Data Bases, pp. 144-155, 1994.
- [2] Knorr, E. and Ng, R., "Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining", IEEE Trans. on Knowledge & Data Engineering, 8, 6, pp. 884-497, Dec. 1996.
- [3] Johnson, T., Kowk, I. and Ng, R., "Fast Computation of 2-Dimensional Depth Contours," Proceedings of 4th International Conference on Knowledge Discovery & Data Mining, pp. 224--228, August 1998.
- [4] Knorr, E. and Ng, R., "Algorithms for Mining Distance-based Outliers from Large Datasets," Proceedings of 24th International Conference on Very Large Data Bases, pp. 392--403, August 1998.
- [5] Ng, R. and Tam, D., "An Analysis of Multi-level Filtering for Image Databases", to appear in: IEEE Trans. Knowledge & Data Engineering (accepted December 1998).

### **3.13 Jonathan Raper**

At present, the creator/curator of any collection of 2D vector and raster graphics, static photographic imagery, surface models, solid models, audio, video, real time GPS data streams, panoramas, spreadsheets and text must marshal a whole series of applications to georeference, store, retrieve and analyse such data. Yet such collections will be found in most libraries of the future and many research data collections of the present. The lack of any integrated data store or cross-data type indexing system capable of handling georeferenced data types is currently a major limit to our ability to explore such collections for space-time patterns and coincidences.

There seem to be two distinct strategies for creating searchable multi-data type collections. Firstly, the expansion of traditional databases from relational to object-relational and the development of multidimensional query languages such as SQL3. The advantage of this 'universal server' approach is that the data store would be based on existing DBMS technology and that hybrid multidimensional queries could be based upon methods bound to each data type. However, this approach requires complex queries that are hard to optimise and the semantics of such queries across data types with georeferencing would not always make sense.

Alternatively, since multi-data type collections are classically what the originators of the hypertext concept envisaged, these collections could be structured by user defined portable pointer structures based upon text-based metadata associated with all the data types. The advantage of this approach is that the arrival of XML as a new standard for mark-up on the web and the flexible architecture of the next generation web browsers makes the storage and access to such data types more and more straightforward. If data types can be stored with user-entered

metadata in RDF/Dublin Core format, then hyperlink pointers can be computed by querying the metadata collection or created by browsing. Directed graphs of hyperlinks forming 'link sets' can be created and stored as access structures for a collection. The disadvantage is that such 'link sets' carry no guarantee of completeness and might not exhaust the data space.

Access structures based on 'link sets' would be more akin to tours or thematic collections, although multidimensional range queries over the georeferencing metadata fields would allow the geographic exploration of collections. Spatial agents could also be developed to 'roam' such collections looking for correspondences. I have a postgraduate student (Armanda Rodrigues) working on the implementation of intelligent spatial agents for spatial data access whose work is close to completion.

The potential of this second approach is currently being investigated using the Virtual Field Course hub and client test bed and some preliminary indications would be presented at the meeting.

### **3.14 Hanan Samet**

The proper treatment of spatial data requires more than just storing spatial data. It also demands proper treatment of nonspatial data such as that associated with the spatial object or location. For example roads have speed limits, regions have people living there with income levels, cities have populations, etc. All of this data, although inherently nonspatial, is said to be spatially-referenced in the sense that it is associated with locations in space. This data can also be referenced by name in the sense that aggregates of space have been given a name to facilitate reference to it.

Since spatial and nonspatial data are so intimately connected, it is not surprising that many of the issues that need to be addressed are in fact database issues. Some deal with performance while others deal with how to interface spatial algorithms with large databases. In our work, we have observed that the algorithms that we developed are responses to spatial queries. These queries can be classified into two principal classes. The first is location-based. In this case, we are searching for the nature of the feature associated with a particular location. For example, "what is the feature at location X?", "what is the nearest city to location X?", or "what is the nearest road to location X?" The second is feature-based. In this case, we are probing, in part, for the presence or absence of a feature based on another feature, as well as seeking its actual location. For example, "what type of vegetation or land cover is found within 10 miles of the Mississippi River?"

Dealing with location-based queries is quite straightforward using spatial data structures such as the quadtree as they involve descending the tree until finding the object. Dealing with feature-based queries is more difficult. The problem is with pure quadtree-like methods (i.e., based on variable resolution) is that there is no indexing by features. The indexing is only based on spatial occupancy. For example, if we want to find the locations where wheat is grown, we need to examine each location (or block) and test if it is indeed a site where wheat is grown. Our goal is to process the query without examining every location in space.

The pyramid is a regular decomposition representation based on multiple resolution that is useful for such queries since the nodes that are not at the maximum level of resolution (i.e., at the bottom level) contain summary information. The pyramid is best characterized as a space hierarchy. Thus we could view these nodes as feature vectors which indicate whether or not a feature is present at a higher level of resolution. Therefore, by examining the root of the pyramid

(i.e., the node that represents the entire image) we can quickly tell if a feature is present without having to examine every location.

There are numerous other representations for spatial data most prominent of which are those based on the use of bounding boxes (e.g., members of the R-tree family [Gutt84]). They are object hierarchies in the sense that objects are aggregated into groups of objects until there is just one object group left. The bounding boxes are used as filtering devices to speed up search. While these methods have been used extensively, they are not necessarily appropriate for all applications. In particular, these representations are designed to distinguish between occupied and unoccupied space. However, these representations, unlike the quadtree and the pyramid, are not based on a regular decomposition. Therefore, when we want to combine different data sets such as terrain, roads, and hydrography etc., representations that use bounding boxes are not very attractive as they are not in registration. The problem, as mentioned above, is that these representations are primarily designed for distinguishing between occupied and unoccupied space in a particular map rather than correlating occupied space in the different maps which serve as input to the operation. The result is that the computation is more complex.

The research that I am interested in conducting is the development of efficient representations to support spatial data mining and algorithms. Some additional research involves the incorporation of statistical information such as averages, means, ranges, etc. The pyramid is idea for the representation of such information. I am also interested in exploring further the difference between space and object hierarchies. I believe that we can learn quite a bit from the way these issues have been dealt with in computer graphics research where the distinction is analogous to the one that is made between image space and object space methods, respectively.

Other research interests involve knowledge discovery through the exploration of databases containing spatial and nonspatial data. My goal is to provide users primitives so that they can explore the data on their own rather than have explicit functions. As an example, we have developed a method for finding nearest neighbors incrementally. This is in contrast to finding the  $n$  nearest neighbors. An example of the utility of this operation which we call ranking is a query that seeks the nearest hotel to a particular amusement park where rooms cost less than \$20 per night. One strategy for obtaining the answer is to incrementally locate the nearest neighboring hotels of the amusement park until finding one with the appropriate nightly rate.

The alternative solution, finds the 10 nearest hotels and then checks their nightly rates and then sorts them by distance. If none of the hotels meet the conditions of our query, then we must restart the search from scratch and now look for the 20 nearest hotels, etc. Clearly, the ranking approach is superior as only the minimum number of neighbors must be obtained.

We have recently applied the ranking idea to compute clusters in a spatial database. In particular, suppose we are given a set of locations of nuclear facilities  $N$  and a set of locations of monitoring stations  $F$ . We wish to calculate the following:

1. For each nuclear facility, find the closest monitoring station.
2. For each monitoring station, find the closest nuclear facility.

The result of these two queries is very much like a Voronoi Diagram (also known as a Thiessen polygon). Executing such queries is not possible using existing database technology. First of all, it is difficult to express this query using SQL. Second, we have to be able to calculate it (i.e., we need an algorithm). A more common example of the same query is one where we are given the

locations of stores and warehouses and we wish to determine which warehouse should deliver to which store.

It turns out that these queries are really repeated instances of ranking where we want the closest (i.e., the element ranked number 1). In particular, in the case of query 1, we are ranking the monitoring stations with respect to the nuclear facilities, while in the case of query 2 we are ranking the nuclear facilities with respect to the monitoring stations. In technical database terminology, the ranking operation is really like a join of the nuclear facilities and the monitoring stations based on the values of the spatial (i.e., locational) attributes (also known as a spatial join).

Obtaining answers to queries 1 and 2 is a semi-join operation that just retains one value for each pair, which is the closest one. We term the general operation a rank join (or distance join) and the operation that retains the closest value a rank semi-join (or distance semi-join). This operation is pretty straightforward to execute when we have a ranking primitive. In fact, this operation was discovered by a user of the ranking primitive. We believe that there may be other operations that can be discovered through the appropriate use of built in spatial operations. We are interested in the further study of such techniques (e.g., finding shortest paths).

### **3.15 Shashi Shekhar**

#### *Clustering via Hypergraphs*

Clustering is the process of identifying regions of high density data points in a non-uniformly distributed multi-dimensional data space based upon a “distance” measure. In the spatial context the Euclidean distance is a natural measure of proximity but “distance” could also be related to the statistical notion of correlation.

A Hypergraph is a graph in which an edge can connect more than two nodes. There is a one-to-one correspondence between hypergraphs, binary matrices and bipartite graphs. We have recently reported a clustering based min-cut hypergraph-partitioning method for spatial join processing using a join-index [1]. Experiments performed on the Sequoia 2000 data sets show that our method outperforms methods based on sorting and other graph heuristics.

Recently, methods based on hypergraph partitioning have been successfully employed to cluster high-dimensional data [3]. These methods use the data mining concepts of frequent itemsets and high confidence to generate a weighted hypergraph and then use a partitioning heuristic to cluster the hypergraph.

We briefly describe how a weighted hypergraph can be constructed from a database of transactions [3]. Later on we will define the concept of spatial transaction and show how hypergraph partitioning can be used even when the data has a significant spatial component.

Given a database of transactions a data mining algorithm is used to generate association rules which satisfy the constraints of minimum support and high-confidence. This process in effect identifies transaction items which are highly correlated and statistically significant. For an overview of efficient mining algorithms see [4].

The output from the mining algorithm is used to generate the edges and weights of a hypergraph. The original set of items is mapped onto the nodes of the hypergraph and the frequent itemsets represent the edges of the hypergraph. Since an itemset can contain more than two items they can

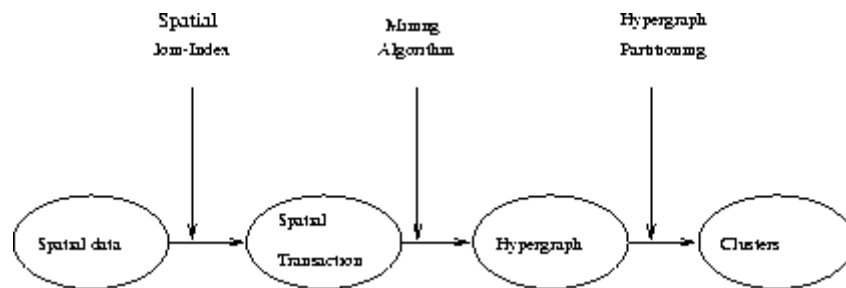
be neatly mapped onto a hypergraph edge. For the weights of the hypergraph the average confidence of the essential rules spanned by the itemset are used. Essential rules are rules in which all the items of the itemset are involved but in which the implication set of the rule, the right hand side, is a singleton. The hypergraph being generated, a partitioning algorithm can be used to cluster the graph.

### *Spatial Transactions*

How can we map spatial data and relationships into a weighted hypergraph? Spatial data are characterized by complex data types like points, lines and regions. Spatial relationships on the other hand depend upon the “granularity” of the underlying physical space: topological, directional, metric or Euclidean. We define the concept of a spatial transaction to capture this complexity.

A spatial transaction is defined in terms of a reference spatial object which acts as a unique transaction identifier(tid). The items of the transaction include spatial objects or characteristics of spatial objects which satisfy a user-defined spatial relationship with respect to the reference object.

For example consider the database of all counties in the United States. The name of the county and its spatial footprint will play the role of tid. The items which constitute the transaction could be symbols representing a range of number of hospitals and indoor shopping malls which lie within the county. For example if there are less than three hospitals in the county a symbol H0-3 could be reserved to denote this relationship. Then H0-3 will be an item in the transaction. Similarly if there are between 5 and 10 shopping malls then a symbol S5-10 could be used. Notice that the hospital and the shopping malls satisfy the topological relationship within with respect to the county. The items need not necessarily symbolize the cardinality of spatial objects. Spatial relationships between objects can also be represented as items. For example if the county courthouse lies to the North-West of the county airport then a symbol C-NW-A could be a symbol and thus an item in the transaction. Having constructed the database of spatial transactions we could then model it as a hypergraph and pass the hypergraph as an input to the clustering algorithm. It is interesting to note that because spatial data enjoys low updates, a spatial join-index would be an ideal candidate to construct the database of spatial transactions. The whole process is summarized in Figure 1.



**Figure 1: Flow diagram for clustering algorithm.**

### *References*

- [1] S.Shekhar, C.T. Lu, S.Ravada, S.Chawla Optimizing Join Index Based Spatial-Join Processing: A Graph Partitioning Approach. Symposium on Reliability in Distributed Software, October 1998.

- [2] S. Shekhar , D-R. Liu. A Connectivity-Clustered Access Method for Networks and Network Computation. IEEE TKDE, 9(1), 102-118, January 1997.
- [3] E. Han, G.Karypis, V.Kumar, B.Mobasher Hypergraph Based Clustering in High-Dimensional Sets: A summary of results. Bulletin of the Technical committee on Data Engineering, 21(1), March 1998.
- [4] R.Agrawal, R.Srikant. Fast Algorithms for mining association rules. Proceedings of the 20th VLDB Conference, Santiago, Chile, 487-499, 1994.

### **3.16 Monica Wachowicz**

Geographical Information Science has become a scientific body of knowledge built from a base in the domains of geographic visualisation, geographic information systems, and spatial analysis (with strong ties to exploratory data analysis efforts in statistics). The volume of basic and applied research carried out in each of these domains over the past decade is responsive to the changing problems and issues that concern government, industry, and society.

In the NCGIA Specialist Meeting, I would like to focus on the issues related to large environmental data sets since they represent a major challenge for both domain and information sciences. The domain sciences, most of which developed under data poor conditions, must now adapt to a world that is data rich - so data rich that large volumes of data often remain unexplored while the media they are stored upon deteriorate or become obsolete. The information sciences, most of which developed in a pre-computer era or when batch processing by computer was the norm, must now adapt to a world that is not only digital but highly dynamic - in which there is a potential for information systems to produce answers in real time as an analyst explores data and poses "what if" questions. It is in this context of both increasing data availability and rapidly evolving computing technologies that I see a substantial challenge for research in spatio-temporal methods. The challenge is two-fold: to extend geographic visualisation, GIS, and spatial analysis methods (currently developed for static data) into a spatio-temporal realm and to integrate these domains of Geographic Information Science to produce innovative methods (and associated tools) that facilitate policy decisions, assessment evaluations, planning and development strategies.

Key among these is the view of the overall challenge as a multi-step process oriented approach of manipulating data (which can include quantitative and qualitative information) to arrive at valid, novel, potentially useful, and ultimately understandable spatio-temporal patterns and processes in data. Geographic Visualisation (GVis) methods emphasise human visual thinking, graphical data manipulation, and human computer interaction. In contrast, GIS methods offer tools for developing domain specific models, query languages for data and meta-data manipulation, and data interoperability for distributed computer environments. From Exploratory Data Analysis (EDA), quantitative methods provide the statistical modelling used for developing data mining algorithms designed to produce sets of statements about local dependencies among variables from very large data sets. The strength of the development and integration of GVis, GIS, and EDA methods lie in the powerful integrated system that they can provide. A system from which to store, explore, and evaluate very large amounts of observational and model simulation data, and subsequently understand and communicate this understanding. Particularly when applied to environmental data, designing an integrated GVis- GIS-EDA system is definitely a non-trivial task.

Some envisaged capabilities are:

- a scheme for creating conceptual hierarchies of information (e.g. GeoVRML worlds that support varied levels of detail that react to a user's virtual distance from a location in the world);
- graphical interactive data exploratory tools such as linked and brushing techniques using space-time cubes, multivariate glyphs, and parallel coordinate fplots (e.g. Visage, SAGE, and SDM);
- flexible database protocols supporting spatio-temporal data types and operators that allow rule-based spatio-temporal reasoning (e.g. LDL++, extended with spatial and temporal construct suitable for the application domain of atmospheric sciences);
- proactive tools such as animation and hypermedia to direct database queries and mediate database navigation (e.g. tools developed for the Alexandria Digital Library Project),
- data mining tools for tracking and detecting evolving physical phenomena in very large data sets (e.g. CONQUEST and GeoMiner tools for clustering, feature extraction, and simulation);
- dynamic space-time modelling capabilities of object-oriented or object-relational databases (e.g. inheritance, polymorphism, and complex database objects and scenarios).

Having suggested some capabilities for an integrated GVis-GIS-EDA system, I have attempted to illustrate my working knowledge background as well as the interdisciplinary topics that can be discussed in NCGIA Specialist Meeting.

## 4 SUMMARY OF WORKSHOP PRESENTATIONS AND DISCUSSION

### 4.1 Thursday, March 18

#### 4.1.1 **Plenary I - Status and trends in geographic information science**

Presented by Jonathan Raper.

This plenary session reviewed a wide range of topics related to the worldview of geographic information science, the role of technology in generating structured, semi-structured and unstructured digital geographic data and the role of geographic information in generating geographic knowledge. Some of the key points of the presentation were:

- i) There is a sharp contrast between the worldviews of geographic information science and the attempts at non-scientific understanding and problem-solving in post-positivism and within political contexts. Specifically, GIScience involves a rigid view of processes (logic), space (geometry) and time (absolute). It is unclear that the scientific worldview can lead to useful knowledge for the non-scientific and cognitive-based perspectives on understanding the world as well as in non-scientific decision processes.
- ii) Information is conceived and generated from particular ontological and epistemological perspectives. Thus, a key question is "what is this information *about*?" when practicing GIScience and GKD. Also, intellectual property rights vary across nations. The institutional and community settings of information collection and use can confound GIScience and GKD.
- iii) Geographic data collection technologies are generating multi-media georeferenced databases that involve structured, semi-structured and unstructured information. Can we develop GKD techniques that can handle semi-structured and unstructured data as well as "mine" across multi-media?

#### 4.1.2 **Plenary II - Methods for geo-spatial data mining**

Presented by Jiawei Han.

This plenary session provided a survey of various methods for discovering knowledge in large geo-spatial databases from the point of view of spatial data mining. The survey covered the following major themes:

Motivation. Necessity is the mother of invention. The computerization of our society has led to the collection of huge amounts of spatial data. To facilitate efficient analysis of such data, new tools must be built. Spatial warehouse and spatial OLAP are tools and infrastructures to facilitate spatial database integration/consolidation and on-line analysis of multidimensional spatial data. In addition to the spatial data warehouse, spatial data mining tools for extracting interesting patterns and knowledge from spatial data should be systematically developed.

Methods for spatial data warehousing. A spatial data warehouse requires all the features of a nonspatial data warehouse plus spatial dimensions and spatial measures. The materialization of certain spatial measures, such as region merging, etc. may lead to the explosive growth of storage space requirements. Thus selective materialization of spatial data cubes and spatial measures and

methods for fast computation of spatial region merges, and other spatial operations applying to massive number of spatial objects need to be developed.

Integrated spatial OLAP and mining. An architecture called spatial OLAM (on-line analytical mining) systematically supports spatial data mining mechanisms in large spatial databases. A system prototype, GeoMiner, which implements such an architecture is introduced. Previous research has addressed methods and experimental results for implementation of such integrated OLAP and data mining techniques.

Overview of geo-spatial data mining methods. An overview of recent research and developments on geo-spatial data mining covered the following issues.

a) *Characterization and comparison of spatial objects*

Spatial objects that share similar characteristics can be grouped together and be summarized at various abstraction levels. Each group of spatial objects can also be compared to see the differences in their general characteristics.

b) *Spatial association*

Methods for association between spatial and non-spatial predicates can discover rules such as, "80% of gas stations outside of the city are near highway intersection." Methods exist for progressive refinement of data mining quality.

c) *Spatial classification and prediction*

Given a set of target classes, spatial classification select the most relevant set of attributes and attributes values which determine which target class a spatial object belongs to. Also, spatial classification can be used for classify some unknown spatial objects and predicting the class labels or certain unknown values.

d) *Spatial clustering and outlier analysis*

Spatial clustering group spatial objects into clusters so that objects in the same cluster are quite similar whereas objects in different clusters are rather different from each other. The five categories of clustering methods are partitioning algorithms, hierarchical algorithms, density-based algorithms, grid-based algorithms, and model-based algorithms. Spatial objects that do not belong to any cluster are called outliers; distance measures can be used to identify outliers and analyze their properties.

Future research topics in spatial data mining. These include:

a) Establishing a theoretical foundation for spatial data mining.

b) Performance improvements of methods for mining different kinds of spatial knowledge.

c) Integration of cartographic modeling and spatial data mining.

- d) Integration with existing spatial data analysis techniques.
- e) Mining spatio-temporal data, raster/image data, and Web data.
- f) Development of an integrated, intelligent geographic information systems.
- g) Development and exploration of spatial data mining applications.

#### **4.1.3 Panel: Tasks for spatial data mining in large geo-spatial databases**

Panel coordinator: Hans-Peter Kriegel. Panel members: Dick Muntz, Raymond Ng, Hanan Samet and Shashi Shekar.

Spatial trend detection. One form of geographic knowledge discovery is spatial trend detection. Spatial trends can consist of both global and local components. GKD can help reveal the factors that lead to local deviations from global or expected trends. A motivating example is the role of physical features in attracting unusual populations and activities, in particular, retirees to the Bavarian region of Germany. This leads to unusual local deviations in economic activities from the expected global trend. A thematic map example indicated the need for spatial data mining across heterogeneous spatial databases.

Spatial clustering and outlier detection. With respect to spatial cluster detection, the relevant questions are: i) given 1 cluster, what are the interesting properties of the cluster? ii) given  $n$  clusters, what do they have in common? and; iii) given 2 clusters, what are the “discriminating” properties on which the clusters are different? Methodologies for answering these questions include thematic map correlation, determining proximity relationship using reference maps and boundary shape matching.

Although outliers may be “noise” to one person, they can also be another person’s “signal.” This is particularly true in surveillance applications such as environment monitoring. Traditional statistical tests for outlier detection are univariate and distribution-based. More recent statistical methods are based on the concepts of “depths.” This method organizes data points as layers in the data space. Points in outer layers are more likely to be outliers. This approach is not computationally efficient in  $k$ -dimensional space for  $k \geq 4$ . An alternative methodology for spatial outlier detection is based on distances.

An example system that uses clustering is the Conquest Scientific Query Processing System. This is a data mining system that locates cyclic storms and their trajectories from weather and climate data. Derived associations among attributes and clustering techniques identify outliers that may indicate cyclonic activity. The system also has various normalization techniques and can generate calendar-oriented statistics.

Open issues for spatial clustering include:

- Capturing direction and connectivity relationships in addition to distance;
- Capturing other types of geographic space, including spherical and attributed space (i.e., geographic space with spatially-varying properties that affect interaction);
- Constraint-based clustering (i.e., clustering when interaction is restricted, e.g., travel across a river limited to bridges);

- Methods that search for voids among clusters.

Spatial querying in GKD. Efficient spatial querying is central to spatial data mining. Geographic knowledge can be directly acquired through efficient query support that allows formulation of complex queries. Relevant classes of queries include *queries about objects*, such as i) all objects that contain a given point or set of points; ii) all objects that have non-empty intersection with a given object; iii) all objects that have a partial boundary in common; iv) all objects that have a boundary in common; v) all objects that have any points in common; vi) all objects that contain a given object; and, vi) all objects that are included in a given object. Another class are *proximity queries*, such as: i) the nearest object; and, ii) objects within a given distance. Finally we consider *queries involving non-spatial attributes of objects*, such as "given a point, find the minimum enclosing object of a particular type."

Central to efficient query support are spatial data structures. Quadtrees and R-trees are best suited for location-based queries while pyramids are best suited for feature-based queries.

#### General research questions for geographic knowledge discovery

- How to analyze or mine semi-structured (map, vector data) or unstructured data (image raster data)?
- How to represent unstructured data for mining? (video data to 2D trajectories)
- How can feature extraction interact with knowledge discovery?
- What is the objective of feature extraction?
  - For content-based querying, these are features intended to maximize discriminating power.
  - For data-mining, it is not so obvious because maximizing discriminating power may destroy patterns.
- How to cope with the high dimension of image data?
- How to support the integrated mining of structured and unstructured spatial data?
- Building domain-specific geographic theory into spatial data mining as a means of "directing" the GKD process;
- Use of GKD in geographic research and decision support. What do geographers want to do that they couldn't do before?

#### **4.1.4 Breakout groups - Setting the GKD agenda**

The discussion topic for the breakout groups was the major issues and questions that should be discussed during the workshop. The following agenda items emerged from this discussion:

##### 1. GKD theoretical issues

- *A general theory* of data, information, knowledge for the geographic domain.
- *Use of geographic concept hierarchies in GKD* to handle the complexity and volume of geographic data by starting at general levels with coarse spatial units and use a refinement process as interested patterns are discovered. The major questions are: i) can we develop or identify a widely supported spatial concept hierarchy for different geographic domains, and; ii) can we develop a standard process for refining geographic concept hierarchies?
- The use of *existing geographic theory and domain knowledge to direct the GKD process*, for example, looking for interesting and unexpected patterns in the residuals between a postulated geographic process model and the empirical data

## 2. Representational issues

- *Representation of spatial objects*, including: i) maintaining the integrity of polygons and other spatial objects (e.g., interpreting a list of coordinates as an integral unit representing the polygon boundary); ii) decomposition of spatial objects into component entities; iii) regular vs. non-regular representations; iv) disjoint vs. non-disjoint spatial objects, and; v) spatial data indexing methods.
- *Capturing the full range of geographic relationships among spatial objects*, including distance, direction, connectivity, possibly mitigated by attributed geographic space.
- *Time in geographic knowledge discovery*, including: i) mismatch between time as a discrete, non-directional entity in databases and a continuous, unidirectional entity in the real world; ii) different types of time, including seasonal, cyclic, multi-level/resolution; iii) capturing attribute rates with respect to time; iv) time as process not just an attribute; v) general shortcomings of the relational data model for modeling temporal data, and; vi) different levels of granularity when recording temporal events.
- *Imprecision and uncertainty in geographic concepts and data*, including: i) fuzzy sets versus probability as the formal model, and; ii) representing spatial data with different levels of uncertainty and imprecision.
- *Impact of converting spatial attributes to non-spatial attributes* for mined results, including loss of continuity information and the correlation among the spatial dimensions.
- *Developing spatial data models* to support GKD, including: i) robust, effective and efficient models at the conceptual, logical and physical levels, and; ii) resolving semantic heterogeneity issues in geographic data
- *Developing an effective design for spatial data warehouses*, including developing selection criteria for warehousing spatial data.

## 3. Analytical issues

- *Frame independent spatial analysis*, in particular, spatial data mining methods that are not dependent on the map scale or spatial resolution of the data.
- *Generalization operators for spatial objects*, including abstraction operators, data modeling techniques, logical/semantic operators and methods for dealing with the *modifiable areal unit problem* (MAUP; in brief, statistical results being an artifact of arbitrary spatial reporting units such as census tracts).

## 4. Application issues

- *Human interaction/visualization* in all stages of GKD.
- *Determining benchmark GKD problems and applications* both with respect to a testbed for spatial data mining algorithms as well as illustrative applications where spatial data mining generates novel results or decisions.

## **4.2 Friday, March 19**

### **4.2.1 Plenary III - Geocomputational tools and GKD**

Presented by Mark Gahegan

GeoComputation has its origins in the 1970's. In the aftermath of the quantitative revolution, computers were seen as the tools/slaves of the quantitative geographer. This involved lots of messy FORTRAN code. The 1980's witnessed the GIS revolution. Geographers were forced to adopt the poor representational and analysis capabilities of GIS. In a sense, geographers became the slaves of the computer. GeoComputation is a conscious effort to explore the middle ground between geography and computer science from a doubly-informed perspective. It is about not compromising the geography, nor enforcing the use of unhelpful or simplistic representations. Its goal is to enrich geography with a toolbox of useful computational methods. GeoComputation attempts to enriching geography with a toolbox of useful computational methods.

Improvements in computing architecture & performance and progress in pattern recognition, search, classification, and function approximation have enabled GeoComputation. The former enables previously intractable problems to be addressed via deterministic means. The latter provides sophisticated solutions for a range of non-deterministic problems: e.g. data mining, knowledge discovery, inductive learning. Other supporting developments include advances in spatial statistics (e.g. K & L functions to test for spatial clustering, adaptive spatial filters), visualization methods for performing exploratory analysis and agent technology as an implementation mechanism.

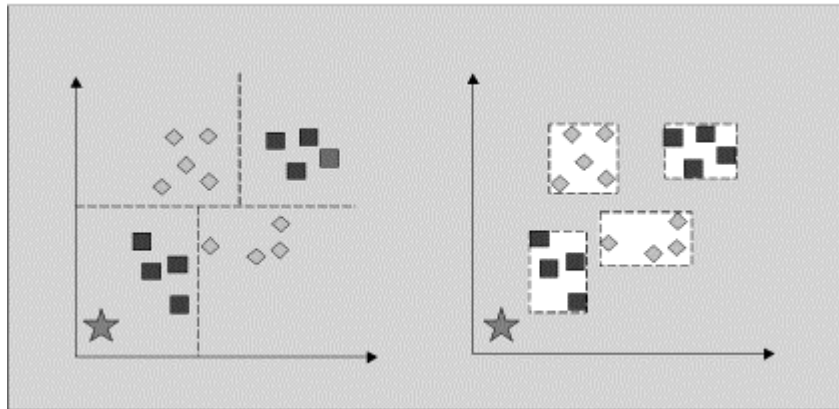
GeoComputation research challenges center on the knowledge gap between the theoretical working of the tools and their successful application to geographical analysis. These challenges are conceptual (how can the tools be applied effectively to geographic problems?) and methodological (sophisticated tools require sophisticated set-up and operation). Specific challenges include:

- How can we 'map' the geographic problem successfully to the search space used by the tools?
- How can we be certain that the tools examine all relevant portions of this space?
- How do we know when a good solution has been found, and can we report on statistical reliability?
- How do we 'mine' the solutions to provide insight or knowledge of the underlying geographic domain? The solutions provided are often highly parameterized with little obvious connection to the problem domain.

*Classification tools* provide illustrative examples of GeoComputational challenges and problems. Classification is useful as a form of data reduction, as a mapping to a different phenomenological domain and a means of recognizing and imposing structure. Classification can be viewed as a search of a hypothesis space defined from all possible classifiers. The hypothesis space is searched to locate the classifier that performs best. In supervised classification, performance is measured in terms of conflict with the training data. In unsupervised classification, performance is usually a function of cluster definition and separation.

Many of the newer analysis tools in computer vision / machine learning are types of classifiers that are scalable and can handle highly dimensional, noisy, sparse and non-Gaussian data. These include neural networks (most types), decision trees, Bayesian networks for supervised learning and k-means classifier and the self-organizing map for unsupervised learning.

Perhaps the most difficult problem to overcome when configuring geographical classifiers is *generalization*. Generalization is a form of bias imposed by the analyst when formulating general classes from the observed data. High generalization occurs when the entire data space is partitioned exhaustively into classes based on the training data (left half of figure below). Low generalization occurs when classes do not extend beyond the dataspace occupied by the training data (right half of figure below). This conservative approach creates "gaps" between the learned classes.



**Figure 1: High versus low generalization in classification**

While a trained inductive learning tool may contain knowledge, extracting this knowledge may be difficult due to the challenges outlined above. To meet these challenges, classification tools require the following properties:

- *Robustness*, or tools that will converge, learn and report reliably
- *Scalability* to very large input vectors, hypothesis spaces and datasets
- *Generalization* methods that treat the "spaces" between classes appropriately
- *Configuration* of architecture, problem setup and control from the data and not the user
- *Knowledge extraction* techniques that determine the knowledge obtained by the inductive learning technique

#### **4.2.2 Panel: Research frontiers in geocomputation and GKD**

Panel coordinator: John Roddick. Panelists: David Bennett, Heikki Mannila, and Elizabeth Wentz.

Lessons from KDD. Knowledge Discovery from Databases (KDD) is a strategy that does not slice the analytical process into subproblems prematurely. In other words, data mining is equivalent to exploratory data analysis. Data mining is also useful for decision support in applied problems.

Data mining and KDD tools for aspatial data include descriptive statistics, rule discovery, predictive modeling and probabilistic model. The criteria for these methods include *simplicity* (simple to use and produces easily understood results), *soundness and robustness* in a statistical sense and *scalability*. A challenge for GKD is a huge gap between the current, complex and non-scalable spatial statistical techniques and the data mining requirements for simple, robust and scalable techniques.

Additional lessons from KDD include the following:

- Interpretability is important
- "Interestingness" varies from user to user
- Deriving rules and patterns from datasets requires effective browsing and querying tools
- Techniques should recognize the difference between global models and local patterns
- Knowledge discovery should occur at multiple levels of analysis
- Data management is important
- Probabilistic modeling is useful
- Knowledge discovery cannot be automatic

Geocomputation and GKD. There are different perspectives on the linkages between geocomputation and GKD. Two useful metaphors are the *shell game* and the *big shovel*. The shell game perspective views data mining as processing a "mountain" of dirty, "useless" oysters to find a few precious pearls. In other words, we can attempt to automate the analysis of a mountain of geographical data to identify a few pearls of wisdom of geographic phenomena. This suggests that information has a high initial value and can be returned from the discovery process in near-final form. Conversely, the big shovel perspective views GKD as processing a "mountain" of low-grade material and transformation of this material into high-grade form. Therefore, we should automate the transformation of a mountain of geographical data into a small volume of information useful for the analysis. This assumes that the information has a low initial value and the product returned from mining process requires substantial value-added processing.

The latter perspective suggest that it is often necessary to apply a sequential set of complicated transformation to produce database that contain spatial and attribute information in an consistent and usable form. Methods for accomplishing this are common knowledge structures and intelligent agents. Intelligent agents can identify relevant databases and construct the sequence and transformation needed to construct new datasets that satisfy the user's request. These agents use expert knowledge and stored knowledge structure about available geographic datasets, different types of available transformation, data needs of specific geographic models, preconditions for applying the transformation and performance characteristic of tools, data and model.

GKD and GIS. Geocomputation and GKD also implies a rethinking of our conceptualizations of GIS. Currently, analysis processes are different between geography and computer science since GIS is viewed as a rigid sequential process rather than a flexible toolkit. To fit data mining fit into geographical analysis, we need a more flexible conceptualization of GIS. A re-conceptualization of GIS will dissolve the boxes around the traditional functions of *data input*, *data storage*, *data analysis* and *data reporting*. Instead, we should view GIS a toolkit whose techniques can be applied in a flexible manner tailored to the problem at hand.

GKD in geographic research. A question for geographers is "What do geographers want to do now that they couldn't do before?" Computational biology provides a useful analogue. Computational biologists have approached computer scientists and said: "We have this kind of data. We would like to find out xx (unspecified)." Can geographers form similar, useful questions?

At present, it appears that the geography can perform classification and visualization for simple phenomena. The geography community apparently cannot:

- Analyze moving patterns (single or pair)
- Visualize dynamic spatial pattern (moving rates in space through time)
- Detailed analysis of urban landuse dynamics
- Analyze interaction patterns at a detailed level of spatio-temporal resolution
- Handle high dimensionality in large-scale analyses
- Characterize uncertainty in geographic patterns
- Analyze space-time changes in rates and processes

Desirable properties of geocomputational techniques for GKD. Geocomputational techniques should have certain properties to be appropriate for GKD. These properties include:

- Simple, sound, scalable types of rules and patterns discovery techniques
- Multiple concepts of distance
- Efficient spatial database querying
- Incorporating geographic domain knowledge
- Easily interpreted

#### **4.2.3 Panel - GKD and domain-specific research**

Panel coordinator: Monica Wachowicz. Panelists: Kathleen Hornsby, Myke Gluck, Duane Marble.

Differences between data retrieval, information retrieval and geographic information retrieval. Data retrieval (DR) is deterministic and involves 'trivial' exact match or SQL-type retrieval. Information retrieval (IR) is indeterminate and open-ended. Examples include retrieving general information on the future Asian market for widgets, seeking patterns in data and exploratory data analysis (EDA). Multimedia resources require a rethinking of IR, including issues of algorithm development to match users' queries with available information, interpretability of results, query formulation methods and new performance criteria. Geographic information retrieval (GIR) involves indeterminate and open-ended queries from georeferenced data. These data are increasingly in multimedia form.

Supporting GIR requires improvements in metadata, retrieval performance measures and interactive multimedia tools. Metadata should describe how information could meet users' information needs rather than just address transfer or interoperability issues. Metadata should also be at the micro-level (associated with data entities rather than the entire database or thematic layer) to support integrated retrieval across heterogeneous databases. New retrieval success measures should reflect a redefinition of the retrieval problem from measuring the number of query matches to the facilitating the ability to "get close and look around" with feedback based on relevance of found items. Finally, the user should have multiple, interactive tools that can support integrated retrieval from heterogeneous multimedia databases.

Constructing knowledge from spatio-temporal data. Constructing knowledge from multivariate spatio-temporal data involves the following processes. First is *geographic visualization (GVis)* techniques that support human capabilities for visual thinking through human-computer interaction and interactive data manipulation. Second is a *knowledge discovery process* involving an iterative process that relies heavily on the analyst's background knowledge. This is closely related to exploratory data analysis. Third are *geographic information systems* with domain-specific spatial representations. Issues here include query languages for data and metadata and data interoperability.

A possible technique for exploring spatio-temporal data is *seriation*. "Seriation" refers to the process of putting objects into a series. This is a univariate scaling method that dates from the 1890's. Seriation can serve as an EDA tool for spatio-temporal in matrix form. The user can interactively permute, reorder or manipulate icons that reflect data values in the matrix cells. This can allow a type of interactive, visual principal components analysis. Augmented seriation techniques can handle multimedia data and simultaneous display of the data in map form in a manner similar to Monmonier's "geographical brushing" technique.

Shifting Granularities. Working with spatio-temporal data raises issues of *shifting granularities*. Distinct from map scale or spatial resolution, shifting granularities captures the world as perceived at different grain sizes or granules. Shifting granularities enables people to translate the complexities of the real world into simpler representations. Shifting to a more detailed view is also a common user requirement in geographic information systems and geographic analysis. It is important for making meaningful interpretations, effective decision-making, prediction or forecasting and historical views of spatio-temporal data.

There are two orthogonal perspectives on granularity. First is the granularity of *objects*. This relates to the complexity of objects. Object granularity allows the analyst to refine the view of objects. The analyst can reveal additional details about an object or coarsen the view of objects with the effect that the object might disappear all together. The second perspective is granularity of *time*. This relates to tracing objects over time. The analyst can expand the sequence of objects and events over time. The analyst can also collapse a sequence of events over time to a few events.

What do we need to accomplish GKD over shifting granularities? First, we need to identify knowledge rules and constraints for reasoning over objects and time independent of a particular domain. Second, we need to identify the operations that allow shifts in granularity. We should also think about extensions to this basic model, e.g., combinations of operations to capture complex granularity shifts.

GKD from spatial interaction data. A rich and under-analyzed data source that is directly of concern to geographers is spatial interaction data. Spatial interaction data is often recorded as flows between origin and destinations and are often published as origin/destination matrices. While analyzing flow matrices is a long-standing concern in geography, existing techniques are overwhelmed by the huge sizes of the matrices being published by entities such as the United States Bureau of Transportation Statistics, real-time data being collected through intelligent transportation systems and operational data being collected by utility companies. Visualization and browsing tools could discover the hidden knowledge within these rich flow matrices.

Some flow-related research questions that could be addressed better through GKD techniques include:

- Characterizing regional economies and interregional economic relationships at a global-scale
- Understanding the relationships between migration, demographics and socioeconomic characteristics at detailed levels of spatio-temporal resolution
- Understanding and predicting congestion propagation within urban transportation networks
- Capturing  $n$ -order cascade effects among flows (e.g., flow from A to B creating second-order flows from B to C, B to D, etc)

#### 4.2.4 Breakout groups - New tools for a geospatial data-rich environment

The focus of the discussions in the breakout groups was developing new analytical and computational tools for spatial analysis in data-rich environments. The following issues were raised:

1. Theoretical issues
  - Is it possible to formulate a *common framework for geocomputational techniques* in GKD?
  - Is it possible to formulate a *categorization/taxonomy for GKD problems*?
  - *What are the new questions that geographers (and others) would like to answer?*
2. Framework and interoperability issues
  - *Interoperability*, not standardization, between databases is required to support integrated spatial data mining and information retrieval across heterogeneous, multimedia databases.
  - While technical interoperability can be achieved, *semantic interpretability is more difficult*
3. Requirements for new techniques, including techniques for:
  - *Spatial summarization*
  - *Spatial/spatio-temporal data capture*
  - *Content-based retrieval*
  - *Exploring flows*, including: i) flows within transportation networks, and; ii) the relationship between flows and other geographic phenomena.
  - *Exploring spatio-temporal relationships and patterns*
  - *Methods for users to easily extract information from multimedia systems*
4. Need for GKD case studies
  - Generate examples to show the *special nature of spatial data*
  - What *spatial data mining problems can and cannot be accomplished* with tradition data mining techniques?
  - Create an *online library or website* to store the data mining software and research together.
5. User-orientation and user interface issues
  - *Conceptually simple models* are required to make GKD accessible to researchers with varying degrees of technical training
  - Methods for users to easily *extract meaningful information and recognize patterns* derived from GKD while *avoiding "information overload."* This could include: i) a language to indicate the interest and relevance of retrieved information; ii) methods for helping users to distinguish important information, and; iii) user-supplied constraints for reducing search spaces.
6. Database issues
  - *Well-indexed databases* may better than metadata
  - A more accessible *database user inference machine*. This can provide a first-cut information retrieval filter for the user

### **4.3 Saturday, March 20**

#### **4.3.1 Plenary IV - Geospatial data and data warehousing**

Presented by Yvan Bedard.

A data warehouse (DW) is a read-only enterprise-oriented, integrated, time-variant and non-volatile collection of data imported from heterogeneous source and stored at different levels of granularity to support decision-making. The DW is the basis for building the data rich environment used for knowledge discovery. DWs allow processing the heterogeneous data in the form of a seemingly homogeneous dataset. DWs also facilitate the efficient exploration of very large databases, but often in aggregated and summarized form.

A DW supports online analytical process (OLAP) but not online transaction processing (OLTP). OLTP systems are oriented toward the entering, storing, updating, integrating, checking, securing and simple querying of data. Normalized relational DB and most GIS application are examples of OLTP.

A DW is also not equivalent to a legacy database system. Some contrasts:

<b>Legacy</b>	<b>DW</b>
Built for transaction	Built for analysis decision
Original source	Copy or read-only data
Detailed data	Aggregated/summary data
Application-oriented	Enterprise-oriented
Current data only	Current & historic data
Normalized data structure	Denormalized, redundant data structure
Run on DBMS, GIS, web servers, CAD	Run on super-RDBMS, and multidimensional-DBMS

A DW can also be both physical and virtual:

<b>Physical DW</b>	<b>Virtual DW</b>
Persistent data	No persistent data
A priori integration	On the fly integration
Full data integration	Integrate as required
Require warehouse specific DBMS	Require no DW specific DBMS
Fast response	Slow response
OK for large DB	Not for large DB

A DW must also be distinguished from *datamarts* (DM). A DM is a localized and more aggregated implementation of a subset of the data in a DW. Some contrasts:

DW	DM
Built for analysis	Built for high-level analysis
Aggregated or summarized data	Highly aggregated or summarized data
Enterprise-oriented	Subject-oriented
Denormalized redundant data	Highly denormalized redundant data
Large DB	Smaller DB

DW should be build on top of existing systems instead of reengineering the entire database system. Supporting OLAP requires an efficient data structure (a highly denormalized and redundant multidimensional-type of structure) and an efficient indexing method. Since it is quasi-impossible to optimize a single system to satisfy both transaction-oriented and analysis-oriented operations, most DW designs are tiered systems that attempt that are a trade-off between both classes of operations. DW architectures include:

- *Corporate architectures* with a centralized integrated DW that receives cleaned data from heterogeneous, legacy OLTP systems and allows access from distributed clients
- *Three-tier architectures* with DMs between the centralized DW and the clients
- *Multi-tiered architectures* with a detailed DW and an aggregated DW, DMs and clients

Experience indicates that 80% of the KDD effort goes to building the DW.

Although many talk about GKD, a true *geographic data warehouse* (GDW) does not exist at present. GDWs present some unique challenges, particularly with respect to geographic data integration. Geographic data compatibility issues include:

- Identification differences (i.e., the geographic feature identified in the "real world")
- Difference in data formats
- Semantic difference (name vs. meaning)
- Domain differences
- Differences in measurement units, geometries, resolution and reference systems
- Differences in accuracy and precision

Possible solutions include cleaning, scrubbing and integrating tools oriented for geographic data, geographic data standards and interoperability protocols.

Another challenge for data integration for GDWs is temporal incompatibility. Temporal inconsistencies can exist between and within datasets. An example is database schema evolution in which definition of geographic object and attributes. New object geometry can also develop because of remeasurement in the real world. Although similar problems exist with mainstream DWs, these problems can be exacerbated when working with geographic data.

GDW research and development challenges include the following:

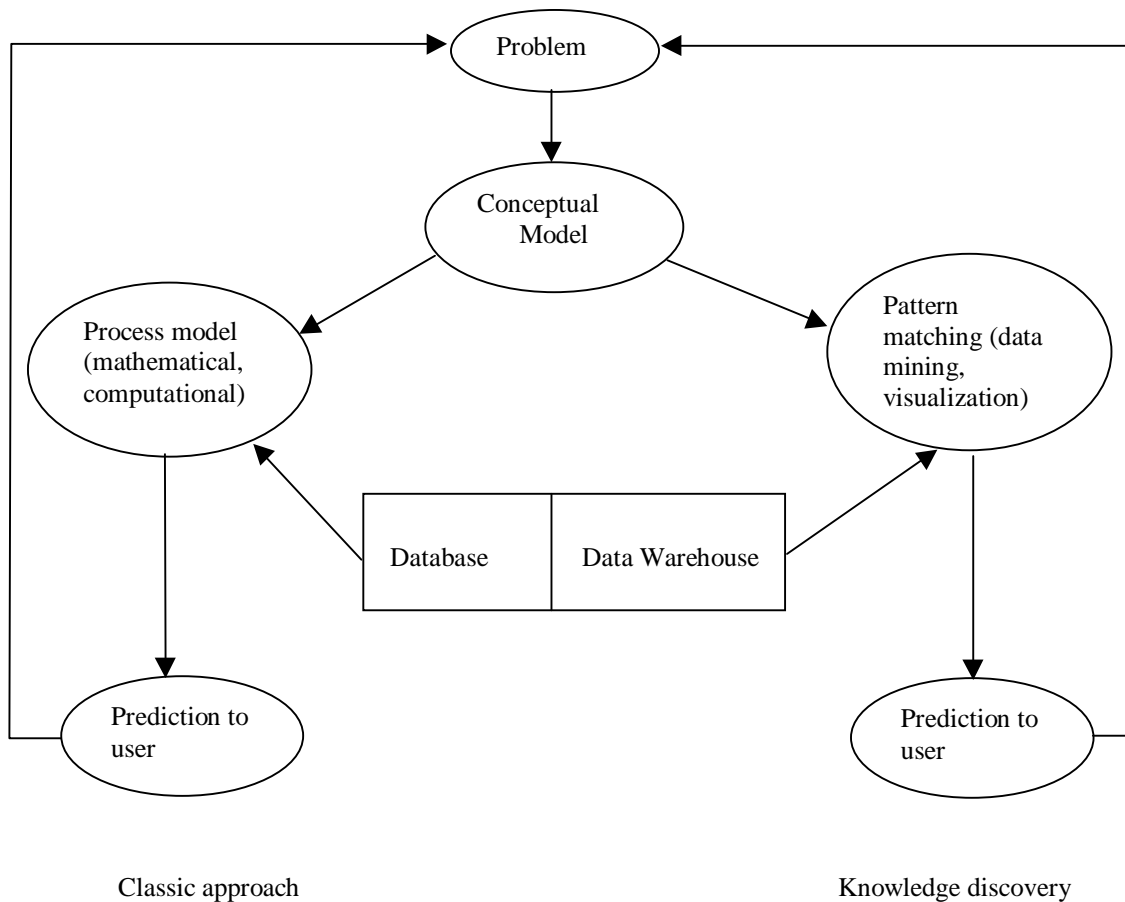
- Better methods and tools to feed the GDW
- Better use and management of metadata
- Better technique to manage very large databases
- Design new tools specifically for GKD, including spatial OLAP techniques
- Develop working prototypes as a proof-of-concept

### 4.3.2 Panel - Application problems and requirements

Panel coordinator: Brian Lees. Panelists: Brian Lees, Steve Smyth and Doug Flewelling.

What are the GKD research and application questions? The fundamental tenet in industry is profit, specifically, the cost of an enterprise must not exceed realizable returns. The knowledge should be unexpected, useful and cost-effective. The latter criterion can strongly affect the research questions asked in application contexts.

The following diagram provides a conceptual model that illustrates the KD process and compares it to the "classic" research approach. The diagram suggests the central role of *domain theory* in data mining.



Example application: Weather and climate data. Weather and climate data has interesting properties that make it a compelling challenge for spatial data mining. Weather is a volatile (sometime dangerous) phenomenon that has universal interest both casually and for economic and market decisions. Weather datasets are very large and of moderate dimensionality (typically consisting of the following attributes: ID, location, time, pressure (at multiple elevations) wind speed, wind direction, humidity, precipitation, snowfall). Weather datasets can also be unstructured (e.g., imagery).

Weather forecasting can be the result of modeling or based on historical records. The former mode requires a form of hypothesis testing based on the accuracy of forecasted weather variables. The latter requires a type of "similarity measure" based on the synoptic qualities of the historical record. This is traditionally a human-centered process based on the intuition and experience of the forecaster. Tools that can support knowledge discovery from historical weather data include visualization, summarization and generalization/abstraction.

Example application: Individual trajectories in space and time. GIS is going mobile. GIS can provide the basis for mobile locational information provision to individuals through cell phones, personal digital assistants or other hand-held devices. Mobile GIS-based information services can enhance the awareness of a person's locality (e.g., where am I? What are the road conditions ahead? What is the best route?). They can also help satisfy needs and wants (e.g., Send help! Where can I get food, fuel, money, etc.?) Mobile GIS services can also facilitate social interactions by dynamically locating and scheduling meetings among individuals.

There is a large potential market for these services. The technology for mobile GIS is becoming rapidly available through breakthroughs in miniaturization, exponentially increasing processing power and improvements in wireless connectivity. Social trends are also leading towards greater acceptance and use of mobile GIS. These include greater acceptance of E-commerce, integration of web-based information and interaction into lifestyles, increases in personal mobility, greater amounts of time spent travelling and the increasing decentralization of organizations. Types of businesses interest in mobile GIS include advertising, "yellow page"-type directory providers, system and application manufacturing, portable geographic content and service providers.

In exchange for information, mobile GIS providers will require users to accept tracking of their virtual and physical activities in space and time. This will generate an ongoing flow of data describing properties of location and travel in the course of human activities. This tracking data flow can be analyzed for information on the planning and execution of human activities. Information can also be fed back to humans who can use it to support desired activities and needs. This will require resolving some difficult privacy issues before these data collection methods are full accepted.

The enormous amount of space-time trajectory data that will be generated through mobile GIS creates tremendous opportunities for data mining techniques to discover knowledge about individuals' activities in time and space. In addition to data mining and visualization techniques, there will also be a need for "alloying" or integrating additional information (such as socio-economic characteristics) to manufacture information. A significant challenge is to build a supporting spatio-temporal framework. Also, there may be a need to modify our current data warehouse configurations to handle events such as potential data overflow.

#### **4.3.3 Breakout groups - GKD research and application frontiers**

The discussion that emerged from these breakout groups focused mostly on specific representation issues, technical issues and application questions at the frontiers of GKD research

##### **1. Representational issues**

- *Levels of granularity* for representing spatio-temporal process and spatial objects

- *Containerization and encapsulation* in tracking space-time trajectories, including when you can stop tracking a space-time trajectory (i.e., when does a pattern become noticeable)?
- Logical and physical structures for *representing flows*
- *Object emergence in dynamic environments*, i.e., when does an "object" emerge from a dynamic field?

## 2. Techniques

- More appropriate *query languages* and *optimization techniques* to support GKD
- *Constraint-based data mining*, in particular, domain knowledge-based constraints
- Developing an *ontology of data mining tasks*
- *Extraction and mining rules vs. mining data*
- *Visualization methods* and *decomposition of visual patterns*
- *Geographic domain knowledge* required to make sense of discovered rules

## 3. Example domain-specific questions for GKD

- Relationships between *individual activity patterns* and *urban fabric* at detailed level of spatio-temporal resolution
- Relationships between *individual activity patterns* and *aggregate flows* in transportation systems
- Relationships between *individual activity patterns* in *physical, virtual and hybrid spaces*
- Interactions between *urbanization* and the *physical environment* at *micro and macro scales*
- *Inferring geographic processes* from changes in geographic patterns
- Associating *spatio-economic behavior at a micro-level to macro-level spatio-economic dynamics*
- *Spatio-temporal patterns in interacting objects*

### 4.3.4 Synthesis

(Harvey Miller, Jiawei Han and Max Egenhofer). Several transcendental, cross-cutting issues in GKD emerged during the three-day workshop. These issues are:

1. What are the new questions for geographic research? A fundamental question for the geographic and related research communities is “What questions do we want to answer that we could not answer previously?” These communities need to form well-structured (but possibly open-ended questions) in order to guide the computer science and related communities in their tool and algorithm development
2. Better spatio-temporal representations in GKD. Current GKD techniques use very simple representations of geographic objects and geographic relationships, for example, point objects and Euclidean distances. Other geographic objects (including lines, polygons and more complex objects) and geographic relationships (including non-Euclidean distances, direction, connectivity, attributed geographic space such as terrain and constrained interaction structures such as networks) should be captured by GKD techniques. Time needs to be more completely integrated into geographic representations and relationships. This includes a full range of conceptual, logical and physical models of spatio-temporal objects. Finally, we need to capture multiple representations (in particular, robust geographic concept hierarchies)

and granularities in spatio-temporal representation in order to manage the complexity of GKD.

3. GKD using richer geographic data types. Geographic datasets are rapidly moving beyond the well-structured vector and raster formats to include semi-structured and unstructured data, in particular, georeferenced multimedia. GKD techniques should be developed that can handle these heterogeneous datasets.
4. User interfaces for GKD. GKD needs to move beyond the technically-oriented researcher to the broader geographic and related research communities. This requires interfaces and tools that can aid these diverse researchers in the GKD process. These interfaces and tools will likely be based on useful metaphors that can guide the search for geographic knowledge and make sense of discovered geographic knowledge.
5. Proof of concepts and benchmarking problems for GKD. There is a strong need for some “examples” or “test cases” to illustrate the usefulness of GKD. This includes a demonstration of GKD techniques leading to new, unexpected knowledge in key geographic research domains. Also important is benchmarking to determine the effects of varying data quality on discovering geographic knowledge. A related issue is research and demonstration projects that illustrate the usefulness of GKD techniques in forecasting and decision support for the public and private sectors.
6. Building discovered geographic knowledge into GIS. Current GIS software uses simple representations of geographic knowledge. Discovered geographic knowledge should be integrated into GIS, possibly through inductive geographic databases or online analytical processing (OLAP)-based GIS interfaces.
- Developing and supporting geographic data warehouses. A glaring omission from current research in GKD techniques is the development and supporting infrastructure for *geographic data warehouses* (GDW). To date, a true GDW does not exist. This is alarming since data warehouses are central to the knowledge discovery process. Creating true GDWs requires solving issues in geographic and temporal data compatibility, including differences in semantics, referencing systems, geometry, accuracy and precision. Supporting GDWs may also require restructuring of transaction-oriented databases systems Supporting GDWs may also require restructuring of transaction-oriented databases systems, particularly for flow and interaction data.

## 5 SEED GRANT PROPOSALS

The Project Varenus Executive Committee, based on the recommendations of the GKD workshop co-leaders and steering committee, funded two seed grant proposals that were submitted after the completion of the workshop.

### **5.1 Ontologies for Spatial Data Mining and Geographic Knowledge Discovery in Large Multimedia Spatial Datasets**

*Principal Investigators:* Jonathan Raper and Myke Gluck

The purpose for requesting this Varenus seed grant is to provide travel and support for the development of a grant proposal to NSF, Bureau of Labor Statistics, EU Fifth Framework research funding in the Information Society Technologies Programme (open to US collaborators), or other funding agencies interested in spatial data mining and spatial content-based retrieval. The full proposal would fund collaboration between Dr. Myke Gluck of Florida State University, an information scientist, and Jonathan Raper of City University, London, a geographer and GIS expert, to develop metaphors, ontologies, and practical interfaces to assist researchers in selecting methods and interpreting results of geographic knowledge discovery in large spatial multimedia datasets.

The Varenus specialists meeting entitled Discovering Geographic Knowledge in Data-Rich Environments held in Seattle developed a rich research agenda that included several components aligned with parts of our previous individual work and bode well for our collaboration. A major issue in the agenda is the need to address the meaningfulness of spatial data mining and in content-based retrieval activities. That is, algorithms and methods are important but how best to use them and under what scenarios they are most useful are rarely clear. In some sense methods such as clustering or association techniques are answers in search of questions. Related to the issue of meaningfulness is the ability to set thresholds and establish useful benchmarks for data mining/content-based retrieval. The computer scientists and statisticians have left the parameterization or threshold settings to the user to declare but few users know a priori what are significant thresholds to set. Thus, the computer scientists reframe the data mining/content-based retrieval problem from finding a method to establishing a threshold or set of weights. Such reframing unfortunately does little for the user of the methods who has expertise in the data but not in the methods or those who have expertise in the methods but little experience in the content domain.

Each of us have separately investigated geographic information retrieval, exploratory data analysis and spatial data mining with multimedia (e.g. Raper 1996). We have found that numerical data must be augmented with other related data such as video and cartographic displays to assist mining tasks to expose meaningful patterns. For example, Gluck et al (1999; Gluck and McCraw, 1997) have used seriation as a mining technique augmented by interactive data visualization and cartographic displays. He is now also pursuing further augmentation with sonification to seek patterns in data that are significant in geographic information retrieval and spatial data mining. He has also begun to explore the role of interface paradigms, interface metaphors, ontologies and meta-ontologies to support data mining.

By contrast Raper has focused on the development of spatial hypermedia databases with hypertext link structures capable of content based querying (Raper and Livingstone 1995). In the Virtual Field Course project he has been exploring the possibility of computed hyperlinks between multimedia spatial resources using spatial criteria. He has also explored the use of spatial intelligent agents to reason about

computable spatial representations stored on networks to improve searching in spatial digital libraries. Recent work has explored the use of virtual environments as interfaces to knowledge exploration.

For the full proposal the research will focus on the deep ontological and epistemological aspects of discovering geographic knowledge; however, we also intend to design interfaces and/or build systems to make manifest the application of these constructs to real world problems. Studies of users interacting with such systems would be integral to our developing deeper understanding of the role the user may play in data mining/content-based retrieval. Most systems provide tools with little guidance for users to employ. We believe discovering geographic knowledge is a process and although it may vary among individuals there are general epistemological, ontological, and cognitive sets of structures that can and must be embedded in systems to assist users. We also believe that spatial data mining and content-based geographic information retrieval have much in common, and that understanding how users seek geographic information impacts how they conceptualize the discovery of geographic knowledge.

Our research will center on observing, cataloging, and organizing these user processes. These processes can then be translated into meta statements on geographic conjunctions and/or patterns which can then be converted into computable statements capable of being used to mine/retrieve data in spatial databases. Raper is part of the Virtual Field Course project which has developed a software architecture where geographic information processing software components can communicate with rich metadata servers (see <http://www.geog.le.ac.uk/vfc/>) permitting the development of prototype tools to capture processes and instantiate meta statements usable for mining/content-based retrieval. Gluck runs a usability testing laboratory that can rapidly test and study how real users interact with systems and data (<http://www.fsu.edu/~lis/usability/index.html>).

A major issue of the research agenda that summarized the Specialists meeting was the need for benchmarks or case studies that illustrate the data mining/content-based retrieval procedures, tools, and effectiveness. As part of the proposal to the funding agencies we would include applications of the developed system in diverse domains such as risk assessment, commodity flows, and digital spatial libraries. As part of the seed grant we would create a rapid prototype to illustrate the central theoretical issues and some tentative solutions to be incorporated into the proposal to funding agencies. We will target funding initiatives such as the Information Ecology call of the EU Information Society Technologies Programme (see <http://www.cordis.lu/ist/fetuie.htm>).

Furthermore, during the Specialists meeting Gluck and Raper began a dialogue addressing how we might empower users to effectively and efficiently employ multiple data mining/content-based retrieval methods in one tool. Such a system would provide a range of current techniques as well as be extensible to incorporate additional methods as they are developed. The result would be a system with multiple methods selected sequentially or in parallel by the users to rapidly try various methods. The system would ease the cumbersomeness of tool selection, invocation, and inform users of each methods limitations. Rapid multimedia visualization of data results and mechanisms to store and replay the history of user actions would, we believe, greatly add to the willingness of users to explore data and to enhance their ability to communicate their results to colleagues and to a broader audience (policy makers, managers, etc.).

## **5.2 Multi-Scale Tools for Modeling Flows in Geographic Databases**

*Principal Investigator:* Eric D. Kolaczyk

Project Description. One of the primary agenda items that emerged from the recent Varenius workshop on “Discovering Geographic Knowledge in Data Rich Environments” was the importance of *flows* to the study of current and anticipated questions in geography, and the need for tools by which to elicit relevant information on flows from geographic databases. In addition, there arose from a number of discussions at the workshop a consensus as to the importance of multi-scale contexts and tools in aiding database users to extract information in a step-wise fashion, of an increasingly localized and detailed nature at each iteration. The PI proposes to conduct preliminary work adapting certain recently developed partition-based, multi-scale statistical models to the analysis of flow data. The work throughout is to be carried out in collaboration with two members of Boston University’s Department of Geography, Professors Sucharita Gopal and Ray Dezzani. Funding is requested to help in providing summer support for one Research Assistant to work with the PI on this project.

It is anticipated that this initial investigation will proceed in two stages, the first involving model development and the second focusing on implementation and testing of the resulting methods. Throughout these two stages we plan to use as our geographical context and preliminary test case a problem in transportation analysis, as described below. In the first stage of model development, the PI’s recent work on hierarchical, partition-based multi-scale models (e.g., Kolaczyk 1999), originally done in the context of standard time series and image analysis problems, will be extended to accommodate the unique nature and indexing of flow data, with its spatially complex “start-point/end-point” structure. These models use factorizations of statistical likelihoods to effectively de-couple the information in the data at a given location across various scales. These factorizations result from a hierarchically defined set of underlying partitionings of the dataspace (which could encode relevant geographic information, in the case of flows), and are crucial in producing tools that are not only mathematically tractable but computationally feasible as well. This latter point is of central importance for the second stage of this project, in which a prototype software implementation of the relevant modeling framework will be carried out. In analogy to the preceding methodology of Kolaczyk (1999), the resulting tools will allow for adaptive response to user queries regarding flow data, in a manner that is sensitive to both spatial and scale variations.

The particular geographical focus accompanying the above-described work is a problem of trip-generation estimation, in which data are collected at two specific scales. Conventional transportation planning analysis models begin with a predetermined traffic analysis zone (TAZ) system. Trip generation data is available at the TAZ scale. Additionally, census block data is often sampled at a larger scale than the TAZ data, and yet this data is often used to infer household trip generation behavior. The dataset consists of the 1994/95 METRO survey of Portland, Oregon and census data relating to the same region. The proposed framework will be used to predict traffic flows in a meaningful manner across a range of aggregations, incorporating information from census data.

Relation to Varenius Research Initiative and Expected Impact. The work proposed herein speaks simultaneously to two of the research initiatives that emerged from the Varenius workshop on “Discovering Geographic Knowledge in Data Rich Environments”: (1) the need for tools in modeling and extracting information from flows in geographic databases, and (2) the potential of multi-scale database tools to improve the ability of users to “zoom in” on knowledge of interest, proceeding from coarse levels of detail to finer levels. Due to the prevalence of flows in problems across the geographical information sciences, the developments proposed herein have the potential for a broad-based impact. At the same time, however, the resulting tools are likely to integrate smoothly with existing and future knowledge discovery systems. For example, the use of hierarchical partition-based structures is a natural way of summarizing information in

many settings, and its use in the proposed modeling strategy to create multi-scale structures parallels recent trends in the use of multi-scale contexts in the spatial database literature (e.g., Rigaux and Scholl, 1995).

The above described project will serve as the first collaborative venture between the PI and members of Boston University's Department of Geography, and it is anticipated that the results of this project would provide the foundation for a more detailed grant proposal(s), with relevance to agencies ranging from the National Science Foundation, to the Bureau of Transportation, to the Environmental Protection Agency.