Specialist Meeting on Spatial Data Analysis Software Tools

Upham Hotel, Santa Barbara, CA
May 10–11, 2002

The Center for Spatially Integrated Social Science (CSISS) is a five-year project funded by the National Science Foundation under its program of support for infrastructure in the social and behavioral sciences. CSISS promotes an integrated approach to social science research that recognizes the importance of location, space, spatiality and place.

One of the CSISS programs is devoted to "**Spatial Analytic Tools**" for the social sciences, that is, the development and dissemination of a powerful and easy to use suite of software for spatial data analysis, the advancement of methods of statistical analysis to account for spatial effects, and the integration of these developments with GIS capabilities. For a more detailed description of the programs and objectives of CSISS, visit our homepage at http://www.csiss.org/.

In order to take stock of the state of the art, assess current impediments and identify promising strategies, a two-day "Specialist Meeting on Spatial Data Analysis Software Tools" will be convened in Santa Barbara, CA, May 10th and 11th, 2002. The meeting is organized by a steering committee, co-chaired by Luc Anselin (University of Illinois, CSISS) and Sergio Rey (San Diego State University) and consisting of Richard Berk (UCLA), Ayse Can (Fannie Mae Foundation), Di Cook (Iowa State University), Mark Gahegan (Pennsylvania State University), and Geoffrey Jacquez (BioMedware).

The meeting will bring together software developers from both the public/academic sector as well as the private sector who deal with tools to visualize spatial data (geovisualization), carry out exploratory spatial data analysis (ESDA) and facilitate spatial modeling (spatial regression modeling, spatial econometrics, geostatistics), with a special focus on the potential for social science applications. These tools include a range of different approaches, such as macros and scripts for commercial statistical packages or GISes, modules developed in open source statistical and mathematical toolkits, and free standing software programs. The focus of the meeting is on software "tools" rather than on the methods per se.

The **objectives** of the meeting are threefold:

- It is an opportunity to demonstrate, showcase, and benchmark state-of-the-art tools and to interact with other specialized developers.

- It will facilitate and promote a dialogue among the wide range of developers about priorities and guidelines for software design, data and model standards, inter-operability, and open environments. It is hoped that this will initiate a discussion of specific open source standards for spatial data analysis.

- The meeting will also serve as a way to introduce CSISS' open source software development initiative, the "OpenSpace" project, and serve as a forum to obtain feedback and comments.

Contributions are invited that:

- describe technical aspects, architecture, design, principles and implementation of specific software tools for spatial data analysis

- compare and review software tools for spatial data analysis in social science applications

- demonstrate the application of new spatial analysis software tools to social science research questions.

All participants are expected to submit an abstract as well as a final paper. The papers will be published on a CD-Rom as a Proceedings Volume, available at the time of the meeting. The meeting will not consist of these paper presentations, but instead the Proceedings are to provide a common background for discussions related to the broader themes.

People interested in attending the meeting should submit a digital abstract for their contribution by e-mail to anselin@uiuc.edu by February 15, 2002. The steering committee will make a decision on the final list of participants by March 1, 2002. The abstract should be two to four pages (including figures, tables and references), in Adobe Acrobat pdf format (10pt Times Roman smallest font). The final paper (for the Proceedings Volume) should be 10 to 15 pages Adobe pdf and will be due by April 15, 2002.

Some funding assistance may be available, subject to NSF rules and prior agreement from CSISS. Please indicate if you require funding to participate.

Position Papers from the CSISS Specialist Meeting on Spatial Tools

Meeting participants were requested to share position papers consistent with the objectives for the meeting. These papers are attached in alphabetical order by authors' last names:

Implementing spatial data analysis software tools in R*

Roger Bivand
Economic Geography Section, Department of Economics,
Norwegian School of Economics and Business Administration, Bergen, Norway
Roger.Bivand@nhh.no

15th February 2002

1 Introduction

This contribution has two equal threads: doing spatial data analysis in the R project and environment, and learning from the R project about how an analytic and infrastructural open source community has achieved critical mass to enable mutually beneficial sharing of knowledge and tools. The challenge is to see whether, and if so how far, we can contribute to the next meeting of the community nurturing R (and other projects) at the Distributed Statistical Computing workshop in 2003. It is fair to say that the statistical and data analytic interests of the community are catholic, rigourous, and enthusiastic, and challenge the perceived barriers between commercial and open source software in the interests of better, more timely, and more professional analysis in the proper sense of the word.

R is an implementation of the S language, as is S-Plus, and often able to execute the same interpreted code; it was initially written by Ross Ihaka and Robert Gentleman (1996). R follows most of the Brown and Blue Books (Becker, Chambers and Wilks, 1988, Chambers and Hastie 1992), and also implements parts of the Green Book (Chambers, 1998). R is associated with the Omegahat project: it is here that much progress on inter-operation is being made, for instance embedding R in Perl, Python, Java, PostgreSQL or Gnumeric. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs out of the box on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux). It also compiles and runs on Windows 9x/NT/2000 and MacOS.

In this abstract, focus is on the second thread, with the first one accessible through the links from the text and the references to reviews elsewhere.

*Abstract of proposed contribution to CSISS specialist meeting on spatial data analysis software tools, Santa Barbara CA, 10-11 May 2002.

2 Spatial data analysis in R: status

One of the most effective and conscientious contributors to the R project is Brian Ripley, who is not only very familiar with spatial statistics as an academic statistician (Ripley, 1981 among other publications), but also contributed an early package to R for point pattern analysis and continuous surface analysis, associated with Venables and Ripley (1999 - third edition). Descriptions of some of the packages available are given in notes in R News (Ripley, 2001, Bivand, 2001b), while a more dated survey was made by Bivand and Gebhardt (2000).

At the time of writing, searching the R site for "spatial" yielded 381 hits. As Ripley (2001) comments, some of the hesitancy that was previously observable in contributions of new packages coming forward has been due to the existence of the S-Plus SpatialStats module: duplicating existing work (including GIS integration) has not seemed fruitful. Over recent months, a number of packages have been released on CRAN in all three areas of spatial data analysis (point patterns, continuous surfaces, and lattice data).

3 What R offers as a programming environment and a project

Above, CRAN (Comprehensive R Archive Network) and packages were mentioned. While R provides a rich language and environment for data analysis and visualization, it is also extendible, not just because the user can write new or customised interpreted functions, and dynamically load compiled C, Fortran or C++ code, but because the project provides tools for checking, building, archiving and distributed user-contributed modules known as packages (like CTAN and CPAN for instance). Each such package is required to document functions, to provide examples which should run without error if the package is correctly installed, and optionally supply test data sets (now including remote online test data sets).

This effectively reduces or removes the barrier between users (with a certain insight into the language and their own problem areas) and core developers, and seems to be a good example of the beneficial consequences of an open source development model. It has been important to maintain a certain conservatism, meaning that hard-won experience (and legacy C and Fortran code) is central, while experimentation continues in parallel, and in part in the Omegahat project. It is also worth stressing that the R project is an open community, with multiple commitments to varying data analysis communities, and a clear willingness to adapt within the possibilities offered by open source development, in particular through inter-operation with other visualization software, databases, languages, and so on (even including R as an Excel Addin).

4 Opportunities for advancing spatial data analysis in R

While a good deal is already going on, there are some clear gaps that need to be filled, over and above making more modern spatial data analysis tools and knowledge available. One is the wish that Ross Ihaka made after the last DSC meeting (at which I had talked about GIS integration, Bivand, 2001a) for mapping capability in R. There is some code around, including topology code, other libraries are available (particularly from Frank Warnerdam's work), and all the current spatial data analysis packages try to solve visualization problems in their own ways. GRASS is also moving to positions from which the use of vector libraries is likely to be possible, also GPL and written in C.

A further area is that of inter-operation, using XML and/or Green Book connections methods, or simple programs writing programs. This could also involve plugging R data computation services into other front ends, say R in PostGIS given that R can already be embedded (experimentally) in PostgreSQL. This is more speculative, but Omegahat seems to be progressing vigourously, and highlights inter-system interfaces. It would however build on any pre-existing spatial data analysis functions in R, which would become available in the environment within which R is embedded if so selected.

References

R. A. Becker, J. M. Chambers, and A. R. Wilks. 1998. *The New S Language*. Chapman & Hall, London.

R. S. Bivand. 2001a. R and geographical information systems, especially GRASS, Proceedings of the 2nd International Workshop on Distributed Statistical Computing, Technische Universität Wien, Vienna, Austria.

R. S. Bivand. 2001b. More on Spatial Data Analysis, *R News*, 1 (3) 13-17.

R. S. Bivand and A. Gebhardt. 2000. Implementing functions for spatial statistical analysis using the R language, *Journal of Geographical Systems*, 2 (3) 307-317.

J. M. Chambers. 1998. *Programming with Data*. Springer, New York.

J. M. Chambers and T. J. Hastie. 1992. *Statistical Models in S*. Chapman & Hall, London.

R. Ihaka and R. Gentleman. 1996. R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics*, 5, 299-314.

B. D. Ripley. 1981 *Spatial statistics.* Wiley, New York.

B. D. Ripley. 2001. Spatial Statistics in R, *R News*, 1 (2) 14-15.

W. N. Venables and B. D. Ripley. 1999 *Modern Applied Statistics with S-Plus.* Springer, New York (book website).

Interactive Linked Micromap Plots And Dynamically Conditioned Choropleth Maps

By Daniel B. Carr[1,2], Jim Chen[1], Sue Bell[2], Linda Pickle[2]
Affiliations: George Mason University[1], National Cancer Institute[2]

Extended Abstract:
This paper describes two recently developed templates for displaying geospatially-indexed estimates: linked micromap (LM) plots and conditioned choropleth (CC) maps. Two common goals in developing these templates were to integrate more statistical information in a display than a traditional choropleth map and to provide for more rapid assessment of statistical and spatial patterns than would be provided by a table. The particular layout and integration of information makes these templates distinct from previous graphical templates.

The primary purpose of this paper is to present recent extensions of the two templates and partial results from ongoing usability assessment. Much of the recent research and Java implementation has centered at the National Cancer Institute (NCI). The particular goal there is to develop, evaluate, improve and deploy methodology for communicating State Cancer Profiles to state epidemiologists and other public health professionals. More broadly, the research is a part of NSF-funded digital government research to develop quality graphics for federal statistical summaries. The templates and extensions are relevant to communication and hypothesis generation efforts of numerous government agencies.

The most current work at NCI on LM plots is in progress and not yet approved for public release as of the writing of this abstract. A Java applet showing trial interactive extensions to LM plots is available http://www.netgraphi.com/cancer4/index.html. (Contact jchen@cs.gmu.edu if the site is down.) The displays are of test data, not of official cancer statistics. The full paper will address many changes emerging from ongoing usability evaluations. CCmaps is a java application and available as shareware from www.galaxy.gmu/~dcarr/ccmaps.

Since the two templates are little known, this abstract tersely describes the basic elements of the two templates as illustrated in Figures 1 and 2. For those wanting more description, a series of papers [1,2,3,4,5] describes the LM plot template and/or illustrates different applications, design variations, and uses. The only published description of CCmaps is in [4].

The primary purpose of LM plots is the communication of statistical summaries but [3] describes its application in data mining. Key features of the LM plot template are present in Figure 1. There are four columns of panels. The left column contains micromaps, the second contains names, the third and fourth contain statistical panels. LM plots have three types of panels (micromaps, names and statistical panels) that can take various forms. For example a micromap can be any "spatial" representation from a human body caricature to a communication network.

Additional key features of LM plots are sorting, perceptual grouping, and linking of multivariate descriptors. The study units in Figure 1 are states. These are sorted by bronchus and lung cancer mortality rates that appear in the third column. After sorting, states are partitioned into small perceptual groups. States are grouped into fives (with one exception) with groups represented as being above, equal to, or below the median. Distinct hues distinguish the five states in each group. The same hue links a state name, its representation in the micromap, and its estimates in statistical panels. Vertical position also links the name and estimates in most LM plots.

The template includes different kinds of statistical panels such as time series and scatterplots. In Figure 1 the statistical panels are dot plots with confidence bounds and reference lines. In contrast

to the choropleth maps, confidence bounds are shown and estimates are displayed with high perceptual accuracy of extraction by using position along a scale. Lines at the edge of the green regions indicate the U.S. Healthy People 2010 targets. The black line is the U.S. reference value.

NCI research is evaluating many interactive options. The more obvious options include selecting different data and sorting (triangle icons appear above the columns.) Additional features include mouseovers and linked blinking, color selection, a fixed header scroll, enlargement for micromaps, and drill down to see the counties of the selected state.

The purpose of CCmaps is to help researchers generate sharper hypotheses about observed spatial patterns. The particular context considered here is mortality mapping. Before conjecturing about the spatial patterns of mortality rates, it is important to produce maps that control for suspected risk factors. In epidemiology, common practice makes separate plots by race and sex to control for differences in study unit populations. Also, an age-adjusted map or a set of age-specific maps controls for study unit differences in age distribution. However, it is uncommon to see maps where efforts have been taken to control for risk factors. Sophisticated regression models provide the best means of controlling the variation due to risk factors. CCmaps provides more rudimentary but widely accessible control by partitioning study units into more homogeneous groups based on two risk factors.

As Figure 2 suggests, CCmaps supports dynamic partitioning of study units into a 3 x 3 layout of maps using newly developed partitioning sliders. The study units in Figure 2 are health service areas (HSAs), counties or aggregates of counties based on where people get their hospital care. HSAs highlighted in a panel have risk variables satisfying row and column constraints. The slider at the bottom partitions HSAs into columns based on precipitation. The slider at the right partitions HSAs into rows based on percent of households below the poverty level. The slider at the top partitions HSAs into three color classes (shown as blue, gray and red) based on lung cancer mortality rate. Hypotheses can be about spatial patterns within panels or differences among panels. Note the red in the top right panel that highlights HSAs with high precipitation and high poverty. One hypothesis might be that Southeastern HSAs have higher cigarette smoking rates. However the strong association with precipitation warrants deeper consideration. The paper will touch on issues of data availability and confounding variables.

Current features of CCmaps include statistical annotation and plots. Sliders show the percent of **people** in each class. The population weighted mean rate for HSA highlighted in each panel appears at the top right. A 3 x 3 layout of dynamic QQplots (not shown) facilitates comparing distributions. The full paper will describe numerous extensions.

[1] DB Carr and SM Pierson. "Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps," *Statistical Computing & Graphics Newsletter*, 1996; **7**(3): 16-23.

[2] DB Carr, AR Olsen, JP Courbois, SM. Pierson, and DA Carr. "Linked Micromap Plots: Named and Described," *Statistical Computing & Graphics Newsletter*, 1998; **9**(1): 24-32.

[3] DB Carr, AR Olsen SM Pierson, and JP Courbois. *Using linked micromap plots to characterize Omernik ecoregions. Data Mining and Knowledge Discovery* 2000; 4:43-67.

[4] DB Carr, JF Wallin, and DA Carr. “Two New Templates for Epidemiology Applications. Linked Micromap Plots and Conditioned Choropleth Maps,” *Statistics in Medicine* 2000; 19:2521-2538.

[5] DB Carr. Designing Linked Micromap Plots for States with many Counties,” *Statistics in Medicine* 2001; 20:1331-1339.

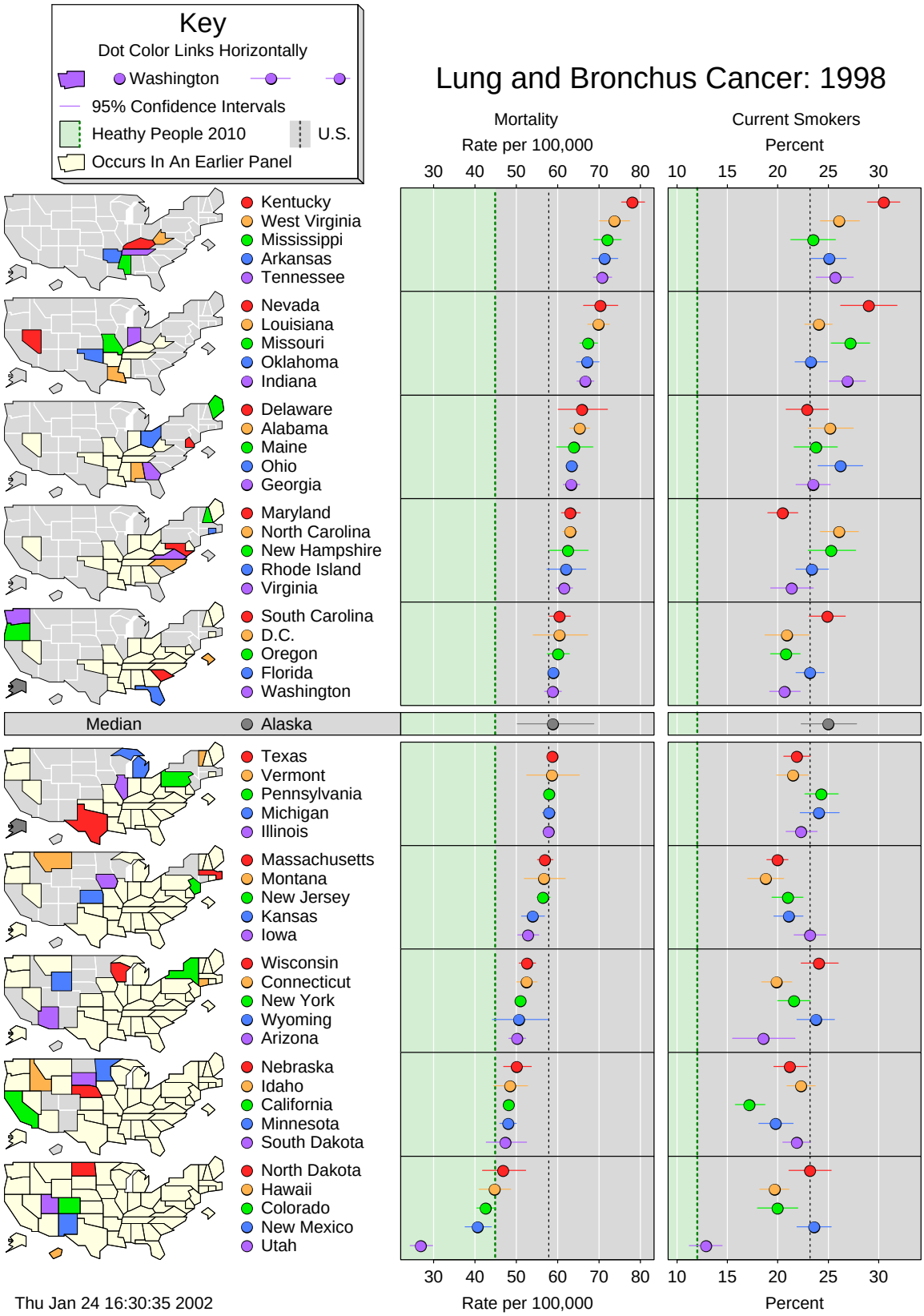


Figure 1. A Linked Micromap Plot

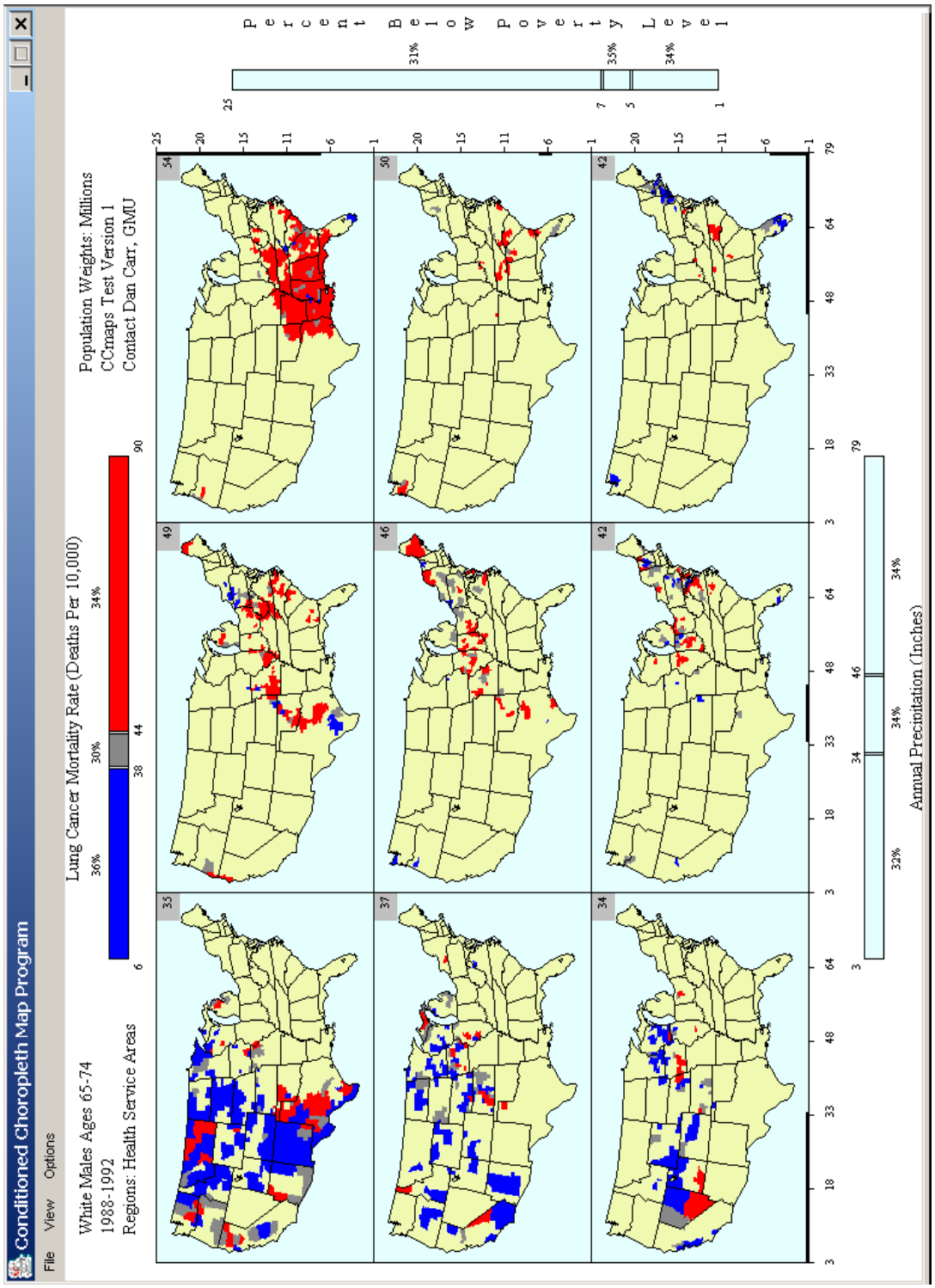


Figure 2. A Conditioned Choropleth Map

Extending Point Pattern Analysis Ability in the Commercial GIS Package

DongMei Chen and Steve Kopp

Environmental Systems Research Institute
380 New York Street
Redlands, CA 92373
Phone: 909-7932853 ext. 2730
Email: dchen@esri.com

Point data, which consist of location and attributes, are commonly used in disease, crime, population, environment, and many other social applications. However, current available tools for supporting point pattern analysis in GIS environment are mean. There have been many blocks along the way, with differences in platform, programming language, data structure and model standards, etc.

With the technology of computer industry shifting toward object-oriented, component-based software, different portions of various spatial models are allowed to be composed as objects and embedded into a GIS or GIS objects embedded into the spatial models. With these capabilities

This paper will present an example of customizing spatial analysis tools by providing a prototype of a set of point pattern analysis tools that are developed using objected-oriented and component-based design standards under a commercial GIS package. These tools will be fully integrated into the GIS and will be a good complement to the current available spatial analysis functions. They provide capabilities for the visualization and analysis of point data by exploring spatial patterns, clustering, spatial autocorrelation and relationships.

Point data analysis tools will enhance the point data analysis capability by adding more functions to support point data exploratory analysis and clustering analysis. The functionality of these tools include:

- Descriptive statistics
- Point density analysis
- Nearest neighbor analysis
- K-function
- Global spatial autocorrelation measures
- Local spatial autocorrelation measures

GEODIVERSITY: SOME SIMPLE GEOPROCESSING INDICATORS TO SUPPORT ENVIRONMENTAL BIODIVERSITY STUDIES

Jorge Xavier-da-Silva, Ph.D.*
xavier@igeo.ufrj.br
Luiz Mendes de Carvalho-Filho, M.Sc.*
mendes@ufrj.br

One of the basic problems of environmental biodiversity studies is the selection of the territorial units to be used to integrate the diversity data. It is upon these units that the identification and corresponding computation concerning the biodiversity must be safely ascertained. Our contention, in this case, is that geomorphological terminology, specifically the names of landforms identified in a thematic map, can be used as the sought territorial basis for the mentioned computations. The reasoning follows:

- Geomorphologists identify landforms, sometimes, by strange names. Some of these names are from colloquial uses, while others have disputed origins. Inselbergs, mesas, spits and beach ridges are some examples. These eventually awkward terms, however, convey relevant information about the geometry, composition and genetic processes associated with the landforms to which they refer. It is also convenient to consider that landforms compose the basis upon which many forms of life (including human) are territorially organized all over the planet.
- Moreover, if geomorphologic terms are used to identify the occurrence of a landform along a part of the earth surface, as is the case in maps, they are defining, for each area identified, an isotropic condition that can be used as a territorial basis for the computation of any variability of environmental characteristics which occur over that same portion of the earth's surface. This possibility of use of the geomorphologic classification, obviously valid as a investigative procedure for the categories of any thematic map, is particularly interesting for biodiversity, since basic environmental information (geometry, composition and origin) is automatically contained in the names of the identified landforms.
- Before disregarding the above considerations as not concerned to GIS and Geoprocessing, some attention can be given to the computational effort needed to identify the environmental diversity associated to each of the landforms composing a geomorphologic map. In effect, to measure the geodiversity of an area consists, essentially, in executing a careful territorial inventory of its prevailing physical, biotic and socio-economical characteristics. This aim requires the existence of a comprehensive environmental database, upon which the mentioned inventory is to be executed. In this regard, the methodological procedure to be used has been named VAIL (Varredura e Integração Locacional, in Portuguese – Scanning and Locational Integration) by Xavier-da-Silva, 2001, and can be used for this purpose of geodiversity investigation as well as for many other procedures of environmental information extraction. This procedure is part of the new semiotics stemming from the massive use of computational resources in environmental investigation (Bonham-Carter, 1996; Xavier-da-Silva, 2001). As a chapter of a book on biodiversity (Garay, 2001) some indicators, named Geodiversity Indexes, have been created (Xavier-da-Silva et alli, 2001a, in Garay, 2001). The present paper reports the use of these indexes in a coastal area near Rio de Janeiro, Brazil, specifically a very sensitive and preserved barrier beach complex named Restinga da Marambaia. As a background for this application, some concepts are discussed.

(*) LABORATÓRIO DE GEOPROCESSAMENTO/UFRJ/CCMN/IGEO/GEOGRAFIA, bloco I sala 1, Cidade Universitária
Rio de Janeiro (RJ), Brasil , ZC: 21.949-900 Phone:55-0XX21-22707773 Fax: 55-0XX21-25989474 www.lageop.ufrj.br

FIGURE 1

LANDFORMS	GEODIVERSITY									
	SPECIFIC - *e (e*)*							MÚLTIPLE		
	ALTITUDE	SLOPE GRADIENT	GEOLOGY	SOILS	VEGETATION COVER	LAND USE	PROXIMITIES	SIMPLE *m (m*)*	AREA *(Km^2)*	WEIGHTED *p (p*)*
Tidal flats	1 (7)	2 (5)	5 (2)	4 (2)	5 (5)	5 (2)	10 (2)	32 (4)	4,44	7,21 (13)
Grassy swamps	1 (7)	2 (5)	2 (5)	3 (3)	6 (4)	2(5)	10 (2)	26 (7)	2,14	12,15 (9)
External beach ridges	1 (7)	3 (4)	6 (1)	5 (1)	6 (4)	4 (3)	10 (2)	35 (2)	4,25	8,24 (12)
Old beach ridges	1 (7)	2 (5)	3 (4)	3 (3)	4 (6)	1 (6)	2 (9)	16 (14)	0,40	40,00 (2)
Naturally filled depression	1 (7)	1 (6)	3 (4)	2 (4)	8 (2)	3 (4)	5 (6)	23 (10)	4,07	5,65 (16)
Depression under filling	1 (7)	1 (6)	3 (4)	4 (2)	7 (3)	3 (4)	6 (5)	25 (8)	6,74	3,71 (18)
Suales, inter beach ridges depression	1 (7)	2 (5)	2 (5)	3 (3)	7 (3)	1 (6)	4 (7)	20 (12)	1,06	18,87 (8)
Dunes	1 (7)	3 (4)	3 (4)	3 (3)	4 (6)	2 (5)	9 (3)	25 (8)	5,03	4,97 (17)
Talus	4 (4)	6 (2)	2 (5)	3 (3)	2 (7)	1 (6)	1 (10)	19 (13)	1,89	10,05 (10)
Rocky structural slopocha	13 (2)	6 (2)	1 (6)	2 (4)	2 (7)	1 (6)	5 (6)	30 (5)	16,28	1,84 (19)
Set of beach ridges	1 (7)	2 (5)	3 (4)	4 (2)	7 (3)	4 (3)	6 (5)	27 (6)	15,96	1,69 (20)
Structural summits	16 (1)	6 (2)	1 (6)	1 (5)	2 (7)	1 (6)	3 (8)	30 (5)	3,10	9,68 (11)
Naturally shoaling lagoon	2 (6)	1 (6)	2 (5)	2 (4)	2 (7)	3 (4)	2 (9)	14 (15)	0,20	70,00 (1)
Wave cut plataform	2 (6)	3 (4)	1 (6)	2 (4)	1 (8)	1 (6)	1 (10)	11 (16)	0,38	28,95 (3)
Beach	1 (7)	5 (3)	5 (2)	3 (3)	9 (1)	4 (3)	7 (4)	34 (3)	5,03	6,76 (15)
Colluvial ramp or slope	1 (7)	3 (4)	4 (3)	2 (4)	6 (4)	3 (4)	5 (6)	24 (9)	0,90	26,67 (5)
Residuals of old beach ridges	1 (7)	1 (6)	3 (4)	2 (4)	5 (5)	2 (5)	2 (9)	16 (14)	2,28	7,02 (14)
Marine-colluvial terrace	3 (5)	7 (1)	4 (3)	3 (3)	4 (6)	6 (1)	11 (1)	38 (1)	2,00	19,00 (7)
Structural elongated summits	7 (3)	6 (2)	1 (6)	1 (5)	2 (7)	1 (6)	3 (8)	21 (11)	0,98	21,43 (6)
Spits	1 (7)	1 (6)	2 (5)	1 (5)	2 (7)	2 (5)	1 (10)	10 (17)	0,37	27,03 (4)

(*) LABORATÓRIO DE GEOPROCESSAMENTO/UFRJ/CCMN/IGEO/GEOGRAFIA, bloco I sala 1, Cidade Universitária
Rio de Janeiro (RJ), Brasil , ZC: 21.949-900 Phone:55-0XX21-22707773 Fax: 55-0XX21-25989474 www.lageop.ufrj.br

Geodiversity is a concept which can be grasped if the environment is considered as the locus of a convergence of causing factors. It is the assemblage of environmental characteristics incident over a particular place. Being axiomatically unique, it can be operationally identified through the use of a digital model of that place (Xavier-da-Silva, 1982). This model (as any model), being a simplification of the reality, must encompass the environmental aspects relevant to the investigation under realization.

Another conceptual aspect deserving attention is the assumption that the biodiversity found in a certain environment is directly associated to its geodiversity. This postulate, which apparently pervades this paper, actually needs to be verified as to the intensity and singularities of the postulated direct association. In this regard, the application of the present or similar indexes can document the presence and degree of this relationship between the overall characteristics of a specific environment – geodiversity - and the variability of its biological components, i.e. its biodiversity.

From the above delineated connection between the concept of geodiversity and the digital model of the environment, which allows the creation of its representative parameters, it follows that some simple geoprocessing techniques can be used to exhaustively measure the geodiversity existing over any area. The results computed over the set of thematic maps (a digital database) can be presented as a table of easy reading (fig. 1). Through the inspection of its rows and columns relevant information can be obtained. This information can assume several formats, some specific, some generalized. The geodiversity indicators presented in the table of fig. 1 are representative of these informational formats. They received denominations believed to be clearly related to their logical content. A systematic presentation of these indicators, to be accompanied by the observation of the table of fig. 1, is made in the following paragraphs.

A - Indexes of Specific Diversity "*e* " and Specific Diversity Rank ("*e")**

The first of these indicators, "e", presents, for each thematic map, the amount of its categories found in each of the classes of the map used as a basis for computation (classes of landforms, in the table of fig. 1). This is the particular diversity of each thematic map found in association with a specific landform. The Specific Diversity Rank ("e*"), namely indicates the ordinal position of the previous index in relation to all other considered classes of landforms.

The two indexes defined above allow comparisons between their values and also along their respective columns, in which are registered the variability of the considered parameter (see the values for Altitude, in fig. 1) in terms of number of categories and respective ranks. It can be seen, for example, that the seventh rank was found for thirteen types of landforms, a validating result totally expected for a low lying coastal area such as the Restinga da Marambaia (see fig. 2).

Non-parametric rank correlation coefficients can be calculated using the values of the rank indexes obtained for each thematic map. These correlations would indicate, to some extent, the degree of independence (or dependence) existing among the parameters chosen for the analysis of the local geodiversity.

B - Indexes of Simple Multiple Geodiversity "*m*" and Simple Multiple Diversity Rank ("*m")**

These indexes are named "simple" because they are non weighted and directly obtained through the sum along the lines of the table of fig. 1. For example, the number 32 found in the eight column of the mentioned table, represents the sum of the classes of each of the seven environmental parameters considered in the performed analysis. It is

accompanied by the number (4), which indicates that, in terms of the simple multiple diversity rank, the landform “tidal shoals” was ranked in the fourth place.

Both indexes immediately above are direct indicators of Geodiversity. Their values may be used for direct comparisons among different results obtained for different areas, but they can be used, basically, to build the next indexes, presented below.

C - Indexes of Weighted Multiple Diversity “*p*” and Weighted Multiple Diversity Rank (“*p**”)

If data concerning the areas occupied by each of the landforms are made available (column 9 of the table of fig. 1), the geodiversity can be expressed through a ratio between the number of classes registered in the simple multiple geodiversity index and the area of the respective landform. This is a measure of the density of the environmental diversity found in association with each landform and the composite index thus generated can also be ranked. The ratio is the Weighted Multiple Diversity Index “p”, and its respective relative position regarding the values associated to the other landforms is the Weighted Multiple Diversity Rank (“p*”).

The main importance of these last indexes is to permit comparisons among different results obtained from different areas. It also obvious that the index “p” can identify areas of great interest, such as small territorial units with great geodiversity. For this type of geographic areas it is expected a great interaction among all the forms of life, generating, in principle, greater biodiversity than the one expected over large areas with the same amount of environmental variability, that is, having smaller density of variability. Selection of areas deserving strict protection against misuses can be based on this type of quantified identification, and rules of environmental management can be devised on the basis of extensive tabulation of values and possibilities of correlations contained in this type of environmental inventory.

REFERENCES

ARONOFF, S. (1991). *Geographic Information Systems: a management perspective*. Ottawa: WDL, 298 p.

BONHAM-CARTER, G.F. (1996). *Geographic Information Systems for geoscientists*. Ontario: Pergamon, 400 p.

CHRISTOFOLETTI, A. (1999). *Modelagem de sistemas ambientais*.São Paulo:Edgard Blücher, 200 p.

GARAY, I. and DIAS, B.F.S. (2001) Conservação da biodiversidade em ecossistemas tropicais – Petrópolis – Editora Vozes – 432 p.

GOES, M.H.B. e outros (2000). *Um modelo digital para a restinga e paleoilha da Marambaia (RJ) para fins militares, investigação científica e ecoturismo controlado*. Rio de Janeiro: LGA/UFRJ, no prelo.

XAVIER-DA-SILVA, J. (1982). A digital model of the environment, na effective approach to areal analysis. *Anais do International Geographic Studies*. Rio de Janeiro: UGI/UFRJ, p.17-22.

XAVIER-DA-SILVA, J. (1994). Geomorfologia: uma atualização de bases e conceitos. *Geomorfologia*/Sandra Baptista Cunha e Antônio José Teixeira Guerra, organizadores. Rio de Janeiro: Bertrand Brasil, p.393-414.

XAVIER-DA-SILVA, J. (2000). Geomorfologia, Análise Ambiental e Geoprocessamento. *Revista Brasileira de Geomorfologia*. Uberlândia: UGB/UFU, volume 1 (1) p.48-58.

XAVIER-DA-SILVA, J. e CARVALHO-FILHO, L.M. (1993). Sistemas de Informação Geográfica: uma proposta metodológica. *Anais da IV Conferência Latinoamericana sobre Sistemas de Informação Geográfica/2° Simpósio Brasileiro de Geoprocessamento*. São Paulo: EDUSP, p.609-628.

XAVIER-DA-SILVA, J.; ALMEIDA, L.F.B. e CARVALHO-FILHO, L.M. (1996). Geomorofologia e Geoprocessamento. *Geomorfologia, técnicas e aplicações*/Sandra Baptista Cunha e Antônio José Teixeira Guerra, organizadores. Rio de Janeiro: Bertrand Brasil, p.283-309.

XAVIER-DA-SILVA, J.; PERSSON, V.G.; LORINI, M.L.; BERGAMO, R.B.A.; RIBEIRO, M.R.; COSTA, AJ.S.T.; IERVOLINO, P.; ABDO, O.E. (2001a). Índices de Geodiversidade: aplicação de SGI em

estudos de biodiversidade. *Conservação da biodiversidade em ecossistemas tropicais*/Irene Garay e B. Dias, organizadores. Rio de Janeiro: Vozes, 299-316

XAVIER-DA-SILVA, J. (2001b) Geoprocessamento para análise ambiental – Rio de Janeiro – Edição do autor – 228 p

Multiresolution Terrain Processing and Visualization with VARIANT (Extended Abstract)

Leila De Floriani, Paola Magillo
Department of Computer and Information Sciences, University of Genova,
Via Dodecaneso, 35, 16146 Genova, ITALY
Email: {deflo,magillo}@disi.unige.it

1 Introduction

In this paper, we present *VARIANT (VAriable Resolution Interactive ANalysis of Terrain)*, a system for processing and visualizing terrains represented through Triangulated Irregular Networks (TINs) at multiple resolutions. The main feature of VARIANT is the possibility of adapting resolution (i.e., the density of the TIN) locally and in a dynamic way, based on the current user needs. Thus, the user can deal with TINs having a high resolution (and, thus, a high accuracy) on interesting domain areas and/or elevation ranges, and an overall small size.

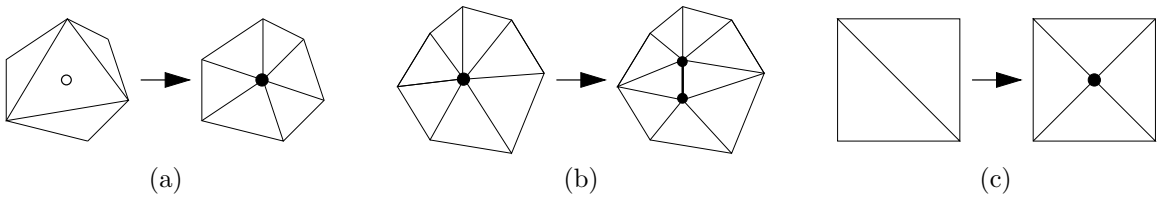
The architecture of VARIANT is modular. It is built on the *Multi-Triangulation (MT) Library* [7], an extensible library for multiresolution geometric modeling, that has been developed by our research group. The kernel of the MT library is dimension- and application-independent, and provides the primitive mechanisms to build, encode, and query multiresolution geometric models. The next level of the MT library contains more advanced functionalities for terrains. The higher levels in the architecture of VARIANT contain the specific operations implemented in VARIANT, and the user interface of the system.

VARIANT directly supports basic queries (e.g., windowing, buffering, computation of elevation at a given point, or along a given line) as well as high-level operations (e.g., fly-over visualization, contour map extraction, view-shed analysis), in real time. The system can be enriched with many other functionalities typical of GIS applications.

2 Multiresolution TINs

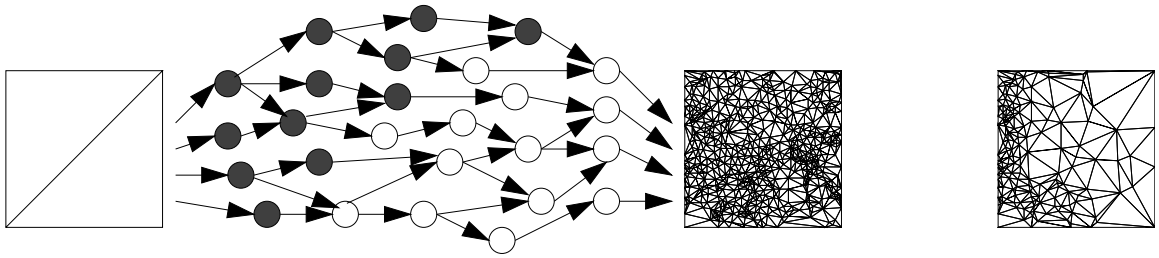
A Triangulated Irregular Network (TIN) is a terrain model which consists of a triangle mesh covering a plane domain, and a set of elevation values associated with the vertices of the triangulation. Elevation values are used to lift the two-dimensional triangulation into three dimensional space. In general, the vertices of a TIN can be at arbitrary positions, but it is worth noticing that triangulations of regular square grids are special cases of TINs.

A *Multi-Triangulation (MT)* is composed of a *base TIN* at a coarse resolution plus a *partially ordered set of updates* that can be applied to progressively refine the base TIN into a TIN at the full resolution. A range of TINs at intermediate resolutions can be extracted from an MT by applying a subset of its updates, while respecting the partial order.



(a) (b) (c)

Figure 1: TIN updates: (a) vertex insertion, (b) vertex expansion into an edge, (c) bisection of a pair of right triangles.



(a) (b)

Figure 2: (a) The graph describing an MT, where nodes represent updates and arcs represent the partial order. The base TIN (just two triangles) and the TIN at the maximum resolution (obtained by applying all updates) are shown. The nodes filled in grey form a consistent subset of updates. (b) The TIN defined by the consistent set of updates shown in (a), which has a variable resolution across the domain.

An *update* consists of replacing a set of triangles with another, larger, set of triangles. Each update has a a *local* action, i.e., it covers a small extension in the domain. Examples of updates are: insertion of a vertex and local re-triangulation (e.g., based on the Delaunay criterion), expansion of a vertex into an edge. For TINs built over a regular grid of points, an example is the insertion of a vertex on the common edge of two adjacent right triangles, which share their longest edge (see Figure 1).

A *partial order* is defined among updates, based on the following rule. Whenever an update u_2 removes some triangles created by another update u_1, then u_1 precedes u_2. An MT is represented by a directed acyclic graph in which the nodes correspond to the updates, and the arcs are induced by the above rule (see Figure 2a).

A TIN at variable resolution is obtained from an MT by considering a subset S of nodes which is *consistent* (i.e., for every node $u \in S$, each node preceding u is also in S), and applying the corresponding updates to the base TIN (see Figure 2b)

The fundamental operation on an MT is *selective refinement*, which consists of extracting a TIN having a resolution that satisfies some user-defined requirements. User requirements are expressed as a Boolean function R on the MT nodes. Such function discriminates updates that need to be applied, and updates that do not need to be applied, in order to achieve the resolution desired by the user. By defining function R in a suitable way, it is possible to focus resolution on certain domain areas and/or ranges of elevations (see Section 3).

Selective refinement is performed through a traversal of the DAG describing an MT, which finds a minimal consistent set S containing all nodes that are classified as *necessary* by function R. The TIN obtained from the base TIN by applying the updates in S is returned. In [1], we have proposed

an incremental algorithm for selective refinement, which is able to modify a current TIN in real time, while the user changes the function R.

3 Terrain Operations at Variable Resolution

The following terrain operations are currently provided by VARIANT: *domain queries* for focusing the attention on a certain area, *interactive visualization*, testing the *mutual visibility of two points*, and computation of *contour lines*. Some examples are shown in Figure 3.

3.1 Queries in the Domain

Basic queries permit to determine the terrain configuration within a certain *region of interest* in the domain, which can be a point (*point location* query), a rectangular axis-parallel box (*windowing* query), a circle centered at a point (*range* query), a polygonal line (which may correspond to a road, a river, etc.).

All above operations reduce to selective refinement from an MT by considering a Boolean function R defined in the following way:

- Any update u which does not interfere with the given query entity does not need to be applied;
- An update u which interferes with the given query needs to be applied if and only if some of the triangles removed by u have an approximation error larger than ε, where ε is a constant set by the user.

For instance, by setting $\varepsilon = 0$, we get the maximum resolution in the interesting terrain portion, while resolution is lower outside, and decreases gradually in order to maintain surface continuity (see Figure 3b and c).

3.2 Interactive Terrain Visualization

In interactive terrain visualization, the user can drive its viewpoint to fly over the terrain. In order to obtain an interactive frame rate, it is necessary to render triangle meshes with a restricted number of triangles, but still having a visual appearance not perceptually dissimilar from the mesh at the maximum resolution (see Figure 3a).

For such purpose, we adopt a function R defined in this way:

- Any update u which does not intersect the view field, does not need to be performed;
- An update u which intersects the view field needs to be performed if and only if some of the triangles removed by u have an approximation error which is larger than $f(d)$, where d is the distance of the update from the viewpoint, and f is an increasing function.

3.3 Point Visibility

Two points on a terrain described as a TIN are mutually visible if and only if the interior of the segment joining them lies completely above the surface (without intersecting any triangle of the TIN). In order to test the mutual visibility of two points p and q on a TIN, we have to find all triangles lying at least partially above segment pq. Points q and q are visible if and only it such triangle set is empty. With a Multi-Triangulation, such test is done by considering the maximum

resolution just on those updates which contain triangles that are candidate to be reported as blocking the view between p and q.

VARIANT provides a module to compute the *visibility graph* of a set P of points, i.e., a graph having its vertices at the points of P, and containing an arc (p, q), with $p \neq q$ if and only if p, q are mutually visible (see Figure 3d).

3.4 Contour Lines

Contour lines are among the most common and natural way to visualize a terrain. Given an elevation h, the contour lines at height h are the intersection of the surface with the plane of equation $z = h$. For a TIN, the contour lines are a collection of closed or open polygonal chains. The projections of several contours (corresponding to a sequence of field values) onto the xy-plane form a *contour map*.

A Multi-Triangulation enhances the generation of contour maps for a terrain by using selective refinement to restrict attention on triangles which contribute to the contour map. For such purpose, we define function R in the following way:

- If the terrain portion covered by an update u cannot contain elevation h, then u does not need to be applied;
- Otherwise, u needs to be applied if and only if some triangle removed by u has an approximation error larger than a user-defined constant ε.

The decision whether the terrain portion covered by an update u can contain value h is based on the elevation range spanned by the triangles of u and on the approximation error associated with such triangles. Next, the contour lines are determined by computing the intersection segments of the triangles returned by the query with the horizontal planes at the given height.

4 Further Developments

In its current version, the MT library user a general-purpose data structure to encode the Multi-Triangulation. Such structure has the advantage to be able to store MTs built through any refinement strategy (e.g., vertex insertion, vertex expansion, refinement of right triangles), but has the drawback of being rather large. We have studied compact data structures for MTs built through specific refinement strategies [4, 3, 2, 6], and we are going to integrate them in the MT library.

VARIANT can be enriched with many other terrain operations, which are suitable to benefit from selective refinement, for instance, *view-shed computation* (see [5]).

The same multiresolution approach used in VARIANT can be adopted for visualizing and processing higher dimensional scalar fields. A prototype system for scientific visualization based on the MT library is described in [1].

References

[1] P. Cignoni, L. De Floriani, P. Magillo, E. Puppo, and R. Scopigno. TAn2 - visualization of large irregular volume datasets. Technical Report DISI-TR-00-07, Department of Computer and Information Sciences, University of Genova (Italy), 2000 (submitted for publication).

[2] E. Danovaro, L. De Floriani, P. Magillo, and E. Puppo. Compressing multiresolution triangle meshes. In *Proceedings 7th International Symposium on Spatial and Temporal Databases, SSTD 2001*, Los Angeles, CA, USA, July 12-15 2001.

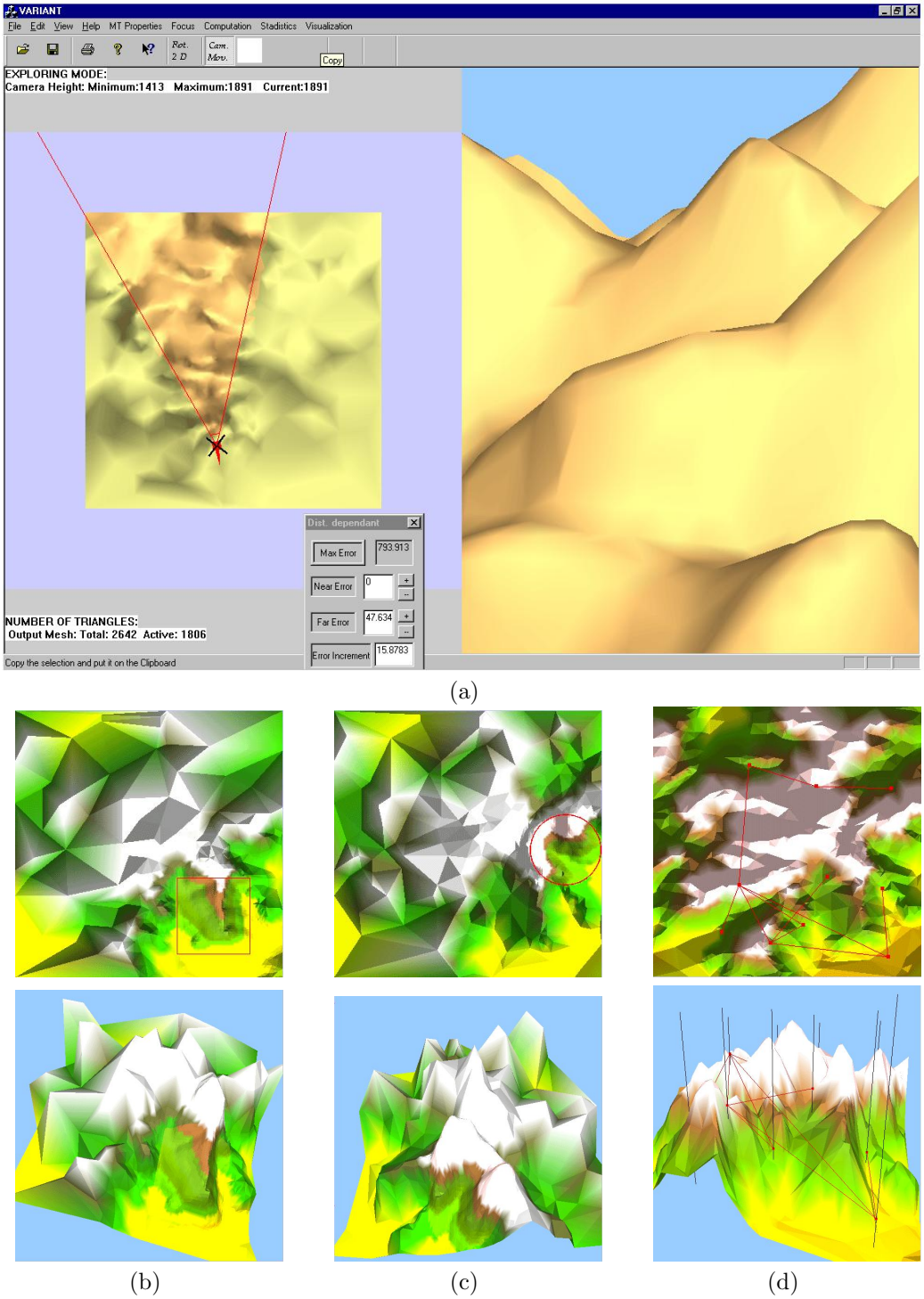


Figure 3: GIS operations with VARIANT: (a) interactive terrain navigation, (b) a box query, (c) a range query, (d) a visibility graph.

[3] E. Danovaro, L. De Floriani, P. Magillo, and E. Puppo. Representing vertex-based simplicial multi-complexes. In G. Bertrand, A. Imiya, and R. Klette, editors, *Digital and Image Geometry, Lecture Notes in Computer Science*, volume 2243, pages 128–147. Springer-Verlag, 2001.

[4] L. De Floriani, P. Magillo, and E. Puppo. Data structures for simplicial multi-complexes. In Guting, Papadias, and Lochovsky, editors, *Advances in Spatial Databases*, volume 1651 of *Lecture Notes in Computer Science*, pages 33–51. Springer Verlag, 1999.

[5] L. De Floriani, P. Magillo, and E. Puppo. A library for multiresolution modeling of field data in gis. In *Int. Workshop on Emerging Technologies for Geo-Based Applications*, pages 133–151, Ascona, Switzerland, May 2000. Swiss Federal Institute of Technology, Lausanne.

[6] M. Lee, L. De Floriani, and H. Samet. Constant-time neighbor finding in hierarchical meshes. In *Proceedings International Conference on Shape Modeling*, pages 286–295, Genova (Italy), May 7-11 2001.

[7] P. Magillo. *The MT (Multi-Tesselation) package*. Dept. of Computer and Information Sciences, University of Genova, Italy, http://www.disi.unige.it/person/MagilloP/MT/index.html, January 2000.

An Enhanced GIS Environment For Multivariate Exploration: a Linked Parallel Coordinate Plot Applied to Urban Greenway Use Survey Data

Robert M. Edsall
Department of Geography
Arizona State University
robedsall@asu.edu

Anna J. Roedler
Department of Geography
Arizona State University
ajroedler@yahoo.com

Abstract

In the winter of 2001-02, a study of use patterns of an urban greenway in Scottsdale, AZ, was carried out using on-site surveying. This study is designed to assess the influence of the greenways on the perception of the quality of life for users of the park. This paper will discuss methods of geovisualization and exploratory data analysis applied to the display and interpretation of the data that comes from these surveys. The representation of survey results poses challenges because the data are both multivariate and involve variables of several different measurement levels, including nominal-, ordinal-, and ratio-level. These variables include diverse but potentially related indices such as distance traveled, primary use type, age, and perception of safety. The character of the survey data makes visualization of the database both necessary and difficult: maps and graphs from ESDA and geovisualization that show many variables at a given time were required. For this task, a modification of a customized ArcView interface, enhanced with a parallel coordinate plot representation, was created. The parallel coordinate plot application in ArcView, heretofore used for visualization of physical (climate) and epidemiological (cancer mortality) data, proved insightful for the display of the social and cultural data contained within the surveys.

The overall project was inspired by a study conducted in Texas based on a user survey of three different urban greenways in Texas (Shafer et al. 2000). Through a three-day survey of trail users this group collect an astounding 1004 on-site surveys. After analysis of the survey results the three greenways were found to have very different use patterns. Although this study focused on how overall the three greenways were perceived to contribute to quality of life, there was little discussion of how the use of the greenway affected this perception. The results were mainly displayed using conventional tables of statistics. If the results could have been displayed not only by greenway, but also by use or by another variable, the researchers might have been able to gain insight about whether use patterns affected answers or whether it was another variable that had a strong correlation with certain quality of life answers. Our study is based on their methodological approach to greenway survey collection, but also incorporates the use of methods of geovisualization and ESDA to analyze the survey results.

The use of such methods in such a social science application, however, poses interesting and challenging issues. Geovisualization (and visualization in general) has traditionally been applied to the numerical data sets commonly found in the physical sciences (Orford et al. 1999). The adaptation of these techniques to handle qualitative data will be important for its application to the social sciences. Visualization methods in the social sciences often are directly adapted from those designed for quantitative data, as nominal or ranked data is converted, possibly arbitrarily, into quantitative ratio-level variables to fit methods of

visualization (Dorling 1992). This is often the case in information visualization applications, where text and other apparently qualitative data types are converted and plotted into a similarity or possibility space based on some numerical index applied to each observation (Rose 1999).

The challenge, thus, is to devise a scheme to place values of a variable in some way that reflects their relative value to one another. This creates the danger of prompting false impressions of order in (non-ordered or arbitrarily ordered) qualitative data. Our paper will discuss attempts at tackling this challenge. One solution involves associating a nominal data set with a ordered numerical variable: for example, the survey asked participants which park feature they most frequently used. We decided that the park features would be ranked in terms of their mobility. That meant that fishing was given the lowest number of 1 (since it was the most dependent on a small area) and biking on paths was given the highest rank of 9. Another possible solution explored in the development of our application is the "mapping" of nominal data in an arbitrary order with a user-friendly interface element that allows – indeed encourages – the rearrangement of the values, thus emphasizing to a user the fluidity and subjectivity of the default ordering.

Aside from the inherent and important issues involved with visualizing nominal data, the survey results are also difficult to represent because of their multivariate nature. A simple and elegant solution to this problem is the parallel coordinate plot (PCP), first described by Inselberg (1985). On the parallel coordinate plot, observations are represented on a PCP as a series of unbroken line segments instead of as points (as in a scatter plot). These lines pass through parallel axes, each of which represents a different variable. Each line passes through an axis at a location that indicates the observation's value relative to all other values. The ends of the axis represent the maximum and minimum values of the axis variable for all observations under consideration. The result is a (possibly unique) multivariate *signature* for each observation, and a visual representation of relationships among many variables.

Using the PCP, interactions among variables can be quickly identified. Observations with similar data values across all variables will share similar signatures; clusters of like observations (survey respondents in this case) can thus be discerned. Two variables directly related to one another would appear on the PCP as two axes connected by a series of parallel (or at the least non-crossing) line segments. Conversely, an inverse relationship between two variables would be displayed as a series of line segments that cross each other between the axes. In a PCP representation of automobile specifications, gas mileage and vehicle weight is shown by Wegman (1990) to exhibit this relationship, and the PCP representation resembles, vaguely, a "bow tie" of intersecting line segments. Another advantage of the PCP is the easy visualization of distributions and characters of many variables at once; the PCP design resembles a series of connected histograms, where distributions that are normal or skewed, and variables that are continuous or discrete, can easily be distinguished.

Although this technique has been applied successfully to physical phenomena and to large data sets (MacEachren et al. 1999), its use in the social sciences seems to be lagging (Fotheringham 1999). One reason is that the PCP generally requires qualitative data to be subjectively ordered, since the order of the possible values on each axis has an effect on the resulting patterns of the lines and thus on the outcome of the visualization. The PCP has been modified by several developers to accomplish certain goals and represent certain types

of data (Wegman 1990; Chang and Yang 1996; ; Fua et al. 1999; Edsall 1999), and this paper will represent another PCP modification that is designed to encourage creative thinking about the relationships among both qualitative and quantitative variables.

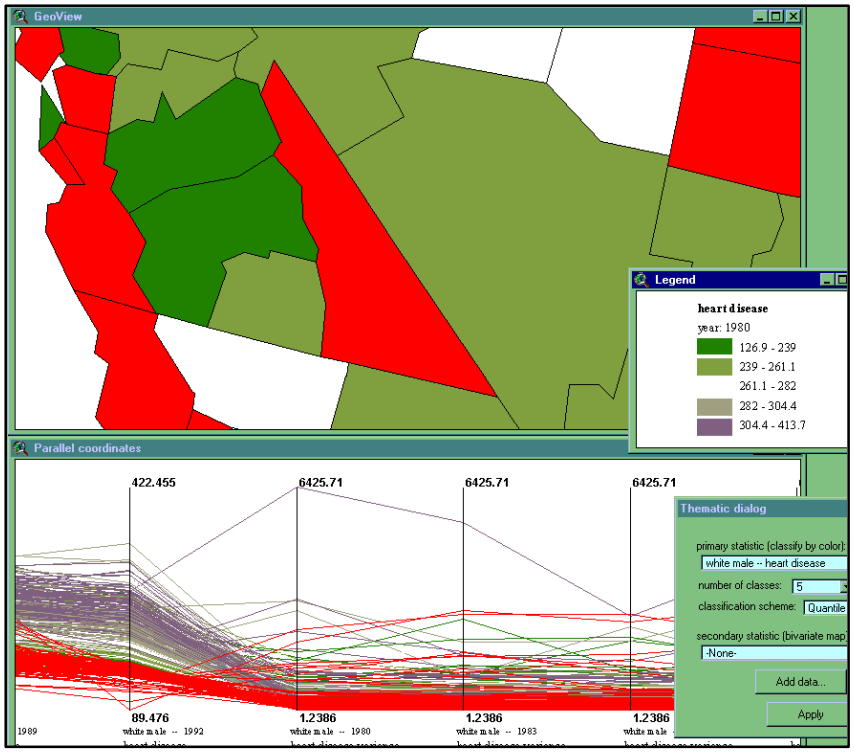


Figure 1. The modified ArcView interface on which the tool described here is based (from Edsall 2002).

The environment to be presented is an enhanced version of ArcView's interface. The enhanced version is itself a modification of an environment developed for the visualization of health statistics (Figure 1, Edsall 2002). Geographic information systems have been linked to exploratory data analysis software in the past (Cook, Majure et al. 1996). This version was created using ArcView's internal interface development language. The GIS environment was selected for its facility to link newly constructed graphical representations to maps, clearly a priority in geovisualization applications. In this application, each participant's trace in the PCP will be linked to the corresponding location of their interview and/or the recreational area being used depicted on the survey map. A variety of interactive elements, including but not limited to those that are standard in GIS interfaces (like a pan and zoom), were added to the PCP (and where appropriate, linked to the map) to facilitate the visual analysis of the data. These elements will be described in detail in the full paper.

By adapting methods of visualization to fit qualitative and quantitative data collected in the social sciences it is hoped that more effective analysis can take place. Through the example of visualizing survey results with the parallel coordinate plot we will show how visualization can be used with data collected to describe human behavior and perceptions.

References

Chang, D. H. and S. J. Yang (1996). "Dynamic Parallel Coordinate Plot and its Usage." *Journal of the Korean Society of Applied Statistics* 9 : 45-52.

Cook, D., J. J. Majure, J. Symanzik and N. Cressie (1996). "Dynamic Graphics in a GIS: Exploring and Analyzing Multivariate Spatial Data Using Linked Software." *Computational Statistics* 11 : 467-480.

Dorling, D. 1992. Visualizing People in Time and Space, *Environment and Planning B: Planning and Design*. 19: 613-637.

Edsall, R. (1999). Development of Interactive Tools for the Exploration of Large Geographic Databases. *Proceedings of the 19th International Cartographic Conference*, Ottawa, University of Victoria. 34B.

---------- (2002) The parallel coordinate plot in action. *Computational Statistics and Data Analysis* (in press).

Fotheringham, S. 1999. Trends in qualitative methods III: stressing the visual. *Progress in Human Geography* 23(4): 597-606.

Fua, Y.-H., M. O. Ward and E. A. Rudensteiner (1999). Hierarchical parallel coordinates for exploration of large datasets. *Proceedings of the IEEE Symposium of Information Visualization*, San Francisco, CA, IEEE Computer Society.

Inselberg, A. (1985). "The Plane with Parallel Coordinates." *The Visual Computer* 1 : 69-97.

MacEachren, A., Wachowicz, M., Edsall, R., and Haug, D., 1999. Constructing knowledge from multivariable integrating geographical visualization with knowledge discovery in database methods, *International Journal of Geographical Information Science*. 13 (4): 311-334

Orford, S., Harris, R., and Dorling, D. 1999. Geography: Information Visualization in the Social Sciences., A state-of-the-art-review. *Social Science Computer Review*. 17(3): 289-304.

Rose, S. (1999). The sunflower visualization metaphor, a new paradigm for dimensional compression. *Proceedings of the IEEE Symposium on Information Visualization (InfoVis '99)*, San Francisco, CA, IEEE Computer Society

Shafer, C.S., Koo Lee, B., and Turner, S., 2000. A tale of three greenway trails: user perception related to quality of life. *Landscape and Urban Planning* 49: 163-178.

Wegman, J. J. (1990). "Hyperdimensional Data Analysis Using Parallel Coordinates." *Journal of the American Statistical Association* 85 (411): 664-675.

Spatial Analysis Software Tools for Community Decision Support

Chris Fulcher*, Yan Barnett** and Chris Barnett***
Center for Agricultural, Resource and Environmental Systems (CARES)
Community Informatics Resource Center (CIRC)
University of Missouri-Columbia

Local governments are increasingly faced with making decisions that were once delegated to the Federal government. However, with this devolution of power, local governments often lack the resources or information to make effective decisions that impact their communities. To further exacerbate this issue, rural communities in the United States are typically resource-poor compared to their more affluent urban counterparts. Specifically, rural communities typically lack the expertise and infrastructure (i.e., access to information, equipment, computer hardware and software) required to make more informed decisions. At the same time, local governments are increasingly employing Participatory Action Research (PAR) methods to address local issues. Participatory Action Research increases the need for accessible, user-friendly, interactive decision support tools to evaluate socio-economic and environmental impacts of group decision making at the local level.

Information science, coupled with emerging information and communications technology (ICT) including geographic information systems (GIS), remote sensing, and data visualization, are increasingly being used to address policy options at the local level. However, there are several drawbacks to using these technologies: (1) limited access – the tools are often developed on stand-alone computers (i.e., not Internet-based); (2) expensive - the cost of the required hardware and software may preclude less affluent communities from using the tools; and (3) resource poor communities may lack the expertise to use the tools and interpret the results. Rural access to the Internet is not considered a limitation in the long run as ICTs continue to evolve (e.g., satellite connectivity to the Internet).

This paper will highlight the spatial decision support tools developed at CARES and CIRC at the University of Missouri-Columbia. These Internet-based tools overcome the shortcomings of traditional decision support tools by increasing access via the Internet, reducing costs and minimizing the expertise required to use the tools; thereby leveling the playing field between primarily resource rich urban and less affluent rural communities in the United States.

There are several advantages to using an Internet-based GIS: (1) Cost effective: The Internet is an efficient and affordable way to distribute information; (2) No GIS software is required: Users only need a browser such as Netscape or Internet Explorer to interact with the Internet-based GIS; (3) No data distribution required: All data and GIS functionality is updated via a centralized server, thus avoiding significant data management and distribution issues; (4) Effective Participatory Action Research tool for engaging stakeholders in their communities; (5) Effective research collaboration tool for engaging scientists in distributed data collection, synthesis, analysis and dissemination.

The Center for Agricultural, Resource and Environmental Systems (CARES) is an intercollegiate research and education center within the College of Agriculture, Food and Natural Resources at the University of Missouri – Columbia. CARES was established in 1992 with the purpose of helping people better understand and address agricultural, natural resource and environmental issues using knowledge and information technologies. The Rural Policy Research Institute (RUPRI) and CARES recently established the Community Informatics Resource Center (CIRC), which is housed in CARES. This new center builds on the Missouri state-level applications offered by CARES and applies the concepts and Interactive Mapping tools to the national level.

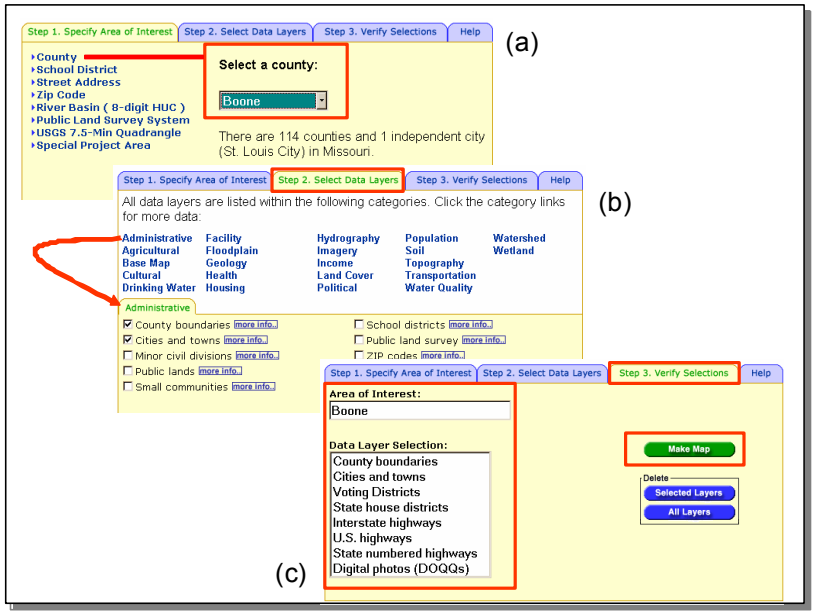
A Holistic Framework for Decision Support

An agency-based framework for disseminating data is not a desirable structure for local governments and citizen groups concerned with issues, such as land use, since they require GIS layers and other data from several agencies. For example, if a citizen steering committee appointed by county commissioners is addressing a land use issue in their county, several layers would be needed, including: land cover satellite imagery which may be obtained from the Missouri Department of Conservation, road networks from the Missouri Department of Transportation, socio-economic and demographic data from the U.S. Census Bureau, soils from the U.S. Natural Resources Conservation

* Assistant Professor, Truman School of Public Affairs, Co-Director, CARES, Director, CIRC; ** Research Associate and *** Associate Director, CARES

Service, and stream networks from the U. S. Environmental Protection Agency or the Missouri Department of Natural Resources. This steering committee, most local governments, and citizen groups will likely not have the resources to readily integrate data from Agency-based Internet Mapping websites and other clearinghouses of GIS data into a meaningful decision support system.

Therefore, CARES coordinated with state and federal agencies to integrate agency-based Internet Mapping websites into a holistic framework for decision support. "Holistic" is often defined as an approach that emphasizes the importance of the whole and the interdependence of its parts. In other words, the "whole" reflects the decision support framework, while the parts are the agency-based data that contribute to the whole.



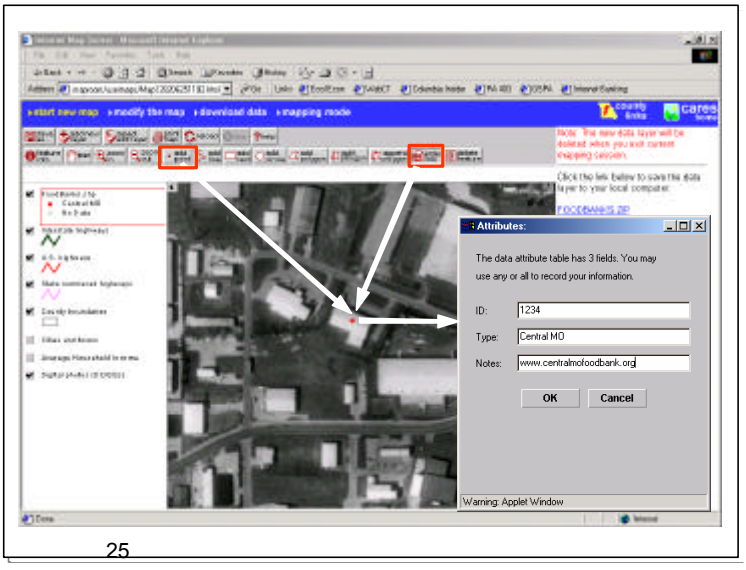
CARES proceeded to develop an Internet Mapping website using focus groups that had no background in GIS or Internet Mapping. These focus groups provided feedback on menu interface development, navigation around the website, and improving the functionality of the mapping tools. Based on focus group input, a three step process was developed for accessing the Internet Mapping website (see figure to right).

Development and Implementation of a Digitizing Tool for Adding Local Knowledge

Engaging citizens and the local government in a Participatory Action Research process is quite valuable in that they provided information, due to their extensive knowledge of the area that may otherwise be missing from agency databases. CARES therefore developed a suite of tools that allows users to digitize (create) their own GIS layers through the Internet. These unique tools enable researchers to engage communities in a participatory research framework by creating "living maps" through the Internet.

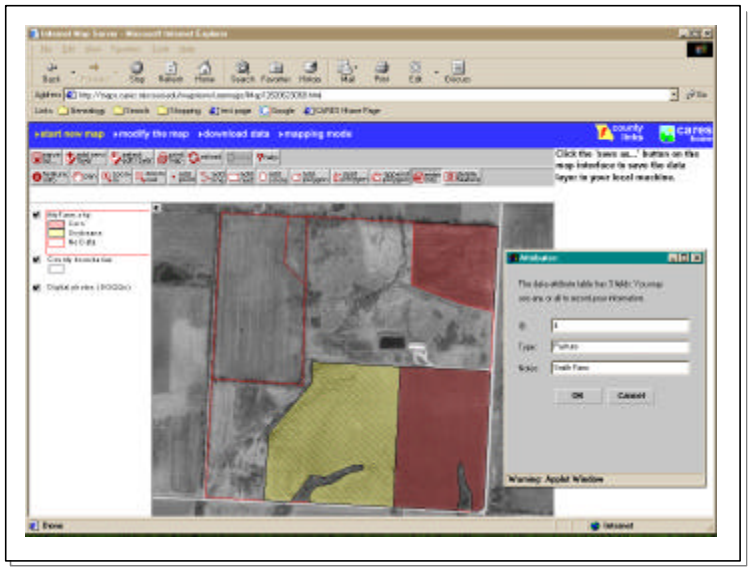
Specifically, the digitizing tools were created to enter in real-time spatial data, including attribute information, via the Internet or a secured Intranet. Using a standard web browser, a user can zoom in on any given location in the Missouri and add or edit geographic features such as points (e.g., households, well locations), lines (e.g., roads, streams), and polygons (e.g., fields or jurisdictional boundaries). 1-meter imagery or other detailed maps often serves as a reference layer to digitize these geographic features.

The figure to the right illustrates the process of entering in a point on top of a house using 1-meter DOQQ imagery and attributing, or adding intelligence to that point using a "proof of concept" menu interface. Assume a user wishes to create a new GIS layer for community foodbanks to show locations of where food is distributed to hungry people. One purpose of adding foodbank locations may be to determine whether they are in optimal locations based on changing demographics (i.e., proximity to areas in poverty based on income census data). The figure shows the process of first creating a



new layer called “foodbanks”; second, clicking on the “add point” button and then clicking on top of the house to generate a point. The user then clicks on the “enter info” button to add attributes to that point. Attributes may include volume of food distributed in pounds, website address, phone number, etc.

In another example, the same digitizing procedure can be used by a farmer to enter data related to crop management practices. The farmer can use the digitizing tool to first delineate field boundaries using the “add polygon” tool, then attribute those fields with information about what crops will be planted the following year (see figure to the right).



The tools served as a foundation for developing a national-level Internet Mapping tool for the 4-H National Technology Conference, which was held at the University of Maryland, College Park from July 8-12, 2000. The objective of the mapping tool was to enable 4-H youth from around the U.S. to zoom from a national map of the U.S. down to their community. The user can then put a “pin” on the map by clicking on their neighborhood with a mouse. A “Community Technology Self-Assessment” survey then appears on the screen and prompts the user to answer questions about their level of Internet access and other technology related questions. The 4-H mapping prototype served to illustrate the potential of providing agencies with a means for distributed data entry of GIS layers and attributes via the Internet. This national website, in turn, served as a prototype for the establishing CIRC.

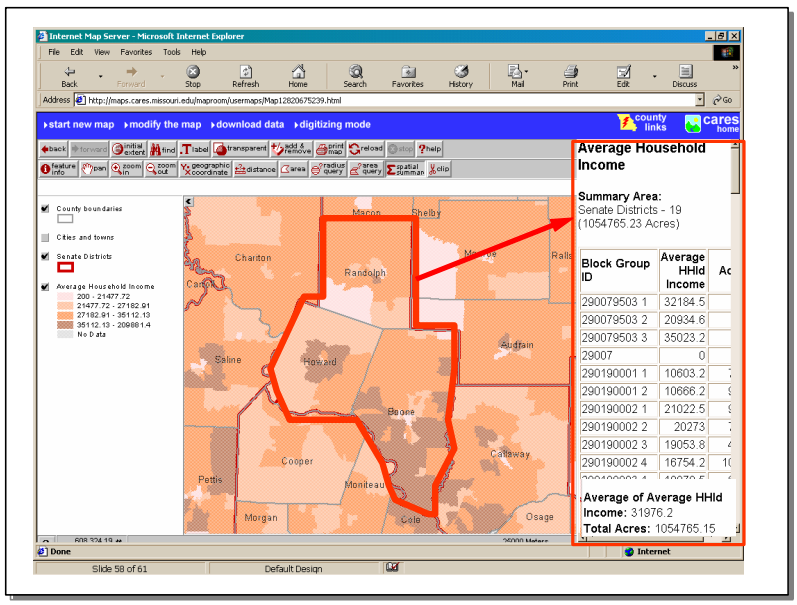
Spatial Summary Tool

A Spatial Summary Tool allows the user to summarize spatial data (e.g., land use) by any geographic feature (e.g., watershed boundary). Essentially, the Spatial Summary Tool serves as a cookie cutter that summarizes data by a specified geographic boundary (county, zip code, school district, congressional district, etc.).

Assume a State Senator wants to summarize “Average Household Income” by senate district, based on the 1990 census data. “Average Household Income” is aggregated to the senate district level by following this three step process:

- Step 1. Click on the “Spatial Summary” icon above the map and then click inside a given senate district boundary on the map.
- Step 2. Select a data layer to summarize: A popup menu opens up for selecting the data layer to summarize; in this case, “Average Household Income”.
- Step 3. Select a data layer to summarize by: A popup menu opens up for selecting the data layer to summarize by; in this case, Senate Districts.

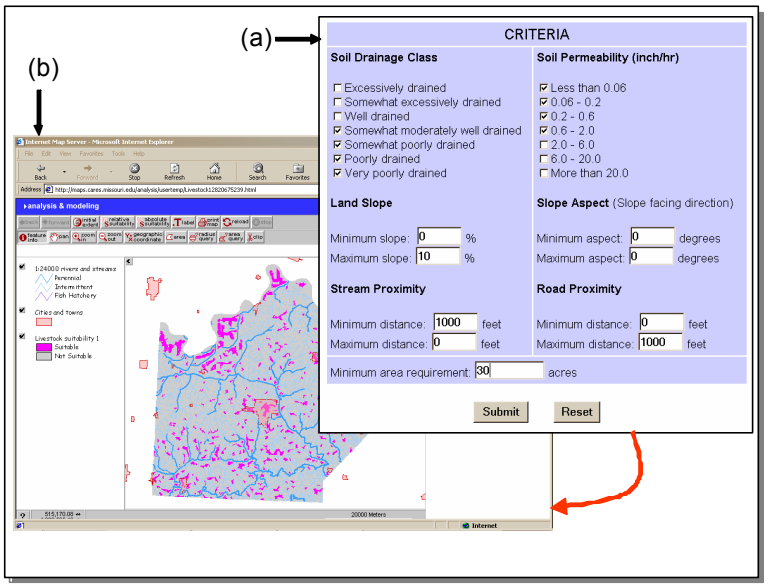
The Internet-based GIS calculates average household income In Senate District 19 to be $31,976 Based on the 1990 Census (see figure to right).



Internet-based Livestock Site Selection Tool

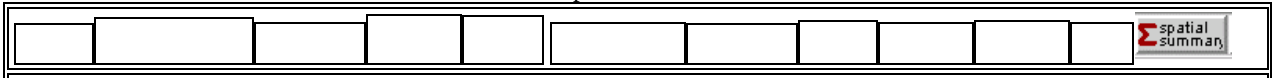
The impact of citizen participation in the local decision making process using Internet Mapping tools is highlighted by a study recently completed for Saline County, Missouri. County commissioners in Saline County approached the University of Missouri-Columbia for assistance in addressing the economic and environmental impacts of Confined Animal Feeding Operations (CAFOs) and their potential expansion in the county. This volatile issue was charged with emotion rather than sound information. Therefore, the University Outreach and Extension put together a team of researchers and Saline County extension staff to work with a citizen steering committee that reflected the varied interests in the county.

As stated earlier, the CAFO issue was volatile; however the site selection tool forced all members of the steering committee to focus on the list of criteria rather than relying on their emotions to drive the process. The first "what if" scenario was based on suitable livestock locations being: somewhat to very poorly drained soil; low in soil permeability; on land less than 10% slope; no closer than 1,000 feet from a stream; no farther than 1,000 feet from a road; and a minimum area of 30 acres (see figure to right).



After selecting values for each criterion, the "Submit" button was clicked to generate a livestock site suitability map for Saline County. The steering committee immediately saw that the City of Marshall was designated as a suitable location for a livestock operation. Both pro- and anti-CAFO steering committee members requested that CARES modify its list of criteria to include an urban setback option whereby a user can enter a minimum distance from an urban area. Other suggestions for adding to the criteria list included providing setbacks for tourist areas, state parks, and rural residences.

For brevity's sake the table below provides illustrations and definitions of additional tools developed at CARES that are located in the menu bar above the interactive map on the CARES Website.



- **Label:** Label features on the map.
- **Transparent:** Make the selected polygon data layer see-through. A popup menu opens for selecting the transparent data layer and level.
- **Add & Remove:** Add or remove data layers from the map.
- **Print Map:** Generate a map for printing out or saving to disk. Opens a separate browser window and displays a page with the current map display, a title, and the legend.
- **Help:** Open this help page.
- **Geographic Coordinate:** Click on a location to display the latitude / longitude and UTM coordinates.
- **Distance:** Draw a line to measure distance. Double click to end the line.
- **Area:** Used to draw a polygon to measure area and perimeter. Double click to end the polygon (planimeter).
- **Radius Query:** Click on a location to get information about features within a radius. A popup menu opens up for selecting the query data layer and entering radius.
- **Area Query:** Draw a polygon to get information about features within an area. A popup menu opens for selecting the query data layer.
- **Clip:** Draw a polygon to cut out features of data layers. Opens a popup menu for selecting the data layers to clip. The clipped data layers are listed as hyperlinks for download.
- **Spatial Summary:** Allows the user to summarize spatial data (i.e., census data) by any geographic feature.

CARES Website: www.cares.missouri.edu; CIRC Website: www.rupri.org/circ; e-mail: FulcherC@missouri.edu

Point Pattern Analysis in the ArcInfo 8.0 Environment

Arthur Getis and Jared Aldstadt

San Diego State University

For a number of years in the Department of Geography at San Diego State University, we have been working on the development of software for a variety of ESDA routines, especially in the field of spatial association. Arthur Getis has had a number of students work on the development of programs that make it possible to easily apply such analytical devices as nearest neighbor, general pattern, and local statistics. We have called our routine Point Pattern Analysis, but the actual product is constantly in flux as we attempt to make the approach as user friendly and comprehensive as we can. Students and former students who have worked on these programs over the years include Marc Armstrong, Laura Hungerford, DongMei Chen, Lauren Scott, and Jared Aldstadt.

The recent work of Aldstadt would be of interest to the community that will attend the Specialist Meeting. Currently he is adapting some of the materials of PPA into an ArcInfo 8.0 environment. He is using Visual Basic for Applications to move forward on the visual representation of several PPA routines. He is now able to demonstrate the use of certain local statistics in a fully functioning, uncoupled system. One of the features of this work is his user-friendly approach to the O statistic, a newly created statistic of Ord and Getis (*Journal of Regional Science* 2001, 41, 411-432) which identifies local clustering in a non-stationary setting.

To better understand what it is that we are attempting to do, I call your attention to our on-line software materials which represents an earlier approach to 14 different ESDA tools with documentation. It can be seen at:

http://xerxes.sph.umich.edu:2000/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi

Building on that work, Lauren Scott recently included in an ArcView 3.2 environment several of these routines. We used these successfully in exercises assigned to social science students and young faculty in

CSISS workshops on pattern analysis the last two summers. Now, we are moving further along by translating these routines into an ArcInfo 8.0 environment. Through visualization, we now show various attributes of patterns utilizing, point, line, and polygon coverages. One can see distance qualities in analysis by stages. This provides an outstanding tool for those wanting to learn about the relevant spatial statistics.

It should be mentioned that PPA has a substantial following. We have received over one hundred requests for the software. The software has been used in at least a dozen research projects which, we understand, have had results published. In addition, *SpaceStat*, the well-known spatial econometrics package, has a few of the PPA statistics included. Just recently, *ClusterSeer*, from BioMedware, the health-related software company, has adopted Aldstadt's approach for certain cluster analyses.

Getis and Aldstadt require funding to participate. It would be more important and useful for Aldstadt to attend than it would be for Getis.

VGE-A New Communication Platform for the General Public

Hui Lin, Jianhua Gong, Yibin Zhao, Zongyu Zhang, Jinyeu Tsou

Joint Laboratory for Geoinformation Science

The Chinese University of Hong Kong

Tel: (852)-2609-6528 Fax: (852)-2603-5006

Email: huilin @ cuhk.edu.hk

Homepage: http://www.jlgis.cuhk.edu.hk

Keywords: Virtual geographic environments (VGEs), public participation, GIS, virtual environments, VRML, Java, urban landscape assessment, wildfire behavior, simulation, 3-D avatars

Extensive Abstract

The Internet and the World Wide Web enable the public convenient access to geographic data, to easily implement spatial data analyses, to conduct model simulation and visualization, and to participate in resource management, environmental protection, regional planning and decision making.

In this paper, we introduce the concept of VGEs. VGEs are environments pertaining to the relationship between post-humans and 3-D virtual worlds. Post-humans are defined as a combination of humans in the real world with 3-D avatars in 3-D virtual worlds. VGE can be a platform for the public participation in spatial data analyses, in model computation and simulation, and in environmental planning and decision making. Three cases are presented for the public participation in terms of human-data, human-model, and human-human relationships. The first case is to study the web-based public participation in visual impact assessment of urban landscape through the application of spatial data analytical functions, especially the viewshed analysis, of distributed Arc/View GIS (see Figure 1). The second case focuses on the public participation in an ecological planning via a wildfire model-driven virtual studio on the Internet (see Figure 2a-2b). The last case is to discuss the communication and interaction among distributed multiple users in a 3D virtual shared space over the Internet via the experiment of a VRML/Java based virtual environment for the country parks in Hong Kong (see Figure 3).

Throughout the design, development, and applications of three system prototypes, we explore the key techniques to the integration of online-GIS, model geo-computation, 3D visualization, and distributed virtual environments for the public participation. We believe that the distributed, multi-user, geographic model-driven 3D virtual geographic environments are powerful platforms for the general public participation in urban and ecological planning and decision-making, and in supporting regional sustainable development across the Internet.

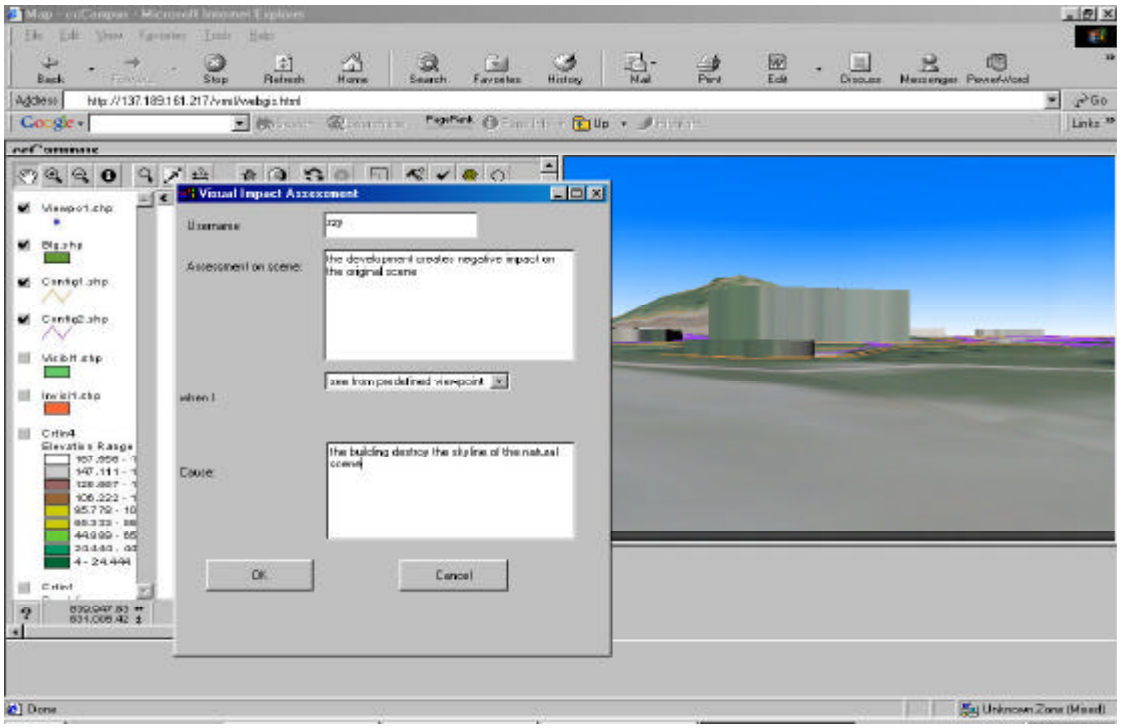


Figure 1. The interface of the prototype system for visual impact assessment

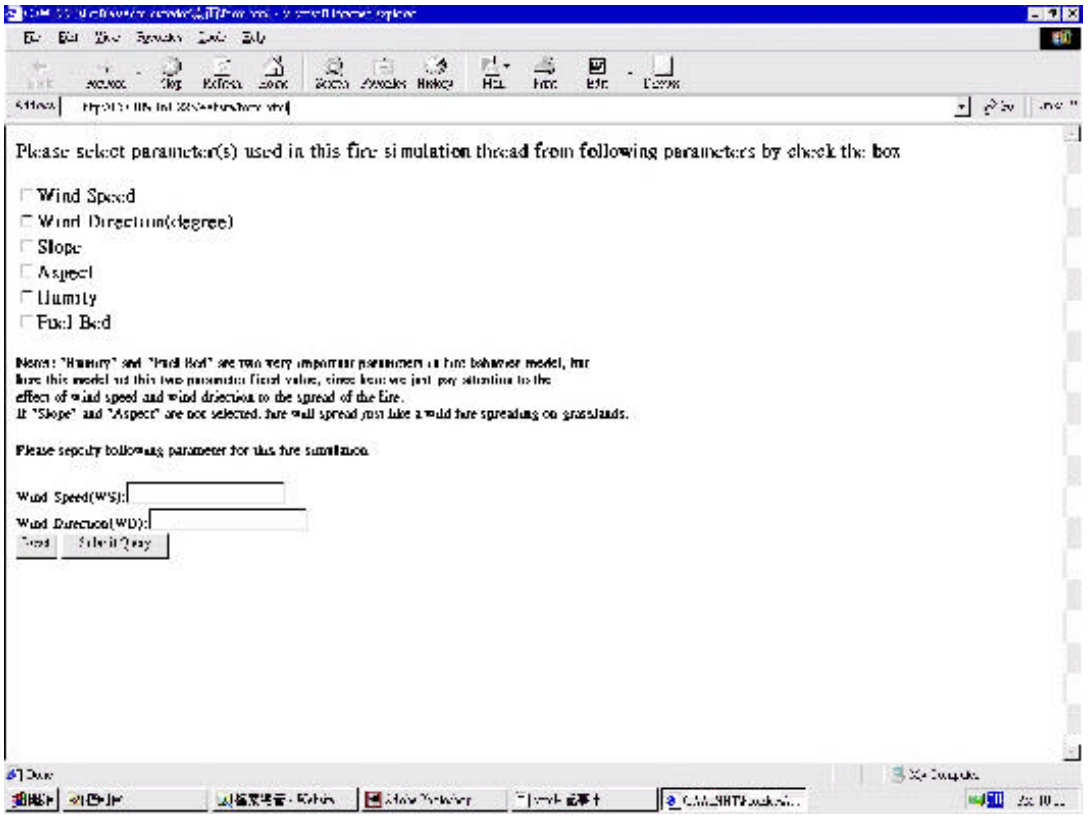


Figure 2a. Forest fire simulation runtime environment parameters initializing page.

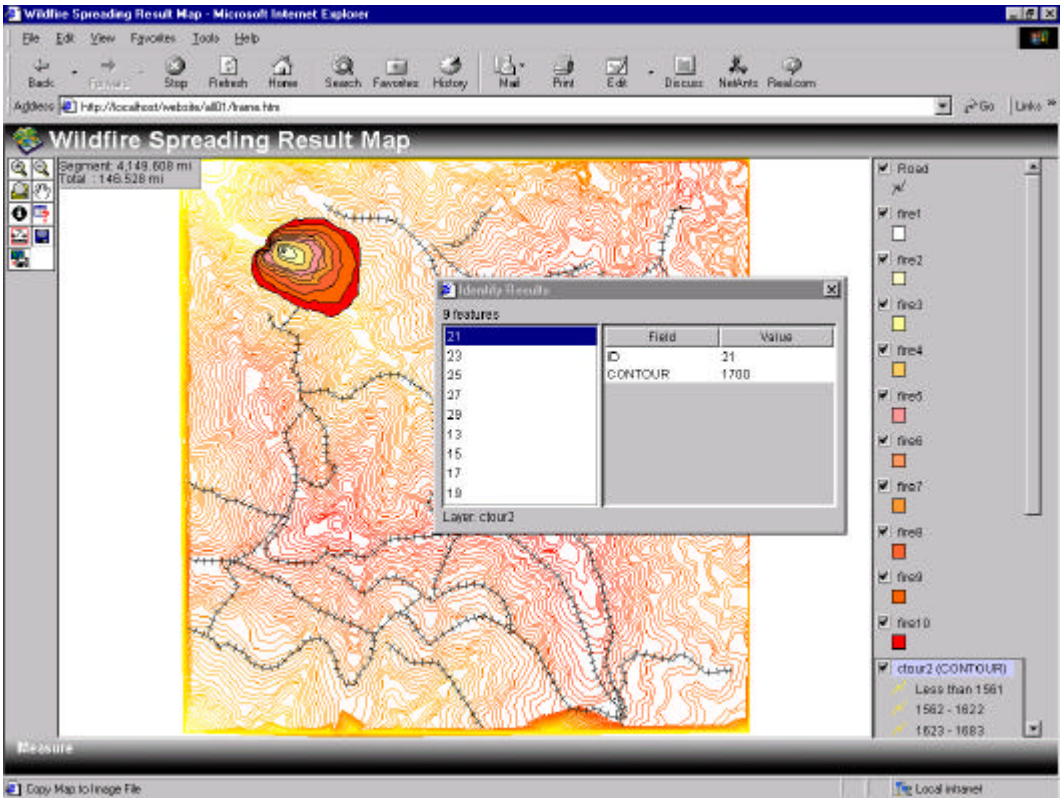


Figure 2b. Forest fire simulation result map displaying and feature query interface.

Figure 3. A distributed, 3-D, multi-user virtual environment, VirtualPark

Main References

Al-Kodmany, K., 1999. Using visualization techniques for enhancing public participation in planning and design: process, implementation, and evaluation, *Landscape and Urban Planning*, 45 :37-45.

Arthur, L.M., Daniel, T.C., and Boster, R. S., 1997. Scenic assessment: an overview, *Landscape Planning*, 4.

Lin, H., and Gong, J.H., 2001. Exploring Virtual Geographic Environments. *Geographic Information Sciences. Vol. 7, No.1, 1-7.*

Lin,H., Gong, J. H., and Wang, F., 1999. Web-Based Three-Dimensional Geo-Referenced Visualization, *Computers&Geosciences*, 25(10): 1173-1181.

Oh, K.,1994. A perceptual evaluation of computer-based landscape simulations, *Landscape and Urban Planning*, **28**.

Plewe, B., 1997. GIS-Online: Information Retrieval, Mapping, and the Internet, Santa Fe, NM, OnWord Press,215-252.

Rothermel, R.C., 1971. A Mathematical Model for Predicting Fire Spread in Wild Land Fuels. USDA Forest Service, Research Paper INT-115, Intermountain Forest and Range Experiment Station, Ogden, Utah.

Rohrer, R. M., and Swing, E., 1997. Web-based Information Visualization, *IEEE Computer Graphics and Applications,* July/August, 53-59.

Susskind, L. E., 1994. Director, MIT-Harvard Public Disputes Program. MIT, Cambridge, MA. from *Public Participation in Environmental Decision making*.

Wang, F. Y., 1998. Functions of country parks in the development of Hong Kong, in *Proceedings, China and the World in the 21st Century, an International Workshop on Geography* ,the Chinese University of Hong Kong, Hong Kong, China,17.

Yu, K., 1996. Security patterns and surface model in landscape ecological planning, *Landscape and Urban planning*, 36: 1-17.

A Set of GIS-based Tools for Spatial Structure Analysis

Bin Jiang
Division of Geomatics, Institutionen för Teknik
University of Gävle, SE-801 76 Gävle, Sweden
Email: bin.jiang@hig.se

Abstract. This paper presents a set of tools we developed with ArcView GIS for spatial structure analysis. The tools are based on the fundamental principles of space syntax (a set of methods for urban morphological studies), and integrated with a new approach. We review related applications mainly from space syntax research community to show how such a set of tools can be used to explain social functions of space and predict human spatial behaviour. Such a set of tools extends existing GIS analytical function for the social sciences.

Keywords: structure analysis, space syntax, social functions of space, and spatial analysis

1. Introduction

Space has a certain structure. The structure refers to that of an urban street network or an architectural layout in this context. Human moving behaviour (on foot or by car) in a space is affected to great extent by the spatial structure. Thus through an effective structure analysis we are able to predict human moving behaviour in a space. On the other hand, social function of a space is likely based on the effective structure analysis. Over the past two decades, space syntax (Hillier and Hanson 1094) has evolved into an important method for such a structure analysis with considerable amount of empirical studies applied to both architecture and city systems. Originated from seminal work in urban morphology (March and Steadman 1973, Steadman 1983), space syntax proposed a graph theoretic method for spatial structure analysis that has special applications in urban studies such as pedestrian flow modelling (Hillier et al. 1993), urban crime analysis (Jones and Fanek 1997, Shu 1999) and spatial cognition research (Kim 1999, Penn 2001). Therefore integration of such a method into GIS can improve GIS functionality in spatial analysis.

We have introduced our initial package named Axwoman elsewhere (Jiang et. al. 1999, Jiang et al. 2000), which consists of partial implementation. More recently we have proposed a new approach (Jiang and Claramunt 2002) to improve the method from a point of view of integrating such a structure analysis into GIS. Now we provide an improved version of the package, as we believe such a set of tools could have special applications in social science studies concerned about spaces and places. So the aim of this paper is to introduce the set of GIS-based tools we developed for spatial structure analysis, and furthermore to stimulate cross analysis with socio-economic data.

The remainder of this paper is organised as follows. Section 2 presents the fundamentals of space syntax on which the set of tools is based. Section 3 describes the details of implementation based on ArcView GIS, involving major components, functions and algorithms. To show how the tools can be applied to social sciences, section 4 presents a selective overview of applications for urban studies identified from space syntax research community. Finally section 5 concludes the paper.

2. Spatial structure analysis based on space syntax

A good way to understand structure analysis is the use of the concept of graph. A graph consists of two basic elements: nodes and links. How nodes are interconnected through links give a sense of structure of a graph (undirected and unweighted graph is adopted in this context). Within a graph, each node has different status from a structural point of view. There are two perspectives to look at the status of a node, namely local and global perspectives. Local perspective focuses on how each node connects to other nodes within a few steps, while global perspective on how each node connects to every other node within a graph.

Let's introduce some important parameters to measure the status of a node within a graph. For any particular node in the graph, the shortest distance far from the node is denoted by s (s is an integer), the number of nodes with the shortest distance s is denoted by N_s, and the maximum shortest distance (diameter of the graph) is denoted by l. Using the expression $\sum_{s=1}^{m} s \times N_s$, we can describe the following space syntax parameters* for the node:

$$\sum_{s=1}^{m} s \times N_s = \begin{cases} connectivity \quad iff \ \ m = 1 \\ local \ integration \quad iff \ \ m = k \\ global \ integration \quad iff \ \ m = l \end{cases}$$

Where k meets the condition $1 < k < l$.

Spatial structure analysis based on space syntax uses the graph view and apply it to different spatial situations. Let's take a look at how a space is actually organised with the situations. At an architectural level, a space is partitioned into (or perceived as) small units (or places) such as rooms and corridors. At a city level, there is no obvious partitions as at an architecture level, but often a street can be perceived as a form of vista spaces. These two observations lead to two basic approaches of space syntax, namely area-based and axial line-based approaches. Both approaches take the perceived units as nodes of a graph, i.e. rooms and corridors in the former case, and vista spaces in the latter case (figure 1). As for the links of the graph, it depends on the visibility among the perceived units. If a room has a door linking to a corridor, then the two perceived units are considered to be visible each other, and there is a link between the corresponding nodes. Or alternatively if two vista spaces are intersected each other, then they are considered to be visible as well, and a link between the corresponding node is needed.

* It should be noted that the integration here is a simplified one: Actual integration is based on a value of real rational asymmetry; refer to Jiang and Claramunt (2002) for details.

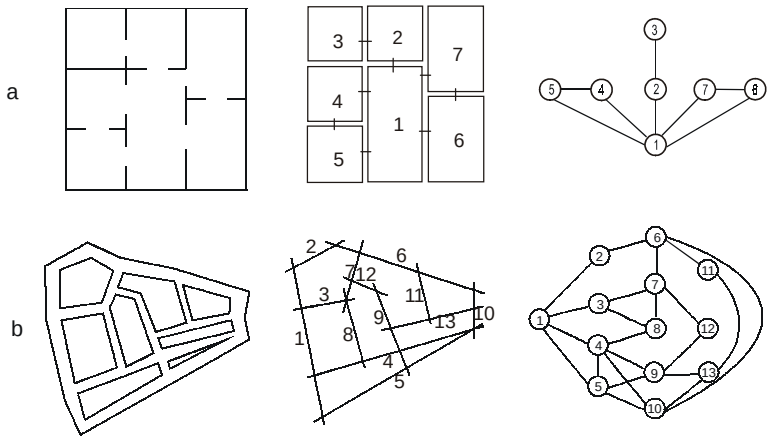


Figure 1: Illustration on area-based (a) and line-based (b) space syntax

It should be noted that the area-based approach, initially called convex representation (Hillier and Hanson 1984), could be applied to a city level as well. We argued that whether or not a perceived unit is convex is not so important, key point is that whether or not it is small enough to be perceived from a single vantage point of view. If it is perceivable from a single vantage point of view, then we can represent a place as a node, even though it is not convex. With the representations, a whole space consists of perceivable places, and the status of each place can be computed from the corresponding graph.

Let us take the example of the axial line identified as number one (equivalent to node one in the dual graph) in figure 1b. This line intersects four lines, so the connectivity of the axial line one is 4. The immediate neighbourhood of line one are lines 2, 3, 4, and 5, and their respective *connectivity* values are 2, 3, 5, and 4. Overall this axial line number one has 4 neighbourhoods one step away, 5 neighbourhoods two steps away, and 3 neighbourhoods three steps away (herein the concept of steps is equivalent to that of shortest distances). So *global integration* is equal to $4 \times 1 + 5 \times 2 + 3 \times 3$ if $k = 3$. If $k = 2$, then local depth with two steps away is equal to $4 \times 1 + 5 \times 2$.

A valid representation is said to be the least number of the largest axial lines needed to cover the entire space (Hillier and Hanson 1984). Otherwise, the structural analysis would be less meaningful, because the overall number of lines will not be representative of the spatial structure. So far the derivation of an axial map still relies on human judgement to draw individual lines, so no automatic solution has been identified, particularly for large cities, within the space syntax research community (c.f. Peponis et al. 1998). To overcome this problem, we have proposed a point-based approach (Jiang and Claramunt 2002) that is based characteristic points identified from a space. These points include road junctions and turning points (i.e., a turning point is defined as the peak of a curve). Based on how these points are visible, we can derive a visibility graph as shown in figure 2 for an example. Our experiments have shown that the approach is at least equivalent to the line-based approach in illustrating spatial structure. However the approach has several advantages over the line-based approach: the point representation is completely computable; and the structure analysis is more suitable for cross-analysis with socio-economic data that is mostly point-based.

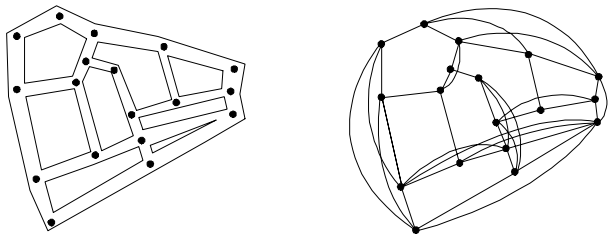


Figure 2: Characteristic points and visibility graph

3. The GIS-based tools for spatial structure analysis

The tools were implemented as an extension of ArcView GIS, thus provides an easy-to-use interface for the end users. ArcView is a desktop GIS with components including the view, table, chart, layout and script. Each of these serves different purposes for spatial data processing and presentation. Among others, Avenue script permits the customization of the ArcView interface, and provides some additional spatial functions. Avenue possesses important programming facilities such as lists and looping constructs for graphics manipulation, spatial queries, and basic arithmetic calculations. ArcView also has a set of extensions such as Spatial Analyst and 3D Analyst. The Spatial Analyst presents generic spatial analysis functionality on grid and feature themes such as proximity analysis, cell statistics and multi-map calculation, while the 3D Analyst allows users to create, analyze, and display surface data.

The GIS-based tools consist of three major components: identification of spatial units that could be point-, line- and area-based, computation of all structure parameters, and analysis of computing results and cross analysis with socio-economic data. The three components were implemented as toolbar as shown in figure 3. The identification of spatial units can rely on ArcView drawing tools manually, but for point-based approach the end users can directly import node file from ArcInfo and convert it from theme to graphics. Once the spatial units have been identified, users can compute all structure parameters and put them in an attribute table linked to a new created theme. With the analysis component, the users can explore data from different perspectives, or import observed data like pedestrian flow rates, and socio-economic data for cross analysis.

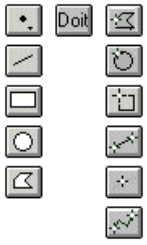


Figure 3: Toolbar with Axwoman3.0

Worth noting is the software's exploration capability. In order to conduct the structure analysis discussed above, a range of analytical components such as spatial units (namely axial maps, polygon maps, or characteristic points), tables, and charts are provided. All these components are dynamically linked to each other, so any action applied to one of the components will be propagated to any other. Figure 3 shows a typical interface.

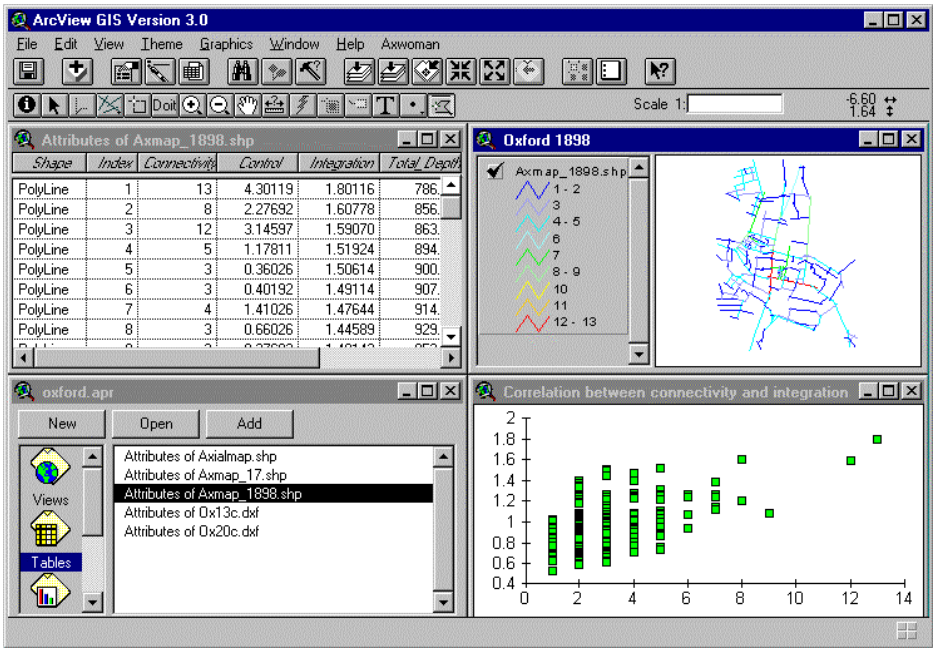


Figure 4: Axwoman3.0 extension to ArcView interface

Now let's take a detailed look to major algorithms with the implemented tools. The first algorithm is to derive a matrix of the graph on which the subsequent computation is based. A key point is to determine how spatial units are interconnected or visible each other, i.e. CONNECT(*v*,*w*) in the following algorithm. For the line-based approach it can be achieved through request *Intersect*, i.e. if one line intersects another line, then the two lines are interconnected. For the area-based approach, if two polygons are connected to a common line segment, they are interconnected. For the point-based approach, we adopt such a rule, i.e. one point is visible from another point if the line linking the two points does not intersect with any building polygons. The above process is time consuming, as it has to check for every spatial unit against every other. The algorithm can be described as follows:

Algorithm CREATE_MATRIX
// *V* is a set of spatial units, assume that there are totally *n* spatial units
// that are indexed from *0* to *n-1*
// *M* is the output adjacency matrix for the representation of the graph
begin
  $A \leftarrow [0\ldots n\text{-}1][0\ldots n\text{-}1]$ // the $n \times n$ zero matrix
  $V' \leftarrow V$ // target set of vertices
  for every $v \in V$ do
    for every $w \in V$ do
      if CONNECT(*v*,*w*) then
        $M[v][w] \leftarrow 1$
        $M[w][v] \leftarrow 1$ // the matrix is symmetric
      end if
    end for
  end for
end CREATE_MATRIX

Based on the above matrix, we can derive a range of structure parameters such as connectivity, local and global integrations. The integration computation is based on the algorithm for calculating the status of a node using Breadth-First Search (BFS) technique (Buckley and Harary 1990). For a connected graph, the algorithm begins at the first node and finds its neighbours, and then their neighbours, and so on until the algorithm has spanned throughout the whole graph and reached all nodes within the graph. Then this process continues with the second and third nodes until all nodes have been exhausted. Overall the tools were implemented as an extension and can be used together with some ArcView extensions. It complements spatial analyst and 3D analyst in spatial structure analysis.

4. Applications in urban studies

The spatial structure analysis can be used as both a design and an analysis tools for the built environments. As a design tool, it facilitates urban and architecture design to achieve best or optimal design scenarios. As an analysis tool, it can be used to determine social functions of space, and predict pedestrian and vehicle flows in a space. In this section, we focus on how the analysis results can be used to predict pedestrian flows, and crime analysis in complex built environments. By doing so, we hope to set up an image as to how the tools can be used for social science study, and diffuse the implemented tools within a GIS community.

The structure analysis is very important for understanding social functions of space. The title "the social logic of space" (Hillier and Hanson 1984) implies a logic or relationship between human activity and spatial structure. Let's take a look at a simple example as illustrated in figure 5. It is a floor from a building complex functioning now as an educational institution. All rooms and corridors are represented as geometric shapes, and a short line segment represents their possible connections with a door or entry. Through the structure analysis, we found that the long corridor is most integrated or connected and it tends to attract more people moving in it. Indeed, crosscheck in reality confirms this observation; both teachers and students are likely to move along the corridor. On the other hand, rooms coloured with the lighted red are most segregated places, which are being as teachers' offices. Other rooms and places with intermediate integration or segregation are used for other purposes. For example, the six parallel rooms are used for classrooms or labs. From the example, we can see that space has different functions in terms of how they are connected each other. Well-connected and well-integrated rooms are often used for meeting, while less connected and less integrated spaces for offices. Considerable amount of case studies have been carried out in this direction and they provide a good insight into urban and architecture design (c.f. Hanson 1998).

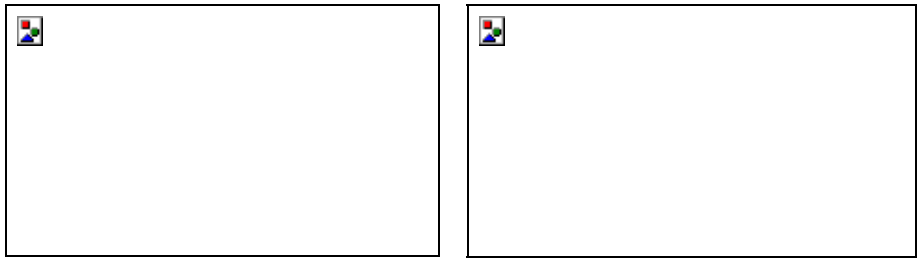


Figure 5: A building space (a) and its space syntax analysis result (b)

Prediction of pedestrian vehicle flows in urban systems is another important application aspect for the structure analysis. Some well established case studies have illustrated that local integration with three steps can be used to predict pedestrian flows (Hillier et al. 1993), i.e. pedestrian flow rate and local integration hold a significant correlation in a scatter plot. The same observation has been tested for vehicle flows as well. These findings about human movement behaviour are interesting, as it is based on spatial structure analysis without knowing individual behaviour. It should be noted that by movement behaviour we mean the average movement choice of a population, rather than individual ones.

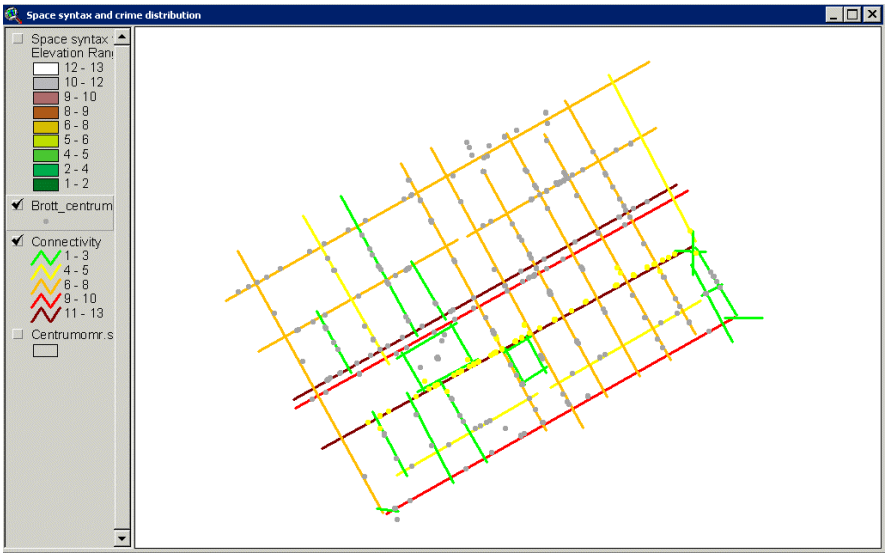


Figure 6: Crime distribution and street connectivity

Not only human movement, but also crime activities such as burglary in dwelling, criminal damage and car crimes of all kinds are affected by spatial structure. A basic hypothesis is that spatial property is one of important factor to determine spatial distribution of crime activities in urban system, i.e. crime activities are more likely happened in well-segregated areas (Jones and Fanek 1997, Shu 1999). In this respect, we have made some case studies as well using crime data collected from local police station. Due to confidential nature of the crime data, we could not make a detailed report here. However our study confirms the above hypothesis, i.e. high density crimes are located at well-integrated areas or streets. Figure 6 shows a crime distribution, where the majority of crimes are around the three most connected streets.

5. Summary and outlook

This paper has introduced the integrated tools and the basic principles behind them. Through a selective overview on the applications of the structure analysis in urban studies, we have shown how human spatial behaviour can be analysed by structural parameters. It provides a solid evidence to diffuse the structure analysis into social sciences that are particularly concerned about spaces and places.

References

Buckley F. and Harary F. (1990), *Distance in Graphs*, Addison-Wesley Publishing Company, Reading, MA.

Hanson J. (1998), *Decoding Homes and Houses*, Cambridge University Press.

Hillier B. (1996), *Space is the Machine: a Configurational Theory of Architecture*, Cambridge University Press, Cambridge.

Hillier B. and Hanson J. (1984), *The Social Logic of Space*, Cambridge University Press, Cambridge.

Hillier B. Penn A. Hanson J. Grajewski and Xu J. (1993), Natural Movement: Configuration and Attraction in Urban Pedestrian Movement, *Environment and Planning B*, Vol. 20, pp. 29-66.

Jiang B. and Claramunt C. (2002), Integration of Space Syntax into GIS: New Perspectives for Urban Morphology, *Transactions in GIS*, Blackwell Publishers, accepted for publication.

Jiang B., Claramunt C. and Batty M. (1999), Geometric Accessibility and Geographic Information: Extending Desktop GIS to Space Syntax, *Computers Environment and Urban Systems*, Vol. 23, pp. 127 – 146.

Jiang B., Claramunt C. and Klarqvist B. (2000), An Integration of Space Syntax into GIS for Modelling Urban Spaces, *International Journal of Applied Earth Observation and Geoinformation*, Vol. 2, No. 3, pp. 161 – 171.

Jones M. A. and Fanek M. F. (1997), Crime in the Urban Environment, in: Hillier B. (ed.), Proceedings of First International Symposium on Space Syntax, London, pp. 25.1-11.

Kim Y. O. (1999), *Spatial Configuration, Spatial Cognition and Spatial Behaviour: the role of architectural intelligibility in shaping spatial experience*, Ph.D thesis at University College London.

March L. and Steadman P. (1971), *The Geometry of Environment: An Introduction to Spatial Organisation in Design*, Methuen & Co Ltd: London.

Penn A. (2001), Space Syntax and Spatial Cognition, Peponis J. Wineman J. and Bafna S. (eds.) *Proceedings of the Third International Symposium on Space Syntax*, George, Atlanta, May 7-11, 2001, pp. 11.1-16.

Peponis J., Wineman J., Bafna S., Rashid M., and Kim S. H. (1998), On the Generation of Linear Representations of Spatial Configuration, *Environment and Planning B: Planning and Design* 25, 559 – 576.

Shu S. (1999) Housing layout and crime vulnerability, *Proceedings of Second International Symposium on Space Syntax*, 29 March-2April 1999, Universade de Brasilia, Brasilia, pp. 25.1 - 25.12.

Steadman P. (1983), *Architectural Morphology: An Introduction to the Geometry of Building Plans*, Pion: London.

Bridging the Gap Between GIS and Solid Spatial Statistics

Konstantin Krivoruchko, Environmental Systems Research Institute, 380 New York Street, Redlands, CA 92373-8100, kkrivoruchko@esri.com

Over the past few decades GIS and statistical spatial data analysis tools have been successfully applied in the fields of meteorology, environmental monitoring, ecological analysis, and mineral exploration (to name but a few). While these synergistic tools operate on spatial data, as yet they have not been incorporated into a single 'user-friendly' software environment. For this reason statistical methods remain somewhat of a mystery to mainstream GIS practitioners. To widen the use of spatial statistics, it is argued that software packages, such as SAS, SPLUS and GSLIB should incorporate three essential components that are provided by a GIS: a robust spatial database (with associated geographic coordinate systems), spatial models and visualization algorithms. An alternative solution is to incorporate statistical algorithms into GIS software. The *Geostatistical Analyst*, an extension to ESRI's ArcGIS 8.1, is an example of the latter approach. This represents a new level of integration of geostatistical techniques within a GIS framework. The software provides a simple windows-based interface for users that are unfamiliar with spatial statistics.

The Geostatistical Analyst software was released in May, 2001, and has two main components, namely, the Exploratory Spatial Data Analysis toolbox and the Interpolation and Statistical Modelling Wizard. The views in Exploratory Spatial Data Analysis tools are interactive with all of the other tools provided with ArcGIS. The algorithms and functions incorporated into the software make it a suitable processing tool for both the expert and novice users. In its most simplistic application users can select default values to create maps from point samples. As the level of knowledge improves, users are provided with a wide range of processing options to explore the properties of the data and hence create a more accurate map. The Geostatistical Analyst creates geostatistical layer, which naturally interacts with other GIS features and options, such as projection change, clipping, querying, exporting, etc.

The Geostatistical Analyst represents a major step in bridging the gap between GIS and geostatistics. Future software developments will focus on widening the range of statistical tools that are required by the GIS practitioners. This will be successfully achieved through consultation and assessment of the areas of need. It is hoped that such developments will widen the use of statistics in the GIS community and encourage statisticians to use GIS. Detailed information about the Geostatistical Analyst features can be found at <http://www.esri.com/software/arcgis/arcgisxtensions/geostatistical/index.html>.

Many enhancements to the Geostatistical Analyst extension to ArcGIS can be made, including the use of non-Euclidean metrics, conditional simulations, nonstationary models, models for locational errors, and space-time geostatistical models. Although users with strong statistical background are awaiting these additions, it is still unclear how many general GIS users are really interested in advanced geostatistical methods. The same is true in regards to lattice and point pattern analysis methods and tools.

Logistic Regression for Modeling Discrete Phenomena in a Continuous Surface: A Study of Crime Probabilities in Savannah City, GA

Naresh Kumar[1]

Abstract: It is easy to generalize natural phenomena on a continuous surface such as elevation, temperature, air pressure etc. even without their controlling variable(s). Socio-economic phenomena, however, are discrete in nature and difficult to generalize in a continuous surface without a number of controlling parameters such as land-use pattern, locational dependency etc. If sample data on controlling parameters are available a logistic regression model can be used to generalize a binary response variable in a continuous surface.

In this paper, I used logistic regression to model crime probabilities in relation to alcohol serving establishments in the city of Savannah, based on 12,458 crime spots and 247 alcohol-serving establishments for the year 2000. Crime probabilities, in this study, were controlled by two factors: land-use patterns and the distance of crime and random spots from the alcohol serving establishments. A similar approach can also be used to regress continuous surfaces for a number of other binary response variables.

This type of application(s) cannot be performed entirely in most of GIS software packages, and one needs to depend on a statistical package to derive the equation parameters. Therefore, there is need to integrate complex statistical models into GIS packages. I have developing an AML to perform the logistic regression and to estimate its probability in a user-friendly manner.

Keywords: Logistic Regression, Continuous Surface, Crime, Probabilities and GIS.

[1] Assistant Professor, Department of Geography and Planning, University of Toledo, Toledo, OH-43606.

The *CrimeStat* Program: Characteristics, Use, and Audience

Ned Levine, PhD
Ned Levine & Associates and Houston-Galveston Area Council
Houston, TX

In the paper and presentation, I will discuss the *CrimeStat* program, its potential uses, and the audience for whom it is intended. In addition, I will also comment on the need for transparency among statistical software, but will distinguish between clear documentation, open data and conversion standards, and open source code.

The *CrimeStat* Program

CrimeStat is a stand-alone spatial statistics program for the analysis of crime incident locations that can interface with most desktop GIS programs. It was developed by myself and Long Doan under research grants from the National Institute of Justice. The program is Windows-based and interfaces with most desktop GIS programs. The purpose is to provide supplemental statistical tools to aid law enforcement agencies and criminal justice researchers in their crime mapping efforts. The National Institute of Justice is the sole distributor of *CrimeStat* and makes it available for free to analysts, researchers, educators, and students.[1] The program includes a manual/textbook that describes each of the statistics and gives examples of their use. The manual is also available for download.

The program is written in C++ and is multi-threading. It will take advantage of multiple processors in a computer which, for a large data set, will considerably cut down on calculation time. The program also includes Dynamic Data Exchange code to allow another program to call up *CrimeStat* and pass the dataset name and variable parameters to it. Such a use was developed by the Criminal Division of the U.S. Department of Justice in developing the Regional Crime Analysis GIS (RCAGIS). That application used the Internet to link weekly crime databases from jurisdictions in the Baltimore metropolitan area to a common interface and set of analytical tools. *CrimeStat* was one of the tools.

CrimeStat is being used by many police departments around the country as well as by criminal justice and other researchers. From what we can tell, it has been used in many courses and has been a tool in a number of Masters and PhD theses. Three versions have been released. The first (1.0) was released in November 1999 and an update version was released in August 2000. The new version is 2.0 and will be released during the spring.

Data Input and Output

[1] The program is available at:

http://www.ojp.usdoj.gov/cmrc (under 'Mapping tools') or
http://www.icpsr.umich.edu/NACJD/crimestat.html

The program inputs incident locations (e.g., robbery locations) in 'dbf', 'shp' or ASCII formats using either spherical or projected coordinates. It can also treat zones as pseudo-points (or points with intensities). The program calculates various spatial statistics and writes graphical objects to ArcView®, ArcGis®, MapInfo®, Atlas*GIS™, Surfer® for Windows, ArcView Spatial Analyst©, as well as programs that follow the ODBC standard.

Program Sections

CrimeStat is organized into five sections:

Data Setup

1. **Primary file** - this is a file of incident or point locations with X and Y coordinates. The coordinate system can be either spherical (lat/lon) or projected. Intensity (Z) values and weight values are allowed. Each incident can have an associated time value.

2. **Secondary file** - this is an associated file of incident or point locations with X and Y coordinates. The coordinate system has to be the same as the primary file. Intensity and weight values are allowed. The secondary file is used for comparison with the primary file in the risk-adjusted nearest neighbor clustering routine and the duel kernel interpolation.

3. **Reference file** - this is a grid file that overlays the study area. Normally, it is a regular grid though irregular ones can be imported. *CrimeStat* can generate the grid if given the X and Y coordinates for the lower-left and upper-right corners. Several routines utilize the reference file.

4. **Measurement parameters** - this identifies the type of distance measurement (direct or indirect) to be used and specifies parameters for the area of the study region and the length of the street network.

Spatial Description

5. **Spatial distribution** - statistics for describing the spatial distribution of incidents, such as the mean center, center of minimum distance, standard deviational ellipse, Moran's I, Geary's C, or the directional mean.

6. **Distance analysis** - statistics for describing properties of distances between incidents including nearest neighbor analysis, linear nearest neighbor analysis, and Ripley's K statistic.

7. **'Hot spot' analysis I** - routines for conducting 'hot spot' analysis including the mode, the fuzzy mode, hierarchical nearest neighbor clustering, and risk-adjusted nearest neighbor hierarchical clustering.

8. **'Hot spot' analysis II** - more routines for conducting hot spot analysis including the Spatial and Temporal Analysis of Crime (STAC), K-means clustering, and Anselin's local Moran.

Spatial Modeling

9. **Interpolation** - a single-variable kernel density estimation routine for producing a surface or contour estimate of the density of incidents (e.g., burglaries) and a dual-variable kernel density estimation routine for comparing the density of incidents to the density of an underlying baseline (e.g., burglaries relative to the number of households).

10. **Journey to crime analysis** - a criminal justice method for estimating the likely location of a serial offender given the distribution of incidents and a model for travel distance.

11. **Space-time analysis** - a set of tools for analyzing clustering in time and in space. These include the Knox and Mantel indices, which look for the relationship between time and space, and the Correlated Walk Analysis module, which analyzes and predicts the behavior of a serial offender.

Options

12. Parameters can be saved and re-loaded.

13. Tab colors can be changed.

14. Monte Carlo simulation data can be output.

CrimeStat is accompanied by three sample data sets and a manual that gives the background behind the statistics and examples.

Audience

Any program has to have an audience to whom it is addressed. Crudely, a distinction can be made between four audiences for whom a statistical package would be appropriate, recognizing that in reality there are always mixtures of the four:

1. Statisticians
2. Researchers
3. Students
4. Analysts

CrimeStat was developed primarily for researchers and analysts, and secondarily for students. It was aimed at providing statistical tools to allow criminal justice researchers and analysts to quickly identify patterns in the distribution of incident locations. This is important in crime analysis. Thus, an emphasis is placed on pattern identification while

statistical significance is treated as a secondary issue.

The program does have a number of formal statistical tests and includes six different routines where a Monte Carlo simulation can produce an approximate statistical test. Nevertheless, the emphasis was placed on graphical representation that can be displayed on a GIS. The integration of *CrimeStat* with a GIS package is an essential part of the program. The underlying philosophy behind the program is as a GIS-based tool to help researchers and analysts in their work. It also differs from SAS, SPSS, S-Plus, and other statistical programs in that it has a detection and identification philosophy that underlies it. No attempt has been made to produce a comprehensive collection of tools. Instead, the tools that have been included were chosen because they are useful to crime analysts and criminal justice researchers (plus, hopefully, others).

Among the uses of the program are hot spot identification (i.e., clusters of crimes that concentrate), visualizing temporal shifts in spatial patterns, the identification of crime 'risk', and the analysis of serial events (e.g., a serial offender). There are many others uses for which the program can be, and has been, used, but those are the primary ones. For a police department, hot spot classification helps in the allocation of police officers. The recognition of temporal shifts in crime allows a more flexible police and community response. The identification of high risk areas is useful for crime prevention purposes while the analysis of serial events is important in apprehending dangerous offenders.

Transparency

Finally, I will provide some thoughts on transparency in statistical software development based on five years of experience in developing this software package.

Clear Documentation

I believe that open and clear documentation is essential for the development of knowledge by others. It's important for a software developer to provide as clear and comprehensive information as possible on the algorithms and methods used in the program. Only if users clearly understand what a routine does will they be able to properly use it. Too often, 'homemade' software has very cryptic documentation. The result is that it is more likely to be ignored than used.

Common Data and Conversion Standards

Similarly, I believe that common standards are essential for the widest development of GIS and spatial information systems. Common data standards allow data to be converted and passed from one program to another. Common program 'hooks' allow third-party developers and researchers to write new routines that can be used by a number of software programs. Ideally, if every statistical software developer or manufacturer used a common set of communication links, then small developers could see their routines used by a variety of users on different program platforms. Common standards can allow users to tailor their analysis to what they need to accomplish, knowing that they can take advantage of a variety of tools, rather than have to rely on those made available by a single producer. Finally,

from an economic perspective, common standards allow many producers to emerge and encourage creativity and innovation.

Open Source Code

On the other hand, I have doubts about the value of open source code. On the one hand, making code open could improve it since it would no longer just be the province of a single producer. But, it could also lead to abuse and breaches of security that could be dangerous to a wide range of users. Computer code, particularly low-level languages, can be very cryptic and personal. Different programmers have their own style and it may take a new programmer a long time to figure out that style. Variations on the original code could turn out to be wrong. In theory, if there are lots of people looking at the code, one would expect a greater 'wisdom' to emerge out of the collective efforts than if a single producer only controlled the code. On the other hand, many program applications, particularly statistical ones, have a small audience; there are few programmers with enough knowledge or motivation to dig into someone's code. If a third-party programmer makes modifications that turn out to be wrong, people may use the routine thinking that it's right and it may take a while to discover that it is not.

Currently, developers produce and own the existing code. The good ones will periodically fix the problems that emerge and update the program. This process can lead to trust by users. With open source code, that trust may not emerge. There are also some security concerns about open code. It is too easy to hide viruses, 'trojan horses', and other types of destructive agents within a cryptic code. Again, unsuspecting users may stumble upon these destructive processes and have damage done to their data or storage media.

In short, while I'm not rejecting the idea of open source code, I believe there are some serious issues that need to be addressed before we fully commit to such a course.

A Performance Comparison Between FRAGSTATS and APACK, Two Landscape Analysis Software Programs

Andrew Lister, Rachel Riemann, Mike Hoppus and James Westfall
USDA Forest Service, Northeastern Research Station, Newtown Square, PA
alister@fs.fed.us

FRAGSTATS has been an extremely popular software application for conducting landscape fragmentation analysis for nearly a decade. Developed as a shareware software utility for analyzing and quantifying landscape structure (McGarigal and Marks 1995), it has since been incorporated into other shareware as well as commercial landscape analysis programs. The vector implementation of the program was written as a suite of Arc/Info Arc Macro Language (AML) programs (ESRI, Redlands, CA 92373). The raster implementation was written in the C programming language, and is frequently compiled for DOS. There is also currently a newer, GUI-based version of FRAGSTATS produced by programmers at the University of Massachusetts available as shareware on the web.

APACK was created in the C programming language at the University of Wisconsin's Forest Landscape Ecology lab as a shareware tool for analyzing large raster datasets consisting of landscape information (e.g., classified satellite imagery). It calculates many or most of the landscape statistics available from FRAGSTATS and some additional ones (see Mladenoff and Dezonia (2001) for details).

The goal of the current project is to perform an empirical assessment of the capabilities of both programs, as well as to assess their efficiency in terms of processing speed. We will do this by setting up a series of 150 randomly chosen test images, using subsets of the North American Landscape Classification's (NALC's) Multi-Resource Land Classification (MRLC) datasets, which consist of classified Landsat TM imagery. The experimental design consists of 5 levels of image size (50, 100, 150, 200, and 250 m^2, each with 30 representatives). The null hypothesis of no difference in mean processing time across the levels of area will be evaluated using ANOVA. In addition, we will qualitatively compare the ease of parameter entry, the quality of the user manual, the usefulness of the output format, and some of the differences between the statistics calculated by both programs. Our goal is to provide the landscape ecology community with some quantitative as well as qualitative basis for choosing a specific landscape analysis software application.

Spatial Analysis of Alcohol-Related Motor Vehicle Crashes in Southeastern Michigan

Jaymie R. Meliker*[1], M.S., Ronald F. Maio[2,3], D.O., Marc A. Zimmerman[4], Ph.D., Hyungjin Mrya Kim[5,6], Sc.D., Sarah C. Smith[2,3], M.P.H., Mark L. Wilson[7,8], Sc.D.

[1] Dept of Environmental Health Sciences, Sch of Public Health, Univ of Michigan, Ann Arbor, MI; jmeliker@umich.edu

[2] Dept of Emergency Medicine, Univ of Michigan, Ann Arbor, MI

[3] Univ of Michigan Injury Research Center, Ann Arbor, MI

[4] Dept of Health Behavior and Health Education, Sch of Public Health, Univ of Michigan, Ann Arbor, MI

[5] Ctr for Statistical Consultation and Research, Univ of Michigan, Ann Arbor, MI

[6] Dept of Biostatistics, Sch of Public Health, Univ of Michigan, Ann Arbor, MI

[7] Dept of Epidemiology, Sch of Public Health, Univ of Michigan, Ann Arbor, MI

[8] Dept of Ecology and Evolutionary Biology, Univ of Michigan, Ann Arbor, MI

Abstract

Objective: Spatial Analysis was used to examine geographic patterns of alcohol-related motor vehicle crashes (MVC). **Methods**: Data were obtained from MVC drivers (833) presenting to two Emergency Departments (ED) from 1992-1994. MVC address was geocoded into the Geographic Information System using 1.5 mile square grid cells that represented the proportion of crashes with drivers with blood alcohol readings ≥ 100 mg/dL. **Results**: Besag & Newell's test and logistic regression indicated that areas of low population density have more alcohol-related MVCs than expected ($p < 0.05$). As the log of population density increased by 1 unit, the odds of alcohol-related MVCs decreased by 32% ($p = 0.004$). **Conclusions**: Within a two-county area of Southeastern Michigan, areas of lower population density were found to be associated with a higher proportion of alcohol-related MVCs.

PROBLEM UNDER STUDY: Motor vehicle crash (MVC) injuries represent the single largest component of injury mortality and injury costs in the United States[1]. Alcohol was associated with more than 300,000 MVC injuries in 1999, and 38% of MVC fatalities[2]. Epidemiological studies have characterized temporal, behavioral and social risk factors that increase alcohol-related MVCs. Much less is known about the spatial patterns and environmental associations of alcohol-related MVCs. This abstract demonstrates the application of new spatial analysis tools (ClusterSeer) to social science research questions.

OBJECTIVES: Spatial analysis and geographic information systems (GIS) provide a unique perspective for approaching traffic-related health problems and have the potential to generate hypotheses which otherwise may not be considered. Spatial analysis is utilized to determine if there is a geographic pattern of alcohol-related MVCs and to evaluate factors which may be associated with those MVCs.

METHODS: Subjects (833) were recruited from MVC drivers who presented to two EDs: a large University level 1 Trauma Center (April 1992-August 1994) and a large community teaching hospital (April 1993-June 1994). Address location of the MVC is geocoded into the GIS and population density information comes from 1990 census data. The study area is broken into a grid of 1.5 mile x 1.5 mile pixels which represent the proportion of crashes in which a driver had a BAC reading >= 100 out of the total crashes. Besag & Newell's spatial analysis, Cuzick and Edward's test, chi-squares, and logistic regression are performed to evaluate the association between population density and proportion of alcohol-related MVCs.

SPATIAL ANALYTIC METHODS: ClusterSeer ™ software (Terraseer, Ann Arbor, MI) was used to conduct Besag & Newell's test, which scanned the two-county area for clusters of grid cells exhibiting high proportions of alcohol-related MVCs. The null hypothesis was that the proportion of alcohol-related car crashes in each grid cell was similar to the proportion of alcohol-related car crashes in the entire study area (139/773 = 0.18). Inputs for the test statistic included the geographic coordinates, number of alcohol-related crashes and number of total crashes for each grid cell. In addition, the proportion of alcohol-related car crashes in the entire study area, and the number (**k**) of alcohol-related car crashes were also used. The geographic coordinates were calculated in the GIS; the value for **k** can range from 0-139. If the value for **k** is too small or too large, then the test has low power because few grid cells have zero or many alcohol-related crashes. An ideal choice for **k** is considered to be between 2 and 10, so we undertook this exploratory analysis with 9 replicates with **k** encompassing this range.
Stat! ™ software (Terraseer, Ann Arbor, MI) was used to conduct Cuzick & Edward's test which scans the two-county area to evaluate whether or not the nearest neighbor MVC of each alcohol-related crash was alcohol-related. The test was conducted on all 139 alcohol-related crashes, as well as on a systematic random sample of 139 of the 634 non-alcohol-related crashes. This sub-sample was chosen by sorting crash drivers according to age and gender, and then using a random number generator to determine a starting point for systematic selection of every 4.56^{th} item in the series (rounding up).

RESULTS: To test whether the spatial distribution of alcohol-related crash prevalence was clustered, Besag & Newell's Test was performed (Table 1). We used this statistic to search for nine different sized grid clusters ranging from two to 10 alcohol-related MVCs. From these nine analyses, 18 grid cells were identified as centers of local clusters ($p<0.05$). In 15 of the 18 grid cells, population density was smaller than the median value (151). These local clusters were

primarily in rural areas.

Table 1

Besag & Newell's Results: 18 Grid Cells Significant as the Center of Local Clusters				
Population per Square Mile in a Grid Cell	Total Crashes in a Grid Cell	Crashes with BAC>= 100 in a Grid Cell	Proportion of Alcohol-Related Car Crashes	# of Times Grid Cell was Significant
111	1	1	1	1
150	1	1	1	1
116	1	1	1	1
138	15	4	0.27	1
62	2	0	0	1
574	3	2	0.67	1
71	1	1	1	1
199	1	0	0	2
62	5	1	0.2	2
99	1	1	1	2
99	1	1	1	3
96	1	1	1	3
110	1	0	0	4
110	1	1	1	5
110	1	0	0	6
665	2	2	1	6
108	1	0	0	7
47	1	1	1	7

*median=152

**range=0-8003

Fifteen of the eighteen significant grid cells have less people per square mile than the median value (151). These 18 grid cells are identified in Figure 2.

Whereas Besag and Newell's Test evaluated the proportion of alcohol-related crashes in grid cells, we also tested for geographic proximity (i.e., nearest neighbor effects) of *individual crashes* using the Cuzick & Edward's Test but no proximity effects were detected. Thus, alcohol-related crashes seemed not to cluster in different areas than non-alcohol-related crashes. This suggested that the nearest neighbor of an alcohol-related crash was no more likely to be alcohol-related than to be non-alcohol-related.

The spatial analyses indicate that individual alcohol-related crashes do not occur in separate areas

from non-alcohol-related crashes. However, higher proportions of alcohol-related crashes in grid cells tend to cluster in rural areas.
Both chi-squares and binomial regression found a significant association between pixels with low population density and high proportion of alcohol-related MVCs. As the log of population density increased by 1 unit, the proportion of alcohol-related MVCs decreased by 0.39 ($P < 0.05$).

CONCLUSIONS: Within a two-county area of Southeastern Michigan, low population density was found to be associated with a high proportion of alcohol-related MVCs.

LIMITATIONS: The participating hospitals are situated in a relatively suburban setting, with patients coming from semi-urban and semi-rural areas. The findings may not be applicable to areas that are more strictly urban or strictly rural. This analysis involved only those who agreed to be tested, or whose next of kin agreed to testing, for BAC and there could be potential problems with selection bias; although it is difficult to imagine how compliance to participate would be associated with the population density of the MVC location. Since 7% of the MVC addresses did not match to the GIS, 60 MVC patients were not included in the analysis and it is possible that these 60 patients have some different characteristics from the other 773 patients. The zip codes of the 7% which did not match, however, appear to be spatially distributed in a similar manner to the 93% which did address match. Therefore, it may be safe to assume that there is no systematic bias in the addresses which are not matching to the GIS. Because not all drivers of a particular MVC were sent to the ED, it is possible that some alcohol-influenced drivers were missed by this study. According to the police reports, only 7 drivers of alcohol-related crashes did not test positive with BAC >= 100. These seven drivers may have been missed by the study or they may not have actually been alcohol-related, since the BAC test is a more reliable test than the police report. Potentially mis-assigning 7 drivers is a fairly minor misclassification, out of 773 total drivers and 139 alcohol-related drivers.

CONTRIBUTION OF THE PROJECT TO THE FIELD: This manuscript provides a unique contribution of spatial analytic techniques to a major preventable public health issue in several ways: (1) Uses state-of-the-art spatial analysis and mapping techniques to study alcohol-related motor vehicle crashes; (2) Demonstrates and discusses the benefits of integrating spatial and traditional analyses; and (3) Concludes that within an urban metropolitan statistical area, rural regions have higher proportions of alcohol-related car crashes, than do urban regions.

REFERENCES:

[1] U.S. Department of Health and Human Services. Healthy People 2010: Understanding and Improving Health. 2nd ed. Washington, DC: U.S. Government Printing Office, November 2000.

[2] Traffic Safety Facts 1999. Washington, DC: US Department of Transportation, National Highway Traffic Safety Administration, 1999, p.92.

A SPACE-TIME VERSION OF GEARY'S C (Abstract)

Myers D. E.[1]

1 Introduction

Geary's C function is a statistic used to detect spatial correlation in a data set. It is the quotient of two statistics, one of which is nearly the same as an omnidirectional sample variogram, the other is simply the sample variance. Unlike Geary's C, the sample variogram is used primarily as an estimator of the theoretical variogram which quantifies dis-similarity for a random function as a function of the separation vector. However the sample variogram does not directly estimate the variogram as a function but rather only estimates values of the variogram for a finite number of choices of the separation vector. It remains to select a valid variogram model (or models in the case of a nested structure), then to "fit" the sample variogram to the selected model, this may be done in several ways. To be a valid variogram model, the function must satisfy several conditions and the most important one is conditional negative definiteness. In practice one chooses from a small list of known valid models, each with some number of parameters to be fitted. The problem is that these are all isotropic models, i.e., functions which depend only on distance. Directional dependence can be incorporated via an affine transformation. However when extending from the spatial context to space-time, there is the problem of no natural choice of a distance function on space-time and at least to some degree space and time must be considered separately.

2 Product-Sum Models

Unlike Geary's C or the sample variogram which need not satisfy any theoretical conditions, the variogram must be conditionally negative definite and when used in the kriging equations,

[1]Dept of Mathematics, University of Arizona, 85721 Tucson AZ - USA; e-mail: myers@math.arizona.edu

i.e., for interpolation, it must be strictly definite. While the sum of a spatial variogram and a time variogram, i.e., a construction of the form

$$\gamma_{st}(h_s, h_t) = \gamma_1(h_s) + \gamma_2(h_t)$$

will be conditionally negative definite if $\gamma_1(h_s)$, $\gamma_2(h_t)$ are valid variogram models in space and time respectively,$\gamma_{st}(h_s, h_t)$ will not be strictly definite as shown by Myers and Journel (1990). The product of two variograms will generally not satisfy the growth condition, i.e., increase less rapidly than quadratic as a function of distance. However, the product of two covariances will be a valid covariance and more generally

$$C_{st}(h_s, h_t) = k_1 C_s(h_s) C_t(h_t) + k_2 C_s(h_s) + k_3 C_t(h_t)$$

will be a valid space-time covariance provided that $k_1 > 0, k_2 \geq 0, k_3 \geq 0$. In turn this can be re-written in the form

$$\gamma_{st}(h_s, h_t) = \gamma_{st}(h_s, 0) + \gamma_{st}(0, h_t) - K\gamma_{st}(h_s, 0) \times \gamma_{st}(0, h_t)$$

where $\gamma_{st}(h_s, h_t)$ is the variogram corresponding to $C_{st}(h_s, h_t)$.

$$\gamma_{st}(h_s, 0), \gamma_{st}(0, h_t)$$

can be estimated from data separately. That is,$\gamma_{st}(h_s, 0)$ will be estimated by averaging over time the sample space variograms for each time data point and $\gamma_{st}(0, h_t)$ can be estimated by averaging over space the sample time variograms for each spatial data point. This then is the motivation for defining a space-time version of Geary's C.

3 Space-Time Geary's C

Let s denote a point in space and t a point in time so that (s, t) denotes a point in space-time. Of course the crucial aspect of space time data is that in general for any given time there may be many points with different spatial coordinates and conversely for any given spatial location there can be many different times. For convenience in describing the proposed Space-Time Geary's C the data will be indexed in two different ways. Let the data values in general be denoted by $Z(s, t)$. Note that while it is convenient to use the two different indexing schemes when describing the formulas it is not necessary when coding the algorithm.

3.1 Spatial Component

Suppose that there are n distinct time data points, there are m_k spatial data points associated with each of these time points $k = 1, \ldots, n$. Define

$$C_s = (1/n)\Sigma_{k=1}^{n}(1/2N(t_k))\Sigma_{i,j=1}^{m_k}[Z(s_i, t_k) - Z(s_j, t_k)]^2 \quad (1)$$

$N(t_k)$ is the number of pairs as is usual in the definition of the variogram.

3.2 Temporal Component

Suppose that there are p distinct data points in space and there are q_k time points associated with each of these, $k = 1, \ldots, p$. Define

$$C_t = (1/p)\Sigma_{k=1}^{p}(1/2N(s_k))\Sigma_{i,j=1}^{q_k}[Z(s_k, t_i) - Z(s_k, t_j)]^2 \quad (2)$$

$N(s_k)$ is the number of pairs as is usual in the definition of the variogram.

3.3 The Variance

The space-time sample variance can be computed using either indexing scheme

$$S_2 = (1/N)\Sigma_{k=1}^{n}\Sigma_{i=1}^{m_k}[Z(s_i, t_k)]^2 \quad (3)$$

where

$$N = \Sigma_{k=1}^{n} m_k$$

.

$$S_1 = (1/N)\Sigma_{k=1}^{n}\Sigma_{i=1}^{m_k}[Z(s_i, t_k)] \quad (4)$$

where

$$N = \Sigma_{k=1}^{n} m_k$$

. Then the sample variance becomes

$$S^2 = S_2 - [S_1]^2$$

3.4 The Space-Time Coefficient

Define

$$C_S T = [C_s + C_t - KC_s \times C_t]/S^2$$

where $K = 1/[maxC_s, C_t]$. C_s is the analogue of $\gamma_{st}(h_s, 0)$ and C_t is the analogue of $\gamma_{st}(0, h_t)$.

4 Software

De Cesare, Myers and Posa (2002) have given a program for computing $\gamma_{st}(h_s, 0)$ and $\gamma_{st}(0, h_t)$. This will be easily modifiable to compute C_{ST}.

5 REFERENCES

- 2001, L. De Cesare, D.E. Myers and D. Posa, Estimating and Modeling Space-Time Correlation Structures, Statistics and Probability Letters 51/1, 9-14
- 2002, L. De Cesare, D.E. Myers and D. Posa, FORTRAN Programs for Space-Time Modeling. to appear in Computers Geosciences
- 2000, S. De Iaco, D.E. Myers and D. Posa, Space-time analysis using a general product-sum model. Statistics and Probability Letters 52, 1, 21-28.
- 2000, S. De Iaco, D.E. Myers and D. Posa, Total Air Pollution and Space-Time Modeling. in Monestiez P., Allard D. and Froidevaux R., Eds, geoENV III, Geostatistics for Environmental Applications, Kluwer Academic Publ., 45-56
- 2002 S. De Iaco, D.E. Myers and D. Posa, Nonseparable space-time covariance models:some parametric families. to appear in Mathematical Geology
- 1990, D .E. Myers and A. Journel, Variograms with Zonal Anisotropies and Non-Invertible Kriging Systems. Math. Geology 22, 779-785
- 2002, D.E. Myers, Space-time correlation models and contaminant plumes. to appear in Environmetrics (papers from the Fourth Int. Conference on Environmetrics and Chemometrics, Las Vegas, NV Sept. 2000)
- 2002, D.E. Myers, S. De Iaco, D. Posa and L. De Cesare, Space-Time Radial Basis Functions, to appear in a special issue of Computers and Math. Applications

COMPOSITION AND DECOMPOSITION OF THE WEIGHTED VORONOI DIAGRAM

Mu, Lan and Radke, John

Authors:

Mu, Lan
PhD Candidate
Geographic Information Science Center
University of California, Berkeley
102 Wheeler Hall
Berkeley, CA 94720-1870
(510) 642-8641
Email: mulan@gisc.berkeley.edu

Radke, John, PhD
Director and Associate Professor
Geographic Information Science Center
University of California, Berkeley
102 Wheeler Hall
Berkeley, CA 94720-1870
(510) 643-5995
Email: ratt@gisc.berkeley.edu

Abstract:

Weighted Voronoi diagrams enhance the strictly location criteria of ordinary planar Voronoi diagrams by considering weights associated with sites understudy. These weights are most often derived from what is commonly referred to as attributes, data stored in tables within a Geographic Information System. Based on different underlying processes, weighted voronoi diagrams can be classified into: multiplicatively weighted, additively weighted, compoundly weighted, power, and sectional diagrams (Okabe, Boots, Sugihara and Chiu, 2000). The purpose of this paper is to present and demonstrate algorithms that compose and decompose the multiplicatively weighted Voronoi diagram, constructed from point data defined as:

$$\text{region}(p) = \{x \mid d_w(x, p) \leq d_w(x, q), q \text{ in } S\},$$

where the weighted distance $d_w(x, p)$ is the Euclidean distance $d_e(x, p)$ divided by the weight: $d_w(x, p) = d_e(x, p) / w(p)$ (Aurenhammer and Edelsbrunner 1984).

Our motivation and purpose to compose and decompose is to characterize landscape, record and detect environmental change over time. We develop a program "**WVD**" in Visual Basic, which distinguishes itself from other approaches found in the literature. The concepts and algorithms we propose are based in the field of Geographic Information Science, the implementation process is straightforward for the transition from a model environment (virtual space) to the real world (physical space), it supports an easy exchange with common GIS software, and it serves as a teaching tool helping to vision the abstract model.

We target and overcome obstacles to using weighted Voronoi diagrams which include weak links between theory and operational models, and the lack of off-the-shelf tools in GIS to generate such diagrams. We develop two methods to construct the multiplicatively weighted Voronoi diagram: 1) a *growth simulation* model, and 2) a *vertex calculation* method.

Figure 1 illustrates the graphic user interface (GUI) of the program **WVD**:

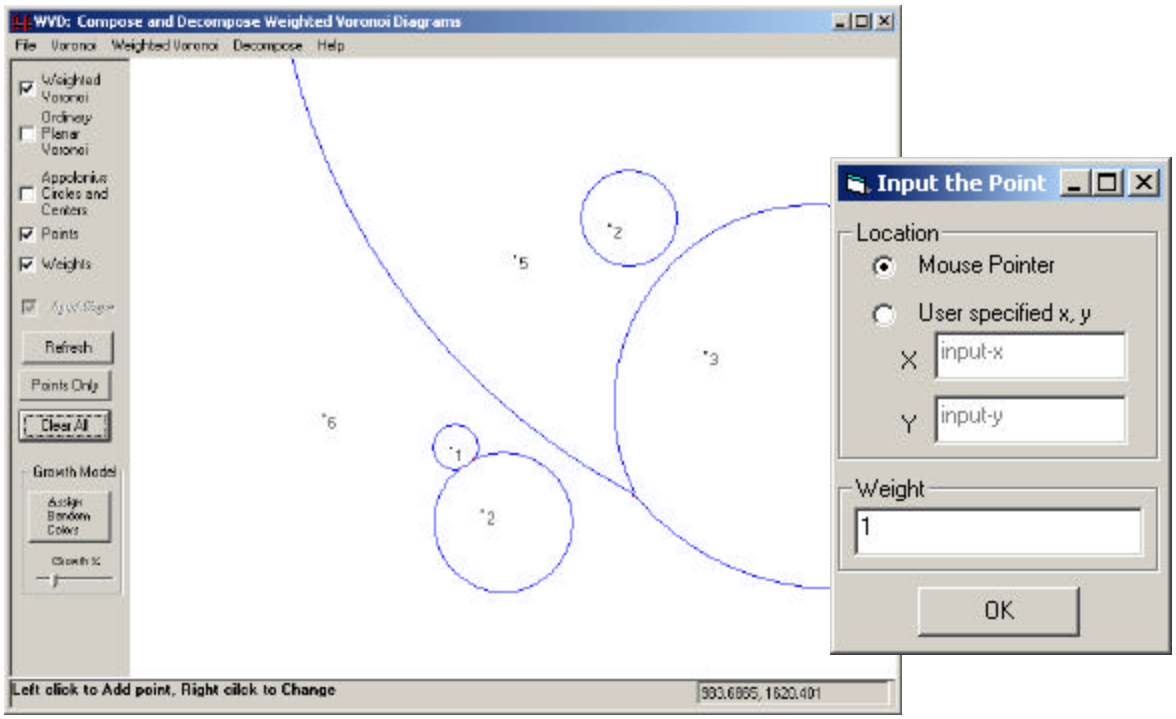


Figure 1 **WVD** Interface

The main features of **WVD** :

1) inputs of point locations and weights can be either interactive using a mouse (Figure1) (freehand or specify x, y coordinates) or read from a text file in a common format (weight, x, y);
2) the points can be easily added, moved or removed, and the weights can be interactively modified which allows continuous tuning of the model;
3) features such as weighted Voronoi, ordinary Voronoi, Appolonius circles, weights, and points can be easily switched on or off;
4) the extent of growth can be controlled (for growth model only);
5) the points and weights can be output as text files, and the weighted Voronoi and ordinary Voronoi (vector) diagrams can be output as an AML file for easy input into an Arc/Info coverage;
6) raster images can be loaded into the background to facilitate the composing and decomposing process;
7) vector data (in ESRI ungenerate format) can be input to approximate polygons as circular arcs and line segments after which they can be decomposed with points and weights following the weighted Voronoi process;
8) the decomposed results can be output as text tables and provide a highly compressed and alternative method for representing polygons as points and weights;
9) batch processing is accommodated for multiple inputs, composing the weighted Voronoi and outputting the results as AMLs.

The **growth simulation method** dynamically simulates the weighted Voronoi diagram on the user's computer screen. It visually allows one to observe how points compete with one another and partitions the space following the weighted Voronoi model. The **vertex calculation method** uses the geometric properties and relations in the Apollonius circles, their generator point weights and locations, their interaction and vertex points to calculate and determine valid vertices and edges (arcs and lines) in the final diagram. Although this method is mathematically more complicated it is computationally less intensive, is faster, can be monitored during the process, which provides an instructional visualization aid, and is easily modified.

The decomposition feature is another powerful tool in this program. Given the boundary shape of a 2D object, we decompose it to a set of points with weights associated with each segment. The properties and rules obtained from the composition process play a key role in this feature. The result of this process helps one understand the spatial relationships in terms of form and process.

Reference:

Aurenhammer, F., and H. Edelsbrunner. (1984). "An Optimal Algorithm for Constructing the Weighted Voronoi Diagram in the Plane." Pattern Recognition **17**(2): 251-257.

Okabe, A., B. Boots, K. Sugihara, and S. Chiu. (2000). Spatial Tessellations, Concepts and Applications of Voronoi Diagrams. New York, John Wiley & Sons Ltd.

SOFTWARE FOR SPATIAL ANALYSIS ON A NETWORK

Atsuyuki Okabe*, Keiichi Okunuki**, Shino Funamoto*** and Kimie Okano****
* Center for Spatial Information Science, University of Tokyo
** Department of Geography, Nagoya University
*** Department of Urban Engineering, University of Tokyo
**** Mathematical Systems Inc.

This software is designed as an analytical toolbox for spatial analyses on a network. The software is implemented with an extension of ArcView8.X, and it will be open to public without charge in 2002. The software is developed under the project entitled by Spatial Information Science for Human and Social Sciences (SISforHSS) supported by the Ministry of Education and Science, Japan.

(1)**Potential Demand for Spatial Analysis on a Network**

Recently locational competition of retail stores in a densely inhabited region becomes very hard. For instance, in the central region of Tokyo (23 Wards; 621 squared km with 8 million inhabitants), almost five thousands convenience stores are competing their locations. A distinctive feature of this type of stores is small market areas. Because of this nature, marketing is concerned with microscopic geographical factors with a street network. Such marketing is called micro-marketing. GIS provide a powerful tool for micro-marketing, but a problem with that is poor analytical tools. The most traditional analytical tools are based upon the assumption that the market areas are homogeneous plane, and distance is measured in terms of the Euclidean distance. In a small area, however, irregular streets produce a heterogeneous plane and consumers access to stores through a street network. This suggests that there be great potential demand for analytical tools for micro-spatial analysis on a network where distance is measured in terms of the shortest-path distance.

(2)The Network Space

We suppose that the real world is represented by a network space formed by connected line segments, and that every geographical objects are

assigned on the network (Figure 1). For instance, a store is assigned on a point on the network in terms of its access point, i.e. the gate or the entrance of the store. A house of a consumer is also assigned to a point on the network in the same manner. Thus the distributions of stores and consumers are represented by the distributions of points on the network. A line segment of the network may have attribute values, such as the width of a street, and a traffic volume.

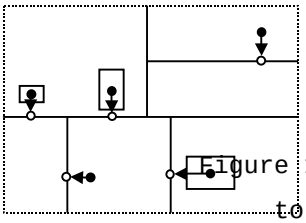


Figure 1: Assignment of (representative) points to the points on a network

(3)Functions of the Software

This software has the following functions.

i) Generate a point for a polygon

This function is to generate a representative point for a polygon representing a geographical object (such as a house) (see the black circles in Figure 1).

ii)Assign a point to a point on a network

This function is to assign a point that is not on a network to the nearest point on the network (see the white circles in Figure 1).

iii) Generate random points on a network

This function is to generate points on a network according to the Poisson point process on the network (i.e. a point being placed on a unit line segment on the network is the same regardless of the location of the segment on the network). This function may be used for Monte Carlo simulations.

iv) Generate a Voronoi diagram on a network

This function is to construct a Voronoi diagram generated by a set of points on a network (Okabe, Boots, Sugihara and Chiu, 2000).

v) Test by the cell count method

This function is to test randomness of points distributed on a network by use of the cell count method as an extension of the quadrat count method on a plane with Euclidean distance.

vi)Test by the nearest neighbor distance method

This function is to test randomness of points distributed on a network by use of the nearest neighbor distance method as an extension of the nearest distance neighbor distance method on a plane with Euclidean distance (Okabe, Yomono and Kitamuara, 1995).

vii) Test by the conditional nearest neighbor distance method

This function is to test if points (of Type A, say convenience stores) are independently and randomly distributed with respect to a set of fixed points (of Type B, say stations) (Okabe and Miki, 1984).

viii) Test by the K-function method

This function is to test randomness of points distributed on a network by use of the K-function method as an extension of the K-function method on a plane with Euclidean distance (Okabe and Yamada, 2001)

ix)Test by the cross K-function method

This function is to test if points (of Type A) are independently and randomly distributed with respect to a set of fixed points (of Type B) (Okabe and Yamada, 2001).

x) Estimate market areas using the Huff model

This function is to estimate the choice probability of choosing stores when consumers' behavior is described by the Huff model (Okabe and Kitamura, 1996).

xi)Estimate the demand for a store using the Huff model

This function is to estimate the demand of a store when consumers' behavior is described by the Huff model (Okabe and Okunuki, 2001).

(4)Implementation

Since the environment for developing programs with ArcView8.X is not ready at present, we have not yet developed a user-friendly interface, but we are now developing the fundamental programs with Visual Basic and Visual C++. Part of the outcome is shown in Figure 2, where the estimated demand for a store is indicated by colored marks.

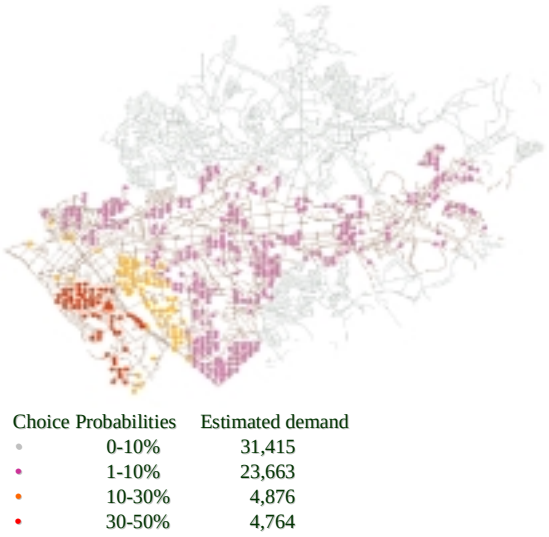


Figure 2: Demand estimation on a network with the Huff model

References

Okabe, A. (2000) "Spatial Analysis on a Network with GIS" presented at the International Conference on Geographic Information Science October 28-31, 2000. Savannah, Georgia, USA.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S. N. (2000) Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, Chichester: John Wiley.

Okabe, A. and Kitamura, M., (1996) "A Computational Method for Market Area Analysis on a Network", *Geographical Analysis,* Vol.28, No.4, pp.330-349.

Okabe, A. and Miki, F. (1984) "A Conditional Nearest-Neighbor Spatial Association Measure for the Analysis of Conditional Locational Interdependence", *Environment and Planning A*, Vol.16, pp.163-171.

Okabe, A. and Okunuki, K. (2001) "A computational method for estimating the demand of retail stores on a street network and its implementation in GIS", *Transactions in GIS,* Vol.5, No.3, pp.209-220.

Okabe, A. and Yamada, I.(2001) "The K-function method on a network and its computational implementation" *Geographical Analysis,* Vol.33, No.3, pp.271-290.

Okabe, A., Yomono, H. and Kitamura, M., (1995) "Statistical Analysis

of the Distribution of Points on a Network", *Geographical Analysis*, Vol.27, No. 2, pp.152-175.

Spatial Data Mining Research by the Spatial Database Research Group, University of Minnesota

Shashi Shekhar and Ranga Raju Vatsavai
Spatial Database Research Group
Department of Computer Science and Engineering
EE/CS 4-192, 200 Union Street, SE., Minneapolis, MN 55455.
[*shekhar*|*vatsavai*]@cs.umn.edu
http://www.cs.umn.edu/research/shashi-group/

Abstract

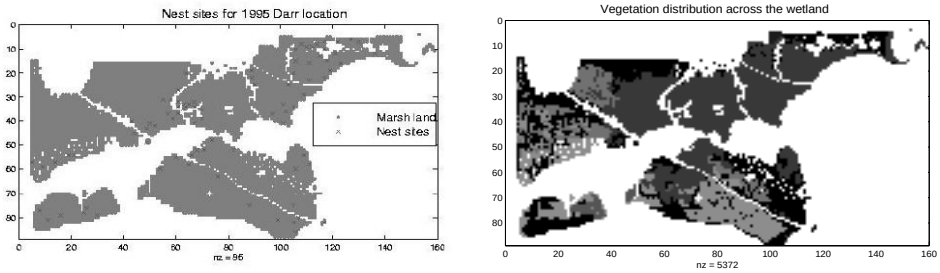
Explosive growth in geospatial data and the emergence of new spatial technologies emphasize the need for the automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. In this paper we describe the ongoing spatial data mining research by the Spatial Database Research Group, University of Minnesota. We discuss several computationally efficient and scalable techniques for analyzing large geospatial data sets and their applications in location prediction, spatial outliers detection and co-location association rules mining.

Keywords: co-location mining, spatial outliers, spatial context, SAR, MRF

1 Introduction

Researchers in the Spatial Database Research Group [22], University of Minnesota, have recently focussed their reserach in the field of spatial data mining, a field whose importance is growing with the increasing incidence and importance of large geo-spatial datasets such as maps, repositories of remote-sensing images, and the decennial census. Applications of spatial data mining can be found in location-based services in the M(mobile)-commerce industry, in the military (inferring enemy tactics such as Flank attack), at NASA (studying the climatological effects of El Nino, land-use classification and global change using satellite imagery), at the National Institure of Health (predicting the spread of disease), at the National Imagery and Mapping Agency (creating high resolution three-dimensional maps from satellite imagery), at the National Institute of Justice (finding crime hot spots), and in transportation agencies (detecting local instability in traffic).

The differences between classical and spatial data mining are similar to the differences between classical and spatial statistics. First, spatial data is embedded in a continuous space, whereas classical datasets are often discrete. Second, spatial patterns are often local whereas classical data mining techniques often focus on global patterns. Finally, one of the common assumptions in classical statistical analysis is that data samples are independently generated. When it comes to the analysis of spatial data, however, the assumption about the independence of samples is generally false because spatial data tends to be highly self correlated. For example, people with similar



(a) Nest Locations (b) Vegetation Durability

Figure 1: (a) Learning dataset: The geometry of the wetland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the marshland

characteristics, occupation, and background tend to cluster together in the same neighborhoods. In spatial statistics this tendency is called spatial autocorrelation. Ignoring spatial autocorrelation when analyzing data with spatial characteristics may produce hypotheses or models that are inaccurate or inconsistent with the dataset. Thus classical data mining algorithms often perform poorly when applied to spatial datasets. Thus new methods are needed to analyze spatial data to detect spatial patterns.

The roots of spatial data mining lie in spatial statistics, spatial analysis, geographic information systems, machine learning, image analysis, and data mining. The main contributions made by computer science researchers to this area include algorithms and data-structures that can scale up to massive (terabytes to petabytes) datasets as well as the formalization of newer spatio-temporal patterns (e.g. colocations) which were not explored by other research communities due to computational complexity. Spatial data mining projects in our group at the Department of Computer Science include location prediction, detection of spatial outliers, and discovery of spatial co-location patterns.

Location prediction is concerned with the discovery of a model to infer locations of a spatial phenomenon from the maps of other spatial features. For example, ecologists build models to predict habitats for endangered species using maps of vegetation, water bodies, climate, and other related species. Figure 1 shows maps of nest location and vegetation durability to build a location prediction model for red-winged blackbirds in the Darr and Stubble wetlands on the shores of Lake Eries in Ohio. Classical data mining techniques yield weak prediction models as they do not capture the auto-correlation in spatial datasets. We provided a formal comparison of diverse techniques from spatial statistics (e.g. spatial autoregression) as well as image classification (e.g. Markov Random Field-based Bayesian classifiers) and developed scalable algorithms for these techniques [28].

Spatial outliers are significantly different from their neighborhood even though they may not be significantly different from the entire population. For example, a brand new house in an old neighborhood of a growing metropolitan area is a spatial outlier. Figure 7 shows another use of spatial outliers in traffic measurements for sensors on I-35W (north bound) for a 24 hour time period. Sensor 9 seems to be a spatial outlier and may be a bad sensor. Note that the figure also shows three clusters of sensor behaviors namely, morning rush hour, evening rush hour, and busy day-time. Spatial statistics tests for detecting spatial outliers do not scale up to massive datasets, such as the Twin Cities traffic dataset measured at thousands of locations in 30-second

intervals and archived for years. We generalized spatial statistics tests to spatio-temporal datasets and developed scalable algorithms [29] for detecting spatial ouliers in massive traffic datasets.

The co-location pattern discovery process finds frequently co-located subsets of spatial event types given a map (see Figure 2) of their locations. For example, the analysis of the habitats of animals and plants may identify the co-locations of predator-prey species, symbiotic species, and fire events with fuel, ignition sources etc. Readers may find it interesting to analyze the map in Figure 2 to find the co-location patterns. (There are two co-location patters of size 2 in this map.) Our group has provided one of the most natural formulations as well as the first algorithms [26] for discovering co-location patterns from large spatial datasets and applying them to climatology data from NASA.

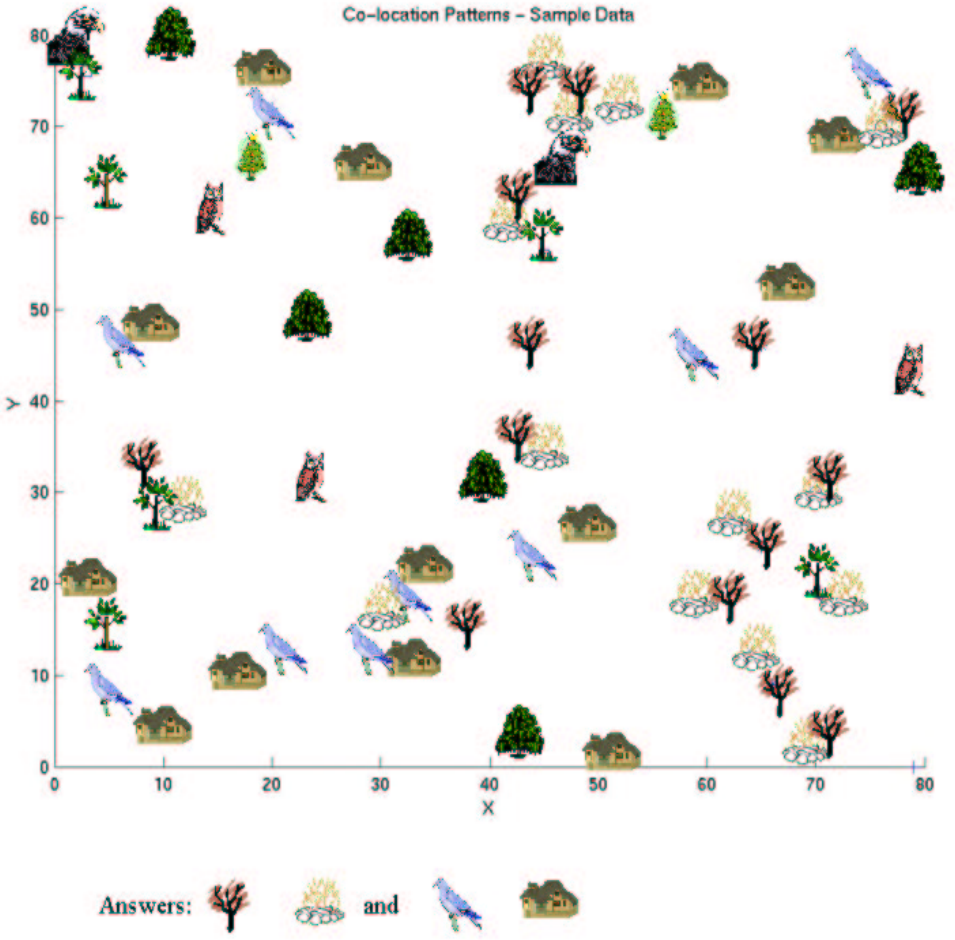


Figure 2: Sample co-location patterns

Paper Organization

We describe each of these techniques in the following sections. In Section 2, we present SAR and MRF techniques for predicting bird nest location using wetland datasets. In Section 3, we introduce spatial outlier detection techniques and their use in finding spatio-temporal outliers in traffic data. Section 4 presents a new approach called co-location mining, which finds the subsets

Figure 3: A spatial framework and its four-neighborhood contiguity matrix

of features frequently-located together in spatial databases. Finally, we conclude with a summary of techniques and results.

2 Location Prediction

The prediction of events occurring at particular geographic locations is very important in several application domains. Crime analysis, cellular networks, and natural disasters such as fires, floods, droughts, vegetation diseases, earthquakes are all examples of problems which require location prediction. In this section we provide two spatial data mining techniques, namely the Spatial Autoregressive Model (SAR) and Markov Random Fields (MRF) and analyze their performance in an example case, the prediction of the location of bird nests in the Darr and Stubble wetlands.

2.1 Modeling Spatial Dependencies Using the SAR and MRF Models

Several previous studies [13], [30] have shown that the modeling of spatial dependency (often called context) during the classification process improves overall classification accuracy. Spatial context can be defined by the relationships between spatially adjacent pixels in a small neighborhood. The spatial relationship among locations in a spatial framework is often modeled via a contiguity matrix. A simple contiguity matrix may represent a neighborhood relationship defined using adjacency, Euclidean distance, etc. Example definitions of neighborhood using adjacency include a four-neighborhood and an eight-neighborhood. Given a gridded spatial framework, a four-neighborhood assumes that a pair of locations influence each other if they share an edge. An eight-neighborhood assumes that a pair of locations influence each other if they share either an edge or a vertex.

Figure 3(a) shows a gridded spatial framework with four locations, A, B, C, and D. A binary matrix representation of a four-neighborhood relationship is shown in Figure 3(b). The row-normalized representation of this matrix is called a contiguity matrix, as shown in Figure 3(c). Other contiguity matrices can be designed to model neighborhood relationship based on distance. The essential idea is to specify the pairs of locations that influence each other along with the relative intensity of interaction. More general models of spatial relationships using cliques and hypergraphs are available in the literature [31].

2.2 Logistic Spatial Autoregression Model(SAR)

Logistic SAR decomposes a classifier $\hat{f}_C$ into two parts, namely Spatial autoregression and logistic transformation. We first show how spatial dependencies are modeled using the framework of logistic regression analysis. In the spatial autoregression model, the spatial dependencies of the error term, or, the dependent variable, are directly modeled in the regression equation[2]. If the dependent values y_i are related to each other, then the regression equation can be modified as

$$y = \rho W y + X\beta + \epsilon. \tag{1}$$

Here W is the neighborhood relationship contiguity matrix and ρ is a parameter that reflects the strength of the spatial dependencies between the elements of the dependent variable. After the correction term ρWy is introduced, the components of the residual error vector ϵ are then assumed to be generated from independent and identical standard normal distributions. As in the case of classical regression, the SAR equation has to be transformed via the logistic function for binary dependent variables.

We refer to this equation as the ***Spatial Autoregressive Model (SAR).*** Notice that when $\rho = 0$, this equation collapses to the classical regression model. The benefits of modeling spatial autocorrelation are many: The residual error will have much lower spatial autocorrelation (i.e., systematic variation). With the proper choice of W, the residual error should, at least theoretically, have no systematic variation. If the spatial autocorrelation coefficient is statistically significant, then SAR will quantify the presence of spatial autocorrelation. It will indicate the extent to which variations in the dependent variable (y) are explained by the average of neighboring observation values. Finally, the model will have a better fit, (i.e., a higher R-squared statistic). We compare SAR with linear regression for predicting nest location in Section 4.

Solution Procedures
The estimates of ρ and β can be derived using maximum likelihood theory or Bayesian statistics. We have carried out preliminary experiments using the spatial econometrics matlab package[1], which implements a Bayesian approach using sampling-based Markov Chain Monte Carlo (MCMC) methods[21]. Without any optimization, likelihood-based estimation would require $O(n^3)$ operations. Recently [24], [25], and [15] have proposed several efficient techniques to solve SAR. The techniques studied include divide and conquer, and sparse matrix algorithms. Improved performance is obtained by using LU decompositions to compute the log-determinant over a grid of values for the parameter ρ by restricting it to $[0, 1]$.

2.3 Markov Random Field based Bayesian Classifiers

Markov Random Field-based Bayesian classifiers estimate the classification model $\hat{f}_C$ using MRF and Bayes' rule. A set of random variables whose interdependency relationship is represented by an undirected graph (i.e., a symmetric neighborhood matrix) is called a Markov Random Field [16]. The Markov property specifies that a variable depends only on its neighbors and is independent of all other variables. The location prediction problem can be modeled in this framework by assuming that the class label, $l_i = f_C(s_i)$, of different locations, s_i, constitute an MRF. In other words, random variable l_i is independent of l_j if $W(s_i, s_j) = 0$.

[1] We would like to thank James Lesage (http://www.spatial-econometrics.com/) for making the matlab toolbox available on the web.

The Bayesian rule can be used to predict l_i from feature value vector X and neighborhood class label vector L_i as follows:

$$Pr(l_i|X, L_i) = \frac{Pr(X|l_i, L_i)Pr(l_i|L_i)}{Pr(X)} \tag{2}$$

The solution procedure can estimate $Pr(l_i|L_i)$ from the training data, where L_i denotes a set of labels in the neighborhood of s_i excluding the label at s_i, by examining the ratios of the frequencies of class labels to the total number of locations in the spatial framework. $Pr(X|l_i, L_i)$ can be estimated using kernel functions from the observed values in the training dataset. For reliable estimates, even larger training datasets are needed relative to those needed for the Bayesian classifiers without spatial context, since we are estimating a more complex distribution. An assumption on $Pr(X|l_i, L_i)$ may be useful if the training dataset available is not large enough. A common assumption is the uniformity of influence from all neighbors of a location. For computational efficiency it can be assumed that only local explanatory data $X(s_i)$ and neighborhood label L_i are relevant in predicting class label $l_i = f_C(s_i)$. It is common to assume that all interaction between neighbors is captured via the interaction in the class label variable. Many domains also use specific parametric probability distribution forms, leading to simpler solution procedures. In addition, it is frequently easier to work with a Gibbs distribution specialized by the locally defined MRF through the Hammersley-Clifford theorem [5].

Solution Procedures
Solution procedures for the MRF Bayesian classifier include stochastic relaxation [9], iterated conditional modes [4], dynamic programming [8], highest confidence first [7] and graph cut [6]. We followed the approach suggested in[6], where it is shown that the maximum a posteriori estimate of a particular configuration of an MRF can be obtained by solving a suitable min-cut multiway graph partitioning problem. Here we briefly provide theoretical and experimental comparisions; more details can be found in [28].

2.4 Comparison of SAR and MRF Using a Probabilistic Framework

We use a simple probabilistic framework to compare SAR and MRF in this section. We will assume that classes $l_i \in (c_1, c_2, \ldots, c_M)$ are discrete and that the class label estimate $\hat{f}_C(s_i)$ for location s_i is a random variable. We also assume that feature values (X) are constant since there is no specified generative model. Model parameters for SAR are assumed to be constant, (i.e., β is a constant vector and ρ is a constant number). Finally, we assume that the spatial framework is a regular grid.

We first note that the basic SAR model can be rewritten as follows:

$$y = X\beta + \rho W y + \epsilon$$

$$(I - \rho W)y = X\beta + \epsilon$$

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\epsilon = (QX)\beta + Q\epsilon \tag{3}$$

where $Q = (I - \rho W)^{-1}$ and β, ρ are constants (because we are modeling a particular problem). The effect of transforming feature vector X to QX can be viewed as a spatial smoothing operation.

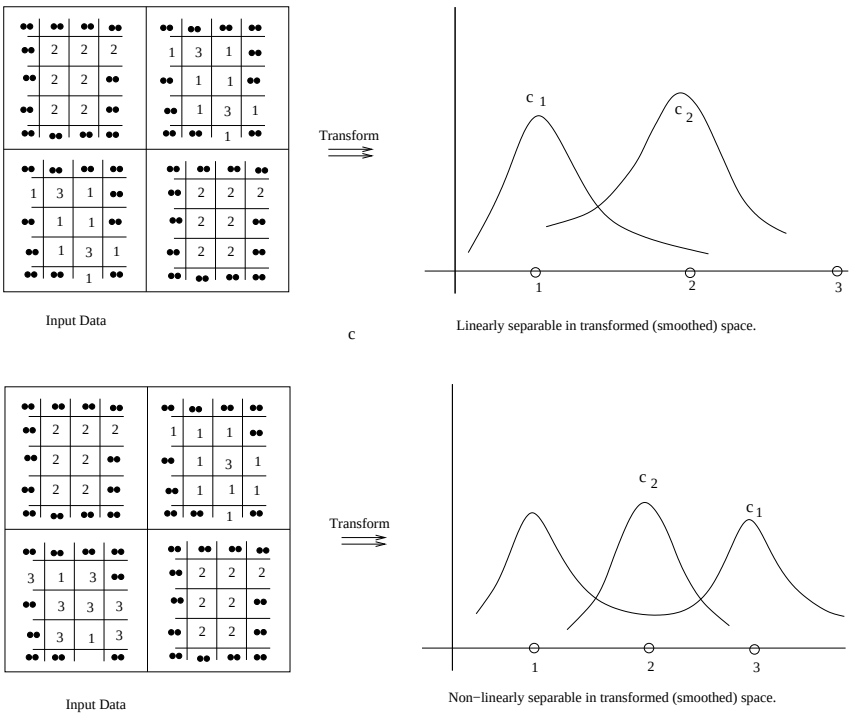


Figure 4: Spatial datasets with *salt and pepper* spatial patterns

The SAR model is similar to the linear logistic model in terms of the transformed feature space. In other words, the SAR model assumes the linear separability of classes in transformed feature space.

Figure 4 shows two datasets with a *salt and pepper* spatial distribution of the feature values. There are two classes, c_1 and c_2, defined on this feature. Feature values close to 2 map to class c_2 and feature values close to 1 or 3 will map to c_1. These classes are not linearly separable in the original feature space. Local spatial smoothing can eliminate the *salt and pepper* spatial pattern in the feature values to transform the distribution of the feature values. In the top part of Figure 4, there are few values of 3 and smoothing revises them close to 1 since most neighbors have values of 1. SAR can perform well with this dataset since classes are linearly separable in the transformed space. However, the bottom part of Figure 4 shows a different spatial dataset where local smoothing does not make the classes linearly separable. Linear classifiers cannot separate these classes even in the transformed feature space, assuming that $Q = (I - \rho W)^{-1}$ does not make the classes linearly separable.

Although MRF and SAR classification have different formulations, they share a common goal, estimating the posterior probability distribution: $p(l_i|X)$. However, the posterior for the two models is computed differently with different assumptions. For MRF the posterior is computed using Bayes' rule. On the other hand, in logistic regression, the posterior distribution is directly fit to the data. For logistic regression, the probability of the set of labels L is given by:

$$Pr(L|X) = \prod_{i=1}^{N} p(l_i|X) \tag{4}$$

One important difference between logistic regression and MRF is that logistic regression assumes no dependence on neighboring classes. Given the logistic model, the probability that the binary label takes its first value c_1 at a location s_i is:

$$Pr(l_i|X) = \frac{1}{1+\exp(-Q_i X\beta)} \tag{5}$$

where the dependence on the neighboring labels exerts itself through the W matrix, and subscript i (in Q_i) denotes the i^{th} row of the matrix Q. Here we have used the fact that y can be rewritten as in equation 3.

To find the local relationship between the MRF formulation and the logistic regression formulation (for the two class case $c_1 = 1$ and $c_2 = 0$), at point s_i

$$\begin{aligned} Pr((l_i = 1)|X, L_i) &= \frac{Pr(X|l_i = 1, L_i)Pr(l_i = 1, L_i)}{Pr(X|l_i = 1, L_i)Pr(l_i = 1, L_i) + Pr(X|l_i = 0, L_i)Pr(l_i = 0, L_i)} \\ &= \frac{1}{1+\exp(-Q_i X\beta)} \end{aligned} \tag{6}$$

which implies

$$Q_i X\beta = \ln(\frac{Pr(X|l_i = 1, L_i)Pr(l_i = 1, L_i)}{Pr(X|l_i = 0, L_i)Pr(l_i = 0, L_i)}) \tag{7}$$

This last equation shows that the spatial dependence is introduced by the W term through Q_i. More importantly, it also shows that in fitting β we are trying to simultaneously fit the relative importance of the features and the relative frequency ($\frac{Pr(l_i=1,L_i)}{Pr(l_i=0,L_i)}$) of the labels. In contrast, in the MRF formulation, we explicitly *model* the relative frequencies in the class prior term. Finally, the relationship shows that we are making distributional assumptions about the class conditional distributions in logistic regression. Logistic regression and logistic SAR models belong to a more general exponential family. The exponential family is given by

$$Pr(u|v) = e^{A(\theta_v)+B(u,\pi)+\theta_v^T u} \tag{8}$$

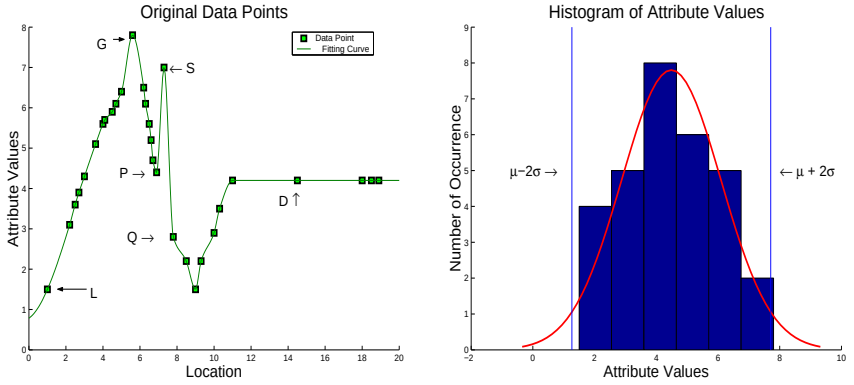
where u, v are location and label respectively. This exponential family includes many of the common distributions such as Gaussian, Binomial, Bernoulli, and Poisson as special cases. The parameters θ_v and π control the form of the distribution. Equation 7 implies that the class conditional distributions are from the exponential family. Moreover, the distributions $Pr(X|l_i = 1, L_i)$ and $Pr(X|l_i = 0, L_i)$ are matched in all moments higher than the mean (e.g., covariance, skew, kurtosis, etc.), such that in the difference $ln(Pr(X|l_i = 1, L_i)) - ln(Pr(X|l_i = 0, L_i))$, the higher order terms cancel out, leaving the linear term $(\theta_v^T u)$ in equation 8 on the left hand-side of equation 7.

Experimental Results: Experiments were carried out on the Darr and Stubble wetlands to compare the classical regression, SAR, and the MRF-based Bayesian classifiers. The results showed that MRF models yield better spatial and classification accuracies over SAR in the prediction of the locations of bird nets. We also observed that SAR predications are extremely localized, missing actual nests over a large part of the marsh lands.

3 Spatial Outlier Detection Techniques

Global outliers have been informally defined as observations in a data set which appear to be inconsistent with the remainder of that set of data [3], or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism [11]. The identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as credit card fraud, athlete performance analysis, voting irregularity, and severe weather prediction. This section focuses on spatial outliers, i.e., observations which appear to be inconsistent with their neighborhoods. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases. These application domains include transportation, ecology, public safety, public health, climatology, and location based services.

We use an example to illustrate the differences among global and spatial outlier detection methods. In Figure 5(a), the X-axis is the location of data points in one dimensional space; the Y-axis is the attribute value for each data point. Global outlier detection methods ignore the spatial location of each data point, and fit the distribution model to the values of the non-spatial attribute. The outlier detected using a this approach is the data point G. On the other hand, S is a spatial outlier whose observed value is significantly different than its neighbors P and Q.



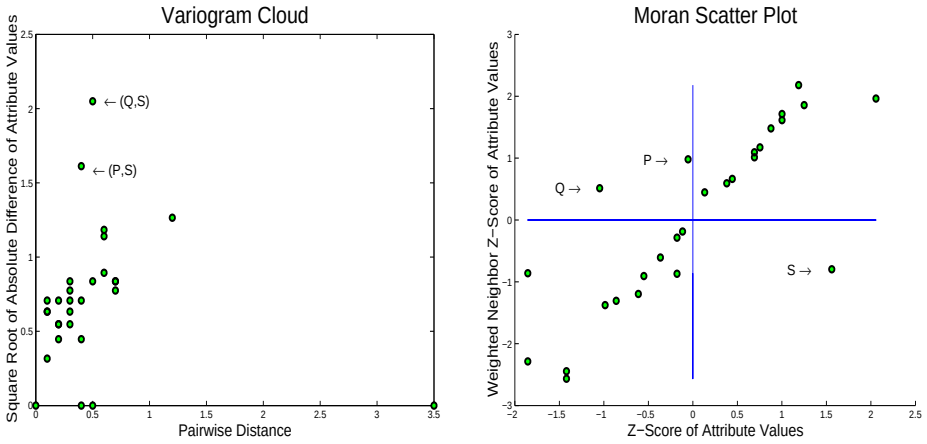
(a) An Example Data Set (b) Histogram

Figure 5: A Data Set for Outlier Detection

3.1 Tests for Detecting Spatial Outliers

Tests to detect spatial outliers separate spatial attributes from non-spatial attributes. Spatial attributes are used to characterize location, neighborhood, and distance. Non-spatial attribute dimensions are used to compare a spatially referenced object to its neighbors. Spatial statistics literature provides two kinds of bi-partite multidimensional tests, namely graphical tests and quantitative tests. Graphical tests are based on the visualization of spatial data which highlights spatial outliers. Example methods include variogram clouds and Moran scatterplots. Quantitative methods provide a precise test to distinguish spatial outliers from the remainder of data. Scatterplots [18] are a representative technique from the quantitative family. Figure 6(a) shows a variogram cloud for the example data set shown in Figure 5(a). This plot shows that two pairs (P, S) and (Q, S)

on the left hand side lie above the main group of pairs, and are possibly related to spatial outliers. The point S may be identified as a spatial outlier since it occurs in both pairs (Q, S) and (P, S). However, graphical tests of spatial outlier detection are limited by the lack of precise criteria to distinguish spatial outliers. In addition, a variogram cloud requires non-trivial post-processing of highlighted pairs to separate spatial outliers from their neighbors, particularly when multiple outliers are present or density varies greatly.



(a) Variogram cloud

(b) Moran scatterplot

Figure 6: Variogram Cloud and Moran Scatterplot to Detect Spatial Outliers

A Moran scatterplot [19] is a plot of normalized attribute value ($Z[f(i)] = \frac{f(i)-\mu_f}{\sigma_f}$) against the neighborhood average of normalized attribute values ($W \cdot Z$), where W is the row-normalized (i.e., $\sum_j W_{ij} = 1$) neighborhood matrix, (i.e., $W_{ij} > 0$ iff neighbor(i, j)). The upper left and lower right quadrants of Figure 6(b) indicate a spatial association of dissimilar values: low values surrounded by high value neighbors(e.g., points P and Q), and high values surrounded by low values (e.g,. point S). Thus we can identify points(nodes) that are surrounded by unusually high or low value neighbors. These points can be treated as spatial outliers.

3.2 Definition of S-Outliers

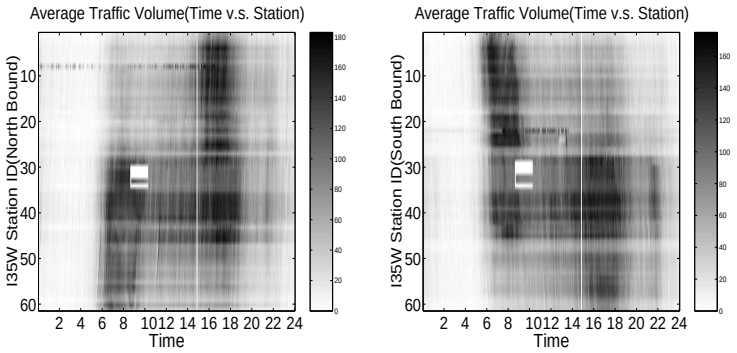
Consider a spatial framework $SF =< S, NB >$, where S is a set of locations $\{s_1, s_2, \ldots, s_n\}$ and $NB : S \times S \rightarrow \{True, False\}$ is a neighbor relation over S. We define a neighborhood $N(x)$ of a location x in S using NB, specifically $N(x) = \{y \mid y \in S, NB(x, y) = True\}$.

Definition: An object O is an S-outlier(f, f^N_{aggr}, F_{diff}, ST) if $ST\{F_{diff}[f(x), f^N_{aggr}(f(x),N(x))]\}$ is true, where $f : S \rightarrow R$ is an attribute function, $f^N_{aggr} : R^N \rightarrow R$ is an aggregation function for the values of f over neighborhood, R is a set of real numbers, $F_{diff} : R \times R \rightarrow R$ is a difference function, and $ST : R \rightarrow \{True, False\}$ is a statistic test procedure for determining statistical significance.

3.3 Solution Procedures

Given the components of the S-outlier definition, the objective is to design a computationally efficient algorithm to detect S-outliers. We presented scalable algorithms for spatial outlier detetection in [29], where we showed that almost all statistical tests are "algebraic" aggregate functions over a neighborhood join. The spatial outlier detection algorithm has two distinct tasks: the first task deals with model building and the second task involves a comparison (test statistic) with spatial neighbors. During model building, algrebraic aggregate functions (e.g., mean and standard deviation) are computed in a single scan of a spatial-join using a neighbor relationship. In the second step, a neighborhood aggregate function is computed by retrieving the neighboring nodes and then a difference function is applied over the neighborhood aggregates and algebraic aggregates. This study showed that the computational cost of outlier detection algorithms are dominated by the disk page access time (i.e., the time spent on accessing neighbors of each point). In this study we utilized three different data page clustering schemes: the Connectivity-Clustered Access Method (CCAM) [27], Z-ordering [23], and Cell-tree [10] and found that CCAM produced the lowest number of data page accesses for outlier detection.

The effectiveness of the $Z_{s(x)}$ method on a Minneapolis-St. Paul traffic data set is illustrated in the following example. Figure 7 shows one example of traffic flow outliers. Figures 7(a) and (b) are the traffic volume maps for I-35W north bound and south bound, respectively, on January 21, 1997. The X-axis is a 5-minute time slot for the whole day and the Y-axis is the label of the stations installed on the highway, starting from 1 on the north end to 61 on the south end. The abnormal white line at 2:45PM and the white rectangle from 8:20AM to 10:00AM on the X-axis and between stations 29 to 34 on the Y-axis can be easily observed from both (a) and (b). The white line at 2:45PM is an instance of temporal outliers, where the white rectangle is a spatial-temporal outlier. Both represent missing data. Moreover, station 9 in Figure 7(a) exhibits inconsistent traffic flow compared with its neighboring stations, and was detected as a spatial outlier. Station 9 may be a malfunctioning sensor.



(a) I-35W North Bound (b) I-35W South Bound

Figure 7: Spatial outliers in traffic volume data

4 Spatial Co-location Rules

Association rule finding [12] is an important data mining technique which has helped retailers interested in finding items frequently bought together to make store arrangements, plan catalogs, and promote products together. In market basket data, a transaction consists of a collection of item types purchased together by a customer. Association rule mining algorithms [1] assume that a finite set of disjoint transactions are given as input to the algorithms. Algorithms like *apriori* [1] can efficiently find the frequent itemsets from all the transactions and association rules can be found from these frequent itemsets. Many spatial datasets consist of instances of a collection of boolean spatial features (e.g. drought, needle leaf vegetation). While boolean spatial features can be thought of as item types, there may not be an explicit finite set of transactions due to the continuity of underlying spaces. In this section we define co-location rules, a generalization of association rules to spatial datasets.

4.1 Illustrative Application Domains

Many ecological datasets [17, 20] consist of raster maps of the Earth at different times. Measurement values for a number of variables (e.g., temperature, pressure, and precipitation) are collected for different locations on Earth. A set of events, i.e., boolean spatial features, are defined on these spatial variables. Example events include drought, flood, fire, and smoke. Ecologists are interested in a variety of spatio-temporal patterns including co-location rules. Co-location patterns represent frequent co-occurrences of a subset of boolean spatial features.

4.2 Co-location Rule Approaches

Given the difficulty in creating explicit disjoint transactions from continuous spatial data, this section defines several approaches to model co-location rules. We use Figure 8 as an example spatial dataset to illustrate the different models. In this figure, a uniform grid is imposed on the underlying spatial framework. For each grid l, its neighbors are defined to be the nine adjacent grids (including l). Spatial feature types are labeled beside their instances. We define the following basic concepts to facilitate the description of different models.

Definition 1 *A* **co-location** *is a subset of boolean spatial features or spatial events.*

Definition 2 *A* **co-location rule** *is of the form $C_1 \rightarrow C_2(p, cp)$ where C_1 and C_2 are co-locations, p is a number representing the prevalence measure, and cp is a number measuring conditional probability.*

The prevalence measure and the conditional probability measure are called interest measures and are defined differently in different models which will be described shortly.

The **reference feature centric model** is relevant to application domains focusing on a specific boolean spatial feature, e.g. cancer. Domain scientists are interested in finding the co-locations of other task relevant features (e.g. asbestos, other substances) to the reference feature. This model enumerates neighborhoods to "materialize" a set of transactions around instances of the reference spatial feature. A specific example is provided by the spatial association rule [14].

For example, in Figure 8 (a), let the reference feature be A, the set of task relevant features be B and C, and the set of spatial predicates include one predicate named "*close_to*". Let us define

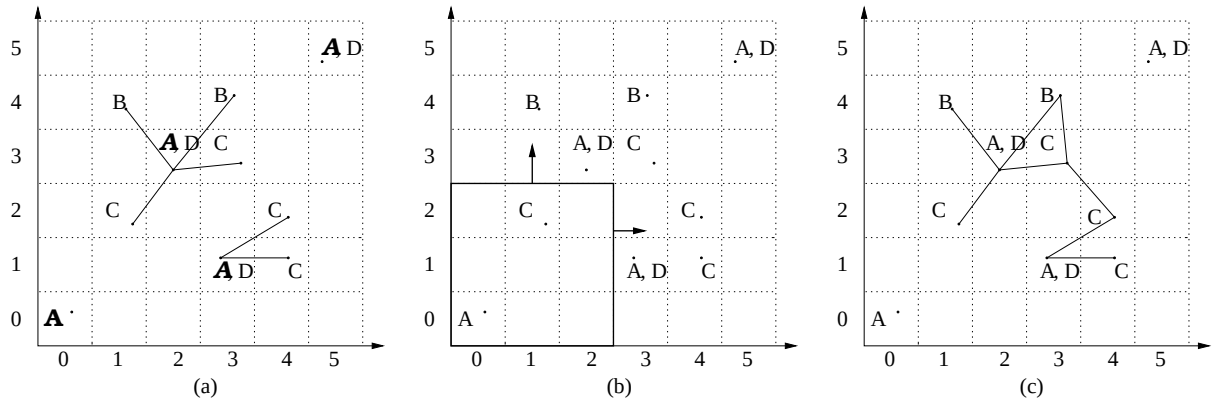


Figure 8: Spatial dataset to illustrate different co-location models. Spatial feature types are labeled besides their instances. The 9 adjacent grids of a grid l (including l) are defined to be l's neighbors. a) Reference feature-centric model. The instances of A are connected with their neighboring instances of B and C by edges. b) Window-centric model. Each 3 X 3 window corresponds to a transaction. c) Event-centric model. Neighboring instances are joined by edges.

$close\_to(a, b)$ to be true if and only if b is a's neighbor. Then for each instance of spatial feature A, a transaction which is a subset of relevant features $\{B, C\}$ is defined. For example, for the instance of A at (2,3), transaction $\{B, C\}$ is defined because the instance of B at (1,4) (and at (3,4)) and instance of C at (1,2) (and at (3,3)) are *close_to* (2,3). The transactions defined around instances of feature A are summarized in Table 1.

Table 1: Reference feature centric view: transactions are defined around instances of feature A relevant to B and C in figure 8(a)

Instance of A	Transaction
(0,0)	ø
(2,3)	$\{B, C\}$
(3,1)	$\{C\}$
(5,5)	ø

With "materialized" transactions, the support and confidence of the traditional association rule problem [1] may be used as prevalence and conditional probability measures as summarized in Table 2. Since 1 out of 2 non-empty transactions contains instances of both B and C and 1 out of 2 non-empty transactions contain C in Table 1, an association rule example is: $is\_type(i, A) \wedge \exists j\, is\_type(j, B) \wedge close\_to(j, i) \rightarrow \exists k\, is\_type(k, C) \wedge close\_to(k, i)$ with $\frac{1}{1} * 100\% = 100\%$ probability.

The **window centric model** is relevant to applications like mining, surveying and geology, which focus on land-parcels. A goal is to predict sets of spatial features likely to be discovered in a land parcel given that some other features have been found there. The window centric model enumerates all possible windows as transactions. In a space discretized by a uniform grid, windows of size kXk can be enumerated and materialized, ignoring the boundary effect. Each transaction contains a subset of spatial features of which at least one instance occurs in the corresponding window. The support and confidence of the traditional association rule problem

Table 2: Interest measures for different models

Model	Items	transactions defined by	Interest measures for $C_1 \rightarrow C_2$	
			Prevalence	Conditional probability
local	boolean feature types	partitions of space	fraction of partitions with $C_1 \cup C_2$	$Pr(C_2$ in a partition given C_1 in the partition)
reference feature centric	predicates on reference and relevant features	instances of reference feature C_1 and C_2 involved with	fraction of instance of reference feature with $C_1 \cup C_2$	$Pr(C_2$ is true for an instance of reference features given C_1 is true for that instance of reference feature)
window centric	boolean feature types	possibly infinite set of distinct overlapping windows	fraction of windows with $C_1 \cup C_2$	$Pr(C_2$ in a window given C_1 in that window)
event centric	boolean feature types	neighborhoods of instances of feature types	participation index of $C_1 \cup C_2$	$Pr(C_2$ in a neighborhood of $C_1)$

may again be used as prevalence and conditional probability measures as summarized in Table 2. There are 16 3X3 windows corresponding to 16 transactions in Figure 8 b). All of them contain A and 15 of them contain both A and B. An example of an association rule of this model is: *an instance of type A in a window* $\rightarrow$ *an instance of type B in this window* with $\frac{15}{16} = 93.75\%$ probability. A special case of the window centric model relates to the case when windows are spatially disjoint and form a partition of space. This case is relevant when analyzing spatial datasets related to the units of political or administrative boundaries (e.g. country, state, zip-code). In some sense this is a local model since we treat each arbitrary partition as a transaction to derive co-location patterns without considering any patterns cross partition boundaries. The window centric model "materializes" transactions in a different way from the reference feature centric model.

The **event centric model** is relevant to applications like ecology, where there are many types of boolean spatial features. Ecologists are interested in finding subsets of spatial features likely to occur in a neighborhood around instances of given subsets of event types. For example, let us determine the probability of finding at least one instance of feature type B in the neighborhood of an instance of feature type A in Figure 8 c). There are four instances of type A and only one of them has some instance(s) of type B in its 9-neighbor adjacent neighborhoods. The conditional probability for the co-location rule is: $spatial\ feature\ A\ at\ location\ l \rightarrow spatial\ feature\ type\ B\ in\ 9-neighbor\ neighborhood\ is\ 25\%$.

4.3 Solution Procedures

Co-location mining is a complex task. It consits of two tasks, schema level purning and instance level purning. At schema level purning, *apriori* [1] can be used. However instance level purning involves neighborhood (i.e., co-location row instance) enumeration, which is a compute intense task. Shekhar et al [26] developed pure geometric, pure combinatorial, hybrid, and multi-resolution algorithms for instance level purning. Experimental analysis shows that the pure geometric algrothm performs much better than pure combinatorial approach. Hybrid algorithm, which is a combination of geometric and combinatorial methods, performed better than both of these approaches. On the other hand, multi-resolution algorithm out performs all these methods when the data is "clumped". It is also shown that co-lcation miner algorithm is complete and correct.

5 Conclusions and Future Work

In this paper we have provided techniques that are specifically designed to analyze large volumes of spatial data to predict bird nests, to find spatial outliers, and to find co-location association rules. We compared the SAR and MRF models using a common probabilistic framework. Our study shows that the SAR model makes more restrictive assumptions about the distribution of features and class shapes (or decision boundaries) than MRF. We also observed an interesting relationship between classical models that do not consider spatial dependence and modern approaches that explicitly model spatial context. The relationship between SAR and MRF is analogous to the relationship between logistic regression and Bayesian Classifiers. The analysis of spatial outlier detection algorithms showed the need for good clustering of data pages. The CCAM method yielded the best overall performance. We showed that the co-location miner algorithm is complete and correct and performs better than the well know *apriori* algorithm.

6 Acknowledgements

This work was supported in part by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014, the content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

We would like to thank to our collaborators Prof. Paul Schrater and Prof. Uygar Ozesmi for their various contributions and group members Yan Huang, Chang-Tien Lu, Weili Wu, and Pusheng Zang for their contributions in developing these techniques and software. The comments of Kimberly Koffolt have greatly improved the readability of this paper.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for Mining Association Rules. In *Proc. of Very Large Databases*, may 1994.

[2] L Anselin. *Spatial Econometrics: methods and models.* Kluwer, Dordrecht, Netherlands, 1988.

[3] V. Barnett and T. Lewis. *Outliers in Statistical Data.* John Wiley, New York, 3rd edition, 1994.

[4] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Statistical Soc.*, (48):259–302, 1986.

[5] J.E. Besag. Spatial Interaction and Statistical Analysis of Latice Systems. *Journal of Royal Statistical Society, Ser. B (Publisher: Blackwell Publishers)*, 36:192–236, 1974.

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts . *Proc. of International Conference on Computer Vision*, September 1999.

[7] P.B. Chou, P.R. Cooper, M. J. Swain, C.M. Brown, and L.E. Wixson. Probabilistic network inference for cooperative high and low levell vision. In *In Markov Random Field, Theory and Applicaitons*. Academic Press, New York, 1993.

[8] H. Derin and H. Elliott. Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, (9):39–55, 1987.

[9] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[10] O. Gunther. The Design of the Cell Tree: An Object-Oriented Index Structure for Geometric Databases. In *Proc. 5th Intl. Conference on Data Engineering*, Feb. 1989.

[11] D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

[12] J. Hipp, U. Guntzer, and G. Nakaeizadeh. Algorithms for Association Rule Mining - A General Survey and Comparison. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.

[13] Yonhong Jhung and Philip H. Swain. Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 34(1):67–75, 1996.

[14] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In *Proc. Fourth International Symposium on Large Spatial Databases, Maine. 47-66*, 1995.

[15] J. P. LeSage and R.K. Pace. Spatial dependence in data mining. In *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, forthcoming, 2001.

[16] S. Li. Markov Random Field Modeling. *Computer Vision (Publisher: Springer Verlag*, 1995.

[17] Z. Li, J. Cihlar, L. Moreau, F. Huang, and B. Lee. Monitoring Fire Activities in the Boreal Ecosystem. *Journal Geophys. Res,. 102(29):611-629*, 1997.

[18] Anselin Luc. Exploratory Spatial Data Analysis and Geographic Information Systems. In M. Painho, editor, *New Tools for Spatial Analysis*, pages 45–54, 1994.

[19] Anselin Luc. Local Indicators of Spatial Association: LISA. *Geographical Analysis*, 27(2):93–115, 1995.

[20] D.C. Nepstad, A. Verissimo, A. Alencar, C. Nobre, E. Lima, P. Lefebvre, P. Schlesinger, C. Potter, P. Moutinho, E. Mendoza, M. Cochrane, and V. Brooks. Large-scale Impoverishment of Amazonian Forests by Logging and Fire. *Nature, 398:505-508*, 1999.

[21] J.P. oeSage. Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, (20):113–129, 1997.

[22] University of Minnesota. Spatial database research group. http://www.cs.umn.edu/research/shashi-group/.

[23] A. Orenstein and T. Merrett. A Class of Data Structures for Associative Searching. In *Proc. Symp. on Principles of Database Systens*, pages 181–190, 1984.

[24] R. Pace and R. Barry. Quick Computation of Regressions with a Spatially Autoregressive Dependent Variable. *Geographic Analysis*, 1997.

[25] R. Pace and R. Barry. Sparse spatial autoregressions. *Statistics and Probability Letters (Publisher: Elsevier Science)*, (33):291–297, 1997.

[26] S. Shekhar and Y. Huang. Co-location Rules Mining: A Summary of Results. *Proc. Spatio-temporal Symposium on Databases*, 2001.

[27] S. Shekhar and D-R. Liu. CCAM: A Connectivity-Clustered Access Method for Aggregate Queries on Transportation Networks. *IEEE Transactions on Knowledge and Data Engineering*, 9(1):102–119, January 1997.

[28] S. Shekhar, Paul R. Schrater, Ranga R. Vatsavai, Weili Wu, and S. Chawla. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transaction on Multimedia (accepted), http://www.cs.umn.edu/research/shashi-group/paper_list.html*, 2002.

[29] Shashi Shekhar, C.T. Lu, and P. Zhang. A unified approach to spatial outliers detection. *TR01-045 (Also to appear in IEEE TKDE), http://www.cs.umn.edu/research/shashi-group/paper_list.html*, 2001.

[30] A. H. Solberg, Torfinn Taxt, and Anil K. Jain. A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transaction on Geoscience and Remote Sensing*, 34(1):100–113, 1996.

[31] C. E. Warrender and M. F. Augusteijn. Fusion of image classifications using Bayesian techniques with Markov rand fields. *International Journal of Remote Sensing*, 20(10):1987–2002, 1999.

An overview of the SAND Spatial Database System*

Claudio Esperança†
Gísli R. Hjaltason‡
Hanan Samet
Frantisek Brabec
Egemen Tanin
Computer Science Department
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland
College Park, Maryland 20742

December 22, 2001

Abstract

An overview is given of the SAND spatial database system, an environment for developing applications involving both spatial and non-spatial data. The SAND kernel implements a relational data model extended with several geometric functions and predicates as well as a spatial index. The main interface to SAND is through an embedded interpreted language. This permits the rapid prototyping of algorithms and makes SAND a useful tool both for applications and research. A graphical user interface that allows for easy database querying, and a client/server approach that simplifies remote access are also outlined.

*This work was supported in part by the National Science Foundation under Grants EIA-99-00268, IIS-00-86162, EIA-00-91474, and IRI-97-12715.

†Author's current address: COPPE, Programa de Engenharia de Sistemas e Computação, Universidade Federal do Rio de Janeiro, Ilha do Fundão, CT, Bloco H, Rio de Janeiro, RJ 21949-900, Brazil.

‡Current address: RightOrder, Inc., 3850 N. First Street, San Jose, CA 95134.

1 Introduction

The dramatic rise in the use of the Internet and the worldwide web has led to a re-examination of traditional ways of looking at data. In particular, we increasingly find the need for applications to be location-aware. This means the incorporation of spatial data such as that found in maps. The first revolutionary step was the development of geographic information systems (GIS). In fact, one of the principal motivations for the development of GIS is to lessen the time necessary to produce a map. Maps have traditionally been formed as a result of a sequence of physical overlay operations and thus a reduction in the time necessary to perform such a task has always been viewed as desirable. Computerizing this physical process was viewed as such an improvement in speed that users were not overly concerned with making the computer execution time optimal. In particular, taking several hours to compute a result was considered acceptable in light of the time needed to tabulate the query manually and generate the output. Unfortunately, users today have been conditioned to get results quickly and do not want to spend much time waiting for them. In other words, they are willing to accept a result that is not 100% accurate and without so much detail provided that they can obtain it sufficiently quickly and with enough detail to make an intelligent decision (see [16] for a similar conclusion in a database environment).

1.1 Direct Manipulation

One of the principal reasons for this change of thinking is the familiarity of GIS users with spreadsheets. The invention of the spreadsheet is generally accepted as the most important factor in demonstrating the utility of the computer to the average user. It enabled her to answer questions such as "what if ..." without large commitments of time and money. One of the powers of the spreadsheet lies in its ability to make a database come alive in a visual sense. Moreover, the method of interaction with the spreadsheet is usually via direct manipulation (e.g., [29]) in the sense that users do not need to know how to write computer programs to get the output they desire. Instead, all operations are in terms of the basic entities and actions of the spreadsheet (i.e., the rows, columns, and drag-drop actions). Thus the spreadsheet enables users to have a very powerful decision support tool that responds to their requests instantly. They do not have to go to the printer to get the output to their queries. The output is in a format that facilitates subsequent queries.

Armed with their experiences with spreadsheets, it is not unreasonable for users to expect the same capability from their other decision making tools. Unfortunately, this is not so easy. For example, in the context of a GIS whose output is a paper map, once we have generated the paper map, we cannot simply annotate it with changes and pose subsequent queries on the modified map. Of course, we could save the output digitally and work directly on the screen. The limited resolution of the screen can be compensated by having a zoom capability (e.g., [5, 6, 21, 28]). Nevertheless, the fact that the output has the resolution usually expected for a paper map meant that the operations also took quite a long time to execute. This is especially troublesome when users do not necessarily require such resolution.

An additional problem for users is that frequently the type of queries that they wish to pose are a combination of spatial and nonspatial data. Spatial data is usually stored in a GIS (e.g., ArcInfo from ESRI) while nonspatial data is stored in a conventional database management system (DBMS) [14, 31, 33]. The drawback of most GIS is that they are usually very good for location-based queries such as "what feature is at location x?" However, they are not so good for feature-based queries such as "where is feature y?" [3]. Such queries are usually best handled by a conventional DBMS. Users want to answer both of these types of queries with equal ease. They do not want to know how the GIS or the nonspatial database are organized.

Handling such queries requires a seamless integration of spatial and nonspatial data [14, 31, 33]. The idea is to interact with a spatial database (GIS) in the same manner as we would interact with a nonspatial DBMS. Moreover, spatial operations should be executed using conventional database primitives. One problem with conventional DBMSs is that they are usually accessed via the aid of SQL (Structured Query Language) which

is rather cumbersome when it comes to nonstandard data such as maps and images [10], although there have been a number of attempts to adapt SQL to the spatial domain (e.g., [11, 13, 25]). Nevertheless, a graphical user interface is more appropriate as it enables users to query the underlying database without having to worry about whether or not a corresponding SQL extension has been defined already.

Conventional spreadsheets get their name from the fact that all the data is spread out in a tabular format and operations are specified in terms of combinations of rows and columns. An analogous problem-solving paradigm in a GIS is the overlay concept (e.g., [12, 32]). In this case, operations are specified as compositions of maps with the output of one or more operations serving as input to other operations. Frequently, in a GIS there is no need for the operation to run to completion to obtain the desired results. Often, we would like to proceed in a pipelined fashion where the first results of an operation are fed as inputs to subsequent operations. We characterize such a solution as being *incremental*. We use the term *browsing* (*browser*) to describe the physical process (processor) of (for) obtaining an answer incrementally.

1.2 Example

As an example of one of the composition queries that we wish to handle, consider "finding the closest county (in terms of distances in the plane) to Cook County (i.e., Chicago) with a bladder cancer mortality rate for white males greater than 7.5 per 100,000 people in the period 1970–1994 and a population greater than 1 million". What makes this query difficult is the presence of the spatial condition involving distances between two-dimensional regions. Conventional DBMSs facilitate retrieval on the basis of a particular attribute by building an index for it (i.e., sorting it). In the case of one-dimensional data like the mortality rate per 100,000 people and the population, this is quite simple as we have a zero reference point with which to sort the data. For example, assuming the existence of an index on the bladder cancer mortality rate per 100,000 people in the period 1970–1994, to find all counties in the order of the closeness of their mortality rate per 100,000 people to that of Cook County for which it is 8.5, we look up the value 8.5 in the index and then proceed in two directions along the index on the population attribute to obtain the nearest counties by mortality rate per 100,000 people in constant time. We do not have to rebuild the index if we want to be able to answer the next query which deals with the mortality rate per 100,000 people in Los Angeles county whose population is about 8 million.

Unfortunately, this strategy cannot be used when dealing with distances in the two-dimensional plane. For example, to find the county closest to Cook County, we could sort the counties according to their distances from Cook County. However, to find the closest county to Los Angeles County, the list of sorted distances from Cook County is not useful to us due to the non-additivity of distances in domains whose dimensionality is greater than one. In other words, the distance from Cook County to Los Angeles County is not equal to the sum of the distance from Los Angeles County to St. Louis County and the distance from St. Louis County to Cook County. Thus we need to be careful in the manner in which we represent the locations of the counties. In particular, we need to use an implicit spatial index that is based on spatial occupancy (e.g., a quadtree, R-tree, etc. [26, 27]) rather than an explicit spatial index that is based on distances from a particular reference point.

The query that we have just described is an instance of a process that we term *ranking*. Ranking is a byproduct of sorting. However, often, we are only interested in the first few values (e.g., the three closest counties to Cook County), in which case sorting the entire set of counties by distance from Cook County is an overkill. Moreover, as we saw, if we want the nearest three counties to Los Angeles County, then we must reinvoke the sort. Thus the most frequent solution is to calculate the k nearest counties to Cook County. The problem with this approach is that if we want the $k+1^{st}$ (e.g., the fourth) nearest county to Cook County, then we have to restart the computation and compute the $k+1$ (i.e., 4) nearest neighbors. Therefore, it is preferable to compute the nearest neighbors in an incremental fashion so that we need not compute more neighbors than are necessary. This is especially useful when we want to respond to queries such as finding

the nearest county to Cook County with a bladder cancer mortality rate for white males in the period 1970–1994 greater than 7.5 per 100,000 people and a population greater than 1 million since the nearest one may not satisfy the mortality rate and population conditions thereby necessitating finding the next nearest, etc.

1.3 Our System

SAND (denoting Spatial And Nonspatial Data) is a prototype spatial database system/GIS developed at the University of Maryland that has many of the above features. In particular, the SAND system contains a browser called the SAND Browser that enables the visual definition and execution of these queries. We are using the SAND system in digital government applications such as FedStats and the National Atlas of Cancer Mortality. The intended purpose of the SAND system is to be a research vehicle for work in spatial indexing, spatial algorithms, interactive spatial query interfaces, etc. The basic notion of SAND is to extend the traditional relational database paradigm by allowing table attributes to be *spatial* objects (e.g., line segments or polygons), and by allowing spatial indexes (such as quadtrees) to be built on such attributes, just as traditional indexes (like B-trees) are built on nonspatial attributes.

The rest of this paper is organized as follows. Section 2 describes the basic structure of the SAND system (i.e., the SAND kernel and the SAND interpreter), and which has recently been extended to function within a client/server environment. Section 3 presents the SAND Browser and the SAND Internet Browser, which provide a graphical user interface to the query facilities of the SAND system, as well as examples of their use in the context of digital government applications. Concluding remarks are drawn in Section 4 in addition to suggestions for future work.

2 SAND

SAND is divided into two main layers, the SAND kernel, and the SAND interpreter (see Figure 4). The SAND kernel was built in an object oriented fashion (using C++) and comprises a collection of classes (i.e., object types) and class hierarchies that encapsulate the various components. SAND adopts a data model inspired by the relational model. Thus, its core functionality is defined by the different types of tables and attributes it supports, and the class hierarchies that encapsulate this functionality are among the most important. Both of these aspects of the SAND kernel are defined in an extensible manner, so that new table and attribute types can readily be added to the system. The SAND interpreter provides a low-level procedural query interface to the functionality defined by the SAND kernel. Using the query interface provided by the SAND interpreter, we have built a number of useful tools. In addition to the interactive spatial query browsers described in Section 3, we have built a prototype for a high-level declarative query interface to SAND, modeled on SQL, and a prototype image database system [30].

The SAND kernel has many of the characteristics of full-featured relational database systems. For example, it has a block-based storage manager that caches blocks associated with the various tables (i.e., relations and indexes) in the system, with an LRU replacement policy. Furthermore, tuples (also termed *rows* and *records*) are laid out in blocks such that a block may contain multiple tuples, while large tuples may span multiple blocks. Nevertheless, due to main emphasis of our research, SAND does not currently support features such as transaction and concurrency support, or query planning and optimization.

2.1 Table Types

The table abstraction in SAND encapsulates what in conventional databases are known as relations and indexes. Tables are handled in much the same way as regular disk files, i.e., they have to be opened so that input and output to disk storage can take place. All open tables in SAND respond to a minimal common set of operators, such as **first**, **next**, **insert**, and **delete**. SAND currently defines three table types: relations,

linear indexes, and spatial indexes. Each table type supports an additional set of operators, specific to its functionality. The function of many of these operators is to alter the order in which tuples are retrieved, i.e., the behavior of **first** and **next**.

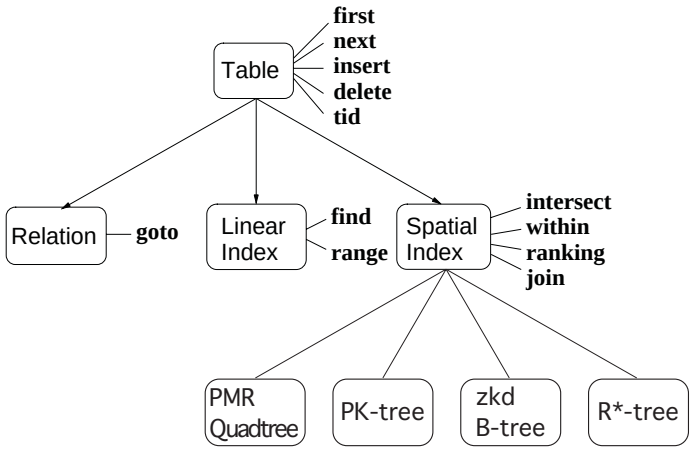


Figure 1: The class hierarchy of tables supported by SAND (the arrows denote class derivation).

Relations in SAND are tables that support direct access by tuple identifier (*tid*, which are composed of a block number and a tuple number). Ordinarily, tuples are retrieved in order of increasing *tid*, but the operation **goto** *tid* can be used to jump to the tuple associated with the given *tid* (if it exists). For access by attribute values, indexes can be defined on attributes or groups of attributes of relations.

Linear indexes for non-spatial attributes are implemented using B-trees [8]. Tuples in a linear index are always scanned in an order determined by a linear ordering relation. Linear indexes support the **find** operator, that locates the first tuple in the index that is greater than or equal to the argument, and the **range** operator, that is used to perform range search.

Indexes can also be defined on spatial attributes in SAND. The SAND kernel defines an extensible framework for spatial indexes that makes it straight-forward to plug an already-implemented spatial access method into the system. Currently, the system supports indexing with the PMR quadtree [22], PK tree [34], zkd B-tree [23], and R*-tree [4, 15]. Spatial attribute values indexed by a spatial index may be represented in up to three ways (at least one of which must be supported by a given spatial index type): 1) *inline*, where the spatial attribute value is stored inside the spatial index, 2) *object table*, where a separate table is created to store copies of the attribute values, and 3) *from relation*, where the attribute values are accessed directly from the tuples of the indexed relation. The advantage of *inline* is that during search, all the information is present in the index, but it has the drawback that it makes the index larger. The advantage of *object table* over *from relation* is that the relation may contain many attributes, all of which would be accessed if the spatial attribute value of a tuple is needed during search. Moreover, the *object table* approach allows *clustering* the spatial attribute values in a way that optimizes I/Os (see [7]), without affecting the relation itself. Nevertheless, it has the drawback that the spatial attribute values are stored in two places, in the relation itself and in the object table.

A number of standard search operators are defined for spatial indexes, some or all of which may be implemented by a particular index type. These include **intersect**, for searching tuples that intersect a given feature; **within**, for retrieving tuples in the proximity of a given feature; and **ranking** [17, 19], for retrieving tuples in order of distance from a given feature (ranking is closely related to nearest neighbor queries). Furthermore, the **join** operator can be applied on two spatial indexes, where the join is either by intersection (i.e., a traditional spatial join) or by distance (termed *distance join* [18]).

1	2	2S	3	*N*
char	point	spoint	point3d	point(N)
char(N)	line	sline	line3d	box(N)
string	rectangle	lrectangle	box3d	
integer	polygon	spolygon	triangle3d	
float	region		triangleStrip3d	
boolean	icon			
array	image			

Figure 2: Attribute types defined for each dimension class.

2.2 Attribute Types

SAND implements attributes of common non-spatial types (integer and floating point numbers, fixed-length and variable-length strings, etc.) as well as various kinds of spatial types. Attribute types have an associated *dimension class*, that group together "compatible" attribute types, and attribute values have an associated dimensionality. The dimension classes currently defined in the system are labeled "1", "2", "2S", "3", and "*N*", where the numeric labels correspond with the dimensionality, "2S" denotes two-dimensional spherical geometry [2], and "*N*" denotes arbitrary dimensionality (i.e., the attribute types of that class support spatial objects of any fixed dimensionality). Thus, non-spatial attribute types are all in the class labeled "1", while each spatial attribute type is in one of the other classes. All the spatial classes (i.e., classes other than "1") contain at least an attribute type for points and another one for axis-aligned hyper-rectangles. The attribute types currently defined in each class are listed in Figure 2.

All attribute types support a common set of operations to convert their values to and from text, to copy values between attributes of compatible types, as well as to compare values for equality. Non-spatial attribute types also support the **compare** operator, which is used to establish a linear ordering between values of the same type. This is required so that non-spatial attributes can be used as keys in linear indexes. Spatial attribute types support a variety of geometric operations, including **intersect**, which tests whether two features intersect, **distance**, which returns the Euclidean distance between two features (used for the **ranking** operator), and **bbox**, which returns the smallest axis-aligned rectangle that contains a given feature (i.e., its minimum bounding rectangle). Some spatial types support additional operations. For instance, the *region* type supports operations like **expand**, which can be used to perform morphological operations such as contraction and expansion, and **transform**, which can be used in the computation of set-theoretic operations.

The attribute types listed in Figure 2 may be thought of as comprising a class hierarchy, with the base class "Attribute", and a derived base class for each dimension class, as partially depicted in Figure reffig-sandattr. However, for performance reasons and for increased flexibility, instead of relying on the object-oriented features of C++, we opted to develop our own attribute type manager that provides an extensible mechanism for attribute types and functions on them. The type manager maintains a registry of types, each of which has an associated string identifier (as listed in Figure 2) and a unique numeric identifier that is used internally. Furthermore, the type manager coordinates the creation and release of spatial objects, and the invocation of the common set of operations mentioned above. The type manager also maintains a function registry, where each function is also identified by a string, and may take an arbitrary number of arguments (each of which may be for a fixed attribute type or for an arbitrary one) and have a return value of several different types. The function registry is used for the spatial operations mentioned above, **intersect**, **distance**, and **bbox**, that are defined for all spatial attribute types, as well as for specialized operations that have been added as needs arose (e.g., **area** for computing the area of two-dimensional polygons). With the type manager it is easy to add new types and functions on them to the system, and the set of types and functions in the system is continually growing.

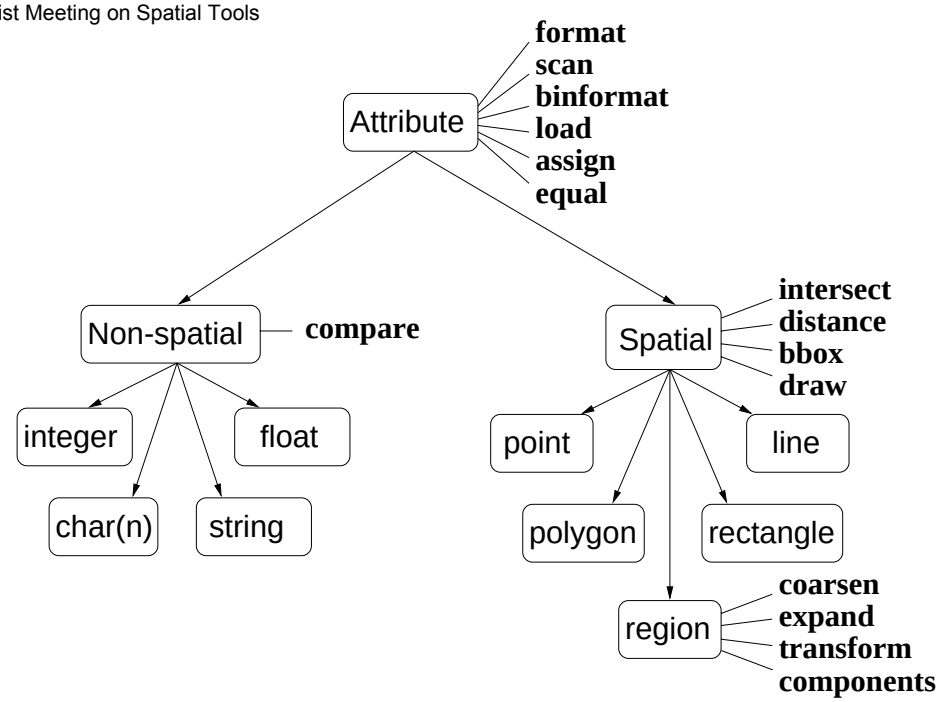


Figure 3: Some of the attributes types implemented in SAND and some of the operations defined on them.

2.3 The SAND/Tcl Interface

The SAND kernel provides the basic functionality needed for storing and processing spatial and non-spatial data. In order to access the functionality of this kernel in a flexible way, we opted to provide an interface to it by means of an interpreted scripting language, *Tcl* [24]. Tcl offers the benefits of an interpreted language but still allows code written in a high-level compiled language (in our case, C++) to be incorporated via a very simple interface mechanism. Another advantage offered by Tcl is that it provides a seamless interface with *Tk* [24], a toolkit for developing graphical user interfaces.

The SAND interpreter provides commands that mirror all kernel operations mentioned in the previous sections. In some cases, a single command may cause more than one kernel operation to be performed. In addition, the interpreter implements data definition facilities. The processing of spatial queries is supported by interpreter commands that operate on spatial attributes and spatial indexes. While some of the commands available in the SAND interpreter are for accessing SAND kernel functionality, most are defined by the underlying Tcl interpreter and the Tk toolkit. In addition, we developed a Tk "widget" for displaying two-dimensional maps, used by the SAND Browser (see Section 3.1), that efficiently handles large data sets and provides zooming and panning facilities. Furthermore, system can be extended through Tcl's scripting capability by writing new methods or query strategies, which are either a standard addition or added by an application developer. In fact, the interpreter can be viewed as the unifying element of the whole SAND system (see Figure 4, which is a block diagram of the SAND system).

2.4 Client/Server Architecture

The SAND interpreter application (i.e., the executable) includes the SAND kernel codebase, since the interpreter directly accesses functionality of the kernel. Thus, any application built on top of the interpreter, such as the SAND Browser (see Section 3.1), runs in the same process as the SAND kernel, and displays maps on

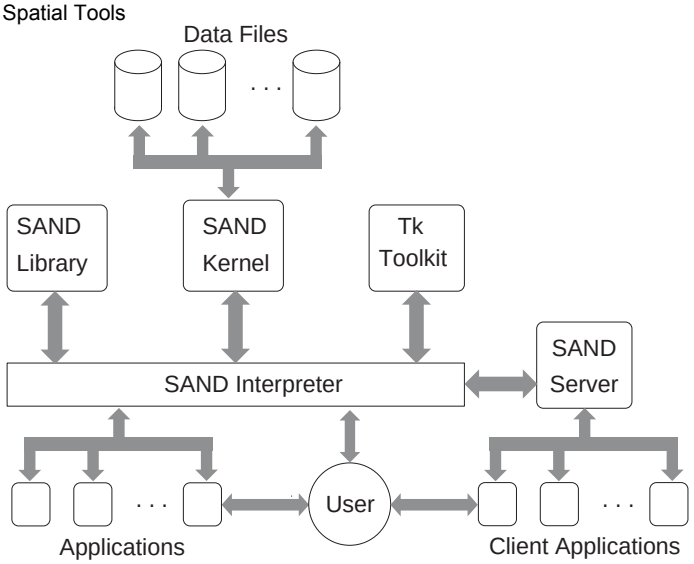


Figure 4: A block diagram of the SAND system.

the same computer[1]. Although for many uses, this is an adequate solution, the proliferation of the Internet makes it increasingly attractive to provide database access over a wide area network. To address this need, many web-based spatial data providers have adopted the approach of delivering maps as images, created by a database server (e.g., www.MapQuest.com and www.MapsOnUs.com). This an appropriate solution for many applications, and requires minimal resources for both hardware and software on the client side. Nevertheless, the resulting product has severe limitations in terms of available functionality and response time, and the transferred images do not have the same flexibility and representational power as the spatial data itself.

Based on the above considerations, we chose to adopt a client/server model where the actual data values (including those of the spatial data types) are transferred between the client and the server, and the client can issue queries to the server. One option would be to expose the full SAND/Tcl interface to the client, but this approach would invite a host of potential security risks. Thus, we designed a protocol that provides a more restricted access to the functionality of the SAND interpreter. The server itself is written in the SAND/Tcl script language, while we have developed a graphical client, written in Java, with functionality that resembles the SAND Browser (see Section 3.2). In an earlier effort, we also developed an OpenMap server interface for SAND, in cooperation with USGS [9].

3 Interactive Query Interfaces

In this section, we describe two ways of interacting with the SAND system in a graphical manner (Sections 3.1 and 3.2), and present examples of their use (Section 3.3).

3.1 SAND Browser

The SAND Browser provides a graphical user interface to the facilities of SAND. It permits the visualization of the data contained in a SAND relation by specifying two types of controls: the scan order in which tuples are to be retrieved, and an arbitrary selection predicate. The tuples satisfying the query are obtained in an incremental order. Users rarely need to wait too long to get visual feedback provoked by an action.

[1]Strictly speaking, this is not true for systems that support the X windowing system. Nevertheless, relying on X is not a viable solution for delivering maps over the Internet.

The class of queries currently implemented in the SAND Browser is restricted to selections and spatial joins (e.g., distance joins and distance semijoins [18]). The user specifies queries by choosing the desired selection conditions from a variety of menus and dialog boxes. Spatial values can either be drawn on the appropriate display pane or typed in by filling forms. Query results can either be displayed interactively using the *First* and *Next* buttons or saved in relations for use in subsequent SAND queries in a manner somewhat analogous to a spreadsheet.

3.2 SAND Internet Browser

As mentioned in Section 2.4, the SAND Browser is not suitable for interactive map delivery over the Internet. In a more recent effort, we have developed another graphical query interface for SAND, that functions as a client to the server interface described in Section 2.4. This client interface, termed the SAND Internet Browser, is built using the popular Java technology, and is a relatively simple and lightweight application. Using Java provides platform independence while reducing installation and maintenance efforts. Like the SAND Browser, the SAND Internet Browser is more than a naive image viewer, but instead operates on vector data and allows the client to perform many operations such as zoom in/out or locational queries without communicating with the server. In essence, the client keeps a local *cache* of a portion of the whole database, which is only updated when additional or newer data is needed.

We see two different types of usages for our Java-based browser. First, the browser can be activated as an applet so that users across various platforms can access a spatial database on a remote machine without having to install the SAND system on their side. Second, the browser along with the SAND kernel can be installed on the client side. In the latter case, the browser can be utilized to view data from remote data sources, while frequently used data can be loaded to the local SAND database on demand, and subsequently accessed locally. In time, users can build up their own local databases and make it available over the network.

3.3 Sample Queries

Figure 5 is a screenshot of what a user interaction with the SAND Internet Browser might look like. It shows the relation corresponding to mortality rates per 100,000 for bladder cancer for white males for the time period 1970–1994. We have also overlaid it with the result of a clustering-like operation that is available in SAND. In particular, we have shown a partition of the underlying space with respect to the 17 counties with the highest mortality rates so that each county in each partition is closer to the county with the high rate in the same partition than to any other county with a high rate. The green dots indicate locations of high chlorine emissions obtained from the FedStats [1] website. The goal is to see if there is some spatial correlation between counties with a high incidence of bladder cancer and large chlorine emissions. As can be seen, locations with a large amount of chlorine emissions are not clustered around these counties. Thus these two events do not seem to be spatially correlated.

The scenario depicted in Figure 5 is analogous to a discrete Voronoi diagram and is a form of clustering. This clustering operation is available in both the SAND Browser and the SAND Internet Browser and can be achieved by executing an incremental *distance semi-join* [18] operation where the input relation corresponding to the high chlorine emissions map is joined with the high incidence of bladder cancer map and the join condition is based on proximity with the closest tuple pairs from the two sets being retained. Once the closest emissions-cancer pair (a,b) has been found, the next closest pair is found from the set of emissions tuples which excludes tuple a from participating. This process is continued until the closest high incidence of bladder cancer county has been found for each of the high chlorine emissions locations.

Figure 6 illustrates another sample query. In this query, nuclear facilities around a certain monitoring station along the northeastern U.S. and the Canadian border (Figure 6a) are computed in the order of their distance to this station. First, we define our query by selecting the location of the station and then the ranking

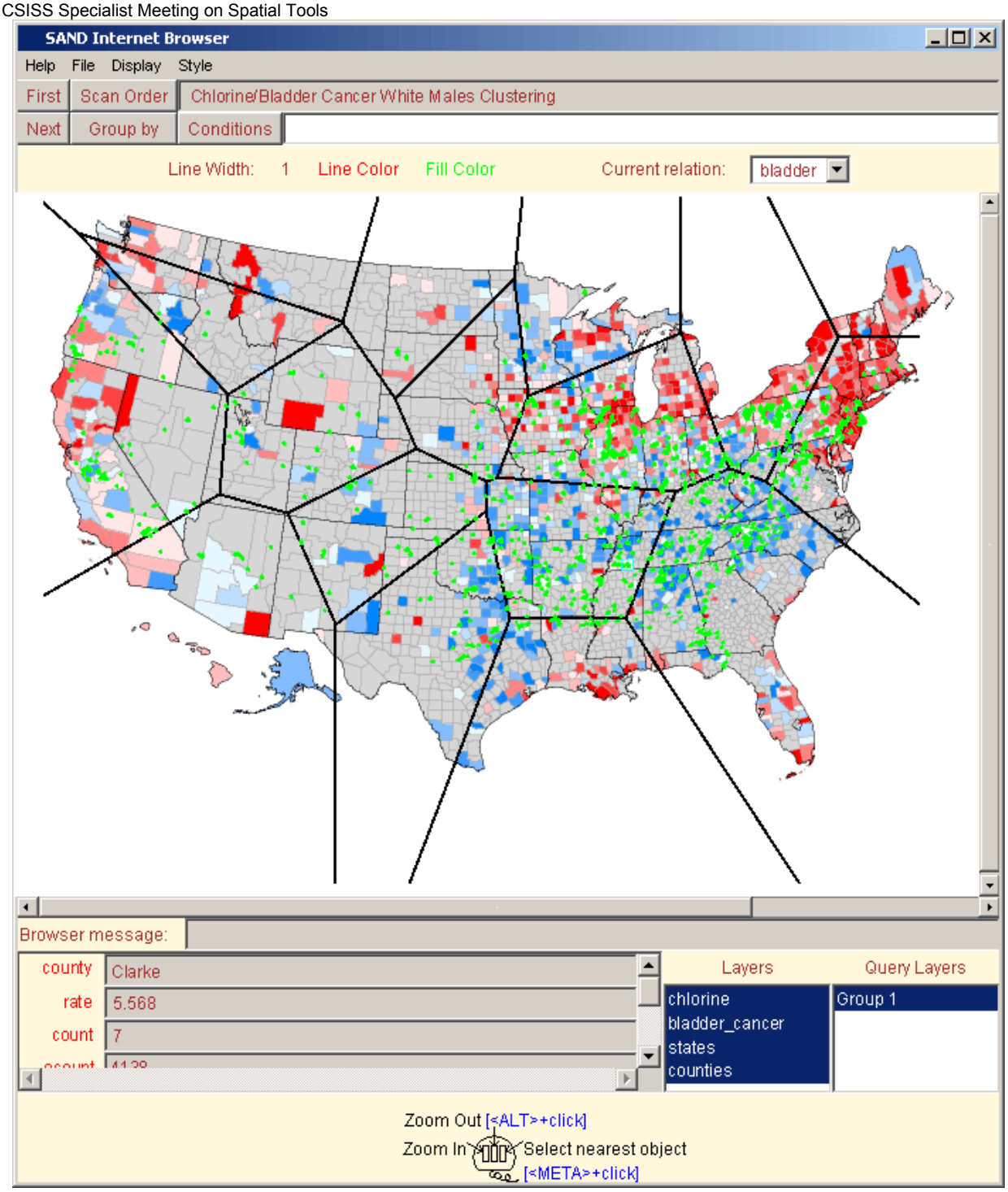


Figure 5: Sample screenshot of a possible user interaction with the SAND Internet Browser. The relation being displayed corresponds to classes of mortality rates per 100,000 for bladder cancer for white males for the time period 1970–1994. It is also overlaid with the result of a partition of the underlying space with respect to the 17 counties with the highest mortality rates so that each county in each partition is closer to the county with the high rate in the same partition than to any other county with a high rate. The green dots indicate locations of high chlorine emissions.

operation starts by displaying our first hit (Figures 6b and 6c). By clicking the "Next" button we can continue this operation as long as we want (Figure 6d). Again, if desired, the nuclear facilities relation can be partitioned with respect to the monitoring stations relation with a discrete Voronoi diagram (Figure 7).

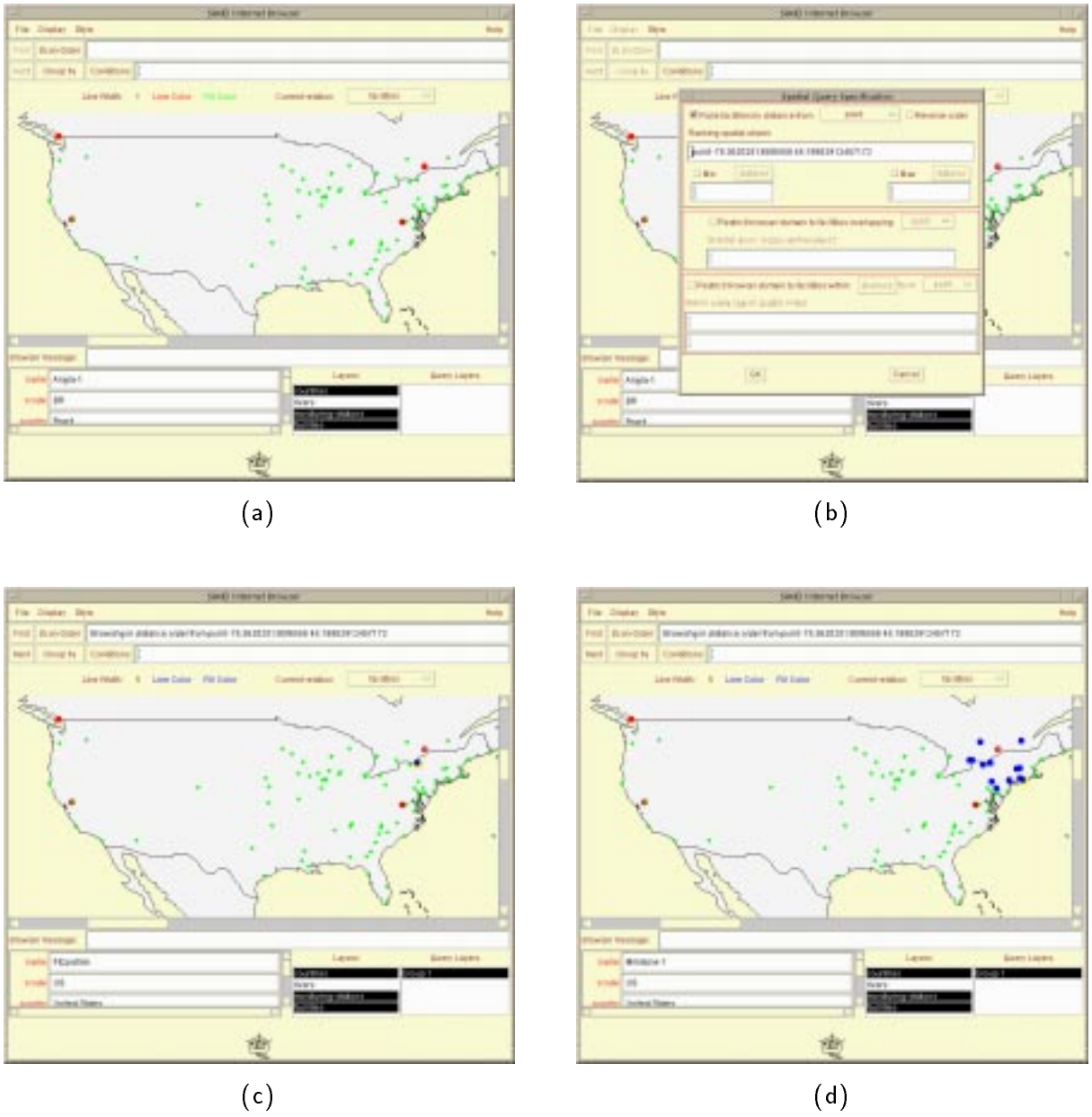


Figure 6: The nuclear facilities around a certain monitoring station along the northeastern U.S. and the Canadian border are computed. The green dots indicate nuclear facilities, the red dots indicate monitoring stations, and the blue dots indicate hits to our query. (a) Displays the two relations, monitoring stations and nuclear facilities; (b) the location of a certain station is chosen for a ranking query by distance; (c) the closest facility is displayed; (d) the query continues with other hits, incrementally.

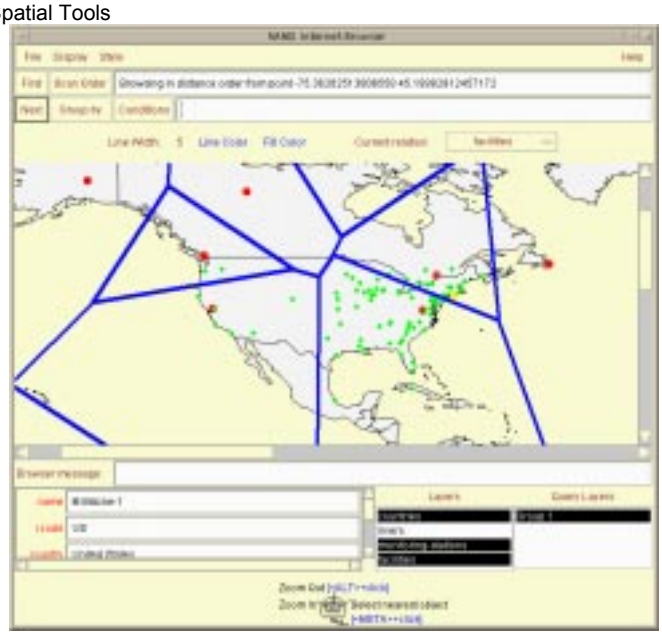


Figure 7: The discrete Voronoi diagram, partitioning the nuclear facilities relation with respect to the monitoring stations relation, is depicted.

4 Concluding Remarks

SAND is an on-going project. At present, we are focusing on the client/server environment for which our efforts are in two directions. The first is on developing efficient caching methods that would balance limited client resources on one side and significant latency of the client/server link on the other while the low bandwidth of this link would be a concern in both cases. The second is to help users that want to manipulate data for prolonged periods of time by developing a peer-to-peer approach to provide them with the ability to download fairly large amounts of data more efficiently by better utilizing the distributed network bandwidth. In addition, we want to use the same mechanism to help them upload the results of their work to a remote server if needed.

The classic client/server approach for transferring data between the two ends of a connection assumes a designated role for each of the ends (i.e., a client + a server). It also ignores the fact that the needs of the two ends may be time dependent (i.e., congested periods of usage for the server). A pure peer-to-peer approach where the two ends/peers can assume both the roles of a client and a server from time to time may improve the overall network performance by resolving the congested situations. At a given time the server may be busy serving other requests (forming a congestion). The common solution to this problem in the realm of databases is to cache or forward results/requests. The novelty of the peer-to-peer approach is converting the static configuration of forwarding requests to a highly dynamic one where a persistent storage is formed from a pool of clients and servers (peers). A request to download/upload data can be performed by a set of selected peers from this pool at a given time that optimizes the network performance. Keeping alive and fresh copies of the data (and hence a directory of active peers) forms a challenging research problem in this area. Hybrid configurations where a main server (e.g., a government/company operated server) exists are also possible.

Other work includes adding a map browsing capability to FedStats using the SAND Browser. This work also involves the construction of a utility that would convert Federal government statistical data in EXCEL format to be compatible with the SAND Browser. In addition, work is ongoing to further develop the concept of a spatial spreadsheet [20] using the SAND Browser. It is interesting to note that SAND has already been used for a prototype image database system [30].

References

[1] Fedstats: The gateway to statistics from over 100 U.S. federal agencies. `http://www.fedstats.gov/`, 2001.

[2] H. Alborzi and H. Samet. Augmenting SAND with a spherical data model. In *International Conference on Discrete Global Grids*, Santa Barbara, CA, March 2000.

[3] W. G. Aref and H. Samet. Efficient processing of window queries in the pyramid data structure. In *Proceedings of the Ninth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 265–272, Nashville, TN, April 1990. Also *Proceedings of the Fifth Brazilian Symposium on Databases*, pages 15–26, Rio de Janeiro, Brazil, April 1990.

[4] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: an efficient and robust access method for points and rectangles. In *Proceedings of the ACM SIGMOD Conference*, pages 322–331, Atlantic City, NJ, June 1990.

[5] B. B. Bederson and J. D. Hollan. Pad++: a zooming graphical interface for exploring alternate interface physics. *Journal of Visual Languages and Computing*, 7(1):3–31, March 1996.

[6] E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose. Toolglass and magic lenses: the see-through interface. In *Proceedings of the SIGGRAPH'93 Conference*, pages 73–80, Anaheim, CA, August 1993.

[7] T. Brinkhoff and H.-P. Kriegel. The impact of global clustering on spatial database systems. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, J. Bocca, M. Jarke, and C. Zaniolo, eds., pages 168–179, Santiago, Chile, September 1994.

[8] D. Comer. The ubiquitous B-tree. *ACM Computing Surveys*, 11(2):121–137, June 1979.

[9] C. B. Cranston, F. Brabec, G. R. Hjaltason, D. Nebert, and H. Samet. Adding an interoperable server interface to a spatial database: Implementation experiences with OpenMap™. In *Interoperating Geographic Information Systems — Second International Conference, INTEROP'99*, A. Včkovski, K. Brassel, and H.-J. Schek, eds., pages 115–128, Zurich, Switzerland, March 1999. Also Springer-Verlag Lecture Notes in Computer Science 1580.

[10] M. J. Egenhofer. Why not SQL! *International Journal of Geographical Information Systems*, 6(2):71–85, March-April 1992.

[11] M. J. Egenhofer. Spatial sql: A query and presentation language. *IEEE Transactions on Knowledge and Data Engineering*, 6(1):86–95, February 1994.

[12] M. J. Egenhofer and J. R. Richards. Exploratory access to geographic data based on the map-overlay metaphor. *Journal of Visual Languages and Computing*, 4(2):105–125, June 1993.

[13] S. Gadia and V. Chopra. A relational model and SQL-like query language for spatial databases. In *Advanced Database Systems*, N. R. Adam and B. K. Bhargava, eds., Lecture Notes in Computer Science 759, pages 213–225. Springer-Verlag, Berlin, Germany, 1993.

[14] R. H. Güting. An introduction to spatial database systems. *VLDB Journal*, 3(4):401–444, October 1994.

[15] A. Guttman. R-trees: a dynamic index structure for spatial searching. In *Proceedings of the ACM SIGMOD Conference*, pages 47–57, Boston, June 1984.

[16] J. M. Hellerstein, P. J. Haas, and H. Wang. Online aggregation. In *Proceedings of the ACM SIGMOD Conference*, J. Peckham, ed., pages 171–182, Tucson, AZ, May 1997.

[17] G. R. Hjaltason and H. Samet. Ranking in spatial databases. In *Advances in Spatial Databases — Fourth International Symposium, SSD'95*, M. J. Egenhofer and J. R. Herring, eds., pages 83–95, Portland, ME, August 1995. Also Springer-Verlag Lecture Notes in Computer Science 951.

[18] G. R. Hjaltason and H. Samet. Incremental distance join algorithms for spatial databases. In *Proceedings of the ACM SIGMOD Conference*, L. Hass and A. Tiwary, eds., pages 237–248, Seattle, WA, Jun 1998.

[19] G. R. Hjaltason and H. Samet. Distance browsing in spatial databases. *ACM Transactions on Database Systems*, 24(2):265–318, June 1999. Also University of Maryland Computer Science TR-3919.

[20] G. Iwerks and H. Samet. The spatial spreadsheet. In *Proceedings of the Third International Conference on Visual Information Systems (VISUAL99)*, D. P. Huijsmans and A. W. M. Smeulders, eds., pages 317–324, Amsterdam, The Netherlands, June 1999.

[21] H. Lieberman. Powers of ten thousand: navigating in large information spaces. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 15–16, Marina del Rey, CA, November 1994.

[22] R. C. Nelson and H. Samet. A consistent hierarchical representation for vector data. *Computer Graphics*, 20(4):197–206, August 1986. Also *Proceedings of the SIGGRAPH'86 Conference*, Dallas, August 1986.

[23] J. A. Orenstein and T. H. Merrett. A class of data structures for associative searching. In *Proceedings of the Third ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 181–190, Waterloo, Ontario, Canada, April 1984.

[24] J. K. Ousterhout. *Tcl and the Tk Toolkit*. Addison-Wesley, Reading, MA, 1994.

[25] N. Roussopoulos, C. Faloutsos, and T. Sellis. An efficient pictorial database system for PSQL. *IEEE Transactions on Software Engineering*, 14(5):639–650, May 1988.

[26] H. Samet. *Applications of Spatial Data Structures: Computer Graphics, Image Processing, and GIS*. Addison-Wesley, Reading, MA, 1990.

[27] H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, MA, 1990.

[28] M. Sarkar and M. H. Brown. Graphical fisheye view of graphs. *Communications of the ACM*, 37(12):73–84, December 1994.

[29] B. Shneiderman. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison Wesley, Reading, MA, third edition, 1997.

[30] A. Soffer and H. Samet. Two approaches for integrating symbolic images into a multimedia database systems: a comparative study. *VLDB Journal*, 7(4):253–274, December 1998.

[31] M. Stonebraker. Limitations of spatial simulators for relational DBMSs. Technical report, INFORMIX Software, Inc., 1997. `http://www.informix.com/informix/corpinfo/zines/whitpprs/wpsplsim.pdf`.

[32] C. D. Tomlin. *Geographic information systems and cartographic modelling*. Prentice Hall, Englewood Cliffs, NJ, 1990.

[33] T. Vijlbrief and P. van Oosterom. The GEO++ system: an extensible GIS. In *Proceedings of the Fifth International Symposium on Spatial Data Handling*, pages 40–50, Charleston, SC, August 1992.

[34] W. Wang, J. Yang, and R. Muntz. PK-tree: a spatial index structure for high dimensional point data. In *Proceedings of the Fifth International Conference on Foundations of Data Organization and Algorithms (FODO)*, K. Tanaka and S. Ghandeharizadeh, eds., pages 27–36, Kobe, Japan, November 1998.

Integrating Cellular Automata and GIS for Dynamic Spatial Modelling of Agricultural Landuse

Tara Sharma and B. Klinkenberg
Department of Geography, University of British Columbia
1984 West Mall, Vancouver, BC, Canada V6T1Z2

Current Geographic Information Systems (GIS) offer a wide range of spatial analytical functions but are inefficient in dealing with dynamic spatial models and the temporal dimensions of some socio-economic and environmental processes. Researchers have sought to address this problem by integrating GIS with models and simulation tools (for example, Batty and Xie 1994, Mikula et al 1996). Integration of Cellular Automata (CA) with GIS has been a very widely used method to address the dynamics of spatial models. CA are dynamic systems based on discrete time and space. In these cell-based systems, the state of a cell for next time step is determined by the state of its neighboring cells according to some transition rules. This makes the CA approach very amenable to handling spatial dynamics. Integration of CA and GIS allows one to exploit the advantages of both systems. CA serve as the analytical engines, enabling dynamic modeling within GIS (Park and Wagner 1997).

While this integration has shown potential, particularly for modeling urban dynamics (Deadman et al. 1993, Batty and Xie 1997, White and Engelen 1997), not much work has been done using it for modeling intra-agricultural landuse conversions. A fundamental difference between modelling urban dynamics and agricultural dynamics is the concept of re-development or re-conversion. Urban systems are considered as self-organizing systems wherein global patterns evolve from interactions of the site with its neighborhood (Wu, 1998). Once the land is converted from some other use to urban use it is generally not re-converted to a non-developed state after that point in time. Dynamics in agricultural landuse, on the other hand, are determined by a mix of factors, such as site suitability, as well as regional factors like urban growth and community preferences for diet, for example. These are also affected by global factors such as export markets. Agricultural land may also be lost due to urban growth. A farmer growing food crops, for

example, may switch to livestock if export markets for meat increase. Interplay of these factors may result in modification or re-conversion of the landuse. This implies that any attempts to use CA for agricultural landuse scenarios must provide for the means to consider these factors in the transition rules.

This research and tool development is placed in a broader context of the development of a simulation tool - QUEST (Quite Useful Ecosystem Scenario Tool) (URL: http://www.basinfutures.net/). QUEST is an innovative computer simulation tool used to explore alternative futures and to facilitate debate and discussions among a variety of diverse groups about regional sustainability in the Georgia Basin. QUEST aims to foster an understanding of sustainability by placing the user in the position of making decisions that impinge upon regional development, and allowing him/her to explore and reflect upon the consequences of those decisions which are presented in the form of future scenarios. User choices are used to develop likely landuse scenarios and their impacts on environmental and socio-economic systems. QUEST consists of 12 sub-models related to various themes; for example, agriculture, forestry, urban growth, and energy. Methods and models developed from this research project will be implemented in QUEST for exploring alternative options for agricultural sustainability.

The objective of this study is to develop agricultural landuse scenarios for the year 2040 in Georgia Basin. Specifically, it simulates the conversions taking place between three major agricultural landuse categories, viz. food crops, greenhouses and livestock operations. The approach used here integrates constrained CA and GIS for the development of realistic landuse scenarios. A constrained CA approach (White et al. 1997, Li and Yeh 2000) embeds some constraints in the transition rules of CA in order to reflect actual development patterns. For example, land converted to urban use is a constraint for the development of agricultural landuse. A combination of GIS and CA methods has earlier been used by Wu (1998) to simulate land conversions. In that study, Multi Criteria Evaluation (MCE) was used to define the relative importance of development factors using an analytical hierarchy procedure. In the present approach, the transition rules are defined using a Multi Criteria Evaluation (MCE) process in GIS.

MCE is used to generate a potential surface of suitability for each landuse. In addition to development constraints, the transition rules defined for conversion of landuse *i* to landuse *j* are based on 4 factors: suitability of a site for landuse *j*, historical precedences of conversion of use *i* to use *j*, demand for use *j*, and the neighborhood uses around the site. The transition rules are translated to push/pull scores for the different factors in order to create a potential surface for each landuse. A potential surface describes the attractiveness of each cell to a given landuse. A cell with a high "attractiveness" value will pull (attract) a land class, whereas a cell with a negative value will push (repel) a landuse class. Potential surfaces are generated for all three uses and summed scores for each use are then converted to probability of conversion from use *i* to use *j*. This approach can connect well to a decision-making process as user choices and preferences can also be used here to assign scores that push or pull a certain landuse.

The constrained CA-GIS approach is implemented in ARC/INFO using Macros. Landuse data are derived from a classification of Landsat TM data. Information on constraints (urban growth) is obtained for different periods using an Urban Growth model developed for QUEST. Information on static constraints (which do not change during the entire simulation period), such as water and rocky areas, is obtained from the classified landuse map. Suitability of land for a particular use is based on its bio-physical factors such as land capability, slope and proximity to markets. Simulation of landuse proceeds in discrete time steps and in each step the conversion probability of use *i* to use *j* is computed. A stochastic disturbance is applied to the probability values in order to determine cells for which conversion will take place. The spatial allocation of land is based on the area requirements for each use. The constraints and the transition rules in each step are updated to reflect the land dynamics in the previous step. Landuse patterns generated in each step serve as the initial state for the next scenario. The final landuse scenario is thus a temporal evolution not only of the initial landuse but also of other local and global factors that affect the development process.

At the meeting, I would demonstrate the QUEST tool and/or this spatial modeler tool that will be integrated into QUEST.

References

Batty,M. and Xie, Y., 1994. From cells to cities. Environment and Planning B, 21: 531-548.

Batty,M. and Xie, Y., 1997. Possible urban automata. Environment and Planning B, 24:175-192.

Deadman,P.D., Brown,R.D., and Gimblett,H.R., 1993. Modelling rural residential settlement patterns with cellular automata. Journal of Environmental Management, 37:147-160.

Li, X., and Yeh, A.G., 2000. Modelling sustainable urban development by integration of constrained cellular automata and GIS. International Journal of Geographic Information Science , 131-151.

Mikula,B., Mathian, H., Pumain,D., and Sanders L., 1996. Integrating dynamic spatial models with GIS. In Longley, Batty (eds): *Spatial Analysis: Modelling in a GIS environment*, 283-295.

Park, S., and Wagner, D.F., 1997. Incorporating cellular automata simulators as analytical engines in GIS. Transitions in GIS, 2 : 213-231.

White,R., and Engelen, G., 1997. Cellular automata as the basis of integrated dynamic regional modeling. Environment and Planning B, 24:235-246.

White,R., Engelen,G., and Uijee, I., 1997. The use of constrained cellular automata for high-resolution modeling of urban landuse dynamics. Environment and Planning B, 24:323-343.

Wu, F., 1998. SimLand: a prototype to simulate land conversion through the integrated GIS and CA with AHP-derived transition rules. International Journal of Geographical Information Systems 12(1), 63-82.

Numerical Evaluation of Local Moran's I_i's p-values for Normal Distributed OLS Residuals under Global Spatial Independence

Michael Tiefelsdorf
Department of Geography – The Ohio State University
Columbus, Ohio 43210
Email: tiefelsdorf.1@osu.edu

Abstract

Global Moran's I and local Moran's I_i are by far the most commonly used test statistics to identify global and local autocorrelation in the residuals of spatial regression models. This discussion will focus primarily on the numerical evaluation of local Moran's I_i's distribution, which is well known to be asymptotically non-normal even under the restrictive assumptions of global spatial independence and normal distributed regression residuals.

The unconditional distribution under global spatial independence as well as the conditional distribution subject to a spatial process of Moran's I, and other quadratic forms as well as ratios of quadratic forms in normal distributed regression residuals, is well understood (see Tiefelsdorf, 2000). Its practical evaluation, however, is limited for single high-end personal computers to moderately sized spatial tessellations up to approximately 3000 spatial objects for global tests and up to approximately 1000 spatial objects for an exhaustive set of local tests. This is mainly due to the fact [1] that a full spectrum of eigenvalues must be calculated, which changes from setting to setting for global tests and from reference location to reference location for local tests, and [2] that numerical integration must be applied to get the individual p-values. Both calculations are mainly time-consuming and to a lesser degree memory demanding, particularly if use is made of the potential sparseness of the spatial relationship matrix and symmetries in the other matrices. For local spatial statistics, which require n independent calculations for an exhaustive set of n reference locations in a spatial tessellation, use could be made of a distributed computing environment in a computer cluster. By allowing several computers to work jointly on the same task, substantial gains in computing time can be acquired. This approach, however, requires besides a computer cluster specially designed software and network administrator skills by the analyst and thus is not feasible in common social and behavioral sciences settings.

Another, more practical approach is the use of accurate approximations and highly specialized algorithms. While the calculation or approximation of eigenvalue spectra is common practice in spatial statistics, numerical quadrature, on the other hand, does not usually belong to the standard repertoire of software developers for spatial statistics. Instead of using numerical integration for the evaluation of the exact significance points of quadratic forms, the saddlepoint approximation first advocated by

Offer Lieberman (1994) is used here. Tiefelsdorf [1](2002) demonstrates that the saddlepoint approximation for local and global Moran's *I* outperforms any alternative approximation methods, such as the normal approximation, the Pearson's distribution family approximations, and the Edgeworth series approximation. The accuracy of saddlepoint approximation is high even under extreme conditions such as those for local Moran's I_i and the saddlepoint approximation is besides the exact approach the only method that has the capability to handle Moran's *I*'s distribution conditional to an underlying spatial process.

The high accuracy and flexibility of the saddlepoint approximation, however, does not come without a price. It requires a full spectrum of eigenvalues as input. In general, the computation of the spectrum of eigenvalues is time-wise the most expensive part of running the procedure. Auspiciously, as shown in Tiefelsdorf (2002) for local Moran's I_i under the assumption of global spatial independence, the exact spectrum of eigenvalues can be given in analytical terms for any OLS regression model. Furthermore, by making explicit use of simplifications based on matrix algebra, the order of computational operations to derive the analytical solution can be reduced substantially. Although this requires direct element-wise programming of the matrix terms.

A prototype SPSS syntax file[1] demonstrates the simplicity and feasibility of the saddlepoint approximation for smaller tessellations *T*. It approximates $\Pr(I \le I_{obs} \mid H_0)$ as well as a complete set of $\Pr(I_i \le I_{i,obs} \mid H_0) \quad \forall i \in T$ for OLS regression residuals under the assumption of spatial independence. In addition, it provides results for either the row-sum standardized *W*-coding scheme, the globally standardized *C*-coding scheme, or the variance stabilizing *S*-coding scheme of the spatial relationship matrix. Also, the moments of global and local Moran's *I* up to the fourth order are given as well as Moran's *I*'s feasible range. Due to restrictions of SPSS's matrix facility to handle sparse matrices, it does not make use of the algebraic evaluation of the underlying eigenvalue spectrum.

References

Lieberman, Offer. 1994. Saddlepoint Approximation for the Distribution of a Ratio of Quadratic Forms in Normal Variables. *Journal of the American Statistical Association* **89**:924-928.

Tiefelsdorf, Michael. 2000. *Modelling Spatial Processes - The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I.* Berlin: Springer Verlag.

Tiefelsdorf, Michael. 2002. The Saddlepoint Approximation of Moran's *I*'s and Local Moran's I_i's Reference Distribution and Their Numerical Evaluation. *Geographical Analysis* **34**, forthcoming[1].

[1] Find at http://geog-www.sbs.ohio-state.edu/faculty/tiefelsdorf/GeoStat.htm

Extending Map Algebra into the Third and Fourth Dimensions

C. Dana Tomlin
Professor of Landscape Architecture and Regional Planning
University of Pennsylvania
Philadelphia, Pennsylvania 19104

"Map Algebra" is a term that has been used over the past two decades in loose reference to a particular way or organizing the analytical and synthetic capabilities of certain geographic information systems. In order to address a wide array of applications in a clear and consistent manner, the map algebraic approach decomposes data, data-processing capabilities, and data-processing control techniques into elemental components that can then be recomposed with relative ease and with great flexibility.

The result is an algebra-like language in which cartographic layers for individual characteristics such as soil type, land value, or population are treated as variables that can be transformed or combined into new variables by way of specified operations. Just as conventional algebraic operations (such as adding, subtracting, multiplying, or dividing) might be combined into a complex system of simultaneous equations, these cartographic operations (such as superimposing one map onto another, measuring distances or travel times, characterizing geographic shapes, computing topographic slopes and aspects, determining visibility, or simulating flow patterns) might be combined into a model of soil erosion or land development potential. This approach is currently employed in most of the world's major raster-based geographic information systems.

While GIS has traditionally offered tools and techniques for the analysis and synthesis of two-dimensional space, a number of developers have begun to consider possibilities for the extension of this functionality into the third (spatial) and/or fourth (temporal) dimensions. In doing so, some have also begun to look to Map Algebra as one way to envision, organize, and implement such extensions.

This paper explores that prospect in terms that are functionally explicit but not specific to and particular GIS. It offers suggestions as to the structure and substance of a map algebraic language which builds on that of Tomlin (1990) in order to encompass maps and map-transforming operations relating to space that may be two- or three-dimensional and (either of) which may also extend over time. The paper considers data, data processing, and data-processing control but focuses on the second of these components. In particular, it describes volumetric and temporal versions of the traditional (planar) map algebraic operations as well as new operations for which there are no two-dimensional analogues.

Tomlin, C.D. 1990. Geographic Information Systems and Cartographic Modeling. Prentice-Hall.

Exploratory Data Analysis and Decision Making with Descartes and CommonGIS

Hans Voss, Natalia Andrienko, Gennady Andrienko

Fraunhofer Institut Autonome Intelligente Systeme, FhG-AIS, Schloss Birlinghoven, Sankt-Augustin, D-53754 Germany, Tel: +49-2241-14-2532, -2329, Fax: +49-2241-14-2072, E-mail: {Gennady.Andrienko, Natalia.Andrienko, Hans.Voss}@ais.fraunhofer.de, URL: http://www.ais.fhg.de/KD/

Abstract. Our thematic mapping software Descartes has been steadily extended and improved over the past eight years. When it became available in the year 1994 it was the first fully interactive, explorative tool for spatial analysis of statistical geo-referenced data available for use in the Internet. Interactivity was achieved by Java-Applet technology. Beside its powerful presentation and exploration functions one primary research and practical goal of its development continuously was to incorporate cartographic knowledge into the software. Thus expert users are freed from routine activities and casual users are enabled to utilize the functions without much practice. In the previous three years, the internal architecture and implementation of Descartes was completely revised. As a result, it is now very easy to build new system configurations including particular sets and variants of functions. One subset of Descartes' functions has been placed in the Internet for free download and non-commercial use. This configuration is named CommonGIS, according to the EU-funded project of the same name, which provided financial support for creating a GIS accessible and usable for expert and common users alike.

1 Introduction

The thematic mapping software Descartes is available as a stand-alone tool for running on a local PC, and as a web version with client-server architecture. Actually, it was the first thematic mapping software that one could interactively run in the web from within a standard Java-enabled browser when it became first available in 1994. In the web version the server performs knowledge-based design of maps, and sends specifications of appropriate visualization methods to the client. The client software, which is a Java Applet, generates maps according to the specifications and provides various related interactive tools.

Funding of the European Union, with project CommonGIS, resulted in a complete revision of the original Descartes software. In particular, it became completely component oriented, configurable, and embeddable into other software systems. A particular assembly of functions, comprising most of the available visualization methods, but not, for example, data base connectivity, was made available for free download and non-commercial use at www.commongis.de. In the following, we will refer to CommonGIS as the "base system", and only mention Descartes when referring to functionality that is not available in CommonGIS.

CommonGIS can be perceived as an interactive Web-GIS, because it provides some standard GIS functionality, but particularly it is a tool for the visualization, exploratory analysis, and decision support based on geographically referenced statistical data.

Working with CommonGIS functions according to the following scenario: Accessing CommonGIS using a standard web browser, a user may select an application (a territory with associated thematic data). Within the application one may repeatedly choose one or more data variables for analysis. Each time, CommonGIS will **automatically** offer one or more adequate thematic maps visually presenting the data. The maps built by the system comply with sound principles of graphical and cartographical representation of information, i.e. they correspond to (1) characteristics of the data (e.g. whether a variable is qualitative, ordinal, or numeric) and (2) relationships between variables (e.g. comparability or inclusion). Moreover, in designing maps the system also may take into account the analytical goals of the user. As a result, one or more map displays will be created which are cartographically sound, on the one hand, and useful, on the other hand.

The digital maps are highly interactive and dynamically transformable, according to the concept of "geographic visualization" and EDA. For every type of representation method (pie charts, bar charts, choropleth maps, etc.) several interactive techniques are realized, all of them specifically designed to increase the analytical power of the respective representation. The interactivity of the map displays on a web client is achieved by Java and Java Applet technology.

2 Functionality

CommonGIS provides many methods for the visual presentation and analysis of data, either based on maps or using various types of statistical graphs. One outstanding feature of CommonGIS is its control of brushing, where the user may inspect related data by simultaneous highlighting in various linked displays. Here is a summary of available functions:

Data transformations ("calculation")

Classification of one attribute

Classification of two attributes (cross classification)

Classification by dominant attribute

Classification by n-dimensional attribute similarity

Ideal point evaluation (Fig. 3)

Ordering of values

Average/Median/Quartiles/Variance

Calculation of arbitrary formula

Filter objects according to specified attribute values

Map visualizations ("mapping")

Choropleth maps for one numeric attribute

Choropleth maps for one qualitative attribute

Choropleth maps for cross classification

Choropleth maps for other classifications (Fig. 2)

Multiple choropleth maps

Diagrams: utility bars

Diagrams: utility pies

Diagrams: normal bars

Diagrams: normal pies

Diagrams: stacked bars

Diagrams: triangles for two attributes

Graph visualizations ("charting")

Dot plot

Scatter plot

S-plot matrix

Parallel coordinate plots (Fig. 3)

Tukey's box plots

Histograms

Basic map functions

Zoom out (to maximal extent)

Zoom out (by factor)

Zoom in (with rectangle)

Zoom in (by factor)

Undo last zoom

Shift the map

Select mouse mode (zoom, pan, select or explore)

Change the order of layers

Change layer properties

Change the size and color of symbols

Windows functions

Open new window with current map

Minimize all

Restore all

Close all

Bring an open window to the front

Support Functions

Help functions

General system help

Task support guide (support based on characteristics of data and intended goal of the user)

General information about system and application

Other functions

Show advanced menu items

Show menu items for application building

Range of Data Input and Output possibilities

Data input

Load attribute data from:

DBF, CSV (Excel), TXT (delimited text), ODBC/JDBC, clipboard.

Load geographical data from:

OVL (GMD Descartes), SHP (ESRI Shapefile), JPEG, GIF, FLT (grid data), WKB, Simple Features and GML (OpenGIS)

Open a pre-defined map description MWI

Data output

Display data records on mouse-over

Edit options for display of data records

Print map

Save map as image

Save application

Select objects by mouse-click

Find objects according to specified attribute values

Display table with all objects

---

3 Architecture

The overall system architecture is depicted in Figure 1. The many dots (…) in the picture symbolize the extensibility of the system in various dimensions. In particular, one may add new visualization methods, which can be put under the highlighting/brushing regime of a central supervisor component, which receives mouse pointing or clicking events from registered displays and broadcasts them to other registered displays.

4 Applications

CommonGIS/Descartes is being used in various applications developed inside and outside our organization. For example, it was the basis for the visualization and analysis of statistical and other thematic data of

- Several cities, such as Bonn, Helsinki, Tilburg;
- Eurostat, the European Statistical Office; here also the visualization of time-series data was addressed.
- The UK National Statistical Office; addressing the visualization of the previous census decades.

- European Forest Institute, including statistical data from various sources such as Eurostat, European National forest institutes.
- German Office for Nature Protection (BfN) in project “Naturdetektive”: German children are requested to enter and analyze observations on interactive maps, and to study the flyways of GPS equipped birds (storcks and cranes).

Some of these and further applications can be found as live on-line examples at http://borneo.gmd.de/and/java/iris/.

Some Publications

Andrienko, G. and Andrienko, N. (1999a) Interactive Maps for Visual Data Exploration. *International Journal Geographical Information Science*. v.**13** (4), pp.355-374.

Andrienko, G. and Andrienko, N. (1999b) Knowledge Engineering for Automated Map Design in DESCARTES. In C.B.Medeiros (ed.) Advances in Geographic Information Systems. Proceedings of the 7th International Symposium ACM GIS'99, Kansas-City, November 5-6, 1999. NY: ACM Press, pp.66-72.

Andrienko, G. and Andrienko, N. (1999c) Data Characterization Schema for Intelligent Support in Visual Data Analysis. In Freksa, C., & Mark, D. M. (eds.) Spatial information theory - Cognitive and computational foundations of geographic information science COSIT'99, Lecture Notes in Computer Science, vol. 1661. Berlin: Springer, pp. 349-366.

Andrienko, G. and Andrienko, N. (1999d) Making a GIS Intelligent: CommonGIS Project View. AGILE'99 Conference, Rome, April 15-17, 1999, pp.19-24.

Andrienko, G. and Andrienko, N. (2001a) Exploring Spatial Data with Dominant Attribute Map and Parallel Coordinates. *Computers, Environment and Urban Systems*. v.**25** (1), pp.5-15.

Andrienko, G. and Andrienko, N. (2001b) Interactive Cumulative Curves as a Tool for Exploratory Classification. In D.B.Kidner, G.Higgs (Eds.) *9th annual conference GIS Research in the UK*, University of Glamorgan, Wales, April 18-20, 2001. University of Glamorgan, 2001, pp.439-442.

Andrienko, G. and Andrienko, N. (2001c) Constructing Parallel Coordinates Plot for Problem Solving. In A.Butz, A.Krueger, P.Oliver, and M.Zhou (Eds.) *1st International Symposium on Smart Graphics*, Hawthorne, New York, USA, March 21-23, 2001. ACM Press, 2001, pp.9-14.

Andrienko, G. and Andrienko, N. (2001d) Interactive Visual Tools to Support Spatial Multicriteria Decision Making. Proceedings UIDIS 2001, Zurich. IEEE Computer Society Press.

Andrienko, G., Andrienko, N., Voss, H., and Carter, J. (1999): Internet Mapping for Dissemination of Statistical Information, Computers, Environment and Urban Systems (Elsevier Science), 1999, v.23 (6), pp.425-441, ISSN 0198-9715

Jankowski, P., Andrienko, N., and Andrienko, G. (2001) Map-Centered Exploratory Approach to Multiple Criteria Spatial Decision Making. *International Journal Geographical Information Science.* v.**15** (2), pp.101-127.

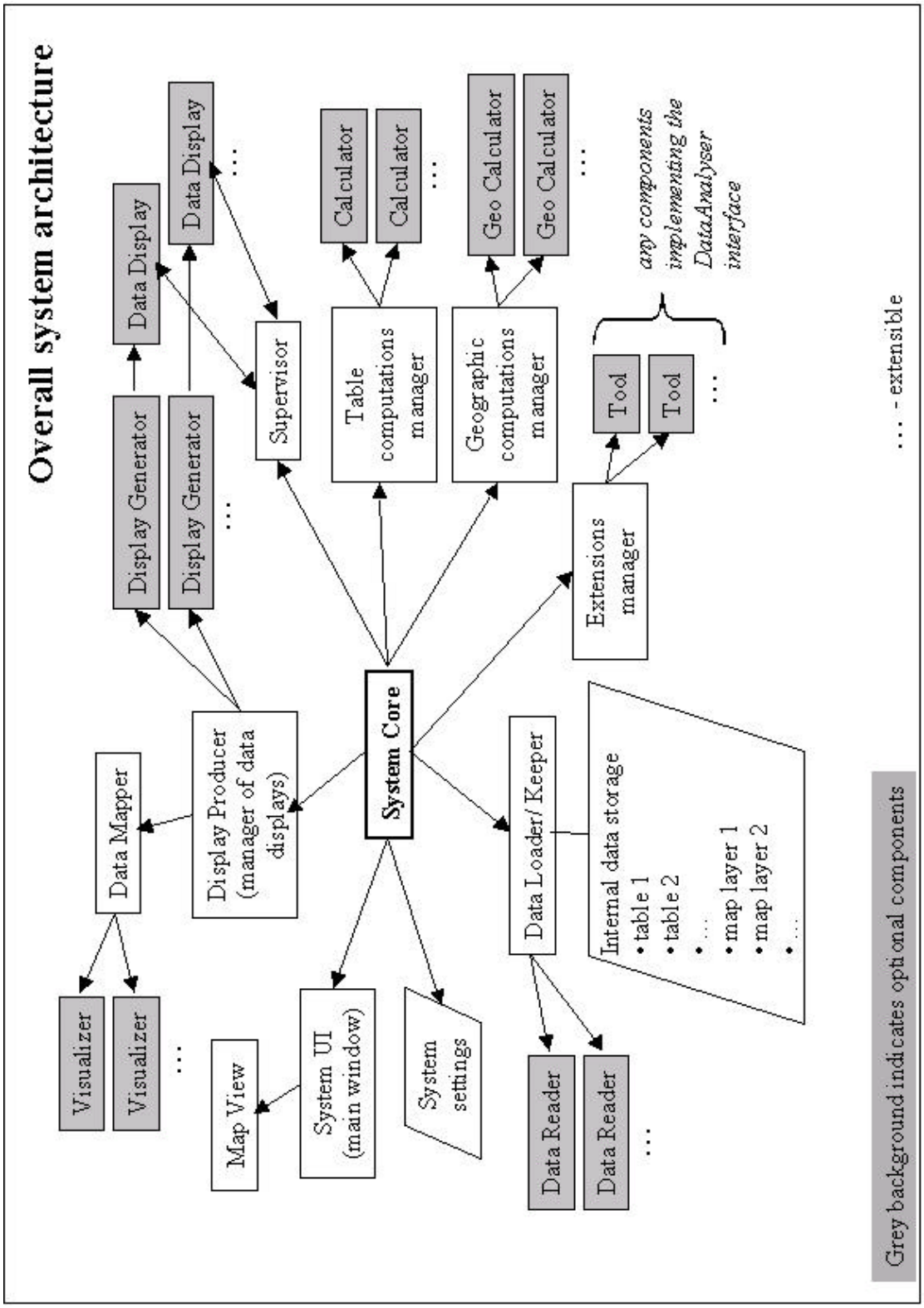


Figure 1. Overall System Architecture

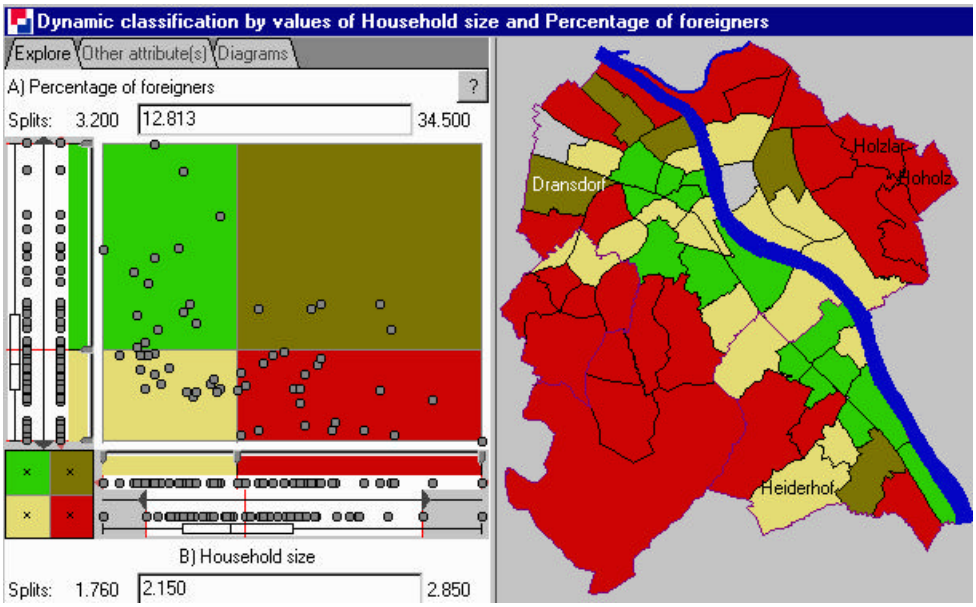


Fig. 2: A scatter plot showing a weak inverse correlation between percentage of foreigners and mean household size for the City of Bonn.

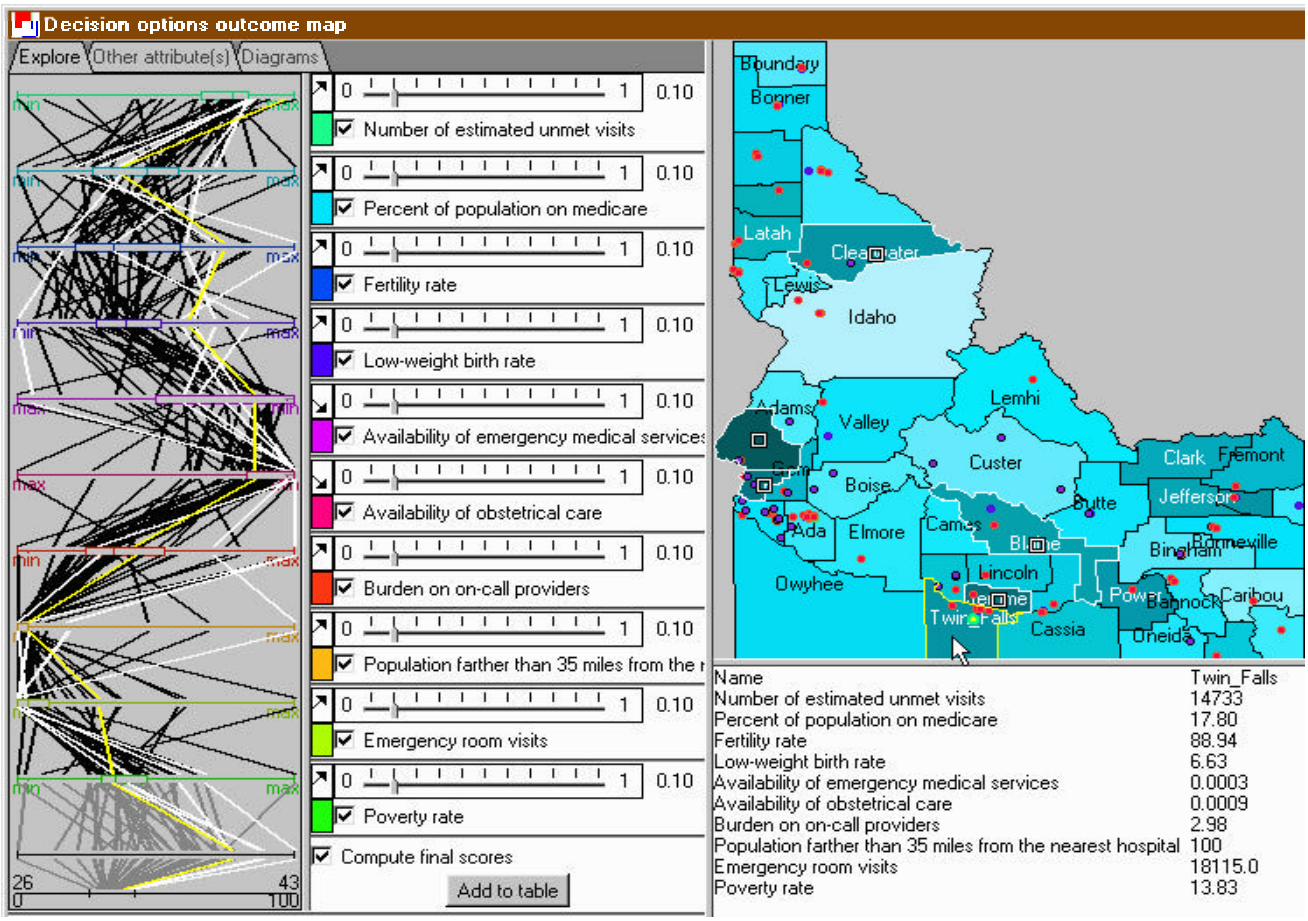


Fig.3: Parallel coordinate plot with prioritized attributes for counties of Idaho. The map shows the ranking of the counties according to the integrated weighted criteria.

LAND USE CHANGE EXPLORER: A TOOL FOR GEOGRAPHIC KNOWLEDGE DISCOVERY

Monica Wachowicz, Xu Ying, and Arend Ligtenberg
Wageningen UR
Centre for Geo-Information

Droevendaalsesteeg 3
PO BOX 47
6700 AA Wageningen
The Netherlands
http://www.geo-informatie.nl/

The main goal of a Geographic Knowledge Discovery (GKD) process is to identify, associate, and understand interesting and unanticipated spatio-temporal patterns in very large data sets that can be used to infer the location, identity, and relationships among spatial objects and events. Every GKD process has its unique goals and characteristics, to which interaction forms and visual representations need to be tailored according to specific user needs for the exploration, synthesis, confirmation, or presentation of spatio-temporal patterns. As with any knowledge discovery process, GKD is characterized as a multi-step process in which data mining plays a central role (Fayyad *et al.* 1996). Data mining refers to the application of algorithms for uncovering patterns in very large databases. Data mining techniques have been developed to perform tasks such as segmentation (clustering, classification), dependency analysis, deviation/outlier analysis, and trend detection (Miller and Han 2001). Mining for spatial dependency involves finding patterns in the form of rules to predict the value of some attribute based on the value of other attributes, taking into account that the values of attributes of nearby spatial objects tend to systematically affect each other (Chawla *et al*. 2001). For example, we may be interested in describing a given area by finding association rules among zones of metropolitan influence on the land market; since they reflect differences in the market status of the land due to its present land uses. On the other hand, mining for temporal dependency involves finding meaningful time-related rules such as the valid time periods during which association rules hold, or the discovery of certain periodicity that association rules have. Roddick and Spiliopoulou (1999) provide a useful bibliography review of spatio-temporal data mining research.

Traditional spatial analysis tools are inadequate for handling the increasing data availability and the complexity of mining spatio-temporal patterns within a GKD process. Geographic Knowledge Discovery represents an important research direction in the development of new generation of spatial analysis tools. Some attempts on developing methods an associated tools for a GKD process have already shown how difficult is to make use of appropriate interaction forms, visual representations and mining tasks in order to allow users to dynamically construct geographic knowledge. Adrienko and Adrienko (1999) have shown the use of dynamic and interactive cartographic representations by different users who are allowed to generate on-the-fly maps according to their individual cognitive abilities and understanding of the problem domain. Another example is the development of taxonomy of GKD operations to facilitate a GKD process for clustering time-series data (MacEachren *et al.* 1999, Wachowicz 2001). The outcomes have shown the potential role of these operations in “process tracking” (using visual operations that display key aspects of the process as it unfolds) and “process steering” (using mining operations of a GKD process as it unfolds, thus changing outcomes on the fly).

The development of methods and tools for supporting a GKD process has introduced new problems resulting from the complexity of the process of exploring data in space and time. From a system perspective, the issues are related to effectively support user-data interactions in both spatial and temporal domains, as well as the development of useful interface metaphors that can support interactive visual representations and data mining tasks in order to amplify user cognition. From a user perspective, the main issue is to make a GKD process very flexible and facilitate intuitive exploration of spatio-temporal patterns using very large data sets.

This paper describes a tool prototype (Land Use Change Explorer), which was developed to allow different users to generate "GKD process tracking” (discovering land use change patterns) and “GKD process steering" (understanding uncertainty in these patterns as the process unfolds). The Land Use Change Explorer was implemented based on the GeoInsignt approach, previously described in

Wachowicz (2001). From a system perspective, the prototype allows users to perform a series of steps of a GKD process, from selecting appropriate data sets; selecting proper representations to visualize land use change information; choosing sequences of mining operations and GIS functionalities. From a user perspective, two types of users are used for the demonstration of the prototype. They represent two stakeholders: an agricultural policy maker and an urban planner.

The prototype was implemented using a Multi-Agent System (MAS) model based on a client/server architecture. At the server side, different resources are available such as different land use databases, land use change rules (evolution rules), data mining algorithms and GIS functionalities. The client side consists of a Graphical User Interface (GUI) that allows users to connect and apply the resources available at the server side. The MAS is able to receive the request from the client, connect to the server to perform the operations using the required resource, and present the results back to the GUI. If we put all these components together as displayed in Figure 1, MAS can be considered as the key technology for the support of a GKD process.

Java programming language was used for the implementation of the GUI for the Land Use Change Explorer (http://jave.sun.com). The Bee-gent system was selected for the MAS implementation (http://www2.toshiba.co.jp/beegent/index.htm). The system is composed of two types of agents: the agent wrapper, which wrappers local functions that can be client application or server resource utility, and the mediation agent that migrates around the network and perform tasks by interacting with the agent wrapper. Both the mediation agent and the agent wrapper can communicate via the Agent Communication Language, which allows the processing by inferring the intention of the requirements of different agents or request information to them. ArcSDE was used for performing the GIS functionalities since includes a Java ODE application programming interface that extends ArcInfo in cross-platform applications (http://www.esri.com). And finally, the SGI Mine Set data mining tool was used to perform the data mining operations (http://www.sgi.com)

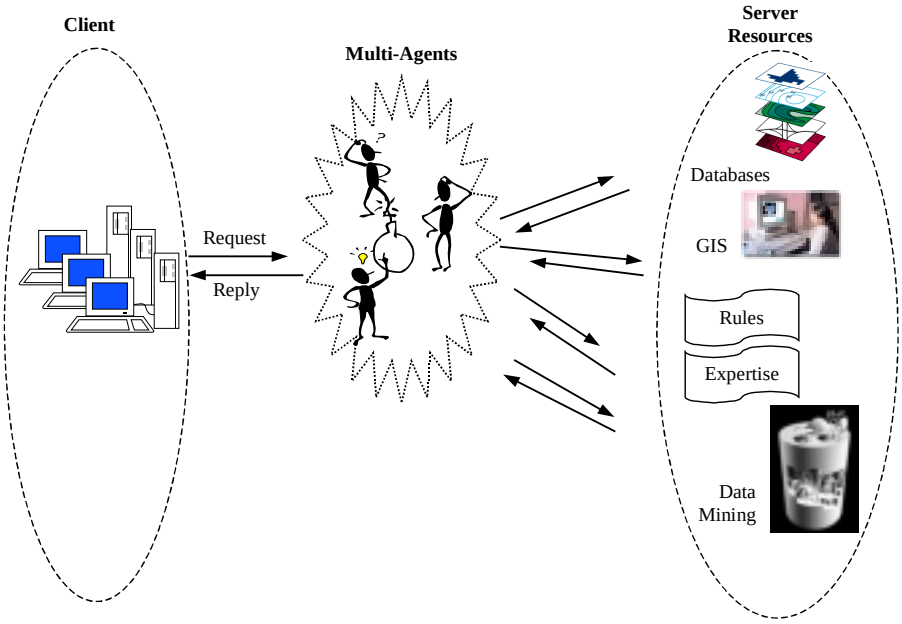


Figure 1 - The client-server architecture for the Land Use Change Explorer

In the Land Use Change Explorer, the agents can:

- Interact with the users;
- Incorporate users preferences in a GKD process;
- Perform queries related to land use changes;
- Display different visual representations of land use change patterns;
- Assist the user in the different steps of a GKD process (select a data set selection, perform data transformation, selection of a data mining task and the respective data mining algorithm, a GIS function or query);
- Provide reliability information of land use change patterns to a specific user;
- Hide cumbersome operations from users;

Six interface metaphors compose the Land Use Change Explorer: the menu bar, database panel, map information panel, more information panel, status panel and the map window (Figure 2). The ***Menu Bar*** provides the following functionalities:

- Database selection. It allows the user to select the databases available at the server. In this prototype, two types of land use databases are available with information about land use in the Netherlands over the last 10 years.
- User identification. Once the agent has identified the type of a user as agriculture policy maker or urban planner, different menus will pop up correspondingly (see Figure 3 and 4).
- Scale selection. At global scale the GKD process focuses on finding changes only at the super-classes level. At local scale, the GKD process uses the data available at class level and sub-class level. Users can decide about which scale is more appropriate to explore land use change, since patterns are scale dependent.

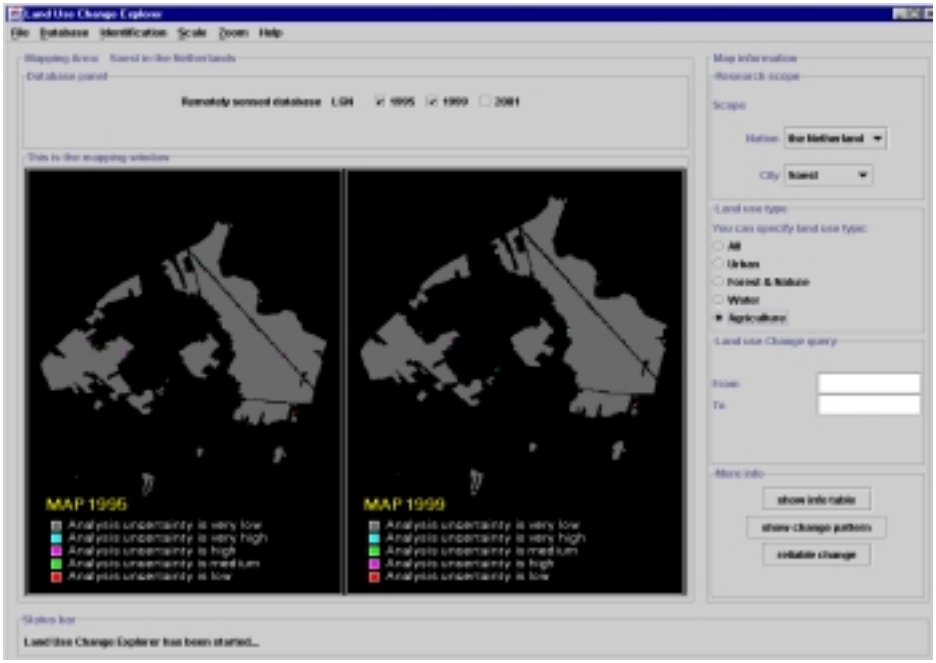


Figure 2 - An example of the GUI implemented for the Land Use Change Explorer

The ***Database Panel*** informs the user what are the "activated" data sets, which are being used in the GKD process. The panel allows the user to select the years for the analysis at any time. The ***Map Information Panel*** provides the user with the selection of the area of interest (using nation and city parameters) and the land use type(s) of interest. All the land use types, which exist in the selected time series, will be displayed in the ***Map Window*** and the GKD process will be carried out using the selected land use types. The ***More Information Panel*** is mainly used to assist the user in the GKD operations such as the formulation of a query, GIS functionality, or data mining operation. There are three buttons:

- 'show info table' button that will trigger an agent to provide available information available about land use;
- the 'show change pattern' button that will trigger an agent to compute and visualize the discovered patterns in the Map Window; and
- the 'reliable change' button that will trigger an agent to compute and visualize the uncertainty of land use change patterns.

The ***Status Bar*** informs the user about the background situation, such as 'ArcInfo is doing analysis operation', 'Interface is displaying map, please wait…' and so on. Once an agent has identified the type of the user, the Agricultural Land Use Change Explorer will be activated for presenting agricultural land use change to agriculture policy makers (See Figure 3). The main metaphors of this interface are the mapping window in the middle and the query formulation panel at the bottom-right side. The mapping window is divided into two parts in order to compare different

changes. Query formulation panel provides information about spatial and thematic changes and their respective reliability.

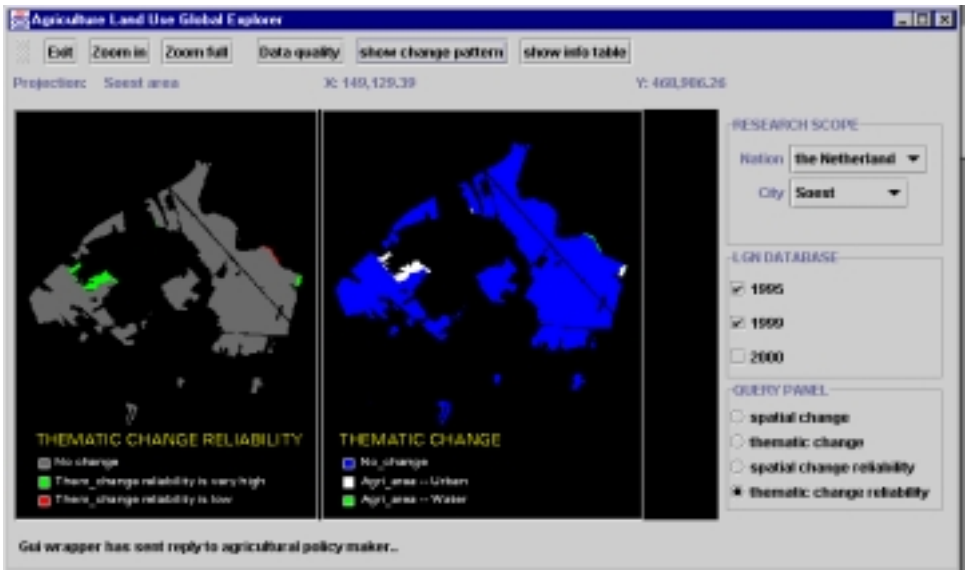


Figure 3 - Agriculture Land Use Change Explorer at global scale

The Urban Land Use Change Explorer (Figure 4) is used to present land use changes that have occurred within an urban area. Therefore, according to the urban planner experience, a specific land use database is used to find land use change patterns. The design of this interface is similar to Agriculture Land Use Change Explorer but much simpler due to preferences of the urban planner. Stakeholders' perspectives were embedded in the GKD process from the selection of suitable data source, mining technique for the change analysis and visualization technique for the land use change patterns.

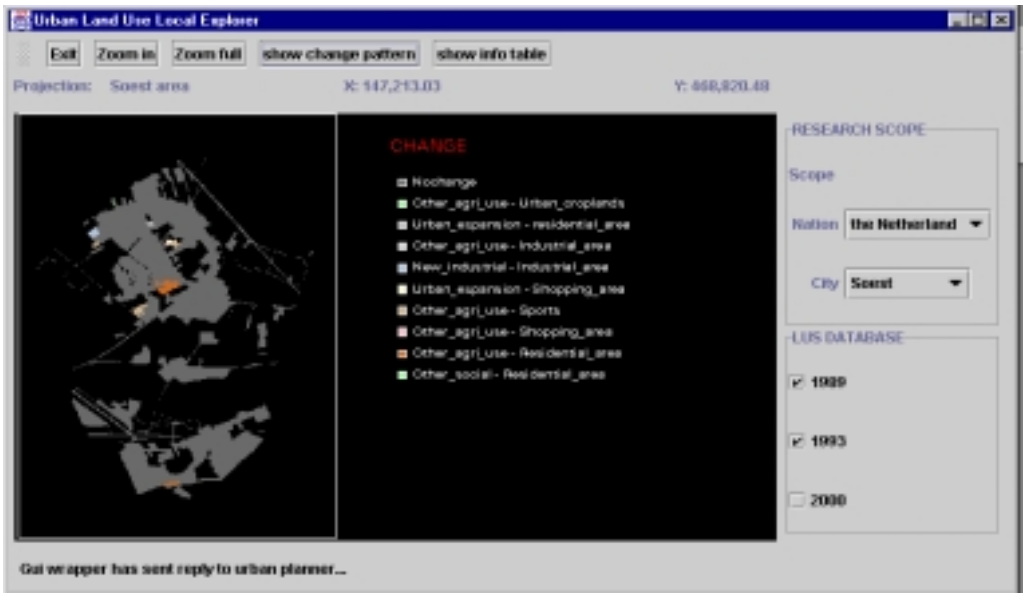


Figure 4 - The Urban Land Use Change Explorer at local scale

Future Directions

In the current prototype, all user-data interactions in both spatial and temporal domains, as well as the interface metaphors developed for the support of visual representations (categorical maps) and data mining tasks actually assume the presence of relational models for the land use databases. As advanced database systems, such as object oriented, deductive, and activate databases are being developed for GIS; methods and tools for supporting a GKD process need to be extended. Therefore,

our next step is to look at how geographic knowledge can be mined from these databases. A key related research issue is the design of a spatio-temporal mining query language that could be supported by the Graphical User Interface and make the process of query creation much easier. The language should be powerful enough to cover the number of data mining algorithms and a variety of formats of existing data sets.

Finally, our next step will be towards the development of multidimensional rule visualization techniques. One of the most effective ways of understanding association, evolution, or classification rules is through interactive visual representations. Multidimensional data visualization techniques have already been proposed in the literature (Inselberg and Avidan 1999, Keim and Kriegel 1994), but multidimensional rule visualization is still in its infancy.

References

Adrienko, G.L. and Adrienko, N.V. (1999). Interactive maps for visual data exploration. *International Journal of Geographical Information Science*,**13**(4), 355-374.

Chawla, S., Shekhar, S. Wu, W. and Ozesmi, U. (2001). Modelling Spatial Dependencies for mining geospatial data. In: *Geographic data mining and knowledge discovery* (Eds. Miller, H. J. and Han, J.), London: Taylor & Francis, pp.131-159.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, Vol. 39, No. 11, pp. 27-34.

Inselberg, A. and Avidan, T. (1999). The automated multidimensional detective. IEEE *InfoVis'99*, Oct 24-29, San Francisco, CA, pp. 112-119.

Keim, D. and Kriegel, H.-P.(1994). VisDB: Database exploration using multidimensional visualization. *Computer Graphics and Applications*, September 1994, pp. 44-49.

MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D. and Masters, R. (1999). Constructing knowledge from multivariate spatio-temporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographic Information Science*, Vol. 13, No. 4, pp. 311-334.

Miller, H.J. and Han, J. (2001). *Geographic Data Mining and Knowledge Discovery*. London: Taylor and Francis.

Roddick, J. F. and Spiliopoulou, M. (1999). A bibliography of temporal, spatial and spatio-temporal data mining research. SIGKDD Explorations. Vol. 1, No. 1, (in press) URL: http://www.cis.unisa.edu.au/~cisjfr/STDMPapers/.

Wachowicz, M.. (2001). GeoInsight: an approach for developing a knowledge construction process based on the integration of GVis and KDD methods. In: *Geographic data mining and knowledge discovery* (Eds. Miller, H. J. and Han, J.), London: Taylor & Francis, pp.239-259.

ChoroWare: A Software Toolkit for Choropleth Map Classifications

Ningchuan Xiao, Marc P. Armstrong and David A. Bennett

Department of Geography
The University of Iowa
Iowa City, IA 52242
E-mail: {ningchuan-xiao; marc-armstrong; david-bennett}@uiowa.edu

When exploratory spatial data analyses are conducted, choropleth mapping is an important means to display 1) the patterns of the spatial observations for enumeration areas, such as counties or Census Tracts, and 2) the resulting spatial statistical outcomes. Choropleth maps group spatial observations into classes and each class is then assigned a particular color that is used to shade the enumeration areas. The classifications can be performed in a large number of ways, with each method focusing on a specific criterion, such as within-class similarity (Jenks and Caspall 1971), geographical structure (Monmonier 1972; Murray and Shyy 2000), and statistical characteristics (e.g., quantiles) (Slocum 1999: 67).

It is unlikely, however, that a particular classification will satisfy all criteria. In practice, a classification that is optimal for one criterion may turn out to be poor from other perspectives. Therefore, an appropriate design goal for a choropleth map is to consider classification as a multicriteria (or multiobjective) problem. That is, instead of looking for a single classification that is best for all criteria, it is more useful for cartographers to examine the classifications along a Pareto-like front where no classification can be considered to be better than, or dominate, others.

The purpose of this paper is to describe the design and implementation of a software toolkit, called ChoroWare, that can be used to help cartographers find a set of class intervals that is suitable for a specific application. Using ChoroWare, users can explore a variety of alternative classifications and select one that they deem most suitable. To achieve this goal, two critical tasks must be carried out. First, ChoroWare must be able to generate the set of nondominated alternatives that forms the Pareto front. Then, a visualization tool must be designed to allow users to interactively display the trade-offs between criteria and the resulting choropleth map for each non-dominated alternative. The overall architecture of ChoroWare is illustrated in Figure 1. The functionality of each module is discussed below.

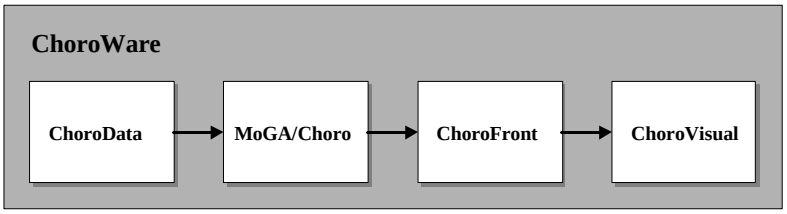


Figure 1. Overall structure of ChoroWare.

Generating nondominated alternatives is difficult for many multicriteria problems (Cohen 1978). Recent developments, however, have shown that genetic algorithms (GAs) are able to generate alternatives for multicriteria problems (Zitzler et al. 2001; Xiao et al. 2002). In a GA, a population of individual solutions (i.e., a classification for a choropleth map) is randomly initialized and then manipulated by a set of iterative operations, including selection, recombination, and mutation. At the end of each iteration, solutions are evaluated according to a set of objectives. To obtain a diverse population of alternatives, which is critical for multicriteria problems, we extended the original GA island model (see Cantú-Paz and Goldberg 2000) and designed a specialized island model. In our approach, the entire population is divided into several sub-populations ("islands") and each subpopulation is processed using a local set of evolutionary operations. A mechanism, called migration, is used to exchange individuals, at a certain rate and interval, among the islands. Inside each sub-population, an individual (i.e., a specific classification scheme) is evaluated on a (partial) set of objectives.

Based on this principle, we designed a module called MoGA/Choro (multiobjective GA for choroplethic classification) using object-oriented techniques (Figure 2). In this context a *population* contains several *subpopulations*, and each *subpopulation* has many individuals represented using class *genotype*. A *genotype* has a full set of genetic operators and a member datum called *chromosome*, which represents a classification scheme by encoding the break points of the classes into a string of integers. A *genotype* also contains an aggregating object called *models*, which contains the models used to calculate the objective values. We have developed four "built-in" objectives: goodness of variance fit (Robinson et al. 1984: 363), equal area of classes, spatial autocorrelation, and boundary accuracy index (Jenks and Caspall 1971); using this pattern of objects (Figure 2), new objectives can be easily defined and incorporated into the source code of the module. Users can specify which objectives are desired in the configuration file of MoGA/Choro. The object *models* contains an aggregating object called *parameters*, which includes the data needed in the calculations. These data are generated from the raw spatial data (in vector or raster formats) by a module called ChoroData. Data generated by ChoroData include an array of all observations (for vector), a two-dimensional array of observations (for raster), an array of non-repeated observations (for both vector and raster), and an array of linked lists of neighbors for each polygon in a vector data structure.

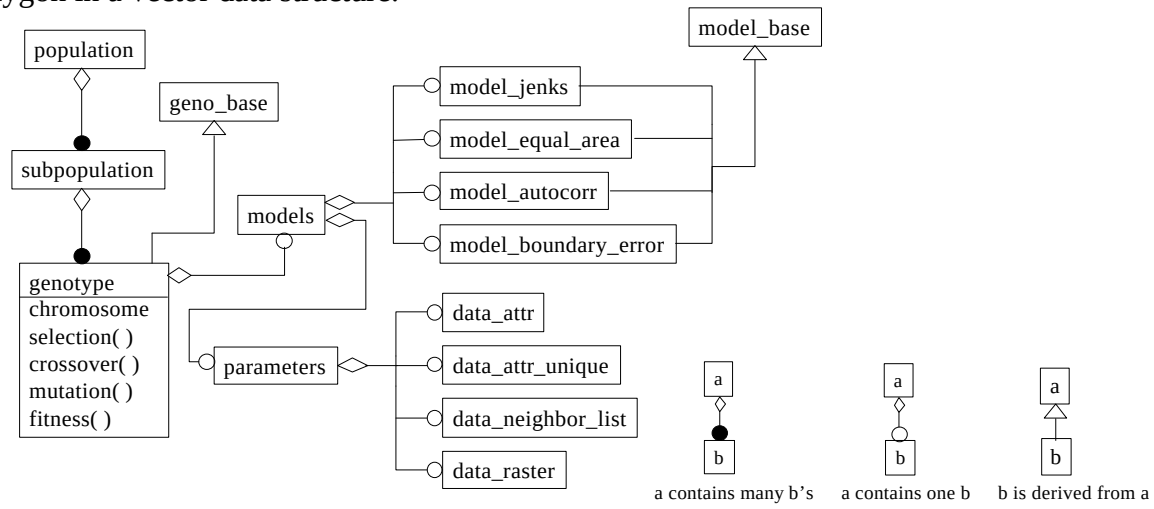


Figure 2. The design pattern of MoGA/Choro.

Within MoGA/Choro, classifications are saved to a file. The nondominated alternatives are then extracted by a module called ChoroFront, and the results are visualized using a module called ChoroVisual. The ChoroVisual tool was developed using GTK+ and GDK (www.gtk.org). It currently consists of six windows to display 1) a map, 2) the attributes of polygons, 3) a value path (also know as parallel coordinate plot), 4) a legend editor (called a classifier, for manual classification), 5) a list of all nondominated alternatives (the window at the upper-left corner marked as “Front List”), and 6) a multivariate plot (Figure 3). In this last window, a user can choose to display a plot formed by any two variables listed in the left panel. Among these variables, the last four are the built-in objectives. Each classification alternative is displayed as a small white square in the solution space drawing area, where a red square is used to indicate the current classification being used to draw the map. The current classification is also highlighted in the front list and in the value path (the red line). Linkages among the map, multivariate plot, and front list enable a user to examine the alternatives on the fly. The purpose of the legend editor is to help users further explore each selected alternative. In the classifier window, a user can examine the histogram of the observations and the break point of the classification, adjust the classification, and change the color of classes. In so doing, a user can start from an acceptable classification and then, based on this classification, try to find a better one.

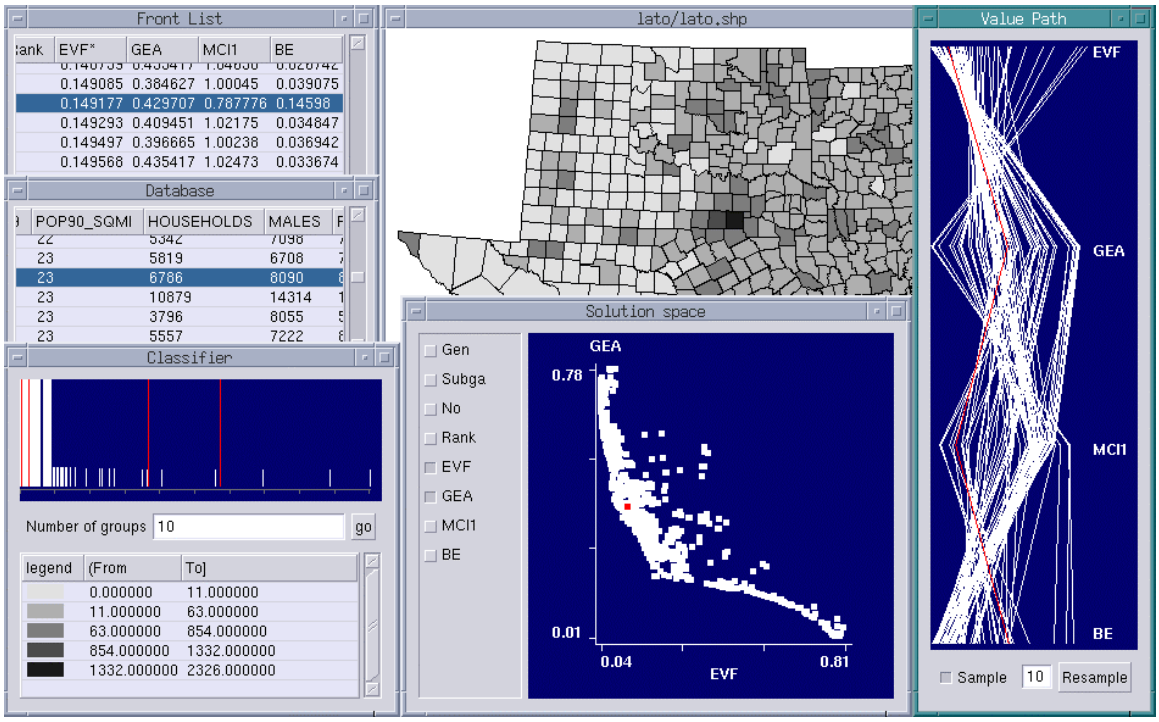


Figure 3. A screenshot of ChoroVisual.

In general, using ChoroWare, cartographers and spatial analysts can have more flexibility in choosing a suitable classification scheme for choropleth maps than current GIS and cartographic software can provide. It is developed using Open Source technology based on Linux. Future developments will focus on integrating ChoroWare with other powerful visualization tools such as GGobi (www.ggobi.org), and developing more modules to help users distinguish alternative classifications.

References

Cantú-Paz, E., and D. E. Goldberg. 2000. Efficient parallel genetic algorithms: theory and practice. *Computer Methods in Applied Mechanics and Engineering* 186:221-238.

Cohon, J. L. 1978. *Multiobjective Programming and Planning*. New York: Academic Press.

Jenks, G. F., and F. C. Caspall. 1971. Error on choroplethic maps: definition, measurement, reduction. *Annals of the Association of American Geographers* 61 (2):217-244.

Monmonier, M. S. 1972. Contiguity-based class-interval selection: a method for simplifying patterns on statistical maps. *The Geographical Review* 62 (2):203-228.

Murray, A. T., and T.-K. Shyy. 2000. Integrating attribute and space characteristics in choropleth display and spatial data mining. *International Journal of Geographical Information Science* 14 (7):649-667.

Robinson, A.H., Sale, R.D., Morrison, J.L., and Muehrcke, P.C. 1984. Elements of Cartography. 5th ed. New York, NY: John Wiley & Sons.

Slocum, T. A. 1999. *Thematic Cartography and Visualization*. Upper Saddle River, NJ: Prentice Hall.

Xiao, N., D. A. Bennett, and M. P. Armstrong. 2002. Using evolutionary algorithms to generate alternatives for multiobjective site search problems. *Environment and Planning A* in press.

Zitzler, E., K. Deb, L. Thiele, C. A. C. Coello, and D. Corne, eds. 2001. *Evolutionary Multi-Criterion Optimization: Proceedings of the First International Conference*. Berlin: Springer.

A GIS prototype for process-based analysis

by
May Yuan
Department of Geography
The University of Oklahoma

Over the years, GIS technology has evolved from automatic mapping tools to a facility enabling complex information queries and computing. In addition to support for direct retrieval of data records, a GIS should be able to provide information that summarizes data characteristics for a better understanding of geographic processes that generate the data. However, most GIS software packages offer spatial analytical functions that are based on geometrical objects, e.g. points, lines, polygons, and cells. While some of these functions are powerful to examine spatial patterns, derive spatial relationships, and transform spatial distributions, they soon reach their limits on investigating processes. Arguably, processes are often the primary concern in social or physical sciences because they are the driving forces that shape our environments. This paper argues that it is of paramount importance to develop analytical tools for process-based analysis. It presents a prototype designed to provide preliminary analysis of process characteristics and behavior. It ends with suggestions for further improvements of process-based analysis.

A geographic process is an integral in space and time, a continuing development involving changes. Analysis of geographic processes has two aspects. First, analysis of individual processes as to how each process evolves and how it influences on the human and physical environments. The analysis should not only examine what happens in a particular place and time, but the transition from one state to another. Therefore, it requires analytical tools that can identify changes and relate them in a meaningful way. Such analytical tools are lacking. While there are software tools developed for the analysis of human behavior as accessibility or travel patterns (Kwan 2000), tools are lacking for tracking land-use and land-cover change, for example, and examining the rate of change, and how that change evolves in space and time. The other aspect of processes relates to the spatial patterns and temporal trends of a certain kind of processes. There are many statistical methods for spatial or temporal analysis, but few, if any, can summarize process characteristics and behaviors. One problem pertains to the fact that many spatiotemporal analysis methods are more or less point-based, such as Geographical Analysis Machine (Openshaw 1987), and many geographic processes are not appropriate to be represented as 0-dimensional objects. For example, wildfire involves burned and burning areas. The front of a wildfire may be conceptualized as a linear feature, but use points to represent wildfire will overlook the importance of heterogeneity in burns and fire-environment interactions.

An emergent challenge is the interplays between individual processes and processes as a whole. Using wildfire as an example again, past fire events bring about the vegetation complexity and influence the fuel types and fuel loading in the environment, which is a determinant to the occurrence of a fire. Once a wildfire occurs, the environment conditions guide the spread of the fire, and the fire in turn further modifies

the environment. The modified environment then results in different fire potential and fire behavior.

Process-based analysis requires tools that can relate spatial extents of a process over time, compute spatial and temporal statistics of these related fields, and summarize spatiotemporal patterns of all processes of the same type. Furthermore, it is desirable to have tools that can relate process of different kinds in space and time, so that spatiotemporal relationships among processes, such as teleconnection, can be discerned. A preliminary prototype has been developed in ArcInfo GIS for proof of concepts. Data of hourly digital precipitation arrays were used to extract processes of storms. Each process consists of sequential precipitation regions that represent how a storm propagates from one area to another. The system stores the data in three hierarchical tiers: states, processes, and events. States represent precipitation distributed at a given time. Areas of precipitation were extracted and connected to form processes. Because a weather event, such as a frontal pass, may result in multiple isolated storms, these storms were organized into an event in the database (Yuan 2001).

The primary characteristics of geographic processes include movement (rate and path) and their interactions with the environment. A simplified algorithm was used to compute the movement of storms using distance between the precipitation-weighted centers of states and the time at each state. Figure 1 represents a sample output of the analysis. It shows that two primary directions of storm movements are north-west to south-east and west to north-east. Rate of movement was up to 61 km per hour while most storms moved at a rate of 40 km per hour. The primary directions can be related to mesoscale weather patterns. A further analysis of the relationships between storm movements and weather patterns can reveal information useful for storm prediction and management tactics.

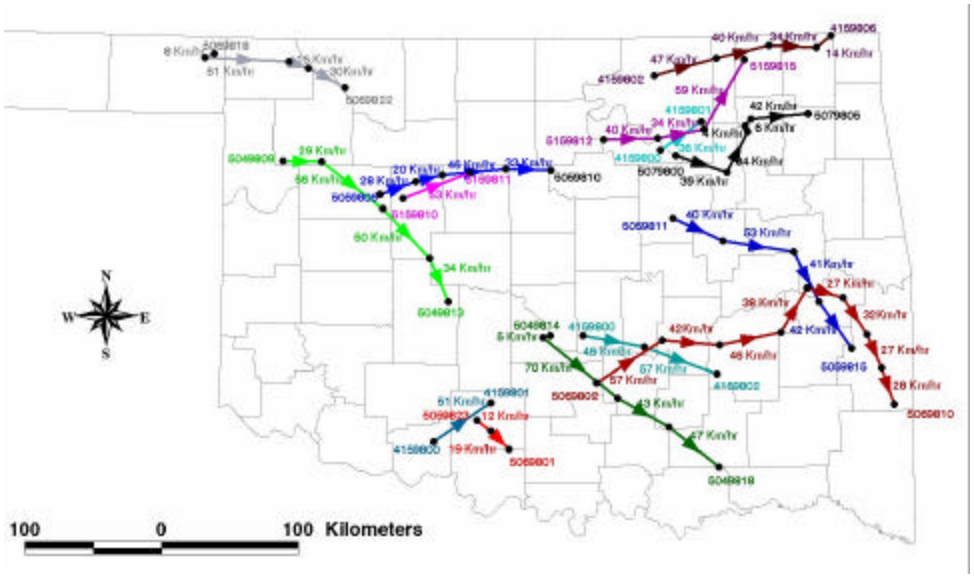
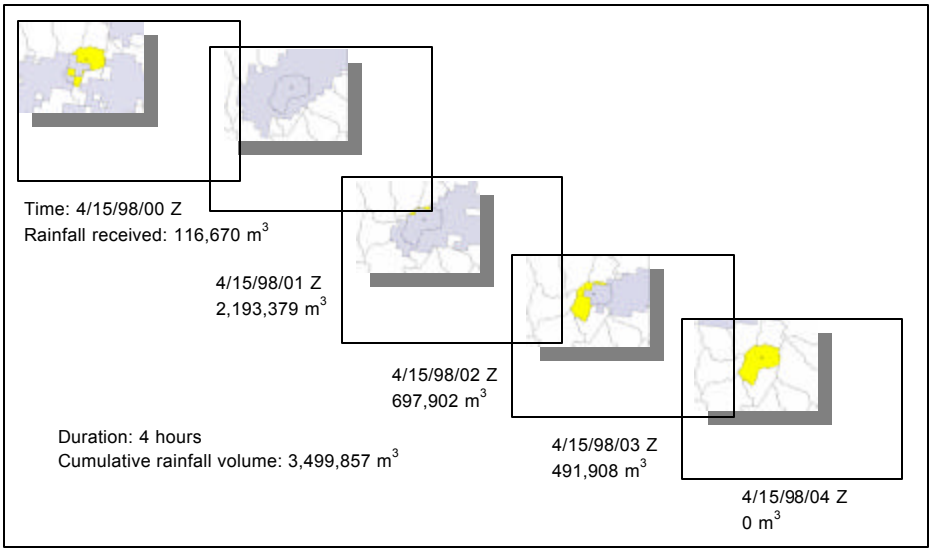


Figure 1: Movements of storms.

Another example to demonstrate the usefulness of process-based analysis tools is summation of process-environment interaction. Figure 2 shows a result of a storm passing a watershed and the amount of precipitation received in the watershed at different time. It also summarizes the total duration of the storm lasting in the watershed and the total accumulative precipitation received. The analysis treats the storm as a process and tracks the storm's behavior (through states that show precipitation variations inside the storm) and interactions with the environment (e.g. the watershed).



The two examples discussed above are simple cases of process-based analysis. There are plenty opportunities to develop more sophisticated tools to facilitate a better understanding of processes that shape our world. Geographic processes are multi-dimensional and interact at multiple spatial and temporal scales. Wildfire presents one of the most challenging cases (Yuan 1997) for its overarching behavior and environmental effects across spatial and temporal scales (Figure 3). The environment settings (fuel, moisture, and weather) suggest fire potential which sets forth a stage to guide fire spread (behavior modeling). Results of fire spread imprint burns in the environment, and with multiple burns over a period of time, wildfire transforms the environment to a new landscape mosaic, which in turn modifies the environment's potential for wildfire. The figure also illustrates information support among wildfire research activities. In this information cycle, fire forecasting generates data useful to simulate fire behaviors, results of fire behaviors (burns) produce data useful to analyze fire effects, and fire effects imprint data useful to construct fire histories, and historical fires reset the landscape of fire potential. Therefore, a comprehensive set of analytical tools for wildfire processes should encompass functions to relate information across different data models.

Furthermore, analysis of geographic processes must be multi-scalar because processes are operating at different scales and they are dynamically and functionally related (Ahl and Allen 1996). Processes at a lower scale are controlled by others at a higher scale, and the dynamics of processes at a larger scale are situated by processes at a

lower scale. Analytical tools that can discern relationships across scales and enable visualization of dynamics among processes can greatly facilitate a better understanding and prediction of geographic worlds.

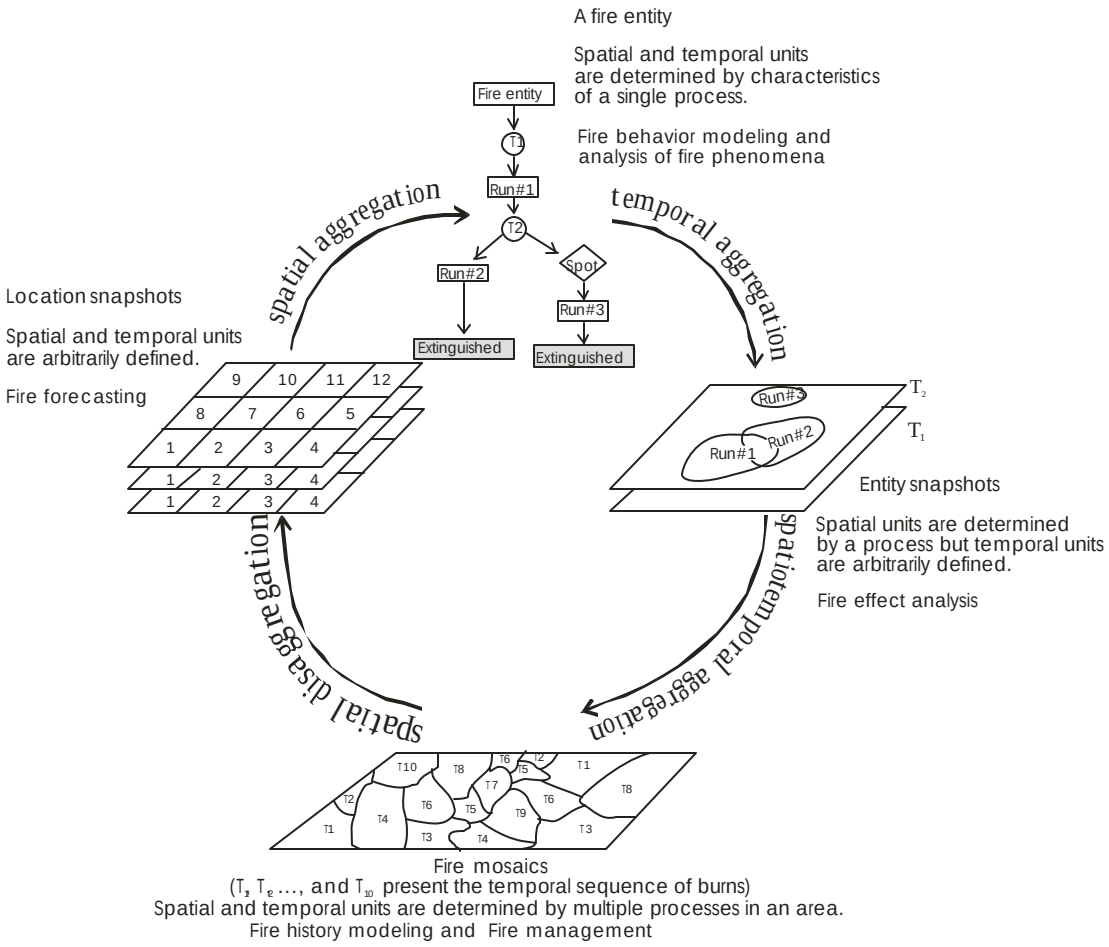


Figure 3: A wildfire information cycle

References

Ahl, V. and T. F. H. Allen (1996). *Hierarchy Theory: A Vision, Vocabulary, and Epistemology*. New York, Columbia University Press.

Kwan, M.-P. (2000). Analysis of human spatial behavior in a GIS environment: Recent developments and future prospects. *Journal of Geographical Systems* **2**: 85-90.

Openshaw, S. (1987). An automated geographical analysis system. *Environment and Planning A* **19**: 431-436.

Yuan, M. (1997). Use of Knowledge Acquisition to Build Wildfire Representation in Geographic Information Systems. *International Journal of Geographical Information Science* **11**(8): 723-745.

Yuan, M. (2001). Representing Complex Geographic Phenomena with both Object- and Field-like Properties. *Cartography and Geographic Information Science* **28**(2): 83-96.

An Object Oriented Framework for Integration of Social Science and Natural Resource Management Tools

Fernanda Zermoglio, John Corbett and Stewart Collis
Mud Springs Geographers Inc.
18 South Main, Suite 718
Temple, Texas, 76501
(fernanda@mudsprings.com).

Delivering information to decision makers is often a bottleneck in turning research results into effective decisions and policies. Information technology in social science research is maturing to provide more sophisticated tools for analyzing, visualizing, managing, and disseminating information. Successful integration and dissemination is dependent on creating flexible and scalable software frameworks that provide both complex analysis tools for advanced users and deliver information to a wider audience. Social science studies are increasingly linked to a specific geographic location (e.g., GPS coordinate) or region (an agro-ecological zone or a political region) and organization of this complex socio-economic data requires a sound data management scheme. Geographic Information Systems (GIS) technology is a logical basis for such a Spatial Decision Support System (SDSS). The principal goal of these systems is to store and manage information for any research project, model, or study in a systematic way that is both useful to researchers and accessible to those other than researchers or technicians.

AWhere™-ACT represents an implementation of an SDSS providing effective analysis and information delivery tools to decision makers who are not GIS specialists. Using object oriented methodologies and Component Object Model (COM) technology AWhere™-ACT provides a scalable framework for efficient integration with other systems (for example, demographic models, epidemiological models and weather generators) and databases (such as geo-referenced documents and remotely sensed time series data). The AWhere™-ACT components (as can all COM compliant objects) can be reassembled into specific applications to meet tailored needs or can be incorporated piecemeal into wholly separate applications. This versatility and flexibility without intellectual property compromise, offers a highly effective and efficient mechanism to share scientific advances and contribute to decision makers information synthesis needs.

Geographic Information Systems (GIS) are an established technology in the environmental management and research arena. However, adoption by non-specialists has not been widespread. The software system AWhere™-ACT[1] brings GIS a step closer to the decision maker whilst providing a sophisticated underlying structure that experts can use for modeling purposes. AWhere™-ACT is a spatial decision support system built using object oriented techniques to manage, analyze, visualize and disseminate information to a broad range of users. The stand-alone software framework for AWhere™-ACT contains data specific modules, which utilize, where appropriate, the same underlying object oriented structure. This approach has proven to be a successful implementation of an information delivery tool that can help users make better policy decisions related to agriculture and natural resource management.

One underlying principle in promoting effective use of any spatial information system is the concept of a *foundation database*. There are volumes of data available to decision makers from various sources, many of which are available via the Internet for free or at minimal cost. This however is of little use to a non-technician who neither has the time to track down such data sources or the expertise to reformat, re-project and load them into appropriate spatial information system tools. As such, an effective spatial decision support tool must come with pre-packaged data that allows at least initial analyses to be done, with no need to seek external data sources. As a user becomes more proficient, or requires more detailed information, the ability to add custom data sets must also be provided. Simply providing data management tools will not address the need of non-GIS specialists. AWhere™-ACT comes packaged with extensive databases, and each layer is thoroughly documented with metadata if further information is required. Databases delivered using AWhere™-ACT have predominately been eastern and southern African countries including Kenya, Ethiopia, Zimbabwe and Botswana and more recently parts of the United States and Mexico. The thematic groupings of data layers are diverse and include climatic, infrastructure, ecological, topographic,

[1] ACT is an acronym for Almanac Characterization Tool.

demographic and hydrographic layers. To manage the diverse nature of these layers an object oriented data model was developed. This design is key to delivering such a variety of information in a usable form.
Essential to the creation of a decision support system that will ultimately have an impact is the framework used to deliver the tools to end-users. This framework must provide a simple interface to end-users but have the potential to provide for the needs of more advanced users. The AWhere™-ACT Shell addresses these issues by utilizing COM technology to package applications and data models in a user friendly graphical user interface (GUI) (client or presentation layer) and provides advanced level access to the underlying components and objects (business and database layers).

To reach a largely non-technical audience the AWhere™-ACT client interface is modeled on the Microsoft Outlook™ interface for familiarity and function (Figure 1). The listbar on the left contains a set of "bars" that represent a separate COM applications or modules. Within each bar is a collection of items that refer to a function within each module. This design has the advantage of delivering a variety of applications to end users with a consistent interface. Interface design considerations are essential for application and information delivery and end user acceptance. There are also significant advantages to using a COM software framework to deliver multiple applications that relate to the interaction between these modules.

Using COM the many common features of the individual AWhere™-ACT applications such as toolbars, listbars, and map controls are handled by the AWhere™ Shell (Figure 3). This allows efficiencies in memory and implementation but also provides a common interface design. More importantly, the object models and data models underlying each individual application can interact with one another. This allows new modules to take advantage of the work previously done in existing modules and leads to faster development. This is an enormous advantage for developers who have the capacity to develop custom modules or stand alone applications using the AWhere™-ACT components.

For users who do not have the capacity or time to develop custom COM based modules for the AWhere™ Shell there is an intermediate yet powerful option available. Incorporating Visual Basic for Applications (VBA) in the shell interface provides an interface to the underlying object models of the applications delivered in the shell. This allows programmatic access to the shell's object models in much the same way Microsoft Access or Microsoft Word can be customized by a user with some programming skills. This offers a powerful tool that allows scientists and researchers to concentrate on developing their own temporal models rather than developing data management tools and data models.

COM (Component Object Model) technology enables great efficiencies in software development. Programmers may integrate existing components (e.g., a graph control) into their applications instead of building them from scratch, thus saving an enormous amount of work. There are thousands of robust, commercial COM controls available, and AWhere™-ACT takes full advantage of these (e.g., map, graph and toolbars). Most COM development tools (Visual Basic, Visual C++ etc.) also allow custom building of COM controls. Hence the flexibility of sharing certain components of an application with other developers is open to all developers. For example, AWhere™-ACT consists of a suite of custom built controls that other developers can access and incorporate into their applications. The ReportWhere module of the AWhere™-ACT suite is a stand-alone control that could be compiled as a separate application or integrated into other applications. There is license protection on these components, so the original developers can maintain control and knowledge of who is using their tools (an asset for version control). COM technology opens up enormous cost efficiencies for software development and opportunities for truly integrated collaboration and are the basis for the development of AWhere™-ACT Shell

In order to take advantage of the benefits object oriented programming and COM technology offers AWhere™-ACT uses a three-tier architecture. Three-tiered architecture is an industry standard that provides a framework for logical components of software to interact and enables flexibility in managing changes and updates in system components. The three tiers consist of the database layer, business layer and presentation layer or *client*. A major advantage of this approach is that if the rules (and subsequently the code) of one tier ever change, the programmer need only modify or replace that layer; there is no need to migrate the changes to the other layers [Sarang, 1999]. At all three levels, object models underlie the components at each layer to help organize code and provide structured models with methods and properties to allow communication between layers and to other applications. Where appropriate

the AWhere™-ACT Shell modules are built on a common data model (e.g. AWhere, CensusWhere and SoilWhere share a common data model).

AWhere™-ACT demonstrates a tool that integrates GIS functionality and aspatial tools and applications. AWhere™ Shell is an object-oriented framework for integrating independently developed applications and providing a common user-friendly interface that can be customized to meet user's needs. Lueng [1997] also points out that "A SDSS without knowledge is a tool of no intelligence and minimal usage. The success of a SDSS in general and knowledge-based spatial information system in particular lies on it's level of intelligence". Whilst the intelligence relevant to environmental analysis and modeling underlying AWhere™-ACT have not been described in detail here [Corbett et al., 2000], the framework presented sets the stage for integration of knowledge and information systems. With the computing power available to researchers and developers, the modeling world sees many diverse and very specific applications. While it is unlikely that a single model or tool will resolve all possible concerns of decision-makers, tools like AWhere™-ACT can provide a common foundation from which to begin more specific investigation.

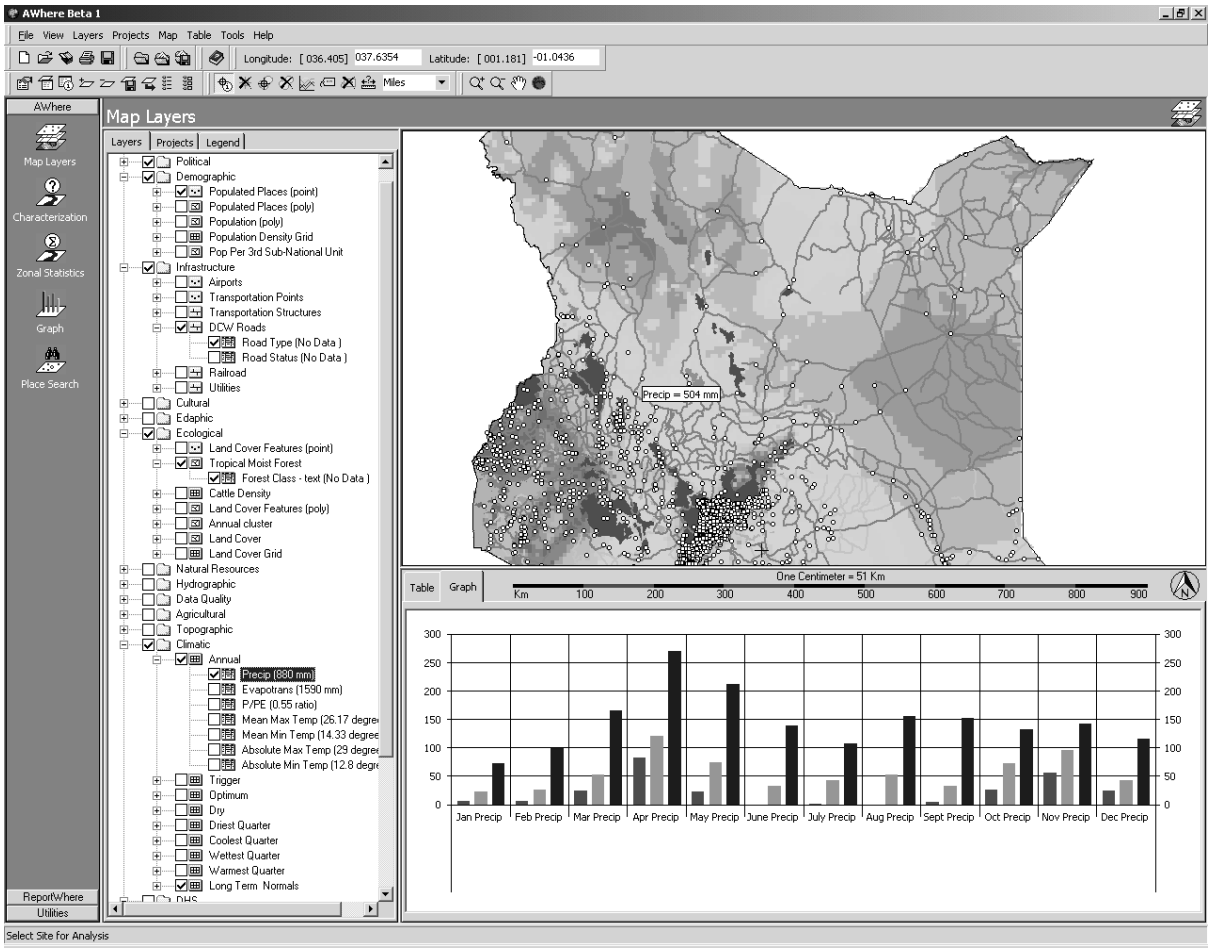


Figure 1. AWhere™-ACT interface.

References

Corbett, J.D., J.W. White, and S.N. Collis, Spatial Decision Support Systems and Environmental Modeling: An Application Approach. Chapter 3, *Geographic Information Systems and Environmental Modeling*, K.C. Clarke, B.E. Parks and M.P. Crane (Eds), 2000.

Corbett, J.D., S.N. Collis, B.R. Bush, R.Q. Jeske, R.E. Martinez, M.F. Zermoglio, Q. Lu, R. Burton, E.I. Muchugu, J.W. White, and D.P. Hodson, Almanac characterization tool. A resource base for characterizing the agricultural, natural, and human environments for select African countries. Texas Agricultural Experiment Station, Texas A&M University System, Blackland Research and Extension Center, Report No. 01-08, documentation and CD-ROM, March 2001.

Kurata, D., *Doing Objects in Visual Basic*, Sams Publishing, Indianapolis, IN, USA, 1999.

Leung, Y., *Intelligent Spatial Decision Support Systems*, Springer-Valeg, Berlin, Germany, 1997.

Sarang, P.G. Develop Three-Tier Applications with VB6 and MTS. ZD Inc. http://msdn.microsoft.com/library/periodic/period99/vb6mts.htm. Accessed July 15 2000.

Vogel, P., *Visual Basic Object and Component Handbook, Prentice* Hall, Upper Saddle River NJ USA, 2000.

Zeiler, M., Modeling our World: the ESRI Guide to Geodatabase Design. ESRI Press, Redlands, CA, USA, 1999

ACT: A GIS TOOL FOR RISK ASSESSMENT OF MALARIA IN KENYA

Guofa Zhou and Guiyun Yan
Department of Biological Sciences, State University of New York at Buffalo, Buffalo, NY 14260
Email: gzhou@icipe.org; gyan@acsu.buffalo.edu

ACT (Almanac Characterization Tool) represents a new generation analytical tool to support decision-makers with an integrated package of spatial database and analytical methods. A key concept behind ACT is that ACT provides not only analytical tools, but also sets of spatial data. Different data layers such as demographic, political, ecological, topographic, climatic, and agricultural data layers are represented in the metadata section of the Map Layer Properties. The primary functions of ACT are available in the Spatial Tools query module. The Spatial Tools available include Map Layers, Characterization, Zonal Statistics, etc. Characterization is a core function of ACT. It can be used to describe areas based upon their biophysical, social, and economic characteristics, or to search for areas that fall within a predefined set of parameters.

The aims of this study are to combine our field mosquito survey data with the ACT database, and to predict the distributions of different species of mosquitoes in Kenya using the Characterization tool of ACT. Malaria, transmitted by anopheline mosquitoes, is one of the most fatal infectious diseases in the world. In Africa, three mosquito species are considered as major malaria vectors: *An. gambiae*, *An. arabiensis*, *An. funestus*.

Material and Method

Study area and mosquito sampling. Adult anopheline mosquitoes were collected in four ecological zone. The first zone was in coastal Kenya. Nine sites were sampled in this region, and the elevation of the sampling sites ranged from 0m to 400m above sea level (a.s.l.). The second zone included eight sampling sites in the Great Rift Valley, the elevation of the sampling sites ranged from 910m to 1,970m a.s.l.. The third zone was in high-elevation area (ranged from 1,500m to 2,400m a.s.l.). The 4th zone is in the basin region of Lake Victoria with an elevation between 1,140m-1,500 a.s.l. (Fig. 1).

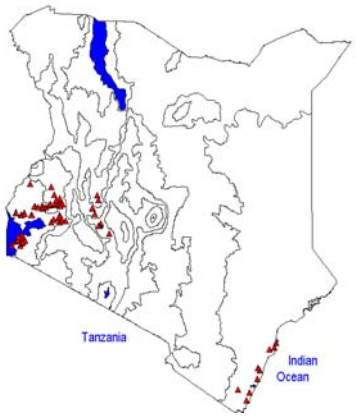


Fig. 1 A Kenya map showing mosquito sampling sites

Anopheline mosquitoes were collected from 10-15 houses at each site in May-July 1998-2000 during the long rainy season. Coordinates of each house were recorded using a hand-held GPS unit. Indoor resting mosquitoes were sampled using the pyrethrum spray catch method, and the collections were preserved in 95% ethanol for subsequent species identification.

Species identification. Female mosquitoes were examined microscopically to distinguish *Anopheles gambiae* species complex from *An. funestus* by morphology. Individual species within *An. gambiae* species complex were identified using the rDNA - PCR method.

Getting Site Info. GPS-reading geographic coordinates, i.e., decimal degrees latitude and longitude, for each sampling site were entered manually. The climatic, ecological, edaphic, and geographic data are displayed in the Tree-view for the particular sites. The climatic data used in the analysis included precipitation, evapotranspiration, the ratio of precipitation over potential evapotranspiration (P/PE), and maximum and minimum temperatures. These climatic variables were analyzed for a season or yearly average, i.e., annual mean, optimal five months, driest season, coolest season, wettest season, and warmest season. The ecological data were mainly land cover, and they all are qualitative data. The edaphic data were FAO soils WHC. Elevation was the only geographic parameter used in this study. The data were then transformed into tabular form using the Site Info to Table tool function, and then exported to MS Excel.

Correlation Analysis. The single variable forward stepwise regression analysis was used to determine the significance of correlations between climatic and other independent parameters and relative abundances of mosquito species. The critical value of correlation coefficient was chosen using the significance level of >99%. Forty-six sampling sites were use for correlation analysis.

Predicting Distribution of Mosquitoes. The variables selected by the correlation analysis were used for predicting the distribution of mosquitoes. Prediction was done using the Characterization tool. The selected variables were highlighted for Characterization Setting, and Intersect Method of Overlay Option was used to generate the predicted distribution map. Characterization results are automatically saved as new Data Layers. This new data layer can then be displayed and/or exported as BMP files from the Map Layers of the Spatial Tools.

Validation. The validation was done by retrieving the predictions from ACT and compared with the field observations. Fifteen sampling sites were used for validating the predictions. Both linear regression and paired t-test between observed mean values and predicted mean values was used.

Result and discussion

Correlation Analysis The significantly correlated variables with both *An. arabiensis* and *An. gambiae* are precipitation and P/PE for both annual values and wettest season values. Precipitation and P/PE were negatively correlated with *An. arabiensis,* but positively correlated with *An. gambiae*. The factors affect abundance of *An. funestus* were temperature and elevation. *An. funestus* relative abundance was positively correlated with temperature, but negatively correlated with elevation.

Distribution Predictions. Predictions of relative abundance of the three mosquito species are shown in Fig. 2. The predicted distribution of *An. arabiensis* is primarily in southern and coastal Kenya. Lake Victoria region is not within the upper quartile (>75%) area. The predicted upper quartile for *An. arabiensis* is the transitional zone – a zone between semi-arid area and lower moist area.

The predicted distribution of *An. funestus* is in a wilder geographic area than An. arabiensis except semi-arid to arid areas in eastern Kenya and Rift Valley. The area above median quartile is in a small area in coastal Kenya.

For *An. gambiae*, the areas above the upper quartile is in the basin region of Lake Victoria, western Kenya and a narrow zone along the coastal line.

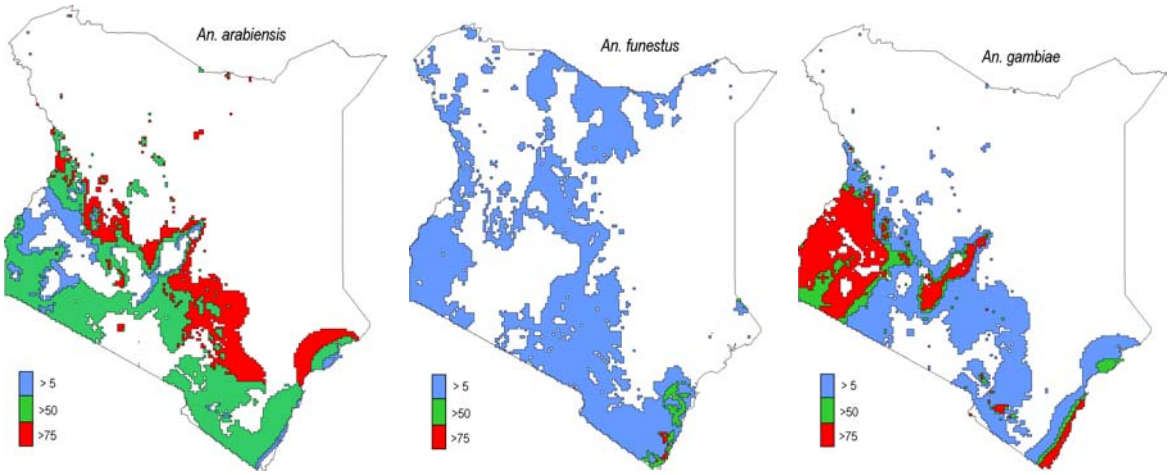


Fig. 2. Illustrations of predictions

Validation T-test results indicated that the mean predicted relative abundance for all three species are consistent with the observed abundances (Table 1). The linear regression between observed and predicted abundance was significant at p>0.99 level, and all the slopes were not different from 1 unit, indicating prediction error was very small. Two-five points were outside of the 95% confidence interval for the 3 species (Table 1, Fig. 3).

Table 1. Validation statistics

Species	t-test for means (t values/p values)	Linear regression		
		R (p value)	Slope b ± 1 SE	N. points (percentage)*
An. arabiensis	0.77 (0.46)	0.98 (>0.99)	1.00 ± 0.05	5 (10.8%)
An. gambiae	0.08 (0.94)	0.98 (>0.99)	0.98 ± 0.03	4 (8.7%)
An. funestus	0.94 (0.37)	0.99 (>0.99)	0.98 ± 0.04	2 (4.3%)

* Number of points that outside of the 95% confidence interval of regression

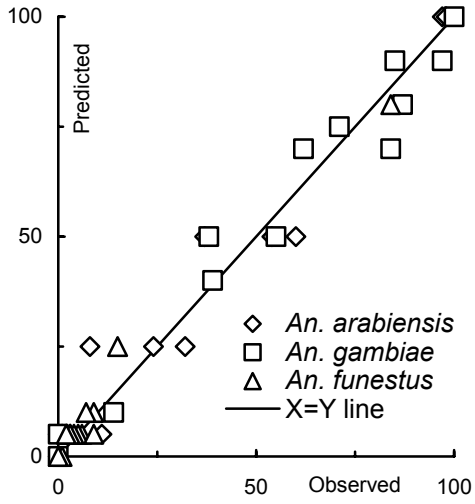


Fig. 3 Correlation between observed abundance and predicted abundance of anopheline mosquito species in Kenya.

The results indicated that ACT is a powerful tool for characterizing distribution of malaria vectors. Our study showed that *An. funestus* was the least important malaria vector in most part of Kenya except a small portion in southern coast where *An. funestus* had a high relative abundance. The majority of mosquitoes were *An. gambiae* and *An. arabiensis*. *An. gambiae* was dominant in high precipitation area but *An. arabiensis* was dominant in the transition area between arid and low moist areas. Understanding malaria vector species distribution is valuable for designing rational vector control strategies.

Reference

Collis S.N. et al. 2001. Almanac Characterization Tool. V. 3.0. Texas Agricultural Experiment Station, Texas A&M University System. Blackland Research and Extension Center. Report No. 01-08