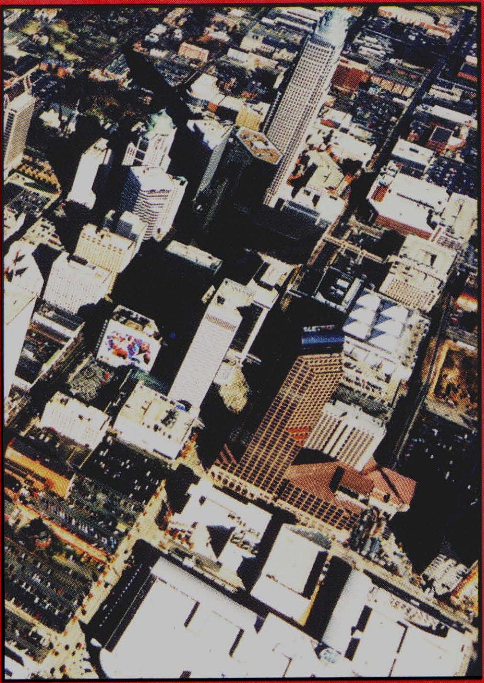
1995

ACSM/ASPRS

Annual Convention & Exposition Technical Papers



Charlotte, North Carolina
February 27 - March 2, 1995

Volume 4
Auto Carto 12

1995

ACSM/ASPRS

Annual Convention & Exposition
Technical Papers

ACSM 55th Annual Convention
ASPRS 61st Annual Convention

Volume Four: Auto Carto 12

Charlotte, North Carolina
February 27 - March 2, 1995

ISBN-1-57083-019-3
ISBN-1-57083-018-5

Published by
American Congress on Surveying and Mapping
&
American Society for Photogrammetry
and Remote Sensing
5410 Grosvenor Lane
Bethesda, Maryland 20814

Printed in the United States of America

Cover image courtesy of Aero-Dynamics Corporation.
This photo was exposed by Aero-Dynamics, Corp. on AGFA H-100 color negative aerial film. This film is distributed, processed, and photographically reproduced by Precision Photo Laboratories, Inc. Dayton, Ohio.

Proceedings

Twelfth International Symposium on Computer-Assisted Cartography

AUTO-CARTO 12

Charlotte, North Carolina
February 27-29, 1995

FOREWORD

This volume contains the papers of the Twelfth International Symposium on Computer-Assisted Cartography. The papers appear in close adherence to the order of their appearance in the program, allowing for parallel sessions, in order to serve the registrants and speakers during the meeting well as to serve the wider audience as a contribution to the literature.

Due to a variety of reasons, including a switch in timing to again coincide with the ACSM/ASPRS spring convention, the Call for Papers did not receive as wide a distribution as was possible for previous Auto-Carto conferences. It therefore came as a pleasant surprise when over 60 extended abstracts were submitted and were of high quality, overall. Perhaps this is an indication that Auto-Carto, which began in the early 1970's, has indeed become a tradition and an anticipated event within the world research community.

All submitted abstracts were reviewed by at least three members of the Program Committee. Full papers submitted on the basis of approved abstracts were then returned to the original reviewers for a final screening. This approach was seen as a "middle ground" between review and acceptance on the basis of abstracts and on the basis of full papers. The experiment did seem to work reasonably well in ensuring an appropriate focus and level of both quality and currentness for the papers. Authors were able to more easily incorporate reviewers' comments while also being allowed a later deadline for full papers than would be possible otherwise.

I wish to thank the many individuals who have helped with the organization and planning for Auto-Carto 12. The Program Committee of Barbara Buttenfield, Nick Chrisman, Robin Fegas, Gail Langran Kucera, Duane Marble, David Mark, Matt McGranaghan, Robert McMaster, and Mark Monmonier carefully reviewed the abstracts and the full papers and provided valuable advice throughout the formulation of the program. I also wish to thank Linda Hachero, ACSM/ASPRS Convention Headquarters, for taking care of local arrangements and the printing and distribution of the program materials, Joann Treadwell, ASPRS Director of Publications, for production of the Proceedings, and Rosemarie Hibbler for helping throughout with voluminous paper and electronic correspondence.

Donna J. Peuquet
The Pennsylvania State University

Table of Contents

Algorithms

Cartography and Decision Making

Cartographic Generalization

Visualization

Spatial Data Representation

System Design Issues

Shared Databases and Formal Frameworks

Data Quality I

Spatial Analysis

Feature Placement and Data Efficiency

Data Quality II

Late Submissions

Author Index

COMPUTATION OF THE HAUSDORFF DISTANCE BETWEEN PLANE VECTOR POLYLINES

J.F. Hangouët
COGIT Laboratory
Institut Géographique National
2, avenue Pasteur
F-94160 Saint-Mandé
France
email : hangouet@cogit.ign.fr

ABSTRACT

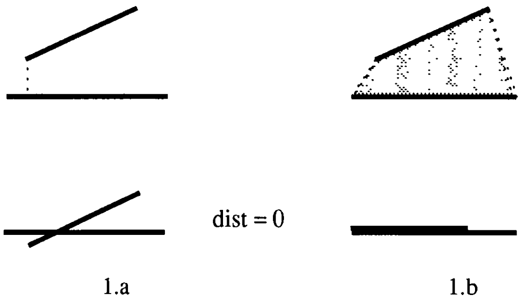
The Hausdorff distance between two objects is a mathematically true distance. When both objects are punctual, it does not differ from the Euclidean distance between points; otherwise it takes into account the mutual positions of the objects relatively to each other. Its main interest for automated cartography, besides quantifying spatial relations between objects, lies in the fact that it expresses *remoteness*. How far are features from each other ? How far is a generalized feature from its original position ? After spotlighting on some properties of the Hausdorff distance applied to geographical features, the paper describes an algorithm for computing the Hausdorff distance in vector mode between two polylines.

KEYWORDS : Hausdorff distance, spatial relations, algorithm.

INTRODUCTION

While the classical Euclidean distance is of foremost importance in surveying issues, GISs have revealed that the point-to-point relation is too limited for cartographic applications. Most geographical features zigzag across the background or swell and bump against each other - they are definitely not points. Distance between features is a difficult concept which has been approximated through various indicators: minimum Euclidean distance [PEUQUET-92], reworked with ε-bands [PULLAR-92], surface "in between" [MCMASTER-92] ... Even if these measures are well adapted to the applications they are meant for (mainly proximity and accuracy evaluations), they all lack at least one of the three prerequisites for being a true mathematical distance : that of separateness, which means that the distance between two objects is zero if and only if those objects are strictly identical (fig. 1). It can be argued that this condition is pointless in cartography, since features will never be strictly the same: attributes or symbolization, if not geometry, will always differ. However, from a geometric point of view, the fulfilment of all criteria for a "gap-quantification" function makes it a safe and systematic distance. The Hausdorff distance is such a mathematical distance - surely not the best, but anyhow a very convenient one.

Figure 1.



Non separateness of some distance functions
1.a minimum Euclidean distance
1.b inner area

HAUSDORFF'S DISTANCE

Definition

Felix Hausdorff (1868-1942) was a German mathematician whose contribution is most remarkable in the field of topology (pure "abstract" topology would be clearer in our GIS and spatial relations context). He built up a distance between objects in finite space as:

$$DH(A,B) = Max\left(sup_{x\in A}\, d(x,B)\, ,\, sup_{x\in B}\, d(x,A)\right)$$

where *A* and *B* are closed sets and *d (x,B)* the classical Euclidean distance from point *x* to object *B* (proof that it is a distance in the mathematical sense can be found in most topology manuals). Two components can be defined (figure 2):

distance from A to B : $D_{A\to B} = \left(sup_{x\in A}\left(inf_{y\in B}\, d(x,y)\right)\right)$

distance from B to A : $D_{B\to A} = \left(sup_{y\in B}\left(inf_{x\in A}\, d(x,y)\right)\right)$

and : $DH = Max\left(D_{A\to B}\, ,\, D_{B\to A}\right)$

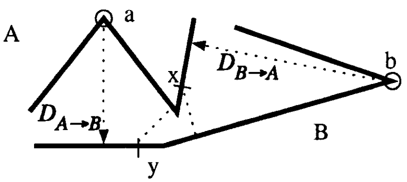


Figure 2.

The two components.

$\forall x \in A,\ d(x,B) \le d(a,B)$

$\forall y \in B,\ d(y,A) \le d(b,A)$

The two components are not necessarily of a same value. This is illustrated in the figure above. Other properties of the Hausdorff distance will now be listed, with the examples of polylines, which are closed objects from a mathematical point of view, and a most common representation of geographic features in vector GIS.

Properties

Asymmetry. (fig. 2) The two components usually have different values.

Orthogonality. (fig. 3) The vector representing the Hausdorff component from one object to the other is perpendicular to the second object (or points onto a vertex of the second object). This is a property inherited from the Euclidean distance from point to line.

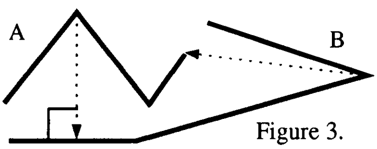


Figure 3.

Sensibility. (fig. 4) Tails make the Hausdorff distance very unstable.

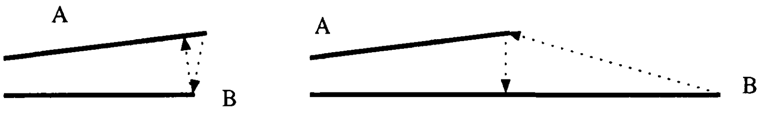


Figure 4.

Tricks. (figure 5) Contrary to intuition, and to what the figures above can suggest, the Hausdorff distance may be achieved between any points of the polylines, and not only on vertices. This makes the computation more difficult.

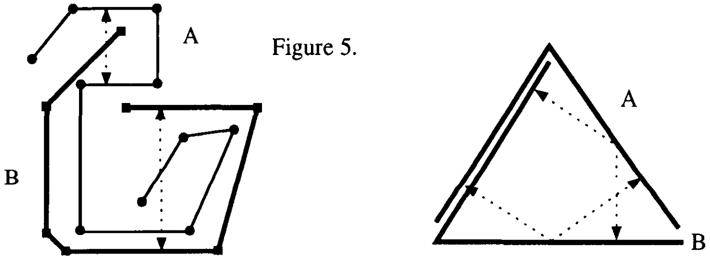


Figure 5.

Tangency. An interesting property of say component *A* to *B* of the Hausdorff distance between polylines is that *when the distance vector starts from a point of A that is not a vertex, there are several distinct points on B which are at the same distance.* In other words, there are several distance vectors, as indicated in figure 5. This result was found while trying to simplify the computation of the distance. An illustration of the proof, rather than the full tiresome demonstration, is given in the appendix.

Applications in automated cartography

The Hausdorff distance between polylines is currently used for feature matching between multi-scale layers [STRICHER-93], for statistical quality controls on linear objects [HOTTIER-92] [ABBAS-94], for the control of generalizing algorithms (current work at our COGIT laboratory).

In cartography, *asymmetry* shows up spectacularly : a generalized line (fig. 6) is closer to the initial line, and the initial line remains far from the generalized line.

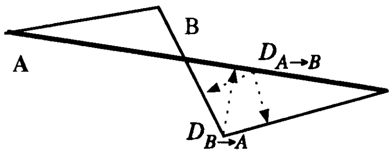


Figure 6. *B is the original line, A its generalization.*

Statistically, [HOTTIER-92],

$$D_{A \to B} < D_{B \to A}$$

Since all maps and models are generalizations of the real world, Hottier [HOTTIER-92] even states that "maps approach reality but reality remains far from maps" - a truth about resemblance that was already sensed, far from any mathematical justifications, by Edmund Spenser in 1590 [SPENSER-90], in the mouth of the False Fox, to Sir Ape :

> *And where ye claim yourself for outward shape*
> *Most like a man, man is not like an ape.*

Sensibility is a critical issue when comparing objects. The problem is similar to that of the delineation of the area "between" two objects when computing the area distance. Where to cut the lines ? Solutions [STRICHER-93] , [ABBAS-94] are often dependent on the applications.

The tricky aspects of some configurations make the computation of the Hausdorff distance ticklish and time-consuming. Hottier has developed a raster algorithm, and Abbas [ABBAS-94] a vector algorithm, the optimization of which is based on the introduction of a likelihood threshold suited to the statistically expected result. The following algorithm also works on vector polylines, first on the vertices, and then if necessary on the inter-lying segments.

AN ALGORITHM

The two components of the Hausdorff distance may have different values, but the way to compute them is the same. The algorithm given here finds one component, from polyline *A* to polyline *B*. To find the Hausdorff distance, the computation must also be applied reciprocally from *B* to *A*, and the final distance is the greatest of the two results.

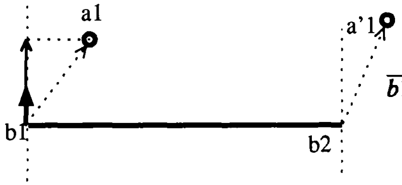
The algorithm to find $D_{A \to B}$ proceeds in three stages :

1. Computation of the distances from the vertices of *A* to polyline *B*.
2. Tests to detect whether further calculation is required or a vertex of *A* bears the greatest distance from *A* to *B*.
3. When no vertex bears the greatest distance, computation of the greatest distance on likely segments.

1. Distances from the vertices of *A* to polyline *B*.

From each segment *sb* = [*b1 b2*] of *B* can be defined a band perpendicular to the segment, with a width equal to the length of the segment (fig. 7). If a vertex *a1* of *A* lies in this band, the distance can be computed from the scalar product between vector **b1 a1** and the unitary vector orthogonal to *sb*. When *a1* lies outside the band, it is closer to a vertex of *B*, with which the Euclidean distance is achieved.

Figure 7



$$\overrightarrow{b1a1}\cdot\overrightarrow{b1b2} \geq 0 \textit{ and } \overrightarrow{b2a1}\cdot\overrightarrow{b2b1} \geq 0$$

$$\overrightarrow{b2a'1}\cdot\overrightarrow{b2b1} < 0$$

For a vertex of *A*, the distances to all segments of *B* have to be calculated, the smallest being the Euclidean distance from the vertex to *B*.

2. Check tests.

When computing the distance of each vertex of *A*, a trace must be kept of the component of *B* on which hits the Euclidean distance (a component being either vertex or segment, identified for example by its position number along the polyline : 1st vertex, 1st segment, 2nd vertex, 2nd segment ...). Thus at this stage, all vertices of *A* have their associated components on *B*. The tests will consist in eliminating all segments of *A* which cannot be farther from *B* than their end-points. The tests check all pairs (a1 , a2) of successive vertices of *A*. Such a segment needs no closer analysis if (fig. 8)

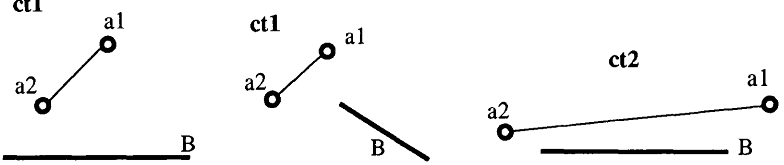
Comp(a1) = Comp (a2) [ct1]

or *Comp(a1) and Comp(a2) are successive vertices of B* [ct2].

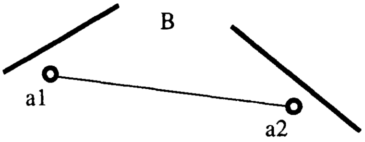
[*Comp(a)* being the number of the component on *B* breeding the Euclidean distance between *a* and *B*].

Such an elimination is justified by the fact that the distance function between two segments is either increasing or decreasing.

Figure 8



This case requires closer observation :



If all successive segments of A are thus discarded, it means that the Hausdorff component is achieved from a vertex of A, that with the greatest Euclidean distance. Otherwise, further calculation is required.

3. Detail analysis

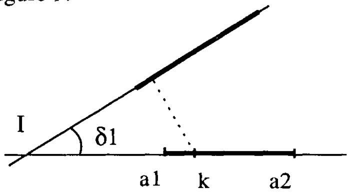
For the remaining segments of A, there is suspicion that points between vertices may be farther from B than the vertices. For each segment there will be computed the greatest Euclidean distance from B - the greatest for all segments being the Hausdorff component.

So now the basic problem is : considering a segment $[a1\ a2]$ of A, the extremities of which are close to different components of B, to find the farthest point in-between. For this we will consider all the distance functions to all segments and vertices of B from $[a1\ a2]$, each point P on $[a1\ a2]$ being identified by its k_p parameter so that :

$$\forall P \in [a1a2], \exists! k_p \in [01] \ / \ \overline{a1P} = k_p \overline{a1a2}$$

The distance function d(k) from $[a1\ a2]$ to a segment is (fig. 9) :

Figure 9.



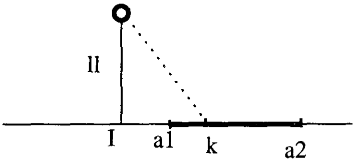
$$d(k) = Abs\left((k - k_I) * d(a1,a2) * \sin\delta_I\right)$$

$$k(a1) = 0$$

$$k(a2) = 1$$

The distance function d(k) from $[a1\ a2]$ to a vertex is (fig. 10) :

Figure 10.

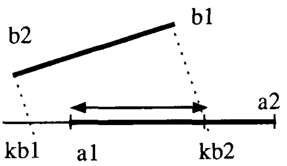


$$d(k) = \left((k - k_I)^2 * d(a1,a2)^2 + ll^2\right)^{1/2}$$

Thus, two numbers have to be computed for each component of B : if it is a segment, the k-parameter for the intersection of the two directions, and the sine of the angle; if it is a vertex, the k-parameter of the shortest distance from (a1 a2) to the vertex, and this distance (ll). Let's call this pair of indicators the *distance representation* of the component.

However, it is not necessary to compute all the representations : thanks to the orthogonality property of the Euclidean distance, this computation is required only for the segments which see part of $[a1\ a2]$ or for the vertex protruding toward $[a1\ a2]$. The test at this stage can consist in computing, for each segment of B, the k-parameters of the intersections with line ($a1$ $a2$) of the perpendiculars at both ends : the *effectiveness interval* (fig. 11). The configurations, and the way to recognize them, are described in fig.12.

Figure 11.



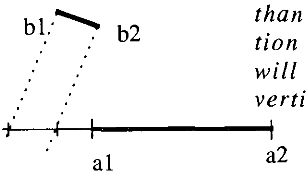
kb is solution of the system :

$$\overline{a1M} = k.\overline{a1a2} \text{ and } \overrightarrow{bM}.\overrightarrow{b1b2} = 0$$

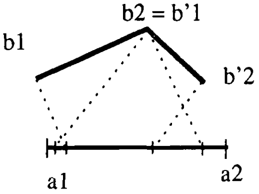
sinδ is the absolute value of the scalar product between a unitary vector of [b1 b2] and a unitary vector orthogonal to [a1 a2].

Let's rename kb1 if necessary, and take kb1 = Min (kb1 , kb2) and kb2 the other one. [Max (0,kb1) Min (1,kb2)] is the effectiveness interval of component [b1 b2].

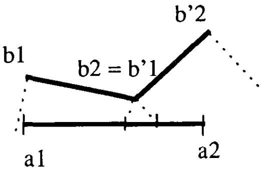
Figure 12.



kb1 and kb2 are both greater than 1 or smaller than 0 : no use trying to find the distance function to segment [b1 b2]. One of the extremities will be nearer anyway, and treated with the vertices.



kb'1 > kb2 : no use trying to find the distance function to vertex b2. Points on [a1 a2] are closer to one of the segments.



kb'1 < kb2 : this vertex points towards [a1 a2], its distance representation has to be kept.

Its effectivness interval is :
[Max (0 , kb2) Min (1 , kb'1)].

So, for one segment [$a1$ $a2$], the steps above provide a list of potential components on B. For each of these, the distance representation has to be computed - and the effectiveness interval stored.

Before describing the algorithm itself, another tool requires description : computation of the intersection between two distance representations.

Intersection of two segment distance representations :

Let the first representation be: (k1, sin δ1), and the second: (k2 , sin δ2).

The resulting *k* are : $k = \frac{k1 - \alpha . k2}{1 - \alpha}$ *or* $k = \frac{k1 + \alpha . k2}{1 + \alpha}$ where $\alpha = \frac{sin\delta 2}{sin\delta 1}$
If a sine is nul, it has to be checked whether the segment is strictly parallel to [*a1 a2*] or not. If it is, its distance is a constant different from zero, and *k* is easy to find. If the segment is on (*a1 a2*), the common part cannot compete for achieving the greatest distance from [*a1 a2*] to *B*.

Intersection of two vertex distance representations :

Let the first representation be: (k1 , ll1), and the second: (k2 , ll2).

The resulting *k* is : $k = \frac{\frac{ll2^2 - ll1^2}{d(a1\,,a2)^2} + k2^2 - k1^2}{2 * (k2 - k1)}$

If k2 = k1, check if ll2 = ll1. If the two values are different, there is intersection. Otherwise (*a1 a2*) is equally distant from both vertices.

Intersection of vertex - segment distance representations :

(k1, sin d1) is the segment distance representation, (k2 , ll2) is the vertex distance representation. *k* is the root(s) of the following equation :

$$\left[sin^2\delta 1 - 1\right] . k^2 + \left[2 . k - 2 . k1 . sin^2\delta 1\right] . k + k1^2 . sin^2\delta 1 - k2^2 - \frac{ll2^2}{d(a1, a2)^2} = 0$$

Core of the algorithm.

Now that the tools are ready, the algorithm is straightforward :

Start from *a1*. Find its associated component on *B*, its distance function *dfc* and effectiveness interval *eic* [which by construction has 0 for lower bound). We are going in fact to follow the lowest possible path from *a1* to *a2* along *B* (fig. 13).

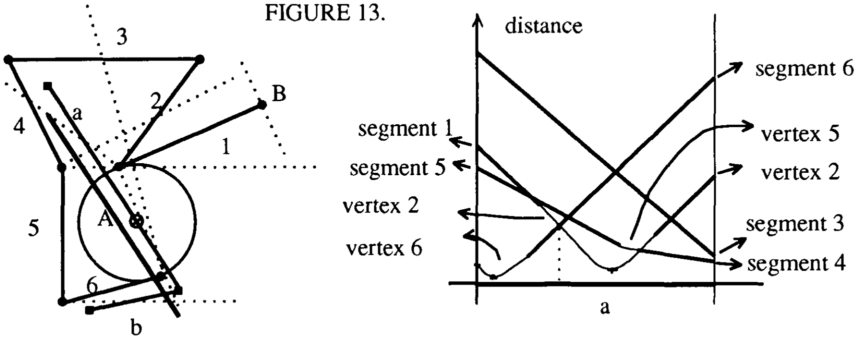
For all other components, find the intersection of their distance function with *dfc*. If the intersection *k* lies without their effectiveness interval or without *eic*, it has to be discarded. For all remaining *k*, the smallest, *kmin*, is the new starting bound. The associated component on *B* is the new component. The search interval becomes [*kmin* , *kb2*] where *kb2* is the upper bound of the effectiveness interval of the new component. The distance between the intersection point and the component has to be stored.

Loop: In this new interval, intersections and new components have to be found on the same criteria.

When no intersection occurs within an interval ending before 1, the following component along polyline *B* has to be found : after a segment, its end-point, after a vertex, the following segment. If the upper bound of the interval is 1, and no intersection occurs, the algorithm has stopped running.

The Hausdorff component is the greatest of all the distance values stored during this pass.

FIGURE 13.



Both extremities of segment b are closer to segment 6 of polyline B : no use investigating on b. However, a has to be looked closely. On the right, the representation of the distance functions from a to the components of B, In this case, there are bound to be intersections, because of the tangency property of the Hausdorff distance. The lowest possible path describes the successive Euclidean distances from a to B; its highest peak is the Hausdorff component.

CONCLUSION

The two components of the Hausdorff distance between polylines give an indication of the mutual remoteness of the polylines, which is a new way of looking at spatial relations. The Hausdorff distance can be an interesting measure on geographical objects - points, lines and contours. The complexity of the algorithm described for finding the component from a polyline with *m* vertices to a polyline with *n* vertices is in most desperate cases O($m.n^2$).

REFERENCES

[ABBAS-94] *Base de données vectorielles et erreur cartographique· problèmes posés par le contrôle ponctuel, une méthode alternative fondée sur la distance de Hausdorff: le contrôle linéaire.* IGN, PhD thesis 1994.

[STRICHER-93] *Base de données multi-échelles: association géométrique entre la BD Carto et la BD Topo par mesure de la distance de Hausdorff.* DESS thesis 1993.

[HOTTIER-92] *Contôle du tracé planimétrique d'une carte* Bulletin d'Information de l'Institut Géographique National, Saint-Mandé, France, 1992 pp. 30-36

[PEUQUET-92] *An algorithm for calculating minimum Euclidean distance between two geographic features* Computers & Geosciences Vol. 18 No. 8, 1992 pp. 989-1001

[PULLAR-92] *Spatial overlay with inexact numerical data* Auto-Carto 10, 1992, pp 313-329

[MCMASTER-86] *A statistical analysis of mathematical measures for linear simplification* The American Cartographer Vol. 13 No. 2, 1986 pp. 103-116

[SPENSER-90] *Mother Hubbard's Tale* 1590

APPENDIX : CLUE TO THE TANGENCY PROPERTY

Here is illustrated the fact that when a Hausdorff component does not start from a vertex, it reaches the other polyline in several distinct points.

[$S_{1\,i}$, $S_{1\,i+1}$] is a segment of polyline *A*, and component $D_{A\to B}$ is achieved between point P1 on] $S_{1\,i}$, $S_{1\,i+1}$ [and point P2 on polyline *B*.

First consequence (A1) : $\forall P \in \,]S_{1\,i}\,;\, S_{1\,i+1}[\,,\ d(P, B) \leq d(P1, P2)$

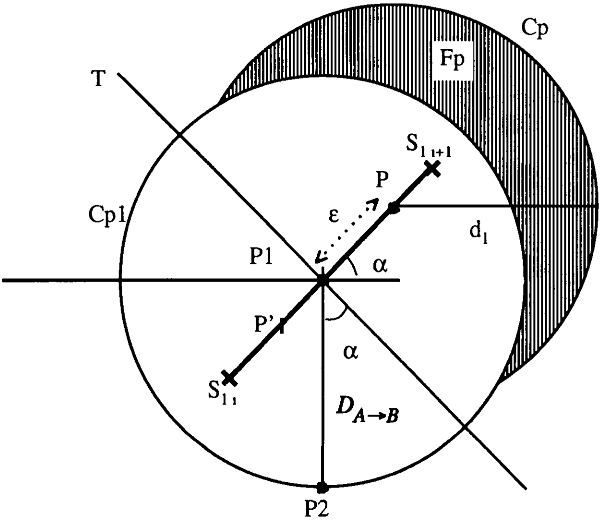
Second consequence (A2) : $\forall Q \in B,\ d(P1, Q) \geq d(P1, P2)$

This means that for every point P on] $S_{1\,i}$, $S_{1\,i+1}$ [, and especially in the most remote part of the segment, there have to be parts of *B* both out of the disc centered on P1, which has $D_{A\to B}$ for radius (because of A2, disc *Cp1* in fig. 14), and inside the similar disc centered on P (because of A1, disc *Cp* in fig. 14). In other words, there have to be parts of *B* inside the grey crescent *Fp* illustrated in figure 14 below.

This is especially true when P comes close to P1. When P draws on to P1, *Fp* will fuse with the semi-circle of *Cp1* limited by *T*, the perpendicular to ($S_{1\,i}$, $S_{1\,i+1}$) in P1. P2 does not belong to it (P moves on the most distant part of the segment), so there has to be at least one other point of *B* on this semi-circle : one other point at the same distance.

$\alpha = 0$ is a special but not revolutionary case (when dealing with polylines).

Figure 14.



AN ALGORITHM FOR GENERATING ROAD CENTER-LINES FROM ROAD RIGHTS-OF-WAY

Naven E. Olson
Digital Mapping Services
Computer Services Division
University of South Carolina
Columbia, SC 29208

ABSTRACT

A totally vector geometric algorithm was developed for generating road center-lines from road rights-of-way The algorithm finds pairs of parallel arc sections within a given distance. The user passes the name of the right-of-way layer, the name of the new center-line layer, a road width, and an expanded multiplier. Pre-processing is done on a new copy of the right-of-way layer linework The main algorithm then creates a table of segments perpendicular to right-of-way arcs. This table is used to generate a table of intersection points which can be sorted and used to generate the basic linework for center-lines Post-processing then joins nodes and center-lines to create road topology and windows out any "center-lines" outside of rights-of-way. The algorithm was developed as part of Digital Mapping Services' in-house mapping system and tested on City of Columbia right-of-way data. The source code is copyrighted by the University of South Carolina.

INTRODUCTION

One problem that is often encountered in computerized mapping is that of generating road center-lines from road right-of-way data. For instance it may be desired to generate center-lines from right-of-way and parcel data generated from a tax mapping program.

Traditional approaches have been of three types:

a) standard manual digitizing;

b) center-lining in a CAD or COGO environment;

c) vector to raster conversion, followed by skeletonization, followed by raster to vector conversion.

This paper sets forth a new, strictly vector, approach which does an automatic generation based on fundamental geometric and trigonometric analysis.

LITERATURE REVIEW

Various skeletonization and other algorithms related to this approach have been developed in the past. Blum(1967) first proposed "medial axes" or skeletons for regions. Raster algorithms for medial axis skeletons have been developed by Rosenfeld and Pfaltz(1967) and Montonari(1969). Brassel and Heller(1984) propose a vector algorithm for bisector skeletons. This could be used to develop "line segment proximity polygons" dividing a region into polygons closest to polygon line segments and having a rough equivalence to Thiessen polygons. It does not, however, produce a line network which would expand to the right-of-way area upon appropriate buffering.

Brassel (1985) has listed area-to-line conversion as a desirable map generalization operator. McMaster (l991) lists Beard (1987), DeLucia and Black (1987), Nickerson and Freeman (1987), and McMaster and Monmonier (1989) as others including this function in their "wish list".

Brassel and Weibel(1988) advocate "processing based on understanding". Mark (1989:76) states that "in order to generalize a cartographic line, one must take into account the geometric nature of the real-world phenomenon which that cartographic line represents." Weibel quotes these and other authors in advocating comprehension of underlying phenomena.

This paper presents an approach to area-to-line collapse suitable for road, railroad, and power-line rights-of-way. It could perhaps be used as a starting point for algorithms for stream collapse or even interpolation between contour intervals.

TRADITIONAL APPROACHES

The problem of generating road center-lines from road rights-of-way has traditionally been handled by standard manual digitizing, center-lining in a CAD or COGO environment, or a rasterization-skeletonization-revectorization approach. All of these approaches, as used in current state-of-the-art practice, have severe weaknesses

Standard manual digitizing is labor-intensive. Due to the generally small distance between road sides there may be significant variation from "parallelism" of the center-line to the rights-of-way.

Center-lining, as in the center-line option under the add arc command in arc edit using ARC/INFO, generally eliminates this non-parallelism problem and will produce a good center-line if used properly It is also less tedious than table digitizing However, it is still labor intensive and will produce errors if bad choices in picking right-of-way points are made and not corrected.

The rasterize-skeletonize-revectorize approach is the most automated of the three. This approach is, however, troubled by several artifact problems. Some well known raster-to-vector artifact problems are aliasing, dimpling, and spurs. (Gao and Minami (1993) and Zhan (1993)) Artifacts common in this rasterize-skeletonize-revectorize approach include incorrect handling of off-center intersections and the gaps illustrated in Figures la and lb.

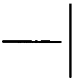


Figure la.

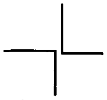


Figure lb.

NEW ALGORITHM

The fundamental approach to the center-line problem involves making a copy of the right-of-way layer and pre-processing the copy, generating the basic line work, and post-processing to close intersections and create topology and window out any "centerlines" that are not inside any right-of-way.

Pre-processing

Step 1. Angle correction. Correct angles on short arcs, as defined in the later section on Arc classification, which resemble Figure 2a so that they resemble Figure 2b. This will prevent good line work from being lost during Intersection points table correction.

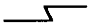


Figure 2a.

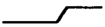


Figure 2b.

Step 2. Arc extension. Extend arcs ending at dangles two standard buffer widths past the dangle node.

Step 3 Arc classification. Classify arcs The original arcs are given attributes of "< *arc number*>,0" if they are standard arcs and attributes of "<*arc number*>,-2" if they are short arcs (less than a buffer width in length) having a sum of angles at their ends greater than about 180 degrees. See Figures 3a and 3b.

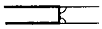


Figure 3a.

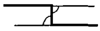


Figure 3b.

Extension arcs would be given attributes of "<*original arc number*>,- 1 " or "<*original arc number*>,1 " for arcs extended from the start (-1) or end (1) of original arc <*original arc number*>.

Main line work generation.

Step 4. Perpendicular segments table generation Generate a table of segments perpendicular to all standard arcs at all discrete points on the arcs and perpendicular to all extension arcs at their extended ends. Perpendicular directions are generated by a process of numeric differentiation The table would consist of the following properties:

a) arc or original arc number;

b) distance along arc from beginning of arc (negative for "<*original arc number*>,-1 " arcs);

c) x- and y-coordinates of center-point on arc;

d) x- and y coordinates of both endpoints, each endpoint being located a standard buffer length times an expansion multiplier away from the center-point along the perpendicular line.

Step 5. Spatial index generation. Generate a spatial index of this table of perpendicular segments.

Step 6. Intersection points table generation. Intersect these perpendicular segments with the right-of-way arcs to form a table (File 6.1) of intersection points calculated by averaging the coordinates of the center point of the perpendicular segment and the coordinates of the point of intersection of the perpendicular segment with the other arc. Each point has the following properties:

a) x- and y-coordinates;

b) arc number of arc the perpendicular segment was generated from;

c) arc number of arc intersected by the perpendicular segment;

d) distance from the beginning of the arc to the perpendicular segment.

An additional table (File 6 2) is also created for the intersection points containing the following properties:

a) the intersection number of the intersection;

b) the segment number of the perpendicular segment;

c) the x- and y-coordinates.

When the tables are generated, intersections that are not close to perpendicular are flagged as invalid. This prevents spurious line work from being created in later steps.

Step 7. Intersection points table correction. Eliminate intersections which are not the first intersection on the ray going out from the center point of the perpendicular segment that they are located on These intersections can cause spurious linework as shown in Figure 4.

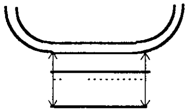


Figure 4.

Sort File 6.2 by perpendicular segment number. Mark each intersection that has any intersections on the same segment between it and the center point of the segment as invalid. Reorganize File 6.1 (the intersections point file) deleting invalid intersection points.

It should be noted that points flagged as invalid because they were generated by non-perpendicular intersections are written out in Step 6 and deleted in Step 7 because such points are sometimes the first intersections on a perpendicular segment. It should also be noted that this step makes Angle correction necessary.

Step 8. Intersection points table sort. Sort File 6.1 (the intersection points file) with field 6.1.b) being the primary key, field 6.1.c) the secondary key, and field 6.1.d) the tertiary key. Sort in ascending order on all fields.

Step 9. Auxiliary intersection points table generation. This is a key step and prevents the loss of a high percentage of the linework.

Generate intersection records with 6.1.b) and 6.1.c) transposed for arcs where the duplicate with lower primary key does not exist or has only one point. Also generate intersection records with 6.1.b) and 6.1.c) transposed for arcs where only one point exists in each duplicate arc.

The first step is necessary because Passing of intersections files uses only intersection points with the primary key less than the secondary key to prevent duplicate arc problems. Due to anomalies some arcs will only appear as the duplicate arc with the primary key greater than the secondary key.

The second step is necessary because other anomalies will cause some arcs to appear as two one-point arcs. The two points will generally be close to the two ends of the arc. The distance of the point from the lower-numbered arc must be converted to the distance along the higher-numbered arc using the following procedure

a)Find the point that distance along the lower-numbered arc.

b)Use a perpendicular segment to find the corresponding point on the higher-numbered arc.

c)Calculate the distance along the higher-numbered arc of this point.

Step 10. Passing of intersections files. Pass the intersections files, starting a new arc each time either arc in the arc pair changes. Use only intersections where the primary key is less than the secondary key to eliminate duplicate arcs.

It should be noted that several indices are created in this program. It is necessary to know what the primary and secondary right-of-way arcs are for each center-line generated. There must be tables of corresponding primary right-of-way arcs, secondary right-of-way arcs, and associated center-lines. One of these tables is sorted by primary arc number, and the other is sorted by secondary arc number.

Post-processing.

This stage proved harder to develop than any other stage of the process. The first algorithm, Ring generation, proved especially difficult. Several tables were used and a great deal of complex indexing was involved.

It should be noted at this time that pairs of right-of-way arcs which surround a centerline do not necessarily have the same directionality. Failure to account for this fact can cause major problems in Ring generation, Tee solution, and Auxiliary two-arc-ring generation.

Step 11. Ring generation. Form "rings" around open intersections as in the hypothetical example in Figure 5.

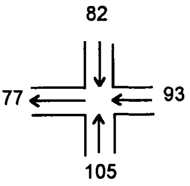


Figure 5.

Here the "ring" has four ends in this order: 77,-82,-93,-105. The procedure for generating a ring was as follows·

a) Pick a one-arc node.

b) Note which two right-of-way arcs were used to generate the arc.

c) Pick one of these right-of-way arcs for the target arc, the finding of which will close the ring.

d) Note which right of-way arc meets the other right-of-way arc at the appropriate end

e) Find which center-line and opposite right-of-way arc pair with this right-of-way arc.

f) Continue this process until the "ring" is closed by reaching the target arc or it is clear the "ring" cannot be closed.

False intersections are a significant problem in Ring generation. Two helpful checks for false intersections are the previous endpoint test (i) and the on-extensions test (ii).

(i). Take the endpoint of the proposed center-line. Find the point on the known right-of-way corresponding to this end point. Use this point to find the point on the proposed right-of-way arc corresponding to the endpoint. Construct a line segment starting at the corresponding point on the known right-of-way and perpendicular to the segment connecting it to the corresponding point on the proposed right-of-way. Intersect this newly constructed segment with the end segment of the right-of-way section which connected to the known right-of-way going backward on the ring. If the intersection is too far from the actual node of the known right-of-way arc, the pairing is invalid.

(ii). Take the intersection of the two end segments of the center-lines. Check to be sure it is on both rays extending outward from the relevant end points. If it is not, the pairing is invalid. This test is especially helpful in "downtown" type neighborhoods where most blocks are square in preventing loss of line work resulting from invalid arcs inside blocks, which will later be deleted, being incorrectly chosen for pairing.

Step 12. Tee solution. Find missed intersections at "T" junctions where there is no pseudo-node at the "top" of the "T" For an example, see Figure 6.

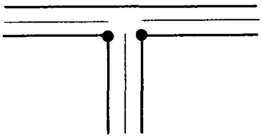


Figure 6.

The procedure for "solving" a "tee" is as follows:

a) Pick a one-arc node.

b) Note which two right-of-way arcs were used to generate the arc.

c) Note which right-of-way arcs meet these right-of-way arcs at the appropriate ends.

d) If either of these arcs is non-existent or these arcs are identical, stop.

e) Find a right-of-way arc which pairs with both of these arcs and the corresponding center-line for each pair.

f) If the center-lines are identical, stop.

g) Check for "false tees" using the on-extensions test used for false intersections during Ring Generation. Both center-line pairs should be checked

h) If the "tee" passes both tests in g), write the "tee" to the "ring" table.

Step 13. Ring closure. Edit the arcs to close the "rings" generated in Step 11 and Step 12. The procedures for closing two-arc and three-arc "rings" are well understood. No attempt has yet been made to handle five-or-more-arc intersections, although the file structure allows up to seven arcs in a "ring".

Four-arc intersections were classed as one of two basic classes. Figure 7 shows several variants of the most common class. Figure 8 shows two variants of the rarer class.

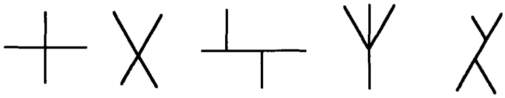


Figure 7.

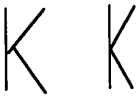


Figure 8.

The rarer case is distinguished by the fact that the sum of the two largest opposite angles minus the sum of the two smallest opposite angles is greater than 1.5 radians and the difference between the two angles in the larger opposite angle pair is greater than 1.5 radians.

For the most common case the procedure was:

a) Intersect the two opposite angle pairs where the sum of the distances from the endpoints to the intersections is smallest.

b) If the difference between these two intersections is smaller than a closure tolerance, move both intersections to their average.

c) Move all "ring" nodes to the appropriate intersections.

d) If the intersections are not within a closure tolerance, add an arc between them.

For the rarer case the following procedure was used:

a) Intersect the arcs in the smaller angle of the large angle pair.

b) Intersect the arcs in the most perpendicular angle of the smaller angle pair to get the intersection for the largest angle.

c) If the intersections are within a closure tolerance, move both intersections to their average.

d) Move all "ring" nodes to the appropriate intersections.

e) If the intersections are not within a closure tolerance, add an arc connecting them.

Step 14. Auxiliary two-arc-ring generation. Generate two-arc rings to cover the cases in Figure 9a and Figure 9b. The case shown in Figure 9a is more common and is basically due to minor mismatch between pseudo-nodes on matching right-of-way arc pairs. The problem in figure 9b arises in situations where a pseudo-node occurs on one side of a road in the right-of-way line work but not on the other. Both of these cases occur at what should be pseudo-nodes in the center-lines and cause minor errors.

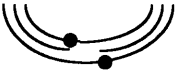


Figure 9a.

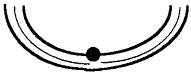


Figure 9b.

Step 15.Auxiliary two-arc-ring closure. Edit the arcs to close the "rings" generated in Step 14.

Step 16. Overlay and windowing. Intersect these center-line arcs with a right-of-way polygon layer and window out "center-lines" which are not inside any right-of-way

CONCLUSIONS

The programs tested well. The great majority of line work was captured. Artifacts were essentially non-existent. Figures 14 through 17 show some rights-of-way and center-line output.

The chief problem not solved at this point is the anomaly illustrated in Figure 10. This results in duplicate one-point arcs which have ends corresponding to the same endpoint.

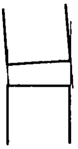


Figure 10.

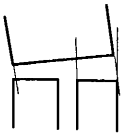


Figure 11.

Compare Figures 10 and 11 with Figures 12 and 13. In these figures translation error, rather than rotation error occurs. Intersections with perpendicular segments occur within the paired arcs and no linework is lost. The translation errors in Figures 12 and 13 and the rotation errors in Figures 10 and 11 are greatly exaggerated to illustrate the point. These errors are not normally visible to the naked eye. In real linework, translation error is greater than rotation error in the great majority of cases. In the rare cases where rotation error is greater than translation error, perpendicular segment intersections are outside paired rights-of-way and linework is lost. Such cases occur most commonly in "downtown" type neighborhoods.

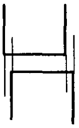


Figure 12.

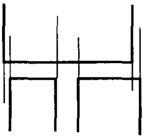


Figure 13.

Work on this anomaly and on the related anomaly shown in Figure 11 is planned This will involve modification of the Auxiliary intersection points table generation step. It is expected that it will shortly be completed.

Other problems observed were:

a) Intersections failed to close in a minority of cases.

b) Problems occurred where paired right-of-way sections deviated significantly from parallel.

c) Problems sometimes occurred at map edges.

d) No provision has yet been made for five-or-more-arc intersections, cul-de-sacs, or filleted intersections.

Future plans include dealing with these remaining problems, including an error report as a final step, and examining what related problems can be treated using similar approaches.

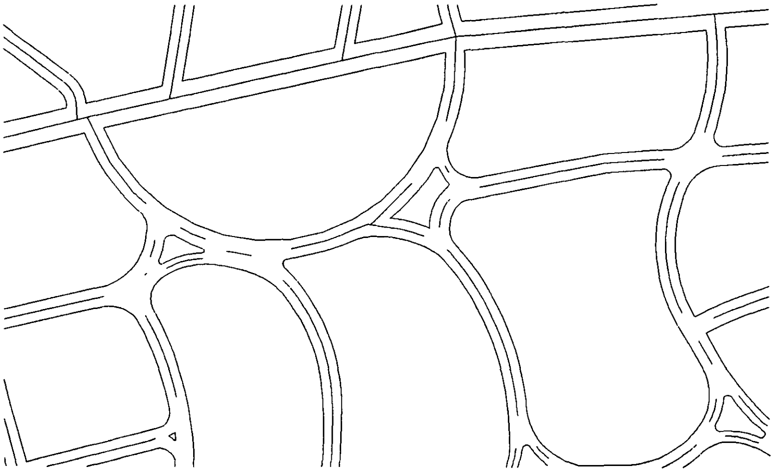


Figure 14.

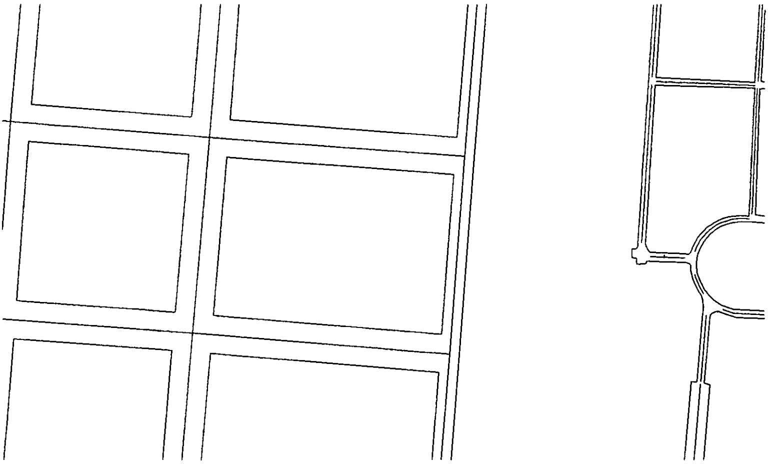


Figure 15.

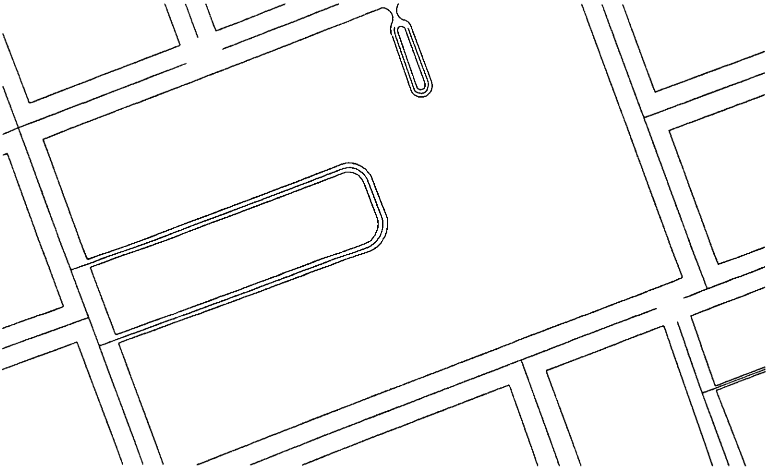


Figure 16.

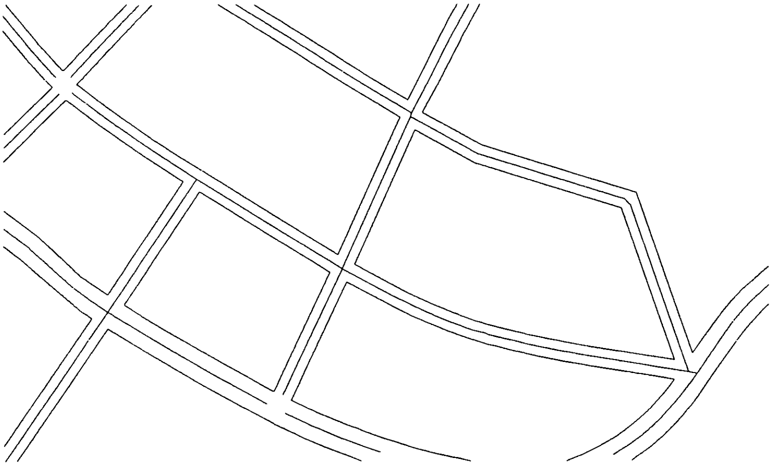


Figure 17.

ACKNOWLEDGMENTS

Acknowledgments are extended to Dr. David J. Cowen of the University of South Carolina Humanities and Social Sciences Computing Lab and his staff who were very helpful in preparing and reviewing this paper. Acknowledgments are extended to my boss, Stan Lawrimore, who encouraged me to write this paper and helped with final obstacles. I would also like to thank Paul Smith, who was very helpful in obtaining clearance, and David Smith I must also acknowledge the City of Columbia, South Carolina which made available the right-of-way data which were used for testing purposes. And I would extend very grateful acknowledgments to Carole Shealy, without whose help this paper could not have been recovered after I lost my disk.

REFERENCES

Beard, M. Kate 1987, "How to Survive on a Single Detailed Database", *Proceedıngs Auto-Carto 8,* Baltimore, pp. 211 -220.

Blum, H. 1967, "A Transformation for Extracting New Descriptors of Shape", *Symposium on Models for the Perception of Speech and Visual Form,* Weiant Whaten-Dunn (ed.), MIT Press, Cambridge, Mass., pp. 362-380.

Brassel, Kurt E. 1985, "Strategies and Data Models for Computer-Aided Generalization", *International Yearbook of Cartography,* 25, pp. 11-29

Brassel, Kurt E., and Heller, Martin 1984, "The Construction of Bisector Skeletons for Polygonal Networks", *Proceedings, First International Symposium on Spatial Data Handling,* Zurich, pp. 117- 126.

Brassel, Kurt E., and Weibel, Robert 1988, "A Review of and Conceptual Framework of Automated Map Generalization", *International Journal of Geographical Information 5ystems,* 2/3, pp. 229-244.

DeLucia, A., and Black, T. 1987, "A Comprehensive Approach to Automatic Feature Generalization", *Proceedings, 13th International Cartographic Association Conference,* Morelia, Mexico, 4, pp. 168-191.

Gao, Peng and Minami, Michael M 1993, "Raster-to-Vector Conversion: A Trend Line Intersection Approach To Junction Enhancement", *Proceedings Auto-Carto 11,* Minneapolis, pp. 297-303.

McMaster, Robert B. 1991, "Conceptual Frameworks for Geographical Knowledge", In Buttenfield, B. P., and McMaster, R. B.(eds.), *Map Generalization: Making Rules for Knowledge Representation,* Longman, London, pp. 21-39.

McMaster, R. B., and, Monmonier, M. S. 1989, "A Conceptual Framework for Quantitative and Qualitative Raster-Mode Generalization", *Proceedings GIS/LIS '89,* Orlando, Florida, 2, pp. 390-403.

Mark, D. M. 1989, "Conceptual Basis for Geographic Line Generalization", *Proceedings Auto-Carto 9,* Baltimore, pp. 68-77.

Montonari, U., "Continuous Skeletons from Digitized Images", *Journal of the Association for Computing Machinery,* 15/4, pp. 534-549.

Nickerson, B G., and Freeman, H. 1986, "Development of a Rule-Based System for Automatic Map Generalization", *Proceedings, Second International Symposium on Spatial Data Handling,* Seattle, pp. 537-556.

Rosenfeld, A., and Pfaltz, J.L. 1966, "Sequential Operations in Digital Picture Processing", *Journal of the Association for Computing Machinery,* 13/4, pp 471-494.

Zhan, Cixiang, 1993, "A Hybrid Line Thinning Approach", *Proceedings Auto-Carto 11,* Minneapolis, pp. 396-405

PASS LOCATION TO FACILITATE THE DIRECT EXTRACTION OF WARNTZ NETWORKS FROM GRID DIGITAL ELEVATION MODELS

David Wilcox and Harold Moellering
Department of Geography
Ohio State University
Columbus, Ohio 43210
Phone: 804/642-7199, 614/292-2608
E-mail: dwilcox@vims.edu, geohal+@osu.edu

ABSTRACT

The topologic arrangement of peaks, pits, ridges, valleys, passes and pales that form a cartographic surface has been called a Warntz Network. Most network extraction algorithms in the literature focus primarily on either the valleys or the ridges and are not easily applied to extract the full connected Network. An algorithm has been implemented that first locates passes and then traces ridges to peaks and valleys to pits to produce the full Warntz Network from a grid DEM. Three critical point location algorithms from the literature have been tested for their ability to accurately locate passes. In addition, a new algorithm has been developed that starts with the maximum elevation on the surface and then steps through each unique elevation, checking for changes in the eight-connectivity of the regions above and below the current elevation. If a change is detected it is checked to see if it represents a pass. The results from testing the system on several mathematical and terrain surfaces show that the new pass location algorithm performs significantly better than the other three algorithms, producing far fewer anomalous passes and missing only a few passes. The resulting network closely approximates the true Warntz Network.

INTRODUCTION

The study of cartographic surfaces including topography has been an active area in the field of analytical cartography and geographic information systems. Important work has been completed to develop methods for surface representation, visualization and analysis. The fairly recent production of medium scale grid digital elevation models (DEMs) has encouraged developments in computer-assisted geomorphology, geology and hydrology. In particular, many large scale hydrological models have been implemented. In addition to this practical work, a smaller group of researchers have been developing a theoretical basis for cartographic surfaces. In 1966, William Warntz, a key member of this second group, extended some of the ideas put forward by Cayley (1859) and Maxwell (1870) to produce a model of the structure inherent in a cartographic surface. The network he described, composed of ridges, valleys, pits, peaks, passes and pales, has been called a "Warntz Network."

This paper describes a system that can directly extract a Warntz Network from a grid digital elevation model. Since no full algorithm for doing this was located in the literature, an experimental approach has been used to design and test an extraction method formed from a combination of existing algorithms. It is based on an idea presented by Mark (1978) that begins by locating the passes on the surface and then uses them as starting points to define the remainder of the network. To assist in the development and testing of this algorithm, an interactive cartographic system has been built that allows the user to execute a particular combination of pass location algorithms and slope line tracing algorithms to build the network. Both visualization and analysis tools are included in the system to assist in assessing the effectiveness of each combination of algorithms.

CONCEPTUAL BACKGROUND

Before describing the elements of the algorithm in more detail and discussing the results of network extraction, it is useful to review the theory upon which this work is based, including a formal definition of the Warntz Network.

When presented with a terrain surface, the human mind quickly identifies the structural elements. For example, mountains, valleys, ridges, lakes and passes may be remembered as distinct elements of the surface. In the late 19th century Cayley (1859) and Maxwell (1870) described how these features could be defined by the pattern they formed on a contour map, but it was not until a century later that the definitions were pulled together by William Warntz (1966) into a combined surface model. Warntz realized that features on a surface do not occur in isolation but must be interrelated. He described a network composed of peaks, pits, passes and pales as nodes and ridges and courses (valleys) as edges. The ridge and valley lines are slope lines (lines of maximum gradient) that flow from a peak through a pass to a peak or from a pit through a pass to a pit respectively. His definition provides the basic concepts that have been used to develop a structural model for continuous cartographic surfaces.

The Warntz Network model has been modified and clarified by several scholars. A rigorous definition using the language of graph theory is given by Pfaltz (1976). Pfaltz also demonstrates how the model can be used to generalize the graph while maintaining its basic structure. Additional research in the use of graph theoretic procedures to generalize the Pfaltz Graph model of a surface is reported by Wolf (1991). In this work, Wolf shows that the surface can be more accurately generalized by incorporating the elevation change along the arcs in the graph, forming what he calls a metric surface network.

Though the development of a graph theoretical model describes the structure of the network, it is also desirable to have firm definitions for the specific features of which it is composed. Pits, peaks and passes are easily described as critical points of the surface (*i.e.* points at which the magnitude of the surface gradient is zero). Critical points can be further classified as pits, peaks and passes by examining the Hessian matrix of second derivatives. A formal description of all the elements of the surface network, including ridges and valleys, is given by Nackman (1984). He restricts the model to Morse surfaces which are twice differentiable with no degenerate critical points. The surface network (called a two-dimensional critical point configuration graph in the paper) is composed of the critical points of a surface (points for which both x and y partial derivatives are zero) and the flow lines that connect them. An ascending slope line through a point is a vector-valued function of one variable from the real numbers to the real plane defined in the following manner:

$$\bar{\sigma}(t): \Re \to \Re^2 \tag{1}$$

$$\bar{\sigma}'(t) = \nabla f(\bar{\sigma}(t)) \tag{2}$$

$$\sigma(0) = x_0 \tag{3}$$

A descending slope line is defined similarly but

$$\bar{\sigma}'(t) = -\nabla f(\bar{\sigma}(t)) \tag{4}$$

On a Morse surface every point has a single slope line that runs through it (at critical points this flow line is a single point). A slope line is said to reach a point if the limit of the path as $t \to \infty$ approaches the point. A ridge is defined to be an ascending slope line from a pass that reaches a second pass or a peak and a valley is defined to be a descending slope line from a pass that reaches a second pass or a pit.

Warntz Networks, or closely related structures, have been studied or used in several different areas of research. Cartography provides the basic theory for spatial surfaces. In hydrology the ridge and channel network structure has a strong similarity to the Warntz Network. In terrain analysis and geomorphology, the Warntz Network has been used to separate study areas into different geomorphic units and has been used to generalize topographic surfaces. In image processing many of the same topographic concepts are applied to identify the "ridges" and "valleys" in a gray scale image. Since the theory and algorithms discussed in this paper were developed within these research areas it is useful to briefly review how each of these areas makes use of surface networks.

The field of cartography has historically been focused on the study of the map as a communication tool. In 1984 the definition of a map was expanded to include 'virtual' maps in additional to the traditional 'real' maps (Moellering, 1984). These virtual maps have either no permanent tangible reality (CRT displays, mental maps, etc.), or are not directly viewable (digitally stored maps, field data, etc.). The field of analytical cartography is based on the analysis of maps using this expanded definition. An immediate benefit gained from this expanded view is the ability to identify 'deep structure' (the structural relationships that can be represented in a virtual map but are not visible in the surface representation of the real map) (Nyerges 1991). It is the deep structure of the DEM (a virtual map) that is captured by the Warntz Network.

Hydrologists use ridge and channel network models as an important tool for studying hydrological processes. A good survey of the many algorithms that have been proposed for extracting drainage networks and basin boundaries from grid DEMs is given by Band (1993). One method uses the local neighborhood to compute the magnitude of the surface gradient and thereby infer surface water flow. The flow direction is assigned towards the neighboring cell with the largest downward slope. Once a unique flow direction has been assigned for each cell, the total number of cells upstream from each cell can be calculated. Channels are determined by selecting cells that drain an area larger than a fixed threshold. The resulting collections of cells are then thinned to single cell-width lines to show the channel centers. A second method uses the morphologic structure of the land to predict the presence of channels. A local operator is used to determine the structure in a neighborhood of the point. If the land is concave upward, the cell is classified as a channel. As with flow line tracing, the cells that are classified as channels will need additional processing. In this case, the networks are not guaranteed to be continuous. To connect the segments of the network, flow lines can be traced downhill from one segment to another to complete the network (Band 1986).

Like hydrologists who are interested in determining the shape and structure of drainage basins, geomorphologists are also interested in the structure of the landscape. In addition to drainage basins, they are interested in the geologic and other morphologic features that exist in the landscape. Several models, including those presented by Dikau (1989) and Feuchtwanger and Blais (1989), have been designed to assist in the storage and analysis of surface structure. Since it is desirable to store a detailed model for use at multiple scales, it is useful to be able to generalize these surface models. A review and conceptual basis for terrain generalization is given by Weibel (1992). For the extended surface network a set of formal generalization rules has been defined based on graph theory. It was shown by Pfaltz (1976) that the Pfaltz Graph can be generalized by contracting a set of peaks and passes into one peak or a set of pits and passes into one pit. Wolf (1991) improved on this work by incorporating the elevation difference into the surface network. In the new structure, the elevation difference and local structure can be used to locate less important elements of the network for contraction.

Many of the concepts developed for terrain representation and analysis have been applied to the interpretation of gray-scale imagery. If one considers the gray levels in an image to represent elevation values, the image can take on many of the characteristics of a terrain surface. One of the first attempts to use surface structure to interpret gray-scale images is described by Toriwaki and Fukumura (1978). The local technique developed by Peucker and Douglas (1975) was applied with a slight modification to separate the image into peak, pit, ridge, ravine, hillside, and pass pixels. The pixels were then thinned and connected into a network much as described by Band (1986). A more complex method for identifying local structure was introduced by Haralick (1983). Instead of directly using the discrete gray-level values, a bicubic surface patch was fit to the 5x5 neighborhood of each pixel. The directional derivatives of this continuous, smooth function were factored to determine the location of zero-crossings, which identify the position of surface specific points. Once these points were located, the sign of the second directional derivative was used to determine if the crossing was a ridge or valley. Pixels containing both ridge and valley crossings were classified as saddle points.

METHOD

An interactive cartographic system has been built to investigate methods for extracting Warntz Networks. The system, which is written in C on a Macintosh personal computer provides 2D and 3D views of the surface and permits the user to control and analyze the network extraction process. Four pass location algorithms were used to generate starting points for the slope line tracing procedure. These were tested on two mathematical and two empirical surfaces.

Three algorithms for locating critical points on a grid DEM were selected from the literature: one from Peucker and Douglas (1975), one from Toriwaki and Fukumura (1978) and one from Haralick et al. (1983). A fourth algorithm has been designed using a significantly different method, but is based on insight gained into the pass location process by studying these three existing algorithms.

The simplest of the three pass-location algorithms is that presented by Peucker and Douglas (1975). It is based on a idea proposed by Greysukh (1967) that determines the local character of a surface by tracing a circle through the neighbors of a point. The number of times the circle crosses the elevation of the center point and the order in which this occurs can be used to characterize the central point. If the point is a pass, it must be a maximum in one direction and a minimum in another. Therefore, the circle must cross the plane of the center at least four times, twice for the maximum and twice for the minimum. The effect of high-frequency noise can be somewhat controlled by requiring the local negative and positive relief to exceed a set threshold. Though the algorithm they describe requires relief in only one direction to exceed the threshold, for this work the algorithm has been slightly modified to require a significant amount of relief in both directions. This modified version of the algorithm will be called the "Peucker/Douglas pass algorithm" for the purpose of this paper.

The second algorithm tested for locating passes is an enhancement of the Peucker and Douglas algorithm and is presented by Toriwaki and Fukumura (1978). Though the emphasis is on the extraction of features from digital images, the technique can easily be applied to DEMs. Two local statistics are calculated for each cell in the image or DEM: a connectivity index (CN) and a curvature index (CC). Since the connectivity index returns the number of connected strings of cells above the central cell, a value of two or above indicates that it is a pass. This algorithm will be called the "Toriwaki/Fukumura pass algorithm" for the purpose of this paper.

The third pass-location algorithm that has been tested also comes from image analysis and is given by Haralick et al. (1983). For each cell a cubic polynomial surface is fitted to its 5x5 neighborhood through the application of a set of ten local filters, one for each term in the polynomial. The Hessian matrix of second derivatives of the polynomial is then computed. If the matrix has two distinct eigenvectors, the roots of the directional derivative in each of these directions are computed to find zero-crossings (if there is only one eigenvector, the Newton direction and the direction orthogonal to the Newton direction is used). If a zero-crossing is found within the area of the cell, the Hessian matrix is recomputed at the crossing point and used to classify the point as a ridge, valley or pass. If a pass or both a valley and a ridge are found in a cell, the cell is marked as a pass. The implementation of this third algorithm will be called the "Haralick pass algorithm" for the purpose of this paper.

The three algorithms described above have one important characteristic in common. They all must make a decision on the character of a cell in the DEM based on the limited amount of information present in a 3x3 or 5x5 neighborhood. This may cause structures with a primary spatial frequency lower than this to be missed. For instance there is often a wide flat area at the center of a pass. If it is large enough, the entire local neighborhood will have equal values and will be interpreted as being flat. The thresholding function found in most image processing systems has been the inspiration for a new pass location algorithm that is able to overcome this problem. If all areas above or below a threshold elevation are grouped and the changes in topology of the groups as the threshold varies are investigated, it is clear that the appearance or disappearance of groups identifies minima and maxima and the connection of groups identifies passes. This is illustrated in Figure 1. If the boundary between groups is viewed as a contour line, the connections correspond to the self-intersecting contours that Cayley, Maxwell and Warntz all used as indicators of the location of a pass.

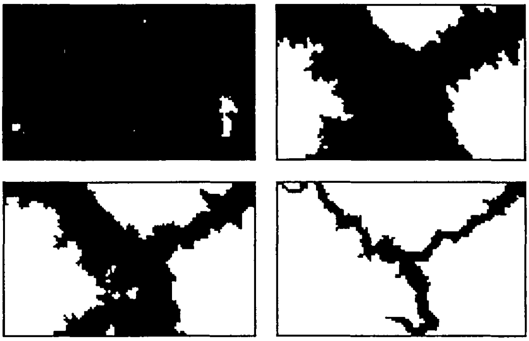


FIGURE 1. Threshold at four different elevations (higher areas are in white)

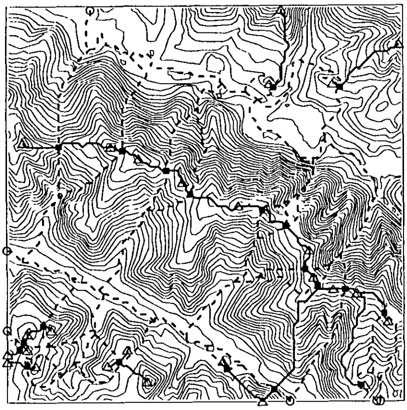
The newly developed algorithm applies this thresholding concept to locate passes. It starts with the set of the cells with the highest elevation. This set is enlarged by systematically adding each layer of lower elevation values to the set until all elevation levels have been processed. As each layer is processed, every cell that is added is checked to see if it changes the eight-connectivity of the higher elevations or the four-connectivity of the lower elevations. If it does change the connectivity, it represents a pass feature; however, some additional checks are necessary to eliminate passes that are too small. This is done by checking that there are no adjacent passes and that there are at least two ridges and two valleys originating from the pass. If the cell passes all of these tests, it is flagged as a pass.

Once the passes are located using one of the four algorithms described above, they are processed by the slope line tracing algorithm which takes a pass location as input and traces the network slope lines emanating from the pass in two steps. The first step is to identify at least two ridge starting points and two valley starting points in the eight neighbors of a cell. The algorithm handles most cases gracefully and forces starting points if necessary (for example in flat areas). Each of these starting points is passed to a second algorithm that attempts to trace a slope line either uphill or downhill starting at the desired neighbor. This second algorithm is complicated by the need to check for intersections and maintain the vector topology as each line is traced. In flat areas, a procedure is used that first identifies the extent of the area and locates any possible outlets. It then builds a distance surface starting at the outlets which is used to direct the slope line to the nearest outlet. In other areas (usually the majority of cases) the slope line is directed to the lowest or highest eight-neighbor using the method described by Jensen and Domingue (1988).

Both artificial and real surfaces have been used to test the system. Two artificial surfaces have been computed from mathematical functions and provide well-defined networks to test the accuracy of the algorithms. The first mathematical surface was generated for a 50 row by 50 column area using a simple cosine function for which the Warntz Network is easily determined. The other mathematical surface was taken from surface III in (Morrison, 1971) and sampled for a 100 row by 100 column area.

The two real surfaces, from California and Colorado, test how well the algorithms perform in a less controlled environment. To provide a model for comparison, a "truth" Warntz Network for each of the surfaces was created by carefully locating the passes on the surface and tracing the slope lines both up and down to capture all the structural elements.

The first terrain data set is a portion of the La Honda California USGS 1:24,000 DEM. This area has been the subject of previous work and therefore its characteristics are well understood. The surface is composed of one main ridge that is flanked by two main



▪ Passes ——— Ridges
△ Peaks ------- Valleys
○ Pits ---------- Connectors

FIGURE 2. "Truth" network for the La. Honda, California surface.

valleys flowing southeast to northwest. The portion selected for testing and shown in Figure 2 is 100 rows by 100 columns of 30 meter cells starting at row 200 and column 75 in the original DEM. The "truth" network that has been determined by the author is shown with 10 m contours in Figure 2. Note that because some of the ridges and valleys leading into the area and leaving the area originate at passes outside the area, they are not captured in the "truth" network.

The second terrain data set is a portion of the Platte Canyon, Colorado USGS 1:24,000 DEM. It is larger and structurally more complex than the La Honda, California surface and was chosen because it contains well-defined terrain and both the DEM and hydrology layer were easily obtained from the USGS. The portion selected for testing is 154 rows by 253 columns of 30 meter cells starting at row 266 and column 87 in the original DEM.

RESULTS

Each of the four pass location algorithms was tested with each of the four surfaces to generate a total of 16 networks. The four networks for the La Honda, California surface are shown in Figures 3 to 6. A thinning function has been applied to all networks to eliminate adjacent passes and produce a more correct network. The passes located by each pass location algorithm are broken down in Table 1 into the number and percentage of correct and anomalous passes and the number of correct passes that were missed. A clear pattern is present in the data. The Peucker/Douglas and Toriwaki/Fukumura algorithms tend to locate the correct passes, but also add many more anomalous passes. The Haralick and the new pass location algorithms are more conservative and add fewer passes. In addition, the new pass location algorithm correctly identifies most passes, giving it the best of both worlds. The advantage of the new pass location algorithm is particularly evident with the empirical surfaces for which the other algorithms identify three to six times too many passes. This large number of anomalous passes almost completely dwarfs the number of correct passes which only represent 14% to 18% of the total. In contrast, the number of correct passes located by the new pass location algorithm make up a full 61% to 69% of the total.

The Peucker/Douglas pass location algorithm performs fairly well on the mathematical surfaces. However, the accuracy on empirical surfaces is much worse. For each correctly located pass, the algorithm also identifies more than five anomalous passes. This is amplified by the many anomalous slope lines generated from each of the passes. A close inspection of the surfaces has identified three cases that are incorrectly classified by the algorithm. The first case occurs when a wide pass is centered on a flat area that is larger than the 3 x 3 neighborhood used in the algorithm, causing it to be missed. The second case is caused by the indeterminate nature of the cell diagonals. The algorithm assumes that it is possible to connect a cell with its diagonal neighbors, but a ridge or valley structure may prohibit this. The third case occurs along passes that are at least two cells in length. The algorithm makes no provision for checking neighboring cells and therefore identifies both pass cells individually creating a duplicate pass. To some extent, the pass threshold can be adjusted to reduce the number of anomalous passes. However, by requiring the local relief to be greater than the pass threshold, important passes with low spatial frequency, that often have little local relief, might be missed. For the results given in Table 1

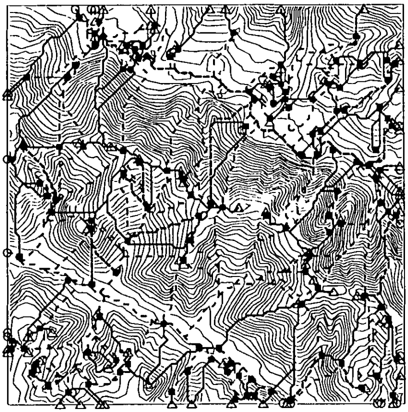


Figure 3. Results from the Peucker/Douglas pass location algorithm

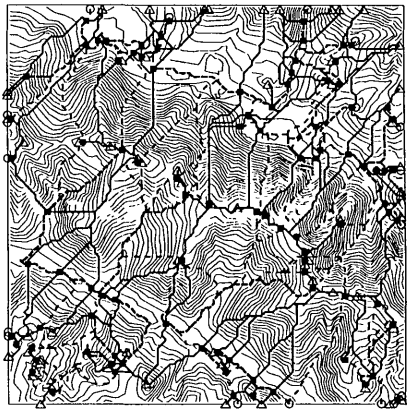


Figure 4. Results from the Toriwaki/ Fukumura pass location algorithm

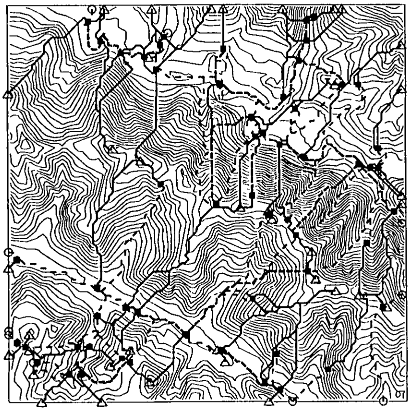


Figure 5. Results from the Haralick pass location algorithm

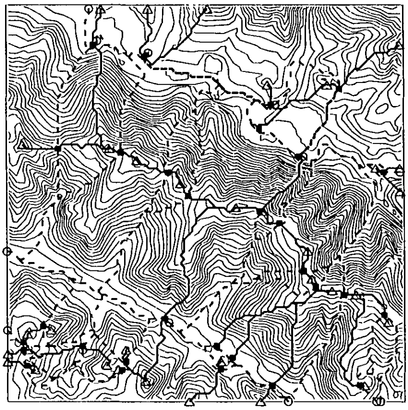


Figure 6. Results from the new pass location algorithm

the pass threshold has been adjusted as high as possible so that no passes are missed. There appears to be no direct method for determining the best threshold value without a prior knowledge of the pass locations.

The Toriwaki/Fukumura pass algorithm has many of the same attributes as the Peucker/ Douglas pass algorithm described above. It is also restricted to a 3 x 3 neighborhood and makes the same assumptions about the position of ridges and valleys that cut across diagonals. Though no threshold is u^sed, fewer erroneous passes are identified. Unfortunately, it also tends to miss more passes. One particular problem is evident in the processing of the double cosine surface. Only one of the five passes is found due to the way equal-value neighbors are treated. Since the algorithm is designed to operate on boolean values, the neighbors must be classified as either above or below the center cell. Therefore, an arbitrary choice must be made to classify equal-value neighbors as either above or below. This choice can cause passes to be missed in even very small flat areas.

The Haralick pass algorithm benefits from its larger 5 x 5 neighborhood that permits it to identify pass structures that are too broad to be captured by a 3 x 3 neighborhood. It is also less sensitive to local variations due to the smoothing effect of the polynomial surface patch. The passes identified by the Haralick algorithm tend to be accurate. However, it is

TABLE 1. Pass Location Algorithm Results

Double Cosine (5 passes)	Total	Correct*		Anomalous*		Missed**	
Peucker/Douglas	5	5	100%	0	0%	0	0%
Toriwaki/Fukumura	1	1	100%	0	0%	4	80%
Haralick	5	5	100%	0	0%	0	0%
New algorithm	5	5	100%	0	0%	0	0%
Morrison III (33 passes)	Total	Correct*		Anomalous*		Missed**	
Peucker/Douglas	35	33	94%	2	6%	0	0%
Toriwaki/Fukumura	39	33	85%	6	15%	0	0%
Haralick	27	27	100%	0	0%	6	18%
New algorithm	33	33	100%	0	0%	0	0%
La Honda, California (20 passes)	Total	Correct*		Anomalous*		Missed**	
Peucker/Douglas	134	19	14%	115	86%	1	5%
Toriwaki/Fukumura	96	15	16%	81	84%	5	25%
Haralick	57	8	14%	49	86%	12	60%
New algorithm	31	19	61%	12	39%	1	5%
Platte Canyon, Col. (88 passes)	Total	Correct*		Anomalous*		Missed**	
Peucker/Douglas	581	85	15%	496	85%	3	3%
Toriwaki/Fukumura	371	68	18%	303	82%	20	23%
Haralick	245	44	18%	201	82%	44	50%
New algorithm	125	86	69%	39	31%	2	2%

* The correct and anomalous passes are also given as a percentage of the total passes

** The missed passes are also given as a percentage of the total correct passes.

often too restrictive and does not identify some of the correct passes. These missed passes cause sections of the Warntz network to also be missed, leaving a disconnected network. In the La Honda surface the algorithm missed more than half of the passes and added 49 incorrect passes, significantly affecting the quality of the final network (see Table 1 and Figure 5). The algorithm also suffers from the lack of continuity between surface patches. This can cause a slight inaccuracy in the surface fit to place the critical point outside the pass cell. If this occurs in all cells in the neighborhood of a pass, the pass is missed.

The new pass location algorithm has been devised to address some of the difficulties faced by the other three algorithms. As shown in Table 1, it performs very well on all the surfaces. It identifies as many or more of the correct passes than the other algorithms and does not create nearly as many anomalous passes. This is made particularly clear if the number of correct passes as a percentage of the total passes identified by the new pass location algorithm is compared to the same percentage for the other algorithms on the two empirical surfaces. For the La Honda, California surface 61% of the passes identified by the new pass location algorithm were correct, far more than the 14% to 16% for the other algorithms. Similarly, 69% of the passes identified by the Wilcox algorithm for the Platte Canyon, Colorado surface are correct compared to the 15% to 18% for the other algorithms. The networks produced from the new pass location algorithm passes are very close to the "truth" networks, deviating only in small details that do not significantly affect the structure.

CONCLUSIONS

Though a full Warntz Network extraction procedure is presented in this paper, the focus has been on the pass location component. Early testing of the system indicated that the correct location of passes is the most critical component of the extraction procedure because the entire network must be based on the identified passes. Testing of the Peucker/Douglas, Toriwaki/Fukumura and Haralick pass location algorithms revealed a tendency to miss broad pass features. The algorithms were also shown to be quite easily confused and can recognize small irregularities in the surface as passes. This has led to the conclusion that a pass location algorithm must consider the global structure of the surface.

The new pass location algorithm was designed with this goal in mind. By considering the structure of the surface at each unique z-level, the algorithm is able to capture the changes in surface topology that indicate the location of a pass. Since the entire surface is processed at each step, even large flat areas can be recognized as being the center of a pass. The advantage of this approach over using an attribute of the cell's local neighborhood is clearly demonstrated by the significantly larger percentage (61% - 69%) of correct passes reported in Table 1, and shows it to provide a better solution than the other three algorithms for which the correct passes form only a small percentage (14% - 18%) of the identified passes.

The combined Warntz Network extraction system, which starts by using the new pass location algorithm to locate passes, then locates ridge and valley starting points and finally uses the deterministic Jensen & Domingue slope line tracing method to complete the network, has been shown to be a reasonably effective tool for the extraction of Warntz Networks. The networks that have been produced by this method closely approximate the true Network and do not contain the many erroneous elements added by the other combination of algorithms that have been tested.

ACKNOWLEDGMENTS

The edited La Honda, California DEM was provided by Prof. Moellering's Spatial Data Display Project which obtained the original DEM from the U.S. Geological Survey.

REFERENCES

Band, L. E. 1986. "Topographic Partition of Watersheds with Digital Elevation Models." Water Resources Research, 22(1): pp. 15-24.

____. 1993. "Extraction of Channel Networks and Topographic Parameters from Digital Elevation Data." in *Channel Network Hydrology*, K. Beven and M. J. Kirkby Editors. Chichester, New York: John Wiley & Sons, pp. 13-42.

Cayley, A. 1859. "On Contour and Slope Lines." *London, Edinbourgh, and Dublin Philosophical Magazine and Journal of Science*, 18(4th Ser.): pp. 264-268.

Dikau, R. 1989. "The Application of a Digital Relief Model to Landform Analysis in Geomorphology." In *Three Dimensional Applications in Geographical Information Systems*, J. Raper, editor. Philadelphia: Taylor & Francis, pp. 51-77.

Feuchtwanger, M. and J. A. R. Blais. 1989. "Phenomena-based Terrain Data Modelling." *Proceedings of the GIS National Conference*, Ottawa, Canada, vol. pp. 1013-1025.

Greysukh, V. L. 1967. "The Possibility of Studying Landforms by Means of Digital Computer." *Soviet Geography, Review and Translation*, 8: pp. 137-149.

Haralick, R. M., 1983. "Ridges and Valleys on Digital Images." *Computer Vision and Image Processing*, 22(1): pp. 28-38.

Jenson, S. K. and J. O. Domingue. 1988. "Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis." *Photogrammetric Engineering and Remote Sensing*, 54(11): pp. 1593-1600.

Mark, D. M. 1978. "Topological Properties of Geographic Surfaces: Applications in Computer Cartography." in *Harvard Papers in Theoretical Geography*: "Geography and the Properties of Surfaces," Series 5.

Maxwell, J. C. 1870. "On Hills and Dales." *London, Edinbourgh, and Dublin Philosophical Magazine and Journal of Science*, 40(4th Series): pp. 421-427.

Moellering, H. 1984. "Real Maps, Virtual Maps and Interactive Cartography." In *Spatial Statistics and Models*, Gaile &. Wilmont, editors, Boston: D. Reidel, pp. 109-116.

Morrison, J. L. 1971. *Method-Produced Error in Isarithmic Mapping*. ACSM Technical Monograph No. CA-5.

Nackman, L. R. 1984. "Two Dimensional Critical Point Configuration Graphs." IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6(4): pp. 442-450.

Nyerges, T. L. 1991. "Analytic map use." *Cartography and Geographic Information Systems*, 18(1): pp. 11-22.

Peucker, T. K. and D. H. Douglas 1975. "Detection of Surface-specific Points by Local Parallel Processing of Discrete Terrain Elevation Data." *Computer Graphics and Image Processing*, 4: pp. 375-387.

Pfaltz, J. L. 1976. "Surface Networks." *Geographical Analysis*, 8: pp. 77-93.

Toriwaki, J. and T. Fukumura. 1978. "Extraction of Structural Information from Grey Pictures." *Computer Graphics and Image Processing*, 7: pp. 30-51.

Warntz, W. 1966. "The Topology of a Socio-economic Terrain and Spatial Flows." *Papers of the Regional Science Association*, 17: pp. 47-61.

Weibel, R. 1992. "Model and Experiments for Adaptive Computer-Assisted Terrain Generalization." *Cartography and Geographic Information Systems*, 19(3): pp. 133-153.

Wolf, G. W. 1991. "A Fortran Subroutine for Cartographic Generalization." *Computers and Geosciences*, 17(10): pp. 1359-1381.

A Case Study for Hypermedia Cartography:
Radial Growth in Trembling Aspen at Waterton Lakes National Park

Christopher R. Weber
Barbara P. Buttenfield
National Center for Geographic Information and Analysis
Department of Geography
State University of New York at Buffalo
Buffalo, New York 14261

Dennis E. Jelinski
Department of Forestry, Fisheries and Wildlife
Institute of Agriculture and Natural Resources
University of Nebraska-Lincoln
Lincoln, Nebraska 68483-0814

Abstract

Hypermedia software developments afford new opportunities for cartographic visualization. There is a burgeoning inquiry into effective techniques and formats within the hypermedia realm Users may now work with visualizations proactively, initiate data queries and steer the presentation in a manner consistent with the associative power of the human thought Hypermedia formats provide advantages of both the distillation of the data and the portrayal of the authored relationships between data chunks. This research describes the creation and implementation of one such visualization. Forty years of radial growth in trembling aspen tree populations from Waterton Lakes National Park, Alberta, Canada are depicted in macro and micro scales. Meta-information about the display and its contents is provided through hyperlinks to graphics, text, embedded animation, and photos.

Keywords: Hypermedia, multimedia, animation, spatiotemporal, cartography, aspen, biogeography.

Introduction

> "Geographers have long practiced and preached that several journeys across a physical landscape were necessary for a full, multi-viewed understanding of the earth as a home of man. ...if we want true learning and understanding in the spatial domain then there is definitely a place for a hypermap" (Laurini and Thompson 1992, p. 681).

Traditional map production generates single static maps offering data that have been distilled for the map reader through a highly focused, authored view of the world (Monmonier 1991). Animated maps provide multiple views of subject matter or present a dynamic portrayal of spatiotemporal data. These have been extant for well over 10 years, though these constructions still present a highly authored view. Comprehensive reviews of this method and its contribution to cartography to date can be found in Cambell and Egbert (1990) and in Weber and Buttenfield (1993)

To overcome the sequential view of the world presented in animation, Monmonier (1991) has proposed a concept called *Atlas Touring*. This format involves the presentation of several maps as well as statistical diagrams and text blocks, organized into *graphic scripts* composed of *graphic phrases*. Each phrase is "a computer generated sequence of focused graphics tailored to the data and intended to explore a distribution, a spatial trend, a spatial-temporal series, or a bivariate relationship" (Monmonier 1991, p. 4). To utilize this format effectively, Monmonier has called for *experiential* maps which allow the map reader to freely explore the scripts within the atlas. It has been proposed that integration of viewer perspectives into the mindset of the cartographer, researcher, or domain expert, is possible by adopting presentation formats which enable *proactive* user involvement, as opposed to *interactive* involvement (Buttenfield 1993, Buttenfield and Weber, 1993). In this manner, map readers may generate their own views and steer the ordering and duration of graphic script equivalents within an atlas format.

Hypermedia software development and multimedia computer platforms provide cartographers with the capabilities for such presentations. Furthermore, due to cascading costs, such systems are fast becoming the norm rather than the exception (Donovan, 1991a). This has resulted in ready access to these new technological standards and working environments. The implications for both cartography and GIS are vast and the topic of much research (see for example Armstrong et al 1986, Koussoulakou and Kraak 1992, Buttenfield and Weber 1993).

This research reviews the design, construction and iterative refinment of a hypermedia animation using forty years of radial growth data from trembling aspen populations from Waterton Lakes National Park, Alberta Canada. The hypermap can be viewed at macro and micro scales, has embedded metadata, proactive user steering, as well as granularized help and encyclopedic reference functions These features are accessed through "hot" icons within the map itself, through hypertext, and through iconic buttons. The document utilizes animation, pictures, graphs, charts, and text Most importantly, the visualization of tree growth data allows viewers to see trends which are counter-intuitive to the climate and soil conditions existing within the park. Without the visualization, these trends are perceptible only through sophisticated statistical analyses. Alone or combined with the original field research account, the visualization reveals the existence of spatial patterns that would not be immediately apparent in a tabular or static map display.

Background

> "There is a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record. The inheritance from the master becomes, not only his additions to the world's record, but for his disciples the entire scaffolding by which they were erected." (Bush 1945, p. 108)

The concept of hypermedia was first introduced by Vannevar Bush in 1945. He envisioned a *Memex* system which would allow for a mechanized associative linking of the vast amounts of information available in the mid 20th century. Memex would free investigators from being bogged down in the ever-growing mountains of research produced by increased specialization into fringe disciplines. Not only would the record of human achievements continue to be enormously extended, it would be accessible

Today, the term hypermedia refers to information structures in which various nodes containing information are associated through direct links much the same way as information is associated within the human mind. It is an automated as opposed to mechanical realization of Bush's vision. Hyperstructures are thus beyond the sequential style of composition found in most books; they are akin to the neural structure of a thesaurus (Laurini and Thompson 1992). Authors of hypermedia documents present information not only as disparate chunks found in the nodes themselves, but also through the structure of the node links which reflect the author's conceptualization of the relationships between the node topics These data nodes are commonly referred to as *hypernodes* and the associative links as *hyperlinks*.

The data residing at a hypernode may appear in various forms including animation, audio, text, video, CD-ROM, static graphics and spreadsheets Thus, hypermedia structures may be multimedia in nature by incorporating divere media within a single structure (Barker and Tucker 1990). Hyperlinks may reflect a hierarchical structure within the data, a consensus of association developed through multi-user input, or perhaps a domain specific ordering which guides experts between intuitively linked nodes For instance, in a GIS, hypernode linkages may be constructed to allow the circular sequence· map layers, landuse, agriculture, erosion, rainfall, cloud cover, greenhouse effect, carbon dioxide levels, pollution, industry, landuse, map layers

Types of links include *inferential* and *organizational* hyperlinks which may be constructed to connect the data to hardware and non-hypermedia programming languages. *Implication* links can connect hypernodes in inference trees. *Execute* links can be sliders or buttons which are used in high level programming interfaces for steering computation. *Index* links may connect to a relational database. *Is-a* links may set up semantic network structures, and has-a links can describe properties of hypernodes. A complete discussion of hyperlink types and their relationship to semantic networks, neural networks, and relational data structures may be found in Parsaye et al (1989).

Because of the seemingly unlimited possibilities presented by mental association of subject matter, navigation through hypermaps and hyper-databases becomes problematic. Properly structured, maintained, and security governed hyperdocuments should rely upon several modes of navigation There may be a novice mode for newcomers to the realm, and an expert mode for users familiar with the data domain (Laurini and Thompson 1992). Co-authorship may also be possible, through editable text and importation of additional data. This is demonstrable in the hypertext help functions familiar to users of Window's^{TM} (Microsoft Corp. 1992) products

The cartographer developing a hypermedia document must decide whom to target as viewers and what information and meta-information they should have access to. Design considerations for the graphical symbols representing hyperlinks along with their visual placement, and the organization of hyperlinks within hyperdocuments are pressing issues for developers Hyperlinks should be designed and placed so as to provide the user with information about the probable nature of the destination hypernode. The very existence of links in hypermedia conditions the user to expect purposeful, important relationships between linked materials (Landow 1991). Hyperstructures should stimulate the user to explore through stylized iconography and color schema which highlight active hyperlinks. It has been suggested that query capabilities should include devices that allow users to see where they have been, to see new paths to a destination they have previously visited, to review paths taken to a particular hypernode, and to put the user into a previous context (Oren 1987) These design considerations should prevent users from becoming lost or disoriented when perusing a document A review of the psychological implications of becoming lost in hyperdocuments can be found in Harpold (1991)

Hypermedia lends itself to easy and rapid prototyping The lure of its capabilities for popular use has insured authoring of transparent interfaces and scripting languages by product developers. Hypermedia documents can be in themselves high-level, flexible representations that require minimal familiarity for new users to adopt (Woodhead 1991) Their effectiveness for data portrayal and learning is becoming apparent. Yager (1991) contends that hypermedia / multimedia solutions enhance audience immersion and that multi-sensory presentations speed and improve understanding, and increase attention spans. The ability of pictures to enhance recall of textual information has been demonstrated (Kozma 1991). For spatio-temporal data, "there is a statistically significant difference between the time it takes to answer a question (at any reading level) looking at an animated map [shorter time] and the time it takes to answer the same question looking at a static map [longer time] displaying the same spatio-temporal phenomenon" even though the quality of answers is not significantly different (Koussoulakou and Kraak 1992, parenthetic comments by this author) Lastly, Beer and Freifield (1992) report that the US Department of Defense finds that learning assisted by hypermedia / multimedia is cost effective. Interactive videodisc instruction takes a third less time, costs about a third less and is more effective than conventional methods of learning even though the initial outlay costs are high.

With the continued evolution of operating environments and accompanying software which allow dynamic linking of various software packages, personalized hyperlinked GIS may arrive in the near future. Such modular architecture has been proposed for Spatial Decision Support Systems as an alternative to the proprietary data, display and report GISs commonly marketed (Armstrong et al. 1986). Hyper-authoring tools are also readily available to cartographers, especially for desktop computers. Implementation of hypermedia is fast becoming common on these platforms through multimedia extension standards in Microsoft's^{TM} Windows 3.1^{TM} running under DOS, and Apple's^{TM} Quicktime^{TM} running under the Macintosh^{TM} operating system Through these technologies, multimedia and hypermedia are predicted to provide the most popular interface tools for these systems within a few years (Donovan 1991a). Currently these systems allow for embedding of external code into multimedia and hypermedia software through Object Linking and Embedding and Dynamic Linking and Embedding (OLE and DDE in Windows 3.1^{TM}) and through External Commands and External Functions (XCMDs and XFCNs in the Macintosh^{TM} environment). Most importantly, these hyper-desktop environments allow for seamless integration of numerous software packages. Microsoft's^{TM} Office^{TM}, now in distribution, is an example of such an environment at work. Reviews of hypermedia authoring products may be found in Bove and Rhodes (1990), Donovan (1991b), Yankelovich et al. (1991), Bielawski and Lewland (1991), Chua (1991), Scisco (1992) and Kindelberger (1993).

There has been a recent surge of exploration into hypermedia by cartographers, and each production has a unique focus. The earliest of these, the BBC's "Domesday Project", was conceived as a geographical commemorative reflecting the United Kingdom of the 1980's (Openshaw and Mounsey 1987, Ruggles 1992). Its title reflects the 900th anniversary of William the Conqueror's survey of England in the year 1086 A series of Ordnance Survey maps at scales between 1:625,000 and 1:10,000 serve as the base reference for a collection of text, pictures, photos, and relevant data reflecting the culture of the United Kingdom, Isle of Man and the Channel Islands in 1986 The system allows for limited geographic data manipulation including statistical and Boolean comparisons of pairs of maps. The design challenge for this project was the encoding and rapid retrieval of approximately 54,000 pictures, 500 analog maps, and sufficient data and algorithms to produce digital maps from the database in real time. Storage requirements are one of the major hurdles facing multimedia cartographic displays. In this case the solution was a combination of data compression techniques and the selection of videodisk technology which can hold 320 Mbytes of digital data in the sound track in addition to the 54,000 video images.

The advent of Apple Computer's Hypercard^{TM} in the past decade and the potential of hypermedia for educational applications has fostered several pedagogical productions. Tau et al (1988) produced "Map Projections" for the US Department of the Interior Geological Survey This Hypercard^{TM} stack is a tutorial on map projections, their transformations, and their uses. Black and white text and pictures serve as the visual media, though the software's scripting language allows for external commands to various media and multimedia hardware

Geoff Dutton has created two rather interesting Hypercard^{TM} stacks The first is a demonstration of data-quality / data-uncertainty mapping principles which provide viewers with fuzzy boundary views of vector data based on probabilities of generalization during digitization (Dutton 1992). The second is an overview of Dutton's triangular tessellation system for the Earth (Dutton 1991) Like the Map Projections project, the graphics and text in Dutton's stacks are limited to black and white colors, though a smattering of sound bites from popular sources alert the user to transitions and aid in steering Additionally, Dutton has relied heavily on Hypercard's^{TM} scripting language Hypertalk^{TM} to avoid redundant storage problems and allow for flexibility in display parameters

An interactive flip-art animation entitled "A Cartographic Animation of Average Yearly Surface Temperatures for the 48 Contiguous United States: 1897-1986" (Weber and Buttenfield 1993) allows for temporal steering through interactive "hot" regions of the display screen, as well as user initiated viewing of meta-information. The hypermedia aspects of this project are not intricately developed. The notion of allowing the map itself to act as the interface was the experimental focus of the interactivity. Exposure of the animation to viewers around the U.S , and consideration of their comments, resulted in an iterative design process which significantly improved the quality of the display and the comprehensibility of the interactive links. The project continues to accept revisions in design. Voice-over and other sonifications are being considered. The topical motivation for the production adheres strictly to the title.

The "Interactive Multimedia Cartography Project" at the University of Wisconsin-Milwaukee (Tilton and Andrews 1993) is a closely structured hypermedia production in which a major focus is a new paradigm for user navigation through the hyperspace The database is a large one, and typically a user can become lost in such huge hyper-databases. Tilton and Andrews are exploring an alternative approach to navigation through display environments within which information is brought to the user and made available for analysis. The project is still under construction, and the efficacy of the navigation system has yet to be reviewed. The authors are utilizing Macromind^{TM} Director^{TM} software to control external C code and access videodisc stored of images.

"Hyperkarte" (Armstrong 1993) is an intentional pun on the popular software, though this particular hyperstack was written in Supercard^{TM}. This software alternative allows color images, animation, and a more refined stack production system than that afforded in Hypercard^{TM}. The motivation for the project is pedagogic The stack is a tutorial of dot and choropleth mapping techniques. At first glance the linkages between nodes appear to be based on a flow chart structure. Upon closer inspection, the "where am I" node allows radial access to all levels of the stack Because of the voluminous amount of conceptual information provided in the system, Armstrong has focused a good deal of attention on navigational aids One particularly effective solution is provided by "postage stamp" icons which pictorially preview the destination nodes of various hyperlinks

Hypermedia's penchant for interactivity and it's multimedia capabilities made it the ideal choice for this researcher's explorations into sonification for cartographic displays (Weber 1993) Embedded sounds were rated in map-task situations by subjects who mouse-clicked on and dragged "sound coins" to desired destinations on digital maps. Scripting commands were used to record subject responses to external files The topical focus of the research necessitated a flexible multimedia testing environment Macromind^{TM} Director^{TM} software again provided a convenient and malleable set of tools to complete this human-computer interface task, though the resulting product was not, in itself, a hypermedia production.

It can be seen that the development of hypermedia software and hardware has propelled the fulfillment of Vannevar Bush's dream. Educational productions are appearing in nearly every discipline including cartography. Hypermedia's rapid prototyping capabilities and multimedia interfaces are allowing the completion of high level research to be possible in less time and at less expense than ever before. Hypermedia's accelerating evolution is a result of its survival as a freely adaptable species Its continued survival will be a test of its ability to fulfill the expectations of researchers who see it as the connecting web of the information age. For the Waterton Lakes project, it certainly performed as expected

Building the Hyper Document

"Radial Growth in Trembling Aspen at Waterton Lakes National Park" was motivated by curiosity into the efficacy of hypermedia for cartographic display, and a need to efficiently visualize numerical ecological data (Jelinski 1987). The ultimate conceptualization and continuing refinement of the document is an interative process involving cartographers, ecologists, computer scientists and non-expert viewers.

Previous experience animating spatio-temporal data in flip-art form (Weber and Buttenfield 1993) indicated an efficient method of visualization, encapsulating 40 years of forest data. It was decided to reduce the storage volume by using a software scripting language (Macromind^{TM} Director's^{TM} Lingo^{TM}) to compute image frames in real time during the animation. Tree-growth data for cumulative yearly growth at a macro scale (the entire park) and at micro scales (individual aspen populations) is embedded in the animation within unseen text cast members Aspen populations at the macro level are represented by leaf icons whose size and color reflect the cumulative growth and yearly growth respectively. At the micro level, cloned tree stands are likewise represented by dynamic tree stand icons. In the visualization, dynamic icons change size from one year to the next giving the impression of continuous growth.

The dynamic icons are positioned on 3-D terrain maps of the park at both scales. Orientations of these surface maps necessitated cartographic design decisions for each population map and the map of the park as a whole Thus, the positions of time lines, map legends, titles, interactivity buttons, as well as the surface map itself is customized for each scale of portrayal. A binding thread is necessary to thematically link each of the animated maps. A color sequence from yellow to deep green represents the relative yearly growth of populations and clone stands, and is diffused throughout the maps' designs. Hot buttons are dithered with this sequence and embossed with neutral gray borders. Hot text and hot labels are presented in a median green tone. Each map's time line, a histogram of cumulative yearly growth over the depicted 40 years, is "revealed" with present and past years in appearing in median green, and upcoming years remaining neutral gray.

Yearly growth classes are both chromatically intuitive and discernible from the legend. For each map, an information / help button labeled with a "?" links the viewer to granularized help displays. These displays describe the map region, the data being portrayed, and link the viewer to text and pictures describing trembling aspen in the study area. Instructions on how to use the interactive display are also available. In addition, the map legend is embossed by a button border. Clicking on it links the viewer to an explanation of its symbology and data classes. Lastly, at the site level, a "ZOOM OUT" button allows the viewer to return to the macro park level

In a structured hypermedia document, the design implies that a topical structure ought to be reflected in the placement of hypernodes and hyperlinks (Jonassen and Wang 1990). During the early stages of development, the hyperstructure of the aspens hyperdocument was basically a flowchart in nature. That is, micro site animations were accessible only by passing down from the macro park level, population statistics were viewable only through the overview map. No horizontal movement through the hyperspace was possible between micro site displays. Horizontal links were created to interconnect all site level maps (see the gray links in figure 1) and meta-displays These links are accessed through stacked at the edge of the display, labeled with abbreviated names of the study sites. The buttons retain their positions between views but change their labels which always appear alphabetically from top to bottom. This consistent web structure is intuitive and simple for users to comprehend.

Other details of the hyperstructure include links between help screens and field guide descriptions of the trembling aspen species, explanations of interactivity, stored data values, and a map showing the location of the park From the field guide, links provide color pictures of trees, leaves, bark and flowers, as well as a map of the species' range A common explanation of the study area legend is linked from each population node. Conceptual and semantic links interconnect similar information chunks. Users can browse supplementary information nodes from each animated portion. This is consistent with interactive conventions which allow expert users to rapidly glean information from the animated overview, and at the same time provide for interrupted and explanatory views for the novice user (Chignell et al , 1991)

It is important to stress that the iterative design process has played a crucial role in the document's construction. There is very little science to guide designers in creation of hypermedia (Laurel et al. 1990). The nature of the medium necessitates designs that reflect the features of the topic at hand, otherwise the associative value of hyperlinks would be lost. The granularity and content of hypernodes must be determined by the subject. The hyperlinks accessing these nodes must be designed with the subject in mind as well. Constant revision within a conceptual framework has allowed creative solutions to navigation and query problems for this document and have led to innovative cartographic solutions for data portrayal and interactivity

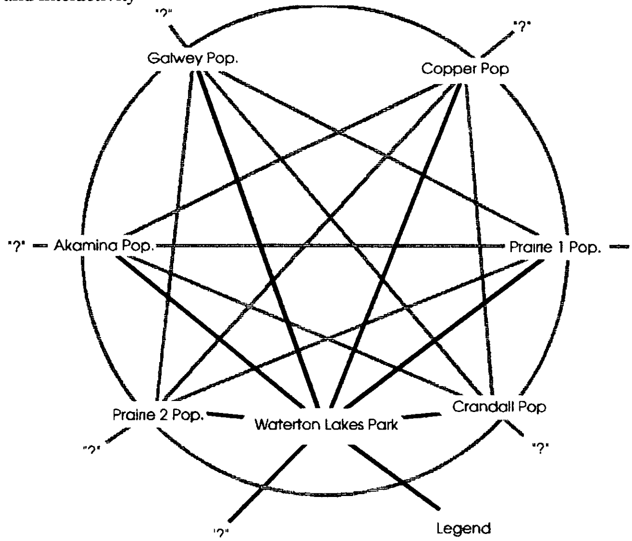


Figure 1: Part of the hyperstructure of "Radial Growth in Trembling Aspen at Waterton Lakes National Park" Gray links lie below black links if conceptualized in 3 dimensions.

The Cartographic Perspective

One pressing issue in cartography today is interface design for computer displayed maps. Computer display real estate, and subsequently map quality, is often traded for interface accommodations Users of GIS and other computer mapping software are confronted with arrays of buttons, slider borders, blocks of text fields, and elusive pop-up menus. "Radial Growth in Trembling Aspen at Waterton Lakes National Park" takes a stab at the traditional screen interface by promoting the map itself as both the interface and the query language

Meta information about the trembling aspen data portrayed is accessed through hot labels Clicking on the portrayal itself (the dynamic icon) allows the user to take a closer look at the clones comprising a population. When the legend is activated, it becomes self-explanatory Clicking on the park map gives rotating 3-D views of the surface. The histogram time-line allows for pause and continue functions, as do each map title. Whenever possible, map elements or hot text provide control, query, and steering functions in lieu of buttons and GUI devices. Users are guided to interface controls through both color scheme and iconography

Traditional cartographic methods pose new questions when viewed in an interactive dynamic map user environment (Koussoulakou and Kraak 1992) The added dimension of spatio-temporal display has afforded the opportunity for bivariate and bi-functional map feature designs in this production The yearly and cumulative radial growth of aspen populations are both able to be depicted with a single icon whose visual variables of color and size serve separate purposes. As a comparison of 40 years of growth is the subject of interest, size for cumulative growth takes visual precedence over color which represents yearly incremental growth. This design is consistent with the hierarchy of visual variables as outlined in Bertin (1983). Position in time is portrayed through color highlighting of a temporal histogram whose remaining visual variables remain constant. Both data icon and histogram designs serve hypermedia steering functions.

Lastly, this animated product circumvents the GIS layer model of data depiction. In most previous animations described in the cartographic literature (see for example DiBiase et al. 1992, Monmonier 1992, Weber and Buttenfield 1993) flip-art and linearly structured animations have been constructed from map pictures whose data had been previously analyzed and portrayed through external systems "Radial Growth in Trembling Aspen at Waterton Lakes National Park" is a hybrid *real time -- real-time later* animation In a real-time animation, "all calculations necessary to produce a frame are immediately followed by its display. In the second situation (real-time later) the results of single frame calculations are written to a file, which is used later for display" (Koussoulakou and Kraak 1992, p. 102). The surface maps for this animation have been previously computed. So have the cumulative yearly growth data, but these are accessed and interpreted in real-time to control the size and color of dynamic icons at multiple scales. This circumvention of the frame-by-frame or flip-art format allows for tight encapsulation of the animation elements, continued expandability, and a subsequent ease of implementation with significant computer memory savings. The temporal growth attributes of each population mimics a primitive object-oriented spatio-temporal data structure. A model such as this could ultimately be used to create a general purpose cartographic animation system in which the user downloads base maps and spatio-temporal data, and then selects or creates dynamic icons for data display.

The Ecological Perspective

Understanding the dynamics of tree growth can be advanced with developments in computer visualization and especially animation. The power of the animation described herein is two-fold. First, the animation is a visualization tool that effectively assists in communicating patterns of growth. Because most forest trees are long-lived, and large sample sizes are required, the data sets become voluminous. Thousands of data values were used in this simulation. Animating radial growth helps clarify complex patterns of development and improves the scientist's understanding of the overall growth structure of the populations. Growth data can be represented at a range of hierarchical levels. The lowest level (individual tree) provides the greatest detail. However, one reaches the limits of knowledge for which lower level data can be useful. Emergent properties may only be visible from an animation of higher levels in the hierarchy (e.g., the growth "behavior" of clones or entire populations). The animation of radial growth in aspens also adds an additional dimension to visualization by dynamically assembling a series of still images or frames which allows the animation to be stopped at any point. The second powerful advantage of animating environmental data such as tree growth is that in addition to being useful as an explanatory tool, it can be used as an exploratory device. In this case it permits evaluation of the effect of environmental controls on the patterns of growth. In a heterogeneous environment such as Waterton Lakes National Park, the animation becomes an important exploratory tool for formulating hypotheses of cause-effect and the effect of environmental changes (e.g., how interannual changes in climate affect tree growth).

Summary

Traditionally, cartographic displays have been designed for static illustration. Reference and navigational maps illustrate the geography of a place at a single time. Thematic maps depict statistical measures for a single sample, or the results of modeling and analysis for a single iteration or set of model parameters. Geographical analysis has begun to address problems and issues that occur in more than a single time slice, and to demand software and hardware technologies that can provide visual tools to assist their research. The display elements traditionally used in geographical analysis are changing from the static map, chart, or table of numeric information to dynamic displays, for example, video, computer simulation, and map animation. With the developments in these display capabilities in hand, one is able to find answers to more complicated and more realistic research questions. The animation of aspen tree growth in Waterton Lakes National Park provides an example of the types of dynamic elements that should become commonplace in geographical analysis and in GIS software very soon.

Readers who are interested in obtaining a copy of the hypermedia document can find it at the anonymous ftp address *alpha2.csd.uwm.edu* with the user name *anonymous* and the password being your entire internet email address. The hypermedia document is one of several archived in the directory *pub/cartographic_perspectives* and is archived under the filename *aspecs.sea.hqx.*

Acknowledgements

This research forms a portion of research Initiative 8, "Formalizing Cartographic Knowledge" and Initiative 10 "Spatio Temporal Reasoning", funded by the National Center for Geographic Information and Analysis, supported by a grant from the National Science Foundation (SBR 88-10917); support by NSF is gratefully acknowledged.

Bibliography

Andrews, S , Tilton, D. (1993) Alternative Interface Designs for Hypermedia Cartography, Annual Meeting, Association of American Geographers, April 6-10, Atlanta, Georgia

Armstrong, M.P., Densham, P.J., Rushton, G. (1986) Architecture for a Microcomputer-based Decision Support System, *Proceedings of the Second International Symposium on Spatial Data Handling*, International Geographical Union, Williamsville, New York, 1986, 120-131

Armstrong, M.P. (1993) Hyperkarte, unpublished hypermedia map design program University of Iowa, Department of Geography.

Barker, J., Tucker, R.N. (1990) The Interactive Learning Revolution, Nichols, New York.

Bertin, J (1983) Semiology of Graphics, The University of Wisconsin Press, Madison

Bielawski, L , Lewland, R. (1991) Intelligent Systems Design, John Wiley and Sons, New York

Bove, T , Rhodes, C. (1990) Que's Macintosh Multimedia Handbook, Que Corporation, Carmel, Indiana

Bush, V. (1945) As We May Think, *Atlantic Monthly*, 176(1), 101-108

Buttenfield, B.P., Weber, C.R., (1993) Visualization in GIS, *Human Factors and GIS,* ed. D. Medyckyj-Scott and H Hearnshaw, London, Belhaven Press

Buttenfield, B.P. (1993) Proactive Graphics for GIS: Prototype Tools for Query, Modeling and Display, *Proceedings Auto-Carto 11*, Minneapolis, MN, 1993, 377-385.

Buttenfield, B.P., Weber, C.R., (1994) Proactive Graphics for Exploratory Visualization of Biogeographical Data, *Cartographic Perspectives*, 19(3), 8-18.

Cambell, C.S., Egbert, S.L (1990) Animated Cartography: Thirty Years of Scratching the Surface, *Cartographica*, 27(2), 24-46.

Chignell, M.H , Valdez, J F., Waterworth, J A. (1991) Knowledge Engineering for Hypermedia, *Multimedia Technology and Applications*, J A. Waterworth, ed , Ellis Horwood, New York.

Chua, T. (1991) Issues in Hypermedia Research, *Multimedia Technology and Applications*, John A Waterworth ed., Ellis Horwood, New York.

DiBiase, D., MacEachren, A.M , Krygier, J.B., Reeves, C. (1992) Animation and the Role of Map Design in Scientific Visualization, *Cartography and Geographic Information Systems*, 19(4), 201-214, 265-266.

Donovan, J W. (1991a) Multimedia Solutions Anticipating a Market, *BYTE*, December, 151

Donovan, J W (1991b) Multimedia Software Sampler, *BYTE*, December, 198-202.

Dutton, G (1991) Zenithal Orthotriangular Projection *Proceedings Auto Carto 10*, Baltimore, MD, March 1991, 77-95.

Dutton, G. (1992) Handling Positional Uncertainty in Spatial Databases. *Proceedings Spatial data Handling Symposium 5*, Charleston, SC, August 1992, 2, 460-469

Harpold, T. (1991) Threnody. Psychoanalytic Digressions on the Subject of Hypertexts, *Hypermedia and Literary Studies*, Delany, P and Landow, G P ed.s, The MIT Press, Cambridge, 171-184

Jelınski, D E. (1987) Intraspecific Diversity in Trembling Aspen ın Waterton Lakes National Park, Alberta: A Biogeographical Perspective, unpublished Ph.D. dissertation, Simon Fraser University, Department of Geography.

Jonassen, D., Wang, S. (1990) Hypertext, Learning and Instructional Design, *Educatıonal Media and Technology Yearbook*, Libraries Unlımited, Inc., Englewood, Colorado, 16, 156-169.

Kındelberger, C. (1993) Multimedia -- The Next Big Wave, *URISA Journal*, 5(1) 121-133.

Koussoulakou, A., Kraak, M.J. (1992) Spatio-temporal Maps and Cartographic Communication, *The Cartographıc Journal*, 29(2), 101-108.

Kozma, R.B. (1991) Learning with Media, *Revıew of Educatıonal Research*, 61(2), 179-211.

Landow, G.P. (1991) The Rhetoric of Hypermedia: Some Rules for Authors, *Hypermedıa and Lıterary Studıes*, Delany, P. and Landow, G.P. ed s, The MIT Press, Cambridge, 81-104.

Laurel, B., Oren, T., Don, A. (1990) Issues in Multimedia Interface Design: Media Integration and Interface Agents, *CHI '90 Conference Proceedings*, Seattle, April 1-5, 133-139.

Laurini, R., Thompson, D. (1992) Fundamentals of Spatial Information Systems, Academıc Press A.P.I.C. Series Number 37, Harcourt Brace Jovanovich Publishers, New York

Monmonier, M. (1991) Ethıcs and Map Design, *Cartographıc Perspectıves*, (10) Summer, 3-8.

Monmonier, M. (1992) Authoring Graphic Scrıpts: Experiences and Principles, *Cartography and Geographic Informatıon Systems*, 19(4), 247-260, 272.

Openshaw, S., Mounsey, H. (1987) Geographic Information Systems and the BBC's Domesday Interactive Video Dısk, *Internatıonal Journal of Geographıcal Informatıon System*, 1(2), 173-179

Parsaye, K., Chignell, M., Khoshafian, S., Wong, H. (1989) Intelligent Data Bases, John Wıley and Sons, Inc., New York

Ruggles, C.L.N. (1992) Structuring Image Data wıthin a Multi-media Information System, *Internatıonal Journal of Geographıcal Informatıon Systems*, 6(3), 205-222.

Scisco, P. (1992) Multimedia Presents, *PC World*, May, 198-200.

Tau, R.A , Vigil, J.F., Buchholz, L. (1988) Map Projections, Open File Report 88-364, United States Department of the Interior Geological Survey.

Tilton, D., Andrews, S.K. (1993) Alternative Interface Designs for Hypermedia Cartography, Annual Meeting, Association of American Geographers, April 6-10, Atlanta, Georgia.

Weber, C.R., Buttenfield, B P. (1993) A Cartographic Animation of Average Yearly Surface Temperatures for the 48 Contiguous United States: 1897-1986, *Cartography and GIS*, 20(3), 141-150.

Weber, C.R. (1993) Sonic Enhancement of Map Informatıon: Experiments Using Harmonic Intervals, unpublished Ph.D. dıssertation, State University of New York at Buffalo, Department of Geography.

Woodhead, N. (1991) Hypertext and Hypermedia: Theory and Applications, Sigma Press, Wilmslow, England

Yager, T. (1991) Information's Human Dimension, *BYTE*, (12), 153-160.

Yankelovich, N., Meyrowitz, N., van Dam, A. (1991) Reading and Writing the Electronic Book, *Hypermedıa and Lıterary Studies*, Delany, P., Landow, G.P. ed.s, The MIT Press, Cambridge, 53-79.

Intelligent Interactive Dynamic Maps

Suguru Ishizaki and Ishantha Lokuge

Visible Language Workshop, Media Laboratory
Massachusetts Institute of Technology
20 Ames St, Cambridge, MA 02139

ABSTRACT

This paper presents an experimental intelligent map system—GeoSpace—which allows information seekers to explore complex spaces of geographic information using dialogue-like interaction. GeoSpace progressively and selectively provides information as an information seeker enters queries while visually maintaining the larger context. Domain knowledge is represented in a form of information presentation plan modules, and an activation spreading network technique is used to determine the relevance of information based on information seeking queries. The reactive nature of the activation spreading network, combined with visual design techniques, such as typography, color, and transparency, enables the system to fluidly respond to the continuous changes in the information seeker's goals and intentions.

INTRODUCTION

The exploration of complex geographic data spaces in an age where both technology and information are growing at exponential rates is a challenging task. Recent developments in interactive computers with high-quality visual displays have provided the GIS designer an opportunity to create more comprehensive environments for presenting complex information. However, most of existing systems fail to support an information seeker's continuous exploration of information and gradual construction of understanding. In other words, although they provide highly sophisticated functionality and displays, they do not relate one presentation to another in response to an information seeker's goals and intentions.

We have applied an activation spreadmg network technique as a representation scheme of domain knowledge, along with an abstraction of information seeking goals and presentation plans, in order to provide interactive maps which embody the following characteristics:

Continuity: We assume that an information seeker's goals are achieved through a series of input queries, or a *dialogue.* The system should be able to consider previous queries as well as to anticipate the forthcoming queries in order to respond to a current input query. We adopted dialogue as a fundamental model of interaction in our system.

Fluid response: When interaction is taking place in a form of dialogue, the system should be able to generate a map, or visual response, in a fluid manner—*as if it was a continuous conversation.* We have used an activation spreading network to create a responsive display which allows the map display to progressively respond to a series of information queries over time.

Visual clarity: Dialogue-based interaction allows the system to limit the range of information that need to be displayed simultaneously, so that the display becomes visually clarified and highly comprehensive, as opposed to a dense display which contains many information elements. GeoSpace uses an activation spreading network to determine the levels of importance for each information element so that an appropriate set of information elements can be chosen to be more visually dominant than the others.

Context preservation: In addition to the visual clarity requirement, we also must visually preserve the larger context in a map, so that an information seeker does not get lost during interaction. GeoSpace achieves this by using various degrees of transparency, type size, and color according to visual design rules.

Two main areas of research have influenced the work presented in this paper. The first area of research involves visual techniques and direct manipulation as a means of exploring complex information space. One such approach is the use of overlapping multiple layers of information in which individual layers are accessible (e.g., Belge 1993, Colby 1991). Most multi-layer approaches provide users with an interface that controls the display based on layers and regions in order to visually simplify the map display. However, this type of interaction becomes cumbersome when the volume of information is large, or when the information seeker does not have prior knowledge about a particular geographic database.

While the above approaches emphasize direct manipulation and visual techniques, other interface displays have been proposed that incorporate domain and presentation knowledge (e.g., Feiner 1993, Maybury 1993, Roth 1993). Maybury introduces an interactive visual presentation method that considers visual presentation as communicative acts (e.g., graphical, auditory, or gestural) based on the linguistic study of speech acts (Maybury 1993). A multimedia explanation is achieved by using rhetorical acts, which is a sequence of linguistic or graphical acts that achieve a certain communicative goal such as identifying an information entity. Rhetorical acts are represented in a form of a plan, which is similar to our representation. Although the system introduced by Maybury enables sophisticated presentation based on a user's single query, it does not have a mechanism to maintain a model of the user's information seeking goals from one query to another.

In this paper, we propose a software architecture for creating intelligent and responsive geographic display systems that allows an information seeker to incrementally asks questions in order to gradually achieve his/her information seeking goals. In the following sections, we first outline the basic functionality of GeoSpace using a simple interaction example. Then, we present a technical framework for ımplementıng the software architecture of GeoSpace. Finally, we discuss potential directions in which our research can be extended.

A TYPICAL SCENARIO

Most GIS users often find it difficult to formulate their information seeking goals in one request. Hence, we believe that an information display that gradually augments this process would greatly enhance the user's comprehension.

As a consequence, we have used the following simple scenario of conversation between an information seeker (IS) and information provider (IP) as an interaction model for GeoSpace. The first query by the IS makes the IP guess what is important to show. After the IP provides information based on the first query, the IS may ask the second query based on what is provided. The IP then determines what is important to show next considering both the first and the second queries. The information seeking dialogue may continue until the IS is satisfied.

An IS's information seeking process can be top-down, bottom-up, or a combination of both. For example, imagine a situation where an IS is trying to locate a new apartment. The IS may start a dialogue by stating that s/he is looking for an apartment. This can be considered top-down since the IS provided the ultimate goal of the dialogue. In this case, the IP is not certaın about what kind of detailed information the IS is aware of. On the other hand, the IS may ask for a particular location (e.g., "Where is Cambridge?"). This can be considered bottom-up since, it targets a specific item of data. In this case, the IP is not certain about what the IS's ultimate goal is.

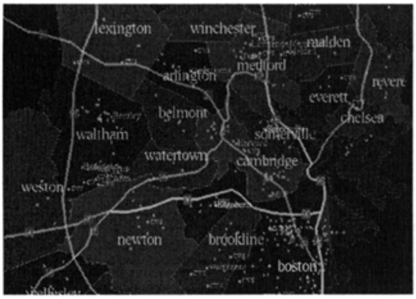


Figure 1. Map of Boston area showing the dense nature of the display.

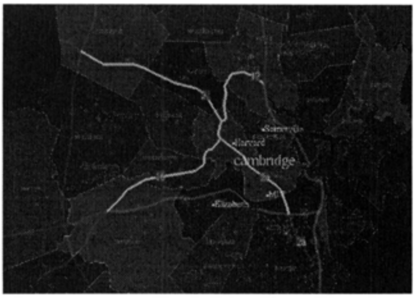


Figure 2."Show me Cambridge."

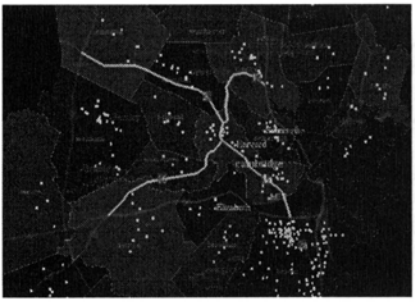


Figure 3. "Show me crime distribution."

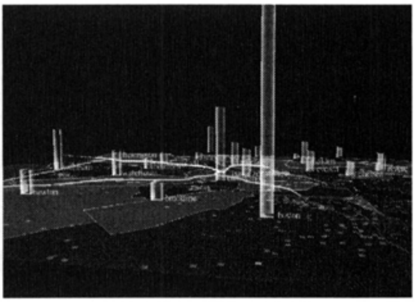


Figure 4. "Show me crime statistics."

In GeoSpace, we consider the IP to be an expert in both domain information and visual presentation and the IP's knowledge is canonicalized in a form of reactive patterns. Instead of deliberately reasoning about what to present every time the user asks a question, the IP simply reacts to it by using canonical presentation techniques.

Based on this scenario, we have developed: (1) a knowledge representation scheme for representing domain knowledge together with visual design knowledge, (2) a computational mechanism whereby the system reacts to a series of user requests (i.e., information seeking goals) while maintaining overall context.

There have been interface approaches to interactive map systems which use queries coupled with graphical displays both for narrowing down information to be presented (e.g., Ahlberg and Shneiderman 1994, Goldstein and Roth 1994) and for supporting users' exploration of the data space (Egenhofer 1990). The information seeking scenario used in GeoSpace emphasizes the latter in its purpose.

The rest of this section presents a simple interaction example of GeoSpace in order to introduce its basic functionality which was developed based on the scenario described above. Figure 1 shows a snapshot of the initial state of the display. The visual complexity of this map display makes it hard for users to discern specific information while interacting without getting lost. The following interaction examples illustrate how GeoSpace dynamically generate a map display according to a series of user queries. Imagine a person new to the Boston area tries to explore the information around the area so as to look for a place to live. Having heard of the perilous life styles of people in Boston, suppose that the person is interested in crime distribution statistics and accessibility to hospitals in the neighborhood.

First, the IS asks the system "Show me Cambridge." Then, the type size of the text label Cambridge and its opacity value gradually increases resulting in a sharper focus of Cambridge (Figure 2). Notice also that related information, such as hospitals, highways around Cambridge became visually prominent, but to a lesser degree compared to the label for Cambridge. This first query exemplifies the reactive nature of discerning visually complex map display into a relatively simple and comprehensible design.

The power of using an activation spreading network to control the visual dynamics is exemplified in Figure 3, where the user requests to see crime distribution following the previous query. This shows a spatial distribution of crime data for the greater Boston area, while maintaining the Cambridge context from the previous query. Now, crime date represented by a collection of small read dots is most prominent in the display, while information related to Cambridge has become secondary but it can be still distinguished from the rest of the data. If the IS asks the relational statistics of the crime data instead of its distribution, the user can obtain a three dimensional view of crime data in the form of a bar graph as shown in Figure 4.

GeoSpace is implemented in C++ and GL graphics language on a Silicon Graphics Onyx workstation with Reality Engine graphics.

DOMAIN KNOWLEDGE

Information seeking goals and presentation plans are the basic components of this approach. A plan consists of a list of sub-plans, a list of conflicting plans, and a list of effects. The effect-list contains a set of goals that are achieved by executing the presentation plan. The sub-plan list contains a set of goals that must be achieved in order to accomplish goals in an effect-list. The conflict list contains a set of goals that are either semantically irrelevant or visually conflicting with the plan. Knowledge about semantic conflicts helps the system to identify a shift of interest. When large amount of data exist in a database, it is often the case that same visual features (such as color, typeface, orientation, or motion) are used by more than one visual element. Knowledge about visual conflicts helps the system to identify visually confusing situations.

Figure 5 shows a typical presentation plan. The plan (a) says, in order for a user to know about transportation, a user must know about bus routes, subways, and place names. The plan also indicate that hospitals and bookstores are not relevant when a user wants to know about transportation. In the current knowledge representation, semantic and visual conflicts are not distinguished. Plan (b) is much simpler; it has neither sub-plans nor conflicts. The activation level specifies the threshold energy required to realize the plan.

Plan:	**{Show_Transportation}**	
Sub-Plans:	**{Know_Place_names, Know_Bus_routes, Know_Subways}**	
Conflicts:	**{Know_Hospitals, Know_Bookstores}**	
Effects:	**{Know_Transportation}**	
Realization:	**ø**	
Activation:	**0.8**	(a)
Plan:	**{Show_Bus_map}**	
Sub-Plans:	**ø**	
Conflicts:	**ø**	
Effects:	**{Know_Transportation}**	
Realization:	**#<bus_map-object>**	
Activation:	**0.3**	(b)

Figure 5. Typical presentation plan.

The domain knowledge-base is independent of the geographic information databases. Figure 6 shows the relationship between a database and the domain knowledge-base, which can be created either manually by a cartographic designer or automatically by the system. In the current implementation of GeoSpace, the domain knowledge is encoded manually by the designer. In the future implementation, we expect to partially automate this process by developing rules that can infer relationships among information in a database. Spatial proximity relationships, information types and visual design principles will provide a criteria for creating the rules.

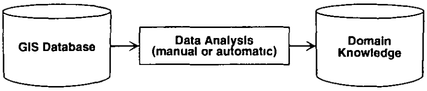


Figure 6. A process of generating the domain knowledge.

ACTIVATION SPREADING NETWORK

The system uses an activation spreading network (Anderson 1983, Maes 1990) to determine priorities of plans based on the user's request. The activation spreading network can be viewed as a graph in which plans are viewed as the vertices, and sub-goals, conflicts, and effects are viewed as the edges. A plan module's activation level is changed by the user's immediate goals, and when their activation level exceed the threshold, positive and negative activation energy is sent to other plan modules connected by hierarchical links and conflicting links respectively. The current system iteratively injects a constant amount of energy to fluidly change the overall activation state. In every iteration, activation levels of all the plan modules are normalized to the most active plan. This also results in the gradual decay of plans whose links are not explicitly specified. In addition, a presentation plan can also be activated by the dynamic changes of information it is representing. For instance, a presentation plan for a highway section can be activated by the dynamic changes of traffic information (assuming it is available).

When the user specifies a query such as "Show me transportation", *know_transportation* becomes the current information seeking goal. The system then injects activation energy to the plans that contain *know_transportation* in the effect-list. When a plan module's activation level reaches a certain threshold, it spreads energy to the plans which contain the sub-goals in their effect list. A plan also spreads activation energy upwards to the higher level plans whose effect-list contains *know_transportation* as sub-goals. This upwards activation results in activating indirectly related information. Figure 6 shows a simple example of an activation spreading process. Every iterative activation spreading

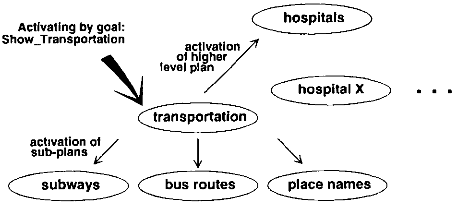


Figure 6. Schematic diagram of typical activation spreading.

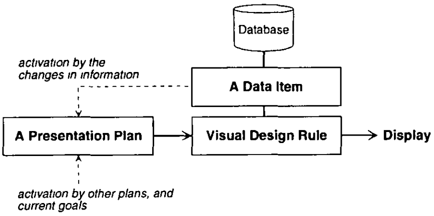

Database
activation by the changes in information
A Data Item
A Presentation Plan
Visual Design Rule
Display
activation by other plans, and current goals

Figure 7. Shows the method for selecting graphical style of data item.

change of the display. Figure 7 shows a schematic diagram of how an activation level of a presentation plan are used to determine the graphical style of a data item.

An activation spreading network not only maintains the immediately relevant information, but it can also preserve the history of a user's exploration process. When a user requests new information, the system seamlessly transforms the previous state into the new state. The network can also prepare for the user's future request by activating plan modules that are potentially relevant in the following interactions. This could greatly assist users to formulate subsequent queries towards satisfying a particular goal.

VISUAL DESIGN

The map display involves many layers of information each of which corresponds to a different set of data. The system is intended to incorporate various visual techniques, such as translucency and focus which helps clarify visual information without loosing overall context. We have incorporated these new techniques along with traditional graphical design techniques in the design of the map display. Most important information is displayed with a higher level of opacity, and related information is displayed with medium translucency. Irrelevant information is displayed almost transparent. Since, the display can show related information using relative transparency, the user has a chance of realizing a new question to ask next. Also, previously displayed information can be shown with medium to high transparency so that the user can maintain a continuos dialogue.

Plans may or may not have a graphical presentation. For example, a plan to show highways does not have a graphical representation, but each highway has a graphical representation. Those plans that have a graphical representation change their graphical style according to their activation levels. Currently, the energy levels are scaled and mapped to transparency values and/or typographic sizes on the cartographic display. The mapping from the activation levels to graphical styles is achieved by simple procedures that are implemented according to design principles. In other words, visual design knowledge is embedded in those procedures and presentation plans. Thus, the quality of visual presentation, such as legibility, readability, and clarity are significantly enhanced.

Intelligent dynamic maps guide user' attention to regions in the display that are important in a fluid manner. The mechanism described above can implicitly chain presentation plans by hierarchically spreading activation energy, and can respond to an immediate shift of interest by spreading negative energy to conflicting plans. This spreading of energy can be driven by temporal information such as weather and traffic data which makes the display truly dynamic. In such cases it is critical that users are focusing their attention at the relevant regions of the map display to comprehend the data being presented. This is accomplished by the visual techniques described previously. The result is an intelligent and highly reactive cartographic display.

CONCLUSION AND FUTURE DIRECTIONS

We have presented an intelligent and responsive map display for interactively exploring complex geographic data space. We have shown that the knowledge representation scheme which uses the activation spreading network, along with an abstraction of information seeking goals and presentation plans, provides the map display with a reactive capability. The mechanism can chain presentation plans by hierarchically spreading activation energy, and can respond to an ımmediate shift of interest by spreading negative energy to conflicting plans. The system can also maintain the context of a continuous dialogue in a fluid manner by gradually changing the states of activation. Dynamic use of various visual techniques, such as translucency, type size and color, are associated with activation levels of plans in order to visually maintain overall context during a course of information seeking dialogue.

Having completed the first generation of GeoSpace presented in this paper, we have begun to develop the next generation in order to further enhance its functionality. The following issues are currently being investigated: First, GeoSpace currently uses relatively small amount of data in order for us to carefully examine the behavior of the activation spreading network. We are in a process of increasing the size of database so as to further examine the potential of this technique. Second, in the first generation, the domain knowledge is built manually by a designer. The next generation will include a graphical interface for building domain knowledge, and a mechanism that automatically constructs the initial domain knowledge base for certain types of geographic information. Third, the activation network is relatively sensitive to the amount of activation energy spread. We are continuing to experiment with varying energy levels to find the optimal network configuration. Fourth, we are experimenting with the use of weighted links (Maes 1992) to support varying degrees of relationship among presentation plans. Finally, the second generation includes a learning mechanism which allow a particular IS to customize the domain knowledge base. Both an explicit learning and implicit learning mechanisms are being experimented.

ACKNOWLEDGEMENTS

We would like to thank Prof. Muriel Cooper, Ron MacNeil, and Dave Small for their continued support and advice at all times during the course of this project. Special thanks to Prof. Whitman Richards, Louis Weitzman, Hasitha Wickramasinghe, and Yin Yin Wong for commenting on the earlier drafts of this manuscript. We would also like to thank the other members of the MIT Media Laboratory's Visible Language Workshop for providing valuable suggestions as we developed our ideas.

This work was in part sponsored by ARPA, JNIDS, NYNEX and Alenia.However, the views and conclusions expressed here are those of the authors and do not necessarily represent that of the sponsors.

REFERENCES

Anderson, John. A Spreading Activation Theory of Memory. Journal of Verbal Learning and Verbal Behavior 22, pp.261-295, 1983.

Ahlberg, Christopher and Shneiderman, Ben. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. Proceedings of SIGCHI Human Factors in Computing Systems, 1994.

Belge, Matt, Lokuge, Ishantha and Rivers, Dave. Back to the Future: A Graphical Layering System Inspired by Transparent Paper. INTERCHI'94 Conference Companion, 1993.

Carberry, Sandra. Plan Recognition in Natural Language Dialogue, A Bradford Book, 1990.

Colby, Grace and Scholl, Laura. Transparency and Blur as Selective Cues for Complex Visual Information. International Society for Optical Engineering Proceedings, Vol 1460, 1991.

Egenhofer, Max J. Manipulating the Graphical Representation of Query Results in Geographic Information Systems, Proceedings of the IEEE Workshop on Visual Languages, 1990.

Feiner, Steven and McKeown, Kathleen. Automating the Generation of Coordinated Multimedia Explanations. In: Intelligent Multimedia Interfaces, ed. Mark T. Maybury, AAAI Press/The MIT Press, 1993.

Goldstein, Jade and Roth, Steven. Using Aggregation and Dynamic Queries for Exploring Large Data Sets. Proceedings of SIGCHI Human Factors in Computing Systems, 1994.

Maes, Pattie. Situated Agents Can Have Goals, Designing Autonomous Agents: Theory and Practice from Biology to Enginering and Back, ed. P. Maes, MIT Press/Bradford Books, 1990.

Maes, Pattie. Learning Behavior Networks from Experience, Towards a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life, ed. F.J. Varela & P. Bourgine, MIT Press/Bradford Books, 1992.

Maybury, Mark. Planning Multımedia Explanations Using Communicative Acts. In: Intelligent Multimedia Interfaces, ed. Mark T. Maybury, AAAI Press/The MIT Press, 1993.

Roth, Steven and Hefley, William. Intelligent Multimedia Presentation Systems: Research and Principles. In: Intelligent Multimedia Interfaces, edited by Mark T. Maybury, AAAI Press/The MIT Press, 1993.

Small, David, Ishizaki, Suguru., and Cooper, Muriel. Typographic Space. SIGCHI Conference Companion, 1994.

CARTOGRAPHIC SUPPORT FOR COLLABORATIVE SPATIAL DECISION-MAKING

Marc P. Armstrong
Departments of Geography and
Computer Science and Program
in Applied Mathematical and
Computational Sciences,
The University of Iowa,
Iowa City, IA 52242, U S.A
Tel. (319) 335-0153
FAX: (319) 335-2725
email: marc-armstrong@uiowa edu

Paul J. Densham
Department of Geography,
University College London,
26 Bedford Way,
London WC1H OAP, U.K
Tel. 0171-387-7050
FAX· 0171-380-7565
email. pdensham@geog.ucl.ac.uk

ABSTRACT

Collaborative spatial decision-making environments in which group members individually and collectively pursue solutions to ill-structured problems have a unique set of cartographic visualization requirements. In this paper we restrict our focus to the domain of facility location problems and describe several new map types that are designed to support the process of making comparisons among alternative scenarios. Facility frequency maps depict the stability of sites chosen in a collection of scenarios. Allocation consistency maps show the stability of allocations from demand to supply locations. Alternative forms of these maps are described and examples are presented.

Keywords: Collaborative spatial decision-making, computer supported cooperative work, cartography, group displays

1.0 INTRODUCTION

Complex public policy problems often are addressed by groups Because group members may represent a diverse set of constituencies and may come from different disciplines, they will have considerably different perspectives on the way questions should be addressed. To proceed, decision-makers also may need to integrate knowledge and data from a variety of sources For example, a residential development proposal that impacts a wetland, might arouse considerable interest among existing landholders as well as environmental advocates, including ecologists and geologists Each of these stakeholders, however, might support or attack different aspects of the proposal. When such groups are brought together specifically to address a particular problem, group members may be unfamiliar with these different perspectives and with the different decision-making styles that may be pursued To complicate matters further, we can expect different group members to have varying levels of expertise and training in the use of computers. In such cases, it is common to abandon computer-based analyses because current software is unable to support the types of computation that are required by group members when they begin to search for solutions to complex, ill-structured public policy problems Moreover, methods to resolve a divergence of views on what constitutes a good solution are not widely available.

Computer supported cooperative work (CSCW) environments have been developed to help groups of individuals work together to address ill-structured questions (Armstrong, 1994). In the geographical domain, collaborative spatial decision-making (CSDM) environments serve the same purpose. CSDM environments use CSCW principles to implement interactive, group-based spatial modeling and decision-making tools. Such environments include methods for eliciting, capturing and manipulating knowledge bases that support individual and collective development of alternative solutions to spatial problems. Other capabilities are used to manage spatial models and support the use of multicriteria decision-making methods to evaluate alternative solutions to ill-structured problems. Because of their spatial orientation, an essential characteristic of CSDM environments is the use of maps to present the geographical aspects of ill-structured problems. The purpose of this paper is to develop a conceptual framework and a set of illustrations that describe the types of cartographic displays that are required to support decision-making in CSDM environments. A series of prototypical examples from location selection problems are used to illustrate the discussion.

2.0 THE ROLE OF MAPS IN CSDM

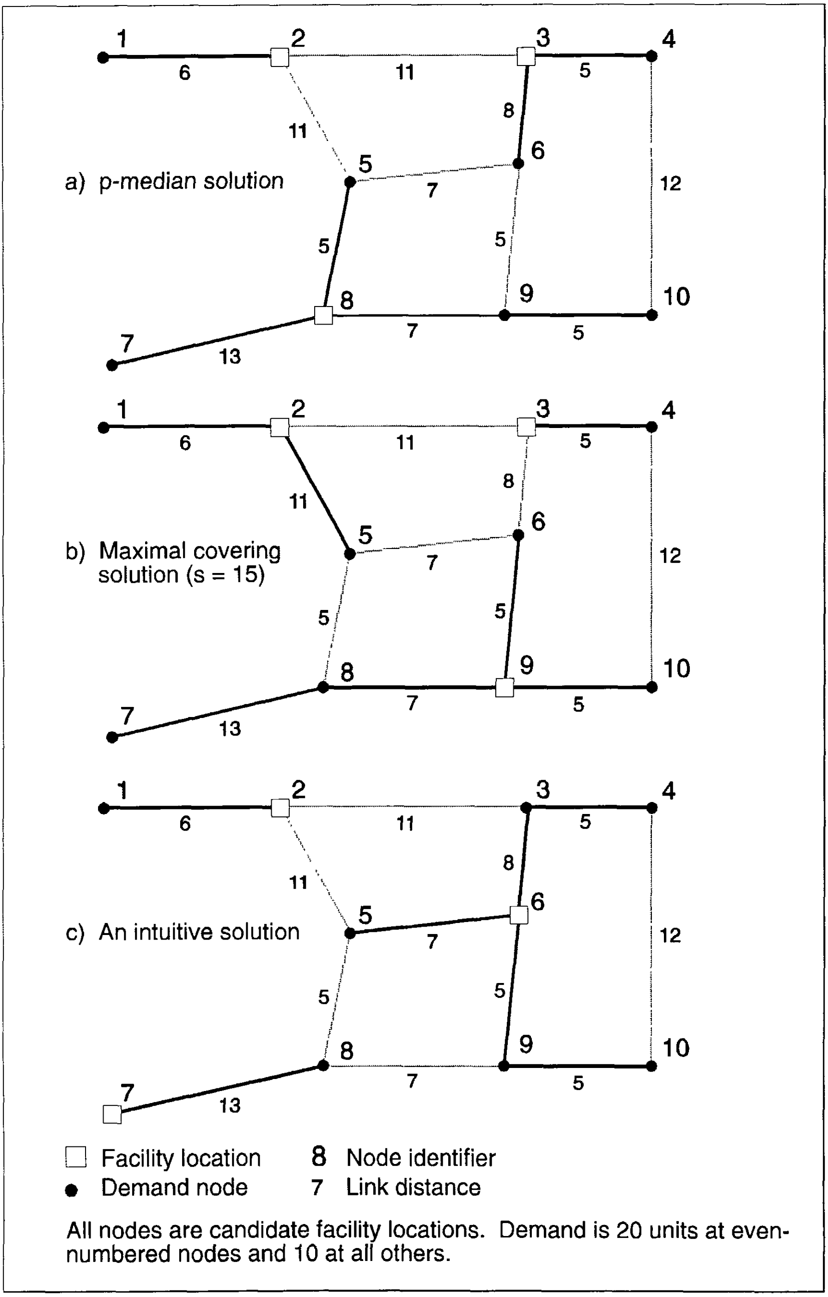
As decision-makers struggle with ill-structured problems, they often generate and evaluate a large number of possible solutions In fact, we have found that decision-makers often have an incomplete understanding of a problem and that it is only after they go through this process of exploration that they begin to gain insights into its true nature. In this process of solution generation and evaluation, decision-makers discover relationships among the various components of a problem and see how criteria may conflict. Given the complexity of supporting even a single user in this iterative process, it is clear that supporting a group of decision-makers introduces additional levels of complexity (Armstrong, 1993) The number of ways in which alternative solutions are evaluated and compared, for example, may increase greatly in a group context because each person may have a different level of training and experience as well as a different perspective on the problem-solving process to be followed

In CSDM environments, maps serve as a basic token of exchange among group members. Although they communicate the form and structure of a scenario, maps are the metaphorical tip-of-the-iceberg In a location selection context, for example, maps are constructed from the contents of a set of data structures that support locational analysis operations (Densham and Armstrong, 1994). The contents of these data structures are themselves derived from the system's database and users' knowledge Although maps depict many characteristics of a scenario in an accessible form, a map is not a scenario. to explore a scenario, and possibly to modify it, users require access to its underlying data. Consequently, while maps are used as tokens, and may be thought of as scenarios by users, to the system designer a scenario consists of a set of user inputs and analytical operations, the contents of the analytical data structures and their linkages back to the database, and scenario displays

Several types of maps that have been designed to support locational decision-making (Armstrong *et al.*, 1992) can serve as tokens of exchange.

- *Location maps* supply basic information about the geographical context of a decision; they may include, for example, highway maps and political maps.

- *Supply maps* show a set of facility locations.



a) p-median solution
b) Maximal covering solution (s = 15)
c) An intuitive solution
1
2
3
4
5
6
7
8
9
10
6
11
5
8
11
7
12
5
5
7
5
13
□ Facility location
● Demand node
8 Node identifier
7 Link distance
All nodes are candidate facility locations. Demand is 20 units at even-numbered nodes and 10 at all others.

Figure 1. Network-based spider maps of three solutions to a location problem

- *Demand maps* depict the geographical distribution of demand for services.
- *Spider maps* show linkages between supply and demand, either for an existing scenario or for one of change - see Figure 1.
- *Delta maps* depict differences between a pair of scenarios, one of which may represent the existing system.

When these map types are used to explore ill-defined location selection problems, they are often used privately, by an individual, during the process of generating a scenario. Supply maps, for example, may enable a decision-maker to identify an underserved area; this problem would then be addressed in subsequent scenarios, leading to the generation of further maps When assembled into a collection, these maps are used to convey the form and substance of a scenario-based decision process to other members of the group. Consequently, for each of the tasks that must be completed at each step during problem-solving, CSDM environments must provide users with appropriate types of maps.

Although the five types of maps described above are well-suited to individual use, the need to compare and evaluate scenarios during collaborative decision-making has generated a new set of display requirements. Thus, an additional set of displays are required to convey the degree of similarity among a set of scenarios· these *summary* displays are specialized forms of delta maps. While considerable amounts of work are required to refine the tools needed to create individual cartographic displays used by individuals, in this paper we will focus on these summary displays.

3.0 CREATING GROUP MAPS

In the context of group locational decision-making, we have identified two new summary map types: *facility frequency* maps and *allocation consistency* maps. Given two or more scenarios that are to be compared, a facility frequency map communicates two important dimensions of the robustness of facility locations - the number of times a given candidate location is selected as a facility site under differing criteria, and the volume of demand that it serves. In contrast, an allocation consistency map depicts the linkages among supply and demand points across two or more scenarios Several dimensions can be displayed on different forms of an allocation consistency map, including the volume of demand served by each facility, the demand that is allocated to the same facility under a range of criteria, and the areal extents of facility service areas.

A map can be thought of as an organized collection of geographical features that are represented using symbols. Different hierarchical levels of organization of these primitive elements are used to communicate information to the map reader For example, symbols can be manipulated to define regions or to ensure that data are perceived at a higher visual level than the base material. These same primitive map elements also can be used to support the development of summary maps that define the geographical dimensions of the degree of similarity that exists among alternative solutions to problems. If maps are viewed in this way, the level of decomposition determines the methods that can be used to make comparisons among different maps. When digital maps are disassembled into collections of constituent primitive elements, these cartographic components can be manipulated to determine similarity among alternatives simply by performing basic arithmetic operations on them. This "map display algebra" can be made operational in a simple accounting framework. For example, the number of linkages between demand and supply locations that are held in common across a set of alternative solutions can be summed to generate a spider map that shows the degree of similarity in the allocation of demand to supply locations In general, we can focus on various aspects of a scenario, enumerate these components and symbolize them to depict similarity using a standard set of visual variables, including symbol size and intensity.

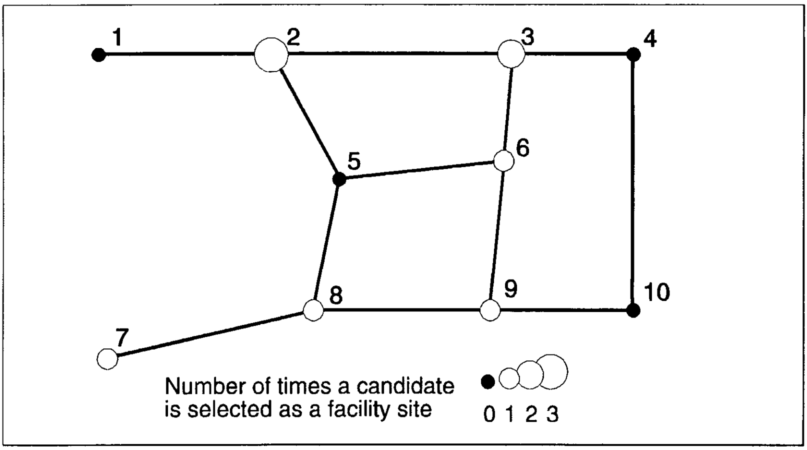


Figure 2. A facility frequency map for scenarios a, b and c (Figure 1).

3.1 Facility Frequency Maps

A facility frequency map indicates the degree of similarity among the facility locations in different scenarios. Decision-makers often wish to change policy variables as they explore an ill-structured problem, including the number of facilities to be located and the maximum distance that is traveled to a facility. In each scenario, therefore, a different set of facility locations might be selected from the set of candidate locations. As the number of scenarios grows, the set of choices can become confusing. To facilitate fruitful discussions among group members about the merits of alternative solutions, a parsimonious means to synthesize alternatives is required We have developed two generic map types for comparing and evaluating scenarios

The first map type focuses on each candidate location and the number of times it is selected as a facility location. Using size to symbolize this value leads to the creation of a typical graduated symbol map (Figure 2). While monochrome versions of this map can be included in laser-printed reports, color could be used to reinforce the size variable or indicate a second dimension to the solution, such as the amount of service provided. The use of classed or continuous (classless) color depends primarily on the number of scenarios to be compared. Because a facility may be located at the same candidate location in every scenario, there must be at least as many classes as there are scenarios to be compared. Consequently, classed maps might be used when the number of scenarios is small and continuous color when the number is large. The point at which one type is chosen over the other depends on, amongst others, the color range of the display device and the preferences of the user An alternative approach to symbolizing this aspect of the collaborative process is best used when a small number of scenarios is being considered. This might occur when a group has narrowed its focus to a small set of choices, for example, and individual scenarios must be identified. Each scenario is assigned a color and bars with the requisite colored segments are built at each facility location. A variation on this approach can be used when the number of individuals participating in the decision is small (fewer than 10) and inter-personal differences are to be highlighted. Each individual is assigned a color that is used to identity the scenario that they wish the group to consider. If each individual is told only which color has been assigned to them, such displays facilitate a form of group Delphi process If the full range of color assignments is made available, individuals could determine where points of difference arise with some stakeholders, and possibly form coalitions with others.

3.2 Allocation Consistency Maps

An allocation consistency map shows which allocations of demand to facilities recur in two or more scenarios. The resulting map is similar to a delta map and, when only two scenarios are to be contrasted, two variants are possible In the first case, the actual allocations of demand to supply locations are suppressed so that emphasis is placed on the demand locations. The map is created simply by coloring green (white) everything that goes to the same facility, and all the demand nodes that go to different facilities in red (black). Unless at least one facility location is common to both scenarios, however, it will consist entirely of red nodes. Although an allocation consistency map highlights those areas that are most affected by the differences in facility locations, it may not provide any information about the service areas of different facilities Consequently, the second type of map provides information about both service areas and the consistency of demand allocations. One scenario is chosen as the "base scenario" and its allocations are depicted using a spider map. Each facility and its allocated demand nodes in the "alternative scenario" are represented by a colored symbol, to differentiate demand nodes and facilities, the latter are assigned larger symbols Since the base scenario may be given an inordinate amount of "weight", it may be wise to reverse the ordering. One possibility is to display the two maps side-by-side (Figure 4), a second is to "overlay" the two maps and fade continuously between them - providing a dynamic display.

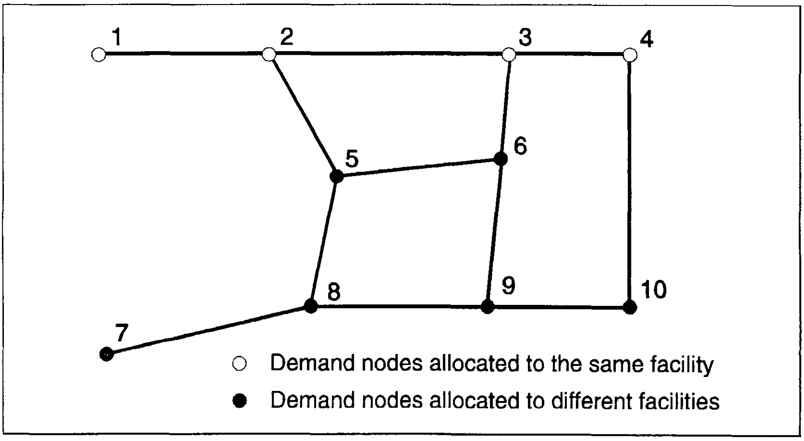

1
2
3
4
5
6
7
8
9
10
○ Demand nodes allocated to the same facility
● Demand nodes allocated to different facilities

Figure 3. Allocation consistency map comparing scenarios a and b in Figure 1.

It is important to note that allocation consistency maps quickly become complex if dissimilar scenarios are compared, this occurs because the different allocations often cross one another. Consequently, while all allocations can be symbolized using an accumulator logic - similar to that used in facility frequency maps - a better approach is to symbolize only salient allocations, such as those that occur in all, or a large number of, scenarios. Where allocations are depicted as routes through a network, for example, attributes also can be symbolized on this type of map. When coupled with allocation volumes, these maps might help decision-makers identify problems caused by traffic congestion or negative externalities, including levels of noise and particulate emissions.

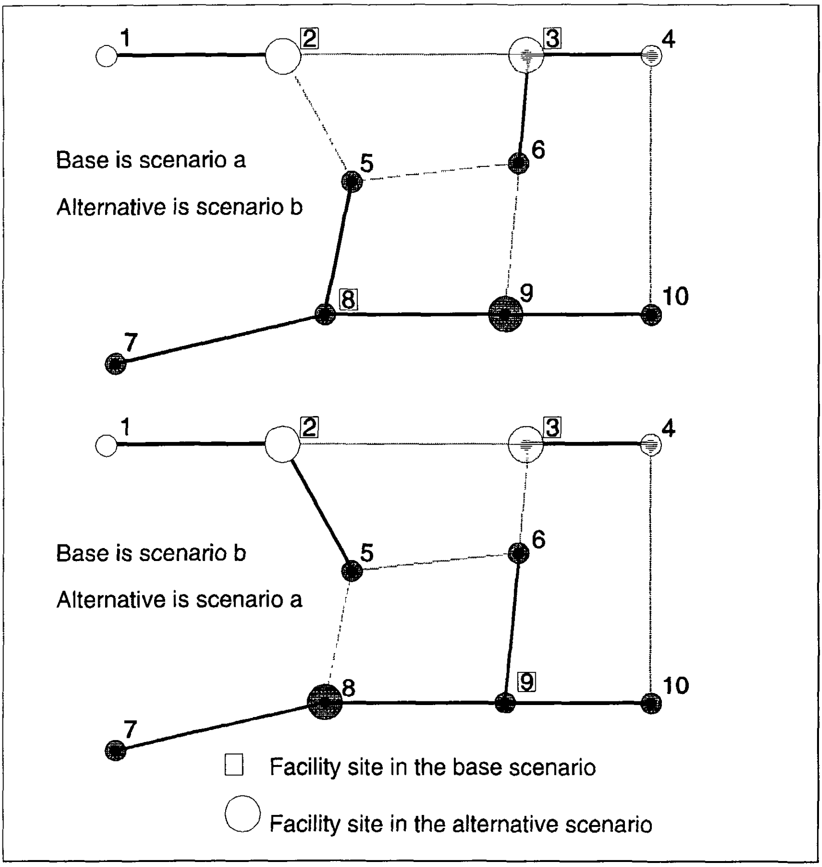

1
2
3
4
Base is scenario a
Alternative is scenario b
5
6
7
8
9
10
1
2
3
4
Base is scenario b
Alternative is scenario a
5
6
7
8
9
10
□ Facility site in the base scenario
○ Facility site in the alternative scenario

Figure 4 Allocation consistency map for scenarios a and b (Figure 1).

A further pair of map types are used when three or more scenarios must be compared. The first uses graduated symbols to depict the number of different facilities to which a demand node is allocated. Such a map can be produced using either monochrome or color symbols to enhance its effectiveness. The second type of map uses a network-based spider map to represent the impact on the transportation system of the selected scenarios (Figure 5). This display depicts the number of times that a transport link is traversed in the scenarios being compared Maps of this type are useful for identifying potential choke points in the transportation network. This is particularly important if the activity being examined can alter the existing patterns of movement of large numbers of people. For example, the bridges linking Manhattan to other parts of New York would have high values in such displays.

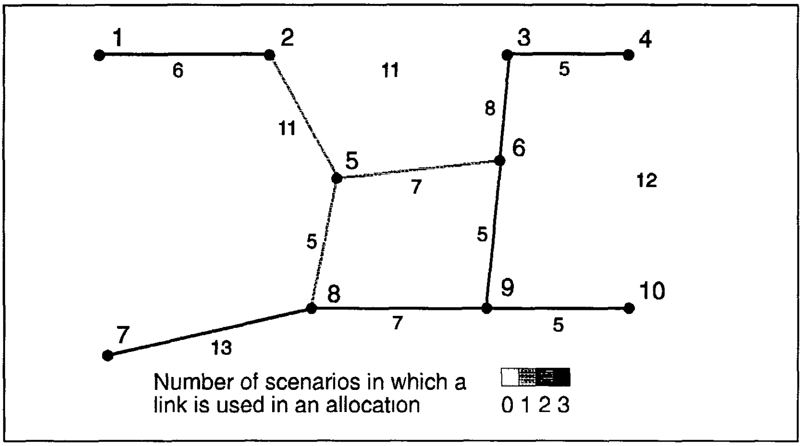

1
2
3
4
6
11
5
8
11
6
5
7
12
5
5
9
10
8
7
7
5
13
Number of scenarios in which a link is used in an allocation
0 1 2 3

Figure 5. An allocation consistency map showing the number of times a link is used in allocations for scenarios a, b and c in Figure 1.

4.0 OTHER TYPES OF CARTOGRAPHIC SUPPORT

When an exploratory and iterative group problem-solving style is employed, it is often useful to be able to track and, if necessary, re-create the process of exploration Two capabilities are useful in this regard. First, users need scenario management tools to trace the decision-making process. The sequence of solutions to a problem might be generated and captured so that it could be replayed as a slide show under user control. In this way, it would be possible to understand where points of divergence or convergence among group members occurred Maps also must be supplemented with lineage information and meta-data to support this capability.

Lineage data tracks the process of decision-making. The capture and maintenance of lineage data can be achieved using a collection of data structures that supplement conventional cartographic data structures. In an object-based representation, each map would have several lineage objects to define the name and creator of the mapped scenario, the identity of the person who last modified it, and the date and time of these activities. In addition, space would be made available for comments - if multimedia capabilities are supported, these comments could be spoken rather than typed.

Structure: Group_lineage
Slot: Name_of_scenario
Slot: Name_of_creator
Slot: Creation_time_and_date_stamp
Slot: Name_of_last_modifier
Slot: Modification_time_and_date_stamp
Slot: Comment

Scale of analysis and display is an important characteristic of complex locational problems Because such problems often contain aspects that are straightforward, it is only in selected locations that potential conflicts occur. In such cases, system users normally wish to examine the different dimensions of a problem in greater detail. Consequently, the pan and zoom capabilities of the system assume a prominent role in user needs. Furthermore, this may necessitate real-time suppression or uncovering of map features, and possibly their generalization

The use of map decomposition strategies also can provide insights into the problem solving process and the design of user interfaces (Armstrong and Densham, 1994). In visual interactive modeling (VIM; Hurrion, 1986, Densham, 1994), it is possible to enumerate the number of times that display and analytical objects are manipulated. Areas that are not manipulated can be treated as stable· there is a high-degree of similarity among alternatives (both within an individual's scenarios and across several individuals' scenarios) By summing the number of times objects are moved, it is possible to determine which areas are investigated by users. Users may be focusing their attention on these areas because they are poorly-served in a series of scenarios, for example; whatever the cause, such areas are potential sources of disagreement To resolve situations of this kind, users typically apply judgment that is exogenous to the system, including personal, *ad hoc* knowledge about the problem.

To meet the needs of a disparate range of users –some group members may have formal cartographic training while others have none– the system must provide a set of cartographic design and production tools that accommodates the needs of novices and experts alike. Thus, not only must the system provide a rigorous set of defaults for use by novices and a much less constraining set of tools for experts, it also must integrate these tools with the system's analytical capabilities (Densham and Armstrong, 1994). For example, iconic map types similar to chart types in Microsoft Excel could be used to structure the map creation process for novice users. In addition to these multiple levels of access and support, the system also can provide users with a customizable map-production environment, rather like the use of style-sheets in a word processor– the use of "themes" in GisPlus (Caliper Corporation, 1992), for example, goes some way towards meeting this need Thus, users can apply their favorite style of colors and symbolism to information, even to maps produced by other group members. While individuals create displays using standard cartographic tools, group displays will be manipulated using a set of specialized group tools. Such tools include generic whiteboard highlighting tools (lasso and pointer), as well as editing tools (scissors), that enable users to capture and edit privately scenarios that have been submitted to the group.

5.0 CONCLUSIONS

The support of interactive, group decision-making processes requires the development of new kinds of cartographic displays. We have developed two new kinds of summary display that are used to compare scenarios. facility frequency maps and allocation consistency maps By synthesizing the characteristics of two more scenarios in a single display, the various forms of these maps enable decision-makers to understand the differences that separate scenarios. As with other forms of displays developed for use in locational problem solving (McKinney, 1992), further research is required to evaluate the utility of both virtual and hard-copy versions of the various forms of these maps to decision-makers.

ACKNOWLEDGMENTS

We acknowledge support for this paper from the National Center for Geographic Information and Analysis (NCGIA). This paper is a contribution to Initiative 17 (Collaborative Spatial Decision-Making) of the NCGIA which is supported by a grant from the National Science Foundation SBR-88-10917.

REFERENCES

Armstrong, M P , P.J. Densham, P. Lolonis, and G. Rushton 1992, Cartographic displays to support locational decision-making *Cartography and Geographic Information Systems*, 19(3), pp. 154-164

Armstrong, M.P. 1993, Perspectives on the development of group decision support systems for locational problem-solving: *Geographical Systems*, 1(1), pp. 69-81.

Armstrong, M.P. 1994, Requirements for the development of GIS-based group decision support systems. *Journal of the American Society for Information Science*, 45(9), pp. 669-677.

Armstrong, M.P. and P.J. Densham 1994, A conceptual framework for improving human computer interaction in locational decision-making. Forthcoming in T.L. Nyerges (editor), *Cognitive Aspects of Human-Computer Interaction for Geographical Information Systems*, Dordrecht Kluwer Academic Publishers.

Caliper Corporation 1992, *GisPlus User's Manual, Version 2 1*, Newton, MA: Caliper Corporation.

Densham, P J 1994, Integrating GIS and spatial modelling. visual interactive modelling and location selection· *Geographical Systems*, 1(3), pp. 203-219.

Densham, P J. and Armstrong, M P. 1993, Supporting visual interactive locational analysis using multiple abstracted topological structures: *Proceedings*, Eleventh International Symposium on Computer-Assisted Cartography (Auto-Carto 11), Bethesda, MD: American Congress on Surveying and Mapping, pp. 12-22

Densham, P J. and Armstrong, M P. 1994, Human-computer interaction aspects of visual-interactive locational analysis Forthcoming in T.L. Nyerges (editor), *Cognitive Aspects o, Human-Computer Interaction for Geographical Information Systems*, Dordrecht: Kluwer Academic Publishers.

Hurrion, R.D. 1986, Visual interactive modelling: *European Journal of Operational Research*, **23**: 281-287.

McKinney, J.V. 1991, *Cartographic Representation of Solutions to Location-Allocation Models*, Unpublished Master's Project, Department of Geography, State University of New York at Buffalo, Buffalo.

MEASUREMENT, CHARACTERIZATION AND CLASSIFICATION FOR AUTOMATED LINE FEATURE GENERALIZATION

Corinne Plazanet

IGN - Service de la recherche - Laboratoire COGIT
2, Avenue Pasteur (B.P. 68) 94160 St Mandé Cedex FRANCE

plazanet@cogit.ign.fr

ABSTRACT

This paper proposes to describe geometrical characteristics of linear features by segmenting lines and qualifying sections detected according to different criteria at several levels ("hierarchical" process). Such a kind of line representation should provide basic and complementary information in order to guide global and local decisions. The ultimate aim is to choose and sequence adequate generalization tools and parameter values.

KEYWORDS: Automated linear feature generalization, segmentation, characteristic point detection, measurement, clustering

INTRODUCTION

Generalization is often described as an holistic process, due to the large number of constraints and interactions that must be taken into account. This paper addresses issues related to linear feature intrinsic generalization. At first, the main issue for such a generalization is that every linear feature seems to be a particular case (at least for a significant level of generalization). This may come out from the fact that a lot of complex, and difficult to formalize, cartographic constraints intervene in the generalization process, and *shapes at different levels of analysis* must be considered and processed (omitted vs retained, simplified vs enhanced ...). Therefore, an identification and analysis of these shapes, prior to the application of generalization operations, as well as a study of the relationships between shape configurations and generalization operations and tools, are crucial for proceeding towards more effective generalization solutions.

LINEAR FEATURE GENERALIZATION AND CARTOGRAPHIC CONSTRAINTS

A cartographer, when generalizing manually, has a global and continuous feeling of each line [Plazanet et al, 94]. He/she reacts as a human being who fully uses his/her artistic and aesthetic judgement and his/her experience. Most of the time, he applies cartographic knowledge that has not been formalized in rules form anywhere. For instance in mountainous areas, many roads need to be enhanced. To eliminate a bend from a series of laces on a given road, when a spatial conflict with a river or other feature occurs, French cartographers generally proceed by maintaining the first and last bends, choosing *the* bend to eliminate (often the smallest) and emphasizing the others.

Looking at the cartographers' own way to proceed should hopefully provide basic knowledge and should help to express and formalize cartographic constraints and rules for correct automated generalization.

What kind of constraints do cartographers take into account when generalizing ? What rules guide manual generalization ? And further *Can we produce objective rules in order to obtain consistent generalization ?* [McMaster, 89] So far, some cartographic rules which guide intrinsic linear feature generalization have been identified such as:

Respect geometrical shapes. Within a road section containing a series of hairpin bends for instance, at least one hairpin should be retained, even at small scales. More generally, it seems important to take into account and preserve as well as possible the geometrical characteristics of bends such as size, amplitude, type of curvature, bend main direction, etc, and to maintain the main directions of linear features.

Preserve intrinsic topology.

Take into account symbol width. Linear feature symbolization is an important factor for decisions and strength generalization operations. The larger the sign, the faster geometrical details on shapes are lost due to line intrinsic conflicts and the legibility limits on the map.

Take into account semantical nature of the features. French cartographers distinguish 2 types of linear lay-out: straight ones observing *soft sinuosity* (rail-roads, freeways), and sinuous ones (roads, foot-paths, rivers). The semantical nature of the features contains a first indication on geometrical properties. Not only this indicates the type of geometry which may be expected from semantics, but generalization results have to respect these relationships.

Focus on the message. Each geographical object contains a particular message in accordance with its semantical nature, the map resolution and the cartographic theme. When producing a road map for instance, roads intrinsic value becomes greater. If a globally straight road has a compact bend, it should be kept and emphasized for user's information purposes such as for instance: "be careful, the road at this point turns abnormally".

Shape importance has to be considered in each case according to its linear environnment. Thus the ϕ shape (see Fig. 1) has to be maintained in the second case, while in the first case it looks like another one and then should be smoothed.

Scale and shape environment notions are essential, introducing some kind of subjectivity into local generalization decisions.

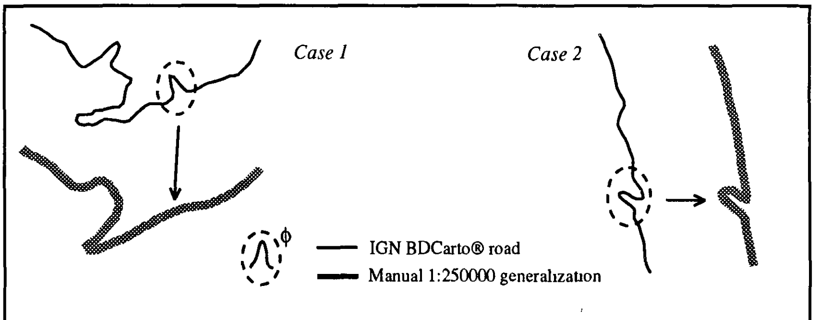


Fig. 1: The shape ϕ is the same, but in some case the generalization decision won't be the same

AUTOMATED GENERALIZATION

A lot of studies dealing with linear generalization have shown the lacks and limits of generalization algorithms, especially for simplification [Beard, 91] [Herbert et al, 92] [Jenks, 89] [McMaster, 89]. The quality of the result seems to vary according to the line characteristics [Buttenfield, 91]. These studies have clearly revealed the need to describe and qualify linear features. Checking on the significant geometric features of the line will hopefully guide the choice of tools and parameters.

More precisely the word "description" raises some questions such as:

- *What kind of information is important in order to generalize a line (Should we describe a line that we want to generalize as we perceive it visually ?) ?*
- *How to extract useful information from a set of points ?*
- *How to organize the extracted information ?*
- *How to take cartographic constraints into account ?*
- *How to take good decisions according to target scale, cartographic constraints, map theme and line description ?*

The automated intrinsic linear generalization process (first including simplification and enhancement) may be schematically divided into 4 tasks:

A - Linear features description

B - Analysis of the description

C - Decision of the generalization process according to analysis

D - Assessment of the generalization according to constraints and targeted scale

The present paper deals with tasks A and B proposing a way to build up a line feature description wich preceeds an approach of the generalization stage proper. The first part deals with linear feature description while the second addresses technical methods for description. The last section proposes first experimental results before we conclude and introduce next steps of the research.

LINEAR FEATURE DESCRIPTION

PRELIMINARY

When asked to describe a line, anybody will use the words *straight*, *sinuous*, *regular* (meaning homogeneous in some specific sense), etc. It is worth noting that such appreciations may be based on a more or less global (or local) inspection of the line. If the line is not regular, local bends are often described, for instance *"this straight road with 2-3 sharp U bends one mile away from the fork"*.

What are the criteria accounting for the global qualification of a line ? Is there nothing more than sinuosity vs. straightness ? According to Mc Master [McMaster, 93]: *"Individuals seem to judge the shape of the line on two criteria: the directionality of the line and the basic sinuosity of the line"*. This observation induces him to use the sinuosity / straightness criterion in order to know where to cut the line.

Sinuosity may be qualified according to the semantical type of objects. For instance a river trace is quite often irregular, rough, while a road may follow "zigzags", hairpins, closed bends. Road features are human constructions, designed from regular mathematical curves. Modern roads are often composed of straight lines, circle arcs and clothoids [Affholder, 93].

One will instinctively cut the line of Fig. 2 into segments which look homogeneous. From the left runs first a hairpin asymmetric bend series followed by a more or less straight segment and a sinusoid series while the right part seems unhomogeneously softly sinuous. The sinuosity criterion itself is built up of several other criteria. In the meandering part on Fig. 2 the amplitude is much larger on the right. Bends in a homogeneous section can be described according to the kind of curvature (spirals, sinusoids...), and to amplitude and symmetry, etc. Local important details may also be noteworthy.

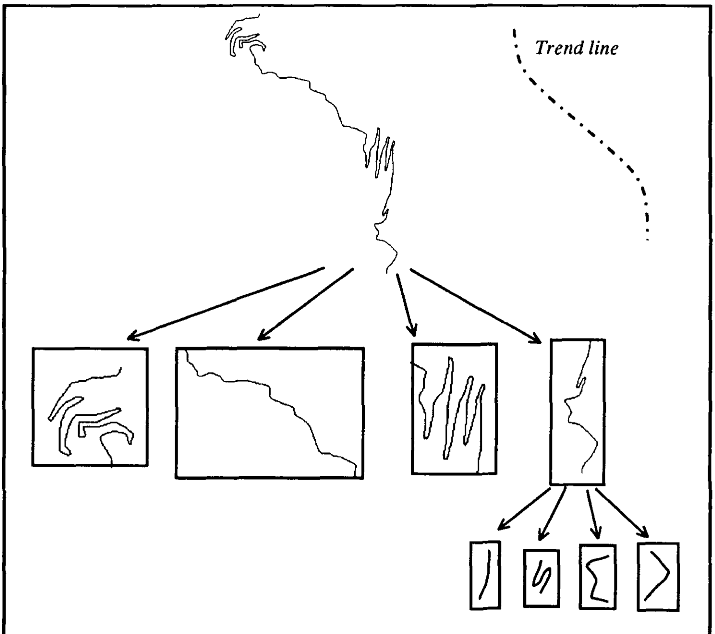


Fig. 2 · An example of natural line segmentation (IGN BDCarto® road)

Our assumption is that, in fact, different levels of perception will consider different shapes of a line. Then the trend line can be seen as a first level of sinuosity (that might be called the global level). For instance on Fig. 2, it can be seen that the trend line has a convex shape. Clearly, the analysis of the trend line could provide us with additional indication of the line complexity.

SEGMENTATION AND ANALYSIS AT EACH LEVEL

At each level of perception, the line or a section of the line is segmented and analysed. If the line or section is unhomogeneous in some specific sense, it has to be segmented again.

A low level analysis is required to describe linear features at different levels. It consists in classifiying the considered line section and judging if it is homogenous or not, using some measures (see below in the "methods" section).

Further, besides a line description at several levels, a deepest analysis is required to extract complementary informations of prior importance for the generalization process itself such as:

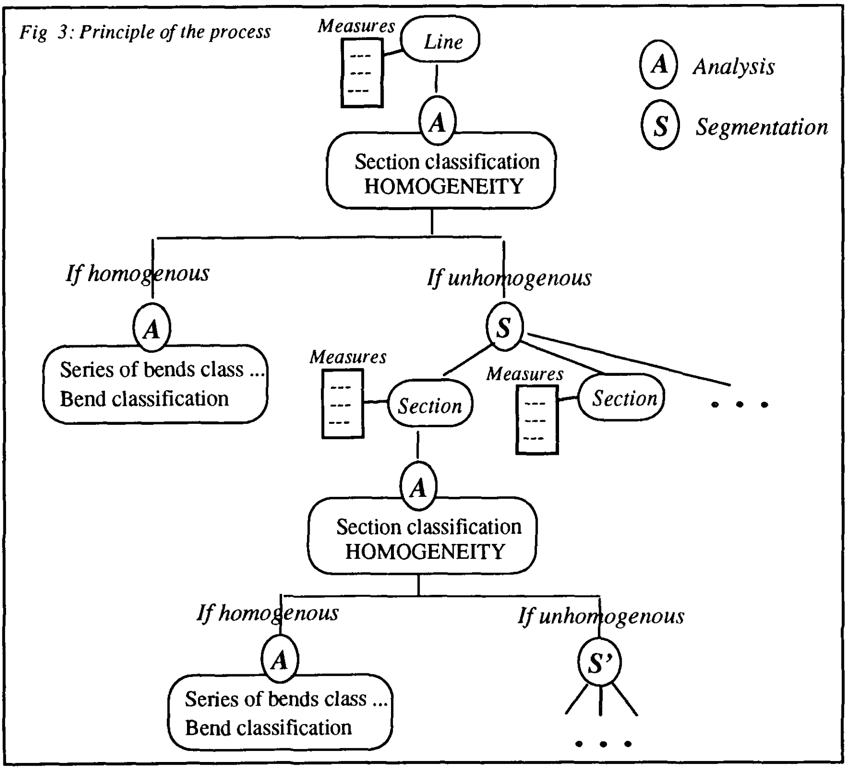
- shape levels (relative levels of shapes within shapes),
- shape environment (relative positions of shapes within shapes),
- shapes or bends repetitions,
- intrinsic conflicts areas of the line.

ORGANIZATION OF THE PROCESS

Our approach is a hierarchical segmentation process analogous to *Ballard strip trees* and to the method proposed in [Buttenfield, 91] where series of measures are computed for each line. B. Buttenfield in [Buttenfield, 87] already proposed a very similar process for classification of cartographic lines also dealing with segmentation and measurement. Starting from the observation that *"a cartographic line is composed of a trend line, and features that bifurcate from it"*, she isolated several levels of details and elaborated measurements on the first level including the number of levels.

Principle of the process

Fig 3: Principle of the process



First the line to be described is analysed using a first set of measures. If it is homogeneous, series of bends are qualified and classified if possible. If it is unhomogeneous, it has to be segmented. Then each section may in turn enter in the process recursively. (See Fig. 3)

Leaves of the tree

The first occurence of "hierarchical" description corresponds to a segmented section of the line with attributes and possibly a shape class code. The set of attributs is not always the same for every level of the tree; Most likely measures for segmentation won't be the same as for characterization.

At the finest analysis level, it is important to know two things about each of the final homogeneous sections: What are their geometrical characteristics ? And how do they fit into the global shape of the line ? The local analysis (between two successive inflection points) of curvature, amplitude and symmetry will decide the characterization of each section.

The lowest node may be a single bend defined as a constant sign curvature section delimited by two inflection points. It can be qualified by its width (large to small), amplitude (tight to open), bend direction (straigth to convex), symmetry and type of curvature (sinusoid, rectangle, spiral).

Potential outcomes

Together segmentation, analysis and classification will hopefully be useful for taking local decisions in order to generalize according to cartographic rules. For instance, for a particular bend or series of bends delimited by 2 inflection points, if the distance between these points is smaller than symbolization width, then a conflict area is detected. Or if a section of hairpin bends inside a straight section from the higher level node is considered as an accident then it should be preserved. If the two rules are verified then the compact series of bends has to be amplified. Then, when the generalization operation is chosen, simplification and enhancement algorithms corresponding to the lowest level class section are applied. Other potential outcome may be the quality assessment in terms of shape maintenance, still badly missing so far.

METHODS FOR SEGMENTATION AND ANALYSIS

CHARACTERISTIC POINTS DETECTED FOR SEGMENTATION AND ANALYSIS

The detection of the characteristic points is a fundamental operation for the process. Characteristic points are useful both for segmentation and analysis operations:

- Hoffman and Richards [Hoffman et al, 82] suggest that curves should be cut at the minima of curvature, which are particular characteristic points.
- The line geometry description is based on the characteristic points detection: Attneave in [Attneave, 54] has proved that "*information is further concentrated at points where a contour changes direction most rapidly*" i.e. vertices. As a rule, in literature, be it psychology or computer vision or more recently cartography, shape detection is based on the detection of characteristic points [Hoffman et al, 82] [Thapa, 89] [Muller et al, 93].

In Euclidean geometry, characteristic points are often confined to curvature extrema: Vertices and inflection points. As Freeman said, this definition is too restrictive: "*We shall expand the concept of critical points to include also discontinuities in curvature, end points, intersections (junctions) and points of tangency*" [Freeman, 77]. According to our needs, we define the following points as characteristic points:

- discontinuities in curvature,
- start or end points,
- maxima of curvature i.e. vertices,
- minima of curvature i.e. inflection points,
- critical points (chosen among of inflection points), which demarcate 2 line sections after a segmentation stage.

Detection of inflection points

The first step consists in detecting the inflection points. Because of the acquisition process, lines frequently contain spurious micro-inflections which are not characteristic. Moreover, according to the analysis level, only the main inflection points are of interest. A process has been implemented and tested at the IGN COGIT laboratory to detect characteristic inflection points [Affholder, 93]:

A smoothing using the convolution product of points with a Gaussian filter will delete details which are not significant for a given analysis level. According to McMaster, *smoothing routines relocate or shift coordinate pairs in an attempt to "plane" away small perturbations and capture only the more significant* [McMaster, 89]. Studying the vectorial product variation in each point of the smoothed line allows for the detection of characteristic inflection points: At these points, there is a significant change in the sign of the vectorial product. The smoothing may be more or less heavily applied, depending on the analysis level. The more global the analysis, the strongest the smoothing.

Detection of vertices

We approximate the actual vertex by computing the point which stands at the greatest perpendicular distance from the segment joiging the two consecutive inflection points (equivalent to the Douglas routine computation). We may go further to compute secondary pseudo vertices of a particular bend (running a more specific Douglas routine).

MEASUREMENTS FOR SEGMENTATION AND ANALYSIS

For a bend (Fig. 4), we may compute:

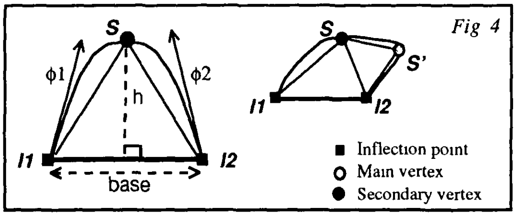


Fig 4

- the height *h* or surface of the triangle (S,I1,I2)
- the euclidean distance between the inflection points, *base*
- the curve length between inflection points I1 and I2, *l*
- the ratio between the curve length from I1 to I2 and the Euclidean distance, *l/base*
- the angle between I1 and I2
- the number of secondary vertices per bend
- the area of the main triangle (S,I1,I2) and secondary triangles (S,S',I2)

The number of secondary vertices or the sum of the area of triangles for instance gives an idea of the type of curvature fo the bend, while the variance of the curve length *l* between the inflection points indicates the regularity of line.

For a section, we may compute the mean (or median) value, variance, minimum and maximum values of each of these measures. Some more measures based on the line joining inflection points seem to be interesting:

- the total number of bends
- the total absolute angles of the inflection points line
- the ratio of the curve length of the line joining inflection points and the curve length of the original line

For instance the variance of distances between inflection points gives an appreciation of some kind of regularity of the line. The inflection points line in a sense may be seen as an approximation of the trend line.

Many of such measures may be easily computed. Also we need to choose and normalize a subset of significant and non-correlated measures that will account for each particular operation of segmentation and analysis of any line.

MEASUREMENTS FOR A FIRST LEVEL SEGMENTATION

With a heavy smoothing, it can be hoped that the strongest and most meaningful inflection points will be detected, and from them the critical points for this first analysis level.

The obvious fact that a very sinuous line has many inflection points in close succession (which is not the case for a straight line) will help us to divide lines into sinuous / straight segments. Looking at the curve representing the Euclidean distances between successive inflection points along the line, one can deduce sections on the line where these distances varie gently.

A FIRST EXPERIMENT OF LINE FEATURE DESCRIPTION

The description process has been attempted on a set of lines taken from the BDCarto® IGN data base, based on the measurements described above and using a classical cluster analysis software. The first objective is to split lines into straight / sinuous / strongly sinuous sections. The different parts of the experimentation are:

1 - INFLECTION POINTS AND VERTICES DETECTION

Hereafter is presented an example of inflection points detection on a 5 m resolution road, Fig. 5. The smoothing is rather strong as the neighbourhood involves a fifth of the total number of points ($\sigma = 40$).

2 - PARTITIONING INTO STRAIGHT / SINUOUS SECTIONS

Once the inflection points are detected as shown on Fig. 5, the line is cut into straight / sinuous parts and the critical points are selected.

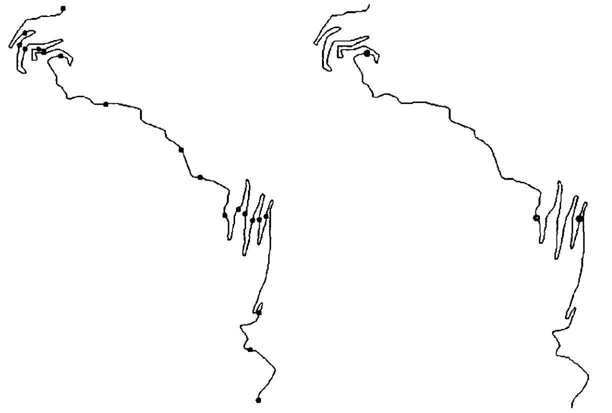


Fig 5 : Inflection points detected on the left side
Critical points retained on the right

These three critical points determine the four sections observed. Through closer analysis, the line will be cut recursively and even the sinuous sections, according to their characteristics: In the first sinuous section (top left), the bends are as sharp and close to each other as in the second sinuous section, but they are not symmetric.

3 - CLUSTERING HOMOGENEOUS SECTIONS

For this first experiment, chosen measures from the described set of measures are: the median ratio between the curve length *l* from I1 to I2 and the Euclidean distance *base* and median ratio between the *h* distance and the Euclidean distance *base*. A set of 40 lines have been classified using a S-PLUS cluster analysis package (See Fig. 6).

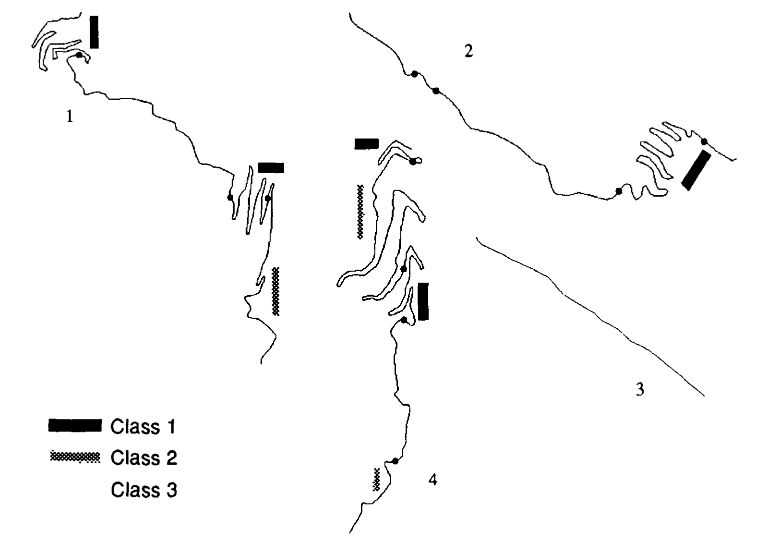


Fig. 6: An example of classification results

DISCUSSION

The results of the segmentation operation seem quite promising. In order to go further, the segmentation process has to be improved by adding complementary information especially about bend amplitudes. As to characterization and classification, the current results show the need for a better understanding of the measures and their interrelations. An interesting approach would consist in devising a large set of measures, and then trying to reduce this set to smaller sets of significant and non-correlated measures for the different purposes of our description tasks. The question which arises then is wether we will be able or not to choose among th initial measures and normalize the initial ones.

The aim of this study is to provide a method for linear shapes identification before generalization in order to choose adequate and effective generalization solutions. Other future work will focus first on the study of the effects of generalizing operations on line shapes, and on the selection of the best representation (cubic curve arcs for instance) for a given homogeneous section. Such kind of work could probably become useful in order to assess the quality of generalization in terms of shape maintenance.

It would be hopeful that a classification of measures, as well as a common terminology, be established and agreed by the research community. This would allow for a greater exchange of research results and could accelerate joint research in this crucial area of generalization.

ACKOWLEDGEMENTS

The author wishes to thank Robert B. McMaster and Jean-Philippe Lagrange for the numerous enriching discussions and their valuable suggestions.

REFERENCES

[Affholder, 93] J. G. Affholder. Road modelling for generalization.
NCGIA Initiative 8. Spec. Meet. Buffalo 1993

[Attneave, 54] F. Attneave. Some informational aspects of visual perception.
Psychological Review. Vol. 61, No. 3, 1954

[Buttenfield, 87] B. Buttenfield. Automating the Identification of Cartographic Lines.
The American Cartographer Vol. 14, N.1, pp. 7-20, 1987

[Buttenfield, 91] B. Buttenfield. A rule for describing line feature geometry.In Map
Generalization (B. Buttenfield & R. McMaster Eds) Part 3. P. 150-171
Ed. Longman Scientific & Technical. London 1991

[Herbert et al, 92] Herbert G., Joao E. M. et Rhind 1992 Use of an artificial intelligence
approach to increase user control of automatic line generalization.
EGIS 1992 p 554-563

[Hoffman et al, 82] D.D. Hoffman, W.A. Richards. Representing smooth plane curves
for visual recognition AAAI Proc. p. 5-8. 1982

[McMaster, 89] R. B. Mc Master. The integration of simplification and smoothing
algorithms in line generalization. Cartographica 26 p.101-121 1989

[McMaster, 93] R. B. Mc Master. Knowledge Acquisition for Cartographic Generalization:
Experimental Methods. ESF GISDATA Workshop Compiègne France
1993. To appear in "GIS and Generalization: Methodological and
Practical issues" Taylor & Francis, London. ESF GISDATA series

[Muller et al, 93] J. C. Muller, Z. Wang. Complex coast-line generalization.
Cartographic and Geographic Information Systems.
Vol 28-2,p. 96-106. 1993

[Plazanet et al, 94] C. Plazanet, J.P. Lagrange, A. Ruas, J.G. Affholder 1994
Représentation et analyse de formes pour l'automatisation
de la généralisation cartographique. EGIS'94 Proc. Vol 2. p 1112-1121

[Ruas et al, 93] A. Ruas. JP Lagrange, L. Benders 1993
Survey on generalization. IGN Internal Report

[Rosin, 93] Rosin P.L. 1993 Non-parametric multi-scale curve smoothing
SPIE Conference, XI: Machine Vision and Robotics, p. 66-77

[Thapa, 89] K. Thapa. 1989 Data compression and critical points detection using
normalized symmetric scattered matrix. AUTOCARTO 9, P. 78-89

MULTIPLE PARADIGMS FOR AUTOMATING MAP GENERALIZATION: GEOMETRY, TOPOLOGY, HIERARCHICAL PARTIONING AND LOCAL TRIANGULATION

Anne RUAS
IGN COGIT Laboratory
2 avenue Pasteur 94160 Saint-Mandé France
Fax: (33) 1 43 98 81 71 ruas@cogit.ign.fr

ABSTRACT

Generalization may be defined as a controlled reduction and simplification of geographical data. Despite the knowledge of basic generalization operators, the automation of generalization remains a complex issue. Actually the change in resolution induces numerous spatial conflicts and the use of generalization operators such as smoothing or aggregation may accidentally degrade geographical data. The aim of this paper is to propose a set of paradigms necessary to automate generalization. Firstly a space partitioning computed from structuring objects allows to define some working areas and to give a pre-sequence of transformations in the process of generalization. Then some local Delaunay triangulations allow to compute proximity relations between non-connected objects and to propagate displacements due to some geometric transformations. To ensure consistency during different geometric changes, both paradigms are integrated with a classical topo-metric structure.

INTRODUCTION

Generalization may be defined as a controlled reduction and simplification of geographical data. Despite the knowledge on basic generalization operators, the automation of generalization remains complex. Among theorical possibilities to automate generalization, one can either predefine the sequence of actions (e.g. "smooth roads then aggregate buildings and simplify building shapes then ..") or use an expert system that finds the next action according to the data and thanks to a rich set of rules. As Mackaness pointed out, the sequence of operations is context dependant and can not be ideally pre-defined: "varying the sequence of a set of operators will generate different solutions [..] any solutions are acceptable depending on the task and the intended message" (Mackaness,94). It seems clear that a batch process is not a good solution as generalization is not a deterministic process. On the other hand, experimental generalization expert systems are not successful (Herbert,91). One can argue that this methodology is not adapted to generalization but from our point of view the knowledge in generalization is not analysed and formalised enough to be successful. An expert system is only the means to this end.

This text aims at proposing the first process to automate generalization, based on conflict detection and resolving. It integrates multiple paradigms such as topology, geometry, hierarchical space partioning and local triangulation. We call "paradigm" each different knowledge category necessary for generalization in the sense given by (Herring,90).

As this study is closely related to IGN-France needs and research in generalization, our starting point is to generalize 2D medium-scale topographic data i.e. general geographic data, with a resolution of approximately 1 meter.

GENERAL PROCESS TO AUTOMATE GENERALIZATION

Let us take the implementation of an expert system for granted. The outline of the process of generalization based on conflict detection and resolving may be the following:

1. global conflict detection,
2. selection of a type of action,
3. conflict identification,
4. choice of a conflict to be solved on an object or a on set of objects,
5. choice of a specific operator such as simplification (Shea,89),
6. choice of an algorithm and a parameter value,
7. setting off the chosen algorithm,
8. internal validation: checking the effect of the algorithm on the object
9. propagation of the effect of the algorithm on the neighborhood,
10. contextual validation: checking the effect of the propagation,
11. going back to the first or third step.

The first two steps may be classed as a high level decision necessary to guide the overall process of generalization. They required a qualitative view on the data and sequencement decisions (e.g. global reduction of line points before more accurate line simplification, or selection of hydrographic objects before displacements due to object proximity).

This process of generalization is based on the "view, choose, act and control" approach. Most of the choices depend on semantic, geometric and spatial data characteristics. Consequently, to automate generalization it is necessary:

- to represent semantic, geometric and spatial information,
- to add specific attributes that allow to describe the behavior of different types of objects according to specific operators (e.g. add an "authorized displacement threshold: *ADT*" attribute to semantic objects and give a specific value according to object nature. In this way, a rule to control a displacement might be "respect *ADT* value of each object".)

Among the necessary studies, we have decided to give the highest priority to the following:

- finding the first method to sequence the action of generalization by means of a <u>hierarchical space partitioning</u>,
- finding a way of propagating a transformation, such as line simplication or displacement, on non-connected objects in order to maintain data consistencies by means of <u>Local Delaunay Triangulations</u>

COMMENTS ON CONFLICTS

A conflict is an infringement produced by the existence of some principles. When geographic information is concerned, a conflict occurs whenever information is not perceptible. One can argue that, in a data base, existing information is always perceptible. If generalization is defined as being a process of data simplification, it includes semantic and geometric simplification, i.e. the data base has a new geometric resolution which means that the objects must be big enough to be visible and recognizable and far enough from each other to be distinguishable. This change in resolution induces numerous geometric transformations such as shape-simplification, object displacement or, worse, change in object dimension (i.e. collapse), or even worse, the replacement of a set of objects by another set of representative objects (e.g. aggregation, typification). A *conflict,* according to us, is each spatial situation that does not meet visibility criteria. Being aware of our lack of knowledge and definitions of conflicts, we nevertheless propose some preliminary ideas:

- a conflict may involve different semantic objects,
- a conflict may involve different geometric objects (point, line, area),
- a conflict is closely related to the new required resolution, and in case of cartographic generalization it is also related to object symbolization,
- a conflict is seldom isolated. We can distinguish between:
  - — area of conflicts, qualified by a density of information such as a *congestion criterion,*
  - — inter-objects conflicts, that occur between identified objects and can be qualified by a *proximity criterion* (e.g. a conflict of proximity between a road and a house),
  - — intra-object conflicts, qualified by a set of *size criteria* (e.g. areas too small, too thin, lines too short, too detailed..).
- a conflict has a kind of "pass on" property: if O_a and O_b are in conflict and if O_b and O_c are in relation, then this relation has to be taken into account during the O_a and O_b conflict resolution otherwise a new conflict might appear between O_a and O_c or O_b and O_c.

As Mackaness noticed, next research needs to focus clearly on "a very comprehensive understanding of the conflicts" (Mackaness,94).

BASIC DATA MODELING: SEMANTIC AND TOPO-METRIC PARADIGMS

In the following we use "object formalism". Objects that share close behaviors are gathered into classes. The static part of an object is described by a set of attributes.

- *att'C denotes the attribute att of the classe C,*
- *att(o) denotes the value of the attribute att of the object o,*
- *Δatt denotes the potential values of att.*

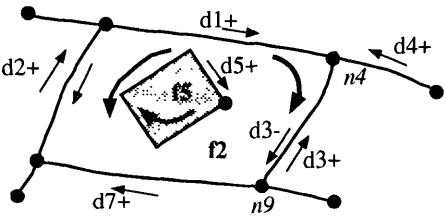
A geographic entity is characterised by a set of descriptives that define its nature and function (semantic part) and its location (geometric part). To structure and constrain the information, we differentiate the semantic and geometric parts into two main classes of information. So a geographic entity is represented by one or more geographic objects and one or more geometric objects.

We distinguish between complex geographical objects *Geo-C* and simple geographic object *Geo-S*. A complex object is composed of geographic objects. A simple object is linked with one or more geometric objects.

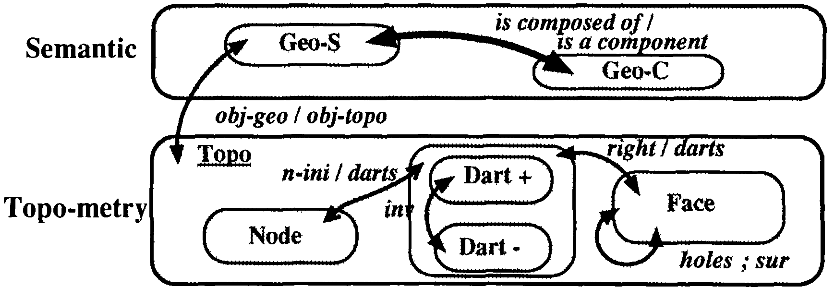
During the process of generalization it can happen that a complex object acquires a specific geometry.

As generalization induces numerous geometric changes, it is essential to manage spatial relations between objects (Ruas,93) (Lagrange,94). One solution is to project the geometry of the objects on a single plane and to split lines as soon as an intersection occurs. This representation of geometric information by means of a planar graph allows to describe connexity and inclusion relations easily. We use the concept of <u>topologic maps</u> (David,91) to represent the ***topo-metric*** information:

- a geometric object, called topo-metric, is either a *Node*, a *Dart* or a *Face*,
- a topo-metric object is linked to one or more geographic objects *obj-geo'Topo*,
- a dart is an oriented arc. It has a reverse dart *inv'Dart*, a right face *right'Dart*, an initial node *n-ini'Dart* and, if it is a positive dart, a list of coordinates *lxy'Dart+*,
- a node is linked with a set of output darts *darts'Node*
- a face is bounded by a set of darts *darts'Face*, and eventually a set of holes *holes'Face* or a surrounding face *sur'Face*



holes(f2)= {f4}
darts(f2)= {d1+,d3-, d7+, d2+}
darts(f5)= {d5+}
n-ini(d3+)= n9
n-ini(d3-) = n4
right(d1+)= f2
right(d5+)= f5
abj-geo(d1+)= road8
obj-topo(road8) = {d4+;d1-;..}

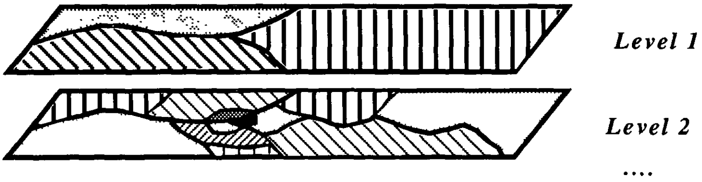


HIERARCHICAL SPACE PARTITIONING

1- Definition and objectives:

The hierarchical space partitioning is a set of areas of different sizes that divide the working space and lie on topo-metric objects. At each level a space partition exists in the mathematical sense. The aim of space partitioning is to find some logical working areas. It is a kind of "divide and conquer" approach which is useful for sequencing generalization actions and controlling the propagation of displacements.

The general idea is to detect objects which structure the space and are maintained during generalization, and to use these objects to compute partitions. The choice of structuring objects depends on the data base and the generalization purposes. The hierarchy allows to define the appropriate working area. Practically, 3 levels seem to be a convenient choice.



2- Structuring object choice

To construct a hierarchical space partitioning, it is necessary to find structuring objects. They require the following desirable properties:

- easy ordering,
- being appropriate to create cycles (necessary to partition the space),
- being maintained during generalization,
- having a density related to general information density.

As, at IGN-France, we are using medium-scale topographic data, we have tested the space partitioning paradigm on this data. A study of different medium-scale maps shows that the road network is the most appropriate structuring information as it is already classified, it is preserved during generalization (communication becomes more important as the scale decreases) and its density is related to the general information density, as roads are closely linked with human activity. This choice is also confirmed by a NCGIA study which concludes that the road network is well-preserved during generalization (Leitner,93), and by other research efforts to automate generalization (Peng,92) (Lee,93).

From a theoretical point of view the choice of structuring objects is important to assess the feasibility and utility of the paradigm. Practically, other networks such as main rivers or railways might be added to improve space partitioning. Moreover, if this method is helpful to define a sequence in conflict resolution, it does not mean that structuring objects are unchanged or even always preserved.

3- Space partitioning modeling

Given P1, P2.. Pn sub-classes of the generic class Partition, and E the total working area. E is limited with a set of darts whose hierarchy is the highest one (i.e. level 1). Let *niv* denotes the level attribute of a part.

For each given level:

R11: E is split into a set of connected and not overlapping parts:

$\forall$ i ε Δniv,

- $(p_a,..,p_j,...,p_n)$ ε Pi, $\Sigma_g\, p_j = E$ $\quad \Sigma_g$ = *geometric union*
- $\forall(p_a, p_b)$ ε Pi , $p_a\, \Lambda_g\, p_b = \phi$ $\quad \Lambda_g$ = *geometric intersection*

R12: Each part is bounded by darts that create a cycle:

$\forall\ i\ \varepsilon\ \Delta niv,\ \forall\ p\ \varepsilon\ Pi\ ,\ \exists\ (d_1, \ldots ,d_n)\ \varepsilon\ Dart\ /$

- $Boundary(p) = (d_1, \ldots ,d_n)$ *darts(p)={dj}*
- $\forall\ d\ \varepsilon\ [d_1,..,d_n],\ hierarchy(d) <= i$

Inclusion relations:

R21: Each part is composed of a set of sub-parts or a set of faces.

$\forall\ i\ \varepsilon\ \Delta niv,\ \forall\ p\ \varepsilon\ Pi\ ,$

If $\exists\ d\ \varepsilon\ Dart\ /\ hierarchy(d)>i\ \&\ dC_g p^o$

where C_g*=geometric inclusion and* p^o *= Interior of p*

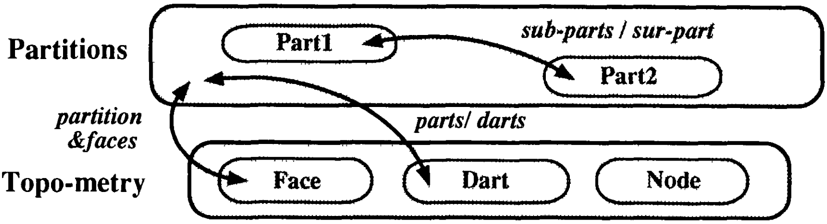
then $\exists\ \{q_j\}\ \varepsilon\ Pi+1\ /\ p = \Sigma_g\, q_j$ *sub-parts(p)={qj} & faces(p) = ϕ*

else $\exists\ \{f_j\}\ \varepsilon\ Face\ /\ p = \Sigma_g\, f_j$ *sub-parts(p)= ϕ & faces(p) =* $\{f_j\}$

R22: Each dart, which is a part boundary, is also one of its sub-part boundary if the part admits some sub-parts.

$\forall\ d\ \varepsilon\ Dart$ if $hierarchy(d) = i$ then $\exists\ p\ \varepsilon\ Pi\ /\ p\ \varepsilon\ parts(d)$

and if $(i+1)\ \varepsilon\ \Delta niv$ then $\exists\ q\ \varepsilon\ Pi+1/\ q\ \varepsilon\ sub\text{-}parts(p)\ \&\ q\ \varepsilon\ parts(d)$



4- Computation method

1. Class level hierarchy acquisition to compute cycles. (e.g. Level 1: Highway, Level 2: Main roads, Level 3: Secondary roads & Railways)
2. Propagation of semantic hierarchy on topo-metric objects (i.e. a dart inherits the highest level from its related geographic object)
3. Partition computation: A part is a cycle composed of darts of the same or higher level. For a given level, the partition computation is the computation of a cycle from a selected set of darts. The computation starts from the highest level. Each dangle-dart is detected and its hierarchy is decreased.
4. Inclusion relation computation between topo-metric faces and partitions: Each topo-metric face is linked with its smallest part which contains it: Either they partially share the same boundary (set of darts) or the face is surrounded by faces that are already linked with a partition (i.e. the inclusion relation is propagated between connected topo-metric faces).
5. Inclusion relation computation between partitions: This is the same process as mentioned before, starting from the lowest partition level and using darts and propagation between connected parts.

5- Conclusions on hierarchical space partitioning

Even if this space partition is not optimal, it constitutes the first method to sequence generalization actions. Thus it should be possible to solve geometric intra and inter-conflicts between the partition boundaries (i.e. a defined and limited set of darts) and then to solve conflicts inside each part. Moreover, in this way different conflicts should be generalized in parallel fashion. The number of hierarchy levels required depends on the density of information and the complexity of the data to generalize.

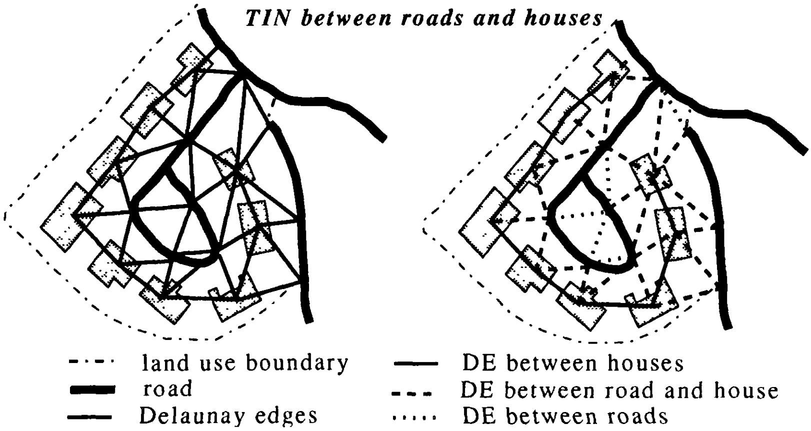
LOCAL DELAUNAY TRIANGULATION

1- Definition and objectives:

The topo-metric paradigm presented above allows to partition space onto topological faces and to represent connexity and inclusion relations, but it does not describe the relative location of non-connected objects that are included in these faces. Yet this knowledge is essential in generalization for operations such as conflict detection and displacement propagation (Lagrange,94). The use of TIN, that also represent topo-metric information, has already been proposed in literature (Gold,94) (Jones,92) but their definition and use are quite different from the ones proposed hereafter, named LDT for Local Delaunay Triangulation:

1. LDT does not hold topo-metric information,
2. LDT is a local triangulation i.e.
   a. it is computed when necessary and deleted just afterwards,
   b. there are many LDT at a time,
3. LDT is used
   a. to detect spatial characteristic relations,
   b. to detect proximity conflicts,
   c. to propagate displacements on-non connected objets.

This triangulation is mainly a means of storing different kinds of information according to the generalization operations. Thus the choice of triangulation points depends on the application.

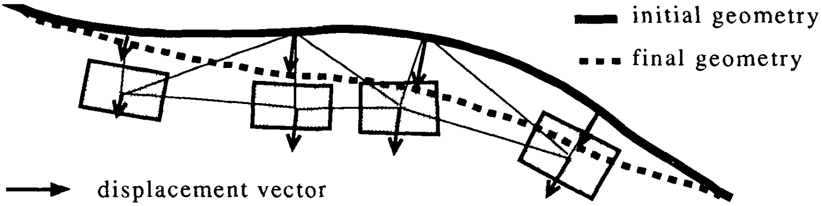


For displacement purposes, the LDT may be viewed as a representation of interaction forces between close objects. These forces allow to compute a decayed propagation from initial displacements. The LDT edges are classified in order to dinstinguish the propagation behaviors according to object nature.

2- The use of LDT and the triangulation node selection

Let us consider the problem of isolated object displacements, within a topological face, induced by the transformation of a nearby line. The sequence of operations could be the following:

1. From the dart d_α that will be modified, selection of the isolated close objects (these objects are included in the two faces connected to d_α and its reverse dart $d_{\alpha-}$),
2. Computation of the LDT between the isolated objects and d_α,
3. Classification of the LDT edges according to the nature of connected objects,
4. Computation of the new geometry of d_α ,
5. Computation of displacement vectors δdisp(d_α) along d_α,
6. Computation of displacement vectors δdisp-i on isolated objects by adding up the successive δdisp(d_α) values weighted according to LDT edge classification and euclidean distances,
7. Isolated object displacements,
8. Application of the transformation on d_α.

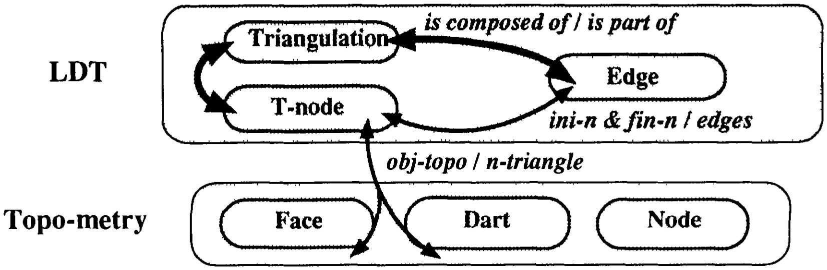


Actually the triangulation nodes are essentially anchor points on topometric objects. Thus for the above example we chose house centers and their projections on the initial road as LDT nodes. It is also possible to select the necessary points according to geometric criteria (e.g. anchor points on the modified line should be far enough one from another).

3- LDT Modeling

A LDT node is defined by:

1. Its membership in a particular triangulation,
2. A link between some specific LDT edges (*ini-n & fin-n <-> edges*),
3. Its coordinates,
4. An "anchor link" with a topo-metric object (*obj-topo <-> n-triangle*),
5. A displacement value.



4- Computation method

Whenever a dart is going to be displaced, the isolated objects within the connected faces are automatically selected by means of topological relationships. Triangulation nodes related to isolated objects are created. They are located at their centers. These nodes are then projected on close darts in order to create new, appropriate anchor nodes.

Then the local LDT is computed according to Tsai's *method (i.e. convex* hull computation and addition of other nodes, one by one) (Tsai,93).

The resulting edges are classified according to the relations between nodes and topo-metric objects, and relations between topo-metric objects and geographic objects.

5- Conclusion on LDT

The LDT have already been implemented. Current work is being done on the aggregation of displacement values on non-connected objects (i.e. decayed propagation according to euclidean distances). The next study should be on the use of LDT for other purposes such as conflict detection.

USE OF NEW INFORMATION STRUCTURES

We can already imagine a generalizing sequence that integrates the previous structures even if it is still somewhat sketchy. This sequence would involve the following:

1. Computation of topo-metric relations,
2. Detection of characteristic shapes and characteristic spatial relations,
3. Acquisition of specific object-behavior according to its nature and to the geometric transformation such as displacement, simplification, aggregation..,
4. Space hierarchy specification acquisition,
5. Space partitioning computation,
6. Intra-object conflict detection and resolution on the partition boundaries with propagation on close objects if necessary by means of LDT,
7. Inter-objects conflict detection and resolution on the partition boundaries with propagation on close objects if necessary by means of LDT,
8. Detection and resolution of intra and inter objects conflicts within partition by means of operations such as simplification, aggregation, collapse, displacements

Steps 6 and 7 are repeated for each hierarchical level. Each transformation should be decided upon and verified. According to specific test results, a backtracking to a previous state should be possible. This kind of mechanism may be implemented in a straight forward way with the expert system shell that we use.

CONCLUSION

Theoretically, spatial relations may be described and modeled in different ways (Kainz,90). Most of the time the choice of a paradigm depends on the kind of geometric data queries. Considering the complexity of generalization, it seems necessary to use different paradigms simultaneously. Topo-metric, space partitioning and local Delaunay triangulation paradigms have been developed within an object-oriented expert system in order to automate generalization. Links between these categories of information have also been implemented in order to ensure consistency during various geometric transformations. Further research, especially in conflict definitions, validation tools and behavior understanding is necessary in order to use these structures dynamically and to improve them. So far, the current state of this work provides us with a good basis for implementing first sequencement decisions.

REFERENCES

David, B. 1991 Modélisation représentation et gestion d'information géographique PhD Paris VI.

Gold, C. Three approaches to automated topology, and how computational geometry helps SDH'94 pp 145-158

Herbert, G. & Joao E.M. 1991 Automating map design and generalisation: A review of systems and prospects for future progress in the 1990's SERRL Report

Herring, J. Egenhofer, M. & Franck, A. Using category theory to model GIS applications SDH'90 Vol. 2 pp 820-829

Jones, C. Ware, J. & Bundy, G. Multiscale spatial modelling with triangulated surfaces SDH'92 Vol. 2 pp 612-621

Kainz, W. Spatial relationships - Topology versus Order. SDH'90 Vol. 2 pp 814-817

Lagrange, J.P. & Ruas, A. Geographic information modelling: GISs and Generalisation SDH'94 Vol. 2 pp 1099-1117

Lee, F. & Robinson, G. Development of an automated generalisation system for a large scale topographic maps. GISRUK'93

Leitner, M. 1993 Prototype rules for automated map generalization Master of Geography- Buffalo

Mackaness, W. Issues in resolving visual spatial conflicts in automated map design SDH'94 Vol. 1 pp 325-340

Peng, W. 1992 Automated generalization of urban road-networks for medium scale topographic data-bases. PhD ITC

Ruas, A. 1993 Modélisation des données pour la généralisation IGN Report

Shea, K.S. and McMaster, R.B. 1989: Cartographic generalisationin a digital environment: when and how to generaliralise Auto-Carto9 pp. 56-67

Tsai, V. 1993 Fast topological construction of Delaunay triangulations and Voronoi diagrams. Computers & Geosciences Vol. 19 N 10 pp. 1463-1474

A New Approach to Subdivision Simplification*

Mark de Berg
Dept. of Computer Science
Utrecht University
P.O.Box 80.089
3508 TB Utrecht
The Netherlands

Marc van Kreveld
Dept. of Computer Science
Utrecht University
P.O.Box 80.089
3508 TB Utrecht
The Netherlands

Stefan Schirra
Max-Planck-Institut für Informatik
Im Stadtwald
D-66123 Saarbrücken
Germany

Abstract

The line simplification problem is an old and well-studied problem in cartography. Although there are several efficient algorithms to compute a simplification within a specified error bound, there seem to be no algorithms that perform line simplification in the context of other geographical objects. Given a polygonal line and a set of extra points, we present a nearly quadratic time algorithm for line simplification that guarantees (i) a maximum error ϵ, (ii) that the extra points remain on the same side of the output chain as of the original chain, and (iii) that the output chain has no self-intersections. The algorithm is applied as the main subroutine for subdivision simplification.

1 Introduction

The line simplification problem is a well-studied problem in various disciplines including geographic information systems, digital image analysis, and computational geometry (see the references). Often the input is a polygonal chain and a maximum allowed error ϵ, and methods are described to obtain another polygonal chain with fewer vertices that lies at distance at most ϵ from the original polygonal chain. Some methods yield chains of which all vertices are also vertices of the input chain, other methods yield chains where other points can be vertices as well. Another source of variation on the basic problem is the error measure that is used. Well known criteria are the parallel strip error criterion, Hausdorff distance, Fréchet distance. areal displacement, and vector displacement. Besides geometric error criteria, in geographic information systems one can also use criteria based on the geographic knowledge, or on perception [Mark '89].

The motivation for studying these simplification problems is twofold. Firstly, polygonal lines at a high level of detail consume a lot of storage space. In many situations a high level of detail is unnecessary or even unwanted. Secondly, when objects are described at a high level of detail, operations performed on them tend to

*This research is supported by ESPRIT Basic Research Action 7141 (project ALCOM II: *Algorithms and Complexity*), and by a PIONIER project of the Dutch Organization for Scientific Research N.W.O.

be slow. An example where this problem can be severe is in the animation of moving objects.

Our motivation for studying the line simplification problem stems from reducing the storage space needed to represent a map in a geographic information system. We assume the map is modelled as a subdivision of the plane or a rectangular region thereof. In this application the main consideration is the reduction of the complexity of the subdivision. The processing time may be a little higher, but within reason. The size of the subdivision is a permanent cost in a geographic information system, whereas the processing time is spent only once in many applications.

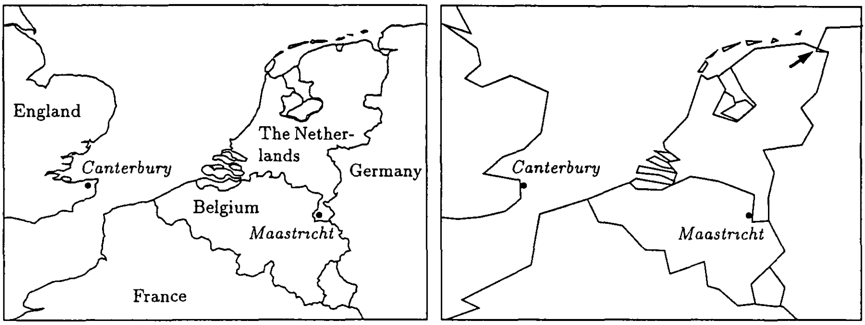


Figure 1: Part of a map of Western Europe, and an inconsistent simplification of the subdivision.

One of the most important requirements of subdivisions for maps is that they be simple. No two edges of the subdivision may intersect, except at the endpoints. This poses two extra conditions on the line simplification method. Firstly, when a polygonal chain is reduced in complexity, the output polygonal chain must be a simple polygonal chain. Several of the line simplification methods described before don't satisfy this constraint [Chan & Chin '92, Cromley '88, Douglas & Peucker '73, Eu & Toussaint '94, Hershberger & Snoeyink '92, Imai & Iri '88, Li & Openshaw '92, Melkman & O'Rourke '88]. The second condition that need be satisfied is that the output chain does not intersect any other polygonal chain in the subdivision. In other words, the simplification method must respect the fact that the polygonal chain to be simplified has a context. Usually the context is more than just the other chains in the subdivision. On a map with borders of countries and cities, represented by polygonal chains and points, a simplification method that does not respect the points can yield a subdivision in which cities close to the border lie in the wrong country. In Figure 1, Maastricht has moved from the Netherlands to Belgium. Canterbury has moved into the sea, and at the top of the border between The Netherlands and Germany, two borders intersect. Such topological errors in the simplification lead to inconsistencies in geographic information systems.

In this paper we will show that both conditions can be enforced after reformulating the problem into an abstract geometric setting. This is quite different from the approach reported in [Zhan & Mark '93], who have done a cognitive study on conflict resolution due to simplification. They accept that the simplification process may lead to conflicts (such as topological errors) and try to patch up the problems afterwards. We avoid conflicts from the start by using geometric algorithms. These algorithms are fairly easy to implement.

The remainder of this paper is organized as follows. Section 2 discusses our approach to the subdivision simplification, and identifies the main subtask: a new version of line simplification. Section 3 describes the approach of Imai and Iri for the standard line simplification problem. In Section 4 we adapt the algorithm for the new version of line simplification. In Section 5 the conclusions are given.

2 Subdivision simplification

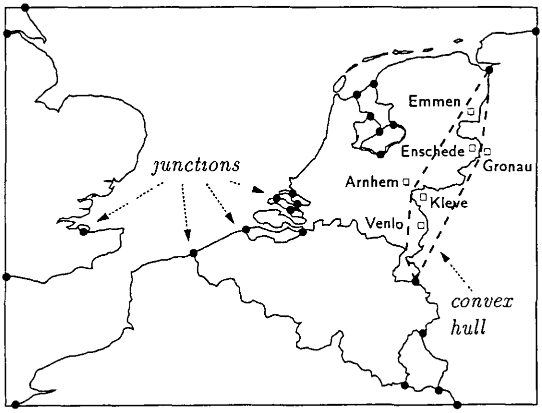


Figure 2: A subdivision with its junctions indicated.

Let S be a subdivision that models a map, and let P be a set of points that model special positions inside the regions of the map. The subdivision S consists of vertices, edges and cells. The degree of a vertex is the number of edges incident to it. A vertex of degree one is a *leaf,* a vertex of degree two is an *interior vertex,* and a vertex of degree at least three is a *junction.* See Figure 2. Generally the number of leafs and junctions is small compared to the number of interior vertices. Any sequence of vertices and edges starting and ending at a leaf or junction, and with only interior vertices in between, is called a *polygonal chain,* or simply a *chain.* For convenience we also consider a cycle of interior vertices as a chain, where we choose one of the vertices as start and end vertex of the chain.

Subdivision simplification can now be performed as follows. Keep the positions of all leafs and junctions fixed, and also the positions of the points in P. Replace every chain between a start and end vertex by a new chain with the same start and end vertex but with fewer interior vertices. If C is a polygonal chain, then we require from its simplification C':

1. No point on the chain C has distance more than a prespecified error tolerance to its simplification C'.
2. The simplification C' is a chain with no self-intersections.
3. The simplification C' may not intersect other chains of the subdivision.
4. All points of P lie to the same side of C' as of C.

Let's take a closer look at the last requirement. The chain C is part of a subdivision that, generally, separates two cells of the subdivision. In those two cells there may be points of P. The simplified chain between the start vertex and the end vertex will also separate two cells of the subdivision, but these cells have a slightly different shape. The fourth requirement states that the simplified chain C' must have the same subsets of points in those two cells.

The first requirement will be enforced by using and extending a known algorithm that guarantees a maximum error ϵ. The other three requirements are enforced by the way we extend the known algorithm. Roughly spoken, the simplified chain consists of a sequence of edges that bypass zero or more vertices of the input chain. We will develop efficient tests to determine whether edges in the simplified chain leave points of P to the wrong side or not. The second requirement, finally, doesn't add to the complexity of the algorithm. When applying the simplification algorithm to some chain of the subdivision, we temporarily add to the set P of points all vertices of other chains of the subdivision. One can show that—since C' has the vertices of other chains to the same side as C—the simplified chain C' won't intersect any other chain of the subdivision. A simplified chain that has the points of P to the correct side and doesn't intersect other chains in the subdivision is a *consistent simplification.*

A disadvantage of adding the vertices to the point set P is that P can become quite large, which will slow down the algorithm. There are two observations that can help reduce the number of points that need be added to P. Firstly, we only have to take the vertices of the chains that bound one of the two cells separated by the chain we are simplifying. Secondly, it is easy to show that only points inside the convex hull of the chain that is being simplified could possibly end up to the wrong side. So we only have to use points of P and vertices of other chains that lie inside this convex hull. In Figure 2, the chain that represents the border between the Netherlands and Germany is shown with its convex hull (dashed) and some cities close to the border (squares). No other chains intersect the convex hull, and only the cities Emmen, Enschede, Kleve and Venlo must be considered when simplyfing the chain.

It remains to solve a new version of the line simplification problem. Namely, one where there are extra points which must be to the same side of the original and the simplified chain. For this problem we will develop an efficient algorithm in the following sections. It takes $O(n(n+m)\log n)$ time for a polygonal chain with n vertices and m extra points. This will lead to:

Theorem 1 *Given a planar subdivision S with N vertices and M extra points, and a maximum allowed error $\epsilon > 0$, a simplification of S that satisfies the four requirements stated above can be computed in $O(N(N+M)\log N)$ time in the worst.case.*

The close to quadratic time behavior of the algorithm is the time needed in the worst case. Therefore, the algorithm may seem too inefficient for subdivisions with millions of vertices. A better analysis that also incorporates some realistic assumptions will show that the time taken in practice is much lower. It will also depend on the sizes of the chains in the subdivision, the number of extra points inside the convex hull of a chain, and the shapes of the chains themselves.

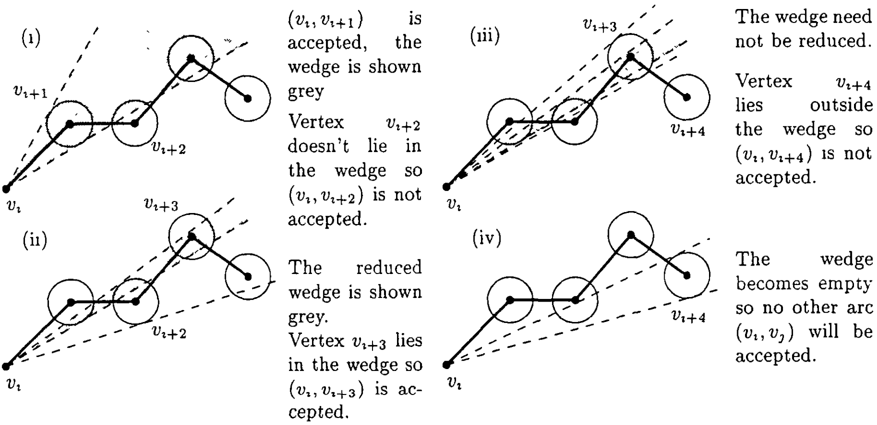
3 Preliminaries on line simplification

We describe the line simplification algorithm in [Imai & Iri '88], upon which our method is based. Let $v_1, \ldots, v_n$ be the input polygonal chain C. A line segment $\overline{v_i v_j}$ is aclled a *shortcut* for the subchain $v_i, \ldots, v_j$. A shortcut is *allowed* if and only if the error it induces is at most some prespecified positive real value ϵ, where the error of a shortcut $\overline{v_i v_j}$ is the maximum distance from $\overline{v_i v_j}$ to a point v_k, where $i \leq k \leq j$. We wish to replace C by a chain consisting of allowed shortcuts. In this paper we don't consider simplifications that use vertices other than those of the input chain.

Let G be a directed acyclic graph with as the node set $V = \{v_1, \ldots, v_n\}$. The arc set E contains (v_i, v_j) if and only if $i < j$ and the shortcut $\overline{v_i v_j}$ is allowed. The graph G can be constructed with a simple algorithm in $O(n^3)$ time and G has size $O(n^2)$.

A shortest path from v_1 to v_n in G corresponds to a minimum vertex simplification of the polygonal chain. Using topological sorting, the shortest path can be computed in time linear in the number of nodes and arcs of G [Cormen et al. '90]. Therefore, after the construction of G, the problem can be solved in $O(n^2)$ time. We remark that the approach can always terminate with a valid output, because the original polygonal line is always a valid output (though hardly a simplification). The bottleneck in the efficiency is the construction of the graph G. In [Melkman & O'Rourke '88] it was shown that G can be computed in $O(n^2 \log n)$ time, reducing the overall time bound to $O(n^2 \log n)$ time. In [Chan & Chin '92] an algorithm was given to construct G in $O(n^2)$ time. This is optimal in the worst case because G can have $\Theta(n^2)$ arcs. We explain their algorithm briefly.

One simple but useful observation is that the error of a shortcut $\overline{v_i v_j}$ is the maximum of the errors of the *half-line* starting at v_i and containing v_j, and the *half-line* starting at v_j and containing v_i. Denote these half-lines by l_{ij} and l_{ji}, respectively. We construct a graph G_1 that contains an arc (v_i, v_j) if and only if the error of l_{ij} is at most ϵ, and a graph G_2 which contains an arc (v_i, v_j) if and only if the error of l_{ji} is at most ϵ. To obtain the graph G, we let (v_i, v_j) be an arc of G if and only if (v_i, v_j) is an arc in both G_1 and G_2. The problem that remains is the construction of G_1 and G_2 which boils down to determining whether the errors of the half-lines is at most ϵ or not. We only describe the case of half-lines l_{ij} for all $1 \leq i < j \leq n$; the other case is completely analogous.



(v_i, v_{i+1}) is accepted, the wedge is shown grey

Vertex v_{i+2} doesn't lie in the wedge so (v_i, v_{i+2}) is not accepted.

The reduced wedge is shown grey.
Vertex v_{i+3} lies in the wedge so (v_i, v_{i+3}) is accepted.

The wedge need not be reduced.

Vertex v_{i+4} lies outside the wedge so (v_i, v_{i+4}) is not accepted.

The wedge becomes empty so no other arc (v_i, v_j) will be accepted.

Figure 3: Deciding which arcs (v_i, v_j) with $j > i$ are accepted to G_1. Only (v_i, v_{i+1}) and (v_i, v_{i+3}) will be accepted.

The algorithm starts by letting the vertices $v_1, \ldots, v_n$ in turn be v_i. Given v_i, the errors of all half-lines l_{ij} with $j > i$ are determined in the order $l_{i(i+1)}, l_{i(i+2)}, \ldots, l_{in}$ as follows. If we associate with v_k a closed disk D_k centered at v_k and with radius ϵ, then the error of l_{ij} is at most ϵ if and only if l_{ij} intersects all disks D_k with $i \leq k \leq j$. Hence, the algorithm maintains the set of angles of half-lines starting at v_i that intersect the disks $D_i, \ldots, D_j$. Initially, the set contains all angles $(-\pi, \pi]$. The set of angles will always be one interval, that is, the set of half-lines with error at most ϵ up to some vertex form a wedge with v_i as the apex. Updating the wedge takes only constant time when we take the next v_j, and the algorithm may stop the inner iteration once the wedge becomes empty.

With the approach sketched above, the graph construction requires $O(n^2)$ time in the worst case [Chan & Chin '92].

4 Consistent simplification of a chain

In this section we generalize the line simplification algorithm just described to overcome the two main drawbacks: it doesn't necessarily yield a simple chain and it doesn't leave extra points to the correct side. We only discuss the simplification of x-monotone chains. There are several ways to generalize our algorithms to the case of arbitrary chains. At the end of this section we sketch one method briefly; for details and extensions we refer to the full version of this paper.

A polygonal chain is *x-monotone* if any vertical line intersects it in at most one point. In other words, an x-monotone polygonal chain is a piecewise linear function defined over an interval. It is easy to see that any simplification of an x-monotone polygonal chain is also an x-monotone polygonal chain. Let C be an x-monotone simple polygonal chain with vertices $v_1, \ldots, v_n$. We denote the subchain of C between vertices v_i and v_j by C_{ij}. Let P be a set of m points $p_1, \ldots, p_m$. From the definition of consistency we observe:

Lemma 1 *C' is a consistent simplification of C with respect to P if and only if no point of P lies in a bounded region formed by C and C'.*

Let Q_{ij} be the not necessarily simple polygon bounded by C_{ij} and the edge $\overline{v_i v_j}$, so Q_{ij} contains $j-i$ edges of C and one more edge $\overline{v_i v_j}$. This last edge may intersect other edges of Q_{ij}. The general approach we take is to compute a graph G_3 with $\{v_1, \ldots, v_n\}$ as the node set, and an arc (v_i, v_j) whenever the bounded regions of Q_{ij} contain no points of P. So we don't consider the error of the shortcut $\overline{v_i v_j}$. This is done only later, when we determine the graph G on which the shortest path algorithm is applied. The graph G can be determined from the graphs G_1 and G_2 from the previous section, and the graph G_3 defined above. G has an arc (v_i, v_j) if and only if (v_i, v_j) is an arc in each of the graphs G_1, G_2, and G_3.

To compute arcs of the graph G_3, we consider for each vertex v_i the shortcuts $\overline{v_i v_j}$. We keep v_i fixed, and show that all arcs (v_i, v_j) with $i < j \leq n$ can be computed in $O((n+m)\log n)$ time. The first step is to sort the shortcuts $\overline{v_i v_{i+1}}, \ldots, \overline{v_i v_n}$ by slope. Here we consider the shortcuts to be directed away from v_i. Since C is x-monotone, all shortcuts are directed towards the right. The shortcuts are stored in a list L.

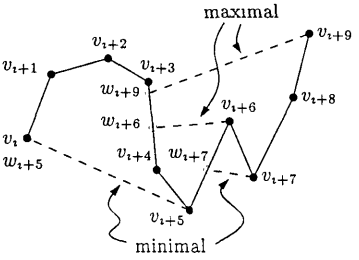


Figure 4: A part of a chain with four tangent splitters.

The second step of the algorithm is to locate all tangent segments from v_i. We define a shortcut $\overline{v_i v_j}$ to be *tangent* if v_{j-1} and v_{j+1} lie in the same closed half-plane bounded by the line through v_i and v_j, and $i+1 < j < n$. The shortcut $\overline{v_i v_n}$ is always considered to be tangent. The tangent shortcuts in Figure 4 are $\overline{v_i v_{i+5}}$, $\overline{v_i v_{i+6}}$, $\overline{v_i, v_{i+7}}$, and $\overline{v_i v_{i+9}}$. A tangent shortcut $\overline{v_i v_j}$ is *minimal (in slope)* if v_{j-1} lies above the line through v_i and v_j. If v_{j-1} lies below that line, then it is *maximal (in slope)*, and if v_{j-1} lies on the line it is *degenerate* (it has length zero). The *tangent splitter* is the line segment $\overline{w_j v_j}$ defined as the maximal closed subsegment of $\overline{v_i v_j}$ that does not intersect C in a point interior to $\overline{w_j v_j}$. So the point w_j is an intersection point of the chain C and the shortcut $\overline{v_i v_j}$, and the one closest to v_j among these, see Figure 4. If v_{j-1} lies on the shortcut $\overline{v_i v_j}$ then $\overline{w_j v_j}$ degenerates to the point v_j. A tangent splitter is minimal, maximal, or degenerate when the tangent shortcut is.

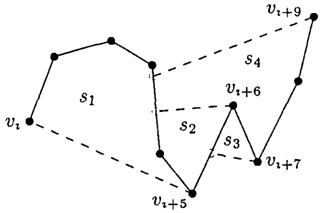


Figure 5: The corresponding subdivision S_i with cells $\gamma(1) = i + 5$, $\gamma(2) = i + 6$, $\gamma(3) = i + 7$, and $\gamma(4) = i + 9$.

Let $\overline{v_i v_{\gamma(1)}}, \ldots, \overline{v_i v_{\gamma(r)}}$ be the nondegenerate tangents. The corresponding set of tangent splitters and C together define a subdivision S_i of the plane of linear size, see Figure 5. The subdivision has r bounded cells, each of which is bounded by pieces of C and one or more minimal or maximal tangent splitters.

For every cell of S_i, consider the vertex with highest index bounding that cell. This vertex must define a tangent splitter, so it is one of $v_{\gamma(1)}, \ldots, v_{\gamma(r)}$. Assume it is $v_{\gamma(b)}$. Then we associate with that cell the number b. The subdivision and its numbering have some useful properties.

Lemma 2 *Every bounded cell of the subdivision S_i is θ-*monotone *with respect to v_i, that is, any half-line rooted at v_i intersects any bounded cell of S_i in zero or one connected components.*

Lemma 3 *Every bounded cell of the subdivision S_i has one connected subchain of C where half-lines rooted at v_i leave that cell.*

Lemma 4 *Any directed half-line from v_i intersects cells in order of increasing number.*

The points w_j can be found in linear time as follows. Traverse C from v_i towards v_n. At every vertex v_j for which $\overline{v_i v_j}$ is tangent (and non-degenerate), walk back along C until we reach v_i or find an intersection of $\overline{v_i v_j}$ with C. In the latter case, the fact that C is x-monotone guarantees that the point we found is the rightmost intersection, and thus it must be w_j. Then we continue the traversal forward at v_j towards v_n. This approach would take quadratic time, but we use the following idea to bring it down to linear. Next time we walk back to compute the next tangent splitter, we use previous tangent splitters walk back quickly. For a new maximal tangent splitter we only use previously found maximal tangent splitters, and for a new minimal tangent splitter we only use minimal ones. One can show that the skipped part of C never contains the other endpoint of the tangent splitter we are looking for.

The total cost of all backward walks is $O(n)$, which can be seen as follows. During the walks back we visit each vertex which is not incident to a splitter at most twice (once when locating w_j for a maximal tangent splitter $\overline{v_j w_j}$, and once for a minimal tangent splitter). Each splitter is used as quick walk backwards only once. So we can charge the cost of the backwards walks to the $O(n)$ vertices of C and the $O(n)$ tangent splitters.

The third step of the algorithm is to distribute the points of P among the cells of the subdivision S_i. Either by a plane sweep algorithm where a line rotates about v_i, or by preprocessing S_i for point location, this step requires $O((n+m)\log n)$ time [Preparata & Shamos '85]. All points of P that don't lie in a bounded cell of S_i can be discarded; they cannot be in a bounded region of the polygon Q_{ij} for any shortcut $\overline{v_i v_j}$. But we can discard many more points. For every cell of S_i, consider the tangent splitter with the vertex of highest index. If that tangent splitter is minimal, we discard all points in it except for the point p that maximizes the slope of the directed segment $\overline{v_i p}$, see Figure 6. Similarly, if the tangent splitter with highest index is maximal, we discard all points in the cell except for the point p that minimizes the slope of the directed segment $\overline{v_i p}$. Now every cell of S_i contains at most one point of P.

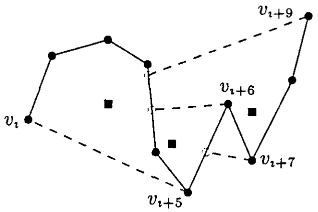


Figure 6: In each cell, only the point indicated by a square is maintained.

Lemma 5 *Any shortcut $\overline{v_iv_j}$ is consistent with the subchain C_{ij} with respect to P if and only if it is consistent with respect to the remaining subset of points of P.*

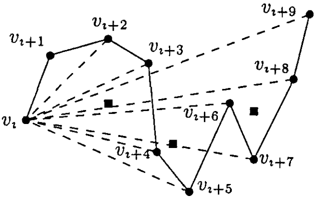


Figure 7: Only the shortcuts $\overline{v_iv_{i+1}}$, $\overline{v_i, v_{i+2}}$, and $\overline{v_iv_{i+3}}$ are accepted.

In the fourth step of the algorithm we decide which shortcuts $\overline{v_iv_j}$ are consistent and should be present in the graph G_3 in the form of an arc (v_i, v_j). We treat the cells of S_i in the order of increasing associated number. When treating a cell, we will discard any shortcut $\overline{v_iv_j}$ that has not yet been accepted and is inconsistent with respect to the one remaining point of P in that cell (if any). Then we accept those shortcut $\overline{v_iv_j}$ that have v_j on the boundary of the cell and have not yet been discarded. For discarding shortcuts, we use the order of shortcuts by slope as stored in the list L in the first step. For accepting shortcuts, we use the order along the chain C.

In more detail, the fourth step is performed as follows. Iterate through the cells $s_1, \ldots, s_r$ of S_i. Suppose that we are treating s_b. If there is no point of P in the cell s_b, then we skip the discarding phase and continue immediately with the accepting phase. Otherwise, let p_b be the point of P that lies in s_b. Assume first that the tangent splitter $\overline{w_{\gamma(b)}v_{\gamma(b)}}$ is minimal. Consider the list L of shortcuts starting at the end where shortcuts have the smallest slope. Repeatedly test whether the first shortcut at that end of the list L has larger or smaller slope than the line segment $\overline{v_ip_b}$. If the shortcut has smaller slope, then discard that shortcut by removing it from L. If the shortcut has larger slope, stop the discarding. In Figure 7, the shortcuts that are subsequently discarded when cell s_1 is treated are $\overline{v_iv_{i+5}}$, $\overline{v_iv_{i+4}}$, $\overline{v_iv_{i+7}}$, $\overline{v_iv_{i+6}}$, and $\overline{v_iv_{i+8}}$. If the tangent splitter is maximal then similar actions are taken, but on the end of the list L where the shortcuts have largest slope.

Lemma 6 *Every discarded shortcut $\overline{v_iv_j}$ is inconsistent with the subchain C_{ij} with respect to the points of P.*

After the discarding phase the accepting phase starts. For all vertices v_j with $\gamma(b-1) < j \leq \gamma(b)$ on C, if the shortcut $\overline{v_iv_j}$ is still in L, accept it by removing it from L and letting (v_i, v_j) be an arc in the graph G_3.

Lemma 7 *Any accepted shortcut $\overline{v_iv_j}$ is consistent with the subchain C_{ij} with respect to the points of P.*

The fourth step requires $O(n)$ time, which can be seen as follows. For each cell, we spend $O(d+1)$ time for discarding if d segments in L are discarded. This is obvious because discarding is simply removing from an end of the list L. To accept efficiently, we maintain pointers between the list L and the chain C so that shortcuts—once they are accepted—can be removed from L in constant time. Then we spend $O(a+1)$ time if a shortcuts are accepted. Since any shortcut is discarded or accepted once, and there are a linear number of cells in S_i, it follows that the fourth step takes linear time.

If we perform the above steps for all vertices v_i, then combine the obtained graph G_3 with the graphs G_1 and G_2 (as defined in the previous section) to create the graph G, we can conclude with the following result.

Theorem 2 *Given an x-monotone polygonal chain C with n vertices, a set P of m points, and an error tolerance $\epsilon > 0$, it is possible to compute the minimum link simplification of C that is consistent with respect to P and that approximates C within the error tolerance ϵ in $O(n(n+m)\log n)$ time.*

The simplification is also simple, but this is automatic because every x-monotone polygonal chain is simple. There are, however, several ways to generalize our results so that they can be applied to arbitrary, not x-monotone chains. Let C be such a chain, and let v_i be a vertex for which we wish to compute good shortcuts. One can determine a subchain $v_i, \ldots, v_k$ of C that is x-monotone after rotation of C. To assure that shortcuts $\overline{v_i v_j}$ with $j \leq k$ don't intersect edges before v_i or after v_k in the chain C, we add the vertices before v_i and after v_k to the set P of extra points. Then we run the algorithm of this section. One can show that any shortcut $\overline{v_i v_j}$ with $j \leq k$ that is consistent with respect to the extra points must be a consistent shortcut for the whole chain C, and it cannot intersect any edges of C. The generalized algorithm also runs in close to quadratic time.

5 Conclusions

This paper has shown that it is possible to perform line simplification in such a way that topological relations are maintained. Points lie above the original chain will also lie above the simplified chain, and points that lie below will remain below. Futhermore, the line simplification algorithm can guarantee a user specified upper bound on the error, and the output chain has no self-intersections. The method leads to an efficient algorithm for subdivision simplification without creating any false intersections. To obtain these results, we relied on techniques from computational geometry. We have also developed more advanced algorithms for simplifying arbitrary chains that allow of more reduction than the algorithm based on the idea described here. These extensions are given in the full paper.

With ideas similar to ours, some other line simplification methods can also be adapted to be consistent with respect to a set of tag points. In particular, the algorithm in [Douglas & Peucker '73] can be extended.

The given algorithm takes $O(n(n+m)\log n)$ time to perform the simplification for a chain with n vertices and m extra points. This leads to an $O(N(N+M)\log N)$ time (worst case) algorithm for simplifying a subdivision with N vertices and M extra points. There are many ideas that can be used to speed up the algorithm in practice. Therefore, we expect that the algorithm performs well in many situations, but probably not in real-time applications. Much depends on whether the quadratic time behavior of the method will actually show up on real world data.

The study in this paper has been theoretical of nature. Yet the given algorithms should be fairly straightforward to implement. We plan to implement our algorithm and run it on real world data. This way we can find out in which situations the efficiency of the method is satisfactory.

References

[Asano & Katoh '93] T. Asano and N. Katoh, Number theory helps line detection in digital images – an extended abstract. *Proc. 4th ISAAC'93*, Lect. Notes in Comp. Science 762, 1993, pp. 313–322.

[Buttenfield '85] B. Buttenfield, Treatment of the cartographic line. *Cartographica* **22** (1985), pp. 1–26.

[Chan & Chin '92] W.S. Chan and F. Chin, Approximation of polygonal curves with minimum number of line segments. *Proc. 3rd ISAAC'92*, Lect. Notes in Comp. Science 650, 1992, pp. 378–387.

[Cormen et al. '90] T.H. Cormen, C.E. Leiserson, and R.L. Rivest, *Introduction to Algorithms,* MIT Press, Cambridge, 1990.

[Cromley '88] R.G. Cromley, A vertex substitution approach to numerical line simplification. *Proc. 3rd Symp. on Spatial Data Handling* (1988), pp. 57–64.

[Douglas & Peucker '73] D.H. Douglas and T.K. Peucker, Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer* **10** (1973), pp. 112–122.

[Eu & Toussaint '94] D. Eu and G. Toussaint, On approximating polygonal curves in two and three dimensions. *Graphical Models and Image Processing* **5** (1994), pp. 231–246.

[Guibas et al. '93] L.J. Guibas, J.E. Hershberger, J.S.B. Mitchell, and J.S. Snoeyink, Approximating polygons and subdivisions with minimum-link paths. *Int. J. Computational Geometry and Applications* **3** (1993), pp. 383–415.

[Hershberger & Snoeyink '92] J. Hershberger and J. Snoeyink, Speeding up the Douglas-Peucker line simplification algorithm. *Proc. 5th Symp. on Spatial Data Handling* (1992), pp. 134–143.

[Hobby '93] J.D. Hobby, Polygonal approximations that minimize the number of inflections. *Proc. 4th ACM-SIAM Symp. on Discrete Algorithms* (1993), pp. 93–102.

[Imai & Iri '88] H. Imai and M. Iri, Polygonal approximations of a curve – formulations and algorithms. In: G.T. Toussaint (Ed.), *Computational Morphology,* Elsevier Science Publishers, 1988, pp. 71–86.

[Kurozumi & Davis '82] Y. Kurozumi and W.A. Davis, Polygonal approximation by the minimax method. *Computer Graphics and Image Processing* P**19** (1982), pp. 248–264.

[Li & Openshaw '92] Z. Li and S. Openshaw, Algorithms for automated line generalization based on a natural principle of objective generalization. *Int. J. Geographical Information Systems* **6** (1992), pp. 373–389.

[Mark '89] D.M. Mark, Conceptual basis for geographic line generalization. *Proc. Auto-Carto 9* (1989), pp. 68–77.

[McMaster '87] R.B. McMaster, Automated line generalization. *Cartographica* **24** (1987), pp. 74–111.

[Melkman & O'Rourke '88] A. Melkman and J. O'Rourke, On polygonal chain approximation. In: G.T. Toussaint (Ed.), *Computational Morphology*, Elsevier Science Publishers. 1988, pp. 87–95.

[Preparata & Shamos '85] F.P. Preparata and M.I. Shamos, *Computational Geometry – an introduction.* Springer-Verlag, New York, 1985.

[Zhan & Mark '93] F. Zhan and D.M. Mark, Conflict resolution in map generalization: a cognitive study. *Proc. Auto-Carto 13* (1993), pp. 406–413.

GRASS AS AN INTEGRATED GIS AND VISUALIZATION SYSTEM FOR SPATIO-TEMPORAL MODELING

William M. BROWN, Mark ASTLEY, Terry BAKER, Helena MITASOVA

Spatial Modeling and Systems Team
U.S. Army Construction Engineering Research Laboratories
P.O. Box 9005
Champaign, IL 61826-9005, USA
and
GIS Laboratory
University of Illinois at Urbana-Champaign
220 Davenport Hall, Urbana, IL 61801, USA.

ABSTRACT

Increasingly, geoscientific data is being measured in three dimensional (3D) space and time for studies of spatial and temporal relationships in landscapes. To support the analysis and communication of these data, a new approach to cartographic modeling is emerging from the integration of computer cartography and scientific visualization. This approach, which we call in this paper multidimensional dynamic cartography (MDC), is based on viewing data processed and stored in a Geographical Information System (GIS) in 3D space and visualizing dynamic models of geospatial processes using animation and data exploration techniques. Such techniques help researchers refine and tune the model in addition to making the model easier for others to understand. We describe various aspects of MDC implementation within GRASS GIS and illustrate its functionality using example applications in environmental modeling. Images, animations and other work associated with this paper may be viewed on the World Wide Web at URL: http://www.cecer.army.mil/grass/viz/VIZ.html

INTRODUCTION

Current standard GIS offer sophisticated tools for creating high quality cartographic output in the form of standard 2D maps. While such maps fulfill the requirements of GIS as an information retrieval and maintenance system, they are not adequate for application of GIS as a tool for analysis of spatio-temporal data and simulation of landscape processes which occur in 3D space and time. (McLaren 1989). To support such applications, GIS has been linked to various visualization systems. However, such links require the user to learn two different environments – one for processing the data and a second one for visualization. To increase efficiency and convenience for the researcher in using visualization for routine modeling and analytical work we have chosen to build visualization tools within GIS and design them to fulfill specific needs of geoscientific data.

Three levels of GIS and visualization integration are defined by (Rhyne 1994): rudimentary, operational and functional. The rudimentary level is based on data sharing and exchange. The operational level provides consistency of data and removes redundancies. The functional form provides transparent communication between the two software environments. We would add that at the functional level, data manipulation and derived data from visualization should be able to be written back to the GIS as new data layers, so that visualization can be used for data development.

Our MDC environment, integrated at the operational level, uses GIS to easily do transformations between coordinate systems, manage spatial data, run simulations, and derive statistical or other visually meaningful layers such as cross validation error or profile convexity (Wood 1993). The visualization tools use known characteristics of geographic data types to speed data exploration algorithms and provide more meaningful responses to interactive data queries. Customized visualization techniques are incorporated to meet common needs of researchers doing cartographic modeling.

An integrated approach also encourages verification of modeled scenarios with actual measured data by providing a common computational/display environment. Well tested models developed in this way are less likely to be dataset–specific because it should be easier to test the model with a variety of datasets under various geographic conditions.

MULTIDIMENSIONAL DYNAMIC CARTOGRAPHY

MDC may be thought of as a special case of scientific visualization; while more confining as to its requirements of data types (to conform to types supported by the GIS), MDC offers a specialized palette of data manipulation and viewing tools and symbolic representations that are meaningful to cartographers.

MDC can be used as either a process of research and discovery or a method of communicating measured or modeled geographic phenomena. As a process of discovery, the MDC process is cyclical in nature, with visualizations feeding a refinement of the model. As a method of communication, MDC is used to demonstrate complicated processes through the use of images and animations.

Combinations of various graphical representations of raster, vector, and point data displayed simultaneously allow researchers to study spatial relations in 3D space. At the same time, visual analysis of data requires the capability to distort this spatial relation by changing vertical scale, separating surfaces, performing simple transformations on point or vector data for scenario development, etc.

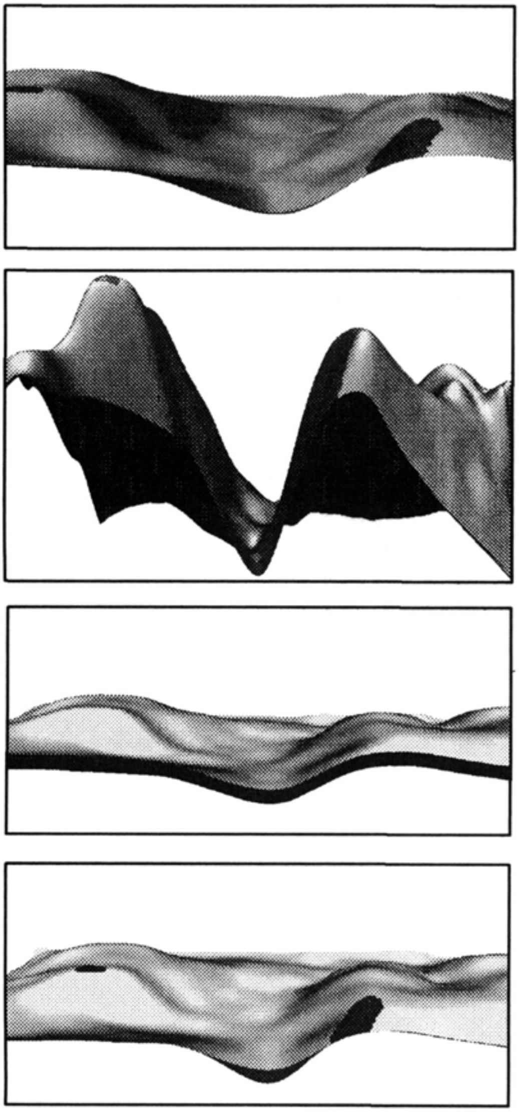
Multiple surfaces are quite useful to visualize boundaries of layers. For example, surfaces may be created that represent soil horizons so that thickness of layers may be displayed. This presents a technical challenge in terms of dimensional scale, as demonstrated by figure 1. The vertical dimension is often quite small relative to the other two, so is often exaggerated when a single surface is displayed. This exaggeration is usually adequate to add relief to an otherwise featureless surface, but in order to separate close stratified layers, the required exaggeration grossly distorts the modeled layers. If vertical translation of a surface is used to separate surfaces enough that they may be viewed separately, intersections between surfaces and relative distances are misrepresented. This is unacceptable since these may actually be the features we are interested in viewing. To study differences between two similar surfaces, we use a scaled difference approach where only the spatial distance between surfaces is exaggerated, maintaining correct surface intersections.

Animation is an important tool for exploring large and complex data sets, especially when they involve both spatial and temporal dimensions. Recent progress in graphics technology and emerging standards for animation file formats have made desktop animations easier to produce and share with colleagues. Animation is useful for representation of change in time, change in a modeling parameter (Ehlschlaeger 1994), change in viewer position (fly–bys) or change in visible data (fence cuts, slices). Figure 2 shows several frames from an animation where a fence cut is moved through data to better view underlying surface structure.

IMPLEMENTATION

General Description

GRASS (Geographic Resources Analysis Support System) as an open, public domain system with a full range of GIS capabilities has provided a sound basis for testbed development of visualization tools for MDC. Each GRASS data type (raster, vector, and site) plus our own 3D grid format may be used for visualization in a single 3D space. In our implementation, there are four object types and various ways to represent



No exaggeration yields little information about spatial differences between surfaces.

Uniform 10X exaggeration overly distorts surfaces.

No exaggeration, surfaces separated – surface intersections are lost.

10X exaggeration of difference between surfaces yields better visualization of relative differences.

Fig. 1 Attempts to visualize two similar surfaces

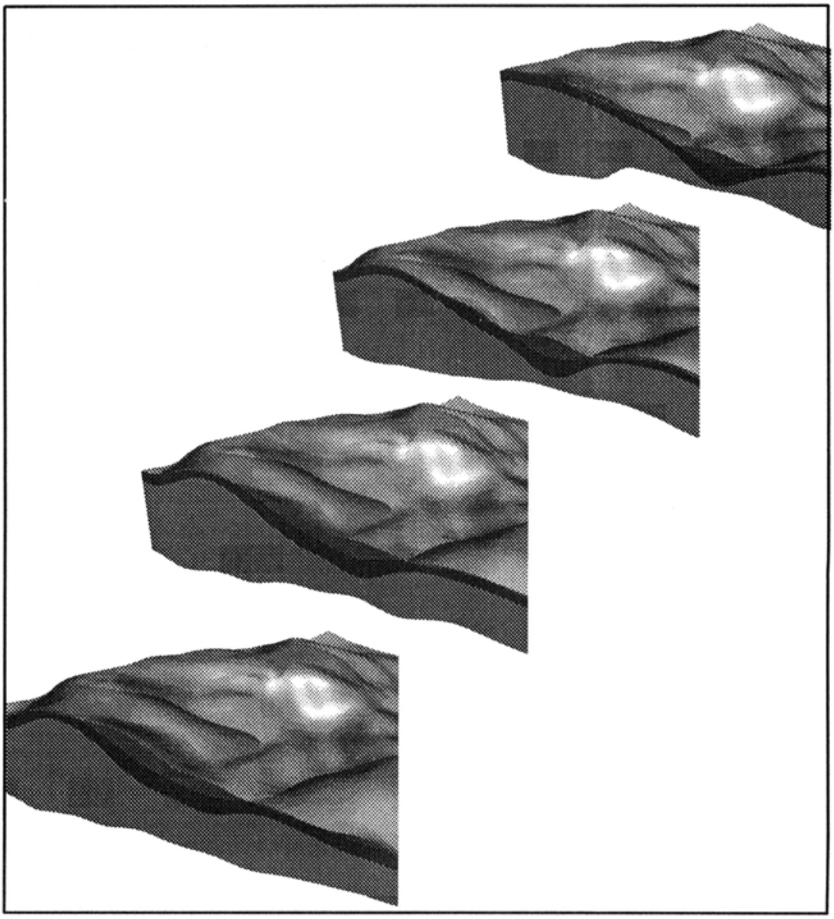


Fig 2. Moving a cutting plane through a group of surfaces
to examine underlying spatial relationships

each.

Surfaces. A surface requires at least one raster data set to represent topography and may use other raster data sets to represent attributes of color, transparency, masking, and shininess. These data sets may have been derived from vector (e.g., countour) or scattered point data using tools within the GIS. Users are allowed to use a constant value in the place of any raster data set to produce, for example, a flat surface for reference purposes or a constant color surface.

Vector sets. 3D vector sets are not currently supported, so in order to display 2D vector sets in 3 dimensions, they must be associated with one or more surfaces. The 2D vector sets are then draped over any of these surfaces.

Site Objects. Point data is represented by 3D solid glyphs. Attributes from the database may be used to define the color, size, and shape of the glyphs. 2D site data must be associated with one or more surfaces, and 3D site data may be associated with a surface (e.g., sampling sites measured as depth below terrain).

Volumes. 3D grid data may be represented by isosurfaces or cross sections at user–defined intervals. Color of the isosurfaces may be determined by threshold value or by draping color representing a second 3D grid data set over the isosurfaces.

For interactive viewer positioning, scaling, zooming, etc., we use custom GUI widgets and a "fast display mode" where only wire mesh representations of the data are drawn. When rendering a scene, the user may select various preset resolutions for better control over rendering time. For positioning we also chose to use a paradigm of moving viewer rather than moving object because we think it is more intuitive when modeling a reality of generally immobile geography. To focus on a particular object, the user simply clicks on the object to set a new center of view.

Scripting is used to create animations from series of data (e.g. time series or a changing modeling parameter), automatically loading the data sets and rendering frames for the animation. A keyframe technique is used to generate animations when there is no change in data, e.g., to create fly–bys or show a series of isosurfaces in volumetric data.

Interface

In addition to providing the functionality necessary for viewing various forms of data, it is also necessary to provide a usable front–end. In particular, we require the following features:

Intuitive: should employ user's preconceived notions of menus, buttons, etc.

Flexible: should be available on a variety of platforms and terminal configurations.

Extensible: should provide a natural mechanism for incorporating user supplied code.

Scripting: should be able to recreate complex interaction in order to automatically produce animations.

To satisfy these requirements our current interface is being developed using the Tcl/Tk toolkit written by John K. Ousterhout, and available via anonymous ftp. Tcl/Tk is an X windows based scripting and widget environment built on Xlib. Tcl/Tk provides a standard library of common widgets such as menus, buttons and the like, as well as a mechanism for developing arbitrarily complex custom widgets. Moreover, Tcl/Tk provides a natural mechanism for extension by allowing the creation of new commands via scripts or C code. Using this mechanism, we have incorporated our visualization library as an extension to Tcl/Tk. Thus, user's have access to visualization tools at the Tcl/Tk level, as well as the ability to extend Tcl/Tk with custom code.

Graphics Library

GRASS is designed to use various output devices for graphic display, via use of a display driver. The most commonly used is the X driver, for output to any X Window System display device. The GRASS driver interface is not currently capable of using 24–bit "true color" mode, as it was originally designed for display of a limited number of grid category values as a pseudocolor raster image. The GRASS driver only supports a limited set of graphic primitives, and only for a two dimensional, flat screen. While it is possible in such a system to display 3D objects and surfaces using a perspective projection, as was implemented in d.3d (USACERL 1993), all projection calculations have to be done explicitly and it is impossible to take advantage of specialized accelerated graphics hardware. Therefore we chose to use a separate display mechanism for visualization.

Fast graphics are required for effective interactive use of MDC. These graphics capabilities were identified early on as being integral in the development of MDC tools:

direct color as opposed to using a color lookup table

fast matrix manipulations for establishing perspective views and transforming object geometry

simulated lighting and material reflectance characteristics

depth buffer for hidden surface removal

double buffering: a scene is drawn in an invisible frame buffer (back buffer), then when the scene is complete, the entire frame buffer is quickly swapped with the visible frame buffer (front buffer).

As our work progressed into rendering multiple surface and volume data, we identified these additional capabilities that are highly desirable:
simulated transparency
user–defined arbitrary clipping planes

For our early tools we used IRIS GL, a software interface specific to Silicon Graphics hardware which took full advantage of hardware acceleration to provide the required functionality and offered an easy–to–use programmers' interface.

As the importance of hardware acceleration for common 3D graphics calculations is being recognized by more hardware vendors, standard software interfaces become necessary in order to increase portability of software and standardize expected rendering behavior. Two such standards are the PHIGS extension to X (PEX), and OpenGL. While PEX is a library, OpenGL is a specification for an application programming interface (API)(Neider 1993). As such, it is up to each hardware vendor to provide an implementation of OpenGL which takes advantage of proprietary hardware acceleration. For current development we chose to use OpenGL and GLX, the OpenGL extension to X, partly because of the ease of porting IRIS GL to OpenGL (enabling re–use of earlier software development) and partly because of our impression that OpenGL implementations would be more likely to provide superior performance on a wide variety of platforms.

EXAMPLES OF IMPLEMENTATION DETAIL

Computer Memory Considerations

Environmental data sets are very large, often involving several million data values for a single time step (Rhyne 1993). In order to effectively interact with this data, much of it needs to be loaded into computer memory to make access faster. The rendering process also adds additional memory overhead. During development of our tools, we tried to limit memory requirements for a typical application to 16–48 Mbytes, which is a common configuration for an average workstation.

In our implementation, 2D data sets are loaded into memory at the resolution defined by the user within the GIS. 3D data sets are not kept in memory, and the user may define whether specific vector or point data should be loaded into memory or read from the GIS whenever needed.

As an example, consider a simple 2D grid data set which is to be rendered as a surface with another 2D data set used for pseudocolor of the surface. In order to render a regular polygonal surface from this data, each data value will need to be located in 3D space by three coordinates with floating–point accuracy to represent a vertex of the surface. In order to simulate lighting, a three–component surface normal also needs to be calculated for each vertex. Assuming the data used to represent color is four byte data and floating point numbers occupy four bytes, the storage requirements for each surface vertex would be 28 bytes. So in the case of a 1000 X 1000 surface, we would already be using 28 Mbytes.

To reduce this requirement we can use known information about the data. Since we know the data represents a regular grid, with offset and resolution obtainable from the GIS, the east and west coordinates may be quickly calculated on–the–fly. We can interrogate the GIS to find the range and accuracy of the data and use the smallest data type possible to store the elevation and color of the surface. We could also calculate the surface normals on–the–fly, but the normals only change when the user adjusts resolution or exaggeration so we chose to pre–calculate and store surface normals to avoid repetitive calculations and speed rendering. However, we found through experimentation that the resolution of the surface normals could be reduced so that each normal may be stored in a single packed 4 byte field instead of three 4 byte floating point numbers. In the case of surface normals, we use 11 bits for each of the east and west components of the normal and 10 bits for the vertical component (taking advantage of our knowledge that a normal to a regular gridded surface will never be negative in the vertical component). Using these figures, the estimated requirements for a 1000 X 1000 surface reduce to 6–12 Mbytes.

When viewing multiple surfaces it is often common to use a data set more than once (perhaps topography for one surface and color for another). Our implementation shares data in such cases, freeing memory when a surface is deleted from the display only after checking that other surfaces aren't using the same data.

Mesh optimization was considered and rejected as a method of reducing memory requirements due to several conflicts: 1) In order to realistically render a mesh with widely varying polygon size, Phong shading is required for lighting calculations so that the surface normals are smoothly interpolated across the surface of each polygon. OpenGL only supports Gouraud shading, which interpolates color across the polygon from the color calculated at each vertex. 2) Draping 2D vector data on a surface requires finding the intersections of the vector lines with the edges of surface polygons. Using a regular mesh enables us to use faster intersection algorithms with less memory overhead. 3) In order to use a second data set for surface color, polygons of the optimized mesh would either have to be texture mapped, a feature fewer platforms support, or the two optimized meshes would need to be intersected, again a time consuming and overhead intensive process. 4) Use of an optimized mesh results in some degradation of data which is not easily quantifiable for accurate representation to the user.

Even using the lowest estimates, a surface that is 3000 X 6000 cells would require 108 Mbytes of main memory. So in practice, the user specifies a lower resolution and the data is automatically resampled by the GIS when loading. Given that the average workstation display only has about one million pixels, the whole data set at full resolution can not be completely displayed. Although there is an advantage to being able to browse the full data set, then zoom in on a small area at full resolution, in general we have found that users can be made aware of memory constraints and use the regular GIS display to either define a smaller region or reduce the resolution as necessary.

In the future we hope to explore the possibilities of using a surface with variable resolution. For example, a user could identify a region of special interest which would be drawn at full resolution within a broader region displayed at lower resolution.

Surface Querying

Querying a 2D data set displayed as a raster image can be thought of as a scale operation and a translation operation. When a user clicks on a pixel, the relative position in the image is scaled by the resolution of a pixel and the north and east offsets are added to obtain the geographical position. When displaying surfaces in 3D space with perspective, however, clicking on a pixel on the image of the surface really represents a ray through 3D space. The point being queried is the intersection of this ray with the closest visible (unmasked) part of one of the surfaces in the display.

One method for 3D querying provided by OpenGL is a "selection" method, where objects are "redrawn" without actually drawing to the screen, and any objects drawn at the query point are returned as the "selected" objects. This method is slow and at best the returned object is a polygon rather than a specific point on the polygon. Therefore, we require the user to specify the type of data they are querying (surface, point, or vector) and then use our knowledge of the geometry of that data to perform a geometric query in 3D. Figure 3 shows that using cutting planes can allow the user to query a specific location on a surface that is covered by another surface.

This specialized point–on–surface algorithm can be outlined as follows:

1) transform point on view plane to a view ray
2) intersect view ray with convex polyhedron defined by the intersection of the parallelepiped view region with any user defined cutting planes.
3) if ray enters this polyhedron, trace ray to find any intersections with visible (unmasked) parts of any surfaces
4) choose closest intersection to viewer (or return an ordered list)

Such point–on–surface functionality is useful for 3D data querying, setting center of view and setting center of rotation for vector transformations.

EXAMPLES OF APPLICATION

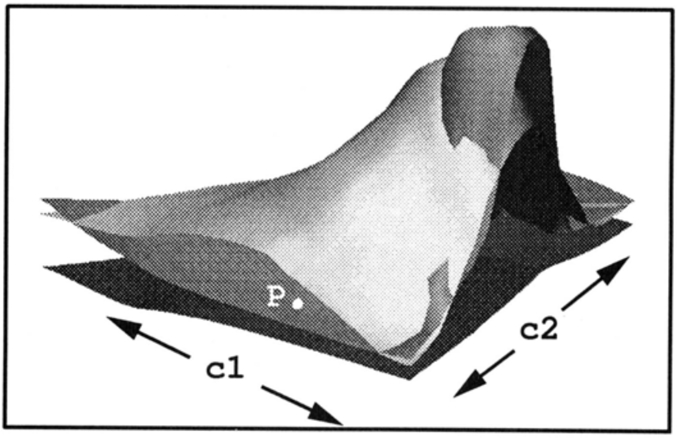
Examples of applications illustrating how a multidimensional dynamic approach to cartography increases its power for study of complex landscape processes are presented in a World Wide Web (WWW) document at http://www.cecer.army.mil/grass/viz/VIZ.html, accessible through Internet via browsers such as NCSA Mosaic. We have chosen to use the WWW to present and communicate applications of integrated GRASS GIS and visualization tools because of the opportunities to include full color images and animations which vividly illustrate the use of techniques discussed in this paper. The document covers several aspects of integration of GIS and visualization as they relate to applications in environmental modeling.

The first 3 documents ("Surface Modeling", "Multidimensional Interpolation", and "3D Scattered Data Interpolation") represent modeling of spatial and spatio–temporal distribution of continuous phenomena in 2D and 3D space. Animations illustrate the properties of interpolation functions (Mitasova & Mitas 1993) used to transform the scattered site data to 2D and 3D rasters and show the difference between trivariate and 4–variate interpolation for modeling spatio–temporal distribution of chemicals in a volume of water (Mitasova et al. 1993).

Animation based on changing data is illustrated by a dynamic surface representing change in monthly precipitation in tropical South America and by changing color map representing monthly temperatures on a static terrain surface in the same area. "Multidimensional Interpolation" also includes examples of a method for visualizing predictive error of interpolation along with the interpolated and animated surfaces using glyphs in data points. Images in "3D Scattered Data Interpolation" show the combination of all data types (2D, 3D raster, vector and sites), allowing users to study the spatial relationships between high chemical concentrations and terrain, railroads and wells (Fig 4). Animation is used to change the viewing position for better perception of spatial distribution of well sample sites in 3D space.

Terrain analysis and simulation of landscape processes related to fluxes of water and material are illustrated by "Terrain Analysis and Erosion Modeling" and "Rainfall Runoff", which include dynamic surfaces and multiple dynamic surfaces representing waterflow accumulation, change in sediment transport capacity and results of hydrologic simulation (rainfall, runoff and infiltration)(Saghafian 1993). Use of multiple surfaces with animated cutting planes for modeling of soil geomorphology and for evaluation of an erosion/deposition model by comparing it with the measured depth of colluvial deposits is illustrated by "Soil Geomorphology".

All documents include references to manual pages and tutorials in experimental HTML form. References to scientific papers and ftp sites for software distribution are also included.



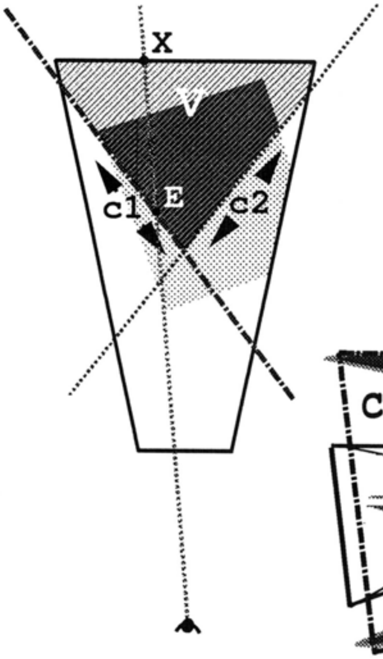
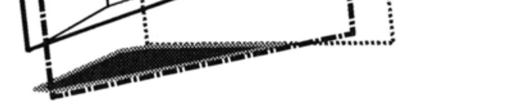


Fig. 3 Querying multiple surfaces. View of surfaces as seen by user (above). Two cutting planes, (c1) and (c2), are used to see lower surfaces better. When the user querys the image at point (P), the view ray intersects with the visible space enclosed by the convex polyhedron (V) at points (E) and (X). The resulting line is then traced for intersections with surfaces and the intersection nearest to the viewing position is used to query the database.



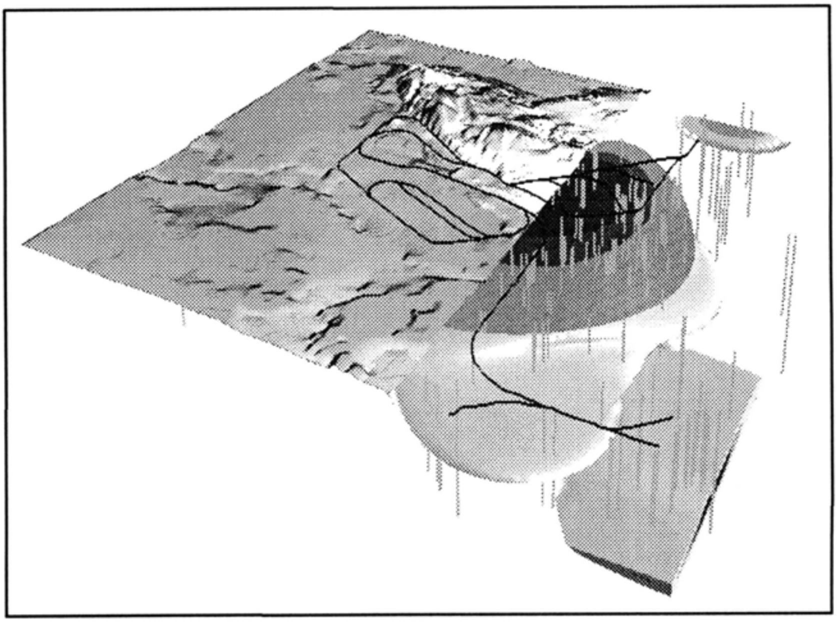


Fig. 4 Several georeferenced data types in one 3D space.

CONCLUSION

As the quantity of spatial data has grown in recent years due to more efficient data collection techniques such as remote sensing and Global Positioning Systems, GIS has played a vital role in managing this data. GIS capabilities such as user querying of data to obtain such metadata as coordinate system, scale and accuracy, or names of geographical features are sometimes taken for granted by modern cartographers, yet such features are often missing from generic modeling & visualization application programs.

When we speak of integration of GIS and scientific visualization, we are not talking about improved ways of accessing spatial data for visualization. Rather, we have identified opportunities for creating specialized custom visualization tools built for detailed analysis of georeferenced data. By integrating advanced visualization capabilities and modeling tools with the traditional spatial query and analysis functions of GIS, researchers are better able to evaluate a model's validity, explore possible causes of unexpected exceptions, tune modeling parameters, and re–visualize the results in a methodical, intuitive way.

ACKNOWLEDGEMENTS

This research was partially supported by the D.O.D. Strategic Environmental Research and Development Program, basic research, Digital Terrain Modeling and Soil Erosion Simulation.

REFERENCES

Ehlschlaeger, C., Goodchild, M. 1994, Uncertainty in Spatial Data: Defining, Visualizing, and Managing Data Errors. Proceedings, GIS/LIS'94, pp. 246–53, Phoenix AZ, October 1994.

McLaren, R.A., Kennei, T.J.M. 1989, Visualization of Digital Terrain Models, Three Dimensional Applications in Geographic Systems, Jonathon Raper Ed. Taylor and Francis Ltd.

Mitasova, H., Hofierka, J. 1993, Interpolation by Regularized Spline with Tension: II. Application to terrain modeling and surface geometry analysis. Mathematical Geology, Vol. 25:657–669.

Mitasova, H., Mitas, Lubos. 1993, Interpolation by Regularized Spline with Tension: I. Theory and Implementation, Mathematical Geology, Vol. 25:641–655.

Mitasova, H., Mitas, L., Brown, W.M., Gerdes, D.P., Kosinovsky, I., Baker, T. 1993, Modeling spatially and temporally distributed phenomena: New methods and tools for Open GIS, Proceedings of the II. International Conference on Integration of GIS and Environmental Modeling, Breckenridge, Colorado

Neider, J., Davis, T., Woo, M. 1993, OpenGL Programming Guide, Addison–Wesley Publishing Co., Reading, MA Robertson, P K. and Abel, D.J. 1993, Graphics and Environmental Decision Making IEEE Computer Graphics and Applications, Vol. 13 No. 2.

Rhyne, T. M., Bolstad, M., Rheingans, P. 1993, Visualizing Environmental Data at the EPA, IEEE Computer Graphics and Applications, Vol. 13 No. 2.

Rhyne, T. M., Ivey, W., Knapp, L., Kochevar, P., Mace, T., 1994, Visualization and Geographic Information System Integration: What are the needs and the requirements, if any ?, a Panel Statement in Proceedings IEEE Visualization '94 Conference, IEEE Computer Society Press, Washington D.C.

Saghafian, B. 1993, Implementation of a distributed hydrologic model within GRASS GIS. Proceeding of the II. International Conference on Integration of GIS and Environmental Modeling, Breckenridge, Colorado

U.S. Army Corps of Engineers Construction Engineering Research Laboratories 1993, GRASS 4.1 Reference Manual, Champaign, Illinois

Wood, J.D., Fisher, P.J. 1993, Assessing Interpolation Accuracy in Elevation Models IEEE Computer Graphics and Applications, Vol. 13 No. 2.

Exploratory Data Visualization for Reliable Integration

Eva-Maria Stephan
University of Zurich, 8057 Zurich, Switzerland
stephan@gis.geogr.unizh.ch

Abstract

Integrating different kinds of data is a typical feature and basic requirement of Geographic Information Systems (GIS). In particular, the modeling of environmental data and processes with GIS requires integration of many heterogeneous data sets. Required compatibility for integrating such data is often not given a priori. Data integration therefore often requires preprocessing of data, such as up- and downscaling, or interpolation, to establish compatibility. Such preprocessing, i.e., transforming of data sets may introduce error and uncertainty to the data. This paper focuses on visual exploration of heterogeneous, incompatible data for integration in order to arrive at greater reliabilty. Interactive investigation and visualization techniques for multiple data sets in a user-friendly environment are discussed.

1 Data Integration and Integrated Analysis

1.1 Introduction

Integrating different kinds of data is a typical feature and basic requirement of Geographic Information Systems (GIS). In particular, the modelling of environmental data and processes with GIS requires integration of many heterogeneous data sets. Driven by the quest for an increased understanding of the working and interrelationships of ecosystems, for example under the conditions of global climate change, combination of data from various sources is called for to derive at new data, e.g., data that is hard or expensive to measure. Through the combination of multiple data sets, *integrated analysis*, e.g., based on some deterministic or probabilistic model (Isaaks & Srivastava, 1989), new information (data) about other phenomena can be modellled. Integrated analysis based on multiple data sets allows one, for example, to calculate measured data of originally sparse resolution, into higher spatial or temporal resolutions.

1.2 Heterogeneity of Data

While GIS are suitable tools for bringing together data from various sources and different thematic fields they are usually devoid of the intelligence for *how* these complex (multidimensional and heterogeneous) data should be properly combined (Burrough & Hazelhoff, 1994). Heterogeneity of data sets involved in an integrated analysis need to be resolved beforehand. Those and problems of incompatibility arise from different characteristics and data quality of the data sets involved, i.e., different spatial and temporal references, various preprocessing of the data, validity regarding time, different classification of values, etc. Integrated analysis hence often requires transformation, a *pre-modelling* of data, which requires knowledge about the different data set's characteristics.

1.3 Virtual Data Set Approach for Integration

The Virtual Data Set approach, as presented by Stephan et al.(Stephan *et al.*, 1993), offers a flexible framework for data integration. Its basic idea follows an object oriented concept which guarantees a high flexibility and reliability for data integration. The concept proposes an enhancement of original data sets to so-called *Virtual Data Sets (VDS)* with persistently stored methods for prediction and data quality assessment. Those methods can be applied to transform a data set's spatial or temporal resolution, or attributes, i.e., to derive or predict new data values. A VDS thus contains in addition to the original data values, information that is not persistently stored, so-called *virtual data*. This data shall be derived at any time, e.g., to meet the requirements of a particular integrated analysis. Accordingly, the data set is able to adapt itself easily and flexibly to any requirements asked by the user, by performing the prediction methods which are stored jointly with the original data set.

The Virtual Data Set concept relies heavily on a sophisticated enhancement of the data set, i.e., on selecting and associating a data set with suitable methods for prediction and data quality assessment (see also (Bucher *et al.*, 1994)). One approach to determine the appropriate methods for the enhancement of data sets is supported by *visual data exploration*.

1.4 Data Enhancement through Visualization

The VDS concept requires methods of data visualization to gain rapid insight into the characteristics of spatial data, prevent false interpretation, and correlate data in order to promote their correct integration. Interactive visualization techniques that support simultaneous viewing of multiple data sets, correlations, trends, and outliers, allow visual exploration of data prior to their use and lead to new insights which may avoid wrong crucial decisions. For example, there are situations where false interpretation of data leads to false decision making and model-building, e.g., false assumptions about stationarity. The following paper will concentrate on visual data exploration for the enhancement, e.g., for reliable integration of data sets.

2 Visualization

2.1 Data Visualization

As mentioned above, natural science researchers are often confronted with large and heterogeneous collections of data. In order to analyze and integrate these, visual presentation of the data can be used for giving an overview, validating and understanding complex data sets. Combining the ability to both, visualize and interact with a spatial data set allows analysts to see complex patterns and trends associated with the given phenomenon. Data visualization can be applied for two main purposes:

- First, visualizing data can yield additional knowledge about the data's characteristics and its underling natural phenomenon. Evaluating the relevant data characteristics and their quality prior to their modelling leads to more reliable evaluation of an appropriate model. The additional insight can be used in an exploratory way to improve and validate accurate modelling of data and natural processes.
- Second, graphical depiction of data and its quality is an immediate way to communicate the uncertainty. It can communicate errors inherent to data and propagated in subsequent processing, e.g., integrated analyses.

2.2 Explorative Data Analysis

Data often are collected as observations of the real world rather than as the result of experiment. In the absence of experiment, appropriate probability models against which to test and evaluate the data are not well defined (Haining, 1990). Before any formal model can be set up, e.g., before integrated analysis can be carried out, it is useful to examine data for patterns, outliers, symmetry, homogeneity and distributional properties. In this sense *exploratory data analysis (EDA)* (Tukey, 1977) has attracted attention particularly for analysis of spatial data (Hasslett *et al.*, 1990). Graphical tools can serve to initially explore the data through hypothesis generation, modelling and analysis of residuals (Tukey, 1977). For spatial data, graphical methods of data portrayal are enhanced by recent advances in computer graphics systems, especially the facility to link windows dynamically and permitting several views of the data to be active at the same time (Hasslett *et al.*, 1990), (Aspinall, 1993).

2.3 GIS and ViSC

Although cartography has a long tradition in graphical depiction and analysis of spatial data (Bertin, 1967), (Tufte, 1983), GIS and computing environments offer a completely new working environment for geographical research. Emerging trends from computer graphics and scientific visualization have led to a whole new research topic of *Visualization in Scientific Computing (ViSC)*. These trends try to bring the measured and numerically modelled data to our eyes through the mediation of computer graphics and through use of the human visual sensory system and its ability to steer and interpret complex processes, to analyze trends and extract patterns in graphs and images. Actually, the goals are similar to the cartographic goals, sometimes cartography is referred to as the root of ViSC.

Unfortunately today's GIS visualization techniques are often designated for presentation purposes and designed to produce maps as output, rather than to provide functionality to interactive data visualization. Dynamic graphical data representations are mostly shortcomming and often require bigger efforts programming efforts. Data structures were primarily aimed at recreating the digitized map and could not respond to questions about the elements of the database (Wood & Broodlie, 1994).

Usually GIS offer visualization techniques for single scalar functions in 2 dimensional (sometimes 2 1/2 dimensional) space under stationary or time-dependent conditions. But the real world isn't that simple, and neither are the simulations that attempt to model it. Exploring the complexity and interactions of natural systems more appropriately can be achieved by combining an interplay of multidimensional visualization techniques. Multiple levels of presentation could facilitate the study of large data sets by providing, first, a quick look and then, if desired, a more detailed look with related position and scalars to compare their characteristics and quality in spatial, temporal and attributional dimensions.

With respect to providing a easy to access and environment for explorative data analysis, the following aspects of GIS visualization functionality need to be further improved:

- better user interface design with graphical interactive access to data and functionality,
- graphical programming and application building,
- 3-dimensional viewing of data coupled with the ability to interactively manipulate viewing angle and display parameters,
- providing visual techniques for presentation and interactive manipulation of data in *realistic* views, i.e. rendered images and complex scenes of the environment,
- providing techniques for continuous and contiguous data, such as surface rendering and animation,
- presenting data at multiple resolutions and allowing real-time up- and down-scaling,
- visualizing data quality,
- combined viewing of multiple data layers,
- linking surface and map-based viewing and statistical plots.

The next section will present a new concept of data visualization and example from a visualization tool, which was developed as an interactive working environment for EDA and user-friendly data viewing of environmental data. Its development was oriented along the above listed items, and aims at improving GIS' data viewing capabilities.

3 Interactive Visual Data Exploration

3.1 Data Scaping

Data Scaping [1] is a new concept for *interactive visual data exploration* of environmental data. It is based on the combination of scientific and cartographic presentations of data. The principal idea is to exploit possibilities offered by interactive graphic systems to support the search for patterns and peculiarities in environmental data. It is designed to support integrated analysis in a GIS computing environment.

This section will outline the requirements and specifications of this new visualization concept and present some examples from a first implementation of a Data Scaping-tool.

[1] The term was derived from *landscape viewing*. But rather than photo realistic images of landscapes, visualization techniques of Data Scaping present *data values* in rendered scenes and graphical plots.

3.2 Two Categories of Visualization

Two different categories of data visualization techniques help to gain insight into data and to to explore data for environmental modelling.

First, exploratory data analysis, based on (geo-)statistic plots shall yield at sampled data's characteristics, and thus increase the ability to reconstruct the spatio-temporal behavior of the phenomenon and consequently estimate attributes at un-sampled locations. I.e., exploratory data analysis supports and improves the interpolation of data for applications where continuous and contiguous data are lacking.

Second, once a data set represents continuous information, i.e., is interpolated on a regular grid or on a contiguous time-scale, then different techniques of *surface visualization and animation* of the data set can demonstrate the spatio-temporal behavior of the real-world phenomenon. Surface rendering, for example, shows the data in such a way, that they corresponds very closely to their real-world appearance, and thus allow for better understanding of its real world behavior. If multiple surfaces are stacked, each representing a different attribute layer, correlation and interdependencies of different attributes and model parameters can be explored.

These two categories of visualization form the bases for exploring the characteristics implicit to data in images. One can either represent the *numeric values* of measured point data using techniques of EDA, or visualize continuous data in their very natural appearance (scaling real world dimensions, space and time on regular grids), by using so-called *realistic* techniques. The level of abstractness (MacEacheren *et al.*, 1992) is very low when data is presented in an image or photo realistic scene, and it is very high, when presented in graphics symbolic form. Representing the two extreme points on an abstractness scale, the combination of both approaches helps to exploit the full range of data presentation.

3.3 Requirements to Data Scaping

The requirements of interactive data visualization are closely tied to explore the characteristics of environmental data. They are in particular:

Real time interaction: a graphical user interface which allows interactive selection and manipulation of

- display parameters, i.e., axes, colors, height, shading, surface texture, legend, etc.,
- viewing angle and object position,
- objects and attributes to be displayed
- resolution and version of an object.

Interactive probes: interactive querying of data values by selecting points on a map or surface with the mouse indicate the numerical values which are color or symbol coded.

2 dimensional views of data: contour plots, images, graphical plots, histograms, scatterplots, boxplots, scattermatrices.

3 dimensional views of data: representing gridded data as color surfaces with different display options (color, shading, transparency), as wire frames, or as combination of both; for surface views the x and the y dimension are always fixed by the data's geographical reference; the z dimension however, also can be given by other attributes than elevation; digital elevation models usually determine the basic shape (x, y and z dimension) for rendered surfaces, and can be overlaid (color-coded) by various attribute layers. The surface's actual z expansion in the display needs to be scalable. This is necessary to view multiple data surfaces with different value ranges in one display.

Animation: viewing time series or rotations of complex scenes;

Multiple linked windows: enables different views of data to be active at the same time and data manipulations in one window imply immediate updates of the representations in other windows.

Accepting different data types: regular or irregular grids, continuous and point data, time series and georeferenced data.

Open architecture: the visualization tool shall provide a port to a GIS or database system.

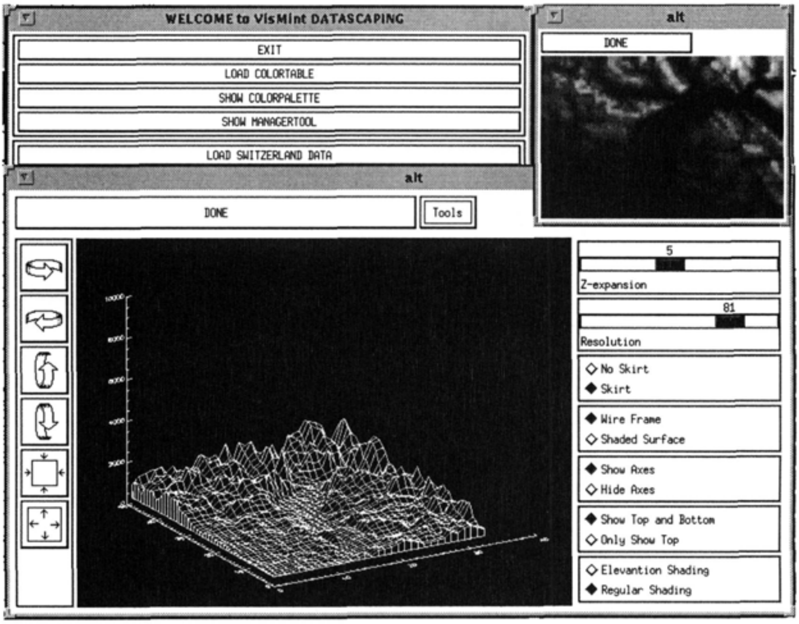
3.4 VisMint Data Scaping Tool

The following images (screen shots) from the explorative Data Scaping tool VisMint (Visualization Methods for Data Integration) are to illustrate how different visualization techniques of 3 dimensional surface visualization may help to reveal structures and correlations of dynamic natural processes. The data used for the illustrations are gridded (environmental) data sets, various attribute layers and a digital terrain model. The data sets represent environmental information about Switzerland and have different resolutions. The interface and visualization tool were implemented as application framework with the advanced visualization system IDL (RSI, 1993).

Various data techniques supported by VisMint and an efficient user interface provide viewing and interaction of regular gridded data in two to four dimensions. In the first figure one can see parts of the user-interface, where data sets can be selected. Further it shows two windows with different views of the *altitude* (elevation) data layer. One is showing a color coded image and one presents the data as rendered surface. In this case the surface representation was chosen to be a wire frame, with 81 per cent of the points of the full resolution. Many other choices of display parameters and and manipulations are possible with the shown surface viewer tool. Choice by mouse and direct feedback is provided.

Figure 2 shows a combined view of the yearly temperature distribution. In the above window, an image view, wire frame surface and a contour plot are combined. Below the same data set is represented as a surface and as a histogram. The windows can be linked. Selecting data points in the histogram plot will highlight the selected points in the other images, so that their spatial distribution can be looked at.

The third screen shot shows a combination of the two data sets, mean annual temperature and yearly precipitation sum. Both data sets are combined as shaded and as wire frame surface in the above window with the so-called multiple surface viewer. Sliders and buttons allow the user to rearrange and manipulate the display, e.g., moving and scaling objects, viewing angel and surface representation. The profile



5
Z-expansion

81
Resolution

◇ No Skirt
◆ Skirt

◆ Wire Frame
◇ Shaded Surface

◆ Show Axes
◇ Hide Axes

◆ Show Top and Bottom
◇ Only Show Top

◇ Elevantion Shading
◆ Regular Shading

Figure 1: VisMint Data Scaping: main menu and surface viewer

plot in the window below provides a linked viewing of the two surfaces, The two profile plots show a vertical cut through the two surfaces and changes dynamically when the mouse is positioned in the window with the two surfaces. The simultaneous viewing yields at the different characteristics of the two data layers. The two smaller windows show again a color image and a shaded surface of the temperature data.

The program provides many more selected data visualization techniques for displaying one or more geographically referenced variables and some techniques for animating time-series. These techniques enable many interrelated natural phenomena to be shown in a format that fosters comparison. By offering data visualizable in images that correspond very closely to the data's real world appearance one can better explore spatial phenomena and validate and compare preprocessed data. This kind of visualization, in the sense of real world simulation is particularly interesting, as the visualized data first have to be mathematically computed. As they can be altered, different scenarios can be carried out and compared with one another.

Usually, all images can interactively be altered according their viewing angle, color, and scaling, which help to perceive 3 dimensions (and time) on a 2 dimensional screen. Interactive graphics has to be experienced, static prints of sceendumps can only give an impression of their potential.

The presented visualization tool VisMint Data Scaping is still under development and being further completed. It currently being use in an environmental modelling project and the evaluation of its significance to data integration is further investigated.

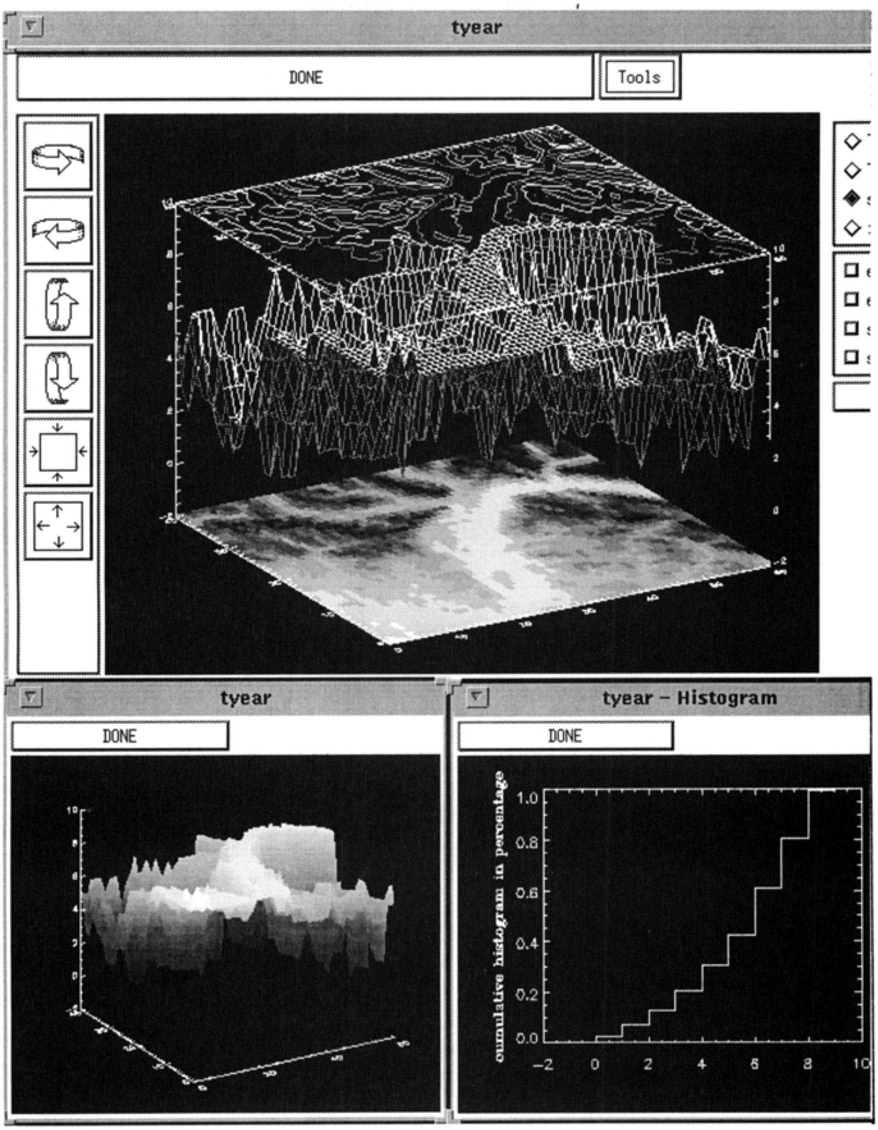
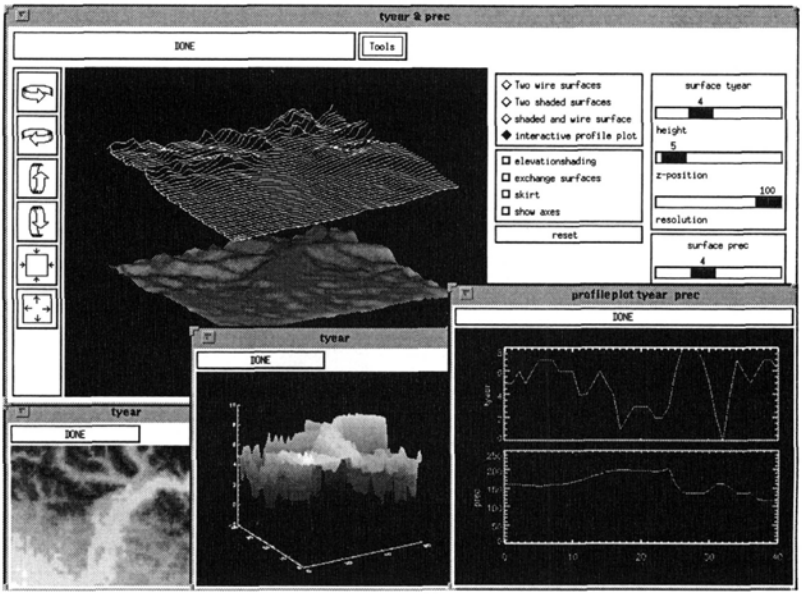


Figure 2: VisMint Data Scaping: combined viewing of a data set



◇ Two wire surfaces
◇ Two shaded surfaces
◇ shaded and wire surface
◆ interactive profile plot

□ elevationshading
□ exchange surfaces
□ skirt
□ show axes

reset

surface tyear
4
height
5
z-position
100
resolution

surface prec
4

Figure 3: VisMint Data Scaping: two surface viewer in interactive profile plot mode

4 Conclusion

The previous sections seek to stress the use of interactive multidimensional visualization techniques from the computer graphics and scientific visualization community for *interactive explorative analysis of environmental data in GIS*. The concept of *Data Scaping* which was introduced, relies on the idea of heavily linking data and their display to gain more insight into data to study their *natural behavior and interrelations*. A implementation of a viewing tool showed that interactive techniques for the interactive examination of environmental data sets can help to better understand data and its underlying natural phenomena.

Data Scaping is to be seen in connection with the VDS concept. The concept requires a profound understanding of data for defining suitable methods for prediction and data quality assessment, which can obtained by visual data exploration. Visualization techniques can serve for exploring uncertainties in two ways, multidimensional exploratory statistical analysis and a multidimensional viewing of geographic surfaces. This paper outlines both approaches and presents visualization techniques such as graphical statistics, dynamic graphics and visual cue techniques for EDA. Using methods for exploring environmental data can contribute much to reveal uncertainties and to gain better insight into accurate modelling.

References

Aspinall, R. J. 1993. Exploratory Spatial Data Analysis in GIS: Generating Geographical Hypothesis Describing Farm Types from Spatial Census Data. *Pages 517–526 of: Conference Proceedıngs EGIS'93.*

Bertin, Jacques. 1967. *La Semıologıe Graphıque.* Gauthiers-Villars.

Bucher, Felix, Stephan, Eva-M., & Vckovski, Andrej. 1994. Standardization and Integrated Analysis in GIS. *Pages 67–75 of: Proceedıngs EGIS'94.*

Burrough, Peter, & Hazelhoff, L. 1994. *Envıronmental Assessment and GIS.* EGIS'94 Workshop.

Hasslett, J., Wills, G., & Unwin, A. 1990. SPIDER - an interactive statistical tool for the analysis of spatially distributed data. *Internatıonal Journal of Geographıcal Informatıon Systems.*

Isaaks, E.H., & Srivastava, M.R. 1989. *An Introductıon to Applıed Geostatıstıcs.* Oxford University Press.

MacEacheren, A.M., M., Alan, Buttenfield, Barbara P., Campbell, James B., DiBiase, David, & Monmonier, Mark. 1992. Visualization. Rudger's University Press.

RSI. 1993. *IDL - Interactıve Data Language.* Research Systems, Inc.

Stephan, Eva-M., Vckovski, Andrej, & Bucher, Felix. 1993. Virtual Data Set: An Approach for the Integration of Incompatible Data. *Pages 93–102 of: Proceedıngs of the Eleventh Internatıonal Symposıum on Computer Assısted Cartography.*

Tufte, E.R. 1983. *The Vısual Display of Quantıtatıve Informatıon.* Graphics Press.

Tukey, J.W. 1977. *Exploratory Data Analysıs.* Adisson Wesley.

Wood, M., & Broodlie, K. 1994. ViSC and GIS: Some Fundamental Considerations. West Sussex, England: Wiley.

APPROACHES TO TRUTH IN GEOGRAPHIC VISUALIZATION

Alan M. MacEachren
Department of Geography
Penn State University
University Park, PA, 16802
e-mail: NYB@PSUVM.PSU.EDU

ABSTRACT

This paper addresses two related issues: how we can judge and represent 'truth' in the context of Geographic Visualization (GVIS) and what 'truth' is in this context. The first issue is approached from an analytical perspective, with emphasis on measurable aspects of validity in geo-referenced information and on representational methods for depicting validity. In relation to the second issue, a philosophical perspective on truth emphasizes the concept of *cognitive gravity* as a factor that leads scientists and policy makers to see only what they expect to see. The goal of this essay is not to provide specific guidelines for dealing with aspects of truth in GVIS, but to introduce a framework for exploring the relevant issues. The framework proposed is grounded in a semiotics of geographic representation.

INTRODUCTION

Truth of spatial information, and of decisions based on that information, is a significant issue for both scientific research and policy formulation. As we enter an era in which visual evidence is gaining (regaining?) prominence (due to technological advances associated with scientific visualization and multimedia), we should take a critical look at the implications of our visual tools. Geographic visualization systems (particularly when linked to GIS) provide powerful prompts toward insight by scientists and serve as facilitators of policy decisions by a range of analysts, resource managers, planners and others. Most GVIS research thus far has focused on getting new tools into the hands of potential users (for example, see: Dorling, 1994; Kraak, 1994; Asche and Herrmann, 1994). It is time that we begin to give some attention to the implications of those tools. These implications are integrally related to the truth of the representations that GVIS provides and to the truth of mental visualizations that it generates. Scientists and policy makers will inevitably make judgements about GVIS truth. At this point, we are just beginning to understand the range of technical and social factors that have an impact on these judgements – and the ways in which these judgements influence scientific thinking and public policy.

A discussion of truth in GVIS must begin with an understanding of GVIS itself. Visualization is a concept with a variety of meanings and the addition of "geographic" as a prefix does not necessarily clarify things. Elsewhere, I have defined GVIS in terms of map use – in contrast to other definitions that have focused on the technology that underlies visualization (MacEachren, 1994). The emphasis of this use-based characterization is on the nature of interaction between users and visualization tools. I see visualization and communication as occupying two corners of a three-dimensional visual tool use space (defined by three axes: public–private, high-interaction – low-interaction, seeking unknowns–accessing knowns). Within this tool use space, I argue that the prototypical examples of geographic visualization are those that combine relatively private use of maps (and other spatially-referenced displays), a goal of seeking unknowns in a set of information, and a process that is highly interactive. It is in the context of scientific exploration and the search for unknowns that issues of truth are particularly complex and problematic. How, for example, are we to judge the truth of a science based increasingly on visual evidence for which we have no standards of truth?

Within the realm of highly interactive use of visual tools as potential prompts to insight, issues of truth can profitably be addressed from the perspective of *semiotics* – the science of "signs." Signs are relations among a *referent* (an entity in the world), a *sign-vehicle* (a representation that stands for that entity), and an *interpretant* (the shared meaning that links the two) (figure 1). As such, signs exist at multiple levels in a visualization context. Each individual mark on a display can be the visible component of a sign relation (i.e., the sign-vehicle), groups of these marks can be seen as a whole forming the visual part of a higher order relation (a compound sign-vehicle), while entire displays can "stand for" complex entities (referents) and the concepts (interpretants) they are associated with.

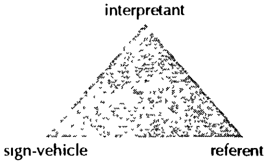


Figure 1. A triadic model of sign relations

Starting from this semiotic base, truth in GVIS can be approached from two perspectives. The first is an analytical one that equates "truth" with validity of the sign relations. From this perspective, the possibility for truth is assumed, and the goal becomes that of increasing the chance for valid signs by making more reliable displays, by measuring and representing display reliability, and by developing display forms that decrease the potential for interpretation error. The second perspective from which GVIS truth can be approached is a conceptual or philosophical one in which the very notion of truth is questioned. The goal is to explicate the standards against which truth of sign relations might be judged.

ANALYTICAL APPROACHES TO "TRUTH" IN GVIS

In practice, the concept of truth is often simplified to one of validity (or reliability). From the semiotic perspective described above, this restricted view of truth has two subcomponents. First, validity can be assessed in terms of the direct sign-vehicle to referent link (dealing with what is referred to as the *semantics* of the sign) (figure 2a). Second, validity can be assessed in relation to the indirect link through the interpretant (i.e., sign *pragmatics*) (figure 2b). In the first case, sign-vehicle validity is defined in relation to the reliability with which sign-vehicles and referents are matched (e.g., how certain can we be that a location on a 100 inch isohyet received 100 inches of precipitation). In the second case, interpretant validity is relative to the meaning we attach to the sign relation, thus putting emphasis on the cognitive models and knowledge that the user brings to the GVIS environment.

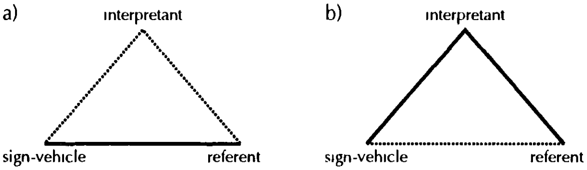


Figure 2. The emphasis of semantics (a) is on the sign-vehicle to referent link. That of pragmatics (b) is on the link through the interpretant.

GVIS semantics – sign-vehicle validity

A semantics of GVIS should address three components of sign-vehicle validity, that associated with data quality, that associated with the representation form, and that associated with the representation process (i.e., how data are manipulated in order to allow a particular representation form to be applied). Validity issues associated with representation form involve the selection of "data models" through which data (generally treated as referents) are signified. Validity issues associated with the representation process center on assumptions that underlie that process and on the analysis of what has been called "method produced error" inherent in spatial data processing.

The data quality component can, perhaps, be addressed most effectively through direct representation of reliability estimates (assuming, of course, that such estimates can be derived). As digital data begin to comply with the Spatial Data Transfer Standard (FIPS 173), reliability assessments should become a more common component of digital spatial data. On the assumption that reliability information will begin to accompany most data sets, considerable attention has been directed to developing appropriate representation methods. Among the issues addressed are: symbolization (i.e., sign-vehicle) schemes for signifying data reliability (MacEachren, 1992; McGranaghan, 1993; Fisher, 1994a), the "syntactics" (or internal logic) of sign-vehicle sets designed to include data and reliability on the same map (van der Wel, et al., 1994), user interface styles for depicting data reliability (MacEachren, et al., 1993), and impediments to representing data quality (i.e., reliability) (Buttenfield, 1993).

Issues of data model validity (and validity of sign-vehicles associated with those data models) are perhaps best dealt with through the use of multiple representations. Jenks (1967) brought attention to the range of interpretations that might be generated due to data model choice. In relation to GVIS, MacEachren and Ganter (1990, p. 78) suggested that visualization systems should "... permit, perhaps even demand that the user experience data in a variety of modes." Pointing to the advances in technology that make production of multiple representations easy, Monmonier (1991) has taken this idea one step farther by suggesting that our past approach of providing the single "optimal" map should now be considered unethical.

In most exploratory GVIS applications, the representation process has an influence on what can be seen in the display (comparable to that of data model choice). Research on method produced error has demonstrated systematic differences in representations resulting from different methods of interpolation (Morrison, 1971), generalization (McMaster, 1987), classification (Coulson, 1987), etc. With some data manipulation procedures, such as kriging, reliability estimates are part of the result. In these cases, method produced uncertainty can be treated in the same ways as uncertainty in data quality (i.e., using the range of representational techniques mentioned above). An alternative is to adopt the multiple representation approach. This use of multiple displays as a method for signifying spatial reliability has been advocated by Goodchild et al. (1994), and by Fisher (1994b). In the former application, the multiple views are based on a set of side-by-side static representations – with reliability signified by similarity at each location from view to view. In the latter case, a single view is changed dynamically – with stability over time being the sign-vehicle for reliability.

GVIS pragmatics – interpretant validity

A pragmatics of GVIS should address validity of interpretations arrived at through use of visualization tools. Pragmatics deals with the sign-vehicle to interpretant link and puts emphasis on the "meaning" embedded in and brought to the sign. The distinction being made here is between a sign's *denotation* and its *connotation*. A sign's denotation is its explicit meaning; that meaning embedded in the display through the cartographic rules used in display creation (e.g., dark red = warm temperatures and dark blue = cool temperatures). Connotations are the implicit meanings brought to interpretation, meanings that may or may not be anticipated by the person making decisions about the display design (e.g., on a temperature map in which red = warm and blue = cool, the break between blues and reds might be assumed to be meaningful, perhaps a representation of the freezing point).

In considering GVIS denotative meaning, an analogy can be made to statistical analysis. Interpretation of a geographic representations is analogous to statistical hypothesis testing. Interpretations based on visual evidence have a potential for two kinds of error: *seeing wrong* (equivalent to a Type I statistical error) and *not seeing* (equivalent to a Type II error)

(MacEachren and Ganter, 1990). As with the statistical counterparts, the likelihood of Type I and II visualization errors is a function of the degree to which the display is a representative "sample" of possible world views, and of the standards applied for acceptance or rejection of what is seen.

'Seeing wrong' occurs when a sign relation is formed using inappropriate links among the three sign components (e.g., when a sign-vehicle is incorrectly matched to referent and/or interpretant) (figure 3). One common GVIS example involves interpretation of coincidence in space as evidence for functional association.

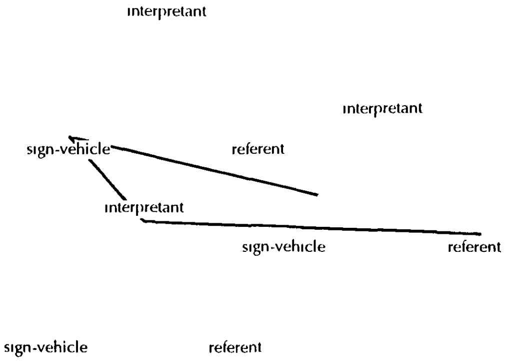


Figure 3. 'seeing wrong' - incorrect links among sign components.

'Not seeing' results when signs are not recognized as such or when sign relations are incomplete (figure 4). In this case there is a failure to recognize a particular component of the display as a sign-vehicle, or a failure to identify the required links to complete the sign relation. With GVIS, such failures are likely to result from a failure to recognize spatial association among subordinate sign-vehicles.

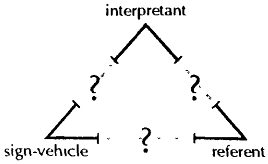


Figure 4. 'not seeing' – a failure to achieve sign component links

There is an obvious standard against which to assess denotative meaning, although it is perhaps difficult to measure. With connotations, however, there is no such standard. Since connotation is 'brought to' the interpretation by the individual, rather than being embedded within the sign relation, there can be no single judgement about the truth of connotative meaning. Connotation associated with GVIS (or other) signs derives from an individual's complex mix of life experience and specialized knowledge. Although connotation is established at an individual level, however, shared life experience leads to some level of shared connotations. Among the most important in relation to issues of truth in GVIS is a *connotation of veracity*, a tendency to assume that 'if it shows up on a map it must be true' (MacEachren, 1995).

TRUTH - BY WHAT STANDARD?

A fundamental flaw in any effort to objectify truth in visualization (or science in general) is that our only standard for truth is what we (science and society) believe to be true – defined by the paradigm within which we are working. Even if we ignore, for the moment, connotative meaning, judgements of "truth" in the semantic and pragmatic components of GVIS must be made against this rather unstable standard. As paradigms change, "scientific truth" also changes.

The strong tendency to believe in established doctrine, and to discount evidence that does not fit the expectations derived from that doctrine, has been pointed to by a number of authors, most notably Kuhn (1970). De Mey (1992) recently introduced the term *cognitive gravity* as a label for the "momentous strength of established conceptual systems." Cognitive gravity can be defined as the pull toward a particular interpretation exerted by the large volume of accumulated facts and rules – and is illustrated by the often strenuous resistance to new ideas (e.g., Copernicus' helio-centric view of the solar system, continental drift, etc.).

To a large extent, cognitive gravity (and its associated regularized patterns of thought) is precisely why science works, and why GVIS tools are so powerful. GVIS facilitates application of our repertoire of learned patterns. These patterns are both visual ones (related to matching what we might see in a display with what we have seen before) and conceptual ones (related to the sequence of views we might elect to look at or the relations we might attempt to find). Applications of these patterns provides the mechanism through which interpretants become part of sign relations.

Using accepted concepts as our gage against which to measure truth, however, has potential negative consequences, particularly in the context of visualization, with its emphasis on a search for unknowns. When the goal is largely that of using the power of vision to let us notice the unusual in an effort to prompt insight, a tendency to disregard the unusual can be a serious impediment to scientific advances.

A consequential example of cognitive gravity associated with visual geographic evidence is provided by the recent discovery of the ozone hole. Initially, the key pattern was not seen in the accepted evidence – NASA produced image-maps from the total ozone mapping spectrometer. Joseph Farman, a British atmospheric scientist was the first to *notice* the ozone hole – because he did not rely exclusively on the visual evidence provided by NASA images (Hall, 1992). For Farman to believe that he had discovered something significant, he had to overcome the "cognitive gravity" of which those images were a part.

Farman was part of a British team that had sensors on the ground. These sensors registered extremely low levels of ozone. The research team initially doubted the measurements (because they did not agree with the images derived by NASA, nor with the assumptions underlying them). Farman and his team checked results by adding a second ground sensor, and after waiting a full year to compare results, the low numbers were finally believed. The paper describing Farman's findings was submitted to *Nature,* where it received the following comment from one reviewer:

> This is impossible! But of course if it's true, we can't wait twenty years to find out that it's true, so publish immediately! (Hall, 1992)

Cognitive gravity nearly prompted the reviewer to discount Farman's evidence in favor of the more compelling and scientifically accepted image maps. It is ironic that maps both delayed discovery of the ozone hole and hastened its acceptance by the general public. The delay resulted because data analysts at NASA decided to 'flag' (and exclude from the images produced) data values that were lower than computer models had predicted – thus only displaying results that were within predetermined expectations. Farman's simple graphs are what ultimately convinced scientists – but the revised digitally produced map (with its connotations of veracity) is what convinced the public.

The tendency to accept what fits our expectations and reject what does not – something that has been called a confirmatory bias – is often coupled with a failure to remember the "respects of similarity" with which representations are linked to the world (i.e., a failure to apply the appropriate interpretant to the sign relation).* Lakoff (1987), for example, contends that chemists analyzing representations produced with a Nuclear Magnetic Resonance device, have come to treat the images as real. Restated in terms of the semiotic framework presented here, the chemists Lakoff refers to seem to treat the sign-vehicle as similar to the referent in *all* respects. It is clear from explanations that accompany isarithmic maps in many earth science and geographic journals that climatologists treat isolines, and their precise position, as equally real. Isolines are assumed to signify not only values at locations, but shape and structure of the underlying phenomenon. GVIS toolkits now allow scientists to use "focusing" to highlight particular isolines and study their position relative to the depiction of a second or third variable (see DiBiase, et al., 1994). We have little conception, however, of the standard that should be applied to judging truth of the interpretations derived.

DISCUSSION

As visual evidence (often facilitated by cartographic tools) becomes increasingly central to the way in which scientists and policy analysts think, we must become increasingly critical of that evidence. The inclusion of data quality specifications in the Spatial Data Transfer Standards, the NCGIA Initiative on Visualization of Data Quality, the efforts by the National Center for Health Statistics to incorporate reliability information in some of their

* see Giere (1988) for a discussion of the "respects of similarity" by which scientific models are linked to the world they "represent."

maps, are all steps in the right direction. Truth in GVIS, however, is more than an issue of data quality or representation reliability. It goes to the heart of the assumptions we make about how the world works and about what is represented in our visual displays. We can devise no absolute measure of truth in GVIS. Truth can only be defined relative to the models or paradigms within which we decide what to represent and how to represent it. If we are interested in issues of truth, we must give as much attention to the implications of these paradigms as we do to the measurement and depiction of validity (as it is judged against a standard determined by the paradigms).

A tradition of critical analysis is developing within cartography. At this point, however, most of the attention of this critique has been directed to the 'communication' corner of my visual tool use–space. Significant contemporary public images such as highway maps (Wood and Fells, 1986) and the Van Sant image–map (Wood, 1992) have been "deconstructed," as have a variety of historical maps and national mapping programs (Harley, 1987; Edney, 1994). Little attention has yet been given to critical analysis of the geographic representations that are becoming increasingly integral to early stages of the research process (see Krygier, 1994 for one such attempt). As the role of cartographers evolves from that of mapmakers (with emphasis on presentation graphics) to that of geo-information facilitators (with emphasis on building tools for interactive data exploration), it is time to take a closer look at the basis for the information we facilitate and at the interaction between the visual display tools we design and the research paradigms they are meant to work within – or break free of.

REFERENCES

Asche, H. and C. M. Herrmann (1994). Designing interactive maps for planning and education. in MacEachren, A. M. and Taylor, D. R. F. (eds.) *Visualization in Modern Cartography.* London, Pergamon. 215-242.

Buttenfield, B. P., Ed. (1993). Representing Data Quality, *Cartographica,* 30 (2): 1-7.

De Mey, M. (1992). *The Cognitive Paradigm.* Chicago: University of Chicago Press.

DiBiase, D., Reeves, C., MacEachren, A., von Wyss, M., Krygier, J., Sloan, J., and Detweiler, M. (1994). Multivariate display of geographic data: Applications in earth system science. in MacEachren, A. M. and Taylor, D. R. F. (eds.) *Visualization in Modern Cartography.* Oxford, UK, Pergamon. pp. 287-312.

Dorling, D. (1992). Stretching space and splicing time: From cartographic animation to interactive visualization. *Cartography and Geographic Information Systems,* 19(4): 215-227.

Edney, M. H. (1993). Cartography without progress: Reinterpreting the nature of historical development of mapmaking. *Cartographica,* 30(1): 54-68.

Fisher, P. (1994a). Hearing the reliability in classified remotely sensed images. *Cartography and Geographic Information Systems,* 21(1): 31-36.

Fisher, P. (1994b). Animation and sound for the visualization of uncertain spatial information. in Hearnshaw, H. M. and Unwin, D. J. (eds.) *Visualization in Geographic Information Systems.* London, John Wiley & Sons. pp. 181-185.

Giere, R. N. (1988). *Explaining Science: A Cognitive Approach.* Chicago, University of Chicago Press.

Goodchild, M. , Chih-Chang, L. and Leung, Y. (1994). Visualizing fuzzy maps, in Hearnshaw, H. M. and Unwin, D. J. (eds.) *Visualization in Geographic Information Systems.* London, John Wiley & Sons. pp. 158-167.

Hall, S. S. (1992). *Mapping the Next Millennium: The Discovery of New Geographies.* New York, Random House.

Harley, J. B. (1988). Maps, knowledge, and power. The Iconography of Landscape: Essays on the symbolic representation, design and use of past environments. Cambridge, Cambridge University Press. 277-311.

Jenks, G. F. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography,* 7: 186-190.

Jenks, G. F. (1973). Visual integration in thematic mapping: fact or fiction? *International Yearbook of Cartography,* 13: 27-34.

Kraak, M.-J. (1993). Cartographic terrain modeling in a three-dimensional GIS environment. *Cartography and Geographic Information Systems,* 20(1): 13-18.

Krygier, J. (forthcoming). *Visualization, Geography, and Derelict Landscapes,* unpublished Ph. D. dissertation, The Pennsylvania State University.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions.* Chicago, University of Chicago Press.

Lakoff, G. (1987). *Woman, Fire, and Dangerous Things: What Categories Reveal about the Mind.* Chicago, University of Chicago Press.

MacEachren, A. M. (1992). Visualizing uncertain information. *Cartographic Perspectives,* (13): 10-19.

MacEachren, A. M. (1994). Visualization in modern cartography: Setting the Agenda. in MacEachren, A. M. and Taylor, D. R. F. (eds.) *Visualization in Modern Cartography.* Oxford, UK, Pergamon. pp. 1-12.

MacEachren, A. M. (1995). *How Maps Work: Representation, Visualization and Design.* Guilford Press.

MacEachren, A. M., Howard, D., von Wyss, M., Askov, D. and Taormino, T. (1993). Visualizing the health of Chesapeake Bay: An uncertain endeavor. *Proceedings, GIS/LIS '93* Minneapolis, MN, 2-4 Nov., 1993, pp. 449-458.

MacEachren, A. M. and J. H. Ganter (1990). A pattern identification approach to cartographic visualization. *Cartographica,* 27(2): 64-81.

McGranaghan, M. (1993). A cartographic view of spatial data quality. *Cartographica,* 30(2 & 3): 8-19.

Monmonier, M. (1991). Ethics and map design: Six strategies for confronting the traditional one-map solution. *Cartographic Perspectives,* (10): 3-8.

Morrison, J. L. (1971). *Method-Produced Error in Isarithmic Mapping, Technical Monograph* CA–5. Washington, DC: American Congress on Surveying and Mapping, Cartography Division.

van der Wel, F. J. M., Hootsman, and Ormeling, F. (1994). Visualization of data quality. in MacEachren, A. M. and Taylor, D. R. F. (eds.) *Visualization in Modern Cartography.* London, Pergamon. pp. 313-331.

Wood, D. (1992). *The Power of Maps.* New York, The Guilford Press.

Wood, D. and J. Fels (1986). Designs on signs / Myth and meaning in maps. *Cartographica,* 23(3): 54-103.

HANDLING IDENTITIES IN SPATIO-TEMPORAL DATABASES

A.A. Roshannejad, W. Kainz
Department of Geoinformatics, ITC
P.O. Box 6, 7500 AA Enschede, The Netherlands
Phone: +31 53 874449, Fax: +31 53 874335
Email: {Roshan, Kainz}@itc.nl

ABSTRACT

Rapid advances in the technology of Geographic Information Systems (GIS) have prompted a wide range of spatially referenced applications. However, conventional GISs are dealing with a static view of a continuously changing real world. Taking a snapshot of the world, can be quite useful for a number of applications, however, many other applications are suffering from the lack of incorporation of time in the GIS analysis. Therefore, the increasing demands for temporal capabilities of GISs have started a stream of research in this field. Undoubtedly, handling time is a direct consequence of handling changes. Therefore, change handling, as well as keeping track of tangible data are the most important conceptual and technical problems in spatio-temporal databases.

1. INTRODUCTION

The most desirable way of dealing with geographical phenomena is to model them as they exist in reality. Therefore, the closer a data model represents real-world phenomena, the more comprehensive it is. In other words, all applications are looking for a data model that is able to provide a better and more real perception of the real-world. At the same time, interesting developments in hard and software as well as increasing demands for more updated geo-data from a variety of spatially referenced applications pose a number of questions for current database technology.

Traditionally, maps are produced as an underlying concept which is "A 2 dimensional graphic image which shows the location of things in relation to the Earth surface at a given time" (Keates 1989). What makes a distinction between GIS and traditional line maps is the way of handling geographical data, how to retrieve data and the manner of analyzing data and extracting required information. One of the important questions, in this regard, is to incorporate time in spatial databases. Besides many conceptual inconsistencies between time and space (which makes this combination rather difficult), in order to reach a successful spatio-temporal database, a number of points need special attention and consideration.

One of the basic and fundamental technical problems is handling dynamic data. In fact, time can only be felt by observing changes in real-world phenomena. Therefore, handling time means handling changes. Without the existence of a change, passing time is meaningless. In a static spatial database, by data we mean the most recent snapshot of the world, while in spatio-temporal databases, the most recent data and their histories are not detachable. Therefore, a strong algorithm must be available to keep track of data histories.

At a first glance it seems that handling object identities is a technical problem, however, since we are dealing with concept of change (to be tracked) the problem turns to have a conceptual nature. Particularly, when the database is going to be used for a wide range of applications (for example nation wide), this becomes clearer, because the definition of change and the criterion to create a new object after an event are not unique.

This paper discusses the importance of object identity in spatio-temporal database design. The second section of the paper describes components of spatio-temporal data. Having explained the data and their components we proceed to define change by means of object components. Section 3 gives a general framework to describe change. Section 4 then distinguishes between object as such and the influence of applications to viewing the object. In section 5 a distinction between concept and representation of geographical objects is drawn. Section 6 looks at some research that has been done till now and, while describing the main problems of previous approaches, explains a new conceptual framework for spatio-temporal database design and defines guidelines to handle object identities in spatio-temporal databases. The paper concludes with a discussion of a different perspective to view objects.

2. COMPONENTS OF SPATIO-TEMPORAL DATA

The most noticeable underlying principle of the real world is its dynamic nature. Geo-referenced phenomena are carrying time as an undecomposable component. In fact, each spatial object is known as a composition of three components, called *spatial, temporal* and *non spatio-temporal* (or *aspatial*). This is why to point at each object correctly, we have to define a "*W triangle*". This triangle is composed of three W's. Each of the vertices of it points to either "*When*" or "*Where*" or "*What*" (figure 1).

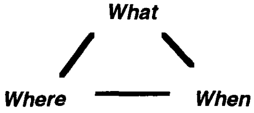


Figure 1. W Triangle

An unambiguous definition of a geographical object has to define these three W's. All the applications (using spatial data) in one way or another, are dealing with geographical object components. Although at the first glance it may be seen that some applications are not interested in all the mentioned components, however, a general and non confusing object identification has to address all three components. In GIS terminology, the geographical object and its components can be illustrated as given in figure 2.

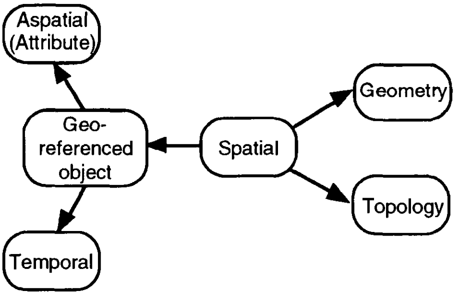


Figure 2. Components of Geo-referenced objects

Among these components, the temporal component is playing a special role. In fact the two others are affected by this one. When we are dealing with changes happening to spatial data, implicitly we are considering the changes in non temporal components. Geometry, topology, as well as attributes of an object are influenced by time. Therefore with a great confidence we can claim that there is no object that is definitely fixed with respect to passage of time. Now, it is clear why time needs to be incorporated in the handling of spatial data, because it is as important as the two other components. As an example, just imagine how the organization of spatial data would look like if we ignore spatial properties and structure of geographic objects, a mass of useless data.

Consideration of time as a component of geographical objects increases the possibility of analyzing objects along the time line and to keep track of their histories. Many applications are referring to time as a valuable criterion. For instance, cadastral applications need to retrieve cadastral data with time stamps. In this application, (for example) time of transferring the ownership of a land can be as important as the spatial attributes of it. Another example is environmental monitoring. Global warming, desertification, deforestation of the rain forest are just few examples of why time is badly required in today's applications. Besides those applications that are dealing with the history of the objects, a considerable number of applications are interested in predicting what will be the next status of the objects, or to determine a trend in the object's development. These applications use the object history to reach the desired goals.

3. WHAT MAKES A CHANGE?

The underlying principle of spatio-temporal databases is the problem of capturing changes in geo-referenced objects. Therefore, at first we have to define what can cause a change in an object. This, in turn, raises the question of what changes are the most desirable ones. This varies from one application to another. A particular application might be interested in aspatial change while the other is dealing with spatial change. Thus the origin of changes has to be identified, first.

As we discussed earlier, a geo-referenced object is a phenomenon that is extended in three directions called *spatial extent*, *temporal extent* and *aspatial extent*. This way of defining an object (as the composition of primitive components) solves many other problems. As long as all the mentioned components are constant in an object, it can be defined or claimed as an unchanged distinguished object with a specific identity. This is the concrete definition of *surrogate* that must, strictly speaking, be unique. As a consequence of this definition, we have to define a new object, with a new object identity, as soon as one of the object's components changes. This is because with respect to the above mentioned definitions, there is a distinction between the object with changed component(s) and the one, that originally holds the object identity. Each of these components is, potentially, subject to change. Some of them, like the temporal component, changes continuously while the others change only irregularly. Since the change in the temporal aspect of a geographical phenomenon is continuous, keeping track of it makes no sense. In fact, considering temporal change leaves us in a paradox, immediately. Therefore, we have to focus on changes of the two other components. Figure 3 gives a view of sources of change.

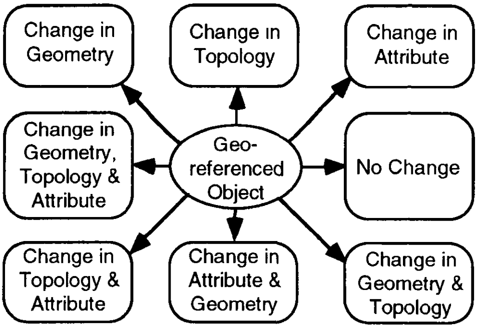

Change in Geometry
Change in Topology
Change in Attribute
Change in Geometry, Topology & Attribute
Geo-referenced Object
No Change
Change in Topology & Attribute
Change in Attribute & Geometry
Change in Geometry & Topology

Figure 3. What can cause a change in a geographical object

This composition can help us to decide which component is the most desirable one for a certain application. Therefore, taking the right component for keeping track of the object's history will be the main issue.

In order to solve this problem, one can assign to the object a specific identity number at the time of creation. This number must be unique in all cases. As soon as a change occurs in one of the components, a new address (or identity number) must be assigned to it. The previous identity number will be kept in a history list and can be accessed through that list. In other words, each object will have a current identity number plus a list of previous identities. The list is arranged sequentially. The identity numbers can be regarded as objects themselves. These identity objects have a value (the current identity number) and also contain two addresses, one refers to the previous object and one refers to the next identity number.

Although this method of decomposing geographical objects into their components seems to be a wise solution, however, since different applications view the same object differently, it will be very difficult to decide if the change in a particular component is capable enough to create a new object. This leads to a conceptual problem of viewing geographical objets that needs to be studied more carefully.

4. DEPENDENCIES OF GEOGRAPHICAL OBJECTS FROM APPLICATIONS

As already mentioned, the definition of geographical objects is very dependent on applications. Therefore, the proposed approach in the previous section has a number of shortcomings in the definition of object creation by changes in its components. On the other hand, and with respect to ever changing components of spatial objects, this method will not be a practical solution to the problem of change tracking. This becomes clearer when we ask questions like *"What kinds of changes in the object's components make a new and distinguishable object?"* or *"When can a particular object at a certain time not be called the same object at the other time?"*.

By asking such questions we are likely to receive as many answers as applications are concerned. One application may regard a new object creation when the object moves few centimeters on the surface of the earth (change in spatial component) while the other is not interested in the precise spatial extent of geographical objects but their aspatial components is more noticeable. Even changing the color of a house (with respect to a particular application) may cause creation of a new object in the database.

In the case that the designed data model is going to serve a wide range of applications, we have to handle all the object component changes. For example the spatial component of a growing city is changing very frequently. At the same time its non spatio-temporal components (such as population) are not constant either. This means having a growing database. Of course this is not what we mean by an optimum approach. But in this example, the ever changing city gives the same perception of the city in people's view. Although, for example, Amsterdam today is different from Amsterdam 50 years ago the city is the same in concept. It seems that in order to define an object we have to deal with the concept and not the object's components. Kuhn (1994) intends to define objects by their own operations and properties instead of their linguistic references. This is exactly how we are approaching from a formal (and low-level) object definition and handling to a more proper way of dealing with objects as a self existent real-world phenomenon.

A great attention, nowadays, has also been paid to the way of communication. The stream of data transfers shows a shift from a fully visualized approach to a more comprehensive and complete communication. This must not be limited to the linguistic areas, but has to be extended to the high level conceptual data modeling. In data modeling, applying this approach means making a distinction between what an object is defined by and what represents the object (the mentioned components). This distinction can be the solution to handling changes, and consequently handling time, in a spatio-temporal database. This way of looking at geographical objects, first of all, avoids expansion of the database due to components' changes. The most important advantage of this approach is to get rid of application views to define geographical objects. This is how handling real-world phenomena should be.

5. CONCEPT VERSUS REPRESENTATION

The discussion about geographical objects and their components in a spatio-temporal database leads us to make a distinction between the concept that every object is defined by, and its different representations. In this respect, e.g., Amsterdam as a concept is a city with certain spatial and aspatial properties at each time stamp and also carries certain non changing characteristics which we call 'Amsterdam' as a concept. This concept is what people refer to when talking about the city of Amsterdam. On the other hand, Amsterdam is extended in three directions, as mentioned, spatial, temporal and aspatial. Each of these components can be used for a certain application and therefore, each of them represents the same object from a different perspective.

Thus if one wishes to obtain a more general view of an object, he/she has to use all the representation components at the same time. Of course, every representation can be used for a different purpose. This is how to perceive the object.

People use to classify objects according to their different representations. For example, all the cities in a certain region, all the 2 or more stories buildings, all the rivers with a certain width and so on. This classification helps to retrieve objects in a more convenient manner, however, creates a dangerous confusion that the objects can be replaced by their representations.

As another example, showing how the traditional way of thinking about the geographical objects limits us, we can refer to the importance of the spatial component of geographical objects in traditional analog line maps. Since in a graphical line map, every feature can be represented by its location (spatial extent), it is implicitly assumed that changing the spatial component of a geographical object causes creation of a new object. Therefore it is very important how to define a geographical object.

In order to more emphasize this topic Kuhn (1994) states that "A general assumption in today's thinking about data modeling for spatial data transfers is that spatial objects get their meaning by linking geometric primitives to feature-attribute catalogues". If we can get rid of object representations in the definition of objects, definitely we will have more freedom to handle objects as they are in the real-world. From this perspective, since it is not the spatial component that defines a geographical object, it does not matter if we find that Amsterdam had another spatial component (this spatial change of course is limited to a certain range) some time ago or some time later in the future, different from a certain snapshot of Amsterdam being used now. Its attributes can be different (at a particular time), but still Amsterdam exists independently of the changes in its components. Clearly an object is something independent of its representations.

6. HANDLING OBJECT IDENTITIES

In the spatial domain, an object has geometric as well as thematic components. When dealing with a static spatial database, any unique identifier given to the spatial object would be quite sufficient to distinguish the objects among each other. In the case of handling spatio-temporal data, as spatial data are subject to change over time, and due to having the possibility of accessing the history of the objects, it is very important to indicate and define the criteria which cause creating a new object and consequently, to define which component of the object has to carry that unique identifier.

This problem becomes critical when splitting or unification of object(s) occur. Obviously, in order to keep track of an object along the time line, one has to give it a unique time-invariant identifier. Clifford and Croker (1988) propose a definition of so-called primitive objects that hold a unique and time-invariant identifier with a non ending lifespan. Therefore, any complex object can have a unique identifier that can easily be constructed from a kind of aggregation of some relevant primitive object identifiers.

This can be applied in aspatial databases. Although this seems to be the only wise solution to the problem of object history tracking, however, in spatial databases, defining such primitive objects is not a straightforward approach. When we deal with the raster domain, this approach – to some extent – can still be applicable (where each pixel can be regarded as a primitive object). Although this seems easy to handle in the raster domain, one has always to consider that the next raster data set will not really register exactly the same as the previous one and furthermore, does not necessarily have the same resolution. Since many spatial databases have a vector-based structure, the problem of defining a unique identifier remains unsolved.

Another solution to this problem is proposed by Worboys (1992). His proposal is based on viewing spatio-temporal objects as finite collections of disjoint right prisms (so called ST-atoms). Bases of these prisms are spatial extents and their heights represent their temporal extents. Figure 4 shows changes that occur to spatial objects at time steps indicated by t_1, t_2 and t_3.

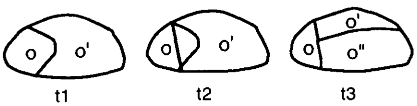


Figure 4. Typical changes occur to spatial objects

The decomposition of spatial objects into ST-atoms is shown in figure 5. Due to Worboy's proposal, the ST-objects O, O' and O" are decomposed as follows:

- O is represented by the collection of two ST-atoms $< S_1,[t_1,t_2] >$ and $< S_2,[t_2,t_\infty] >$.
- O' is represented by the collection of three ST-atoms $< T_1,[t_1,t_2] >$, $< T_2,[t_2,t_3] >$ and $< T_3,[t_3,t_\infty] >$.
- O" is represented by the single ST-atom $< U,[t_3,t_\infty] >$.

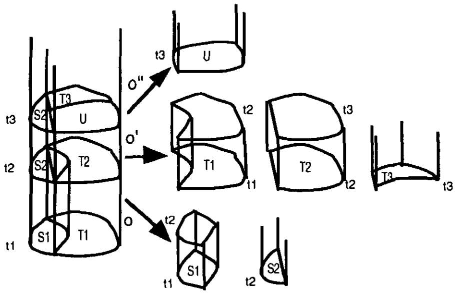


Figure 5. Decomposition of ST-objects to their ST-atoms (after Worboys (1992))

Identification of the objects now becomes confusing. The mentioned decomposition states that T_3 is a new version of T_2 which in turn has been T_1 at time t_1. Considering the spatial extent of T_3, it cannot potentially lead to the spatial extent of T_2 (and the same argument holds for T_2 and T_1). Worse than that is that the history of U is missing. Although U is a completely new object (as far as it is not born from nowhere and always there was something that is modified to make U), ignoring the history of U cannot be accepted. Assume that S_1 and S_2 are bare soil, T_1, T_2 and T_3 are agricultural areas and U is a residential building. Obviously, U is a new object (a building), but as a spatial phenomenon has a history. Simply it has been part of an agricultural area at time t_2.

Another problem can also be observed from the topological point of view. A spatial database must have rigid topological constraints that exactly define, for example, spatial neighbors of an object. This, in turn, creates the problem of maintenance of topological relationships, which is very difficult with the previous approach. These relationships, of course, are not only limited to spatial ones but cover the temporal domain as well. Handling objects as a composition of ST-atoms is not capable enough to keep track of changes in topological relationships. Furthermore, the most important problem in this approach is that ST-atoms are based upon the spatial extent.

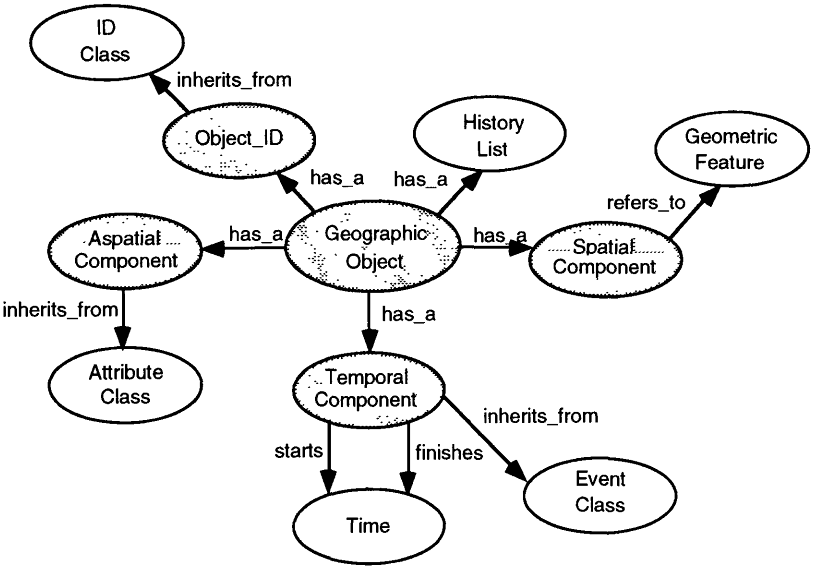


Figure 6. The required structure for defining geographical objects

As it can be seen in figure 6, the identity is attached to the object itself and therefore is independent of object component changes. In terms of object-oriented programming, the following scheme gives a brief description of the structure required to handle geographical objects independently of their components:

```
Geographical_Object
{
Properties:
        Object.Geometry     Refers to geometric features;
         {.....}            Operations on spatial extent of the object
        Object.History      Refers to event class;
         {.....}            Operations on temporal extent of the object
        Object.Attribute    Refers to the list of all possible attributes;
         {.....}            Operations on aspatial extent of the object
}
```

In this structure, the aspatial component inherits from a class called *attribute class*. The attribute class is an enumeration list of all possible attributes that potentially are capable to be attached to the object. The spatial component of the object keeps the geometrical properties of the object and inherits from a more general class that defines *geometric feature*. Geometric feature, in turn, can be defined in a clearer structure which has a proper construction to be able to cover all the required details of handling geometry of the geographic features. It is clear that both aspatial and spatial components are keeping relative histories of the relevant changes in aspatial and spatial properties of the object, respectively. Of course, each status of the object in terms of spatial or aspatial will be recorded in related components.

In this respect, a geographical object has a main body (defined by the concept of the object) and also has two tails, one originating at the spatial and the other at the aspatial component. The temporal component, on the other side, is used for keeping temporal topological relationships. Therefore, object identity, as far as the object's concept is not changed, is unique.

Up to this point, we can define two different types of change. The first type of change is called "*Constructive Change*" and the second type can be named "*Non-constructive Change*". As from the name of these two changes we can understand, constructive change is the change that creates a new object out of an existing one with a new concept different from the previous one. A non-constructive change is one that only changes a certain status of the object in terms of spatial or aspatial components. In non-constructive change, the object identity remains unchanged and only a new part is appended to the end of tail(s) of the object. Obviously, in order to distinguish between every piece of the tail and to be able to keep track of non-constructive changes, all those pieces of the tail(s) are attributed by a relevant time stamp. By this way, if analyzing history of non-constructive changes of the object is required, the mentioned tails can be searched easily.

The object identity of the object inherits from an *ID-class*. In this class all the history of constructive changes is kept. A rough structure of identity can be drawn as:

```
Type identity
value identity_number

Attribute
        identity : Previous_identity_number
        identity : Next_identity_number
End identity;
```

For an implementation of this approach object-oriented tools must be used, because many features such as aggregation, association and multiple inheritance are required. This is exactly what we expected from the definition of the conceptual framework. It is based upon an object-oriented modeling perspective. The base is objects and they are self-existent phenomena. The concept defines the object and not specific representations of the objects.

7. CONCLUSION

The approach described in this paper serves as a basis for a multi purpose data model required for a wide spectrum of applications. The most important advantage of this approach is that it reduces the required change capacity for object histories and that a quick expansion of the database will be avoided.

Viewing geographical objects in the manner described in this paper helps to handle object identities in a more proper and logical way, however, the importance of this approach is not limited to the topic of handling object identities in spatio-temporal databases. In fact, what this approach of perception of geographical objects can do is to get rid of representations of objects as criteria to define the objects. This concept, is a key element in approaching dynamic GIS. Existing GISs are suffering from modeling moving objects, only because the base for definition of geographical objects is the spatial extent of them. This shortcoming of the current GISs can be resolved by this approach. As this paper is part of an ongoing research, the results of applying this approach will be reported in forthcoming international events.

8. REFERENCES

Clifford, J. and Croker, A., 1988, Objects in Time, *IEEE Data Engineering*, Vol. 7, No. 4, pp. 189-196.

Keates, J.S., 1989, *Cartographic Design and Production*, Second edition, Longman.

Kuhn, W., 1994, Defining Semantics for Spatial Data Transfers, *Proceedings of the 6th International Symposium on Spatial Data Handling*, Edinburgh, UK, Vol. 2, pp. 973-987.

Worboys, M.F., 1992, Object-Oriented Models of Spatio-Temporal Information, *Proceedings of GIS/LIS'92*, San Jose, California, USA, pp. 825-834.

Representation of Continuous Fields

Andrej Včkovski
University of Zürich
Winterthurerstr. 190, CH-8057 Zürich, Switzerland
vckovski@gis.geogr.unizh.ch

Abstract

This contribution discusses the representation of continuous fields within Geographical Information Systems. The fields considered are fields of physical properties (observables) defined for every point of a spatio-temporal region. The two major problems discussed are the *inherent* uncertainty present in the samples of the field (field values are measurements) and implications of the discretization when continuous fields are sampled. Fields are usually measured at a finite set of points or regions in space and time. The subsequent use of such data sets, however, often requires field values at other, unsampled locations, or with different temporal and/or spatial aggregation and unit systems. We present the concept of Virtual Data Sets (VDS) which helps to overcome such incompatibility problems. A VDS incorporates the necessary semantics of the data thereby allowing such transformations (i.e., interpolations, aggregations, unit transformations) to be performed transparently and within the domain of the data set. This is achieved by defining a common interface layer which lets the user query a VDS within the GIS and other client applications in a more general way than current GIS allow.

1 Introduction

Due to their strength in analysis and visualization of spatial data Geographical Information Systems (GIS) are frequently used for scientific information processing. Especially research in the broad field of the earth sciences (e.g., oceanography, atmospheric physics, geology, seismology) involves large spatial data sets and their integrated analysis. These data sets usually consist of a set of aggregated samplings (macro data) of natural phenomena (e.g., temperature), where every value is somehow related to a region or subset of space and time, i.e. related to a *spatio-temporal object* which we call *index* for a particular value. Those objects may be points (i.e., samplings at a given point in space and time), squares of a regular partition of a subspace (i.e., pixels of a satellite image) or other objects with a (temporal and spatial) geometry that is usually owing to the sampling and/or preprocessing stage. The samplings together with the related indices are an estimate of the corresponding field, i.e., an estimate of the physical property as a function of space and time.

There is a growing need to use and re-use such data sets for various applications, usually involving many different data sets for specific analyses, simulations and other scientific work. Previous work already dealt with issues of the representation of continuous fields and appropriate data models (Kemp, 1993), (Laurini & Pariente, 1994).

As mentioned before, in most cases data sets entering such integrated analysis projects are already aggregated to some degree, i.e., they represent *macro data* as opposed to *micro data* ("raw" sampling values). The aggregation or postprocessing might be hidden within the sampling apparatus (e.g., temporal aggregation during the exposure time, and spatial aggregation over a pixel in a satellite sensor) or performed explicitly by the data collection organization (e.g., temporal averages over a time period). The aggregation leads to non-point indices[1] one the one hand, since the aggregation is some integral over space and/or time. On the other hand, the aggregation helps the data collectors to assess some information about the quality of the data, e.g., magnitude of measurement errors.

Therefore, a typical data set consists of a set of measurements being expressed as a set of mean values and standard deviations, for example, and a corresponding set of indices. When such data sets are used as an estimate of the corresponding field there are usually two major problems:

- Field values are needed for indices that are not available in the given set of indices. This means simply that one often needs values at "unsampled locations".
- It is non-trivial to include the usually[2] available information about the measurement errors and other sources of uncertainties when using the data for further computations and analyses. This is especially a problem when data are used in multi-disciplinary projects by scientists from domains other than the data collectors'.

Here, an approach is presented to enhance reliability and usability of computer based scientific work when using data that represent fields. The next section will introduce some notation and give definitions of some important terms. Section 3 then gives a short overview of the digital representation of uncertainties (e.g., non-exact values), which is one of the basic building-stones of the Virtual Data Sets presented in section 4. This is followed by a short outlook for our future work.

2 Notation

A *field* is a physical property or observable $z(s)$ that is spatially and temporally indexed (index s). For the sake of simplicity we will just consider properties where the value domain $\mathbb{B}$ is real, i.e. $z(s) \in \mathbb{R}^N$. The field values might be scalars ($N = 1$, e.g., temperature) or vectors ($N > 1$, e.g., wind speed). We will restrict the following discussion to scalars (or 1-dimensional vectors), but the generalization to $N > 1$ should not pose any difficulties. An ideal (undisturbed and deterministic) field is a

[1] There are also cases where the indices can be approximated by points, e.g., measurements of stationary (in time) phenomena. The temporal dimension has not to be considered then. A digital terrain model (DTM) for example represents the (approximately stationary) terrain height field (THF) and its indices usually do not include temporal information.

[2] One might argue, that the "quality information" is exactly the type of information usually *missing*. I think, that information of data quality in general and measurement errors in particular are available in most cases, but this information often gets lost on the way from the data collectors to the data users. This is partly due to missing capabilities of the infrastructure for data exchange and data processing (e.g., data exchange formats, features of the GIS systems used) and partly to some user's ignorance about the importance of error information.

function $z(\cdot)$ which relates every value s of an index domain $\mathbb{I} \subset \mathbb{R}^M$ to a value $z(s)$ from the value domain $\mathbb{B} \subset \mathbb{R}$:

$$s \xrightarrow{z} z(s), \quad s \in \mathbb{I} \subset \mathbb{R}^M, \quad z(s) \in \mathbb{B} \subset \mathbb{R} \tag{1}$$

This model of a field has to be enhanced to allow for uncertainties owing to inherent indeterminism of the physical property and measurement-induced uncertainties. Let $Z(s)$ be such an enhanced description of a field. If the uncertainties can be modeled using probabilistic concepts, then $Z(s)$ is a random variable for every s. $Z(s)$ is sometimes called a *random function* (Isaacs & Srivastava, 1993). Depending on the nature of the phenomenon under consideration $Z(s)$ might also be an interval, fuzzy set, or some other means of modelling uncertainties. These objects usually can be described or approximated by a set of real numbers which are the (estimated) values of corresponding functionals (mappings to $\mathbb{R}$) $\phi(\cdot)$. For random variables, typical functionals are the *expectation value*[3] ϕ_E and the *variance* ϕ_V.

A data set describing such a field consists of metadata and a set of index-value tuples:

$$\mathcal{D} = \{M, \{s_i; z_{i,1}, \ldots, z_{i,n}\}_1^n\}, \quad s_i \subset \mathbb{I}, \quad z_{i,j} \in \mathbb{R} \tag{2}$$

For every index s_i the values $z_{i,1}, \ldots, z_{i,n}$ describe the field at s_i. A typical example might be that $n = 2$ and $z_{i,1}$ is a mean value and $z_{i,2}$ the standard deviation of the field over s_i. M is the metadata describing $\mathbb{I}$ and $\mathbb{B}$, the corresponding unit systems (i.e., units and coordinate systems including appropriate metrics) and any other metadata necessary and available.

It is important to analyze and understand how the values $z_{i,j}$ of a data set are related to the field $Z(s)$. This relation – which mathematically is the relation between s_i and $z_{i,j}$ – consists of two parts:

- The relation between the index s_i and the field $Z(s)$: Since s_i is a non-point set in many cases (i.e., a spatio-temporal region) the values $z_{i,1}, \ldots, z_{i,n}$ in the data set corresponding to s_i are the result of some aggregation A_{s_i} of $Z(s)$ over s_i:

$$\hat{Z}(s_i) = A_{s_i}(Z(s)) \tag{3}$$

  A_{s_i} might be an average over s_i, the value in the "center" of s_i etc.

- The relation between $\hat{Z}(s_i)$ and the data set value $z_{i,j}$: This is basically the fore mentioned functional ϕ_i yielding $z_{i,j} = \phi_j\left(\hat{Z}(s_i)\right)$.

The s_i's are therefore related to the $z_{i,j}$ through $A_{s_i} \circ \phi_j$:

$$s_i \xrightarrow{A_{s_i} \circ \phi_j} z_{i,j} \quad z_{i,j} = \phi_j(A_{s_i}(Z(s))) \tag{4}$$

In order to use a data set $\mathcal{D}$ and to understand its values it is necessary to have an idea about the mappings ϕ_j and A_{s_i}. It would be optimal to know its inverse since this would allow determination of the field values from the samplings.

Mathematical system theory calls the process $A_{s_i} \circ \phi_j$ which the *behaviour* of the system under consideration (Kalman, 1982). $A_{s_i} \circ \phi_j$ depends on the whole measurement process and contains all the transformations involved when measuring a field $Z(s)$ (e.g., measurement apparatus, preprocessing). For the analysis of field measurements, i.e., the reconstruction of the field, the inverse $(A_{s_i} \circ \phi_j)^{-1}$ has to be approximated, i.e., modelled.

[3] To be rigorous one would have to distinguish between those functionals and their estimates. The notation in the following will be a bit sloppy where the interpretation should be clear from the context.

3 Digital Representation of Uncertainties

The previous section has shown that physical properties affected by some uncertainty are usually characterized with a small set of real numbers, e.g., a mean value and a standard deviation. The way to describe uncertain values depends strongly on the nature of the sampling process and the phenomenon investigated. It is, therefore, impossible to define one single way to describe uncertain values applicable to all cases. A digital representation of uncertain information should meet the following requirements:

- Digital encoding means mapping to real numbers[4]. It is desirable that the representation uses a small set of numbers.
- Operators and operations for uncertain values should be defined. For instance, the arithmetic operators $\diamond \in \{+, -, \times, \div\}$ should be availabe along with standard functions (e.g., trigonometric functions).
- It should be possible to convert one representation to another since different data sets and their values will often use different representations.
- A suitable set of functionals (mappings to $\mathbb{R}$) should be available, e.g., $\inf(\cdot)$ (lower bound), $\sup(\cdot)$ (upper bound), $\alpha_i(\cdot)$ (i-th moment), $\beta_i(\cdot)$ (i-th central moment).
- If the representation is based on probability theory it should support *Monte Carlo simulations* (Johnson, 1987). The representation of a random variable A should be able to generate pseudo-random realizations a_i. Actually, this is just another type of functional $\phi(\cdot)$ which we call $\mathrm{rnd}(\cdot)$.[5]

One of the basic decisions when choosing a suitable model for an uncertain value is whether it can be modelled with *probabilistic concepts.* In most cases it will be the primary choice to use probability theory to describe uncertainty within scientific results. Since its formalization (Kolmogorov, 1933) probability theory is quite well understood and a lot of methodology has been developed. Especially mathematical statistics has benefited from this framework and has produced many useful techniques and methods for the description and analysis of data. The random-ness of a property (random variable X) is defined within probability theory with a corresponding probability distribution $p(x)$. Three variants were selected to describe a probability distribution $p(x)$:

Empirical moments: $\alpha_k = \left\langle \int x^k p(x)dx \right\rangle$ and $\beta_k = \left\langle \int (x-\alpha_1)^k p(x)dx \right\rangle$, usually α_1 (mean value) and β_2 (standard deviation).

Empirical distribution function or histogram: The distribution $p(x)$ is described with a set of quantiles, i.e., (x_i, p_i) with $P(X \leq x_i) \approx p_i$.

[4] In fact, digital encoding means mapping to integer numbers. There are, however, means to represent a finite, countable subset of the real numbers with integer numbers (i.e., *floating point numbers*).

[5] The functional $\mathrm{rnd}(\cdot)$ actually has another "hidden" parameter (sequence number) which identifies the realization requested. During a Monte Carlo simulation run, a specific realization might be requested several times so that it is necessary that the value is always the same during one simulation step. Consider for example an expression $C = A + AB$. The simulation would calculate $c_i = \mathrm{rnd}(A) + \mathrm{rnd}(A)\mathrm{rnd}(B)$. The second $\mathrm{rnd}(A)$ must have the same value as the first one

Representation	para-	inf(·)	$\alpha_l(\cdot)$	rnd(·)	$+,-,$	Standard-
	meters	sup(·)	$\beta_k(\cdot)$		$\times,\div$	functions
Interval	2	●	○	○	●	●
Fuzzy Set	⋆	●	○	○	●	●
α_i,β_j,$i \leq n$,$j \leq m$	$n+m$	●	⋆	●	⋆	⋆
Histogram (K classes)	3K	●	●	●	●	●
Uniform distribution	2	●	●	●	○	○
Normal distribution	2	○	●	●	$+,-$	○
Other distributions	⋆	⋆	●	●	⋆	⋆

● = available, ○ = not available, ⋆ = variable.

Table 1: Properties of different representation types for uncertain values

Parametric distributions: A distribution type may be determined when the micro data are aggregated, i.e., the empirical distribution is approximated with a standard distribution and its parameters. Typical parametric distributions are the normal distribution (parameters μ, σ^2), uniform distribution (a, b) or Weibull[6] distribution (p, γ). Sometimes the parameters of a distribution correspond to some moments, e.g., for the normal distribution $\mu = \alpha_1$ and $\sigma^2 = \beta_2$. It is, however, different to describe a distribution solely by moments than by distribution type and a set of parameters.

It is not always suitable to apply probability theory to describe uncertainty. Therefore, two other non-probabilistic ways to describe uncertain values were included:

Intervals: Intervals define lower and upper bounds for a value and are very convenient due to their simplicity. An important advantage are the simple rules for computations using interval values (e.g, (Moore, 1966), (Bauch *et al.*, 1987), (Mayer, 1989),(Moore, 1992), (Polyak *et al.*, 1992)).

Fuzzy sets: Fuzzy sets (Zadeh, 1965) may be seen as an extension to intervals. The basic idea is to define a "degree of membership" for a number in a (fuzzy) set. This is in contrast to intervals where a number is either within the interval or not.

While intervals are a simple yet powerful way to deal with uncertain real-valued data, fuzzy sets have not yet been used widely to describe scientific data[7]. despite the attention it had in the last decades. The major criticism of fuzzy sets is the usually subjective assignment of membership degrees (definition of set membership functions) which often is not acceptable in scientific work where objectivity is an important issue.

Table 1 summarizes the properties of these representations.

[6] Most often the special case with $p = 1$ (exponential distribution) is used.

[7] However, for some applications see (Bandemer & Näther, 1992)

4 Virtual Data Set

4.1 Requirements

The previous section has introduced different methods to digitally represent uncertain values (i.e., physical properties from the real world). In this section an architecture is presented that is suitable for the representation of *fields* consisting of uncertain values.

Although the discussion here is rather theoretical, the final goal is to have an *actual* application. The properties of a representation, therefore, are strongly influenced (or even determined) by the user's requirements. The typical information the user wants from a data set $\mathcal{D} = \{M, \{s_i; z_{i,1}, \ldots, z_{i,m}\}_1^n\}$ describing a field are

Query type A Retrieve the various components of the data set, e.g. the metadata (M), the set of indices $\{s_i\}$ and for every index s_i the related set of parameters describing the field value $z_{i,1}, \ldots, z_{i,m}$. This type of query is available in every system. It simply retrieves available data values.

Query type B Estimate $\phi(Z(s))$, where ϕ is some functional (e.g., mean value) and $s \in \mathbb{D}$ is an "unsampled location", i.e. $s \notin \{s_i\}$. ϕ might be one of the functionals $\phi_1, \ldots, \phi_m$ that define the set of parameters available for every s_i (i.e., $z_{i,j} = \phi_j(Z(s_i))$, so that it is not necessary to transform the representation as shown in the last section. This type is probably one of the most important queries when working with fields.

Query type C An advanced information request is to query the field for the spatial references where the field value has a certain value, i.e., for a specific z_0 and $\phi(\cdot)$, solve the equation $\phi(Z(s)) = z_0$ for s. Consider a digital terrain model, where $\mathbb{D} \subset \mathbb{R}^2$, all s_i are points, every field value (height of terrain) is given by one parameter which is the mean value (i.e., $m = 1$ and $z_{i,1}$ is the mean height of the terrain at s_i). Then, the solution would "compute the set of points s where the mean terrain height is z_0". This type of query is basically an inversion of the field function $Z(s)$.

In the context of this paper we will not discuss information requests of type C (inversion of field function). Instead, we will focus on queries of type A and B. In principle, type A queries can be answered by most of the current GIS and related systems.[8]

Type B queries typically need much more user interaction. The procedure to compute $\phi(Z(s))$ at a location s is not trivial and involves sometimes very complicated computations. Even if $\phi(\cdot)$ corresponds to one of the parameters available in $\mathcal{D}$, i.e. $\phi(\cdot) = \phi_j(\cdot)$, the estimation of the field value parameter $\phi_j(Z(s))$ needs sophisticated inter- and extrapolation methods.

A common approach to handle type B queries is to transform them into type A queries. This means that a set of locations s_i' (and a set of functionals $\phi'(\cdot)$ at which the field might be queried is determined beforehand. The data set $\mathcal{D}$ is then transformed to a new data set $\mathcal{D}' = \left\{M', \{s_i'; z_{i',1}, \ldots, z_{i,m'}\}_1^{n'}\right\}$ so that every query for $\phi'(Z(s_i'))$ will be a type A query. There are, however, many situations where this approach fails due to the long life cycle of some data sets. The transformation

[8] It is not always eays to obtain all of the metadata (M) or the set of indices $(\{s_i\})$ directly, e.g., the coordinates of the pixel cells of a satellite image

to a new data set $\mathcal{D}'$ might not satisfy future queries, so that $\mathcal{D}$ or $\mathcal{D}'$ have to be transformed again into a new data set $\mathcal{D}''$ that conforms to the new requirements. Sometimes the initial data set $\mathcal{D}$ is not available anymore, so that $\mathcal{D}''$ has to be computed from $\mathcal{D}'$. Every transformation $\mathcal{D} \rightarrow \mathcal{D}'$ usually affects the quality of the data, the quality rarely ever increases.

Another problem that contributes to the loss of quality are the methods used to transform $\mathcal{D}$ to $\mathcal{D}'$. Most often, these methods, i.e. the computation of $\phi(Z(s))$ given $\{s_i\}$ and the related $\{z_{i,j}\}$, are not trivial and need a lot of expert knowledge from the domain scientists and data collectors, respectively. When a field is sampled this expert knowledge is available in the data collecting organization. At a later stage, e.g., when transforming from $\mathcal{D}^{(ix)} \rightarrow \mathcal{D}^{(x)}$, this expert knowledge often is not available anymore. It is therefore the user's own choice and responsibility *how* to transform $\mathcal{D}^{(ix)} \rightarrow \mathcal{D}^{(x)}$.

4.2 Concept

The concept of the Virtual Data Set (VDS) described in (Stephan *et al.*, 1993) and (Bucher *et al.*, 1994) tries to minimize quality loss due to subsequent transformations, while maintaining usability of a data set at the same time. The main idea is to allow for queries of type B. A data set $\mathcal{D}$ is therefore enhanced with information P needed to process queries of type B.

$$\mathcal{V} = \{P, \mathcal{D}\} = \{P, M, \{s_i; z_{i,1}, \ldots, z_{i,m}\}_1^n\} \tag{5}$$

An approach would be to formalize the necessary information so that P includes standardized specifications for the answer of type B queries, e.g. interpolation method to use, parameters of the methods, uncertainty modelling methods used. For many data sets this approach would fail, for example because the interpolation method used has some very specific constraints which where not taken into consideration when P's content was formalized.

VDS uses another approach. P is *procedural information* in the sense that it defines the entire method to process and answer queries of type A, type B (and probably type C). In the terminology of object-oriented design (Booch, 1991) P is the *behaviour* of object $\mathcal{V}$ and $\mathcal{D}$ its *state.* The term "virtual" emphasizes the VDS's capability to process queries of type B, i.e., queries that return values not stored on secondary storage.

This VDS-concept is similar to other approaches which try to enhance reuseability of expensive data sets using object-oriented concepts (e.g., the OGIS project (Buehler, 1994)). A computer industry term for P could be *middle-ware*, giving a client standardized access to the data ($\mathcal{D}$).

The next section outlines some implementation and design issues that have been considered as first experiments.

4.3 Implementation

The current design for the implementation of VDS uses a *client-server* architecture. Data requestors (e.g., a GIS application) are clients; a VDS $\mathcal{V}$ or its procedural part P, respectively, is a server. Communication between the clients and servers is based on messages which are queries of type A or B and their corresponding replies. Whenever a client needs data it sends a request to a VDS (server), where the request is

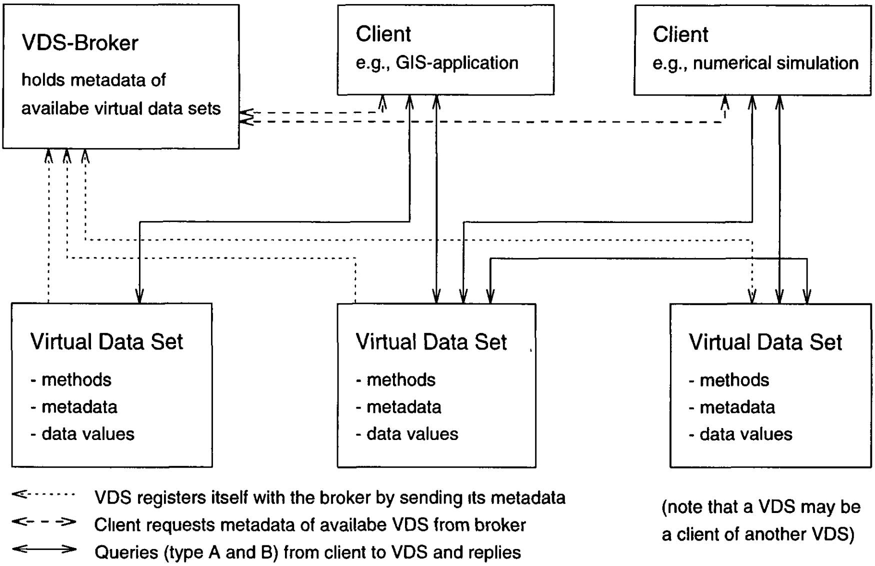

VDS-Broker
holds metadata of
availabe virtual data sets
Client
e.g., GIS-application
Client
e.g., numerical simulation
Virtual Data Set
- methods
- metadata
- data values
Virtual Data Set
- methods
- metadata
- data values
Virtual Data Set
- methods
- metadata
- data values
VDS registers itself with the broker by sending its metadata
Client requests metadata of availabe VDS from broker
Queries (type A and B) from client to VDS and replies
(note that a VDS may be a client of another VDS)

Figure 1: Communication between virtual data sets, broker and clients

processed and the results are returned to the client. The processing on the server-side (the component P of $\mathcal{V}$) might just read the data requested from secondary storage (i.e., type A query), or read data and apply some interpolation methods (i.e., type B query). In addition to clients and VDS-servers there is a specialized server we call *VDS-broker*. The broker is a metadata-base that holds the information of the VDS available[9]. Figure 1 shows the different components involved and the communication between them.

As soon as a VDS becomes available (e.g., "runs" somewhere or is initialized) it sends its metadata M to the broker in a specified format, i.e., the VDS registers itself with the broker. Whenever a client needs some data it may look up if the requested VDS is available and how to access it (i.e., sending a corresponding request to the broker). Once the client knows the "address" of a VDS it sends future requests directly to the VDS. This structure allows the client and server parts to run on different computers, leading to a distributed computing environment. A message sent to a server might be a *local function call* or a "real" message sent over a *network*. Since the queries of type B might include a specific functional $\phi(\cdot)$ a VDS needs to be able to handle different types for representing uncertainty. While the way a VDS handles the data $\mathcal{D}$ internally is not relevant to a client, the results of queries processed by P are. Therefore, the creation of the P-part of a VDS is simplified by a common (class) library which covers the following functionality:

- low-level details for the communication to and from a client or server, respectively

[9] Its existence is technically motivated. It simplifies the communication from clients to servers because clients can easily look up exisiting servers through the broker.

- construction of uncertain values and conversion between different representations, evaluation of arithmetic expressions and standard functions involving uncertain values
- a collection of general purpose interpolation methods and associated helper functions (e.g., spatial neighbour search, etc.).

5 Conclusion And Outlook

This contribution has presented an overview on the questions involved when dealing with data sets that represent continuous fields. The two major issues were:

- The data available is *never* accurate. It is *inherently* necessary to use techniques for modelling uncertainty and errors.
- The field values often are queried for other locations, unit systems, types of aggregation, etc., than available within the data set.

The concept of the Virtual Data Set tries to approach those problems just by the definition of a clear interface for data access. In fact, it does not *solve* the problems but defines *where* the problems should be solved. It is nonetheless a concept that can be translated into a real system as first prototyping approaches have shown. We believe that the more complex systems and applications get, the more important interoperability questions become. The slowly emerging trend in recent years is towards systems built up from a custom set of cooperative modules or services. It has been shown that this is a way to reduce complexity and benefits both users and system providers. The former will have customized solutions instead of large and monolithic ones and the latter can increase the system quality delivered since they systems are more easily maintainable.

Our future work consists of more general prototype implementations and the selection and adoption of suitable industry or scienctific standards for the implementation of distributed, modular systems (e.g., OMG CORBA (CORBA, 1992), OGIS (Buehler, 1994), COSIMA (De Lorenzi & Wolf, 1993)).

References

Bandemer, Hans, & Näther, Wolfgang. 1992. *Fuzzy Data Analysis.* Kluwer Academic Publishers.

Bauch, H., Jahn, K.U., Oelschlägel, D., Süsse, H., & Wiebigke, V. 1987. *Intervallmathematik.* BSB Teubner, Leipzig.

Booch, Gary. 1991. *Object Oriented Design with Applications.* The Benjamin / Cummings Publishing Company, Inc.

Bucher, Felix, Stephan, Eva-Maria, & Včkovski, Andrej. 1994. Integrated Analysis and Standardization in GIS. *In: Proceedings of the EGIS'94 Conference.*

Buehler, K. A. 1994. *The Open Geodata Interoperability Specification: Draft Base Document.* Tech. rept. OGIS Project Document 94-025. OGIS, Ltd.

CORBA. 1992. *The Common Object Request Broker: Architecture and Specification.* Object Management Group. OMG Document Number 91.12.1.

De Lorenzi, Michele, & Wolf, Andreas. 1993. A Protocol for Cooperative Spatial Information Managers. *In: Workshop on Interoperabilty of Database Systems and Database Applications.*

Isaacs, E.H., & Srivastava, M.R. 1993. *An Introduction to Applied Geostatistics.* Oxford University Press, New York.

Johnson, Mark E. 1987. *Multivariate statistical simulation.* Wiley series in probability and mathematical statistics. Applied probability and statistics. John Wiley and Sons, New York.

Kalman, Rudolf E. 1982. Identification from real data. *Pages 161–196 of:* Hazewinkel, M., & Kan, A. H. G. Rinnoy (eds), *Current Developments in the Interface: Economics, Econometrics, Mathematics.* D. Reidel Publishing Company, Holland.

Kemp, Karen K. 1993. *Environmental modelling with GIS: A Strategy for Dealing With Spatial Continuity.* Ph.D. thesis, NCGIA.

Kolmogorov, A. N. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Springer Verlag.

Laurini, R., & Pariente, D. 1994. Towards a Field-oriented language: first specifications. *In: Papers presented at the GISDATA Workshop on Spatial Objects with indetermined boundaries.*

Mayer, G. 1989. Grundbegriffe der Intervallrechnung. *In:* Kulisch, U. (ed), *Wissenschaftliches Rechnen mit Ergebnisverifikation.* Akademieverlag.

Moore, Ramon. 1966. *Interval Analysis.* Englewood Cliffs, N.J.: Prentice Hall.

Moore, Ramon. 1992. Parameter sets for bounded-error data. *Mathematics and Computers in Simulation*, **34**, 113–119.

Polyak, B., Scherbakov, P., & Shmulyian, S. 1992. Circular Arithemtic and Its Applications in Robustness Analysis. *Pages 229–243 of:* Kurzhanski, Alexander B., & Veliov, Vladimir M. (eds), *Modeling Techniques for Uncertain Systems.* Basel: Birkhäuser.

Stephan, Eva-Maria, Včkovski, Andrej, & Bucher, Felix. 1993. Virtual Data Set: An Approach for the Integration of Incompatible Data. *Pages 93–102 of: Proceedings of the AUTOCARTO 11 Conference.*

Zadeh, L. A. 1965. Fuzzy Sets. *Information and Control*, **8**, 338–353.

BUILDING AN OOGIS PROTOTYPE : EXPERIMENTS WITH GEO$_2$

Laurent RAYNAL, Benoît DAVID, Guylaine SCHORTER
IGN / COGIT
2, Av. Pasteur, 94160 Saint-Mandé (FRANCE)
{raynal,david,schorter}@cogit.ign.fr

ABSTRACT

In 1991, an experiment began at IGN at the COGIT laboratory. Its objective was to examine the potential of object-oriented technhology in manipulating and modelling geographical objects. From an experimental point of view, a prototype has been developed on top of a commercial DBMS, namely O$_2$, from O$_2$Technology. The prototype has been called GeO$_2$. This article gives an account of the most prominant experiments that have been carried out at this stage of development. We concentrate on the management of large volume of data and look at processes which could be affected by it : loading, visualizing, accessing to (through a spatial index). A test of portability on C++ storage system is also mentioned. Then results of our experiments allow us to shed new light on object-oriented technology.

INTRODUCTION

IGN-F (the French National Geographic Institute) deals with a large volume of digital geographic information, for example, one data set from the BD Topo® (BDTopo 1992) (a database roughly equivalent to a 1:25,000 scale map), corresponding to an area of 20x28 km^2, represents as many as 60 Megabytes (in ASCII files). The total surface of France is a thousand times larger than this extract.

Furthermore, geographic data is complex as it combines both geometric components and semantic characteristics. Indeed, in handling geometric primitives (points, lines, polygons) it is necessary to describe geographic notions (reference system, projection system, coordinates), to model primitives and often to manage topological relationships (proximity relationships between primitives). However, geometric data handling generates processes that are algorithmically complex and CPU intensive and it puts high demands on the capabilities of computers. Display data on a map with a legend is a simple example of such a process (display order, definition of the legend according to the scale, numerous points to display, etc...).

In the face of these requirements, it appears that standard relational DBMS are not able to manage geographical information efficiently (Frank 1984). Object-Oriented DBMS seem more suitable for geographic data management than relational DBMS. Indeed, OO DBMS is needed most for managing the geometric primitives. However, in order to develop a GIS from an OO DBMS some reliable modules such as spatial indexes, computational geometry operators or topological structures are also necessary. Adding these modules should not require an unacceptable increase in the cost (memory and computation time) of data management. Otherwise, one could not speak of a geographical DBMS as a geographical extension of a DBMS.

Adding new types and associated operations, referencing objects, are here the basic capabilities a user needs in a DBMS or a GIS. An experiment began in 1991 at IGN at the COGIT laboratory to examine the potential of object-oriented technology in manipulating geographical objects. This aim was attained through the implementation of a prototype using a commercial DBMS, namely O_2 (O2 1991), from O_2Technology. This prototype, called GeO_2 (David 1993a), is the core toolbox for a GIS, that means GeO_2 is merely composed of the necessary functions of a GIS and is not designed to meet specific application requirements.

This article will describe the most prominant experiments that have been carried out at this stage of development. In a nutshell, in each experiment, we will briefly focus on activities and performances obtained (section 2). Finally, these experiments will form the basis for estimating the capabilities offered by OO data model, OO languages and OO DBMS (section 3).

EXPERIMENTS : PRACTICAL USE AND PERFORMANCES

Four topics are reviewed : a study on spatial indexes, a data visualization module, a generic data loading module and a portage in a C++/Versant environment. The geographic data model of GeO_2 has already been presented in (David 1993a).

Spatial Indexes

When voluminous geographical databases are handled, most queries are spatial and accessing, scanning data may become a laborious and unrewarding task for the user. Then, specific mechanisms to get access to data quickly, introducing spatial ordering among data, must be added. A spatial index has this function. Three points are extracted from this experiment : the simplicity of our implementation, which allows easy use and easy extension of the module, the variation of window query results according to data sets, which is illustrated with the R*-Tree, one spatial index designed to fit with spatial objects' distribution and the influence of OODBMS in the R*-Tree construction process.

A module for spatial indexes, built on top of O_2, is implemented through a very simple schema so that testing, adding or changing a spatial index can be done very easily in the process of manipulating real geographic data. One generic (or virtual) class called SpatialIndex has been introduced. The following four actions are allowed on it :

insert() : to insert one object in the index (to "index" an object),

delete() : to remove one object from the index,

locate() : to make a point query (objects located on the query point are reported),

search() : to make a window query (objects intersecting the query window are reported).

Three spatial indexes (i.e. a R*-Tree (Beckmann 90), a Quadtree (Samet 90) and a Point-Quadtree (Samet 90)) have been implemented as three sub-classes of the class SpatialIndex. Each function (insert, delete, locate, search) was then defined for each index. Then tests have been made on several kinds of data sets*. However, because of the variety of data sets, spatial indexes can't be compared. Results fluctuate even for the R*-Tree which is designed to fit with spatial objects' distribution. Gradual queries of 10 objects, 100 objects, 500 objects and 1000 objects have been executed and show this diversity of results (Table 1).

Table 1 : Time for querying objects in GeO2 using an R*-Tree

	BDCarto	BDCarto	BDCarto	BDCarto	BDTopo	BDTopo
	ADM	CTA	CTA	OCS		
	1 tile	1 tile	4 tiles	4 tiles	1/36 tile	1 tile
Number of objects	106 lines	688 lines	4173 lines	6023 lines	3381 lines	115 162 lines
~ 10 objects	13 in 2 sec	10 in 1 sec	9 in 0 sec	7 in 1 sec	9 in 3 sec	17 in 13 sec
Average number of object	6 obj/sec	10 obj/sec	-----	7 obj /sec	3 obj/sec	1,3 obj/sec
~ 100 objects	100 in 2 sec	161 in 2 sec	92 in 2 sec	162 in 2 sec	136 in 5 sec	178 in 16 sec
Average number of object	50 obj/sec	80 obj/sec	48 obj/sec	81 obj/sec	27 obj/sec	11 obj/sec
~ 500 objects	------	402 in 4 sec	633 in 5 sec	404 in 7 sec	553 in 10 sec	737 in 31 sec
Average number of object	------	100 obj/sec	126 obj/sec	58 obj/sec	55 obj/sec	24 obj/sec
~ 1000 objects	------	-------	1384 in 11 sec	1113 in 15 sec	1187 in 22 sec	1522 in 61 sec
Average number of object	------	-------	125 obj/sec	74 obj/sec	53 obj/sec	25 obj/sec

Still working with the R*-Tree, one problem raised concerning its construction time. To reduce it, two issues were followed :

- The first issue concerns tuning of O2. By changing the size of the client buffer (O2BUFFSIZE) and the size of the server buffer (O2SVBUFFSIZE) an improvement of about 50% was noted (1h30 instead of 2h30). Nevertheless, this construction process can definitely not be neglected yet.
- Another issue was brought up by using a temporary Unix file outside of the OO DBMS. This file contains all the index nodes while all spatial objects remain into O2. However, at the end of the construction, each index node is copied into the OO DBMS. A considerable improvement was obtained (only 0h30 instead of 2h30 for construction). So, the temporary Unix file seems a very good alternative to speed up the R*-Tree construction process.

Our work on spatial indexes showed us that a major improvement can be expected from OO DBMS. It concerns the building time of the index. Indeed, it is mandatory to produce the indexation of objects in less or just as much time as for the indexation of objects with a temporary file. This process affects also the querying of objects through

* Data sets belong to two geographical databases produced by IGN : the BDTopo®, already mentioned and the BDCarto®. The BDCarto is a database equivalent to a 1:100,000 scale map and contains four main themes : ADM (administrative boundaries), CTA (transportation networks), HYA (hydrography) and OCS (land cover). One tile corresponds to an area of 20x28 km2. 1/36 of a tile corresponds to the BDTopo acquisition unit.

the index. Partly because there is no clustering inside O_2, the search process is still rather slow. We hope that these two limiting factors will be by-passed with the implementation of a spatial index inside the kernel of the DBMS, and of a clustering procedure for objects according to the index structure, which would map out each index node with each physical page of the OO DBMS.

The final result of our experiment is the simplicity of our DBMS environment to test indexes (very simple class hierarchy), easy extensibility (by the addition of SpatialIndex subclasses) and easy implementation on O_2 with O_2C (later experiments were made by Pr. Scholl's team (Peloux 1994) at CNAM using GeO_2 too).

Data Visualization

A GIS without a visualization aid for geographic data is not a GIS. But in 1991, O_2 did not provide vector graphical interface and we decided to develop our own user interface for vector data on top of Xlib functions. We wrote it in O_2C, by the means of O_2 classes.

The first benchmark consists in the display of objects on the screen because the display time is crucial in GIS. The key to produce a rather quick interface is to look at the time of polyline display. Indeed, while many vertices can belong to a polyline, it is important to retrieve them very quickly.

Two kinds of storage means for polylines have been compared :

- the first considers a polyline as a list of point objects (using O_2 list constructor) and the display time is quite long ;
- the second solution considers a polyline as an array of points, ensuring contiguity of storage and no identifier for points. This kind of constructor is not provided by O_2. So, this storage means has been implemented as a C structure (an array of points) and is inserted inside O_2 as a bits chain. This storage means clearly outperforms the first solution as it accelerates display time by a factor 5.

Our best results with the second solution (shown in table 2) reveal that displaying one hundred polyline needs approximately from 0,6 to 0,9 seconds. This value is confirmed even if there are many points per arc (see fifth column BDTopo contour lines tile). The conclusion of this experiment is polyline modelling. The polyline must be seen as an array of points, which means a contiguous storage of points dynamically extensible and no particular identifier for the points.

Table 2 : Display time for geographic data

	BDCarto ADM 1 tile	BDCarto CTA 1 tile	BDCarto CTA 4 tiles	BDCarto OCS 4 tiles	BDTopo contour lines 1 tile	BDTopo 1 tile
Number of objects	106 lines	688 lines	4173 lines	6023 lines	7082 lines	115 162 lines
Number of points	2 765 points	11 665 points	38 514 points	62 042 points	415 636 points	804 788 points
Display time	1 sec	5 sec	25 sec	47 sec	46 sec	15 min 38 sec or 938 sec
Redraw time	< 1 sec	3 sec	18 sec	29 sec	31 sec	12 min 50 sec or 770 sec
Average display time for 100 lines	0,94	0,7	0,6	0,78	0,65	0,81

Generic data loading

A geographic database is designed to be used, exchanged or sold among producers and users. Loading geographic data sets is then the first task requirement in any GIS. Since each geographic data set has a data schema in accordance with an exchange data model, loading a geographic data set consists in the translation of the data schema into another one and in the loading of data. This task appears to be painstaking and critical since many exchange data models are available and the management of geographic data consumes a lot of computer resources.

So, the generic data loader developed for GeO2, is a useful module that automatically reads and translates the schema of geographic data into corresponding O2 classes, and ensures the loading of data, decomposing, if necessary, the loading of geographic data into several transactions. Several stages are necessary to achieve such a purpose (see figure 3).

1. The schema of the geographic data set is translated into our own schema definition file. This is our pivot data model to which translation rules must be specified for each exchange data model.

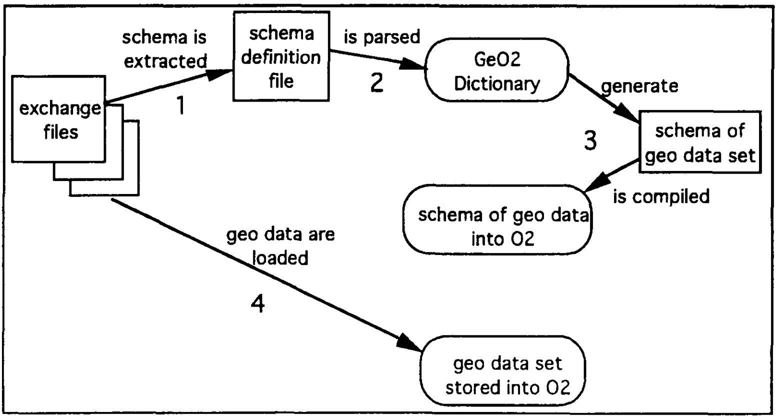
2. Then this file is parsed to constitute our data dictionary. The latter has been stored as O2 objects for easy development reason.

3. By using a temporary file the schema of the data set can be generated following O2 syntax and compiled into O2*

4. Geographic data is effectively loaded (geographic objects, geometric primitives).

Two topics are now tackled : data schema translation and data loading.

Figure 3 : Synthesis of the loading of a data set



The decomposition of data schema translation in three stages ensures us a kind of re-usability for each stage. The first stage only depends on the exchange data model definition. So, if another exchange data model is chosen, only the first part will be modified. Similarly, the second stage depends on our data model while the third stage depends on DBMS. Then, each factor is clearly distinguished and evolution is easier.

* This generation part has been developed before O2 proposes the meta-schema function.

Nevertheless, the fourth task cannot be generic at all and will be modified whatever the evolution.

For the fourth stage, performances have been examined on the BD Topo tile which includes 95 000 nodes, 115 000 lines (with 733 000 vertices), and 37 000 areas. Loading such a volume of data (and building spatial indexes, constructing topology) is consuming computer resources and requires approximately 25 hours as for current commercial GIS. From a DBMS point of view, long transactions would be here convenient. However, this loading can not be done with O$_2$ in a one-transaction process due to a limitation in memory size. No garbage collector, which could act as a background process, can't be made during the loading phase.

The solution we choose, consists in splitting the loading of geographic data (phase 4) into several DBMS sessions, launching O$_2$ and quitting O$_2$ respectively at the beginning and at the end of each step. For example, for loading geometric primitives, one step corresponds, in terms of nodes, to 5000 nodes each, in terms of lines, to 2500 lines and in terms of faces, to 5000 faces. But these values are completely configuration-dependent (for the determination of these thresholds see (David 93b)).

So, two conclusions emerge :

First, regarding time for loading, we have seen the great influence of topology building and spatial index construction. Indeed, they are the main cause of time consumption during loading and it is imperative that they should be optimized. Spatial index construction has already been tackled in the study of spatial indexes. But in topology building, we can only question the optimization of our algorithm.

The last point concerns the generic loading process that translates one schema in one model (the exchange data model) into one schema in O$_2$ and that needs access to the meta-schema in order to create classes on the fly depending on what the user wills. This function is very useful and it must be offered by object-oriented languages.

Portage of GeO$_2$ on C++/Versant

To test the portability of the prototype on a C++ persistent storage system, the portage of GeO$_2$ in C++ using the OO DBMS Versant was decided in 1993. The whole application had to be re-coded in C++. We focused at the beginning on automatic translation between O$_2$C and C++. But, after a pre-study, this task turns out to be too heavy to be done automatically. So, a "manual" rewriting of GeO$_2$ has begun while keeping in mind the idea of an automatic translator (Cuzon 93).

Three levels of translation have been distinguished :

The first level is a translation out of context. It consists in substitutions of keyword blocks into other blocks (class declaration...). This is the easiest part.

The second level is much more context-dependent. Hence some capabilities offered by Versant/C++ have been chosen explicitly and this slightly modifies the meaning of the code. For example, object and object reference can be distinguished in C++ in class definition, in function parameter whereas object reference alone can be used in O$_2$C.

And finally the third level reveals conflicts that have been detected only lately and that have prevented us from finishing and validating the portage.

Conflicts. The first conflict concerns inheritance and its two approaches ; either overloading (in C++), or sub-typing (in O$_2$C). As it is illustrated in figure 4, this mismatch implies that the code was adpated and some supplementary casts were made.

Figure 4 : From O$_2$C to C++

in O$_2$C		in C++
class SimpPoly ... method geom : SimpPoly, ... end class	is translated into	class SimpPoly {... virtual SimpPoly& geom() ; ... }
class CompPoly inherit SimpPoly ... method geom : CompPoly, ... end class	cannot be translated into	class CompPoly : virtual public SimpPoly {... CompPoly& geom() ; ... }
	is translated into	class CompPoly : virtual public SimpPoly {... SimpPoly& geom() ; ... }

But the most important conflict that considerably affects the developer's productivity concerns persistence. Here the greatest difference in object-oriented development has been detected.

In fact the problem occurs each time one decides to create an object. With persistence by reachability systems, the programmer does not have to care about the temporary or persistent character of the object. The system will garbage this object if it is no longer referenced.

On the contrary, with an explicit persistence system, the programmer must decide if this object will be persistent or temporary. This requires a preliminary knowledge of which are the temporary objects, which are the persistent ones. Because the transformation of temporary objects into persistent objects causes a duplication, there is less efficiency and the heap can be saturated (we have come up against this problem). So the right decision is, once again, left to the user's common sense through the adoption of conventions (i.e., always distinguish two kinds of methods : the persistent one and the temporary one).

After this experiment, we can establish the critical aspect of the portage that reveals all the weak points in the design of the prototype. In fact, every unsettled situation, every breakpoint quite often corresponds to the situation of an implicit action in another context (as in object and object-reference differentiation).

Finally, due to the conflicts, we have given up the comparison of OODBMS performance but only felt the great difference in the development with an explicit persistence system.

ON OBJECT-ORIENTED TECHNOLOGY

Thus, on the basis of the experiments described above, several questions about OO model, OO languages and OO DBMS are raised. This segmentation into three packages results in a re-organization and discussion of each conclusion in our experiments.

Object-Oriented Model

One delicate point is the object/value dichotomy found in O_2. And the reasons for choosing an object rather than a value are explained rather scantily. From our experience, two indicators can serve as a guide to choice-making. They are directly related to the use of the data, as it is shown now :

- Must this data be shared between several other objects ?

An object reference is mandatory in case of sharability. Otherwise a value is sufficient.

- Could we attach any behaviour to the data ? Is it dependent on the user's choices ?

In this case, we rely upon the classical encapsulation concept that is the association of data structure and data behaviour into one class. Once again, for active data, it will be better to choose an object rather than a value.

The second extension is the inheritance mechanism. Inheritance is a powerful abstraction for sharing similarities among classes while preserving their differences. However, its meaning can be ambiguous.

First, it can be interpreted as a generalization construct. It facilitates then modelling by structuring classes, capturing the similar external behaviour of different classes. A virtual super-class is introduced and can act as a representative of its sub-classes in other modules. For example, it is the SpatialIndex class in the module of spatial indexes. This is a conceptual approach similar to the sub-typing approach in programming languages.

But it could also be interpreted as a means to re-using services defined in another class, overriding operations for extension. For example, if we want to re-use the coordinate storage capabilities for topological entities, topological entities will inherit from geometric entities. Through this link, several classes tends to be closely joined and consequently, two modelling constructs tend to be joined. A new modelling hierarchy is created which is much more complex and very difficult to master on a long-term basis. For example, evolution inside the geometric domain is limited because it must fit with an evolution in the topological domain.

On the other hand, this approach has the advantage of being more efficient because it avoids to create two objects, to redefine all operations. Here, design options, for optimization, are opposite to long-term development and evolution. Note that this inheritance meaning is close to the overloading approach in programming languages.

Object-Oriented Languages

The function of a language is to easily translate the notions developed in the model and to create some uniformity through syntax rules and efficient processing. An object-oriented language gives the developer a more efficient way of organizing and testing his/her programs by means of classes (a kind of re-usable software component). And they are successfully used. Nevertheless, the management of persistence is still delicate.

Persistence is the property of data that enables it to exist for an arbitrary period of time : as short or as long as necessary. But distinguishing heap-allocated instances and long-term instances generates a considerable programming overhead (Atkinson 1983), a disturbing existence of two distinct type systems ; instead, the type-orthogonal persistence allows each kind of persistence to every data whatever its type. It makes long-term storage completely transparent and is supported by an extension of classical garbage collection (every data that is no more referenced by a persistent object is garbaged). Persistence by reachability system belongs to this group. But, currently, garbage collection is not efficient enough and produces an heavy overhead.

Then the current solution, that represents a backward step in comparison with orthogonal persistence, is explicit persistence. This means of development tends to be close to current developments on RDBMS.

The last point is about the high level concepts that have been enumerated all along the experiments described. Those represents the short-term need for developers : meta-schema functionality (mandatory when loading), array constructor (for polylines).

Object-Oriented DBMS

With OO DBMS, we face the last component that must surround every concept mentioned above. It is therefore all the more difficult to produce and fortunately, we found a rather complete product, which met our requirements, on the current market. So only two subjects are tackled :

As we have already said, the major difficulty in OODBMS was the insertion into the kernel of the spatial index and of the clustering mechanism (associated to the spatial index structure). But a more general problem concerns the extensibility of such a system (Schek 1993) with regard to the user's needs. What is the solution if we discover that however great our optimization efforts, topology building still remains a bottleneck in the loading process for geo-data ? And this question can be extended to each new research subject on geographical information (spatio-temporal database, multi-scale database). These models are, or will be, more elaborate than the current object-oriented model but we do not intend to build a completely new system. Perhaps it is necessary to re-think the DBMS as a module and clearly define what is the outer interface to this module ?

The second point, regarding short-term needs, concerns indicators that are lacking at present among available products. Much information on time, memory management, actual buffer size, I/O requests is needed to explain *where* time is wasted, *where* to search, and this kind of aid is not provided by OO DBMS yet.

CONCLUSION

It is worth noting that these subjects are not examined in a global context. We kept aside the subject of distributed system, of interoperable systems etc.... But results of our experiments allow us to shed new light on object-oriented technology. Indeed, three keywords can be extracted concerning the adequation of OO DBMS for GIS ativities : performances, extensibility and standardization (portability).

GeO$_2$ is now a platorm for further geographical research. Some particular points are being treated, such as, a study on a multi-scale database (Raynal 1994), on a spatio-temporal database (Latarget 1994) and on the modelling of geometric accuracy (Vauglin 1994). We know the capabilities of our prototype and some of its limitations. However, by choosing O$_2$ and O$_2$C, we preserve the qualities of easy development and easy exploratory experimentation that is provided when working with persistence by reachability system. This is a conscious choice even if it does not go hand in hand with present-day industrial processes. But, on the other hand, it is not the prime objective of a research.

REFERENCES

Atkinson, M., Bailey, P., Chrisholm, K., Cockshott, K., and Morrison, R. 1983, An approach to persistent programming: Computer Journal, Vol 26, n°4.

BDTopo 1992, Base de Données Topographiques: Bulletin d'Information de l'IGN, n°59.

Beckmann, N., Kriegel, H.P., Schneider, R. and Seeger, B. 1990, The R*-Tree: An efficient and robust access method for points and rectangles: Proc. ACM SIGMOD International Conference on Management of Data, pp 322-331.

Cuzon, A., and Grelot, C. 1993, Portage d'un système d'informations géographiques en O_2C sur un gérant d'objets compatible C++: Master thesis, DESS Systemes et Communications Homme-Machine, Université Paris XI.

David, B., Raynal, L., Schorter, G. and Mansart, V. 1993a, Why objects in a geographical DBMS ?: Advances in Spatial Databases, LNCS 692, Springer, pp 264-276.

David, B., Raynal, L. and Schorter, G. 1993b, Evaluation of the OO approach for geographical applications: Deliverable of ESPRIT project AMUSING n°6881.

Frank, A. 1984. Requirements for database systems suitable to manage large spatial databases: Proc of First International Symposium on Spatial Data Handling, pp 38-60.

Latarget, S. and Motet, S. 1994, Gestion de l'historique de l'information localisée par un journal: Proceedings of "Les Journées de la Recherche CASSINI", to appear.

O2 1991, The O_2 System: Communications of the ACM, Vol 34, n° 10.

Peloux, J.P., Reynal de Saint-Michel, G. and Scholl, M. 1994, Evaluation of spatial indexes implemented with the O_2 DBMS:Ingénierie des Systèmes d'Information, to appear.

Raynal, L. and Stricher, N. 1994, Base de données Multi-Echelles : Association géométrique des tronçons de route de la BD Carto et de la BD Topo: Proc 5th EGIS/MARI'94, EGIS Fundation, pp 300-307.

Samet, H. 1989, The Design and Analysis of Spatial Data Structures, Addison-Wesley Reading, MA.

Schek, H.J. and Wolf, A. 1993, Cooperation between Autonomous Operation Services and Object Database Systems in a Heterogeneous Environment: Interoperable Database Systems, Elsevier.

Vauglin, F. 1994, Modélisation de la précision géométrique dans les SIG: Pole Bases de Données Spatiales: Journées d'Avignon, pp 43-53.

IMPLEMENTATION OF TRIANGULATED QUADTREE SEQUENCING FOR A GLOBAL RELIEF DATA STRUCTURE

Jaime A. Lugo
Department of Geography
Mankato State University
Box MSU-2 Mankato
MN 56001, USA

and

Keith C. Clarke
Department of Geology and Geography
Hunter College—CUNY
695 Park Avenue
New York, NY 10021, USA

ABSTRACT

A three-dimensional, multilateral data structure based on Morton sequencing, two-dimensional run-length encoding, and Dutton's Quaternary Triangular Mesh (QTM) was developed, implemented and assessed. This data structure, called the Triangulated Quadtree Sequence, or TQS, was developed to map a global raster dataset onto the surface of a three dimensional solid octahedron. TQS provides the means to translate from Morton Codes to QTM and to TQS structures. To implement the triangulated quadtree sequence, a modified interrupted Collignon map projection was developed to project the world into eight equilateral triangular facets. To assess the TQS, the regular latitude/longitude grid of the land portion of the ETOPO5 global relief database was translated and compressed into the TQS triangular lattice. The validity and usefulness of the model is assessed, and its potential uses discussed.

INTRODUCTION

Why do pixels always have to be squares, or why tessellate with squares? Since the first applications of computer science to automated mapping, cartographers have preferred to measure and model the Earth using square-based coordinate grids and to perform analyses on grids. As a result, most of today's research in hardware (and software) design is based on square structures such as quadtrees (Samet, 1990a). A square-based structure, although excellent for planar geometry and two-dimensional coordinate systems (such as the Cartesian Grid), is one of the least suitable geometric models available for developing a data structure to store, link, and aggregate global scale three-dimensional data. A multilateral data structure using a geometric building object other than the square (such as a hexagon, octahedron, or triangle) is almost always better suited for the modeling of three-dimensional data (Wuthrich and Stucki, 1991).

The main objective of this study was the implementation of a hierarchical indexing method for assigning a unique set of geocodes to a global scale database–given a set of coordinates and the level of resolution desired–and to propose a method by which the indexing system can be developed into a powerful analytical tool for rendering global datasets onto a three-dimensional Earth model. This geocoding method is based on the tessellation of equilateral triangular facets after an initial division of the earth into an octahedral structure, as proposed by Dutton (1984). The storage method is hierarchical, and therefore is able to reference different tessellation levels, with each level having a higher resolution (Tobler and Chen, 1986).

A quadtree structure was used for the proposed planetary indexing system, since the planetary relief model is referenced based on a hexagonal/triangular coordinate system. Various alternate quadtree structures were initially compared, including the B+-Tree (Abel, 1984), Morton Sequencing (Morton, 1966), the PM3' quadtree for vector data (Samet and Webber, 1985), the Quaternary Triangular Mesh (Dutton, 1989a), and many others (Samet, 1990a and 1990b). Each of these quadtrees were weighted using Tobler and Chen's (1986) requirements for a global coordinate system.

The Quaternary Triangular Mesh (QTM) is a spatial data model developed by Geoffrey Dutton for global data modeling and measurement of locational uncertainty (Dutton, 1988). Even though the QTM data structure is similar to the quadtree in that it is based on a fourfold branching hierarchical structure (representable by only two binary digits), similarities end there. Dutton's spatial data model is unique because it uses a triangulated, numerical scheme for geocode addressing within a spherical rather than a planar system (Dutton, 1984).

The QTM is especially important to this study because it matches the design goal of a nontraditional data structure, and has remained a model not yet fully implemented. Although other quadtree variations exist for storage of global data, such as the cubic quadtree structure discussed by Tobler and Chen (1986) and Mark and Lauzon (1985), and Lauzon et al.(1984) these models tend to ignore the modeling of global phenomena in favor of accessing and retrieving map and image data (Dutton, 1989a). Goodchild (1988) also pointed out the need for a spherical model for point pattern analysis based on Theissen polygons and for the tessellation of polyhedral skeletons.

In the current work, Dutton's QTM addressing was implemented, with some modifications, and extended into a quadtree sequence based on recursively divided triangles. The initial mapping was onto a global octahedron in a near equal area projection (the Modified Collignon), and the test implementation consisted of building software and testing it using the land portion of the ETOPO5 global terrain data set.

TRIANGULAR TESSELATION AND QUADTREES

Triangles, unlike squares, have corners that match important surface nodes, and in spherical coordinate systems, their edges touch the global great circles. Therefore, triangle edges can represent linear features that are easily translated into a topological vector structure (e.g., the Triangulated Irregular Network (Peucker, et al. 1978; Peucker, et. al., 1986).

Given the task of representing a global data base, the first step is to build an initial set of triangles. The simplest solid figure with equilateral triangles for sides and with nodes on the surface of a sphere is the octahedron. Each of the facets on the octahedron's surface was tessellated into four geometrically identical triangles. If the edge midpoints of a triangle are connected, the result will be four new triangles–with an identical geometric structure to their parent triangle–every time the original triangle is tessellated. Following the Morton Sequencing quadtree method, Dutton assigned values of 0 (center), 1 (apex corner), 2 and 3 (opposite corners) to each subset triangle (Figure 1). Labeling the subsets from 0 to 3 allows the model to store each tessellated triangle into a 2-bit binary code. Since a triangular tessellation always produces four sub-triangles, it is possible to follow any triangular quadtree address from a chosen global quadrant down to any level of resolution desired. Each facet on the octahedron (or octant) after the first tessellation always has its center triangle labeled 0 and its apex labeled 1, while the other vertices are assigned 2 and 3 (counterclockwise).

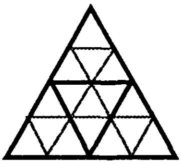


Figure 1a: A tessellated triangle produces more triangles

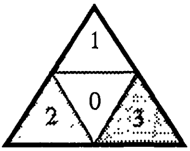


Figure 1b: Facet Breakdown Numbering. (Dutton, 1988)

APPLIED MAP PROJECTIONS

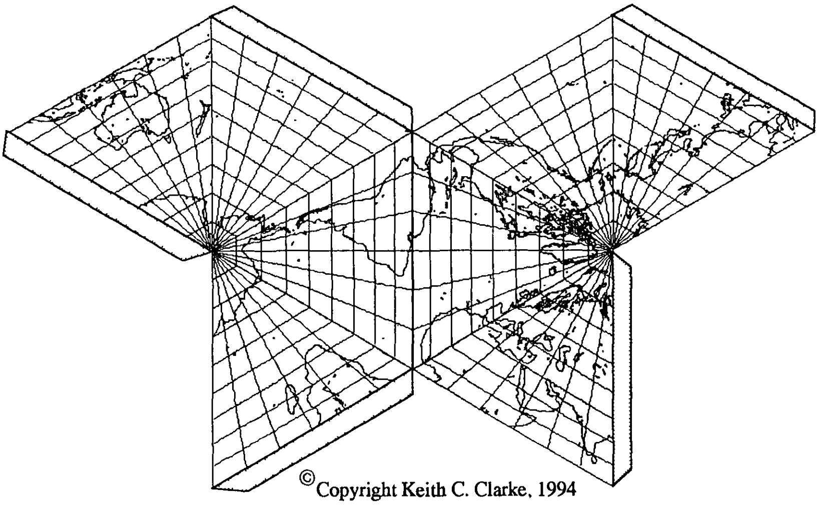
A major concern in developing a planetary relief model from an octahedron is the apparent error that may be introduced if the model does not account for the Earth being a geoid or ellipsoid rather than a perfect sphere. As discussed by Dutton (1984), every time an octahedron is sequentially tessellated it more closely resembles a multilateral globe. Dutton (1984) proposed to solve the geoid dilemma by first tessellating an octahedron to several thousand facets, then projecting it onto a geodesic projection. No specific projection was suggested in any of Dutton's papers (1984 to 1991); he claimed that more research was needed to determine a feasible solution.

To solve the riddle of what projection gives a feasible solution, over one hundred projections were examined, using *An Album of Map Projections* (Snyder and Voxland, 1989). It was determined that only a pseudo-cylindrical, triangular projection with straight parallels and meridians–such as the Collignon Projection–permits the world's outline to be projected onto a three-dimensional octahedron. It is especially important that latitude be represented as near linear steps and orthogonal at all points to the projection's central meridian, for the use of the triangle rotation algorithm in TQS. After several modifications, the Collignon projection was used to georeference both the World Data Bank and the ETOPO5 Global Relief Dataset into eight triangular lattice structures (octants).

The modifications involved on the Collignon require that the globe be first interrupted and that each quadrant be located and tested to determine whether it belongs to the northern or southern hemisphere. If it falls below the Equator, the quadrant is inversely projected (all points y values are multiplied by −1.0). The chosen central meridian is then used to determine the coordinates that define the longitudinal range of the quadrant's base, or equator. A C language program was written using the modified Collignon projection to project the world into eight equilateral triangular octants of an octahedral skeleton. As a result, the octahedral facets may be applied either collectively or individually. An algorithm was written that returned both a global octant number and a projected location within the octant.

The most important modification to this projection was the rescaling of the original projection to fit an equilateral triangle; the Interrupted Collignon is no longer completely equal-area. As a result, the Interrupted Collignon provides the specific Collignon location for any given latitude/longitude (x,y) coordinate in an equilateral triangle lattice. This feature not only locates all coordinate points on the globe, but serves as the method by which Morton, QTM and TQS addresses may be issued to every node on the octahedral skeleton. The quadrants can be seen in Clarke's Butterfly Projection (see Figure 2), where each quadrant is arranged by translation and rotation to resemble a butterfly. This projection can also be assembled into the shape of a three-dimensional paper octahedron.

Figure 2: The Interrupted Collignon(Clarke's Butterfly) Projection.



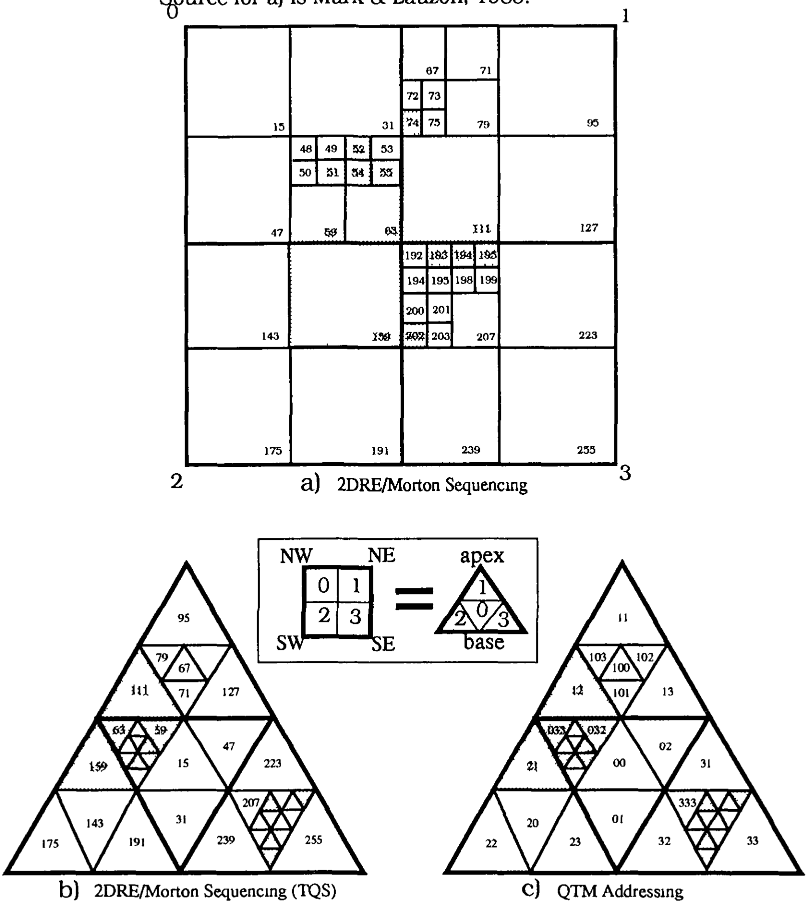
The Triangulated Quadtree Structure (TQS)

The proposed three-dimensional model requires a quadtree structure capable of georeferencing coordinates and elevation data to the triangle nodes within the octahedral structure, up to any level of tessellation. Such a quadtree must also be capable of linking large numbers of entities, attributes and relationships between the dataset and the octahedral facets. The Triangulated Quadtree Sequence is a combination of Morton sequencing, two-dimensional run-length encoding, and the QTM structure, modified to fit the previously defined specifications.

In order to store the outline of the world in the octahedral structure of the Interrupted Collignon Projection, the geocode given to each triangular quadrant must follow a static pattern similar to a rectangular Morton sequence (see Figure 3a). With a static pattern it is then possible to automatically assign Morton numbers and QTM addresses simultaneously. First, the center facet of each tessellated triangle was labeled zero [0] (Dutton, 1988); its north-south orientation is always opposite to the location of the central triangle's base. Second, the apex (labeled one [1]) shares the same base with the center facet (note that both triangles face in opposite directions).

Finally, the remaining facets are labeled by assigning the numbers two [2] and three [3] in a counterclockwise direction beginning at the apex. This step differs from Dutton's labeling scheme, where facets [2] and [3] may be arbitrarily labeled (Dutton, 1988). The result is a static addressing pattern that allows a triangular tessellation to be stored directly into a quadtree structure and vice versa, and which can also be stored using either Morton or QTM numbers. As seen in Figure 3, a raster image was originally stored using Morton sequences in a two-dimensional run-length encoding (2DRE) format. Note that the sequencing scheme used is logically identical to the Morton/2DRE addressing method in Figure 3a. The only difference is the applied geometric base. The next step is to apply the QTM addressing method. Once the raster image has been transferred from a square to a triangular lattice, it is a simple matter of substituting the Morton numbers for QTM addressing numbers. The results are shown in Figures 3b and 3c.

Figure 3: Although both a) and b) and c) are identical tessellations, a) is a 2DRE/Morton Quadtree, b) is a Triangular Quadtree Structure (TQS), and c) is a QTM triangular structure. The original image shown in 3a was directly imposed on 3b following the addressing pattern between Figures 3a and 3b. Each number shown in every one of these figures labels the highest value assigned to each particular quadrant, and based on the original image.
Source for a) is Mark & Lauzon, 1985.



a) 2DRE/Morton Sequencing

b) 2DRE/Morton Sequencing (TQS)

c) QTM Addressing

IMPLEMENTATION AND TESSELLATION OF THE TRIANGULATED QUADTREE STRUCTURE (TQS)

Recursive Tessellation for Determination of a TQS Address

First, each point is transformed from its original (x,y) coordinate (squared) lattice into a triangular lattice with the modified Collignon algorithm. The original raster image of the ETOPO5 dataset was transformed and stored into a set of eight separate octant quadtree files (one octant per quadtree). Each of the quadtrees in turn represents one of the facets of the global octahedron. After each latitude and longitude coordinate was computed from the original image (a 16-bit ERDAS "lan" format file, figure 4) and transformed into a triangular Collignon coordinate; the C program assigned a QTM address to each point within ETOPO5. The Collignon coordinate was then tested recursively to determine where within each triangle (facets 0, 1, 2, or 3) did the point belong. To save computing time, the maximum number of tessellations needed was determined before a TQS address was assigned for each point. Once the desired resolution was known, each point was transformed from a square-based lattice to the Collignon triangular lattice. Figure 5 shows the boundaries of an octant after a point is fitted through the Collignon algorithm. The most prominent feature is the grouping of points as one moves from the triangle base up to the apex (see Figure 2).

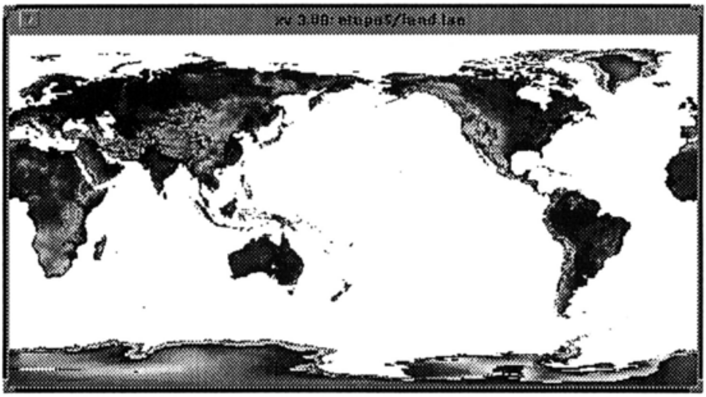
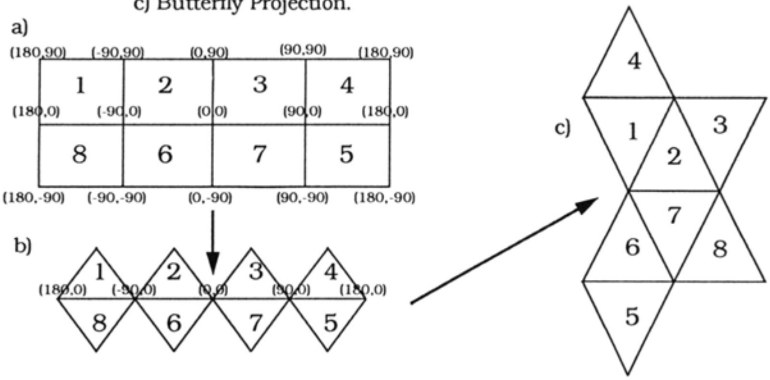
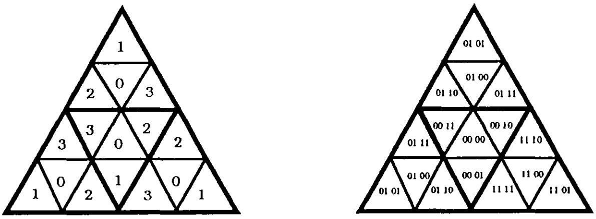


Figure 4: The original ETOPO5 raster image, global topography, land portion.

Figure 5: Transformation from a) square-based raster image, to
b) triangle-based lattice structure, to
c) Butterfly Projection.



In Figure 5a, the ETOPO5 data points are located in a rectangular, latitude-longitude model. Each rectangular octant was then projected into the model shown in Figure 5b, which is later arranged into its proper position within Figure 5c through eight separate rotation algorithms. The implementation of the triangulated quadtree structure began with the assignment of QTM addresses to each ETOPO5 point. The addressing scheme used differs from Dutton's model in that triangle addresses are always assigned in a counterclockwise order–Dutton's schema varied from triangle to triangle (Dutton, 1984; 1991). The order of facets through each tessellation is shown in Figure 6. This method was selected for two reasons: 1) a regulated order is essential if TQS is to be linked to any type of hierarchical structure, and 2) the computer algorithm method used to detect triangle location is based on this ordering scheme.



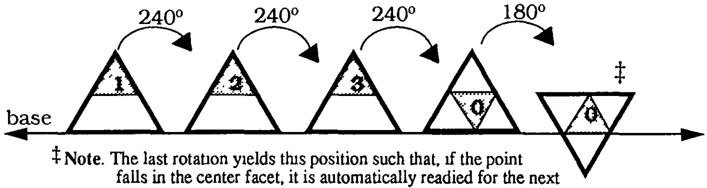
a) QTM: Address Ordermg

b) QTM: Bit-Addressing

Figure 6: TQS Addressing Scheme.

The function assigns TQS addresses to ETOPO5, point by point, by determining whether the y-value falls between the midpoint's y-range and the maximum y- value. If so, this point is 1) assigned a QTM number representing its triangle location, and then 2) its x- and y-values are divided in half, along with the x- and y-ranges. This coordinate division allows for the subsequent tessellation of this point's location until the full QTM address is determined up to the level of resolution desired, or maximum depth is achieved; see Figure 7. If the point fails to fall above the midpoint range, then the point is continually rotated 240° counterclockwise until the correct corner triangle location is determined. And, if it again fails to fall in a corner triangle, then by default it must fall in the center triangle. Therefore, by continually rotating and tessellating each x- and y-value, this method assigns a quaternary hierarchical address to every point in the dataset.

Figure 7: Triangle Rotations and Translations
Triangle rotation and translation for a point falling in the center facet. Note that the last translation inverts the point, then reduces its value by half (literally moving the point to the next tessellation level).



‡ Note. The last rotation yields this position such that, if the point falls in the center facet, it is automatically readied for the next tessellation.

Each time a point gets translated to determine whether it belongs to the facet that is currently above the midpoint range, it must be rotated counterclockwise 240°, its x- coordinate gets subtracted one fourth of the xrange value, and its y-coordinate gets added one half the y-range value. These additions and subtractions translate the octant back to its original position, except that the next triangle in the tessellation order is now at the apex (above the midpoint range). Note that while it seems that a full triangle gets translated, only one point gets translated at a time. Triangles are used to help visualize the area being processed. Since geometric translation and rotation of points requires that all calculations be performed in radians, each point is first translated into radians, and then rotated using standard affine rotation formulae. Each point went through this process twelve times at increasing resolution to determine the full TQS address (12 tessellation levels).

Discussion of the TQS Addressing Method

Dutton's Quaternary Triangular Mesh provided the means to develop and assign a TQS address for every coordinate point on the ETOPO5 global relief dataset. Each TQS address can be stored into a maximum of three bytes (24 bits), for a global resolution of approximately 4,883 meters, or 5 minutes. This is accomplished by assigning a 2-bit value (from 0 to 3) each time a point gets tessellated. This 2-bit value gets bit-shifted to the left (twice) and then appended to the previous value until the TQS address consists of twenty four binary digits (0 or 1). Therefore, TQS addressing eliminates the need for a pointer to each tessellation level since each pair of binary digits represent each level–with the leftmost pair indicating the root node. As a result, this method allows the user to recursively select any level of resolution desired.

The TQS address is now explained. Beginning from the left (root node) of the TQS number, each pair of binary digits represents the image's tessellation level up to the maximum pixel resolution; the two rightmost digits indicate the tessellation depth with the highest resolution. The TQS number is compressed even further by storing it in hexadecimal, rather than binary notation. This method introduces another source of compression into the geocoding process, though. When the binary number gets translated into hexadecimal notation, all leading zeroes get truncated from the number.

The result is a much smaller number than originally anticipated. For example, say that a point falls in TQS binary number `00 00 00 01 01 11 10 01 00 00 00`. Translated into hexadecimal notation, this number becomes [`5E40`]. If this number is again translated into a binary location, it becomes `1 01 11 10 01 00 00 00`. Note that, instead of the original 24 bits, the number now consists of only 15 bits. The missing levels are computed by adding the number of zeroes missing to the left of the binary address, until the number of binary pairs is equal to the maximum depth desired by the user (24 bits in this example). Sorting the TQS strings by length then allows maximal compression and multiresolution resampling.

Results From the Application of the TQS Addressing Method

The raster image of ETOPO5 as an ERDAS 16 bit lan file is 18,662,913 bytes. The small size of this data set comes from the fact that coordinates are implicit by order from the grid, not explicit. As such, reordering or sorting for resolution becomes impossible. With 4320 columns and 2160 rows, this file would expand to 214,617,600 bytes in ASCII with explicit coordinates before all points below sea level were excluded. Using just the land segment of the file, designed to test the multi-resolution aspect of TQS, left an explicit reference coordinate file of 52,254,720 bytes. The TQS, by storing an exact locational code for each point in hexadecimal notation, reduced the amount of storage needed to 48,540,279 bytes at twelve levels of recursion, equivalent to the same approximate resolution, stored as eight individual files (one for each octant).

Since the original coordinates were excluded from the octant files, the resulting files contain implicit location, instead of explicit location (octant number, elevation, and TQS address; no coordinates), and use binary form, not ASCII. Instead of storing the floating point elevation value (plus its link) required for representing each point in ASCII format at a 5-minute resolution, the TQS method stored the same point into a 24-bit binary value (three bytes instead of sixteen bytes). At a 5-minute resolution, the original coordinate and elevation values (in ASCII format) were stored as `+/-DD.MM +/- DDD.MM EEEE[EOL]` Degrees, Minutes, Elevation or the equivalent of twenty bytes. The same value in TQS format ranges from one to nine bytes, depending on resolution. ETOPO5 was further simplified for TQS by eliminating all those points with identical x- and y-coordinates within the ERDAS image file.

Elimination of data redundancy within a raster image was crucial because the original raster image adds identical x, y and elevation values as the image reaches the poles to account for the stretch caused by transposing the Earth's surface onto a flat grid. For example, the x, y and elevation values are identical for every pixel along the latitudes for the north and south poles, while every point is unique along the Equator. The Interrupted Collignon projection, unlike the original image, transposed the (x,y) coordinates closer and closer together as the points reached the poles, thus compressing the high latitudes into the apex of each triangle. Although it may seem as if accuracy is lost by concentrating more points in smaller and smaller areas, actually there is no more space at the poles than anywhere else on Earth. The error of reduction in areal resolution can be solved by implementing a three-dimensional planetary relief model (see Dutton, 1988; Lugo, 1994).

CONCLUSION

In conclusion, the Triangulated Quadtree Sequence (TQS) structure successfully linked the land portion of ETOPO5 to the Interrupted Modified Collignon Projection. The accuracy and precision of the TQS addressing numbers increased with every tessellation, until each address represented each elevation point up to the resolution of the original dataset. Once the georeferenced data is stored in the triangulated quadtree structure, areas with little relief can be generalized, while areas with high relief can be further tessellated up to the highest level of resolution.

The result has been a model that fitted the Earth's surface onto the surface of a map projection that was easily transformed into an octahedron. This accomplishment will eventually lead to a three-dimensional model capable of providing an image of the globe for any number of attributes, including population density, income, meteorological data, spread of disease, and others. In other words, this model may be viewed as a 2.5-dimensional choropleth map of the world. On the other hand, while this method offers simplicity and geolocational precision, it lacks the ability to measure scale throughout the octahedron's surface. As the levels of tessellation change, so does the length of every triangular segment, meaning that scale continues to change throughout the globe's surface and becomes impossible to measure (Dutton, 1989b). In other words, while a coordinate may tell the user where an object is, it fails to reveal the actual size of the object.

Despite this weakness in scale measurement, the achievement of the stated goals makes the proposed application a new step toward the development of a fully functional multilateral data structure for three-dimensional study. Although people today take for granted that a coordinate system is a always a discrete rather than continuous set of points, there will always be feasible alternatives to be researched, developed and implemented.

REFERENCES

Abel, D. J. (1984). "A B+-Tree Structure for Large Quadtrees." *Computer Vision, Graphics and Image Processing*, v. 27, pp. 1-18.

Dutton, G. (1984). "Geodesic Modelling of Planetary Relief." *Cartographica*, v21, Numbers. 2 & 3, pp. 188-207.

Dutton, G. (1988). "Modeling Locational Uncertainty via Hierarchical Tessellation." ed. Goodchild Michael and Sucharita Gopal. *Accuracy of Spatial Databases*. New York: Taylor & Francis, pp. 123-140.

Dutton, G. (1989a). "Computational Aspects of Quaternary Triangular Meshes." Prime Park, MS: Prime Computer, Inc. (unpublished).

Dutton, G.(1989b). "The Fallacy of Coordinates." Prime Park, MS: Prime Computer, Inc. (unpublished).

Dutton, G. (1991). "Improving Spatial Analysis in GIS Environments." *Proceedings Auto-Carto 10,* 1991, pp. 168-185.

Goodchild, M.(1988). "The Issue Of Accuracy In Spatial Databases." *Building Databases for Global Science*, Mounsey, H., and Tomlinson, R. (editors). London: Taylor & Francis, pp. 31-48.

Lauzon, J., D. Mark, L. Kikuchi, and A. Guevara (1985). "Two-Dimensional Run-Encoding for Quadtree Representation." *Computer Vision, Graphics and Image Processing,* v30, pp. 56-69.

Lugo, J. A. (1994). "Implementation of a Triangulated Quadtree Structure for a Three-Dimensional Planetary Relief Model." M.A. Thesis. New York: Hunter College of the City University of New York.

Mark, D. M. and J. P. Lauzon (1985). "Approaches for Quadtree-based Geographic Information Systems at Continental and Global Scales." *Proceedings Auto-Carto 7*, Falls Church, VA: ASPRS/ACSM, pp. 355-364.

Morton, G. M. (1966). "A Computer Oriented Geodetic Data Base, and a New Technique in File Sequencing." IBM Canada Ltd.

Peucker, T. K. Fowler, R. J., Little, J.J., and Mark, D. M. (1986). Digital Representation of Three-dimensional Surfaces by Triangulated Irregular Networks (TIN). Technical Report Number 10, United States Office of Naval Research, Geography Programs.

Peucker, T. K., et al (1978). "The Triangulated Irregular Network." *Proceedings of the Digital Terrain Models Symposium of the American Society for Photogrammetry/*American Congress on Surveying and Mapping. St. Louis, Missouri, USA, pp. 516-540.

Samet, H. (1990a). *Design and Analysis of Spatial Data Structures.* New York: Addison-Wesley Publishing Company.

Samet, H. (1990b). *Applications of Spatial Data Structures.* New York: Addison-Wesley Publishing Company.

Samet, H. and R. Webber (1985). "Storing a Collection of Polygons Using Quadtrees." *ACM Transactions on Graphics,* Vol. 4, pp. 182-222.

Snyder, J. P. and P. M. Voxland (1989). *An Album of Map Projections.* Reston, VA: Department of the Interior; U.S. Geological Survey, Denver, CO.

Tobler, W. and Z-T Chen (1986). "A Quadtree for Global Information Storage." *Geographical Analysis,* v18(4), pp. 360-371.

Wuthrich C. A., and P. Stucki (1991). "An Algorithmic Comparison Between Square and Hexagonal-Based Grids." *Computer Vision, Graphics and Image Processing: Graphical Models and Image Processing.* v53(4), pp. 324-339.

A MULTI-SCALE DAG FOR CARTOGRAPHIC OBJECTS*

SABINE TIMPF AND ANDREW U. FRANK
Technical University of Vienna
Department for Geoinformation E127.1
Gusshausstr. 27-29
1040 Vienna, Austria
{timpf,frank}@geoinfo.tuwien.ac.at

ABSTRACT

Geographic Information Systems manage data with respect to spatial location and that data is presented graphically as a map or sketch. A database of objects with some geometric properties is used to render these objects graphically for different tasks. Typically these tasks require graphical presentations at different levels of detail, ranging from overview screens to detailed views [Herot *et al.*, 1980]. Practically a base map is stored and its scale changed graphically. Without major distortions, only changes to twice or half the original scale are feasible by simple numeric scale change. A function to draw cartographic sketches quickly and in arbitrary scales is needed. We propose a multi-scale hierarchical structure where renderings of spatial objects with increasing detail are stored: a directed acyclic graph (DAG). These are used to compose a topographic map at a particular scale. We assume that the object renderings for the DAG already exist. Methods to select objects for rendering are based on the importance of the object for a particular task and on the principle of equal information density. We propose a method for determining information density, namely measuring the ink content of the map.

1. INTRODUCTION

1.1. Motivation

The traditional view of the cartographic process considers three different models and two transformations which lead to the production of a visual map (Fig. 1). There is first the model of the world (e.g. a GIS), consisting of objects with descriptive data.

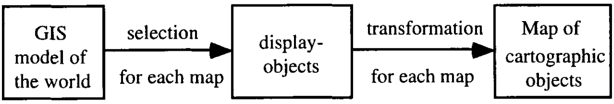


Fig. 1: Traditional cartographic process

From this model, a subset of the objects is selected to be included in a map, resulting in the set of display objects. These objects are then transformed from a geometrical description to a graphical form, applying rules for symbolization and other aspects of graphical encoding producing the set of cartographic objects. This viewpoint includes usually a strong feedback from the graphical rendering.

The traditional view assumes that the database contains the objects at highest resolution and that procedures exist to reduce them in scale and correspondingly in detail etc. (Beard). The selection step is first applied and the resulting objects then generalized to the desired level. This calls for automated generalization, a still difficult problem. Efforts to achieve automatic cartographic generalization were successful for specific aspects

* Funding from the Austrian Science Foundation (FWF) for the ESPRIT-Amusing project and from Intergraph Corporation is gratefully acknowledged.

[Freeman, and Ahn, 1987; Powitz, 1993; Staufenbiel, 1973], but no complete solution is known, nor are there any expected within the immediate future. Buttenfield and McMaster give a comprehensive account of current research trends [Buttenfield, and McMaster, 1991].

This proposal stores object representations for multiple scales in the database and applies only the selection step automatically (Fig. 2), reversing the order of steps.

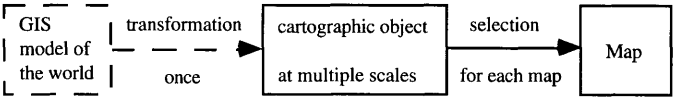


Fig. 2: The proposed selection process

The database will not be much larger than the most detailed database assumed in current proposals (assume that every more generalized representation is a quarter the size of the previous one, then the total storage requires only one third more capacity than the most detailed data set). Generalized representations can be collected from existing maps or for some cases produced automatically.

The transformation of an entity to a cartographic object in this proposal includes the generalization techniques of simplification, aggregation, symbolization and exaggeration (enhancement). The selection technique is explicitly mentioned in the process, whereas the displacement technique has to be considered when placing the objects in the map [Bundy, Jones, and Furse, 1994]. This also implies a feedback from composing the map to the selection technique.

1.2. Proposal

The approach selected here is to construct a multi-scale cartographic DAG (directed acyclic graph), where renderings for cartographic objects are stored at different levels of detail. The output map is constructed as a top-down selection of pre-generalized cartographic objects, till sufficient level of detail is achieved. We assume that the pre-generalized objects are given. Methods to select objects for rendering are based on the principle of equal information density. The dominant operation is 'zoom', intelligently replacing the current graphical representation with the more detailed one, that is appropriate for the selected new scale. We propose a relatively simple method to achieve equal information density, namely measuring 'ink'. The ink content can be determined by counting the number of black pixels per total pixels in the map.

The structure of the multi-scale DAG is based on a trade-off between storage and computation, replacing all steps which are difficult to automate by storage, which would be redundant if automatic deduction were feasible. The resulting DAG structure is more complex than the hierarchical structures proposed in the literature so far [Samet, 1989] since objects can change their appearance considerably e.g. from a single object to a group of objects. These different types of changes are used to derive the structure of the DAG.

The paper is structured as follows: first we examine the different types of changes in object renderings that can occur during scale changes. Then we derive the structure of the multi-scale DAG. After this we explain the selection process and the principle of equal information density. The paper closes with some remarks on further studies.

2. STRUCTURE OF THE MULTI-SCALE DAG

2.1. Types of changes

The multi-scale DAG extends ideas of hierarchies or pyramids and is related to quad-trees [Samet, 1989] and strip-trees [Ballard, 1981]. The DAG structure is more complex than the hierarchical structures proposed in the literature so far, as objects may change their appearance as shown in Table 1. Hierarchical subdivision of special object classes has been dealt with and studied extensively. Strip trees or a very similar design could be used to deal with lines which remain lines over multiple levels. Areal features can be represented by quad-trees as long as they remain areal features. But the multi-scale DAG must also include more dramatic changes. We examined changes that occur during 'zooming into' a representation.

continuous changes	discrete changes		
	slight change	complete change	appears
1. no change in appearance 2 increase of scale	3. change in symbol 4. increase of detail 5. appearance of label	6. change in dimension 7. shift to geometric form 8. split into several objects	9. appears

Table 1: Types of changes of object representations for smaller to larger scale

The table demonstrates that objects may change their spatial appearance in the generalization hierarchy. A particular problem is posed by objects which are not represented at small scale and seem to appear as one zooms in (type 9).

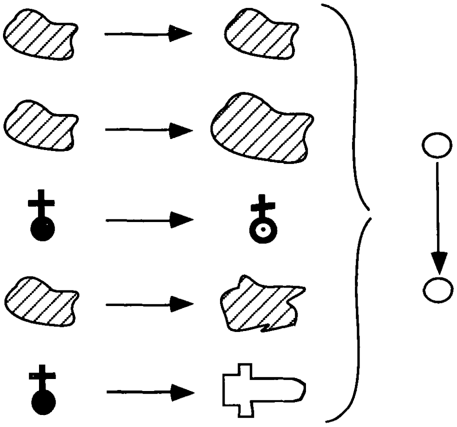


Fig. 3: Examples for the types 'no change', 'increase of scale', 'change in symbol', 'increase of detail', and 'shift to geometric form'

In figure 3, examples are shown for several of the types mentioned in table 1. These changes map to a representation in the DAG, which is shown on the right hand side of figure 3. It is the simplest of all mappings in the DAG, namely from one object representation to another (circles represent nodes with graphical output, pointers represent links between nodes).

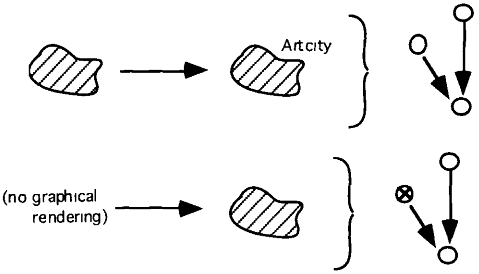


Fig. 4: Examples for the change 'appearance of label' and 'object appears'

The appearance of a label and of an object require an additional branch coming into the DAG-structure. Whereas the label is there and could be shown, a representation of the appearing object does not exist (crossed circles).

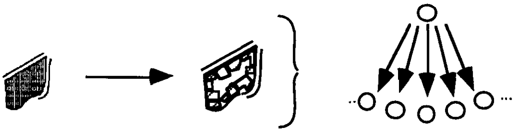


Fig. 5: Example for the change 'split into several objects'

An object, that is split into several objects, requires several links going into the more detailed level in the DAG structure.

Each type of change can be associated with an operation. It is interesting to note, that those operations have inverse operations, which are highly ambiguous. We first defined the changes for zooming in before looking at the changes for generalizing.

2.2. Hierarchical structure

The proposed multi-scale DAG is a method to produce maps of different scales from a single database [Beard, 1987]. It avoids the known problems of cartographic generalization, which cannot be fully automated today, using redundancy. Objects are stored in different levels of generalization, assuming that at least for the difficult cases, the generalization is done by humans, but only once. Building a multi-scale DAG is probably a semi-automated process where automated processes are directed by a human cartographer. All operations where human time is necessary are done once only and the results are stored.

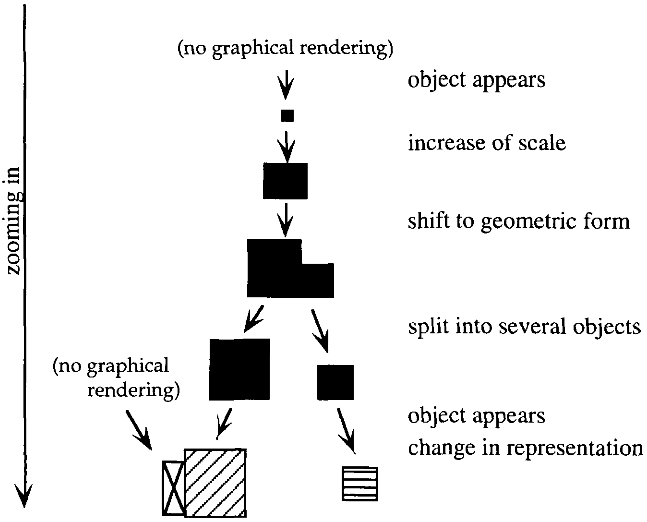


Fig. 6: A multi-scale DAG for buildings

In Graph theory, the structure of a DAG is well known [Perl, 1981]. Applied to our present problem, a multi-scale DAG for buildings might look like the DAG in Fig. 6. While zooming in different changes occur to the representation of the object and affect the structure of the DAG.

It is necessary to note that there will be several DAGs: Traditionally, changes in the map should occur in a certain order [Hake, 1975], namely rivers, railroads, roads, settlements, symbols; areal features, labels. For each of these object groups a different DAG is necessary

3. MAP COMPOSITION

3.1. Selection of objects

The operation applicable to every node of the DAG is a 'rendering' operation, which transforms the geometric data into a graphical picture. The problem to address is the selection of the objects in the DAG which must be rendered. Two aspects can be separated, namely,
the selection of objects which geometrically extend into a window and
the selection of objects to achieve a constant information density.

The selection of objects which extend into the window is based on a minimal bounding rectangle for each object and a refined decision that can be made based on object geometry. In order to assure fast processing in the multi-scale DAG, the minimal bounding rectangles must be associated with the DAG and DAG branching, such that complete sub-DAGs can be excluded based on window limits. This is well known and the base for all data structures which support fast spatial access [Samet, 1989].

The interesting question is, how the depth of descent into the DAG is controlled to achieve an equal information density. In data structures for spatial access, an 'importance' characteristic has been proposed [Frank, 1981; Van Oosterom, 1989]. It places objects which are statically assessed as important higher in the DAG and they are then found more quickly. The method relies on an assessment of the 'importance' of each object, which is done once, when the object is entered into the cartographic database. When a cartographic sketch is desired, from this ordered list the most important objects are selected for rendering.

The usability of this idea is currently being studied for a particular case, namely the selection of human dwellings (cities) for inclusion in a map [Flewelling, and Egenhofer, 1993]. A method based on an ordering of objects is not sufficient for the general case. It lacks provisions to deal with multiple representations of objects and must be extended for a multi-scale DAG, which is designed to deal with multiple representations of the same objects. It is also clear, that the ordering of objects is dependent on the purpose of the map. Nevertheless, substantial contribution to our understanding of the cartographic selection process is expected from the study of selection based on ordered (single-represented) objects.

3.2. The Principle of equal information density

The principle of constant information density can be derived from Töpfer's radix law [Meier, and Keller, 1991; Töpfer, and Pillewitzer, 1964]:

$$n_f = n_a * (m_a/m_f)^{n/2} \quad (1)$$

where m_a is the given scale,
m_f the following scale,
n_a the number of objects at the given scale,
n_f the number of objects at the following scale, and
n is the selection level.

Reformulating this law in terms of window area with n=4 and not in terms of the map scale, results in

number of objects / area = constant.

Retrieving a map at a given scale requires a top down search of the DAG. This search is guided by comparing spatial locations of the objects with the window of interest. Such comparisons are fast (linear in the number of objects compared), and the DAG can be structured such that only the relevant part must be searched.

The depth of the search in the relevant part of the DAG is bounded by the amount of graphical objects which can be shown. The test if an object can be shown is based on testıng that the required space is free; a test requiring constant time per object. Most of the objects tested are also included in the output. Therefore the search process is linear in the number of objects included in the output, which is - following the principle of constant information density - constant.

3.3. The 'ink' notion

A *perfect method* to achieve uniform information density requires a method to measure the information an object contributes to the map. The algorithm would then descend the DAG, within the limits of the window, and refine objects till the desired level is achieved. Note the difference to a rank order selection: all objects within the window are selected initially, and the process is refining or expanding objects till the desired amount of detail is achieved (this requires that all objects are initially included in a highly generalized fashion, which is a zero rendering).

This method is idealized and methods to measure information content of a cartographic object are currently unknown. A simplistic application of information theory [Shannon, 1948] cannot deal with this particular case, where the information content of an individual object must be measured.

A practical method is to measure *'ink'*, i.e. pixels which are black. One assumes that there is an optimal ratio of ink to paper. This ratio must be experimentally determined, measuring manually produced good maps (we expect a ratio of black pixels to total pixels). The expansion of the DAG is progressing from the top to the bottom, accumulating ink content and stopping when the preset value for graphical density is reached. The ink content would be measured not for the full window, but the window will be subdivided and ink for each subdivision optimized.

4. CONCLUSIONS AND FUTURE WORK

A number of applications requires graphical presentations of varying scale, from overview sketches to detailed drawings. The approach used here assumes that generalized versions of a map for an area are available and could be stored in a multi-scale DAG, where features which represent the same object, are linked. From such a multi-scale DAG, a map of arbitrary scale can be deduced by searching in the DAG till sufficient objects within the window of interest are found to produce a map of the same information density than the previous map. The principle of constant information density, which is number of objects/area = constant for the following scale, can be derived from Töpfer's radix law. We measure the information density with the practical method of counting 'ink'. Retrieving a map at a given scale requires a top-down search of the DAG, accumulating the number of black pixels within a given area until a preset value is reached.

The concept ıs based on a trade-off between computation and storage, replacing all steps which are difficult to automate with storage. These difficult steps are performed initially, while the remaining steps, which can be easily automated, are performed each time a query asks for graphical output. All operations where valuable human time is necessary are done once only and the results are stored.

For each of the object groups in a map, a different DAG is necessary. It is subject to further studies how these different DAGs interact. We propose to define the structure formally, e.g. with algebraic specification, and study the interaction possibilities.

The resulting DAG structure is more complex, than spatial hierarchical structures proposed in the literature so far and methods to deal with dangling branches and subbranches have to be found.

Changes in the map correspond to operations in the DAG and these have impact on the structure of the DAG. Operations for zooming in and for generalizing are inverse but highly ambiguous. Further studies on this subject are necessary, before a DAG structure can be fully implemented.

ACKNOWLEDGMENTS

The authors thank Dr. Werner Kuhn and the reviewers for their valuable suggestions. The support from the Austrian Science Foundation and from Intergraph Corporation is gratefully acknowledged.

REFERENCES

Ballard, D. H. 1981. “Strip Trees: A Hierarchical Representation for Curves.” *ACM Communications* 24 (5): 310-321.

Beard, K. 1987. “How to survive on a sıngle detailed database.” In *Auto-Carto 8 in Baltimore, MA*, edited by Chrisman, N. R., ASPRS & ACSM, 211-220.

Bundy, G.Ll., Jones, C.B., and Furse, E. 1994. “A topological structure for the generalization of large scale cartographic data.” In *GISRUK in Leicester, England*, edited by Fisher, P., 87-96.

Buttenfield, B. and, and McMaster, R., ed.1991. *Rule based cartographic generalization.* London: Longman.

Flewelling, D. M., and Egenhofer, M. J. 1993. “Formalizing Importance: Parameters for Settlement Selection.” In *11th International Conference on Automated Cartography in Minneapolis, MN*, edited by ACSM, ACSM.

Frank, A. U. 1981. “Applications of DBMS to Land Information Systems.” In *Seventh International Conference on Very Large Data Bases VLDB in Cannes, France*, edited by Zaniolo, C., and Delobel, C., 448-453.

Freeman, H., and Ahn, J. 1987. “On the Problem of Placing Names in a Geographic Map.” *International Journal of Pattern Recognition and Artificial Intelligence* 1 (1): 121-140.

Hake, G. 1975. *Kartographie.* Sammlung Göschen/de Gruyter: in German.

Herot, C.F. *et al.* . 1980. “A Prototype Spatial Data Management System.” *Computer Graphics* 14 (3, July):

Meier, S., and Keller, W. 1991. “Das Töpfersche Wurzelgesetz im Lichte der Stochastischen Geometrie.” *Wissenschaftliche Zeitschrift der Technischen Universität Dresden* 40 (5/6): 213-216.

Perl, J. 1981. *Graph Theory (in german).* Wiesbaden (FRG): Akademische Verlagsgesellschaft.

Powitz, B.M. 1993. “Zur Automatisierung der kartographischen Generalisierung topographischer Daten in Geo-Informationssystemen.” PhD, Hannover.

Samet, H. 1989. *Applications of Spatial Data Structures: Computer Graphics, Image Processing and GIS.* Readıng, MA: Addison-Wesley.

Shannon, C.E. 1948. *The Mathematical Theory of Communication.* Bell System Technical Journal, July and October 1948, reprint: Illinois Books.

Staufenbiel, W. 1973. *Zur Automation der Generalisierung topographischer Karten mit besonderer Berücksichtigung grossmasstäbiger Gebäudedarstellungen.* Kartographisches Intitut Hannover. Wissenschaftlıche Arbeiten Hannover 51.

Töpfer, F., and Pillewitzer, W. 1964. “Das Auswahlgesetz, ein Mittel zur kartographischen Generalisierung.” *Kartographische Nachrichten* 14: 117-121.

Van Oosterom, P. 1989. “A reactive data structure for geographic information systems.” In *Auto-Carto 9 in Baltimore, MA*, edited by Anderson, E., ASPRS & ACSM, 665-674.

Improving the Performance of Raster GIS: A Comparison of Approaches to Parallelization of Cost Volume Algorithms.

Daniel F. Wagner
Michael S. Scott

Department of Geography
The University of South Carolina
Columbia SC 29208

phone: (803) 777-8976
fax: (803) 777-4972
{dan, mike}@lorax.geog.scarolina.edu

ABSTRACT

Performance evaluation and improvement are increasingly being recognized as critical elements for the success of Geographic Information Systems. This research examines one strategy for improving the efficiency of spatial algorithms: parallelization. The very nature of spatial data and operators is such that decomposition has obvious benefits. However, identifying the most beneficial approach for parallel decomposition of spatial operators is not obvious. This research examines the parallelization of a cost volume operator in raster space. Two distinct approaches to decomposition of the algorithm are undertaken on the 64-node MIMD Intel Paragon Supercomputer at the University of South Carolina. The different approaches are compared on several test workloads. Performance metrics include execution time, speed-up and efficiency. In addition, the relationship between data volume and the benefits of parallelization are considered.

1. Introduction

Performance evaluation and improvement are increasingly being recognized as critical elements for the success of Geographic Information Systems (GIS). As workloads continue to increase, performance constraints become more acute. Further, as system capabilities continue to expand, performance tends to actually *decrease* due to increased overhead of user interfaces, advanced spatial modeling capabilities, etc. Incremental advances in hardware technology can no longer be relied upon to provide the solution to GIS performance problems. Efforts must be made to increase the efficiency of spatial algorithms. Further, alternative hardware solutions must be considered.

As capabilities of GIS systems expand, for example true handling of three spatial dimensions, alternative platforms such as parallel processors will become requirements rather than options. Many spatial algorithms exhibit latent parallelism. This inherent parallelism is identified by Armstrong (1994) as one area of future work in the realm of GIS and High Performance Computing. Specifically, Armstrong asks the question "are there collections of code that can be translated simply into parallel versions - the so called embarrassingly parallel algorithms?" It is our assertion that the nature of many spatial algorithms are inherently parallel. Specifically, geometric strategies such as plane sweeping (Preparata and Shamos, 1985) in vector and roving windows in raster possess a high degree of parallelism. However, even among such 'embarrassingly parallel' algorithms a number of strategies are possible.

Algorithms are parallelized through decomposition. Decomposition may be carried out on the program code itself as well as on the problem space operated on. Further, both may be attempted in the same algorithm. No single strategy is best, and the considerations are many. Particularly acute is the number of processors available. In general, the more processors available the more options for decomposition, however, more processors only come at the expense of power or cost. In addition scalability must be considered. Scalability refers to how well an algorithm can be mapped to different numbers of processors. Algorithms which scale well show consistent improvements in performance as more processors are utilized. However, not all algorithms, nor all spatial algorithms scale well. Thus, the ideal system is not necessarily the one with the most processors.

This paper is arranged in several parts. This section has introduced the importance of parallel processing for overcoming the performance constraints of spatial algorithms, and has reviewed several applications of parallel processing of spatial algorithms. In the next section, the major issues in parallel processing are reviewed. Following this a brief survey of parallel processing of spatial algorithms is presented. The spatial algorithm which forms the focus of this research, 3-D raster cost volume generation is then introduced. Following this, the experimental design for the performance testing is presented. This is followed by the results of our preliminary analysis. Lastly, a discussion of these results and conclusions are provided.

2. Issues in Parallel Processing

2.1 Parallel Architectures

Two principle architectures for parallel processing machines exist. These are the Single Instruction Multiple Data stream (SIMD) and the Multiple Instruction Multiple Data stream (MIMD). In SIMD machines, all processors execute the same instructions in lock-step. Such systems, often referred to as array processors, are only suitable for a narrow set of applications. In MIMD machines each processor executes its own set of instructions. This may be the same set of instructions as other processors - but not executed synchronously - or processors may have different sets of instructions -in effect the processors do different jobs. MIMD machines may be further divided into shared and distributed memory systems. Shared memory has the effect of limiting the number of processors and performance is affected by competition for memory. In distributed memory systems, each processor has its own local memory and data is shared among processors through messaging. Such communication can result in processors being idle for long periods while waiting for data to be passed from other processors.

2.2 Decomposition Strategies

Parallelization strategies are commonly divided into control, domain, and hybrid.
In control decomposition, different functions are assigned to different processors. A common approach is pipe-line processing in which data flows from function to function and processor to processor sequentially. Another approach involves a manager-worker strategy in which one processor (the manager) assigns jobs to other processors (workers). In domain decomposition the data is partitioned among processors and each partition is assigned to a processor. Performance is largely a function of communication and balance and is affected by the granularity or size of the data partitions. Ideally, the partitioning strategy should conform to the problem/data. Since geographic problems are multi-dimensional, they provide a wealth of opportunities for partitioning.

2.3 Communication

Communication represents one potential bottleneck in parallel processing. Through message passing data can be shared across processors. Communication may take place either synchronously or asynchronously. For synchronous messaging both processors must be prepared. If the sending processor is not ready the receiving processor must wait. In asynchronous communication the receiver issues a request for data and can continue processing, however if the data has not been (fully) received by the time the receiver needs it problems will arise. Communication is related to partition size in that finer grain potentially requires more communication and may result in a poor communication/computation ratio. Some inherently parallel algorithms require little or no communication, while some inherently sequential algorithms will always run faster on sequential machines due to the communication bottleneck.

2.4 Balance

Balance refers to the distribution of workload among processors. If some processors are idle while others are active, poor balance exists. The total execution time for an application is equivalent to the execution time of the last processor to complete its task. If all processors complete at the same time, then execution time is optimized and perfect balance is achieved.

2.5 Performance

The most basic measure of system performance is execution time. A common strategy in benchmarking programs is to query the system clock both immediately before and immediately after a block of code. The difference between these two times is one measure of execution time (Wagner, 1991) . In parallel processing, comparisons between execution time for parallel programs running on a single processor (or sequential programs) and parallel multi-processor programs can be made. Speed-up (*S*) is the ratio between the time to execute on one processor (T_1) and the time to execute on *P* processors (T_P):

$$S = T_1 / T_P \tag{1}$$

and under ideal conditions:

$$T_P = T_1 / P \tag{2a}$$

$$S = P \tag{2b}$$

Efficiency is the ratio of speed-up to number of processors:

$$E = S / P \tag{3}$$

and under ideal conditions:

$$E = 1 \tag{4}$$

3. Parallel Processing of Spatial Algorithms

Parallel processing of spatial algorithms has been considered by a number of researchers. A variety of spatial operators have been the focus of parallel processing. These include line intersection detection (Franklin et al, 1989; Hopkins and Healey, 1990), polygon overlay (Hopkins and Waugh 1991), spatial statistics (Griffith, 1990; Rokos and Armstrong, 1993; Armstrong et al, 1994), location models (Armstrong and Densham, 1992); intervisibility and viewsheds (Sandhu and Marble, 1988; Mills et al, 1992), grid interpolation (Armstrong and Marciano, 1993), name placement (Mower, 1993a; 1993b), line simplification and point-in-polygon (Li, 1993), shortest paths (Sandhu and Marble, 1988; Ding and Densham 1992), and hill shading and drainage basin delineation (Mower 1993a).

Almost exclusively a data-parallel or domain decomposition approach has been taken to parallelization of spatial algorithms (Franklin et al, 1989; Hopkins and Healey, 1990; Hopkins and Waugh, 1991; Armstrong and Densham, 1992; Mills et al, 1992; Ding and Densham, 1992; Armstrong and Marciano, 1993; Mower, 1993a; Li, 1993; and Armstrong et al, 1994). Much less common are examples of a control-parallel or control decomposition approach to parallelization of spatial algorithms. Examples include: Mower (1993a) and Rokos and Armstrong (1993) who adopt a master-worker approach.

Parallelization has been undertaken on machines ranging from two (vector) processors (Sandhu and Marble, 1988) to massively parallel processors such as Connection Machines (Mills et al, 1992). Generally, however, the number of processors available has been small.

A number of insights have been noted by many sources: in general it can be observed that most spatial problems, and especially large ones, will benefit from parallelization, but that a variety of strategies, considerations, and liabilities exist. Of particular concern are communication bottlenecks (Mills et al, 1992), and the interdependencies of the particular problem (Rokos and Armstrong, 1993).

In addition to considering the approaches which have been undertaken, it is worthwhile to note that comparisons between control and domain decomposition have not been attempted for the same algorithm or on the same machine. In addition, hybrid decomposition has not been attempted for spatial algorithms. All of the reviewed applications have been restricted to 2-D (plane) or 2.5-D (surface) cases. Finally, while shortest paths have been considered by Sandhu and Marble (1988) and Ding and Densham (1993) only network (vector) paths have been considered, and best paths have not been considered at all. In this research we address all of these points through an attempt to examine and compare different strategies for parallelizing the same spatial algorithm. It is hoped that by comparing parallelization strategies, insights can be gained which can be extended to the parallel decomposition of other spatial algorithms.

4. Best Paths *in* 3-Dimensional Raster Space

Best path problems represent a generalization of the shortest path problem. In shortest path problems the path which minimizes distance between 2 points is determined. Best paths differ in that the minimization function is based not solely on distance but other criteria (e.g. time) as well. The effects of the other criteria are included in the form of impedances or frictions associated with movements in directions or over spaces. Best paths in 2-D raster space have been examined by a number of authors (Goodchild, 1977; van Oosterom, 1993; Wagner et al, 1993) however, extensions of the problem to 3-dimensions is relatively new (Scott, 1994a). Unlike shortest and best paths in network space which can take any direction, paths in raster space, may take only a fixed number of directions.

In this paper, we consider a 3-dimensional 26-direction raster cost volume operator which makes use of a 3-dimensional pushbroom. The algorithm was developed initially on a sequential processor (Scott, 1994a; 1994b) and resulted as a direct extension from an improved 2-D best path algorithm (Wagner et al, 1993). Pushbroom algorithms require multiple *sweeps* through the dataset (Eastman, 1989; Scott, 1994a) - one for each *broom* - and are thus computationally intensive.

The majority of the computation time in such best path algorithms is involved in the generation of the cost surface or volume which is subsequently used to generate the best path. If absolute barriers are allowed, a large number of *sweeps* may be required. In the 2-D case, four *sweeps* - one originating at each corner of the rectangular study area - are required (barrier free case), while in the 3-D case, eight *sweeps* are required - one from each of the eight corners of the rectangular prismoidal study volume (again, barrier free case). In both cases, each *broom sweeps* from a vertex to its diametrically opposed vertex. In the 3-D algorithm examined here, only 26 directions are utilized. In the 26-direction case, only the immediate neighborhood of a focus cell is considered. Such neighbors share a face, edge or vertex. Each of the eight *brooms* must consider 7 directions (Figure 1), thus some redundancy is unavoidable in the (sequential) algorithm. As example output, the x=25, y=10 planes of a 50x50x50 3-D cost volume (origin in center) are shown in Figure 2.

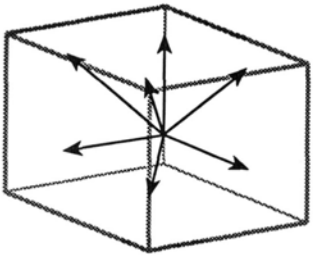


Figure 1. Seven directions considered by one *broom*.

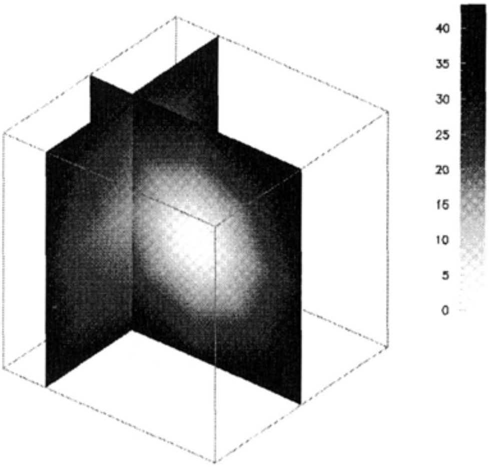


Figure 2. 50^3 Cost Volume, origin in center; planes: x = 10, y = 25

5. Experimental Design

5.1 Introduction

Tests were run on The University of South Carolina's Intel Paragon XP/S4 supercomputer. The Paragon has a distributed memory MIMD architecture using Intel 50 Mhz i860XP processors. The machine used in this research has 64 nodes arranged in a 2-D mesh, 1 gigabyte of distributed RAM, and a 12 gigabyte RAID. The nodes are divided between service and compute partitions. The service partition consists of 8 nodes each with 32 Megabytes RAM and handles logins, editing, compiling, utilities, and launching parallel programs. The compute partition contains the remaining 56 nodes, any number of which (including all) may be owned by a particular application. Each compute node has 16 megabytes of RAM and two i860XP processors, one for computing and the other for communication. Thus, 56 nodes represents the physical limit for this research. The i860XP processor is rated (peak) at 75 double precision megaflops computation and 200 MB/s communication. USC's Paragon has a peak performance of 4.2 gigaflops. All programming was done using C with parallel extensions.

The primary objective of this research is to examine and compare strategies for decomposition of spatial algorithms for parallel processing. We begin with the sequential algorithm, and then decompose this algorithm based on both control and domain. For each decomposition we examine the effect of scalability. Finally we examine a hybrid decomposition strategy which has elements of both control and domain decomposition.

Tests were performed on various sized synthetic datasets. The use of synthetic data allows for the total control needed to systematically examine the performance space of the system (Wagner, 1992). Results are reported in terms of execution time, efficiency, and speed-up.

5.2 Sequential Algorithm

The sequential algorithm can be conceptualized as a quintupley nested loop: For each of the eight *brooms*, and for each cell, the cost of moving from that cell to the neighboring cells is calculated and compared to the current cumulative cost for that neighboring cell and the minimum of the two is assigned. In pseudo-code, the algorithm may be envisioned as follows:

```
Broom
 Row
  Column
   Level
    Neighbor
     cmltv_cost(Neighbor)=
       MIN{cmltv_cst(Neighbor),cmltv_cst(Cell)+offset(Neighbor)}
     :
    :
   :
  :
 :
```

5.3 Parallel Algorithms

Our control decomposition is effected by parallelizing the outermost loop. That is the *brooms* sweep on several processors in parallel rather than each broom sweeping sequentially. In order to avoid the communications bottleneck only the calculation of cumulative cost is done within the loops and comparison and selection of minimum cumulative costs is delayed until a broom has completed *its sweep*. These values are then passed to a collector processor which determines the minimum value from the eight candidates assigned by the processors. Obviously in this approach anything beyond eight processors (or nine if the collector function is assigned to a separate processor) will not produce any performance improvements. For choices of 2, 4, and 8 processors, theoretically perfect parallelism can be achieved.

Control decomposition based on Neighbors rather than brooms is possible. Decomposition based on neighbors would involve assigning different directions to individual processors. All directions could be processed simultaneously (this would involve combining the outermost and innermost loops) using 26 processors. Collection would be required as above. This method of control decomposition is not implemented in this study.

Domain decomposition requires division of the study volume into sub-volumes. A major consideration for push-broom algorithms is that the broom must be able to sweep from corner to corner. In essence, the global effect must be preserved, so that a phenomenon can propagate throughout the entire study volume. For the cost volume algorithm, preserving the global effect requires multiple passes to allow effects to propagate across the sub-domain boundaries.

Sub-volumes are assigned to individual processors and the (sequential) algorithm applied. In order to allow propagation across sub-domain boundaries with a minimum of communication, the concept of 'ghost' cells (Intel 1994) is adopted. Rather than making the domain decomposition a mutually exclusive and collectively exhaustive one, the sub-volumes are designed to overlap. For pushbroom algorithms, the required depth of overlap is one cell. These overlapping cells are termed ghost cells. The ghost cells facilitate communication between passes and reduce the communication bottleneck.

Theoretically, perfect parallelism can be achieved if the sub-volumes are all the same size and shape. Perfect balance is achieved when the decomposition is uniform in all three directions. Balance is a function of the number of passes which must be performed to preserve the global effect. The number of passes required is equal to the maximum number of subdivisions of the problem volume along any axis. That is if the problem volume is bisected along an axis 2 passes are required; if the problem volume is trisected along an axis, 3 passes are required, etc. In effect, perfect balance is obtained when each spatial dimension is decomposed by the same integer (the cube root of the number of sub-volumes). Thus perfect balance and perfect parallelism can be achieved for 8, 27, 64, ... sub-volumes.

Hybrid decomposition, in which both control and domain decomposition are combined, is possible. Such a scenario might involve decomposing the study volume into eight sub-volumes and decomposing the sequential algorithm as described above. In this way better scalability should be achievable. That is, larger numbers of processors can be incorporated effectively. Hybrid decomposition is not attempted in this study.

6. Results

Control decomposition utilizing a collector and between 1 and 8 *broom* processors (2-9 processors) was performed using cubic study volumes ranging from 10x10x10 to 100x100x100 cells. Domain decomposition utilizing 1, 2, 8, and 27 sub-domains was performed on cubic study volumes ranging in size from 10x10x10 to 100x100x100. Results of individual tests were performed several times and averaged in order to lessen residual background effects. The results from the tests are presented in Tables 1-3.

Scott (1994a) developed the original 3-dimensional 26-direction raster cost volume operator in Turbo Pascal on a (sequential) PC-based environment. He reports that using a 33 Mhz 486DX processor with 16 megabytes of RAM of which 8 megabytes were configured as a RAM disk, approximately 1069 seconds were required to generate a 40x40x40 cost volume, and 79,000 seconds were required to generate a 100x100x100 cost volume. He also notes that the performance of his algorithm is very much disk constrained. Our first step in this research was to port Scott's sequential algorithm to C and a RISC/UNIX environment. On a SPARC IPX, the algorithm required 21 seconds to generate the 40x40x40 cost volume and 1200 seconds to generate the 100x100x100 cost volume. Next, the sequential algorithm was run on the Paragon (Table 1). Following this we undertook control and domain decomposition of the algorithm. Tables 2-3 present the results of these experiments.

Table 1. Sequential Program: Execution time in seconds.

Number of Nodes	Problem Size									
	10^3	20^3	30^3	40^3	50^3	60^3	70^3	80^3	90^3	100^3
1	0.10	0.98	3.49	8.50	16.9	29.4	47.0	70.6	102.	139.

The sequential tests show almost constant linearity between test size and execution time. The performance of the Paragon in sequential mode appears an order of magnitude better than the IPX for the 100^3 case, and two orders of magnitude better than the PC. Tables 2-3 illustrate that the sequential times are slightly faster than the parallel times using one node. This is due to the overhead associated with the parallel processing. This disadvantage vanishes as multiple-nodes are used, as expected. In the 100^3 tests, the sequential processor again demonstrates the best performance due to overhead associated with problem size in parallel mode.

Table 2. Control Decomposition: Execution time in seconds.

Number of Nodes	Problem Size									
	10^3	20^3	30^3	40^3	50^3	60^3	70^3	80^3	90^3	100^3
1	0.12	1.02	3.57	8.69	17.2	30.0	47.9	71.9	118.	225.
2	0.06	0.57	1.97	4.76	9.48	16.5	26.4	39.6	66.2	193.
3	0.11	0.47	1.59	3.77	7.44	13.0	20.7	30.7	45.0	238.
4	0.08	0.37	1.25	2.95	5.81	10.1	16.1	23.5	35.1	258.
5	0.23	0.53	1.37	3.11	6.06	10.4	16.3	24.7	35.9	315.
6	0.32	0.56	1.37	2.94	5.56	9.62	15.0	22.7	33.0	341.
7	0.33	0.58	1.30	2.83	5.45	9.61	15.2	23.0	31.9	401.
8	0.39	0.54	1.30	2.73	5.14	8.74	13.8	20.5	29.2	457.

Table 3. Domain Decomposition: Execution time in seconds.

Number of Nodes	Problem Size									
	10^3	20^3	30^3	40^3	50^3	60^3	70^3	80^3	90^3	100^3
1	0.24	2.01	7.08	17.2	34.1	59.4	94.9	142.	215.	374.
2	0.16	1.15	3.93	9.42	18.6	32.3	51.4	77.0	223.	429.
8	0.42	0.92	2.41	5.16	9.74	16.8	26.1	39.2	57.7	455.
27	0.86	2.39	6.43	14.1	26.5	45.0	70.7	105.	157.	1268

6. Discussion

The observed execution times illustrate that the 3-D 26-direction raster cost volume generation algorithm does benefit from parallelization. However, the observations also suggest that the algorithm does not scale evenly and that data volume can have a significant impact on speed-up and efficiency. In general the observations for the smallest problems (10x10x10) are least interesting. This is because the execution times are so small as to be completely confounded by the background processing overhead.

The control decomposition strategy works consistently better than the domain decomposition strategy on this algorithm. Closer examination of the parallel algorithms suggests that this is due largely to the communication required in the domain decomposition. In control decomposition, communication is necessary only at the beginning and ending of the processing. At the beginning, the origin(s) and friction values must be passed to each processor. Upon completion of the *sweep* the cumulative cost values must be passed to the collector processor where the minimum value is selected. Thus, little communication is needed, and what communication there is does not interfere with the individual *broom* processing.

In comparison, the domain decomposition strategy involves much more communication and further, this communication occurs during the sub-domain processing. While attempts have been made to minimize communication through the use of ghost cells, communication between all sub-domains must occur between each pass, and presents a significant bottleneck. The execution times for domain decomposition are on average nearly twice as long when compared to control decomposition.

Another observation centers on performance versus problem size. In both the control and domain tests anomalies in the pattern of performance can be observed as the problem size grows large. Visual inspection of the processor status has shown that these performance drop-offs occur when the workload becomes disk intensive. Problems up to some threshold size can be handled entirely in RAM, but beyond this threshold, disk storage must be utilized. Once this threshold is passed, all processors spend a majority of their time waiting for the RAID to service their requests for data. In addition, performance degradation occurs in the preprocessing stage when large numbers of processors or large amounts of array space (i.e. large data sets) are utilized. Again, visual observation of processor status has shown that while processors and/or memory is being allocated, other processors generally sit idle.

In addition to raw measure of execution time, speed-up and efficiency should be considered. Tables 4-7 present the speed-up and efficiency measures for both experiments. In general, speed-ups increase moderately as processors are added. However, this rate is clearly not proportional to the change in number of processors. Thus, while the greatest speed-ups are gained using the most processors, the greatest efficiencies are gained using the smallest number of processors. In addition, as execution times increase, speed-up and efficiency both drop-off with very large problems or large numbers of nodes.

A final topic for discussion is balance. Near perfect balance is possible in both the control and domain strategies outlined here. In fact, in certain instances the control decomposition for the cost volume algorithm represents the ideal case. While the processor workloads can be equally well balanced in the domain strategy, it is the communication between processors which offsets this advantage.

Imperfect balance occurs in the domain decomposition strategy if the sub-domains are not all the same size and shape. That is balance will be upset if some processors must operate on larger sub-domains. Imperfect balance occurs in the control decomposition case when the number of *broom* processors is not a factor of eight. Thus, examination of Table 2 shows that there is little difference in performance times for 4, 5, 6, and 7 processors. In effect, with more than three but less than eight processors, two passes are still required to process all eight *brooms*. Thus, some of the additional processors sit idle while others are busy on the second pass. In other words, such numbers of processors throw the parallelization greatly out of balance.

Table 4. Speed-Up from Control Decomposition.

Number of Nodes	Problem Size									
	10^3	20^3	30^3	40^3	50^3	60^3	70^3	80^3	90^3	100^3
2	2.00	1.79	1.81	1.83	1.81	1.82	1.81	1.82	1.78	1.16
3	1.09	2.17	2.25	2.31	2.31	2.31	2.31	2.34	2.62	0.95
4	1.50	2.76	2.86	2.95	2.96	2.97	2.98	3.06	3.36	0.87
8	0.31	1.89	2.75	3.18	3.35	3.43	3.47	3.51	4.04	0.49

Table 5. Efficiency from Control Decomposition.

Number of Nodes	Problem Size									
	10^3	20^3	30^3	40^3	50^3	60^3	70^3	80^3	90^3	100^3
2	1.0	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.6
3	0.4	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.3
4	0.4	0.7	0.7	0.7	0.7	0.7	0.7	0.8	0.8	0.2
8	0.0	0.2	0.3	0.4	0.4	0.4	0.4	0.4	0.5	0.1

Table 6. Speed-up from Domain Decomposition.

Number of Nodes	Problem Size									
	10^3	20^3	30^3	40^3	50^3	60^3	70^3	80^3	90^3	100^3
2	1.50	1.75	1.80	1.83	1.83	1.84	1.85	1.84	0.96	0.87
8	0.57	2.18	2.94	3.33	3.50	3.54	3.64	3.62	3.73	0.82
27	0.28	0.84	1.10	1.22	1.29	1.32	1.34	1.35	1.37	0.29

Table 7. Efficiency from Domain Decomposition.

Number of Nodes	Problem Size									
	10^3	20^3	30^3	40^3	50^3	60^3	70^3	80^3	90^3	100^3
2	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.5	0.4
8	0.1	0.3	0.4	0.4	0.4	0.4	0.5	0.5	0.5	0.1
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0

7. Conclusions

We have examined two distinct strategies for parallelization of 3-D raster cost surface generation. In this test, control decomposition appears to hold significant advantages over domain decomposition. While domain decomposition offers better scaling and can be spread over a broader range and greater number of processors, the increased communication necessary for this approach ultimately lessens its performance.

While we are confident our control decomposition is optimal, we are aware of a number of potential strategies for improving our domain decomposition. This research has shown that control decomposition can be superior to domain decomposition for spatial problems. However, additional research is necessary before generalizations can be drawn. As in many other studies, we conclude that the most successful strategy is the one making use of both the greatest number of balanced processors and the least amount of communication. In the example of 3-D raster cost volumes, near perfect balance and near perfect parallelism can be exploited in the control decomposition. While (near) perfect balance can be easily achieved in our domain decomposition, the degree of communication necessary to preserve the global effect and allow values to propagate across sub-domains insures that near perfect parallelism (i.e. little communication) is not possible.

Additional strategies for decomposing this algorithm are also worthy of investigation: these include control based on the 26-directions and hybrid decomposition strategies. In light of the communications bottleneck associated with domain decomposition, hybrid strategies should prove a viable alternative to facilitate increasing the number of processors utilized while maintaining satisfactory perforce. In order to achieve this, the maximum number of *brooms* should be included since this 'control' component of the hybrid decomposition is not so communication bound.

In addition to examining alternate strategies, it is necessary to investigate the effect of architecture on the success of the decomposition strategies. Finally, it is necessary to more rigorously examine the performance space to better determine the existence and extent of efficiency thresholds or discontinuities in the performance space.

Acknowledgements

The authors would like to thank Dr. Patrick Fay, computational scientist at USC's Computer Services Division, for his invaluable assistance with the Paragon.

Bibliography

Armstrong, M., and Densham, P. 1992. "Domain Decomposition for Parallel Processing of Spatial Problems." *Computers, Environment, and Urban Systems.* Vol 16, pp 497-513.

Armstrong, M., and Marciano, R. 1993. "Parallel Spatial Interpolation." *Proceedings*, Auto-Carto 11. Bethesda, MD: ASPRS and ACSM. pp 414-423.

Armstrong, M. 1994. "GIS and High Performance Computing." *Proceedings*, GIS/LIS '94. Bethesda, MD: ASPRS and ACSM; Washington, DC: AAG and URISA; Aurora CO: AM/FM International. pp 621-630.

Armstrong, M., Pavlik, C., and Marciano, R. 1994. "Parallel Processing of Spatial Statistics." *Computers & Geosciences.* Vol 20, pp 91-104.

Eastman, J. R. 1989. "Pushbroom Algorithms for Calculating Distances in Raster Grids." *Proceedings*, Auto-Carto 9. Bethesda, MD: ASPRS and ACSM. pp 288-297.

Ding, Y., and Densham, P. 1992. "Parallel Processing for Network Analysis: Decomposing Shortest Path Algorithms for MIMD Computers." *Proceedings*, Spatial Data Handling 5. Columbia, SC: Department of Geography, University of South Carolina, IGU Commission on GIS. pp 682-691.

Franklin, W. R., Narayanaswami, C., Kankanhalli, M., Sun, D., and Wu, P. 1989. "Uniform Grids: A Technique for Intersection Detection on Serial and Parallel Machines." *Proceedings*, Auto-Carto 9. Falls Church, VA: ASPRS and ACSM. pp 100-109.

Goodchild, M. 1977. "An Evaluation of Lattice Solutions to the Problem of Corridor Location." *Environment and Planning A.* Vol 9, pp 727-738.

Griffith, D. 1990. "Supercomputing and Spatial Statistics: A Reconnaissance." *The Professional Geographer*, Vol 42, pp 481-492.

Hopkins, S., and Healey, R. 1990. 'A Parallel Implementation of Franklin's Uniform Grid Technique for Line Intersection Detection on a Large Transputer Array." *Proceedings*, Spatial Data Handling 4. Columbus, OH: Department of Geography, Ohio State University, IGU Commission on GIS. pp 95-104.

Hopkins, S., and Waugh, T. 1991. "An Algorithm for Polygon Overlay Using Cooperative Parallel Processing." *Working Paper No. 19.* Edinburgh: Economic and Social Science Research Council Regional Research Laboratory for Scotland.

Intel Supercomputer Systems Division. 1994. "Paragon™ User's Guide." Beaverton: OR.

Li, B. 1993. "Suitability of Topological Data Structures for Data Parallel Operations in Computer Cartography." *Proceedings*, Auto-Carto 11. Bethesda, MD: ASPRS and ACSM. pp 434-443.

Mills, K., Fox, G., and Heimbach, R. 1992. "Implementing an Intervisibility Analysis Model on a Parallel Computing System." *Computers & Geosciences.* Vol 18, pp 1047-1054.

Mower, J. 1993a. "Implementing GIS Procedures on Parallel Computers: A Case Study." *Proceedings*, Auto-Carto 11. Bethesda, MD: ASPRS and ACSM. pp 424-433.

Mower, J. 1993b. "Automated Feature and Name Placement on Parallel Computers." *Cartography and Geographic Information Systems.* Vol 20, pp 69-82.

van Oosterom, P. 1993. "Vector vs. Raster-based Algorithms for Cross Country Movement Planning." *Proceedings*, Auto-Carto 11. Bethesda, MD: ASPRS and ACSM. pp 304-317.

Preparata, F., and Shamos, M. 1985. *Computational Geometry: An Introduction.* New York, NY: Springer-Verlag.

Rokos, D., and Armstrong, M. 1993. "Computing Spatial Autocorrelation in Parallel: A MIMD implementation for Polygon Data." *Proceedings*, GIS/LIS '93. Bethesda, MD: ASPRS and ACSM; Washington, DC: AAG and URISA; Aurora CO: AM/FM International. pp 621-630.

Sandhu, J., and Marble, D. 1988. "An Investigation into the Utility of the Cray X-MP Supercomputer for Handling Spatial Data." *Proceedings*, Spatial Data Handling 3. Columbus, OH: Department of Geography, Ohio State University, IGU Commission on GIS. pp 253-267.

Scott, M. 1994a. "Optimal Path Algorithms in Three-Dimensional Raster Space." Unpublished M.S. Thesis. Columbia, SC: Department of Geography, University of South Carolina.

---------. 1994b. "The Development of an Optimal Path Algorithm in Three-Dimensional Raster Space." *Proceedings*, GIS/LIS '94. Bethesda, MD: ASPRS and ACSM; Washington, DC: AAG and URISA; Aurora CO: AM/FM International. pp 687-696.

Wagner, D. 1991. "Development and Proof-of-Concept of A Comprehensive Performance Evaluation Methodology for Geographic Information Systems." Unpublished Ph.D. Dissertation. Columbus, OH: Department of Geography, Ohio State University.

-------------. 1992. "Synthetic Test Design for Systematic Evaluation of Geographic Information Systems Performance." *Proceedings*, Spatial Data Handling 5. Columbia, SC: Department of Geography, University of South Carolina, IGU Commission of GIS. pp 166-177.

Wagner, D., Scott, M., Posey, A. 1993. "Improving Best Path Solutions in Raster GIS." *Proceedings*, GIS/LIS '93. Bethesda, MD: ASPRS and ACSM; Washington, DC: AAG and URISA; Aurora CO: AM/FM International. pp 718-726.

CONSTRAINT BASED MODELING IN A GIS: ROAD DESIGN AS A CASE STUDY[1]

Andrew U. Frank
Dept. for Geo-Information E127.1
Technical University Vienna
Gusshausstr. 27-29
A-1040 Vienna Austria
FRANK@GEOINFO.TUWIEN.AC.AT

Mark Wallace
Imperial College
Wilham Penney Laboratory
London SW7 2AZ, U.K.
MGW@DOC.IC AC.UK

ABSTRACT

Modeling in GIS is limited to the standard geometric data models: vector and raster. Not all problems can be expressed in these models and extensions are requested by applications. To determine precise requirements for extensions, case studies are beneficial. In this paper the focus is on the expressive power required for design applications. A significant part of the road design task is used as a case study to explore if constraint databases can contribute to the solution. Road design is a suitable example for GIS design applications, as road design in the past used topographic maps and map analysis methods.

The general design task for the layout of a highway is presented: Find the technically feasible alternative road layouts between two points. Select the best for further assessment. Given are the design parameters of the road (design speed, minimal radius of curves, maximum slope) and information about the terrain (land cover, elevation, geology).

The problem is sufficiently complex to pose substantial research questions, but simple enough to be tractable. Three observations result from this case study:

- Constraint databases can be used to model continuos variables, and therefore space (not only discrete points in space),
- Representing space with a Delaunay triangulation leads to conceptual simplifications,
- The original constraint of the road design problem, which are of a higher degree, can be linearized to make the problem computationally tractable.

A test implementation is underway, exploring the performance aspects of the use of constraint databases for road design.

1. INTRODUCTION

Modeling in GIS is more limited then what is most often perceived. GIS users are extremely able to select only problems which they can solve with the functionality of today's GIS and other problems are not considered. In the past years, more and more functions have been added to the commercial GIS. The limitations imposed by the models employed, however, remain. These are limitations of 'expressive power' - what kind of problems can be described in the language provided by the GIS (Gallaire, Minker, and Nicolas 1984).

The major limitations arise from the standard geometric data models: the vector and raster model (Frank 1990, Goodchild 1990, Herring 1990). Efforts to combine the two

[1] Funding from Intergraph Corporation and the Austrian Fund for Scientific Research (FWF) is gratefully acknowledged. The case study is a part of an effort supported by the European Community (ESPRIT program) to develop constraint databases (project CONTESSA) Stimulation to consider the problem and additional support for the work was received from the Bundesministerium fuer wirtschaftliche Angelegenheiten (Strassenbau-Sektion)

models (Peuquet 1983) do not overcome all limitations. They combine only the power of the two models, but do not provide what is not included in one or the other. Using constraints is a novel, promising concept, which allows to model areas directly with a set of constraints: the area is the infinite point set which simultaneously fulfill all constraints.

Recently, efforts focused to overcome the requirement that all objects in a GIS must be well defined and have clearly defined boundaries (Burrough and Frank 1994). Many objects important for GIS applications have broad transition zones. Fuzzy logic provides but one model and different application areas use different concepts (an overview gives (Burrough and Frank 1995), for the use of Fuzzy Logic in a GIS system (Eastman 1993, Eastman *et al.* 1993)).

Another limitation is the requirement that locations in the GIS are expressed as coordinate in a single fixed system. Information which is not expressible as properties of determined locations cannot be entered. In emergency situations, partial information is often received. For example, an informant may report: "X was seen on the left bank of the river, between A and B". Several such pieces of intelligence constraints the solution and together they may be sufficient to direct a search action.

In design problems, information is often expressed as constraints. The solution is described in general terms and the task is to find a particular solution which fulfills the general description. For example in road design, the location and general form of the road are restricted by constraints, found in national design standards and technical guidelines of agencies. The designer's art is to find a geometric layout which respects all the restrictions and is optimal as measured on some scale (typically least cost is used). Traditionally, the designer used topographic maps and map analysis methods to determine optimal road layout. Today, GIS seems the appropriate tool.

Road design is a complex, multi phase effort. Here, only one phase of a general layout is considered, namely the production of a small number of major alternative road layouts. These layouts must fulfill the technical constraints. They should be ordered using some simple cost estimates to direct further refinement and assessment to the most promising ones. No potential solution must be left out, as the following process is only selecting among the candidates and does not produce new alternatives. The following assessment and decision process are not considered here.

The case study is undertaken to investigate a typical spatial engineering problem, which is well understood, but not immediately solvable with current GIS. A method sometimes proposed uses raster data model and assesses the cost of building a road across each raster cell as a function of current land use. This technique does not capture the geometric restriction of the road, i.e. minimal radius of curves, slopes etc., and yields results which are very sensitive to the selection of the cost parameters.

The problem can be formalized and expressed with a combination of tools from current spatial data models and constraints programming. Constraint programming is a set of techniques and languages especially suited to deal with design problems (Borning, Freeman-Benson, and Willson 1994). Constraint programming has been combined with database techniques to construct *Constraint Databases* (Kanellakis and Goldin 1994). It builds on the current research in the database community, which extends databases with constraints to provide generic tools to address such problems.

Constraints can express areas as a continuum and, therefore, have been used extensively to solve spatial problems (Borning 1981, Borning 1986, Borning *et al.* 1987) and for a general review (Borning, Freeman-Benson, and Willson 1994). The constraints on the road layout from technical restrictions as stated in design standards are used directly as constraints. They include minimal curve radius, maximal slope, and the regular constraints for smoothness (tangent, transition curvess between straight segments and curves). Terrain (land use, topography and geology) is modeled as a continuos function, interpolated from e.g. raster data. Constraints keep the road outside of prohibited areas, while all acceptable areas have an estimate unit cost for constructing a road there.

The structure of the paper is as follows: The next section details the problem and gives examples for typical constraints. Section 3 describes constraint programming and constraint databases. Section 4 develops the solution and gives the reformulated constraints. The paper concludes with an assessment of the method and the limitations of today's systems.

2. DESCRIPTION OF PROBLEM

Assume that a road should be designed to run from A-town to B-city avoiding a number of restricted areas (figure 1). In this section, the constraints for the road layout are explained and the data describing the particular geographic situation are detailed. The goal is to identify all relevant information for the task and reformulate the problem in these terms.

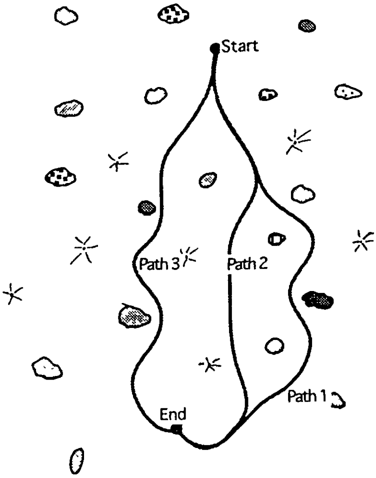


Figure 1 - The situation with three alternative paths

THE CONSTRAINTS FOR THE ROAD

The road designer is given the task to construct a road of a specific road type, e.g. a highway, between two points. This translates to a typical road cross section, with a set of descriptive attributes, namely number and width of lanes, number of breakdown lanes, etc. and the total width of the roadway.

The designer is also given a design speed. Design speed controls most geometric layout parameters, in particular minimal curve radius, maximal slope. Specific values for a highway designed for 100 km/h are listed in table 1 (following Swiss practice (Dietrich, Graf, and Rotach 1975)).

Table 1 Typical Geometric Constraints

Design speed	*100 km/h*
horizontal layout	
minimal length of elements	150 m
maximal length of straight elements	1700 m
minimal radius	425 m
minimal radius for curve connected to straight segment	4000 m
curved and straight segments: ratio of length	0.5 .. 2
sequence of two curves: ratio of radii	0.5 .. 2
clothoid parameter A	180 m
vertical layout	
minimal slope	1%
maximal slope	7%
minimal radius for slope change (for slope differences from 2% to 8%)	
valley:	5 500 .. 30 000 m
hill :	11 000 .. 60 000 m

Rules vary from country to country and the list just demonstrates the types of constraints found in design standards and provide a collection of realistic rules for the case study.

In general, horizontal and vertical layout are independent and can be designed separately. However, certain combinations of situations must be avoided, because they surprise drivers and can lead to accidents:

- no large changes of slopes on short distance (up-down-up-down roads)
- no curve starts behind a hill.

DESCRIPTION OF THE SITUATION

The start and end points are given as fixed locations (coordinate values) and the direction of the road in these points specified. Additional way points may be given.

Certain areas of the terrain are restricted ("forbidden areas"), for example built-up areas, lakes, natural reservation areas etc.. Cost of constructing the road over terrain varies: swamps, for example, cause foundation problems and add much cost. For all areas, cost estimates for building a road there is available ("unit cost").

In order to fit the road to the terrain and to calculate slopes, a digital elevation model gives the terrain height for any location. This can be interpolated from a regular raster or a TIN based Digital Elevation Model.

FORMALIZATION OF THE PROBLEM

The problem can be restated as follows:

Given:

- a digital terrain model, which permits determination of terrain height for any point, $h(x,y)$
- a road unit cost model, which gives the unit cost per road length for any point, $c(x, y)$
- a set of 'forbidden areas', which the road must not traverse,
- a set of way points, through which the road must pass, possibly with the direction the road must have (this set includes at least the start and end point)
- a set of design restrictions for the horizontal and vertical layout, which assure the smoothness of the road and the minimal radii.

Determine:

- substantially different alternative road layouts fulfilling the technical constraints,
- order the alternatives by cost estimates,
- list the best alternatives for further assessment.

The rank ordering must assure that the 'overall best' design is included in the alternatives provided. It is not acceptable that a promising alternative is excluded early, but it does not add much cost, if too many candidates are selected and are later discarded.

CURRENT ROAD DESIGN SOFTWARE FOR THE REFINEMENT

Road design software models the road as a sequence of road design elements. The horizontal and vertical layout is treated separately. For the horizontal layout the elements are straight lines, segments of circles and transition curves between different curves. For the vertical layout, only curves and straight lines are used.

The solution is found using a least squares adjustment technique to determine the unknown elements. Equations in the design parameters and the locations are set up and constraints added. These equations are linearized and solved under an additional optimality constraint using Lagrange multipliers (the sum of the squares of the deviations is made minimal).

This method is very effective to vary a large number of unknown by small amounts to find an optimal solution. The method fails for problems, where non-continuos jumps must be made to find an optimal situation. For example, these least squares based methods will not find an alternate path circumventing an obstacle (as shown as alternate path in figure 1)(Kuhn 1989)

APPROACH

An optimal sequence of actions is

1. determine possible substantially different alternatives as polygon of way points; select the most promising ones based on cost estimates.

2. Refine the candidates to smooth curves and refine cost estimates.

The first step must assess a large number of possible paths, most of which are clearly not respecting the geometric constraints or not close to optimal. Constraint databases are a promising tool to solve this task.

The second step refines the candidates identified in the first step using a least squares approach. Constraints databases could in principle solve this task, but the performance implications are not yet clear (see next section).

3. CONSTRAINT DATABASES

Standard databases allow the storage and retrieval of facts, e.g. the occurence of some species at certain locations, the form and location of an area. Query languages add deductive power, balancing the gain in deductive power against the additional complexity of the retrieval and inference process (Gallaire, Minker, and Nicolas 1984). A relational database can search quicker than a Prolog interpreter, but the expensive power of a relational query language is much less than the expressive power of Prolog. This is the standard trade-off between expressive power and performance. The current standard query languages lack the expressive power to adequately deal with spatial queries of the complexity typically found in a design process (Egenhofer 1992, Frank 1982) (Egenhofer 1991).

Constraint databases offer, within a single integrated database system, support for enhanced handling of data and knowledge. In a constraint database continuos data is directly supported. A single data item (a "constraint tuple") is a constraint on the tuple variables. A traditional data value is a special case of a constraint tuple of the form *Variable List = Value List* or*Var1=Val1 & Var2=Val2 & ...* An example is *<X,Y>* = *<10,5>* or equivalently *X=10 & Y=5.* A region is defined by similar constraints, except that the equality is replaced by an inequality, for example *5 =<X & X+Y =<15 & 3 =<Y.*

Such a simple modeling of continuos data makes the system very easy to extend and modify. Efficient computation of queries such as intersection is made possible by integrating constraint solvers for appropriate classes of constraints (Kanellakis, Kuper, and Revesz 1990, Kanellakis, and Goldin 1994) with novel forms of indexing for multi-dimensional data (Freestone 1993).

Current database query optimizers are designed for relatively simple queries requiring a predetermined number of "joins", "selections" and "projections". Query optimization centers around the order in which these operations should be carried out. By contrast the queries posed in the current application are much more complex. Existing query optimizers are not designed to handle such queries. Experiments with query optimizers (Bressan, Provost, and Wallace 1994) have revealed two limitations: the optimization process itself takes a very long time or fails to terminate, and the resulting query evaluation plan, if it is produced, is very inefficient to execute.

In conjunction with their very general data representation, constraint databases offer intelligent, dynamic query optimization, based on the optimizers developed for constraint programming systems. The uniformity of the data representation, as constraint tuples, simplifies the optimization problem since it is not necessary to treat each different kind of spatial object as another special case.

A standard constraint program comprises of two parts: a declarative part in which the constraints on the problem are stated; and a procedural part in which an algorithm is programmed for searching for (optimal) solutions which satisfy the constraints. As a result of making choices - or during the set-at-a-time evaluation of queries in a constraint database - new constraints are imposed on the problem variables.

Often these constraints are just value assignments *X=N*, where *X* is a problem variable and *N* a value; however, the added constraints can also be more complicated. The simplest way to apply constraints during search is as checks on precise solutions found by the search algorithm. In this case the constraints are delayed until each relevant variable has been assigned a value. In our application this means that each potential route plan for the road is generated, and then it is checked for feasibility. This treatment is, in general, too inefficient for practical applications. The key to constraint programming, and constraint databases, is that the constraints are applied actively to optimize the search for a feasible (optimal) solution.

Constraint programming systems support a number of specialized constraint solvers for handling particular classes of constraints. For handling linear rational constraints, as shown in the above example, a solver based on the Simplex method is built into the system. Thus the query

Find $<X,Y>$ such that $5 =< X \& X+Y =< 15 \& 3 =< Y$ and *region(X,Y)*

can be answered by the system directly, even if region is defined by a set of constraint tuples. The result of the query is, again, a set of constraint tuples, each representing the conjunction of the query condition and a constraint tuple from 'region'.

Constraint programming systems have also been built which incorporate a decision procedure for nonlinear constraints (Monfroy 1992), which would allow direct queries to be stated about the feasibility of roads with curves through regions involving curved boundaries. However, such constraint solvers are computationally very expensive, and are still not viable for real application of this kind.

Nevertheless complex constraints can be handled by "delaying" them, until they can be directly tested or easily solved. For example a constraint limiting the distance from a certain point can be expressed as a quadratic equation $X^2+Y^2 =< Z$. This can be handled efficiently by waiting until the values of X and/or Y and/or Z are known. As an example consider a database in which finitely many $<X,Y>$ pairs were held. The query

Find $<X,Y>$ pairs satisfying $X^2+Y^2 =< 20$

could be processed by calculating for each $<X,Y>$ the value $N= X^2+Y^2$, and returning only those pairs for which $N=<20$.

Applying a decision procedure is the most powerful way to handle constraints, while delaying them until all the variable values are known is the weakest (but computationally cheapest). In between these extremes is a variety of ways of using constraints actively during the search.

To use constraints actively, which is often called "propagation", the query evaluator extracts simple information from them, which can be easily combined with information extracted from the other constraints. The extracted information is often logically weaker than the constraint itself, and so for correctness the original constraint must also be delayed until, at a later point in the search further information can be extracted.

A well-known example of constraint propagation is interval reasoning. Given initial intervals $0<X<10$, $0<Y<10$, $0<Z<10$, and the constraint $X^2+Y^2=Z$, the evaluator can immediately deduce that $0<X<3.17$ and $0<Y<3.17$. These new intervals still do not capture the precise constraint, so the constraint must be delayed until the intervals associated with X, Y or Z become reduced further (as a result of processing another constraint, which may have been added during the search for a solution).

A powerful integration of interval reasoning in constraint programming is presented in (Benhamou, McAllester, and Hentenryck 1994). One form of constraint is a bound on the maximum cost of any solution. This cost can be dynamically reduced as new, better, solutions are found, thus yielding a form of branch-and-bound algorithm. Propagation on this bound constraint corresponds to estimating lower bounds on the cost of the remaining part of the problem. Thus the branch-and-bound algorithm automatically adapts itself to take into account the specific constraints of the problem.

Earlier theoretical work on Constraint Databases focused on embedding constraint solvers in database systems, but more recently constraint propagation has been implemented and tested in a database context (Harland and Ramamohanarao 1993) (Brodsky, Jaffar, and Maher 1993) (Bressan 1994)

4. ROAD LAYOUT WITH CONSTRAINT DATABASE

The road layout problem can be subdivided in several steps, which use different types of knowledge. These steps differ in the complexity of the constraints used and the strength of the restrictions imposed. Performance of the system is affected by the order in which candidates are produced and constraints considered. The following sequence of steps appears optimal:

1. Produce all major path alternatives, using topological reasoning.
2. Select geometrically acceptable path.
3. Assess the path to find a small set of best candidates.
4. Refine approximation to a smooth curve.
5. Refine cost estimate.

The constraint database contains the data consisting of the description of the situation and the constraints and provides an efficient search method to execute the query

find optimal path from A to B.

This reflects the general trend in AI and Computer Science to separate the descriptive specification of a problem from the computational organization to solve it. It makes it easy for the user to change the data and the constraints (e.g. to adapt to a different set of standards) without inference with the generic search engine (Frank, Robinson, and Blaze 1986a, Frank, Robinson, and Blaze 1986b, Frank, Robinson, and Blaze 1986c).

In order to apply this paradigm the representation of the problem must be adapted to the power of the search process. Depending on the representation the search considers a larger or smaller search space and can exclude candidates quicker without detailed exploration. A smaller search space is achieved using a two step process, separating a discrete topological and continuos problem. In figure 1 the major alternative path is immediately recognized. The path can vary in its exact position within a corridor, but the topological relations with respect to the obstacles are fixed.

PRODUCE SUBSTANTIALLY DIFFERENT ALTERNATIVES (CORRIDORS)

Produce a ranked set of substantially different alternative paths which can fulfill the constraints. A substantially different path is represented by a corridor within which the exact path can vary (figure 2). A corridor is acceptable if there is at least one path which fulfills the constraint in it; the paths within one corridor are not 'substantially different', meaning that they can vary without changing the assessment much.

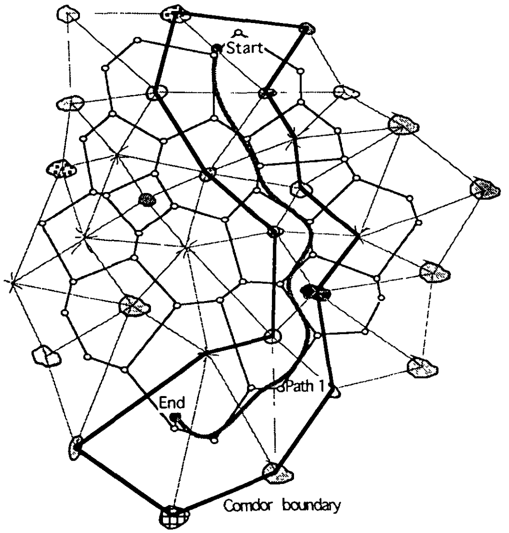


Figure 2 - A path with the corridor for variation

The substantially different alternative paths are created using Delaunay triangulation. The space with the obstacles is triangulated. Then the Voronoi diagram (the dual of the Delaunay triangulation) is created. All substantially different alternative paths consist of the edges in the Voronoi diagram. To each path belongs a corridor, which consists of all triangles from the Delaunay triangulation touched by the path (see figure 2).

The representation of the terrain is based on a triangulation, such that the cost function within the triangle is flat. Variation of the path within the triangle does not affect cost. Considering only the obstacles ("forbidden areas"), as in figure 1, creates a small number of alternatives. If one considers the partition of the cost surface in classes of equal cost and equal height and integrates these boundaries into the Delaunay triangulation (Gold 1994), then the Voronoi diagram contains many more alternative paths. Each triangle contains one way-point and the triangles must be small enough that no more way points are necessary (figure 3).

Forbidden areas can be left out from the triangulation and barriers can be included (constrained Delaunay triangulation). The triangulation must represent the intersection of the cost function with the elevation data, such that variation within one triangle element is bounded.

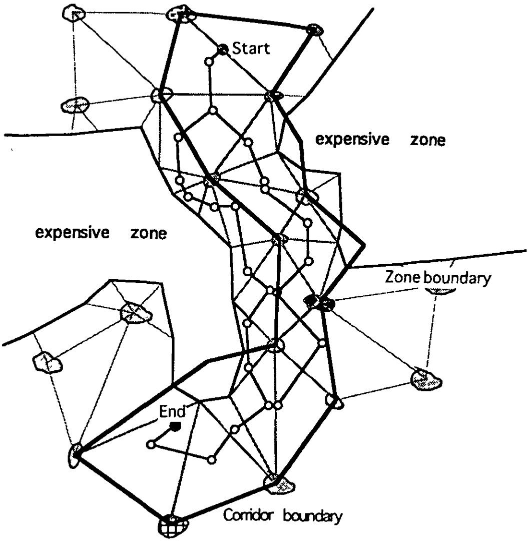


Figure 3 - A finer triangulation, including boundaries of cost zones

ASSESSMENT OF A LINEAR PATH

Roadway alternatives in the Voronoi diagram are described as a sequence of points connected by straight lines (polygon of tangents, Figure 4). It is possible to quickly assess their viability for road design without going to a curve model. Modeling curves is computationally considerably more expensive as it requires to solve problems of higher degree (bi-quadratic or of degree 4).

GEOMETRIC FORM - HORIZONTAL LAYOUT

For each change in direction a segment of a circle of at least minimal radius is necessary. This results in conditions for the length of the two adjoining segments in function of the change in direction. Additional length is necessary for a clothoide to achieve a smooth change in radius (figure 4).

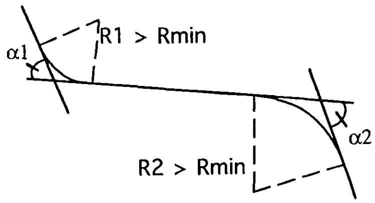


Figure 4 - Minimal length of segment

$$length_{seg}(alpha_1 + alpha_2) = R_{min} * (alpha_1 + alpha_2) + (A^2 * R_{min})$$

Segment length is further restricted by the minimal length. Thus the constraint is

$$length'_{seg}(alpha_1 + alpha_2) = max\ (length_{seg}(alpha_1 + alpha_2),\ minimum_{seg})$$

SLOPE:
Considering the height value for start and end of each tangent together with the length of the line quickly eliminates all paths with slopes much steeper than the maximum slope. For any change of slope the length of the neighboring segments must be long enough to accommodate the gradual change of slope (the calculation is the same as for horizontal layout).

COST OF THE ROADWAY:
The alternative paths are assessed for their cost in order to sort them by increasing cost for further testing. A shortest path within the corridor is constructed by considering the nodes of the triangles and the center points and form a new triangulation, effectively subdividing each triangle in six. In this subnetwork the path with least cost is determined and this value used for the ordering of alternatives. If the road crosses a linear obstacle (river, railway) the cost of special constructions are added.

It is possible to use an A* search algorithm, if we compute a lower cost estimate for the completion of any partial path using Euclidean distance and minimal unit cost. The error of the cost assessment depends on the size of the triangles and can be estimated.

5. IMPLEMENTATION METHODS

The method to construct viable alternatives for further selection described above can be programmed in a standard programming language. The constraints given by the design standards are then coded in by the programmer, possibly allowing changes of the constants in the constraints by the users.

Using a constraint programming language and a database puts the description of the constraints in a format the transportation engineer can understand and change. Nevertheless, the most efficient methods for solving the constraints are used. Most ımportantly, much less programming must be done and therefore the transportation engineer depends less on specialized acquired programs, where the assumptions built in are not visible.

6. CONCLUSION

An interesting and realistic part of the road design process is naturally expressed by the experts as a set of constraints. Using a constraint database for these GIS applications to design applications leads to an expression of the tasks in a formal language understandable to the domain expert.

The road design problem studied here is a realistic example for spatial design problems as many GIS applications include, e.g., in communications (optimal placement of stations for cellular phone system, design of distribution networks for cable based communication etc.), drainage systems and many others.

Fundamental for the formalization of these design problems is the appropriate representation of the problem. In this case study

- space is represented with a Delaunay triangulation, merging the cost surface and the digital terrain model
- the constraınts on the road geometry are simplified to a linear form, leading to much faster constraint resolutions.

The most promising corridors selected in the triangulation are then further refined and assessed to determine the best solution. Constraint programming produces all efficient solutions and assures that none is overlooked.

We are currently working on the implementation of the described methods in a constraints database and to determine performance. Of particular interest is the combination of the first selection based on the triangulation and the second refinement to smooth curve.

REFERENCES

Benhamou, F , McAllester, D , and Hentenryck, P Van "CLP (Intervals) Revisited." In *International Symposıum on Logıc Programmıng ın* 1994.

Borning, A *The Programmıng Language Aspects of ThıngLab, a Constraınt-Orıented Simulation Laboratory* ACM Transactıons on Programmıng Languages and Systems//vol. 3, No. 4, October 1981//pp. 353 - 387: copy, 1981.

Borning, A. *Defining Constraints Graphically* Mantei, M. and Orbeton, P. (Eds.): Human Factors in Computing Systems, CHI'86 Conference Proceedings//Boston, April 13 - 17, 1986//pp. 137 - 143: orig, 1986.

Borning, A. *et al.* . "Constraint Hierarchies." In *Object Oriented Programming Systems, Languages and Applications (OOPSLA'87) in Orlando, Florida, October 1987*, 48-60, 1987.

Borning, Alan, Freeman-Benson, Bjorn, and Willson, Molly. "Constraint Hierarchies " In *Constraint Programming*, ed. Mayoh, Brian, Tyugu, Enn, and Uustalu, Tarmo. 1 - 16. Berlin. Springer Verlag, 1994.

Bressan, S "Database query optimization and evaluation as constraint satisfaction problem solving." In *ILPS Workshop on Constraint Databases in Lincoln*, CSE, Univ. Lincoln Nebraska, 1994.

Bressan, S., Provost, T. Le, and Wallace, M. *Towards the Application of Generalized Propagation to Database Query Processing*. European Computer Research Center, 1994. CORE 94-1.

Brodsky, A., Jaffar, J., and Maher, M. *Towards Practical Constraint Databases.* IBM Yorktown Heights Research Center, 1993

Burrough, Peter, and Frank, Andrew U "Concepts and paradigms in spatial information: Are current geographic information systems truly generic?" *IJGIS* in press (1994).

Burrough, P. A , and Frank, A U. *Geographic Objects with Indetermined Boundaries.* London: Taylor & Francis, 1995.

Dietrich, K., Graf, U., and Rotach, M. *Road Design (Strassenprojektierung).* 2. ed., Zurich. Institut fur Verkehrsplanung und Transporttechnik ETHZ, 1975.

Eastman, J Ron. *IDRISI Version 4 1 Update Manual* Version 4.0, rev. 2 ed., Worcester, Mass.: Clark University, 1993.

Eastman, J. Ron *et al.* *GIS and Decision Making.* Version 4 0, rev. 2 ed., Vol 4 UNITAR Explorations in GIS Technology, Geneva, Switzerland: UNITAR, 1993.

Frank, Andrew U. "Requirements for a Database Management System for a GIS." *PE & RS* 54 (11 1988): 1557-1564

Frank, A. U. "Spatial Concepts, Geometric Data Models and Data Structures." In *GIS Design Models and Functionality in Leicester, UK*, edited by Maguire, David, Midlands Regional Research Laboratory, University of Leicester, 1990

Frank, Andrew U., Robinson, V , and Blaze, M. "An assessment of expert systems applied to problems in geographic information systems." *ASCE Journal of Surveying Engineering* 112 (3 1986a):

Frank, Andrew U., Robinson, V , and Blaze, M "Expert systems for geographic information systems: Review and Prospects." *Surveying and Mapping* 112 (2 1986b): 119-130.

Frank, Andrew U , Robinson, V., and Blaze, M "An introduction to expert systems " *ASCE Journal of Surveying Engineering* 112 (3 1986c):

Freestone, M "Begriffsverzeichnis: A Concept Index " In *11th British National Database Conference in* 1 - 22, 1993.

Gallaire, H , Minker, Jack, and Nicolas, Jean-Marie. "Logic and Databases: a deductive approach." *ACM* 16 (2 1984): 153-184

Gold, Christopher. "Three Approaches to Automated Topology, and How Computational Geometry Helps." In *Six International Symposium on Spatial Data Handling in Edinburgh*, edited by Waugh, T. C., and Healey, R G , AGI, 145 - 158, 1994

Goodchild, Michael F. "A geographical perspective on spatial data models." In *GIS Design Models and Functionality in Leicester*, edited by Maguire, David, Midlands Regional Research Laboratory, 1990

Harland, J , and Ramamohanarao, K. "Constraint Propagation for Linear Recursive Rules." In *ICLP in Budapest*, 1993.

Herring, John R. "TIGRIS. A Data Model for an Object Oriented Geographic Information System " In *GIS Design Models and Functionality in Leicester*, edited by Maguire, David, Midlands Regional Research Laboratory, 1990.

Kanellakis, P., Kuper, G , and Revesz, P. "Constraint Query Languages." In *Principles of Data Systems in* 1990

Kanellakis, P. C , and Goldin, D. Q. "Constraint Programming and Database Query Languages." In *2nd Conference on Theoretical Aspects of Computer Science in* 1994.

Kuhn, W *Interaktion mit raumbezogenen Informationssystemen - Vom Konstruieren zum Editieren geometrischer Modelle.* Dissertation Nr 8897//Mitteilung Nr. 44, Institut fur Geodäsie und Photogrammetrie//ETH Zurich. 1989.

Monfroy, E "Groebner Bases. Strategies and Applications." In Conference on Artificial Intelligence and Symbolid Mathematical Computation in Karlsruhe, BRD, 1992.

Peuquet, D J "A hybird structure for the storage and manipulation of very large spatial data sets." Computer Vision, Graphics, and Image Processing 24 (1 1983): 14-27.

DESIGNING A FLEXIBLE AUTOMATED MAPPING SYSTEM FOR MEETING UNIQUE CUSTOMER REQUIREMENTS

Timothy F. Trainor
Frederick R. Broome
Geography Division
U.S. Bureau of the Census
Washington, DC 20233-7400

ABSTRACT

The methods used for creating maps have changed dramatically during the past ten years. Most map producing organizations have embraced some form of the automated cartographic technology available today. The map production process consists of several discreet components from an initial design concept through the completion of the map in its final form. Technological development has impacted each of the production process components in different ways and at different rates. The effect of this differential application of automation to the cartographic production process is that some parts of the process are fully automated while others, most notably the design component, lag behind in use of automation. A concomitant effect has been a change in the role of the cartographer.

Historically, large map producing organizations printed a standard set of map products for the public and other customers. Tailoring a product for specific customer needs was an expensive activity. The changing nature of cartography and of producing user-specified maps have brought about whole new map production systems.

This paper examines the characteristics of a flexible automated mapping system design for meeting unique customer requirements. Geographic coverage, content, scale, symbology, design, output media as well as efficiency and expense are examples of basic requirements. The issues involved in developing such a system are considered in a theoretic framework. The position of the cartographer also is considered. This leads to some guidelines to aid in training cartographers and designing map production systems to meet present and future map needs.

INTRODUCTION

Historical

Mapping activities are on the rise as more and more data become available and computer software takes over the tasks of processing and output of the data into maps of high cartographic quality. At the same time an apparent shift has occurred from standard map products traditionally produced by large mapping organizations to more user-specified, oftentimes one-of-a-kind maps. While most map producing organizations have embraced some form of the automated cartographic technology available today, the application of automation to the cartographic production process has been differential with automation of some parts of the process lagging behind others. The implication is that mapping institutions are being challenged by the move from established products with cartographic design to customer-need based mapping. This has challenged not only the mapping organization's structure - ways in which they do business - but has effected a change in the role of the cartographer.

Historically, large map producing organizations printed a standard set of map products for the public and other customers. Tailoring a product for specific customer needs was an expensive activity. Usually, tailoring a unique product was left to small staffs that specialized in that kind of mapping. The changing nature of cartography and of producing user-specified maps have brought about whole new map production systems.

Current

The current state of map production is one of flux. Traditional map production philosophy is being challenged on all sides by GIS systems capable of map generation, by desktop computer mapping software, by hand held computer based mapping and data capture systems, and so forth; even by virtual reality systems. The whole concept of what constitutes a map and mapping is being reexamined and redefined. This has resulted in an unclear view of what should constitute a mapping system and what kind of training is necessary for a cartographer, or even the role of a cartographer in the map production process.

Future - The Satisfied Customer

The characteristics of a system to satisfy mapping requirements in today's emerging environment must be founded more on customer satisfaction than on production management. This means developing flexible automated mapping systems designed for meeting a customer's unique requirements. There are five words in this statement that are keys to designing a successful system for the future. Each of these words reflects a design issue in the theoretic framework.

Unique (product type) The map content and design are not necessarily the same as that offered by a standard map series. Need is usually immediate and cannot be easily fit into a regular production schedule. Creative design, perhaps even non-traditional design, is required to effectively communicate the data message. The map may require the use of data from data sets that are not easily or commonly related, and indeed may not even be digital.

Customer (product purpose) The product orientation is toward customer-based need and less toward production organization. The customer is the user of the map production function primarily and the user of the product secondarily; though they may be the same. In neither case is it the emphasis on the producing agent or organization.

Meeting (system purpose) The system purpose is to provide the product when, where, and how needed while maintaining quality.

Automated (means of operation) The map product is produced via any of the three automated means: (1) total batch generation, (2) interactive generation, or (3) batch generation with interactive editing. No traditional manual cartographic operations are performed directly on artwork or product.

Flexible (functionality) The system is able to produce many designs, output to many media, and function within many operating environments. It is not a system that is strongly hardware or operating environment dependent. Nor is the system limited in map design or output media to one or even a few standard formats.

Several of these design issues are present in today's mapping systems. Unfortunately, smartly devised demonstrations coupled with customer imaginations have combined to create an impressive panoply of mapping options that appear to say that the needed mapping systems are here today. This has led to high customer expectations of systems that, in fact, cannot meet their mapping requirements. Customers familiar with traditional map production operations appreciate what can truly be done by today's systems. The experienced customer knows how to test the precise capabilities of systems by asking simple questions such as, "Will you demonstrate how to...?" When this happens, the stark truth of any system's automated flexibility to meet customer's unique requirements is revealed.

MAP PRODUCTION MODELS

At the core of all mapping systems is the map and the operations needed to produce it. In the world of today and tomorrow, the traditional view of what is a map and the terminology applied to map production operations must change. When one examines textbooks and articles about cartography, one is able to find numerous map classification schemes but few definitions of a map. Robinson and Petchenik (1976) described the difficulty of defining a map. Thrower (1972) and Raisz (1938) and others have all attempted to define a map, while acknowledging the difficulty in doing so. Previous definitions, when examined in the light of today, seem hopelessly inadequate and unnecessarily restrictive.

A map is not merely a visual phenomenon. Rather, a map may be considered to exist at the moment that it becomes a unique entity, distinctly separate from the sources and materials needed for its production. Much in the manner that a cognitive map is said to exist, a map today can exist as a series of computer commands that a display device acts upon to produce a visual image.

To a person viewing a map, selected data and their spatial relationships are seen presented in a manner designed to communicate a message. The message may be as mundane as a catalogue of data locations or as complex as a traffic flow map representing vehicle flow-volumes at peak hours. The marginal and other supporting information, such as scale, source note, date of creation, accuracy diagram, graticule, legend, and so forth, are part of the map and add to clarify information in the map image area.

Thus, a map is defined as an entity consisting of a collection of selected data and their spatial relationships, and of marginal and other supporting information, all designed to communicate a specific set of understanding through presentation.

Fundamental Cartographic Operations

Implicit in the definition of a map are the cartographic operations necessary to produce one. Basically, the operations are conceptualizing a design to fulfill the map's purpose, evaluating and gathering source data, manipulating the data to prepare the geographic and thematic map components, map preparation including marginal and other support information, and map production to the desired media and form; each of which is appropriate in both the manual and automated systems.

Conceptualizing a design to fulfill the map's purpose The cartographer relies on experience and skills to conceive of a design that fits the map's purpose. This usually is arrived at by sketching and reworking the basic layout and content of the map. From this is developed a list of what is needed to execute the map and a plan for the production.

Any future systems should allow for just such free-form design work. And the system should be able to identify what sources, symbol libraries, and so forth will be needed to execute the design; or at least assist in the identification.

Evaluating and gathering source data This activity involves an analysis of the content and data relationships of available digital databases. Future systems should have a way of "cruising" through different geographic and thematic databases so the operator can make reasonable judgements as to a database's applicability and difficulty of use for this design. An example from both traditional manual and automated production would be:

> Manual - deciding if a base map must be re-scribed to show only some general background information for use in the final map.
>
> Automated - deciding if selected items must be extracted from a digital geographic database and processed through a generalization operation for use in the final map.

Manipulating the data to prepare the geographic and thematic components Here the cartographer performs the "drafting" operations. These include operations related to, but not limited to, symbolizing, projecting, layout, and so forth. This is the preparation of the basic map image area.

Preparing the map This is where the final touches are added to the layout, where the legend, scale, marginal information, accuracy or quality statement or diagram, source notes, and so forth are added.

Producing the map to the desired media and form This activity includes printing, storage on CD-ROM, display on a screen, or whatever the map's purpose requires.

Large, traditional map production systems have very formal procedures for achieving the steps outlined above. They organize, and schedule them in ways that are quite efficient for the production of large quantities of maps of a standard design. Smaller cartographic units go through the same types of steps, but usually with fewer time consuming formal procedures. Today's automated cartographic software and systems are capable of bringing many of these activities together into even fewer formal, externally managed steps. Indeed, that is one of automation's strengths.

A MAPPING SYSTEM FOR TOMORROW

Automation has witnessed a revolution in output media. Magnetic tape slowly slips to museum displays alongside paper tape and punch cards while CD-ROM, electronic communication of images, and even virtual reality move to the forefront. Crude monochrome display screens are now replaced by ones with much higher resolution and mega-color capability.

The impact of output media on map design limits options and capabilities more than any other component of an automated mapping system. Maps are designed for the final output device (as they were designed for the final print media characteristics in the past). The output media now are plotters, display screens, and printing presses. Digital map files are embedded with the peculiarities of the intended output device. A map file used on multiple output devices carries with it design embellishments and flaws of the primary device for which it was designed.

Design issues

The design of tomorrow's mapping systems must be guided by several key concerns. The three most significant address issues related to learning, using, and map quality.

Ease of learning The software, and whole production system, must be easy to use (intuitive). Having a system that requires the user to read volumes of manuals or attend expensive mandatory training classes is not only user unfriendly, but counter productive to producing quality maps.

Labor dependence of software The number of systems designed around interactive mapping activities far exceeds the number of systems using a batch method for map production. The software tool has replaced the manual pen and scribing tools, but at what benefit? Without question, automation is the better alternative in most situations and by its very nature provides vast opportunities for map production and cost savings. The number and variety of maps produced by automation surpasses those previously produced by conventional methods. The time required to construct the map and the labor and materials costs associated with conventional map production contributed to the push into automation. The issue is to reduce the dependence on human interaction within mapping systems.

Many of the software packages today can and do place text on maps. Few, if any, place the text in cartographically appropriate and ascetically pleasing positions. Most require extensive intervention on the part of the human operator to handle situations such as overlap, rotation, and leader line or arrow placement. Not only are these operations not handled automatically by the systems with only limited human intervention, they are not even recognized by the systems as needing human action. This means humans must search for text placement problems. Exasperating as this is, many systems then require a difficult series of commands that must be executed just to adjust even a single piece of text.

Map quality Future systems must be able to ensure quality to the point where the map message is communicated accurately and clearly. The systems must be able to warn the user when a fundamental cartographic rule is being violated, even if the software cannot automatically fix the problem. The system must be capable of providing alternative solutions to poor design decisions made by the user.

Components

Parts that comprise the entirety of the mapping system extend beyond the mere combination of hardware and software that form the working tools of the map production process. There are distinct program modules, and systems within systems that interact in complex ways to effect a simple solution (procedure). Examples of components include: user requirements (specifications); various mapping program modules; symbol libraries; editing functions; education and training mechanisms; computer hardware; and a production control system.

User requirements Those characteristics that satisfy the customer are the user requirements. The user has a purpose for the map. User requirements are stating how these will be fulfilled.

Mapping program modules Generally, the more modular the mapping functions, the more flexible the mapping system. Mapping programs perform the functions necessary to access and process the data, including any required links with internal and external data sources such as symbol libraries and statistical data sets respectively. Mapping programs prepare the data for cartographic output. Chaining together data that share the same cartographic symbolization is an example of data preparation. Maintaining intelligence about each of the component parts of the chain for future processes exemplifies higher-level programming modules.

Symbol libraries. Symbols are graphic representations assigned to classed data and, if designed well, contribute to the map readers' understanding of the map message. Previously, symbols were designed and produced based on the emphasis of the data item as well as the interaction with other symbols. This action oftentimes is a function of map scale.

Today, symbol design is further complicated by the limitations of the output device. The same dot size and density of an areal screen on a high resolution plotting device is very different on a low resolution screen. Conversely, symbols differentiated by a 256 color palate monitor loose their discriminating meaning on a monochrome laserwriter.

Symbol use also is affected by automation. The order and sequence of applying cartographic symbols are more complicated as the rules of computer systems are fixed, even where multiple iterative attempts are allowed and applied. The result is a stringent interpretation of encoded rules whereas a cartographer, even in an interactive mode, applies "cartographic license" based on what is known, perceived, and felt.

Editing functions In order to accommodate map requirements that exceed the built-in capabilities of the mapping system, or to adjust the map image to maximize quality, editing functions enable the user to improve those capabilities or to alter the map image. Ultimately, the less editing required, the more effective the production capacity of the mapping system.

Education and training Intuitive use of a system is most desirable. As more complex functions are added to systems, there is a tendency to complicate the learning required to simply use the system. Regardless, a mapping system should provide users with whatever information (and examples) are needed to maximize its use. Weeks and months of training should not be a prerequisite for use.

Computer hardware An assumption for current and future mapping system development is open architectures. Systems that are device dependent are doomed to fail over the long term. Both software and hardware must be portable in order to serve the greatest number of users. A user of a mapping system should not have to write a program to convert anything. Those conversions should be embedded as a function of the system. This does not mean a pull-down menu of choices in which the system inevitably fails to support the user's requirement. Rather, the system should recognize the data structure, process the information correctly, and import the map data.

Production control system The creation of multiple maps usually warrants a system that monitors the mapping process. The system not only reports the whereabouts and status of the production flow, but it also reports map errors via automated editing procedures implemented throughout the production cycle. Production control is an excellent weather vane on the efficiency of the mapping system. Automated checks on the process contribute to future development requirements.

THE CARTOGRAPHER'S POSITION

The dramatic change in methods for creating maps has effected a change in the role of the cartographer. Not only has the availability of commercial mapping software given everyone with a desktop computer access to make their own maps, but even in larger map production facilities, the power of automation has increased the number of map makers who are not trained in cartography or related fields. This begs the question, "What is the position of the cartographer in future map making operations?" There are several possible answers.

Possibility 1. Designs, initiates, implements, and uses automated cartographic systems in one or more of the following ways: determines requirements; specifies content and system functionality; writes computer program modules; tests system effectiveness; provides mapping and system development expertise to support staff; instructs mapping system users; and designs, compiles, and produces maps using the system. (This is the most proactive role for the modern mapmaker in which the knowledge, skills, and experience of the cartographer direct and guide mapping system technology and assure it is founded on sound cartographic principles).

Possibility 2. Involved in developing and implementing cartographic design rules within the software. (This is a participatory role in mapping system development but it leaves aside many valuable skills that the cartographer can contribute to the map production operation. An example of a skill in early phases of mapping is a knowledge of source materials and how to integrate them into the final product).

Possibility 3. Developer of guidelines for software users to apply when evaluating the quality of their final map product. (The cartographer is on the outside looking in and is passively participating in a manner which depends heavily upon the software user to take the extra step for quality, an action not common to human nature).

Possibility 4. Withdraw from the active production of maps and deal only with research into the meanings, messages, and methods of cartography, that is, make it an abstract academic research specialty that provides professional criticism of mapping systems. (Cartographers become observers and drop out of the mapping process. This seems to be the line of least resistance, but ultimately costs the most. Inevitably, this could reduce cartography to a sub-topic under graphic arts within the fine arts department).

Possibility 5. Increase the involvement of cartographers in more aspects of education at all levels, even non-geographic fields of study, so that future users will have or understand the need for cartographic principles and quality. (This requires an extensive effort on the part of the field. It also demands a long term commitment. However, the payoffs are highest, indeed greatest, to both the discipline and users of maps as a means of communication. The involvement should be implemented regardless of which position the cartographer assumes in future endeavors).

TRAINING TOMORROW'S CARTOGRAPHERS

Fundamentals, Fundamentals, Fundamentals

Cartography, as art and science, mandates a renaissance approach to a diverse discipline.

Formal cartographic training A well-rounded education in map design and production processes (both historical and modern) provides a sound foundation for map-related work, particularly mapping applications. The evolution of an effective map from creativity of the map theme through the visualization of a symbolized image carries terms that are subjects unto themselves: source; perception; point, line, and area; scale; color; font; ink; file; resolution; printing; accuracy; plot; software; geographic; dimension; projection; label; paper; symbol; placement; (tele)communication; interpretation; and others. Skill in each of these areas is necessary.

Geography Map subjects are infinite. Cartographers should learn basic facts about the map theme in order to design it appropriately. General knowledge of geography complements facts, content, and the map purpose and improves the effectiveness of imparting the map message.

Logic and mathematics The ability to solve complex problems (usually without formulas, theorems, and equations) is a normal mapping responsibility. Identifying a problem and/or a system shortcoming followed by a problem-solving, step-by-step approach that leads to a solution, is a required staff trait for mapping system development. Efficiency also is important but is secondary to problem-solving. Problem-solving at this level is based on the ability to deal with individual details while keeping in view the "whole."

Computer science Automation is the tool. A cartographer must know, appreciate, and respect the capabilities and limitations of the mapping environment. Computer programming skills in one or a variety of forms is highly desirable. Knowing basic functions such as how a machine draws a line or links two discreet elements aids in understanding how those same machines perform what appears to be more complex actions. It allows the cartographer to speak intelligently and effectively about the automation of cartographic operations.

Communication skills. Most of what has been described requires exchange of information, ideas, concepts, plans and so forth. Customer-based cartography, by its very nature, requires extensive interaction on the part of the participants. Effective written and oral communication ability is important. The term "effective" includes the ability to draw out of the customer the map requirements without the user having to learn cartographic technology.

Automation Awareness

Technology continues to evolve. Ten years ago, the computing environment was very different from today, and with time, more options will become available to users and customers. Technology has improved at an increasing rate which limited the scope of specialization on the part of developers.

Cartographers are affected by the same technical pressures. In order to be effective in automated cartography today a cartographer does not necessarily have to be an expert in programming. Rather, it is important that the cartographer knows enough about programming and the capabilities and limitations of different working automated mapping systems. For example, knowledge of the advantages and disadvantages of particular output (display) devices is desirable to make intelligent symbol choice decisions and recommendations. Likewise, a cartographer should know enough about database operations and capabilities as well as how to examine a digital source for content or be able to assess a database's value for a particular cartographic operations.

Cartographers in the traditional mapping environment were skilled in the use of multiple tools and methods for producing artwork. In order to be functional today, cartographers should be skilled in the use of two (2) or more mapping software packages. These skills do not replace formal cartographic education. The intent is not to train cartographers to be operators of two or more packages, but rather to allow the cartographer to experience firsthand how cartographic concepts are executed by the software packages as a result of initiating the program commands. Each package has been designed to perform specific functions, and do so with variable rates of success. The goal for today's cartographer is to design, using the best functionality of systems, an automated mapping system that meets customer demands.

SUMMARY

Designing automated mapping systems is challenging. Altering the goal to include the design of flexible automated mapping systems to meet unique customer requirements causes a rethinking of traditional approaches to current mapping software. The greater the degree of automation, the more complicated the implementation of technical development. Meeting customer requirements is not a static exercise. Continual reassessment of the mapping functions are, and will continue to be necessary.

The field of cartography is experiencing a second wave of change in automation. The shift from traditional mapping techniques to those which used computers was the first phase. A redefinition of the role of cartographers in a world where mapping capabilities are available to users at their desktops currently challenges a profession and an industry in which each is struggling for identity and excellence. This should not be a struggle for dominance, but one of cooperation for quality. The various aspects presented here of an automated system should be the basis of a balanced design that leads to cooperation. Together they provide promise for development of user-friendly, customer-based systems that produce high quality, accurate maps on demand.

REFERENCES

Raisz, E. (1938) General Cartography, McGraw-Hill Book Company, Inc., New York, New York, U.S.A.

Robinson, A. H. and Petchenik, B. B. (1976) The Nature of Maps, University of Chicago Press, Chicago, Illinois, U.S.A.

Thrower, N. J. W. (1972) Maps and Man, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, U.S.A.

NATIONAL AND ORGANIZATIONAL CULTURES IN GEOGRAPHIC INFORMATION SYSTEM (GIS) DESIGN
A TALE OF TWO COUNTIES

Francis Harvey
Department of Geography, DP-10
University of Washington
Seattle, WA 98195

ABSTRACT

How does culture influence GIS design? How is the cartographic discipline involved in GIS design in different cultures? The research presented in this paper builds on comparative information science work on culturally mediated differences in work values and the comparative analysis of information systems. Research from Hofstede and others identifies four dimensions of culture: power distance, uncertainty avoidance, individualism, and masculinity (Hofstede 1980). Power distance and uncertainty avoidance are most important for information systems. Two county GIS designs are examined based on this framework. The research presented here leads to the identification of culturally influenced differences between the county GIS designs of King County, Washington, USA and Kreis Osnabrück, Germany.

INFLUENCE OF CULTURE ON GIS DESIGN

In GIS research, the information science literature is largely read from a cognitive viewpoint (for example Mark 1989; Frank 1992; Nyerges 1993; Frank 1994). Another body of literature in the information sciences has considered the broader context of information system design and use based on comparative cultural analysis (Hofstede 1980; Eason 1988; Jordan 1994; Tate 1994; Watson, Hua Ho et al. 1994).

Authors in the GIS field recognize larger structures that mediate information system design and practice. De Man (1988) explicitly recognizes the role of culture in influencing the meaning and value of information. Aronoff, writing about implementation, identifies the organizational context of GIS as critical not only to the operation, but for the essential identification of functions that the GIS is to fulfill (Aronoff 1989). Chrisman develops the most complete framework for understanding the influence of culture on GIS to date through custodian institutions and mandates (Chrisman 1987).

Information systems are designed to provide information for enlightened decision making. In a wider social context the meaning and value of information are determined by culture (Weber 1946). In their examinations and ruminations about the use and design of maps, many authors clearly refer to culture's importance. Cartographers, examine this from various perspectives (Turnbull 1989; Freitag 1992; Muehrcke 1992). This literature is very general and does not reflect information system work. Specifically developing the influence of culture on information system design and implementation is not possible from this literature. Recent literature from scholars with a surveying and computer science background examines the issues of culture in spatial comprehension and GIS use (Campari and Frank 1993; Campari 1994).

Work examining the culturally mediated influence of the cartographic discipline on GIS design must elaborate the cultural and institutional framework. One path to study cultural influences would highlight the role of disciplines as the vehicle of cultural expression. Chrisman explicitly discusses the disciplinary and organizational context for the design of GIS and understanding the role of cartography in mediating cultural conceptualizations of spatial 'reality' in his 1987 paper. This extension of the cartographic communication model (Robinson and Petchenik 1975; Morrison 1978) describes the connection between map and reality formed by culturally manifest mandates and institutions. Culture is a framework that contextualizes individual and institutional values, meaning, and consequently the processes through which decisions are made. The study of cultural influences in GIS design and implementation should reflect the social role, situation, and power of disciplines.

The key question examined in this paper is how the cartographic discipline mediates national and organizational cultural values and meaning in GIS design. A discipline can function overtly, as a guild, or indirectly, through a network of relationships developed around professional and personal criteria. This examination limits itself to the role of cartography, a core discipline in the GIS field.

Cultural comparisons are complex. Any cultural characteristics of information conceptualization and exchange must be based upon comparatively validated theoretical approaches.

This paper applies the findings of information science research to a comparison between two county GIS designs in King County, Washington, USA and Kreis (County) Osnabrück, Lower Saxony, Germany. The focus is on overall system design and cultural differences manifest through the distinct involvement of the cartographic discipline. Detailed documents on their respective system designs form the basis of this comparison. Experience working at the communal level on GIS projects in both countries augments this comparison.

The next section of this paper will proceed by giving the reader an overview of each county's GIS design. We will follow this by a section that describes the theoretical framework for the comparison of information systems. The results of the comparison are presented in the third section and a summary of the findings and conclusion follows.

TWO COUNTIES — TWO GIS

King County and Kreis Osnabrück started their GIS design work in 1989. Kreis Osnabrück is the smaller of the two countries, both in size and population.

The essential difference in the GIS design documentation lies in Kreis Osnabrück's reliance on national standards (ATKIS, ALK, MERKIS) for the design of their county GIS, whereas King County is developing their GIS from the ground up. ATKIS - *Automatisierte Topographisch-Kartographische Informationssystem* (Automated topographic and cartographic information system) is the most important. It is the object orientated data model for provision of vectorized topographic data at three scales: 1:25,000, 1:200,000 and 1:1,000,000. ALK - *Automatisierte Liegenshaftskataster* (Automated property cadastre) is the automatization of the Grundbuch (Property Book). The *Maßstaborientierte Einheitliche Raumbezugsbasis für Kommunale Informationssysteme* (Map scale orientated uniform spatial coordination basis for communal information systems)describes GIS at the communal level as a "... geographic data base for agency specific, spatial communal information systems based on the national coordinate system, a unified data model for all topographic and agency specific spatial data, . . ." (Landkreis Osnabrück 1990).

King County's GIS is implemented through a project that involves design issues. Its design starts with a needs assessment (PlanGraphics 1992a) followed by a conceptual design document that describes most data layers along with source, conversion method, and maintenance responsibility (PlanGraphics 1992b). Major features and tabular attributes describe shared data base layers, but without any detailed data modeling (PlanGraphics 1992b). The county council supported a trimmed down version of the unified county GIS and a core project to provide the shared base layers was started. When the project is completed, they will pass on most responsibilities for maintenance to the appropriate agencies. However, a central group is foreseen, whose exact functions are not identified. Beyond the basic lists of features and attributes, the project develops the work on data modeling and agreements with the participating agencies. There is no common understanding of what the county GIS is based on or should provide.

The following table summarizes key design differences between the two counties.

Kreis Osnabrück	**King County**
Organizational	
Information system department of the county government is lead agency. Various working groups are coordinated by a newly created position. GIS data base design is carried out in the responsible agency based on the ATKIS model (Landkreis Osnabrück, Der Oberkreisdirektor 1993b).	Information system department of county transit agency (recently merged into County government) is lead agency. Two committees accompany the project. GIS data base design is coordinated with other agencies, municipalities, and corporations (Municipality of Seattle 1993).
Purpose	
Provision of data and information for more efficient administration and planning at the communal level (Landkreis Osnabrück 1990). Needed improvements are identified by agency and function (Landkreis Osnabrück, Der Oberkreisdirektor 1992; Landkreis Osnabrück, Der Oberkreisdirektor 1993b).	The core project provides capabilities for diverse agencies and purposes that are vaguely defined, ie. "better management of". Project goals are limited to development of the county GIS data base.
Budget overview	
DM 2.89 million (app. USD 1.94 million) (Landkreis Osnabrück, Der Oberkreisdirektor 1993b)	USD 6.8 million (Municipality of Seattle 1993)
Data model (Base layers)	

Provided and defined largely by the national standards ATKIS, ALK, and MERKIS. Extensions are for county purposes and already listed in the object catalog. Agencies can extend the data model when needed in a given scheme.	No explicit data modeling in the conceptual design documents. 72 layers in all Core first five of 14: Survey Control Public Land Survey System Street Network Property Political
Disciplinary involvement	
The use of ATKIS/ALK as basic data base model and requirement that data "fit" this data base (MERKIS) make it necessary to involve cartographers. These standards require other specialists to perform data modeling that fits these standards.	Cartographers and GIS specialists hold key positions in project management. These disciplines and surveyors, environmental scientists, and other specialists (foresters, biologists) are spread throughout other levels of the project.
Information collection, analysis, and display	
Documents describe administrative procedures and source maps in detail, but not how they are performed/used in a GIS-based automatization of procedure (Landkreis Osnabrück, Der Oberkreisdirektor 1992; Landkreis Osnabrück, Der Oberkreisdirektor 1993c).	Documents sometimes identify rough costs are(Municipality of Seattle 1993), but no detailed requirements, sources, procedures of any kind are identified. Only general tasks are described (PlanGraphics 1992b).

THEORETICAL COMPARATIVE FRAMEWORK

In the information science field several works have been published that empirically study the involvement of culture in information system design. This work is generally informed by the sociological work of Max Weber. In his broader system of sociology he establishes the role of culture through its shared set of beliefs that influence what we consider to be meaningful and valuable. Disciplines (professions) and institutions in modern bureaucratic society nurture and transmit these values and meanings (Weber 1946; Weber 1947). Obermeyer (1994) recently discussed the role of professions in GIS in Weber's framework. Chrisman, writing about the involvement of different disciplines and guilds in spatial data handling, also identifies disciplines as carriers and transmitters of cultural values (Chrisman 1987).

To establish the critical role of culture and its effects on GIS design a more explicit, comparative study is on order. Essential to this study is the identification of core cultural values that directly affect GIS design. In the information systems field, unique and key work that empirically establishes these factors was published by Geert Hofstede.

Hofstede published the results of a study of 117,000 questionnaires sent to 84,000 participants in 66 countries examining the role of culture in work-related values (Hofstede 1980). Applying theories of culture and organizational structure from Weber

(Weber 1946) and Parsons (Parsons and Shils 1951; Parsons 1951) to Geertz (1973) and Kluckhohn (Kluckhohn 1951; Kluckhohn 1962) to the research findings, Hofstede (1980) establishes four dimensions of national culture:

- **uncertainty avoidance** the extent to which future possibilities are defended against or accepted
- **power distance** the degree of inequality of power between a person at a higher level and a person at a lower level
- **individualism** the relative importance of individual goals compared with group or collective goals
- **masculinity** the extent to which the goals of men dominate those of women.

Uncertainty avoidance is the focus of information systems, decision support systems, etc. (Jordan 1994). It is considered together here with power distance because of interaction effects (Hofstede 1980). Due to the similarity in this variable between Germany and the USA it alone is not particularly significant, but important in a wider context with power distance. It is important to note that Hofstede's findings ascribe ideal typical qualities and ideals to each culture in a Weberian sense: they are the strived for forms, not individual characteristics nor implementations.

Organizational Form, Implicit Model of Organization, Problem Solving Approach, and National Culture

	High Uncertainty Avoidance	Low Uncertainty Avoidance
Low Power Distance	Workflow bureaucracy Well-oiled machine Regulations German speaking, Finland, Israel	Implicitly structured Market Negotiations English speaking, Scandinavia, NL
High Power Distance	Full bureaucracy Pyramid Codified hierarchial Latin Japan some other Asian	Personnel bureaucracy Family Diffuse hierarchy Southeast Asian

Figure 1 Uncertainty avoidance and power distance. Each box lists the organization type, the ideal model for an organization, the ideal problem solving, approach and national culture associations. (After Hofstede, 1980 p. 319)

Uncertainty avoidance and power distance form critical interactions affecting organizations. In Germany and the USA, where power distance is low, there are two possibilities how to keep organizations together and reduce uncertainty. If there is high uncertainty avoidance (German speaking), "people have an inner need for living up to rules, ... the leading principle which keeps the organizations together can be formal rules" (Hofstede 1980 p. 319). With low uncertainty avoidance (Anglo), "... the organization has to be kept together by more ad hoc negotiation, a situation that calls for a larger tolerance for uncertainty from everyone" (Hofstede 1980 p. 319). Figure 1 shows important organizational characteristics based on fourfold division based on uncertainty avoidance and power distance dimensions. Hofstede comments the figure in detail. The "Anglo" cultures "would tend more toward creating 'implicitly structured' organizations" (Hofstede 1980 p. 319). In contrast, German speaking cultures establish 'workflow' bureaucracies that prescribe the work process in much greater detail (Hofstede 1980 p. 319). Hofstede argues that problem solving strategies and implicit organization forms follow: Germans establish regulations, Anglo-Americans have horizontal negotiations. Germans conceive of the ideal type of well functioning organization as a "well-oiled machine," whereas Anglo-Americans conceive of a "village market" (Hofstede 1980). However insightful cultural factors are, they cannot explain all the differences between organizations.

Even proposing explanatory theories of organizational behavior is a manifestation of the cultural values in which they were written. Comparing Weber's (German), Fayol's (French), and Follett's (American) writing on the exercise of authority in bureaucracies the influence of cultural values is clear, "Weber stresses the office; Fayol, the person; Follet, the situation" (Hofstede 1980 p. 323).

Information transaction cost theory (Willamson 1975) provides additional insight into cultural influence on organizational structure and approaches to problem solving. In this theory, all business activity is transaction between individuals and groups. Information serves as the controlling resource (Jordan 1994). In this form the theory is overly reductionist and simplistic. Boisot extended transaction cost theory to include cultural issues, distinguishing two characteristics of information that affects transactions:

- **codification** the degree of formal representation
- **diffusion** the degree of spread throughout the population (Jordan 1994)

Internalizing the transaction in the organization reduces the diffusion of information (Jordan 1994). Centralized information requires a bureaucracy, whereas diffuse information is distributed in a market. These differences correspond to Hofstede's work (Jordan 1994) and are crucial for comparing and assessing GIS designs. How GIS design codifies or diffuses information will depend on the importance of uncertainty avoidance and ideal organization type. Multi-disciplinary, multiple goal orientations (Radford 1988) will have additional hurdles to face in information system design.

Nominally, highly integrated industries and commerce apply the information transaction approach. GIS design approaches often begin with a similar structured systems approach (Gould 1994). When considering heterogeneous public administrations, a different, highly diversified organizational structure is possible. In county governments the multi-disciplinary interests, missions, goals, and perspectives require consideration of the cultural values that influence the information system.

COMPARISON

The mediation of cultural values through cartography in King Count and Kreis Osnabrück is clear from the respective GIS design documents. In Kreis Osnabrück rules and clearly defined components and procedures are implemented, based on the organizational form of workflow bureaucracy. Effective decison making in Kreis Osnabrück comes after regulations. King County documents leave the coordination and many questions of design open, to be determined through a process of negotiations in an implicit organizational structure.

The following overview, based on the structure as the description of GIS design above, provides a condensed description of differences between the two counties GIS designs.

Kreis Osnabrück	King County
Organizational	
The EDP group in the county is "lead"; agencies work on "own" data but coordination is maintained through coordinator, working groups, and administrative procedures.	The core project implements conversion according to negotiated data model and quality standards and hands the converted data over to the respective `stakeholder' agencies.
Purpose/Intent	
Development of communal information system: a collection of geographical data and methods and procedures for the systematic collection, updating, processing, transforming, and distribution of all data	Provision of base layers of county GIS to fulfill agency-identified functions in coordination with county agencies.
Disciplinary influence	
Definition of standards and guidelines for organization of data and training of personnel through national groups (ATKIS/ALK and MERKIS)	Implicit through backgrounds and experiences of key personnel. They have been involved with Wisconsion MPLIS projects and local area projects, including the City of Seattle's GIS data base design.
Data organization	
Defined in object catalog based on ATKIS (Landkreis Osnabrück, Der Oberkreisdirektor 1993a)	Developed through negotiations with key agencies
Information structure, analysis, and display	
Defined by standards, agency, legal requirements, and procedure	Described in documents and developed through negotiations with agencies.
Information formalization and control (codification vs. diffusion)	
Centrally codified in an object catalog, with defined ranges for additions through individual agencies.	Ongoing, defined through a pilot project and negotiations with `stakeholders'
Re-engineering or adoption of new techniques and procedures (uncertainty avoidance)	
GIS is used to automate existing procedures. The exact usage of GIS is not clarified for individual procedures and agencies.	Focus is on products. Procedures are open for change. Implementation is developed through negotiations with agencies

Considering the diffusion/codification of information, low/high uncertainty avoidance, and negotiation/regulation approach we can summarize these differences in problem solving strategies. The GIS design of King County foresees the distribution of data and information among participatory agencies after completion of the project (stakeholders with market shares), low uncertainty avoidance (integration will be worked out later), and ongoing negotiations to clarify fundamental and detail questions (including a pilot project to identify key issues). On the contrary, Kreis Osnabrück bases their GIS design on a codification of information (catalog of geographical phenomena based on ATKIS), high uncertainty avoidance (full description not only of data types, but of their use in administrative procedure), and regulations to prescribe form and function of the GIS.

However, in Kreis Osnabrück some administrative regulations may require special data collections. They should be based on ATKIS/ALK, but a mechanism for checking this may not be in place. Thus, the question of the integration of specific data, ie. animal preserves (*Tiergehege*), with ATKIS/ALK remains. If large conceptual differences remain, integration may require substantial interdisciplinary efforts.

Hofstede also identifies characteristics of information system design strategies in terms of implicit models of organization: the "market" for American culture and the "well-oiled machine" for German culture. King County's GIS design makes the concurrent negotiation of critical design issues during implementation necessary. Kreis Osnabrück, on the other hand, designs around standards, the critical issues in implementation are making the technology do what the regulation and procedure require. This is perhaps overly reductionist, but as a pastiche is highly illuminative of the complex fundamental differences between these two cultures' GIS design approaches.

In these two cases, divergent roles of the respective cartographic discipline are evident. In Kreis Osnabrück, cartographers have been key in defining the national standard for geographical databases, ATKIS. These are the basis for the data model designed by the GIS design group. The core project in King County has no such national standards, but the backgrounds of key management personnel in MPLIS work and local GIS work at the local and regional level, ensure the implicit influence of design strategies and approaches from the broader professional context of cartographers and GIS.

SUMMARY AND CONCLUSION

Culture is a broad framework for understanding important contextual factors in GIS design. At the level of national culture, we identify substantial differences in the conceptualization and design of information strategies to aid decision making and the culturally mediated involvement of cartography in GIS design. The role of cartography in directly or implicitly influencing GIS design is evidenced in the role of respective national standards. In the US, the Federal government implements national standards, primarily to regulate transfer, ie SDTS. In contrast, national groups in Germany prepare standards that effect the detailed aspects of GIS design, ie. ATKIS.

Mediation of cultural values in GIS design, implementation, and use are directly or informally manifested. In Kreis Osnabrück, cartographers and surveyors play the key role through the standardization of basic spatial data in ATKIS and ALK. These form the foundation for much of the county's GIS work. King County's core GIS gets around the issue of standardization by agreeing in committees to develop base layers for the county GIS that best serve the common interest and represent the stakeholders, without predefining the integration of any of the stakeholders' data. Implementation and use of the county GIS in King County hinges on ongoing relationships and negotiations between the stakeholders.

The use of GIS for decision making in both counties will hinge on the success of the county GIS design to implement diverse organizational and cultural conceptualizations of geographical phenomena. The integrative role of GIS hinges on the capability of different disciplines to amalgamate their 'world views'. Over time, will a more predefined topographic model or cartographic representation work more efficiently than a heterogeneous, non-conformal model?

Further research based on case studies (Onsrud et al. 1992) should examine the variance between design and use. Case studies of implementation can provide specific insight into implementation and variance from design, opening important insight into the direct and indirect influence of culture and discipline in GIS.

ACKNOWLEDGMENTS

Mr. M. Balikov and Mr. K. Johansen, King County and Mr. Remme, Kreis Osnabrück graciously provided essential documents. We gratefully acknowledge this support. Discussions with and comments to a draft version from Prof. N. Chrisman were of great value in the writing of this paper.

REFERENCES

Aronoff, S. (1989). *Geographic Information Systems: A Management Perspective.* Ottawa, WDL Publications.

Boisot, M. (1987). *Information and Organizations: The Manager as Anthropologist.* London, Fontana.

Campari, I. (1994). GIS commands as small scale space terms: cross-cultural conflict of their spatial content. Sixth International Symposium on Spatial Data Handling, Edinburgh, Scottland, p. 545-571.

Campari, I., A. F., Frank (1993). Cultural Differences in GIS: A Basic Approach. EGIS '93, Genoa, Italy, p. 1--16.

Chrisman, N. R. (1987). "Design of Geographic Information Systems Based on Social and Cultural Goals." Photogrammetric Engineering and Remote Sensing **53**: 1367.

De Man, W. H. E. (1988). "Establishing a GIS in relation to its use. A process of strategic choices." International Journal of Geographic Information Systems **2**(3): 245-261.

Eason, K. (1988). *Information Technology and Organisational Change.* London, Taylor & Francis.

Frank, A. (1992). "Spatial Reasoning - Theoretical Considerations and Practical Applications." The European Conference on Geographic Information Systems 1992. p. 310-319..

Frank, A. U. (1994). Qualitative temporal reasoning in GIS - ordered time scales. Sixth International Symposium on Spatial Data Handling, Edinburgh, Scottland, p. 410-430.

Freitag, U. (1992). "Societies without maps." *Kartographische Konzeptionen, Cartographic Conceptions*, Berliner Geowissenschaftlich Abhandlungen, Reihe C, Kartographie, Band 13. Berlin, Selbstverlag Fachbereich Geowissenschaften, Freie Universitat Berlin. p. 277-286.

Geertz, C. (1973). The Interpretation of Culture: Selected Essays. New York, Basic Books.

Gould, M. (1994). "GIS Design: A Hermeneutic View." Photogrammetric Engineering and Remote Sensing **60**(9): 1105.

Hofstede, G. (1980). *Culture's Consequences. International Differences in Work-Related Values*. Beverly Hills, Sage Publications.

Innes, J. E. a. S., David M. (1993). Implementing GIS for Planning: Lessons from the History of Technological Innovation. APA Journal, p. 230-236.

Jordan, E. (1994). National and Organisational Culture Their Use in Information Systems Design, Working Paper Series (No. WP94/08), Faculty of Business, City Polytechnic of Hong Kong.

Kluckhohn, C. (1951). "Values and value-orientations in the theory of action: An exploration in definition and classification." *Toward a General Theory of Action.* Cambridge, MA, Harvard University Press.

Kluckhohn, C. (1962). "Universal categories of culture." *Anthropology Today*. Chicago, University of Chicago Press.

Landkreis Osnabrück (1990). *Das Kommunale Raumbezogene Informationssystem (KRIS). Eine Arbeitspapier zur Realisierung*. Osnabrück, Referat A.

Landkreis Osnabrück, Der Oberkreisdirektor (1992). *Situationsbericht*, Der Oberkreisdirektor.

Landkreis Osnabrück, Der Oberkreisdirektor (1993a). *Fachkonzept (Entwurf)*. Osnabrück, Der Oberkreisdirektor.

Landkreis Osnabrück, Der Oberkreisdirektor (1993b). *Losungsvorschlag*. Osnabrück, Der Oberkreisdirektor.

Landkreis Osnabrück, Der Oberkreisdirektor (1993c). *Systemkonzept*, Der Oberkreisdirektor.

Mark, D. (1989). Cognitive image schemata and geographic information: Relation to user views and GIS interfaces. GIS/LIS '89, Orlando, FL.

Morrison, J. (1978). "Towards a functional definition of the science of cartography with emphasis on map reading." American Cartographer **5**: 97.

Muehrcke, P. and M., Juliana (1992). *Map Use. Reading, Analysis, Interpretation.* Madison, JP Publications.

Municipality of Seattle (1993). King County GIS Scoping Project. Seattle, Municipality of Metropolitan Seattle.

Nyerges, T. (1993). "How Do People Use Geographical Information Systems?" *Human Factors in GIS*. eds. Medyckyj-Scott, D. and Hearnshaw, H., London, Belhaven Press.

Onsrud, H. J., Pinto, J. K., et al. (1992). "Case study research methods for geographic information systems." URISA Journal **4**(1): 32-44.

Parsons, T. and Shils, E. A. (1951). Toward a General Theory of Action. Cambridge, MA, Harvard University Press.

Parsons, T. (1951). The Social System. London, Routledge & Kegan Paul.

PlanGraphics (1992a). *King County GIS Needs Assessment/Applications, Working Paper.* Frankfort, KY, PlanGraphics.

PlanGraphics (1992b). *King County GIS Conceptual Design.* Frankfort, KY, PlanGraphics.

Radford, K. J. (1988). *Strategic and Tactical Decisions.* New York, Springer-Verlag.

Robinson, A. H. and B. B. Petchenik (1975). "The map as a communication system." Cartographic Journal **12**: 7-15.

Tate, P. (1994). "Hands across the borders." Information Week. 188.

Turnbull, D. (1989). *Maps are Territories. Science is an Atlas.* Chicago, University of Chicago Press.

Watson, R. T., T. Hua Ho, et al. (1994). "Culture, A fourth dimension of group support systems." Communications of the ACM **37**(10): 45-55.

Weber, M. (1946). *From Max Weber: Essays in Sociology.* New York, Oxford University Press.

Weber, M. (1947). *The Theory of Social and Economic Organization.* New York, The Free Press.

Willamson, O. E. (1975). *Markets and Hierarchies: Analysis and Antitrust Implications.* New York, Free Press.

Evaluating User Requirements for a Digital Library Testbed

Barbara P. Buttenfield
NCGIA
Department of Geography, 105 Wilkeson
SUNY-Buffalo, Buffalo, NY 14261

GEOBABS@UBVMS.CC.BUFFALO.EDU

ABSTRACT

The development of widespread capabilities for electronic archival and dissemination of data can be coupled with advances in information systems technology to deliver large volumes of information very fast. Paradoxically, as greater volumes of information become available on electronic information networks, they become increasingly difficult to access. Nowhere is this situation more pressing than in the case of spatial information, which has been traditionally treated as a 'separate' problem by archivists, due to complexities of spatial ordering and indexing. A research project recently funded by NSF will address these problems and implement a working digital library testbed over the next four years. This paper will focus upon one aspect of the testbed, namely evaluating user requirements to inform interface design. The paper presents the evaluation plan, using hypermedia tools to collect real-time interactive logs of user activities on the testbed under design. Conventionally, interactive logging is analyzed by deterministic measures of performance such as counting keystrokes. In this project, the interactive log data will be analyzed using protocol analysis, which has been shown to provide a rich source of information to formalize understanding about semi-structured and intuitive knowledge.

INTRODUCTION

The development of widespread capabilities for electronic archival and dissemination of data can be coupled with advances in information systems technology to deliver large volumes of information very fast. As greater volumes of information become increasingly available on electronic information networks, they become increasingly difficult to access. This paradox calls for research to implement intelligent software that provides and preserves access to electronic information, creating a digital library for bibliographic and analytical use (Lunin and Fox, 1994). A research project recently funded by NSF (the Alexandria Project) will address these problems and implement a working digital library testbed over the next four years. The project requires assessment of user needs, basic research to address technical impediments, software development, and a rigorous program of evaluation and quality control.

This paper will focus upon one important aspect of the testbed, namely evaluating user requirements to inform interface design. Challenges associated with interface design for distributed data have been reviewed in Kahle et al (1994). This paper argues for a return to empirical evaluation of GIS interfaces using alternative paradigms to the psychophysical designs once popular in cartographic research.. A brief chronology of empirical evaluation research in cartography is presented. Semi-structured interviews are presented as an alternative paradigm for eliciting cartographic knowledge. This is followed by a description of the Alexandria Project, emphasizing user requirements and evaluation.

The paper will present the user evaluation plan, which involves real-time interactive logging of user activities using hypermedia tools to simulate the look-and-feel of the testbed under design. Conventionally, interactive logging is analyzed by deterministic measures of performance such as counting keystrokes, recording time units between system dialog and user response, etc. These types of analyses are useful but limited in their neglect of user cognition. Interactive user logs collected for this project will be analyzed using Protocol Analysis, which has been shown to provide a rich source of information to formalize understanding about semi-structured and intuitive knowledge (Ericsson and Simon 1993). Its application to interactive logs has not been reported in the literature.

EMPIRICAL EVALUATION IN CARTOGRAPHIC RESEARCH

Empirical research has largely disappeared from the cartographic literature following dissatisfaction with the psychophysical methods that figured prominently for several decades (roughly, 1965-1985). To be clear, the dissatisfaction lay not with the analytical methods, which tend to be highly structured, highly deterministic, and highly confirmatory. The dissatisfaction lay rather in the limited gains in understanding intermediate stages in the process of cartographic communication, meaning those stages of reasoning that occur intermediate between identifying visual elements (seeing map items) and reaching an interpretation (inference of pattern). Similar concerns were expressed at this time in other disciplines.

> After a long period of time during which stimulus-response relations were at the focus of attention, research in psychology is now seeking to understand in detail the mechanisms and internal structure of cognitive processes that produce these relations... This concern for the course of the cognitive processes has revived interest in finding ways to increase the temporal density of observations so as to reveal intermediate stages of the processes. Since data on intermediate processing are costly to gather and analyze, it is important to consider carefully how such data can be interpreted validly, and what contribution they can make to our understanding of the phenomena under study. (Ericsson and Simon 1983, p.1)

In cartography, the stimulus-response paradigm brought forth one insight consistently highlighting the sensitivity of experimental results. Estimates of symbol size were shown to vary with symbol clustering and with inclusion of basemap enumeration units (Gilmartin, 1981; Slocum, 1983). Estimates of grayscale were shown to be dependent upon screen texture (Castner and Robinson, 1969; Kimerling, 1975; Leonard and Buttenfield, 1989) as well as the visual ground on which figures appeared (Kimerling, 1985). Color identification has been shown to vary with respect to adjacent colors (Brewer, 1994) and even with viewer expectations that are unrelated to map attributes (Patton and Crawford, 1979). Several articles (Gilmartin, 1981b; Shortridge and Welch, 1980) summarize some of these sensitivities, and articulate the concern of the discipline during this time.

It is unfortunate that after this period many cartographers turned away from empirical evaluation research (but there are exceptions, for example Lloyd, 1994). For a time the published results of map evaluation were trivialized in passing remarks that cartographic researchers had become obsessed with the question "How big IS that graduated circle?". Although a body of research on interface design (for example, Laurel, 1990), on spatial conceptualization and spatial reasoning (for example, Egenhofer, 1992; Golledge, 1991; Lloyd, 1994) has been published, empirical user evaluation is pursued largely outside the GIS discipline (see for example an excellent brief review in Nyerges, 1993).

Concurrent developments in GIS software and user interfaces have not been supported with empirical evaluation. Many user requirements studies precede GIS system development (Calkins and Obermeyer 1991), however there is little or no formal evaluation once system components are implemented. System refinements are often directed towards improving performance and efficiency rather than system use. One reason for this may relate to the complexity of GIS system use. Nyerges (1993, p.48) comments "GIS applications tend to be quite involved ... in rather rich problem settings. Hence realistic task models will most likely be rather complex. Real world analyses of GIS use tasks are needed, in addition to the selective focus of controlled experiments in the laboratory."

Clearly, there is more to knowing how people come to understand what they experience than can be provided by strict adherence to models of stimulus-and-response. In the absence of adopting novel paradigms for evaluative research, user evaluation studies in cartography have languished.

AN ALTERNATIVE PARADIGM FOR USER EVALUATION

The foundation of evaluation research is observation. Observational data analysis in the stimulus-response paradigm is documented by magnitude estimation, determination of equivalence or difference, or other measures. Most often these are metric, and their goal is to uncover perceptual patterns such as visual clustering, contrast, brightness, size, and so on. Other types of observational data analysis capture higher order cognitive patterns, such as selecting one path of action from several options, recounting steps taken to complete a process, and verbalizing a set of criteria used to solve a problem. Associated data measures are less deterministic, less structured, and must be elicited using some degree of introspection. Ericsson and Simon (1983, p.48-62) recount an extensive history establishing their position that semi-structured introspective reports including think-aloud and immediate retrospective reports can in fact reflect intermediate cognitive processes, although they caution (p. 56, 62) on the difficulty of evaluating un-structured introspection (eg. free association).

Several observational methods for evaluating semi-structured information have been presented in the literature (Sanderson and Fisher, in press). Content analysis (Krippendorf, 1980) tends to focus (as its name implies) on the substantive meaning or content of user responses. Sense-making analysis (Dervin, 1983) is used to facilitate user reconstruction of a procedural task from memory. Protocol analysis (Waterman and Newell, 1971; Ericsson and Simon, 1984, 1993) has been chosen for this project as it is often used to study an ongoing process and is of relevance to evaluating computer interface usage. It is of special utility for studying the intermediate stages of decision-making, and for eliciting knowledge about a decision while subjects are involved in a decision-making process.

"Protocol analysis ... is a common method by which the knowledge engineer acquires detailed knowledge from the expert. A 'protocol' is a record or documentation of the expert's step-by-step information processing and decision-making behavior." (Turban, 1990, p. 462) Protocol analysis was developed as a systematic methodology designed to treat semi-structured behavior and dialogs as quantitative data.

Sanderson et al. (1989) report uses for their semi-automated protocol analysis software (SHAPA) to analyze problem-solving styles of doctors and nurses, crew communication during military surveillance missions and navigational tasks, program debugging in robotics, and elicitation of domain knowledge in industrial management. A more recent version of this software (MacSHAPA) has just been released (Sanderson, 1994) which extends previous analyses and incorporates real-time video control.

Examples applying Protocol Analysis to spatial behavior and human-computer interaction can also be identified. Lewis (1982) reported on the use of 'think aloud' Protocol Analysis in support of computer interface design. Mack et al. (1983) report on a study of word-processor interface design using Protocol Analysis. Golledge et al. (1983) connected results of Protocol Analysis to a theory of children's spatial knowledge acquisition. Lundberg (1984, 1989) used Protocol Analysis to analyze various types of marketing and consumer behavior. Sumic (1991) used Protocol Analysis in dissertation research to elicit expert knowledge from a utility company's electrical engineers, to support the development of a knowledge-based system linked to ARC/INFO. The company (Puget Sound Power and Light) was impressed enough with his work that he was hired full-time to continue his research and maintain the company's knowledge base. He reported on this work at an ARC/INFO workshop on expert systems attended by the author (Sumic, personal communication, 1991). Finally, Gould (1993) used SHAPA to analyze geographic problem-solving using maps of Puerto Rico.

ALEXANDRIA - THE DIGITAL LIBRARIES PROJECT

Project Overview

As stated above, the adoption of distributed data archival and retrieval can deliver large volumes of information very fast to any user on the network. Under these conditions, procedures for organizing, browsing, and retrieving such information become increasingly complex. Nowhere is this situation more pressing than in the case of spatial information, which has been traditionally treated as a 'separate' problem by archivists, due to complexities of data volume, spatial ordering, indexing, and spatial and temporal autocorrelation. For this reason, many libraries separate their text and literary archives from map archives, effectively prohibiting the cross-referencing of literary with graphical holdings. NSF, ARPA and NASA recently issued a collaborative solicitation to provide $24 million in research funds for six projects to develop "digital libraries", software testbeds demonstrating intelligent browsing and retrieval methods for digital data of any kind. One of the six awards was made to a research team including all three sites of the NCGIA.

The Alexandria Project will create, implement, and evaluate a distributed, high performance system for examination, retrieval, and analysis of spatial information from digital collections. A major goal is to remove distinctions between text and spatial data archives (or at least make those distinctions transparent to users). Alexandria will continue for the coming four years, building its testbed in two stages. The first stage will be a prototype based on commercial off-the-shelf software, to be completed within the first year of the project. In the second stage, development of the testbed will proceed in parallel with user evaluation studies to inform system engineers and designers.

Details and Plans for User Evaluation

As with any software engineering problem, understanding user requirements is of primary importance to build an effective user interface (Fox et al, 1994; Laurel, 1990). Evaluation of the Alexandria testbed will include several classes of users, including those familiar with the testbed contents (geographers, earth and space scientists, and professionals in public and private sector who work with spatial data) and those familiar with library cataloging and indexing systems (data archivists, research librarians, map librarians, and so on). Either class of users can be characterized as people whose knowledge of either the geographic domain or the archival domain is deep, but whose interest in learning system architecture or command structure may be minimal. For example, a query by the geographic user class might focus upon browsing satellite imagery to learn more about deforestation within a fixed distance of a river channel over several rainy seasons. A query in the archival user class might focus upon browsing through recent map acquisitions to determine if new editions exist for a given map sheet. Evaluation testing will be performed using both classes of potential users. A third class of users will be considered to include system designers, characterized as having a deep interest in geographic content, archival, AND system design. Evaluation testing will be applied to this class as a control group.

One function of the evaluation process is to determine whether characterization of user requirements is appropriate. Initial requirements are straightforward. For successful access and query of maps and satellite data, users will require individual image/map sheet control for custodial functions (acquisition of data), cataloging and index control (collection maintenance), and bibliographic control (for research and map making). A workable spatial data browsing system will require functions supporting georeferencing and fusion of multiple data types, at multiple spatial, spectral and temporal resolutions. Some requirements are important to one but not both classes of testbed users.

It is probable that access to the testbed and to the data will modify user requirements. A second function of the evaluation process is to determine and track such changes. Interface design and evaluation must be fluid and dynamic. Repetitive testing protocols will be developed to address these issues. That is, interface evaluation will begin with completion of the first stage Alexandria prototype. Evaluation will include questions targeting possible changes in user requirements, necessitating re-testing of some but not all test subjects. Results of the first round of evaluation will be returned to system designers to guide refinement and revision of the interface. As each version of the revised interface is completed, its components will undergo empirical evaluation, with results of the analysis informing subsequent revision. Three or four cycles of evaluation are planned through the course of the project.

A third function of the evaluation process relates to metadata and data quality. Users will need to know the reliability of information returned on queries, thus the evaluation must capture user confidence as well as user satisfaction. The bottom line of this part of the evaluation process must answer two questions. The questions are first, "Does the user get the information as requested?" and second, "Does the user avoid information which is not needed?" Answers to these may not be straightforward in all situations, nor for all three user classes, which will challenge the evaluation and analysis.

Data collection for the evaluation will be accomplished by videotaping user behaviors, by interactive logging of keystrokes and response times, and by semi-structured interview, in an electronic version of think-aloud reporting. Videotaping (direct observation) is intended to capture nonverbal responses and to provide insights about user learning styles, frustration and fatigue, for example. Interactive logging will provide quantitative data measuring response times, keystrokes and mouse activity, and indicate aspects of user and system performance. Semi-structured interviewing will provide information on user requirements and on user confidence and satisfaction levels. Testing will proceed early on at the testbed development site, at UC-Santa Barbara, and proceed from there to otehr sites designated as Alexandria partners. These partners include federal and public libraries, including Library of Congress, USGS Headquarters Libraries in Reston Virginia, and the St. Louis Public Library, and various academic map libraries around the nation. Some testing will be scheduled at various GIS conferences, human-computer interface conferences, and library science conferences planned for the coming four years.

The actual mechanism for evaluation will be an operational but simulated interface, in early versions of the system. Screens with look-and-feel capabilities identical with the testbed will simulate the user interface screens and functions. Data subsets will be embedded to experiment with query and response functions for limited portions of the spatial archive. Running 'underneath' the simulated interface will be a set of interactive logging mechanisms recording keystrokes and mouse placement, documenting response times, etc. The logging mechanism will include a dialog function to converse with the user in semi-structured question-answer mode. The dialog function will be triggered throughout the evaluation session, either by the system or by the user. System triggers will initiate dialogs during specific tasks (during file retrieval for example, or on completion of a query formation). User inactivity over a threshold timeframe will also trigger a system dialog, to ask the user about confusion, or task success, for example. Users will be able to initiate dialog as well, and encouraged to keep a journal of their activities and impressions. These dialogs will be saved in a relational database for subsequent processing by Protocol Analysis.

"Protocol analysis is notoriously difficult and time-consuming to perform." (Sanderson et al. 1989, p. 1275). Application of protocol analysis involves the following steps: choice of an encoding vocabulary, definition of encoding rules, encoding of data using Protocol Analysis software, and inter-coder reliability checks. The vocabulary provides components on which user behaviors and dialogs may be categorized, and the encoding rules impose those categorizations on the data. The analysis is conceptually similar but not identical to principal components analysis, in the sense of looking for underlying patterns, rather than in the sense of a data reduction technique. Key patterns will appear consistently in one or more user classes, and inform system designers about how interface components are used by specific user classes, under what circumstances, and conditions and/or user functions that are confusing or problematic. Key patterns identified in early iterations of the user evaluation will be applied and revised for later evaluation experiments.

The software to be utilized for analyzing the Alexandria data is MacSHAPA (Sanderson, 1994) described earlier in the paper. This software can link the dialog protocols with videotape excerpts, which will tie a record of nonverbal with verbal reports of interface use. The final step in Protocol Analysis involves inter-coder reliability checks to determine if the components and categories have been applied consistently to all test subjects. Inter-coder reliability checks insure objectivity and repeatability of analytical results. "Careful protocol analysis is time-consuming, and extensive analyses require automatization. A considerable increase in objectivity may occur, since the analysis will be accomplished with determinate rules..." (Waterman and Newell, 1971, 285).

SUMMARY

Efforts to evaluate user activity and user interface design have been largely absent from cartographic and GIS research, and it is argued this is due in part to dissatisfaction with results of research based on the stimulus-response paradigm once popular in cartography. Alternative paradigms based on observational data analysis can provide a data source for studying higher level cognitive processes involved with learning a GIS interface, including user confidence and satisfaction. One example of this type of paradigm is Protocol Analysis, which will be adopted to evaluate the user interface for the Alexandria Project, a software testbed for intelligent browsing of distributed digital spatial databases. The testbed will be under development for the coming four years, and interface evaluation studies will run concurrent with system development. The evaluation plans have been outliend in the paper, and include videotaping (direct observation) linked with interactive logging techniques underlying a simulated system interface. The logging techniques will include not only the customary deterministic measures (counting keystrokes, etc.) but also incorporate the semi-structured dialog and interview methods practiced in Protocol Analysis. Three classes of users will be evaluated, including users of spatial data (geograpy and earth science professionals), archivists of spatial data (library and information management professionals), and system designers and engineers.
It is felt that such evaluation can only improve the flexibility of system interface design, and additionally assist researchers in formalizing some types of cartographic knowledge.

ACKNOWLEDGEMENTS

This research forms a portion of the Alexandria Project, funded by the National Science Foundation (NSF contract IRI 94-11330), and a portion of Research Initiative 8, "Formalizing Cartographic Knowledge", of the National Center for Geographic Information and Analysis (NSF contract SBR 88-10917). Funding by NSF is gratefully acknolwedged.

REFERENCES

Brewer, C.A. 1994 Guidelines for Use of the Perceptual Dimensions of Color for Mapping and Visualization. **Proceedings,** IS&T/SPIE Symposium on Electronic Imaging Science and Technology, Color Hard Copy and Graphic Arts III Technical Conference, 8 February 1994, San Jose, California (no page nos. on manuscript).

Calkins, H.A. and Obermeyer, N.A. 1991 Taxonomy for surveying the Use and Value of Geogrpahical Information. **international Journal for Geographic Information Systems**, vol. 5(3): 341-351.

Castner, H.W. and Robinson, A.H. 1969 Dot Area Symbols in Cartography: The Influence of Pattern on Their Perception. **Technical Monograph No. CA-4.** Washington D.C: American Congress on Surveying and Mapping.

Dervin, B. 1983 An Overview of Sense-Making Research: Concepts, Methods, and Results to Date. Presented International Communication Association , Dallas Texas, May 1983 (no page numbers on manuscript).

Dervin, B. and Nilan, M.S. 1986 Information Needs and Uses. In M.E. Williams (Ed.) **Annual Review of Information Science and Technology**, vol. 21: 3-33.

Ericsson K.A. and Simon, H.A. 1993 **Protocol Analysis: Verbal Reports as Data.** Cambridge Mass: MIT Press (2nd Edition).

Ericsson K.A. and Simon, H.A. 1984 **Protocol Analysis: Verbal Reports as Data.** Cambridge Mass: MIT Press (1st Edition).

Fox, E.A., Hix, D., Nowell, L.T., Brueni, J., Wake, W.C. Heath, L.S., and Rao, D. 1994 Users, User Interfaces, and Objects: Envision, a Digital Library. **Journal of the American Society for Information Science**, vol.44(8): 480-491.

Gilmartin, P.P. 1981a Influences of Map Context on Circle Perception. **Annals of the Association of American Geographers**, vol 71: 253-258.

Gilmartin, P. P. 1981b The Interface of Cognitive and Psychophysical Research in Cartography. **Cartographica**, vol.18(3): 9-20.

Golledge, R.G. 1991 The Conceptuall and Empirical Basis of a General Theory of Spatial Knowledge. In Fischer, M.M. Nijkamp, P. and Papageorgiou, Y.Y. (eds.) **Spatial Choices and Processes.** Amsterdam: North Holland: 147-168.

Golledge, R. G., Smith, T. R., Pellegrino, J. W., Doherty, S., Marshall, S. P. and Lundberg, G., 1983 The acquisition of spatial knowledge: Empirical results and computer models of path finding processes. Presented at XIX Inter-American Congress of Psychology, Quito, Ecuador, July 1983 (no page numbers on manuscript).

Gould, M. D., 1993 **Map Use, Spatial Decisions, and Spatial Language in English and Spanish.** Unpublished Ph.D.dissertation, Department of Geography, State University of New York at Buffalo.

James, J. M. and Sanderson, P. M., 1991 Heuristic and Statistical Support for Protocol Analysis with SHAPA Version 2.01. **Behavior Research Methods, Instruments and Computers**, vol. 23(4): 449-460.

Kahle, B., Morris, H., Goldman, J., Erickson, T., and Curran, J. 1994 Interfaces for Distributed Systems of Information Servers. **Journal of the American Society for Information Science**, vol.44(8): 453-467.

Kimerling, A.J. 1985 The Comparison of Equal Value Gray Scales. **The American Cartographer,** vol. 12(2): 132-142.

Kimerling, A.J. 1975 A Cartographic Study of Equal Value Gray Scales for Use with Screened Gray Areas. **The American Cartographer,** vol. 2(2): 119-127.

Krippendorf, K. 1980 **Content Analysis: An Introduction to its Methodology**. London: SAGE Publications, vol. 5 (The CommText Series).

Laurel, B. (ed.) 1990 **The Art of Human-Computer Interface Design**. Reading, Massachusetts: Addison-Wesley.

Lewis, C. H., 1982 Using the "Thinking Aloud" Method in Cognitive Interface Design. **IBM Research Report** RC-9265. Yorktown Heights, NY: T.J. Watson Research Center.

Lloyd, R. 1994 Learning Spatial Prototypes. **Annals of the Association of American Geographers**, vol. 84(3): 418-439.

Lundberg, G., 1984 Protocol analysis and spatial behavior. **Geografiska Annaler,** vol. 66B(2): 91-97.

Lundberg, C. G., 1989 Knowledge acquisition and expertise evaluation. **The Professional Geographer,** vol. 41(3): 272-283.

Lunin, L. F. and Fox, E.A. 1994 Perspectives on Digital Libraries. **Journal of the American Society for Information Science**, vol.44(8): 441-445.

Mack, R. L, Lewis, C., and Carroll, J., 1983 Learning to Use Word Processors: Problems and Prospects, **ACM Transactions on Office Information Systems,** vol.1(3): 254-271.

Sanderson, P.M. 1994 **MacSHAPA**. Version 1.0.2. Champaign-Urbana, Illinois: Department of Mechanical Engineering, Univeristy of Illinois. Distributed by CSERIAC, Wright-Patterson Air Force Base, Ohio.

Sanderson, P.M. and Fisher, C. (in press) Exploratory Sequential Data Analysis: Foundations. **Human-Computer Interaction**, vol.9(3).

Sanderson, P. M., James, J. M., and Seidler, K. S., 1989. SHAPA: an interactive software environment for protocol analysis. **Ergonomics** vol. 32(11): 1271-1302.

Shortridge, B.G. and Welch, R.G. 1980 Are We Asking the Right Questions? Comments on Cartographic Psychophysical Studies. **The American Cartographer**, vol.7: 19-23.

Slocum, T.A. 1983 Predicting visual Clusters on Graduated Circle Maps. **The American Cartographer**, vol. 10(1): 59-72.

Williams, R.L. 1958 Map Symbols: Equal Appearing Intervals for Printed Screens. **Annals,** Association of American Geographers 48:132-139.

Turban, E., 1990 **Decision Support and Expert Systems: Management Support Systems**. New York: Macmillan.

Waterman, D. A. and Newell, A., 1971 Protocol Analysis as a Task for Artificial Intelligence. **Artificial Intelligence** vol.2: 285-318.

SPATIAL DATA BASES: DEVELOPING A MEANINGFUL INTERNATIONAL FRAMEWORK FOR SPATIALLY REFERENCED CULTURAL AND DEMOGRAPHIC DATA

Leslie Godwin
Geography Division
U.S. Bureau of the Census
Washington, DC 20233-7400
Tel: (301) 457-1056
lgodwin@census.gov

ABSTRACT

Development of a framework for defining, representing, and ultimately exchanging spatial features has naturally focused on physical or political entities in pursuit of the objective of creating and sharing heterogeneous, spatial data bases. Entity data elements, including their definitions, attributes, and relationships, have been developed and refined for physical entities at national levels. The United States has issued both the Spatial Data Transfer Standards and the Content Standards for Digital Spatial Metadata. Similar standards have been issued in other nations.

Although the importance of physical entities remains unquestioned, another aspect of spatial data bases is beginning to receive much needed attention. The heretofore "forgotten component" of spatial data bases is their cultural and demographic component, i.e. their human dimension. Definitions for cultural and demographic data sets are currently being developed in several countries. As with physical entities, the framework for these cultural and demographic components recognizes the importance of data categorization, topics and characteristics classification, and the geographic unit of coverage and temporal component which are required for identification and use of these data.

As the exchange of information through spatial data bases crossed international boundaries and became worldwide, research at the international level initially focused on the application of sharing national concepts of physical entities across nations. The seemingly straightforward task of defining physical entities has proven to be quite complex in the international arena. For example, one nation's concept of a water body proved to be different than another given cross-cultural, cross-linguistic, and cross-disciplinary comparisons (Mark, 1993). Defining cultural and demographic data at the international level will, at the least, be no less daunting a matter. Addressing the broadening of national cultural and demographic data definitions to meet international needs is a timely issue as the development of most national cultural and demographic data frameworks are at an early stage.

This paper provides some guidelines for achieving an international set of standards for describing cultural and demographic data. Since the majority of the decisions taken on behalf of nations or between nations involves components of the human dimension, removing obstacles to sharing cultural and demographic data in an international environment of heterogeneous spatial data bases is imperative. The success of exchanging complete spatial data sets will enhance the abilities of

researchers and decision makers to achieve more equitable and meaningful decisions.

INTRODUCTION

Standards for spatial data sets serve as the basis for understanding, interpreting, and exchanging the data. Standards for spatially referenced data sets typically address the definitions, attributes, and relationships of and between the entities within a spatial context. Sometimes the standards will include an explanation of the theoretical model used to develop the standard. Sometimes the standards will emphasize definitions by offering detailed glossaries of terms related to applications and/or levels of generalization. The entities are the building blocks which provide the framework for constructing higher level entities such as physical features, cultural features, and so forth. The value of a standard rest in how well it describes and conveys the information contained in a data set.

Two kinds of standards for spatially referenced data are being developed, those that describe how to encode the data for exchange and those that describe the content of the data set. The latter is referred to as a metadata standard and is of primary importance here. A number of national standards for spatially referenced, i.e. geographic, data sets are either complete and have been issued or are nearing completion. To analyze and describe the work worldwide is beyond the scope of this paper. Experience indicates that similar problems exist everywhere when it comes to developing geographic standards. Therefore, the work in the U.S. as represented by the standards issued through the Federal Geographic Data Committee (FGDC) and the work in South Africa as represented by the South African National Standard for the Exchange of Digital Geo-Referenced Information (SANSEDGI) will be the primary examples described herein. The FGDC's Spatial Data Transfer Standard (SDTS) and the Content Standards for Digital Geospatial Metadata will the two examples from the U.S.

One very important and diverse data category of spatially referenced data is cultural and demographic data. Cultural and demographic data center on the "human dimension" and cover an extremely wide range of topics. Topics such as agriculture, business, communications, customs, economics, education, the environment, as well as many other aspects of our daily lives are examples. Standards for describing and exchanging cultural and demographic data are at the early stages. Typically, groups interested in cultural and demographic data are using as their starting point existing national standards for geospatial data.

When comparing the characteristics of spatially referenced cultural and demographic data sets with current geospatial standards, two items of importance arise. Their importance is heightened even more when this comparison includes cross-cultural and cross-linguistic considerations. First is the absence of a theoretical model for clarifying the meaning of cultural and demographic data similar to the model generally accepted for other geospatial data. The development of such a model is a necessary foundation to adequately describing the fundamental entities of cultural and demographic data. Second, given the nature of the data, is how to accommodate the lack of precision in terminology used to describe data sets and data components. This imprecision exists between technical fields, even within the same culture, but is made more imprecise when the terms are from different cultures, languages, and so forth. Any cultural and demographic data set standards must allow data producers the freedom to either reference generic definitions or provide specific definitions while giving data users easy access to the definitions, and do this without placing an undue burden on either the data producer or the data user. This paper proposes a model for cultural and demographic data and suggests a framework for developing an international metadata standard for spatially referenced cultural and demographic data.

THE TIMELINESS OF AN INTERNATIONAL STANDARD ON CULTURAL AND DEMOGRAPHIC DATA

Spatial data is a necessary and integral part of an information system and misunderstandings about the spatial framework can have a major impact on the presentation of associated data. However, the importance of data sets to their users lies in a user's ability to display and/or analyze data about topics of concern and interest against a geographic framework. One type which has received little attention, but has a major impact on each of us, is cultural and demographic data. These data center on the "human dimension." Indeed, the majority of the economic, political, and societal decisions made daily are made based upon cultural and demographic data.

Differing perceptions of cultural and demographic data coupled with the fact the data collector, data producer, and data user may believe they have a precise understanding of this data impacts our lives. Cultural and demographic data are a cornerstone to the decision making process used by policy makers at all levels of governments and in all governments. Further, decisions are no longer based on information gathered just within the borders of the country making the decision, but rather, increasingly many decisions are being made that cross international boundaries and effect the citizens of the world.

Data sets from many cultures are accessible given today's computer and communications technology. Internet provides access to an increasing quantity of data via a few keystrokes. Addressing cultural and demographic data standards which meet international needs is a timely issue as data sets become available internationally. The development of most national cultural and demographic data standards are at an early stage, and attention has remained focused on the geographic framework standards and is only now beginning to shift to the importance of spatially referenced cultural and demographic data. The lack of attention this type of data receives is evident by browsing through the Internet with the assistance of the many available search tools. A recent, informal browse achieved only three matches with "human dimensions" (interestingly, all were related to the human dimensions of global environmental change.) While there were many matches to "cultural" and "demographic," none of the information pertained to classifying, clarifying, and exchanging such data. Recognizing national standards for cultural and demographic data are likely to be produced, it is important that they be developed on the best international model possible. An understanding of the international characteristics will help build a foundation for cultural and demographic data that will cross cultural lines.

DEVELOPING A MODEL

A Conceptual Data Model of Geospatial Entities

The U.S. Content Standards for Digital Geospatial Metadata was developed by the FGDC to provide a common set of terminology and definitions for the documentation of geospatial data (FGDC, 1994). Successful creation of this standard was due in part to the independent but cooperative work in many organizations and the many public discussions occurring prior to finalizing the standard. These actions led to a broad concensus on content and a theoretical starting point. This concensus contains an unwritten understanding or perception of the fundamental units of geospatial data. Most persons working with geospatial data agree the fundamental units are the geographic units of points, lines, and areas (and volume or surface under special circumstances). Base features developed from these units form an important framework for such operations as delineating the boundaries of higher level geographic units. Base features may include roads, rivers, pipelines, etc. The exact features that organizations categorize as necessary base features varies and depends on the organizations' requirements for and use of a data base.

Given that the basic building blocks used to construct both the higher geographic units and the base features are the points, lines, and areas (polygons), then there exists a unit of measurement, position. The building blocks are considered as being positioned through the use of commonly accepted coordinate systems (often latitude/longitude). Different organizations have different terminology for the building blocks as well as different coordinate schemes. The U.S. Bureau of the Census refers to them as 0-cells, 1-cells, and 2-cells and uses latitude and longitude as the unit. The U.S. Geological Survey refers to nodes, lines, and areas. A graphic of this fundamental model and the common building blocks is depicted in Figure 1.

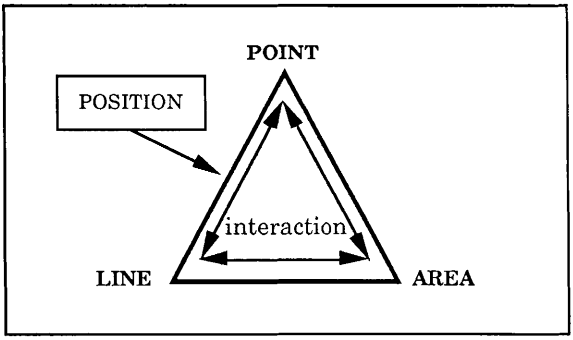


Figure 1. The building blocks of geospatial data.

The Advantages of A Conceptual Model

The fact that common building blocks were accepted did not lead to a single, refined data model nor to a common data structure. Rather, the mechanics of how the data are mentally conceived of in relation to the "real" world and physically organized in terms of structure within a data base varies widely. It also did not lead to a concensus on what constitutes a "correct" classification, or even a "singular," widely accepted set of data definitions for features. An understanding of the building blocks did lead to the additional understanding that there could be different use of the building blocks to create features. For example, roads might be portrayed as lines, perhaps representing road centerlines in one data base, and as polygons in another data base with a width representing say a road's paved surface. Portrayal as road centerlines may be best for the data base user who is concerned with flow or networking data. On the other hand, the data base user interested in siting buildings near highways may need the polygons clearly denoting road widths, right of ways, or some other boundaries to be of the greatest value. The data model provides a common frame of reference for all producers and users of geospatial data. This provides a means by which spatial data can be referenced, even if there is disagreement on terminology assigned to the building blocks.

A Conceptual Model for Cultural and Demographic Data Sets

Whereas for spatial data there exist the fundamental geographic units of points, lines, or areas, a corresponding data unit for cultural and demographic data is elusive at present and difficult to conceptualize. Cultural and demographic data are the data contained in or represented by the geographic framework units. Note, these data and the attempt to clarify them should not be confused with efforts to delineate geographic units which have a cultural aspect as their common denominator.

A number of efforts are underway to delineate culturally-related entities or geographic units. The U.S. military's Tri-Service GIS/Spatial Data Standards (1994 draft) identifies culture as one topic inside the category for delineating geographic units (other topics include landform, geology, soils, hydrography, and climate.) The Tri-Service standard includes identification of such culturally related areas as historical structures, historic maritime sites, prehistoric sites, survey areas, probable and sensitive sites, and native American sites. Although adequate for delineating sites, the standard does not attempt to define, categorize, or classify the cultural and demographic data sets related to these sites.

Just as a fundamental cultural unit has remained vague, a concensus on the building blocks of cultural and demographic data sets has remained equally elusive. Work is currently underway in the U.S. with the purpose of identifying the components of cultural and demographic data. From this effort a conceptual model of the underlying building blocks appears to be emerging.

The identification of the basic cultural and demographic data unit seems to be centered on three questions. Question: Does the data describe a (human) activity (for example an economic activity or a land use activity)? Question: Does the data describe an aspect of humans (for example their health or age)? Question: Does the data describe an aspect of (human) society (for example its political, social, or historical aspects)?

It appears all cultural and demographic data fits into at least one and possibly more of these categories similar to the way in which geospatial data may be categorized as points, lines, and areas. Loosely applying the "geometric" model, one can consider the human as the point, society as the line (linking humans and bounding activities), and activities as taking place over or in relation to an area. And just as there appears to be a common reference to all geospatial features by position, there appears to be a common reference to all cultural and demographic data by count, such as its number or amount. Figure 2 depicts this model graphically.

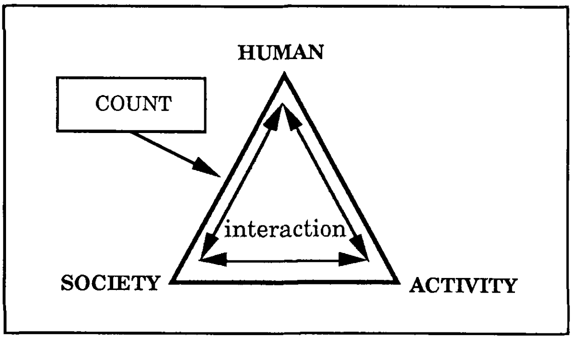


Figure 2. The building blocks of cultural and demographic data.

Though not sufficient for building internationally applicable standards, agreement upon a model such as this can provide a starting point. Once an underlying model is in place, the components for defining the data can be evaluated and various ways of constructing the data proposed and compared. Discussion on the building blocks needs to begin with professional organizations and international associations and conferences providing the forum.

U.S. WORK TOWARD A CULTURAL AND DEMOGRAPHIC METADATA STANDARD

The FGDC's Subcommittee on Cultural and Demographic Data is preparing a standard for describing spatially referenced cultural and demographic data (FGDC, 1994). The Subcommittee is developing this adjunct standard to the geospatial metadata standard because it feels cultural and demographic data are unique from the other more "physically" perceived data and are not adequately represented by using only the geospatial metadata standard. To minimize the time involved in developing metadata, the cultural and demographic data metadata standard is being prepared so that metadata producers and users can easily "crosswalk" between the Digital Geospatial Metadata Standard and the proposed cultural and demographic data metadata standards, as well as the more general Government Information Locator Service (GILS).

The draft cultural and demographic standard identifies a relatively complete description of the contents of a cultural and demographic data set. The principle descriptive parameters in the cultural and demographic data metadata standard are data set identification, themes, geographic framework, temporal framework, source, and data quality. Although all the parameters are needed to obtain a complete picture of the data, the sections on themes, geographic framework, and temporal framework are considered the components most important to cultural and demographic data.

The method for placement of cultural and demographic data sets into themes is the critical operation when applying the standard. Theme prioritization, i.e. nesting, is the most difficult task for the data set producer given the complexity of ideas which can make up a data set. Cultural and demographic data sets often address many major themes and any number of minor themes, all within the same data set. To accommodate a range of combinations of data items and allow the data set producer to prioritize his categories, the Subcommittee developed approximately fifty major themes or categories of cultural and demographic data and approximately two hundred minor themes or categories which can be used hierarchically to further describe the data much in the way attributes are used to describe geospatial features. Although defined in the standards, the major themes are purposefully general and are meant to refer to the general category of the whole data set, even though specific subthemes may be in entirely different categories. Upon determining a major theme, the data set producer may identify as many hierarchies or levels of minor themes as desired depending on complexity and the purpose of the data set.

STANDARD DEFINITION ISSUES

The Geospatial Data Example

Work on spatial data bases has focused on defining the geographic units, the units which provide the base to which additional data may be related. The geographic units may be legal, statistical, natural, or thematic. In the case of legal geographic units, definitions are straightforward and can fairly easily be applied in an international context once the frame of reference is known. For example, if a South African data base is being used by a U.S. researcher the S.A. geographic unit "province" is mentally translated to the geographic unit "state" once the researcher understands a province is the primary legal subdivision of South Africa as a state is the primary legal subdivision of the U.S. The data base user's understanding of the geographic units and their position in the hierarchy in the spatial framework probably easily mirrors that of the data base producer. This argument is valid for the majority of statistical areas.

Although statistical units do not always have boundaries resulting from charters, laws, treaties, or so forth, they are often delineated by governments for data collection and/or tabulation purposes. Their creation for a specific purpose leads to a rather precise definition which can again be translated internationally, though some confusion may arise. If the same South African data base includes enumeration areas, the U.S. researcher might make the mental translation to enumeration district, a low level statistical unit used by the U.S. Bureau of the Census for tabulation. The South African Central Statistics Survey defines an enumeration area as its smallest collection area. In some countries, the basic level of geography for tabulation corresponds to the basic level of geography for collection, consisting of an area one enumerator can cover within a fixed amount of time and containing a limited number of housing units. However, in some countries the basic unit for collection has evolved to being one or a cluster of blocks without regard to number of housing units because the enumeration is conducted electronically. Additionally, in other countries, such as the U.S., both geographic units are used for collection but only the latter for tabulation. The U.S. researcher would have to refer to precise South African definitions to understand that in South Africa the collection unit is also the tabulation unit and probably contains between 200 and 400 housing units[1].

[1] The South African Swuato enumeration area (EA) is one of the exceptions. It is representative of how commonly understood geographic area terms can take on a different meaning to accommodate special situations. Estimates of the Swuato EA population range from one to three million persons and its housing units exceed the planned 200 to 400.

Unfortunately, the simplicity of mental translation becomes more difficult when considering natural or thematic geographic units. Natural or thematic units tend to delineate environmental, physically identifiably similar areas that have boundaries based on the properties or distribution of some variable data. What exactly is a drainage basin? The U.S. researcher may have a definition, but it may not be close enough to another country's definition that the intended use of the data base is not effected. Unlike legal and even statistical geographic units, there are many definitions of a drainage basin and they do not neatly coincide. A drainage basin may be defined as a geographic unit (valley) whose area contributes water to and is drained by a drainage system (one stream and its tributaries). Drainage basins are separated by divides. However, a drainage basin may also be referred to as a watershed; if so, technically the drainage rim is considered to be a part of the watershed (American Geologic Institute, 1976). Inclusion of the drainage rim may, or may not, effect the use of the data base; the total effect depends on two considerations--the extent of the differences in definitions of both the data set producer and data set user and, most importantly, if the data set user is planning to relate additional information to the drainage basin, but use the producer's definition.

Definitions for Cultural and Demographic Data

Definitions, which can be exchanged somewhat successfully about a majority of geographic units, appear to be more complex and elusive in the context of cultural and demographic data. A reason for this is cultural differences must be considered when referencing cultural and demographic data sets, which after all are inherently cultural. Cultural and demographic data sets lack the clarity of their spatial counterparts. Consider as a further component of the South African data base the population count per enumeration area. The U.S. researcher would most likely mentally picture a population count as including a count of the total human population of a given area. The South African data set, however, would not necessarily be clear as to what population of a given area is being reported if the data set were one released prior to 1994.

Definitions may differ uniquely in persons from different cultures due to the cultural bias even in such apparently universal things as time. Just as groups of people or countries' perceptions change over time, individual perceptions also change. A researcher, at age twenty, may include anyone older than fifty in a count of a category of say, "over the hill persons." The same researcher conducting a similar count thirty years later may reevaluate and include anyone older than seventy-five years in the "over-the-hill" count. The resulting data sets and the information which could be gleaned from them would be dissimilar due to changing perceptions, even though the categories stayed the same in title.

There are many examples of varying interpretations of data set items resulting from cultural differences. Cultural misunderstandings are not confined to data analysis; they begin during data collection. Many countries count "households." The U.S. Bureau of the Census considers a household as consisting of a person or a group of persons who make common provision for food and other necessities for living, incorporating a household-housing unit concept (USBC, 1994). In many Moslem communities located in Africa, polygamy is practiced. Multiple wives and their children may live in one compound in a single dwelling, or may live in one compound in multiple closely situated dwellings, or may live in a village in several widely dispersed dwellings. The data collectors may record the counts erroneously based on their cultural interpretation of what constitutes a household. There also may be an error in the population count dependent on whether a husband is counted as spouse to no specific wife, one particular wife, or to several wives. Another example, one effecting data reliability, is age. Generally, persons in some countries know their exact age, perhaps because so many events (beginning school, applying for a driver's license, the right to vote, collecting social security benefits) are based on age. In many other countries people, particularly older people, know their approximate age rather than exact age. The approximation itself may vary, as often the age is approximated around calendars of historical events. The data user may erroneously expects the error associated with age-related data to be as small as the error existent in the data sets more commonly used.

How Standards can Address This Problem

From the FGDC's Subcommittee on Cultural and Demographic Data's experiences certain guidelines for achieving an international set of standards for describing cultural and demographic data are being identified. The first is the need for all those participating in the development of such standards to realize that cultural and demographic data, to a greater extent than other types of data, are susceptible to cross-linguistic, cross-cultural misunderstandings. Because of these cultural differences as well as changes in personal and cultural attitudes which span the temporal dimension, providing a means of referencing a multiple, well-defined, dynamic rather than rigid data definitions becomes extremely important. The task of metadata standards for cultural and demographic data is twofold--they must allow data set producers to easily define their own terms or reference definitions while assuring data set users can easily locate specific definitions.

Many metadata standards allow "free text" entries for terms that are not explicitly defined within the metadata document and allow these terms to be used in domains rather than limiting the data set producer to a closed domain. The freedom "free text" provides to the data set producer is important in accurately describing their data sets. However, with this freedom comes the increased potential for misunderstanding. When "free text" is utilized, either a concise definition of the "free text" or a reference to an easily accessible data dictionary must be assured. Further, data set producers should have the freedom of including within the metadata any additional information they feel to be pertinent to data set use. Standards should not be so rigidly structured that the metadata producers are limited in their ability to provide information they feel is of importance to the data set. This freedom may raise problems in developing a parser for meaningfully accessing the metadata, but the price for overcoming these technical problems are more than compensated for by the value of the resulting increased precision in terminology.

SUMMARY AND RECOMMENDATIONS

This paper did not attempt to build an argument supporting the importance of cultural and demographic data to our daily lives because it was felt to be self evident. Rather, examples demonstrating the need for a standard were presented along with examples of difficulties encountered. A conceptual model of basic units is presented. The model is offered as a starting point only.

If full advantage is to be taken in this age of electronic access to an ever-increasing quantity of cultural and demographic data, a metadata standard is obviously needed. Work must begin now, both within and between countries and cultures. Therefore, the following recommendations are offered:

1. Support research in identifying a robust conceptual model.
2. Undertake cooperative efforts to define and refine a metadata encoding scheme.
3. Make spatially referenced cultural and demographic data sets, their description, availability, and exchange a topic for discussion, presentations, and so forth at both national and international professional meetings and conferences.

It is hoped that the need stated in the paper and the model proposed will encourage development of a truly international metadata standard for spatially referenced cultural and demographic data.

REFERENCES AND ACKNOWKLEDGEMENTS

American Geological Institute, Dictionary of Terms, Revised Edition, 1976, USA.

Broome, Frederick R. and David B. Meixler, 1990, "The TIGER Data Base Structure," Cartography and Geographic Information Systems, 17(1), ACSM, Bethesda, MD-USA, pp.39-48.

Department of Commerce, 1992, 1990 Census of Population and Housing, Summary Population and Housing Characteristics, United States, Department of Commerce, U.S. Bureau of the Census, USA.

Department of Commerce, 1992, Spatial Data Transfer Standard (SDTS) (Federal Information Processing Standard 173), Department of Commerce, National Institute of Standards and Technology, USA.

Federal Geographic Data Committee, 1994. Content Standards for Digital Geospatial Metadata (June 8), Federal Geographic Data Committee, Washington, D.C, USA.

Federal Geographic Data Committee, Subcommittee on Cultural and Demographic Data, 1994, Draft Cultural and Demographic Data Metadata Standards (unpublished), USA.

Lynch, John, September 1994, conversations with Mr. Lynch recorded in meeting notes (unpublished), Central Statistics Service, S.A.

Mark, David M., 1993, "Toward a Theoretical Framework for Geographic Entity Types," prepared for COSIT'93, USA.

(no author), 1994, The Government Information Locator Service (GILS): Report to the Information Infrastructure Task Force (May 2, 1994), USA.

Standards Committee of the Co-ordinating Committee for the National Land Information System, November, 1990, National Standard for the Exchange of Digital Geo-Referenced Information, Version 2, CCNLIS Publication No. 1, S.A.

Tri-Service CADD/GIS Technology Center, 19 November 1993, Tri-Service GIS/Spatial Data Standards (draft for comment), Release 1.2., USA.

The author wishes to thank Mr. Fred Broome and Mr. Leo Dougherty of the U.S. Bureau of the Census for their examples of cultural differences.

TRANSACTION MANAGEMENT IN DISTRIBUTED GEOGRAPHICAL DATABASES.

Britt-Inger Nilsson
Dept.of Environmental Planning and Design
Luleå University of Technology
S-971 87 Luleå, Sweden.
bini@sb.luth.se

ABSTRACT

The purpose of this paper is to study transaction management in a standard distributed data processing environment and to analyze methods for transaction management between independent spatial databases.

In distributed systems concurrency control is even more complex than in centralized systems. A transaction may access data stored at more than one site. Consistency of the databases involved must be guaranteed.

Standard database management systems of today cannot manage spatial data in an optimal manner. For that reason special spatial databases have been developed. Users of geographical information systems (GIS) are demanding the same sort of flexibility as in nongeographical database management systems.

The purpose of this paper is to compare different methods for concurrency control especially considering spatial data. The object is to elucidate restrictions and advantages of a distribution of geodatabanken, which is the national geographical database of the Swedish Land Survey.

In a GIS the user is working with long transactions. The disadvantages of different methods for concurrency control will be even more obvious. Which method to choose for concurrency control should be based on predictions not only about the current use of data but also the future use.

This study is primarly made with respect to geodatabanken. Geodatabanken is currently using the so called optimistic method for concurrency control. Also, a heterogeneous database environment is currently being considered within the organisation.

Preliminary studies of geodatabanken conclude that the optimistic methods for concurrency control seem to be enough at present. In consideration of probable change of the access patterns and the transaction types, a change of transaction management is recommended.

INTRODUCTION

The majority of present-day database systems are relational, and they also tend to be multi-user. These systems are developed to manage administrative data and therefore presume short transactions. Relational systems have well developed methods to guarantee data integrity.

Management of spatial information has requirements which traditional database managers cannot meet today. One problem is how to manage long transactions. Most GIS-vendors offer special solutions developed for spatial information.

Concurrently with the increase of information and the use of computerized systems the demand for service, access, flexbility, security, and local control will also increase. Distributed databases have been developed for nongeographical systems. Distributed systems make it possible to store data at different sites. A user at any site can access data stored at any site, without knowing where any given piece of data actually is stored. This in turn increases the requirements for security controls to maintain consistency between sites.

Today the techniques on how to handle distributed geographical databases are quite undeveloped. However, the problem has been stated before, see for example (Webster, 1988), (Sun, 1989), (Frank, 1992), (Pollock & McLaughlin, 1991), (Edmondson, 1989) and (Devine, Rafkind & McManus, 1991). The interest in distributed geographical databases will increase as the use of GIS technology spreads.

The Swedish Land Survey is currently in the process of establishing a continuous database of Sweden. Data will be stored in Geodatabanken which is the national geographical database developed by the Swedish Land Survey. These nation-wide databases will be very large. It is desirable that parts of these databases be stored at local sites. If these local databases consist of a subset of the central nation-wide database, all databases must have identical information in overlapping areas.

The purpose of this paper is to describe the methods for transaction management and to analyze methods on how to distribute geographical databases, especially geodatabanken.

TRANSACTION MANAGEMENT

The normal assumption in a traditional DBMS is that a transaction is short. In a banking system, for example, a transaction can be to transfer an amount of money from one account into another. Since this process is short, a few seconds at most, other users will not become aware of it.

In a GIS, a transaction can be very long, possibly hours or months. Here a transaction might be to digitize a map sheet or to redesign parts of an old utility network. Other users should not be prevented from working during this time.

Causes of failure

To create a reliable system which is able to recover automatically, all possible failures need to be identified. These are classified under four headings (Bell & Grimson, 1992).

- Transaction-local failure.
  - Transaction-induced abort. The programmer has written code to trap a particular error.
  - Unforseen transaction failure arising from bugs in the program.
  - System-induced abort. For example when the transaction manager explicitly aborts a transaction to break a deadlock or because it conflicts with another transaction.

- Site failures.
  - Occurs as a result of failure of the local CPU or a power supply failure resulting in a system crash.

- Media failures.
  - A failure which results in some portion of the stable database being corrupted. Disk head crash.

- Network failures.
  - Communication failure, for example, resulting in the network being partitioned into two or more sub-networks.

Concurrency control

While most DBMSs are multi-user systems there are tools needed to ensure that concurrent transactions do not interfere with each other. The purpose of consistency control is to guarantee consistency of the database.

A lot of problems can occur if concurrent transactions can interfere with one another, for example:

- Lost updates. A transaction that seems to be successfully completed can be overridden by another user.

- Inconsistent retrieval. A transaction can obtain inaccurate results if it is allowed to read partial results of incomplete transactions which are simultaneously updating the database.

There are three basic techniques in traditional DBMSs for concurrency control. Locking methods and timestamp methods are conservative methods. They check if a conflict occurs every time a read or write operation is executed. Optimistic methods are based on the premise that conflicts are rare and allow transactions to proceed. When a transaction wishes to commit, the system checks for conflicts. Version management is a special form of optimistic method developed to manage long transactions in spatial databases.

Locking methods

Locking methods are the most widely used approach to handle concurrency control. A transaction puts a read or write lock on a data item before a read or write operation is executed. Since read operations cannot conflict, it is permissable for more than one transaction to hold read locks on the same data item at the same time. A write lock gives a transaction exclusive access to the data item, no other transaction can read or update that data item as long as the write lock is hold. The write operations will not be visible to other transactions until the actual transaction releases its write locks.

Some systems allow upgrading and downgrading of locks. This means that if a transaction holds a read lock on a data item, this lock can be upgraded to a write lock on condition that the transaction is the only one holding a read lock on that data item. Upgrading of locks allows a transaction to examine the data first and then decide whether or or not it wishes to update it. Where upgrading is not supported, a transaction must hold write locks on all data items which it might want to update some time during the execution of the transaction.

The most common locking protocol is known as two-phase locking (2PL). The method got its name becuse it operates in two distinct phases. It consist of a growing phase during which the transaction acquires locks and a shrinking phase during which it releases those locks.

The rules for transactions which obey 2PL are:

- Transactions must be well-formed. A transaction must acquire a lock on a data item before operating, and all locks held by a transaction must be released when the transaction is finished.

- Rules for locking must be consistent. No conflicts like write-write locks or read-write locks are allowed to exist.

- Once the transaction has released a lock, no new locks are acquired.

- All write locks are released together when a transaction commits.

Deadlocks can occur when a transaction waits for a lock to be released by another transaction which in turn is waiting for a lock to be released by the first transaction. In a distributed system these deadlocks are even more complex to detect since a number of different sites can be involved in transactions.

Timestamp methods

No locks are involved in timestamp methods, and therefore no deadlocks can occur. Transactions involved in conflicts are rolled back and restarted. The goal of timestamping methods is to order transactions in such a way that older transactions (transactions with smaller timestamps) get priority in case a conflict occurs.

If a transaction tries to read or write a data item, the read or write will only be allowed if the last update on that data item was made by an older transaction. Otherwise the requesting transaction is restarted given a new timestamp.

Timestamp methods produse schedules from the timestamps of transactions which are committed successfully. Timestamps are used to order transactions with respect to each other. Each transaction is assigned a unique timestamp when it is started. No two transactions can have the same timestamp. In a centralized system they can be generated from the system clock or alternatively from a simple counter. When a transaction is started it is assigned the next value from the counter, which is then increased. To avoid very large timestamp values, the counter can periodically be reset to zero.

In a distributed system there is no central system clock or centalized counter. Each node has it own clock, and there is no guarantee that these clocks are syncronized with each other. Two (or more) transactions can start at the same time at different sites in the network. A simple approach is to define global time as the concatenation of the local site clock time with the site identifier.

A transaction must guarantee atomicity. It can easily be done by locking methods. With timestamp methods we do not have any locks, thus it is possible for other transactions to see partial updates. This can be prevented by not writing to the database until the actual transaction is committed. The updates are written to buffers, which in turn are written to the database when the transaction commits. This approach has the advantage that when a transaction is aborted and restarted no changes need to be made in the database.

Optimistic methods

Optimistic methods are based on the premise that conflicts are rare and that the best approach is to allow transactions to continue. Timestamps are only assigned to transactions. To ensure atomicity of transactions, all updates are made to local copies of the data. When a transaction commits, the system checks for conflicts. During detection of conflicts the system checks if there are transactions which overlap each other in time and, if so, if they affect the same data item. In the event of conflict, the local copy is discarded and the transaction is restarted. If no conflicts have been detected, the local copy is propagated to the database.

In environments where read-only transactions dominate, optimistic methods are attractive because they allow the majoriy of transactions to proceed without hindrance.

Version management

The different methods for concurrency control as mentioned above have been developed for short transactions. When using GIS the concurrency problem are even more complex because of long transactions. Version management is a method managing many users involved in long transactions. This method is essentially an optimistic approach on the assumption that, in the majority of situations, there probably will not be any conflicts between versions.

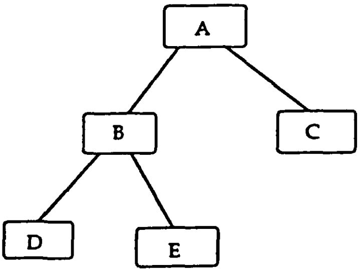
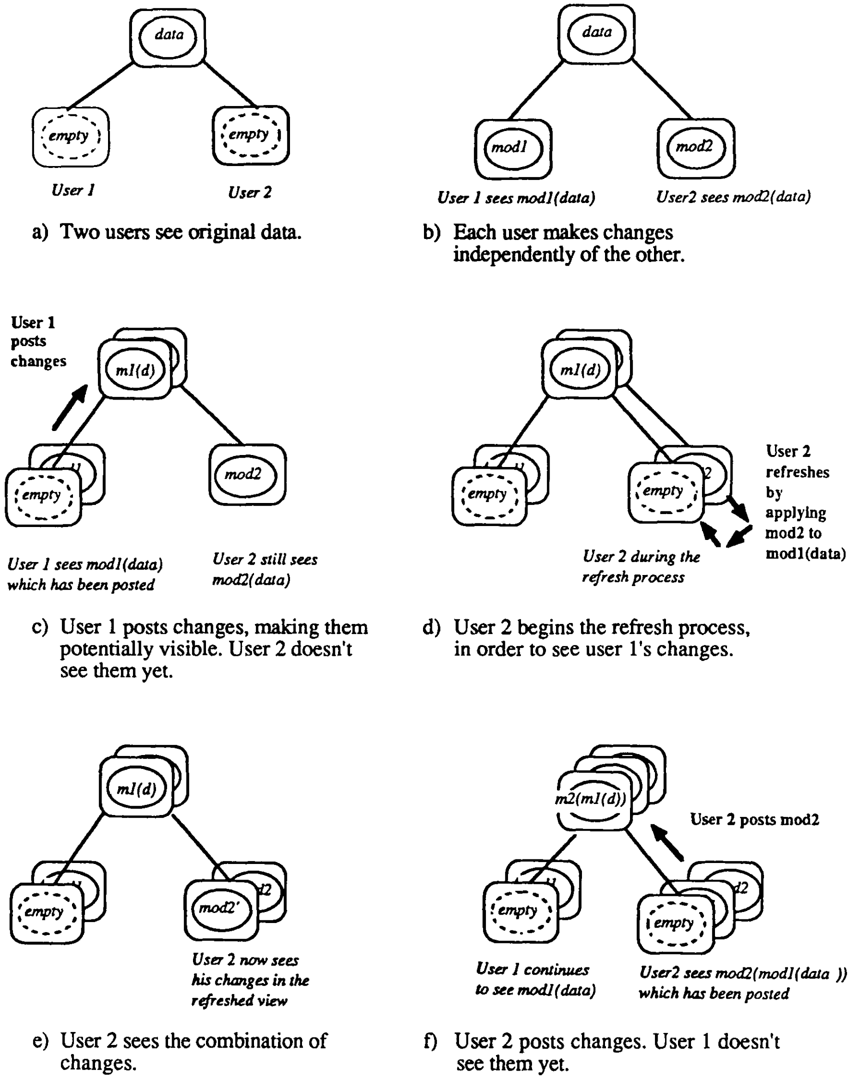


Figure 1. Example of a Version Tree.

Versions are thought of as having a hierarchical arrangement so that later versions are derived from earlier ones. At the top is the master database. At the leaves of the tree, each user has an alternative in which to work without conflicting with others. Each version has a unique route back to the root of the tree. This route can be called an access path since it determines precisely which part of the database each user can see and also what can be modified.

Changes are made visible (posted) one level at a time. Provided that the version at the higher level is the one from which the changed "leaf" version was derived, the changes will be made. But if the version at the higher level has since been updated, the version about to post must first be refreshed with superior updates. The other users will not notice that their versions suddenly change. They will have to make a refresh operation in order to see changes made at the higher level.


data
empty
empty
User 1
User 2
data
mod1
mod2
User 1 sees mod1(data)
User2 sees mod2(data)
User 1 posts changes
m1(d)
empty
mod2
User 1 sees mod1(data) which has been posted
User 2 still sees mod2(data)
m1(d)
empty
empty
User 2 refreshes by applying mod2 to mod1(data)
User 2 during the refresh process
m1(d)
empty
mod2'
User 2 now sees his changes in the refreshed view
m2(m1(d))
User 2 posts mod2
empty
empty
User 1 continues to see mod1(data)
User2 sees mod2(mod1(data)) which has been posted

a) Two users see original data.

b) Each user makes changes independently of the other.

c) User 1 posts changes, making them potentially visible. User 2 doesn't see them yet.

d) User 2 begins the refresh process, in order to see user 1's changes.

e) User 2 sees the combination of changes.

f) User 2 posts changes. User 1 doesn't see them yet.

Figure 2. Illustrating the modify/post/refresh cycle for a pair of users. (Easterfield et al.)

Comparison

The different methods for concurrency control deal with conflicts differently which gives each method its own advantages and disadvantages.

Table 1. Comparison between different methods for concurrency control.

	Locks	Timestamps	Optimistic methods	Version management
How are conflicts dealt with?	The transaction waits.	The transaction is rolled back and restarted.	The transaction is rolled back and restarted.	Every single object is validated manually.
When are conflicts detected?	When a lock is applied for.	When reading/ writing a data item.	When a transaction wishes to commit.	When posted to a higher level.
Advantages	Waiting is not as expensive as restarting.	No deadlocks. No waiting.	No deadlocks. No waiting. Attractive where read-only transactions.	No deadlocks. No waiting. Possible to keep versions.
Disadvantages	Deadlocks. Enormous message overhead . Waiting is unacceptable when long-transactions	Rollback and restart is expensive. In distributed systems analyses will be very complex.	Rollback and restart is expensive. In distributed systems analyses will be very complex.	Rollback and restart is expensive. Inconsistency can occur as a result of manual validation.

GEODATABANKEN

Geodatabanken is a national spatial database developed by Swedish Land Survey. Its purpose is to store spatial data independent of map types or scales. Geodatabanken can briefly be described as:

- Feature oriented, which means that feature type is stored in conjunction with coordinates.
- Seamless, no map sheet partitioning.
- Multi-user system.
- Check-out/check-in routines with a possibility to append data.
- History.
- Optimistic methods for concurrency control.
- Network data structure.
- Inhouse developed security system.

Geodatabanken uses a time-dimension which gives information about the history of an object. It contains information about when the object was created, when it was edited or deleted, and who did the operations. The time is necessary for appending data, and time is also needed for concurrency controls.

Local databases

The Swedish Land Survey is currently in the process of establishing a continuous database of Sweden. Data for the northern region is being collected at the regional office in Luleå. Since the nation-wide database is to be stored in geodatabanken, the office in Luleå is regularly checking data in and out of geodatabanken.

A nation-wide database will be very large. The office in Luleå requires a local database containing a subset of the nation-wide database, more exactly, all data from the local region. A problem that will occur is that all databases must have identical information in overlapping areas.

In addition to the requirements of concurrency controls and consistency controls additional demands have to be fulfilled such as:

- an object has to have an unique identity,
- all objects have to be able to store time.

It has been discussed that the regional office should use a commercial GIS database management system, for instance ArcStorm from ESRI or GeoData Manager from Intergraph, instead of the inhouse developed system Geodatabanken. This would lead to a distributed heterogeneous environment. A requirement would be to maintain consistency. It would be necessary to preserve all information in both systems. It would also be necessary for their concurrency control methods, at least to some extent, to be able to communicate and co-operate. All information needed in a system to maintain consistency has to be transferable and differences in dataformat has to be manageable.

DISCUSSION

The course of checking-out - checking-in represents a transaction in Geodatabanken. Several users can simultaneously work in the same area. Geodatabanken uses optimistic methods for concurrency control when trying to check-in data. If any conflict occurs, the user has to add data from geodatabanken through the earlier check-out operation and then validate the data manually.

Today, work in Geodatabanken is concentrated in building new databases. The nation-wide data collection is divided into regions and thereby conflicts are rare.

Due to the optimistic methods, data is always accessible in geodatabanken. This is an advantage because almost all of the transactions are long, and in these circumstances it is unreasonable to expect other users to be locked out. The disadvantage with this method is the enormous overhead involved in restarting a transaction in case of a conflict. In the worst situation, important decisions can be made on the basis of inconsistent data.

Preliminary studies of work done in theSwedish Land Survey shows that the main part of the work uses local data. If there is occasionally interest to use other than local data, it is almost always dealing with read operations only. These facts lead to the conclusion that the optimistic methods for concurrency control seem to be enough at present.

The nation-wide databases will be very large. It is desirable that parts of these databases will be stored at local sites. By spreading these databases to the regions, better access to regional data is possible,and it might also give better quality data.

Some of the advantages of distributing local copies of Geodatabanken are:

- Local autonomy.
- Response time will be reduced.
- Greater reliability. All work in the region can continue even if a failure at the central database occurs, and vice versa, in case of failure of a local database, all data are available from the central database.
- Communication costs will be reduced.

Producing a true distributed spatial database-management system is difficult. Besides all the complicated problems which occur by distributing traditional databases, the developer, among other things, also has to deal with long transactions. Even in non-spatial database environments there is a lack of experience considering distributed systems.

To be able to make conclusions on whether the optimistic methods are enough for concurrency control in the future, the future access pattern has to be predictable. Due to the limitations of this method, the transaction management might have to be changed if the access pattern changes. Questions to pay attention to:

What will the future use of spatial data look like?
Who is responsible for updating these data?
Who will be interested in using these data?

REFERENCES

AutoKa-PC-projektet, 1993.Datorprogram, Beskrivning, Flyttfil för Geodatabanken och AutoKa-PC. Dnr 162-93-609. Lantmäteriverket, Gävle.

Baker T. and Broadhead J., 1992. Integrated Access to Land-related Data from Heterogeneous Sources. URISA Proceedings. Papers from the Annual Conference, Washington, DC, USA. Vol 3, pp 110-121

Batty P., 1992. Exploiting Relational Database Technology in a GIS. Computer & Geosciences, Vol. 18, pp 453-462.

Bell D. and Grimson J., 1992. Distributed Database Systems. Addison-Wesley Publishing Company.

Brown R. and Mack J., 1989. From the Closed Shop to the Organizational Mainstream: Organizational Impacts of a Distributed GIS. GIS/LIS '89 Proceedings. Annual Conference. Vol 1 pp 152-160.

Ceri S. and Pelagatti G., 1984. Distributed Databases - Principles and Systems. McGraw-Hill Inc. USA.

Date C.J., 1986. An Introduction to Database Systems. Vol 1, Fourth Edition. Addison-Wesley Publishing Company.

Degerstedt K., 1993. Synkronisering av geodatabanker. Utkast nr 2. Funk spec. Lantmäteriverket, Gävle.

Degerstedt K. and Ottoson P., 1993. En introduktion till Geodatabanken. Lantmäteriverket, Gävle.

Degerstedt K. and Ottoson P., 1993. AutoKa-Geodatabanken, Handbok. Lantmäteriverket, Gävle.

Devine H.A., Rafkind C.D. and McManus J., 1991. A New Approach to GIS Implementation: Colonial National Historic Park. Technical Papers - 1991 ACSM-ASPRS Annual Convention, Baltimore, USA. Vol 4, pp 51-60.

Easterfield M., Newell R.G. and Theriault D.G.. Version Management in GIS - Applications and Techniques. Technical Paper 8. Smallworld Systems Ltd., Cambridge, England.

Edmondson P.H., 1989. Transition from a Closed Shop GIS to a True Distributed GIS. GIS/LIS '89 Proceedings. Annual Conference. Vol 1, pp 161-170

Frank A.U., 1992. Acquiring a Digital Base Map: A Theorethical Investigation into a Form of Sharing Data. URISA Journal, Vol 4, pp 10-23.

Lidén J., 1992. Versionshantering av geografiska databaser. Smallworld Svenska AB. Stockholm.

Newell R.G. and Easterfield M., 1990. Version Management - the problem of the long transaction. Technical Paper 4. Smallworld Systems Ltd., Cambridge, England.

Pollock R.J. and McLaughlin J.D., 1991. Data-Base Management System Technology and Geographic Information Systems. Journal of Surveying Engineering, Vol 117, No 1, pp 9-26.

Salen F., 1992. Kartbibliotekssystem - för hantering av geografiska databaser. Högskolan i Luleå.

Sun Chin-Hong, 1989. Development of the National Geographic Information in Taiwan. GIS/LIS '89 Proceedings. Annual Conference, Vol 1, pp 270-275.

Webster C., 1988. Disaggregated GIS architecture - Lessons from recent developments in multi-site database management systems. International Journal of Geographical Information Systems. Vol 2, no 1, pp 67-79.

Özsu T.M and Valduriez P., 1991. Principles of Distributed Database Systems. Prentice-Hall Inc, Englewood Cliffs, New Jersey.

SEMANTIC NET OF UNIVERSAL ELEMENTARY GIS FUNCTIONS

Jochen Albrecht
ISPA, University of Vechta
Postfach 1553, 49364 Vechta, Germany
email jalbrecht@ispa.Uni-Osnabrueck.de

ABSTRACT

Current use of GIS is data-driven, prohibiting the application of model-based procedures The analysis of user surveys and functional taxonomies leads to a classification of data model-independent elementary GIS functions whose relations were examined employing a semantic net An important finding is the notion that there is no universal set of application-independent elementary GIS functions Each application has its own view onto the semantics, expressed as a web of relations, that is determined by domain context and can be made explicit in simple knowledge bases The Virtual GIS shell presented here, supports the knowledge building as well and offers an easy-to-use task-oriented Graphical User Interface (GUI) that allows the domain scientists to concentrate on their tasks rather than on their data.

INTRODUCTION

Currently available GIS technology, while impressive, is isolated from both the actual problem context and the users' conceptualization of the problem It provides 'powerful toolboxes in which many of the tools. are as strange to the user as a robot-driven assembly plant for cars is to the average home handyman' (Burrough, 1992, page 9) The Virtual GIS (VGIS) project currently pursued at the Institute for Spatial Analysis and Planning in Areas of Intense Agriculture (ISPA) is geared to overcome this problem Its goals — the theoretical foundation of which is presented in this paper — are

(1) to develop a taxonomy of fundamental GIS operations with conditions for their application
(2) to help the user to develop a plan for an analysis using available data and operational capabilities of the GIS, as well as mapping a general operation description into a specific set of commands or actions appropriate for the current GIS
(3) to define a model for relating and retrieving information based on the users' description of context This model consists of
    - data or objects broad categories that generalize both the information the user is trying to retrieve and the information already available
    - relations general categories that describe how objects are connected or related, such as sequence, instance, or composition
    - context schemes patterns of knowledge categories that are composed of particular objects and relations

    The model representation will be based on semantic nets or frames that have been employed in artificial intelligence work.
(4) Semantically-assisted information retrieval as described in previous works by Bennett and Armstrong (1993) and Walker *et al.* (1992) where a user's description of a desired (geographic) dataset is matched with metadata stored with that dataset

The VGIS project emphasizes a user-oriented visualization of tasks instead of the commonly used (technically oriented) functions (Albrecht, 1994b) These tasks can be dissected into universal, elementary GIS functions that are independent of any data structures and thereby of any underlying GIS No matter whether data is collected, manipulated or analyzed, all we do with data can be reduced to a small number of functional groups (see Table 1) that are subordinate to goals (see Figure 1). The author examined a suite of taxonomies as they can be found in the relevant literature of the past few years (Burrough, 1992, Goodchild, 1992; de Man, 1988, Rhind and Green, 1988, Unwin, 1990). To be included into the resulting list, each function needs to be system (or vendor) -independent.

In the following, *task* shall be used for all combined actions that requires some (human or machine-based) knowledge about semantic and spatial relations such as 'map updating', 'routing' or 'siting' whereas the term *function* shall be used for singular actions that can be performed automatically, i.e. 'multivariate analysis' While tasks, as used here, are always application-oriented, *functional groups* are the attempt to aggregate functions and tasks on an abstract technical level Following Monckton (1994) one can differentiate between overall goals (i e inventory, reference, prediction), the tasks used to accomplish them and inferior functions which are called while performing a task

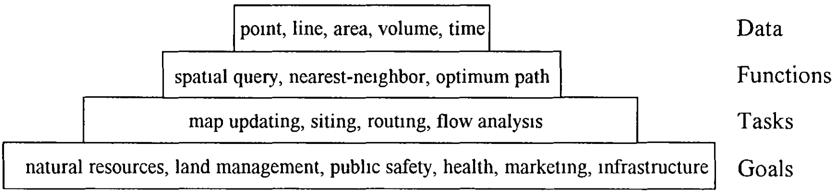

point, line, area, volume, time
Data
spatial query, nearest-neighbor, optimum path
Functions
map updating, siting, routing, flow analysis
Tasks
natural resources, land management, public safety, health, marketing, infrastructure
Goals

Figure 1 The information pyramid consisting of data, functions, tasks and goals *(after Huxhold, 1989, adapted)*

The above mentioned examination of functional taxonomies is condensed in the tables at the end of this paper. Each of the 144 functions screened is related to its neighbors by answering the following two questions· (i) how does a function fit into a thematic context, i e is a operation similar to another one, and (ii) how does a function fit into the work flow, i e. what needs to be done before that function can be called and what other function does it lay ground for? The result is a very complex net of relations that is comprehensible only if transformed to a net of more general *tasks* (Figure 2)

One interesting insight gained here though, is that different users or applications call for different connections (the edges of this semantic net). Figure 3 gives one possible view onto this net of universal GIS functions The most surprising result, however, was that there doesn't seem to be a set of universal, elementary GIS functions that is application-independent

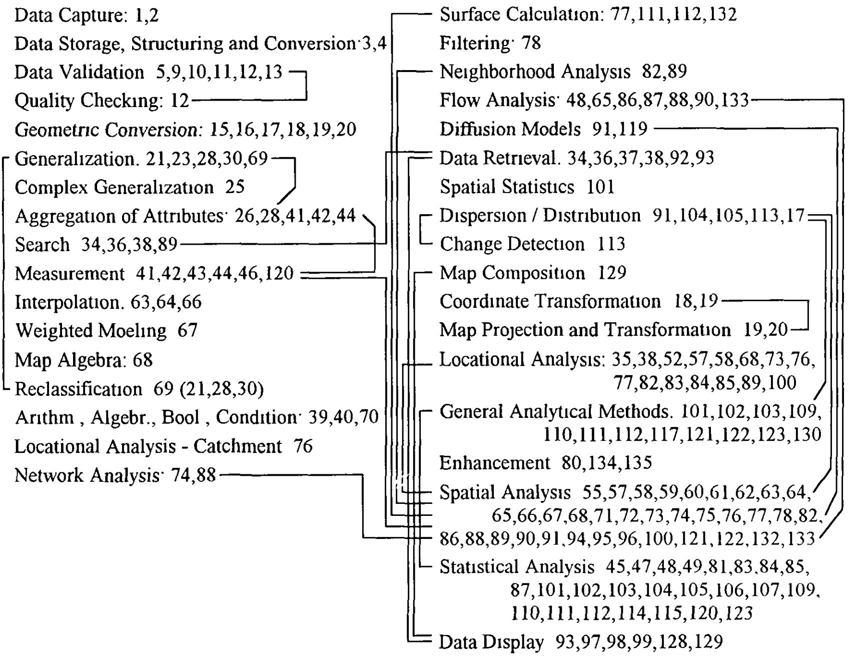

Data Capture: 1,2
Data Storage, Structuring and Conversion 3,4
Data Validation 5,9,10,11,12,13
Quality Checking: 12
Geometric Conversion: 15,16,17,18,19,20
Generalization. 21,23,28,30,69
Complex Generalization 25
Aggregation of Attributes 26,28,41,42,44
Search 34,36,38,89
Measurement 41,42,43,44,46,120
Interpolation. 63,64,66
Weighted Moeling 67
Map Algebra: 68
Reclassification 69 (21,28,30)
Arithm , Algebr., Bool , Condition 39,40,70
Locational Analysis - Catchment 76
Network Analysis 74,88
Surface Calculation: 77,111,112,132
Filtering 78
Neighborhood Analysis 82,89
Flow Analysis 48,65,86,87,88,90,133
Diffusion Models 91,119
Data Retrieval. 34,36,37,38,92,93
Spatial Statistics 101
Dispersion / Distribution 91,104,105,113,17
Change Detection 113
Map Composition 129
Coordinate Transformation 18,19
Map Projection and Transformation 19,20
Locational Analysis: 35,38,52,57,58,68,73,76, 77,82,83,84,85,89,100
General Analytical Methods. 101,102,103,109, 110,111,112,117,121,122,123,130
Enhancement 80,134,135
Spatial Analysis 55,57,58,59,60,61,62,63,64, 65,66,67,68,71,72,73,74,75,76,77,78,82, 86,88,89,90,91,94,95,96,100,121,122,132,133
Statistical Analysis 45,47,48,49,81,83,84,85, 87,101,102,103,104,105,106,107,109, 110,111,112,114,115,120,123
Data Display 93,97,98,99,128,129

Figure 2. Semantic net of GIS tasks. Displayed are only the relations given in the cited references; obviously there are more such as *Map Algebra* and *Spatial Analysis*

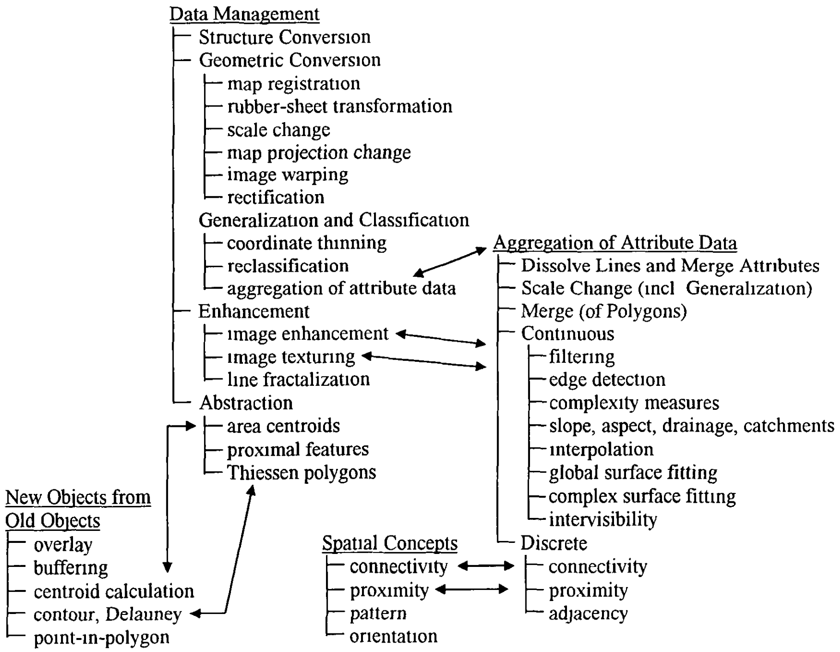


Figure 3. One of many possible views onto the semantic net of GIS functions

GIS Tasks

In addition to this literature-based work the author conducted informal surveys conveying that practitioners of GIS tend to categorize their tasks into the following 15 functional groups:

- Visualize / Show
- Encode
- Find
- Monitor
- Create
- Combine / Relate
- Allocate
- Determine
- Aggregate / Summarize
- Compare
- (partial) Select
- Substitute
- Derive
- Correct
- Evaluate

Table 1 Functional groups of GIS tasks.

Special emphasis was given to the spatial, logical, procedural, and cognitive relations among each of these functionalities. The results compare favorably with the findings in cognitive science where image schemata are used to describe spatial categorizations Of all the schemata summarized by Mark (1989) only '*direction*' and '*center-periphery*' were missing These seem to be more of scientific interest and the findings of this survey suggest that spatially-aware professionals outside geography do not use these relations

Task-Oriented Processing Templates

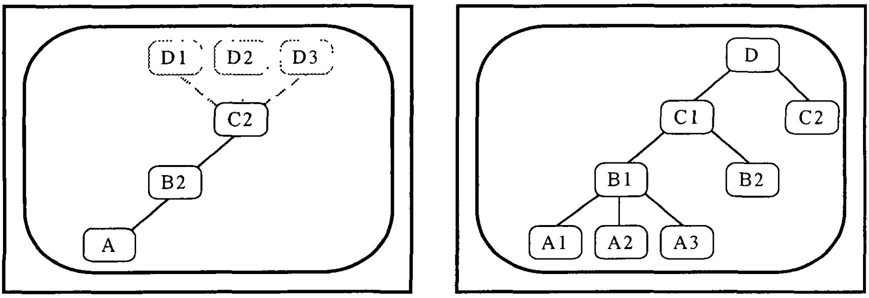
In general there are two possible ways to ascertain the independence from the multitude of data models. The OGIS project (Gardels, 1994) and the 'Virtual Data Set' (Stephan *et al.*, 1993) approach suggest that data objects rather than operations are the primary focus of attention This conforms with the view held by many users of GIS who (driven by the demands of the system they use) tend to think in terms of building data layers or making maps rather than in terms of applying particular GIS operations to manipulate data Especially analytic functionality is distinctively underrepresented in the listing given above This is reflected (or determined?) by the offerings on the vendor's side. less than 10% of the functions listed in ESRI's ARC command reference are analysis-related (ESRI, 1991) But there is also evidence that experienced users (those that have used a couple of different systems and learned to think in tasks rather than in the manipulation of data layers) as well as those who have never encountered a GIS do not think about these layers as isolated, individual entities

A table of elementary methods such as those from the previous section can be regarded as a periodic chart of universal GIS functionalities which allows the building of typical applications much like molecules are made up of atoms (Albrecht, 1994b) Their difference to the generally available GIS functions is in their top-down approach to user needs (as they are derived from general tasks instead of technical conceptions of the manufacturer) and their independence from the underlying system

Each task is associated with a task description file that can recursively call another task This file lists the required input parameters, a pointer to a visualizing icon, meta information and, most important of all, a macro consisting of elementary GIS functions and possibly other tasks. Similarly, there is a file for each class or data type (e g DEM, vector file, point source) describing the methods that are used to create an object instance These files can be seen as a crude form of external knowledge base

The results of both the literature review and the survey were analyzed with a methodology borrowed from speech analysis to construct semantic nets (Birnthaler *et al.*, 1994) For the purposes of the VGIS project, the edges of the semantic net are weighted by the relative importance with which they contribute to a typical GIS procedure These weights are encoded in the task description file Since there are no application-independent paths through the semantic net, each application will require a task tree such as the ones depicted in Figure 4

Almost any situation that requires the presentation of a series of processing steps, especially in its planning stage is best visualized by flow charts. VGIS utilizes flow charts as a graphical means to guide the user through the steps necessary to accomplish a given task By selecting a task from the main menu the first questions for input parameters are triggered If the input data exists in the correct format then the macro of elementary GIS function, stored in the base file, is executed Otherwise the system tries to generate the necessary data based on the knowledge stored in the base files. This might require further input by the user. Data and all operations on them are displayed by icons and connecting edges In the ideal case that all necessary data already exists, the user will get to see the bottom leaf of the task tree only. Otherwise the tree will unfold as in Figure 4a



a) unfolding task tree from user perspective

b) bottom-up creation of a task tree from developer perspective

Figure 4 Task trees

VGIS addresses two classes of users One is the spatially-aware professional who does not want to have to know about GIS internals This kind of user is interested in accomplishing a task and VGIS supports this approach as depicted in Figure 4a The other kind of users that VGIS tries to support are conceptual models Scientists who want to "play" with a variety of models. This latter group utilizes elementary GIS functions to build new task trees. They need to actively rearrange the connections of the semantic net, testing the results and thereby creating new knowledge bases that then can be used by members of the first group Figure 4b describes this bottom-up approach

The base files that store the knowledge about processing procedures can be thought of as preprogrammed model templates (Kirby *et al.*, 1990) that are independent of any data Previous runs are stored in metadata files and can be triggered at any time again, either with links to the old data files or crunching new files after the metadata file has been edited accordingly Lineage information and other metadata are of special importance The methodoloy used for VGIS follows the ideas presented in (Lanter and Veregin, 1992, Lanter, 1994) This usage of metadata information applies to both user classes Model builders may include conditional operators as they are used in formal programming languages The effect is a prototyping and development platform similar to the famous STELLA® (HPS, 1994) but working with real GIS data

CONCLUSION

What is the practical value of yet another taxonomy? The trivial answer to this is that the VGIS project is based upon these universal elementary GIS functions and would have to fail without them. However, there are a number of more universal applications. At larger sites that used numerous GIS in parallel for different applications, a user interface such as VGIS could be used to pick the highlights of whatever functionality a particular system is especially good at One might have an especially intuitive digitizing module while another one shows its strengths in spatial statistics or produces exceptionally good cartographic output. Within a single project it is not advisable to switch from one system to another VGIS enables the GIS manager to build a user interface made of the elementary GIS functions which then invisibly call the functions of the underlying system (see Albrecht, 1994a and Ehlers *et al.*, 1994 for a detailed description of the technical issues).

An additional advantage of such a systemization of GIS functions is the opportunity to use them for software benchmarks Up to now all attempts to objectively compare geographic information systems had to fail because a comparison beyond the boundaries of certain data models is like comparing apples with oranges Since VGIS is based on task-oriented elementary GIS functions that represent the desired functionality without being restrained by the usually historically evolved oddities of a particular vendor system, it is possible to use these as evaluation criteria. New built applications are then the basis for studies as to how much a given system diverges from the desired functionality and thereby how difficult it is to handle this system. VGIS strengthens the toolbox character of GIS by·

- identifying elementary GIS tools (functions),
- providing users a framework (semantic nets) to integrate functions and present different views on task-oriented GIS application design,
- encouraging users to switch from currently dominant data-centered perspectives to procedure-based (task-oriented) GIS project design

This then allows to comply with the demand by cognitive scientist Dan Norman (1991) who wrote 'Good tools do not just extend or amplify existing skills, they change the nature of the task itself, and with it the understanding of what it is we are doing.'

ACKNOWLEDGMENTS

This research is supported under Grant # IIC5-Eh-85/3-1, "Virtual GIS for interaction modeling as a basis for an integrative image analysis in remote sensing" from the German Science Foundation May Yuan contributed valuable hints and helped in summarizing the results of the VGIS project

REFERENCES

Albrecht, J., 1994a. Semantisches Netz Elementarer GIS-Funktionen. In Dollinger, F. and J Strobl (Eds.). *Angewandte Geographische Informationstechnologie VI* Salzburger Geographische Materialien 21·9-19. Salzburg, Austria

Albrecht, J , 1994b. Universal, Elementary GIS Tasks - beyond low-level commands. In Proceedings Sixth International Symposium on Spatial Data Handling, pp. 209-222, Edinburgh.

Bennett, D and M Armstrong, 1993 A Modelbase Management System for Geographic Analysis In *Proceedings GIS/LIS'93*, pp 19-28 ASPRS/ACSM, Falls Church, VA

Birnthaler, T , Kummert, F, Prechtel, R, Sagerer, G, and S Schroder, 1994. *Erlanger Semantisches Netzwerksystem (ERNEST)* Manual version 2 1 Electronically published under ftp://ftp.Uni-Bielefeld.de/pub/papers/techfak/ai/ERNEST/ErnestManual ps

Burrough, P , 1992 Development of Intelligent Geographic Information Systems In *International Journal of Geographical Information Systems*, VI,1 1-11

de Man, E , 1988 Establish a Geographic Information System in Relation to its Use In *International Journal of Geographical Information Systems*, II,3 257.

Ehlers, M , Brosamle, H and J. Albrecht, 1994 Modeling Image Interpretation in Remote Sensing Through a Virtual GIS Shell (VGIS) *Proceedings ISPRS Commission III Spatial Information from Digital Photogrammetry and Computer Vision* Vol. 30 Part3/1 pp 212-218 Munich

ESRI, 1991 A Functional List of ARC/INFO Commands *In ARC/Info User's Guide· ARC command references*, I:1-15. Environmental Systems Research Institute, Redlands, CA.

Gardels, K , 1994 Virtual Geodata Model: the structural view In Buehler, K (Ed), *The Open Geodata Interoperability Specification*, preliminary report for a May 2nd, OGIS meeting at Sun Microsystems in Mountain View CA, USACERL, Champaign, Illinois

Goodchild, M 1992 *Spatial Analysis Using GIS: a seminar workbook* 2° National Center for Geographic Information and Analysis, University of California, Santa Barbara

High Performance Systems Inc. (HPS), 1994 STELLA® II an introduction to systems thinking Hanover, NH

Huxhold, W., 1989 An Introduction to Urban Geographic Information Systems Oxford University Press, New York 337 pp

Kirby, K and M Pazna, 1990 Graphic Map Algebra. In Proceedings Fourth International Symposium on Spatial Data Handling, pp 413-422 Zurich

Lanter, D and H Veregin, 1992 A Research Paradigm for Propagating Error in Layer-Based GIS. In *Photogrammetric Engineering and Remote Sensing*, LVIII,6 825-833

Lanter, D , 1994. A Lineage Metadata Approach to Removing Redundancy and Propagating Updates in a GIS Database In *Cartography and Geographic Information Systems*, XXI,2.91-98

Mark, D , 1989 Cognitive Image-Schemata for Geographic Information relations to user views and GIS interfaces In *Proceedings GIS/LIS'89*, pp 551-560 ASPRS/ACSM, Bethesda

Monckton, P., 1994, May 9 Re. GIS Tasks [Discussion]. Geographic Information Systems Discussion List [Online]. Available email. GIS-L@UBVM,CC BUFFALO EDU

Norman, D., 1991. Cognitive Artifacts In *Psychology at the Human-Computer Interface* Cambridge University Press, Cambridge

Rhind, D. and N Green, 1988 Design of a Geographic Information System for a Heterogeneous Scientific Community In *International Journal of Geographical Information Systems*, II,2 175

Stephan, E , Vckovski, A and F Bucher, 1993 Virtual Data Set an approach for the integration of incompatible data In *Proceedings Eleventh International Symposium on Computer-Assisted Cartography (Autocarto 11)*, pp 93-102 ASPRS/ACSM, Bethesda

Unwin, D, 1990. A Syllabus for Teaching Geographic Information Systems In *International Journal of Geographical Information Systems*, IV,4 461-462

Walker, D. Newman, I., Medyckyj-Scott, D. and C. Ruggles, 1992 A System for Identifying Datasets for GI Users In *International Journal of Geographical Information Systems*, VI,6.511-527.

#	name	reference	requires	necessary for	task or function	raster vector TIN all	related to
1	Data Capture / Import	R&G	0	2,3	T	A	2,3,5,6,7,8,9,17
2	Digitizing, Scanning	R&G	(17)		F	A	1,3,5,6,7,8,9,17
3	Data Storage	R&G	1	4	T	A	1,2,4
4	Data Structuring	R&G	3	5,6,7	T	A	3,5,6,7,8,9
5	Planar Enforcement		1,3	12	F	V	3,4
6	Structure Conversion	R&G	3	7,8	T	A	7,8
7	Raster/Vector	G, R&G	3,6	11	F	R	6,8
8	Quadtree/Vector	R&G	3		F	R	6,7
9	Data Validation	R&G	3	all	T	A	1,2,5,9,10,11,12,13
10	Distortion Elimination	G	15,(16),17		F	A	9,12,15,16,17
11	Sliver Polygon Removal	G	1,2,3,4,5,9,10		F	V	12,13
12	Quality Checking	A, U	3,4,9		T	A	
13	Data Editing	R&G	3	14	T	A	10,11,14,22
14	Generate Graphical Feature	G	0		F	A	
15	Geometric Conversion	R&G	3	10,18	T	A	15,16,17,18,19,20
16	Georeferencing	A, B, R&G, U	0		T	A	19,20,27,28,29,56,124, 125,126,127
17	Map Registration	A, R&G	0	2	F	A	1,2,3,4,5,6,7,8,9,10,11,12 13,14,15,16
18	Rubber-Sheet Transformation	G, R&G, U			F	A	15
19	Scale Change (simple)	B, R&G, U			F	A	16,20,27,28,29,56, 124,125,126,127
20	Projection Change	B, R&G, U			F	A	16,19,27,28,29,56, 124,125,126,127
21	Generalization	R&G			T	A	23,25,26,28,30, 69
22	Line Thinning / Smoothing	G, U			F	A	21,22,25,26,27,29,30
23	Line Generalization	A			F	A	
24	Area Centroid Calculation	G, R&G			F	A	24,31,32,33,42,52, 55,58,59,60,61,66, 114,119
25	Complex Generalization	G			T	A	
26	Aggregation	M			F	A	21, 30,69
27	Dissolve Lines and Merge Attributes	B, G			F	A	16,19,20,28,29,56,124, 125,126,127
28	Aggregation of Attributes	B, M, R&G			T	A	16,19,20,27,29,56,124, 125,126,127
29	Scale Change (incl Generalizat)	B, G			T	A	16,19,20,26,27,2829,41. 42,44,56,124,125,126, 127
30	Coordinate Thinning	R&G			F	A	21,26,69
31	Proximal Feature (Abstraction)	R&G			F	A	24,32
32	Thiessen Polygons	R&G			F	A	24,31,32,33,52,55,58,59, 60,61,66,114,119
33	Line Smoothing with Splines	B			F	V	24,32,33,52,55,58,59,60. 61,66,114, 119
34	Search	A, M			T	A	36,38,39
35	Point-in Polygon	G, M			F	A	
36	Search Attribute	G, M	39		F	A	
37	Select	R&G	39		F	A	
38	Search by Region	G, M	39		F	A	
39	Logical Condition	M			F	A	
40	Suppress (Converse of Select)	G, M	39		F	A	
41	*Measurements*	R&G			T	A	42,43,44,46,120
42	Number of Items	G, U, M			F	A	24,32,33,52,55,58,59,60, 61,66,114, 119
43	Distance	A, G, R&G, U			F	A	41,42,44,46,83,87,102, 107
44	Perimeter, Acreage, Volume	A, G, R&G, M			F	A	41,42,43,46
45	Calculate Bearing (to the north)	G			F	A	

#	name	reference	requires	necessary for	task or function	raster vector TIN all	related to
46	Direction Measurement	B			F	A	41,42,43,44
47	Calculate Height	G			F	A	
48	Calculate Angles and Distances along Linear Features	G		65	F	A	
49	Calculate Endpoint of Traverse	G			F	A	
50	Windowing, Clipping	A			F	A	
51	Overlay	A			F	A	
52	Line-on-Polygon	G, M			F	A	24,32,33,42,55,58,59,60, 61,66,114, 119
53	Graphic Overlay	G			F	A	
54	Line Intersection	A, U			F	A	
55	Polygon Overlay	G, R&G			F	A	24,32,33,42,52,57, 58,59,60,61,73,66, 95,114,119
56	Merge	A, B			F	A	16,19,20,27,28,29, 124,125,126,127
57	Route Allocation	R&G, M			G	A	55,73,95
58	Buffering	A, G, M, U			F	A	24,32,33,42,52,55,59,60, 61,66,114, 119
59	Buffer Zones	B			F	A	24,32,33,42,52,55,58,60, 61,66,114, 119
60	Simple Buffer	B			F	A	24,32,33,42,52,55,58,59, 61,66,114, 119
61	(An-)Isotropic Buffer	B			F	A	24,32,33,42,52,55,58,59, 60,66,114, 119
62	Corridors	A, U			F	A	
63	Interpolation	B			T	A	64,66,73,74,76,83,84,87, 95, 111,112,118
64	Interpolation of Heights	G			F	A	
65	Height Along Streams	G	48		F	A	
66	Interpolate Contour from Points	G, U			F	A	24,32,33,42,52,55, 58,59,60,61,114, 119
67	Weighted Modeling	G			T	A	
68	Map Algebra	A,M			T	R	
69	Reclassification	A, B, R&G			T	A	21,26,28,30
70	Arithm Algebraic, Bool Cond	B, G, M			T	A	39,40,70,71,72,109,110
71	Weighted Boolean Operations	B			F	A	70,72,109,110
72	Bayesian prior Probabilities	B			F	A	70,71,109,110
73	Slope and Aspect	B, G, R&G			F	T,R	55,57,63,74,76,83,84,87, 95, 111,112,118
74	Drainage Network	B			F	T,R	63,73,76,83,84,87,88,95, 111,112,118
75	Watershed Boundaries	G			F	A	
76	Locational Analysis -Catchment	B, M			T	A	63,73,74,83,84,87, 95, 111,112,118
77	Surface Calculation	M			T	A	71,111,112,132
78	Filtering	M			T	A	
79	Geometric Filtering	B			F	A	22,80,103
80	Image Edge Detection/Enhance	B, R&G			F	R	
81	Discrimination	B			F	A	
82	Neighborhood Analysis	M			T	A	89
83	Proximity	A			F	A	43,84,87,102,109
84	Adjacency	B			F	A	83,87
85	Contiguity Analysis	G			F	A	
86	Flow Analysis	M			T	A	48,65,87,88,90,133
87	Connectivity Analysis	B, G, M, U			F	A	43,63,73,74,76,83,84,95, 102,107,111,112,118
88	Network Analysis	A,G, M, U			T	A	74
89	Nearest Neighbor Search	G			F	A	82
90	Shortest Path	G			F	A	
91	Diffusion Models	M			T	A	119
92	Data Retrieval, incl Selection	R&G			T	A	34,36,37,38,93
93	Create Lists and Reports	B, G			F	A	99,113,115,117,123
94	Generate Viewshed Maps	G			F	A	
95	Intervisibility	R&G			F	A	55,57,63,73,74,76,83,84, 87,111,112, 118
96	Line-of -Sight	G			F	A	
97	Generate Block Diagram	G			F	A	
98	Scene generation	G			F	A	
99	Generate Cross-Section	B, G			F	A	93,113,115,117,123

#	name	reference	requires	necessary for	task or function	raster vector TIN all	related to
100	Optimal Location	M			G	A	
101	Spatial Statistics	M			T	A	130
102	Pattern	A, M, U			F	A	43,83,107
103	Complexity/Variation Measure	B, M			F	A	
104	Measures of Dispersion	M, R&G, U	42		F	A	105,109,113,115,117
105	Point Dispersion / Distribution	A, U	42		F	A	104,106,131
106	Point Centrality	A, U	42		T	A	
107	Orientation	A, U			F	A	43, 83,102
108	Multi-Criteria Decision Making	A			G	A	
109	Multivariate Analysis	B, R&G			F	A	70,71,72,104,110,115, 117
110	Regression Models	B			F	A	70,71,72,109
111	Global Surface Fitting (Trend)	B			F	A	63,73,74,76,83,84,87,95, 112,118
112	Complex Surface Fit (Fourier)	B			F	A	63,73,74,76,83,84,87,95, 111,118
113	Change Detection	B, G, M			T	A	93,99,115,117,123
114	Statistical Functions	G			F	A	24,32,33,42,52,55,58,59, 60,61,66,119
115	Histograms	B, R&G			F	A	93,99,104,109,113,117, 123
116	Discrimination	B			F	A	
117	Frequency Analysis	B, M			F	A	93,99,104,109,113,115, 123
118	Kriging	B			F	A	63,73,74,76,83,84,87,95, 111,112
119	Spread over Friction Surface	B, M			F	A	24,32,33,42,52,55, 58,59,60,61,66,119
120	Measurement of Shapes	B			F	A	41,85,87,88,121, 122
121	Topology, Description of Holes	B			F	A	41,85,87,88,120, 122
122	Topology; Upstream Elements	B			F	A	41,85,87,88,120, 122
123	Indices of Similarity	B			F	A	93,99,113,115,117
124	Map Join	B			F	A	16,19,20,27,28,29, 56, 125,126,127
125	Object Join	B			F	A	16,19,20,27,28,29, 56, 124,126,127
126	Snap	B			F	A	16,19,20,27,28,29, 56, 124,125,127
127	Scissors and Cookie Operations	B			F	A	16,19,20,27,28,29, 56, 124,125,126
128	Plotting	A			F	A	
129	Map Composition	A			T	A	
130	Spatial Autocorrelation	M			F	A	101
131	Tessellation from Point Data	U			F	R, T	42,104,105,106
132	Delineation of Homogenous Areas	M			F	A	
133	Flow Between Regions	M			F	A	
134	Image Texturing	R&G			F	R	
135	Line Fractalization	R&G			F	R	
136	Coordinate Transformation	U		18,19	T	A	18,19
137	Map Projection and Transformation	U		20	T	A	19,20
138	Location Analysis	M			T	A	35,38,52,57,58,68,73,76, 77,82,83,84,85,89,100
139	General Analytical Methods	M			T	A	101,102,103,109,110, 111,112,117,121,122, 123,130
140	Enhancement	R&G			T	A	80,134,135
141	Abstraction	R&G			T	A	
142	Spatial analysis	R&G			T	A	55,57,58,59,60,61,62,63, 64,65,66,67,68,71,72,73, 74,75,76,77,78,82,86,88, 89,90,91,94,95,96,100, 121,122,132,133
143	Statistical Analysis	R&G			T	A	45,47,48,49,81,83,84,85, 87,101,102,103,104,105, 106,107,109,110,111, 112,114,115,120,123
144	Data Display	R&G			T	A	93,97,98,99,128,129

Table 2. List of universal GIS functions drawn from cited references (some columns are yet incomplete!).

Topology of Prototypical Spatial Relations Between Lines and Regions in English and Spanish[1]

David M. Mark[2] and Max J. Egenhofer[3]
National Center for Geographic Information and Analysis

Abstract

Thirty-two native-speakers of English drew examples of roads that fit the spatial relations to a park, as indicated in 64 English-language sentence. Also, 19 native speakers of Spanish drew examples for 43 Spanish-language sentences. Then, each of the 2856 drawings (2044 English and 812 Spanish) was classified according to the road-park spatial relation into one of 19 categories of spatial relations defined by the 9-intersection model. For each of the 107 sentences, the proportion of subjects drawing each relation was determined. These counts indicate the prototypical spatial relations corresponding to each sentence. Results confirm our previous work on prototypical spatial relations: 2522 of the 2856 drawings (88 percent) fell into just 5 spatial relations, roughly equivalent to 'inside', 'outside' (disjoint), 'enters', 'crosses', and 'goes to'. Evidently, there are many ways in English and Spanish to express relations approximately corresponding to the English inside, outside, enter, cross, and goes-to, and relatively few verbally compact ways to express other spatial relations between roads and parks, perhaps lines and regions in general. The topological results suggest that English and Spanish are very similar in the ways they express road-park spatial relations in English and Spanish. Several Spanish-English pairs of sentences with similar common-sense meanings also had very similar profiles of response across the 19 spatial relation categories. Future work will examine the geometry of the examples drawn by the subjects in this experiment, and will examine other languages.

[1] This paper is a part of Research Initiative 10, "Spatio-Temporal Reasoning in GIS," of the U.S. National Center for Geographic Information and Analysis (NCGIA), supported by the National Science Foundation (SBE-88-10917); support by NSF is gratefully acknowledged. The work is also supported by a grant from the Scientific Affairs division of the North Atlantic Treaty Organization (NATO). Max Egenhofer's research is also supported by NSF under grant IRI-9309230, and by Intergraph Corporation. The experiments described herein were approved by the human subjects review procedures of the Faculty of Social Sciences, SUNY Buffalo. Terry Pratt assisted with data collection in Spain and Costa Rica. Thanks are especially due to the subjects who participated in the experiment.

[2] Department of Geography, University at Buffalo, Buffalo, NY 14261
Email: geodmm@ubvms.cc.buffalo.edu

[3] Department of Surveying Engineering and Department of Computer Science, University of Maine, Orono, ME 04469 Email: Max@grouse.umesve.maine.edu

Introduction

Spatial relations are a very important aspect of spatial cognition, spatial reasoning, and geographic information systems (Claire and Guptill, 1982; Peuquet, 1986; Abler, 1987; Pullar and Egenhofer, 1988). In a report written in 1991, Egenhofer and Herring (1994) presented the '9-intersection', a new model for characterizing spatial relations between entities in the plane. This model provides a strong formal basis for spatial relations in computational systems, but also represents a hypothesis about spatial relations in cognition and language. We have been using the 9-intersection as a framework for hypotheses, and testing these hypotheses using human subjects. The work so far has concentrated on line-region spatial relations in English, using the example of a 'road' and a 'park' (Mark and Egenhofer, 1992, 1994a, 1994b; Egenhofer and Mark in press). This paper continues that work, reporting results of a new experimental protocol (drawing task), and providing the first systematic comparative results for languages, with parallel experiments in English and Spanish.

The 9-Intersection Model for Spatial Relations

The 9-intersection model (Egenhofer and Herring, 1994) distinguishes the interior, boundary, and exterior of each spatial entity. The spatial relation between two entities is then characterized by a 3 by 3 intersection matrix, that records whether the intersection between each part (interior, boundary, exterior) of one entity and each part of the other entity is empty or not. For a two-dimensional entity (region), the definitions of interior, boundary, and exterior are the intuitive ones; for one-dimensional entities (chains or lines), the boundary consists of the two end nodes, and the interior is every other point on the line (excluding end nodes). The 9-intersection model distinguishes 8 spatial relations between two simple connected regions with no holes, 33 between two unbranched lines, and 19 distinct spatial relations between an unbranched line and a simple region.

Methods

Subjects were presented with outlines of a park, eight to a page, with a sentence describing a spatial relation between a road and a park printed under each. An example is given in Figure 1. Sixty-four sentences were tested in English, and 43 in Spanish. We assembled the sets of test sentences using a variety of techniques. For English, some sentences were group descriptions from a spatial relations grouping task (Mark and Egenhofer, 1994), and others were listed by several native English speakers. Some Spanish-language sentences were elicited from a native speaker of Spanish, and others arose from translation of some of the 64 English-language sentences, again by a native Spanish speaker. Subjects were asked to draw a road on the outline so that it would conform to the spatial relation described in the sentence. In English, the instructions were:

> On each of the following 64 diagrams, the shaded polygon represents a state park. Please draw a line on each diagram to represent a road that the spatial relationship to the park that is described by the sentence other the diagram. Try to draw a road that makes the diagram a 'best example' of the relationship described by the sentence. If you think two sentences indicate the same spatial relation, you can drawn the road in the same place to exemplify each.

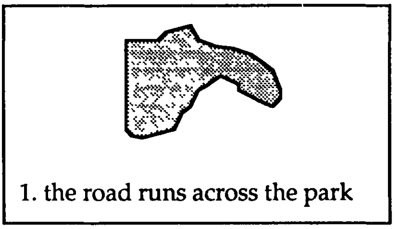


Figure 1. Example of one of the stimuli for the English-language test of production of examples.

And in Spanish:

> En cada uno de los 43 diagramas, el polígono que esta tonado representa un parque. Por favor de dibujar una línea en cada diagrama para representar una carretera que tiene la relación espacial al parque que está describida en el frase debajo del diagrama. Trata de dibujar una carretera que indica el mejor ejemplo de la relación que la frase describe. Si usted piensa que dos o mas frases indican el mismo relación espacial, puede dibujar la carretera en el mismo lugar para cada diagrama.

Once the drawings were completed, we examined each, and classified the topological spatial relation between each 'road' drawn and the 'park' into one of the 19 spatial relations. We then counted the relative frequency with which each of the 19 relations was drawn, for each of the test sentences across all of the subjects of that language.

Results

Table 1 shows the relative frequencies for each of the 19 relations (plus a category for missing or ambiguous) for each language, totaled across all sentences. We compared the relative frequencies of drawing topologies in Spanish and English by using a Chi-square test. The Chi-square test for non-parametric comparison of relative frequencies in two or more samples requires no category to have an expected value of less than five. In order to achieve this, seven topological categories that were very seldom drawn by these subjects (12, 32, 33, 62, 66, 72, and 76) were combined into a single class 'other'. With a class for cases omitted by subject, the Chi-square table becomes 14 by 2, and there are 13 degrees of freedom. The computed value of Chi-square is 100.12, which with those degrees of freedom is significantly different from zero at the % level. Three topological categories contribute 74 units to this Chi-square value: Classes 13 ('goes to'), 44 ('inside'), and 73 (a special case of 'crosses', truncated at the boundary at one end but extending outside at the other). Drawings of class 13 had a relative frequency almost twice as high for the Spanish sentences as for the English, whereas 'inside' prototypes were more than twice as frequent for the English-language sentence collection. Relation 71 occurred far more often in Spanish. Of course, this may be simply an artifact of the sentence collection, but even that could be interesting. The implication is that there are more ways to talk about 'inside' in English, and more ways to say 'goes exactly up to' in Spanish.

Mathematically, each of the 19 relations is equally unique, and distinct from all the others. Cognitively, however, they are far from equal. More than 85 percent of the drawings (88.6 % for English, 86.6 %) fell into just five of the nineteen topological classes distinguished by the 9-intersection. The drawings can be considered to be typical or prototypical of the associated spatial relationships. Interestingly, these five most frequently drawn spatial relations are ***exactly*** the same five relations that were most frequently selected as group prototypes in an open-ended grouping task with road-park drawings (Mark and Egenhofer, 1994b, Figure 11).

TABLE 1: Relative Frequencies of the Spatial Relations for Roads Drawn to Exemplify Locative Sentences, in Descending Order by Total Frequency

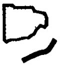
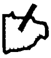
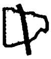
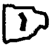
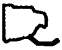
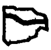
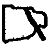
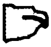
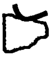
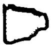
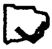
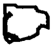
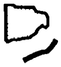
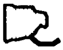
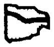
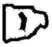
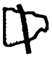
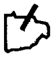
		English		Spanish		
		Number	%	Number	%	Total
11		558	27.3	214	26.4	772
75		448	21.9	196	24.1	644
71		430	21.0	156	19.2	586
44		269	13.2	47	5.8	316
13		110	5.4	94	11.6	204
42		70	3.4	28	3.4	98
73		32	1.6	32	3.9	64
46		34	1.7	7	0.9	41
31		33	1.6	7	0.9	40
22		20	1.0	1	0.1	21
74		8	0.4	3	0.4	11
64		6	0.3	3	0.4	9
12, 32, 33, 62, 66, 72, 76		26	1.3	24	3.0	50
Total		2044	100.0	812	100.0	2856

TABLE 2: Prototype = 11 (Outside)

	Test sentence	 11	 13	 42	 44	 71	 75	all other
2.05	La carretera pasa cerca del parque	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000
2.07	La carretera rodea el parque	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000
2.22	La carretera está fuera del parque	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000
2.23	La carretera va por fuera del parque	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000
2.31	La carretera esparalela al parque	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000
1.13	The road is outside the park	**0.969**	0.000	0.000	0.000	0.000	0.000	0.031
1.31	The road is entirely outside the park	**0.969**	0.000	0.000	0.031	0.000	0.000	0.000
1.43	The road encircles the park	**0.969**	0.000	0.000	0.031	0.000	0.000	0.000
2.33	La carretera está cerca del parque	**0.947**	0.000	0.000	0.000	0.000	0.000	0.053
1.11	The road ends near the park	**0.938**	0.031	0.000	0.000	0.000	0.000	0.031
1.38	The road rings the park	**0.938**	0.000	0.000	0.031	0.000	0.000	0.031
1.39	The road avoids the park	**0.938**	0.000	0.000	0.000	0.000	0.000	0.062
1.56	The road circles the park	**0.938**	0.000	0.000	0.063	0.000	0.000	0.000
1.20	The road bypasses the park	**0.906**	0.000	0.000	0.000	0.063	0.000	0.031
1.37	The road surrounds the park	**0.906**	0.000	0.000	0.031	0.000	0.000	0.063
1.40	The road goes by the park	**0.906**	0.000	0.000	0.000	0.031	0.000	0.063
1.47	The road passes the park	**0.906**	0.000	0.000	0.000	0.000	0.000	0.094
2.06	La carretera pasa lejos del parque	**0.895**	0.000	0.000	0.000	0.000	0.000	0.105
1.03	The road is near the park	**0.875**	0.000	0.031	0.000	0.000	0.000	0.094
1.52	The road encloses the park	**0.875**	0.000	0.000	0.031	0.000	0.000	0.094
1.61	The road starts near the park	**0.875**	0.031	0.000	0.031	0.031	0.031	0.001
2.08	La carretera bordea el parque	**0.842**	0.000	0.000	0.000	0.000	0.000	0.158
1.22	The road runs along the park	**0.719**	0.000	0.000	0.031	0.031	0.000	0.219
1.33	The road runs along the park boundary	**0.719**	0.000	0.000	0.031	0.000	0.000	0.250
2.32	La carretera pasa por el borde del parque	**0.684**	0.000	0.000	0.000	0.053	0.000	0.263
1.10	The road runs along the edge of the park	**0.594**	0.000	0.000	0.031	0.000	0.000	0.375
2.39	La carretera empiéza fuera del parque	**0.579**	0.158	0.000	0.000	0.000	0.263	0.000
1.53	The road starts just outside the park	**0.469**	0.031	0.000	0.000	0.094	0.406	0.000
1.58	The road goes away from the park	**0.469**	0.156	0.000	0.000	0.063	0.250	0.062
1.25	The road ends just outside the park	**0.438**	0.125	0.000	0.000	0.156	0.250	0.031
2.42	La carretera termina nada mas fuera del parque	**0.316**	0.158	0.000	0.053	0.158	0.211	0.104
2.04	La carretera recorre el parque	**0.263**	0.000	0.000	0.211	0.158	0.211	0.157

Tables 2-6 include the 107 sentences tested, grouped by the topology most frequently drawn to represent that sentence. Within each table, the sentences are listed in descending order of the relative frequency of the most typical spatial relation. Relative frequencies for the other prototypical classes are in the columns of each table. (The one sentence dominated by some other spatial relation, a special case of crosses, is included in Table 5.) The first column of each table is a sentence number, beginning with 1 for English and 2 for Spanish.

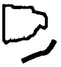
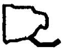
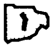
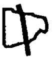
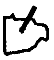
Outside

Thirty-two sentences (20 English, 12 Spanish; see Table 2) were most frequently illustrated by roads drawn entirely outside (disjoint from) the park. Some Spanish-English pairs of equivalent sentences had remarkably similar frequencies for their most frequent classes (for example, Table 2, sentences 2.32 and 1.10, or sentences 1.25 and 2.42).

Inside

Only 14 sentences (Table 3) had most frequently drawn topologies in which the road was entirely inside the park. Whereas seven English-language sentences were exemplified by drawing roads entirely inside the park by more than 80 percent of the subjects, the highest frequency for drawings inside for any Spanish sentence was 52.6 percent (sentence 2.24).

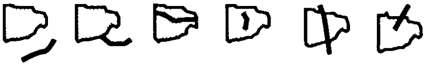
TABLE 3: Prototype = 44 (Inside)

								
	Test sentence	11	13	42	44	71	75	all other
1.50	The road is inside the park	0.000	0.000	0.000	**0.938**	0.031	0.000	0.031
1.08	The road is enclosed by the park	0.000	0.000	0.000	**0.875**	0.031	0.000	0.094
1.24	The road is within the park	0.000	0.000	0.000	**0.875**	0.000	0.000	0.125
1.06	The road is contained entirely within the park	0.000	0.000	0.094	**0.844**	0.000	0.000	0.062
1.26	The road starts and ends in the park	0.000	0.000	0.031	**0.844**	0.000	0.000	0.125
1.29	The park encloses the road	0.000	0.000	0.000	**0.844**	0.000	0.031	0.125
1.42	The road is in the park	0.000	0.000	0.031	**0.844**	0.031	0.000	0.094
1.18	The road is entirely contained in the edge of the park	0.000	0.000	0.063	**0.531**	0.000	0.000	0.406
2.24	La carretera comienza y caba en el parque	0.053	0.000	0.211	**0.526**	0.000	0.000	0.210
1.16	The road connects portions of the park	0.000	0.000	0.188	**0.500**	0.063	0.063	0.186
2.36	La carretera está confinada en el parque	0.053	0.000	0.105	**0.421**	0.053	0.105	0.263
2.20	La carretera está contenida en el parque	0.000	0.000	0.158	**0.368**	0.211	0.158	0.105
2.28	La carretera está dentro del parque	0.000	0.000	0.263	**0.368**	0.053	0.158	0.158
2.30	La carretera está en el parque	0.000	0.000	0.053	**0.368**	0.158	0.000	0.421

Crosses, Enters, Goes To

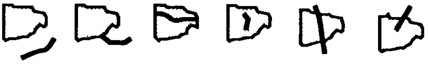
Twenty-four sentences (Table 4) were illustrated by roads that 'crossed' or otherwise 'bisected' the park. Another very basic type spatial relationship is a path from the inside to the outside of a container, which we refer to here as 'enters'; this spatial relation was most-frequent for 30 of the sentences (Table 5). Nine others involved roads which had one end on the park boundary but which otherwise were outside the park (see Table 6). In all of these tables, there are pairs of sentences which seem to be direct translations of each other, and for which the frequencies of different topological examples are also about equal. For example, in Table 5, "La carretera va al parque" was exemplified by a 'goes to' relation by 63 % of Spanish-language subjects, while "The road goes to the park" was drawn that way by 62 % of English-language subjects.

TABLE 4: Prototypes = 42,71 (Crosses)



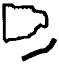
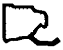
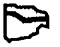
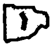
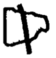
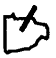
	Test sentence	11	13	42	44	71	75	all other
1.57	The road comes through the park	0.000	0.000	0.000	0.000	**0.969**	0.000	0.031
2.27	La carretera divide el parque	0.000	0.000	0.053	0.000	**0.947**	0.000	0.000
1.48	The road transects the park	0.000	0.000	0.000	0.031	**0.938**	0.000	0.031
1.32	The road divides the park	0.000	0.000	0.031	0.000	**0.906**	0.000	0.063
1.54	The road crosses the park	0.000	0.000	0.000	0.000	**0.906**	0.031	0.063
1.21	The road cuts through the park	0.000	0.000	0.031	0.000	**0.875**	0.031	0.063
2.02	La carretera atraviesa el parque	0.000	0.000	0.105	0.000	**0.842**	0.000	0.053
2.03	La carretera cruza el parque	0.000	0.000	0.158	0.000	**0.842**	0.000	0.000
2.38	La carretera y el parque se cruzan	0.000	0.053	0.000	0.000	**0.842**	0.053	0.052
1.12	The road goes through the park	0.000	0.000	0.156	0.000	**0.813**	0.031	0.000
1.14	The road splits the park	0.000	0.000	0.094	0.000	**0.813**	0.031	0.062
1.17	The road cuts across the park	0.000	0.000	0.094	0.031	**0.813**	0.000	0.062
1.30	The road bisects the park	0.000	0.000	0.031	0.000	**0.813**	0.000	0.156
1.09	The road cuts the park	0.000	0.000	0.156	0.000	**0.781**	0.063	0.000
2.18	La carretera va a través del parque	0.000	0.000	0.105	0.053	**0.737**	0.000	0.105
1.27	The road goes across the park	0.000	0.000	0.156	0.125	**0.688**	0.000	0.031
2.01	La carretera pasa por el parque	0.211	0.000	0.053	0.000	**0.684**	0.000	0.052
1.59	The road traverses the park	0.094	0.000	0.000	0.094	**0.656**	0.031	0.125
1.46	The road intersects the park	0.000	0.156	0.000	0.031	**0.594**	0.094	0.125
1.41	The road and the park intersect	0.000	0.219	0.000	0.000	**0.531**	0.156	0.094
1.01	The road runs across the park	0.000	0.000	0.469	0.000	**0.500**	0.000	0.031
2.17	La carretera conecta partes del parque	0.000	0.053	0.105	0.105	**0.316**	0.158	0.263
2.25	La carretera termina fuera del parque	0.158	0.053	0.000	0.000	**0.316**	0.316	0.157
1.04	The road spans the park	0.125	0.000	**0.438**	0.125	0.219	0.000	0.093

TABLE 5: Prototype = 75 (Enters)



	Test sentence	11	13	42	44	71	75	all other
1.49	The road goes into the park	0.000	0.031	0.000	0.000	0.031	**0.906**	0.032
1.35	The road comes out of the park	0.000	0.000	0.000	0.000	0.063	**0.906**	0.031
1.51	The road comes from the park	0.000	0.063	0.000	0.031	0.000	**0.906**	0.000
2.16	La carretera termina dentro del parque	0.000	0.000	0.000	0.000	0.000	**0.895**	0.105
1.07	The road ends in the park	0.000	0.031	0.000	0.031	0.000	**0.875**	0.063
1.28	The road ends just inside the park	0.000	0.000	0.031	0.063	0.031	**0.875**	0.000
1.36	The road leaves the park	0.000	0.000	0.000	0.000	0.063	**0.875**	0.062
1.45	The road comes into the park	0.000	0.000	0.000	0.000	0.063	**0.875**	0.062
1.15	The road exits the park	0.000	0.125	0.000	0.000	0.031	**0.781**	0.063
1.05	The road enters the park	0.000	0.188	0.000	0.000	0.031	**0.750**	0.031
2.11	La carretera se pierde en el parque	0.000	0.053	0.000	0.000	0.000	**0.737**	0.210
1.63	The road starts just inside the park	0.031	0.000	0.000	0.125	0.000	**0.719**	0.125
1.64	The road runs into the park	0.000	0.063	0.000	0.000	0.188	**0.719**	0.030
1.34	The road starts in the park	0.000	0.000	0.000	0.219	0.000	**0.688**	0.093
2.19	La carretera empiéza dentro del parque	0.053	0.000	0.053	0.000	0.053	**0.684**	0.157
2.21	La carretera entra en el parque	0.053	0.000	0.000	0.000	0.263	**0.632**	0.052
2.34	La carretera sale del parque	0.053	0.211	0.000	0.000	0.053	**0.632**	0.051
2.41	La carretera viene del parque	0.105	0.211	0.000	0.000	0.000	**0.632**	0.052
1.02	The road goes out of the park	0.000	0.156	0.031	0.000	0.031	**0.625**	0.157
2.10	La carretera se aventura en el parque	0.053	0.000	0.000	0.000	0.316	**0.579**	0.052
1.44	The road starts outside the park	0.281	0.031	0.000	0.000	0.125	**0.563**	0.000
2.40	La carretera se mete en el parque	0.000	0.105	0.000	0.000	0.316	**0.526**	0.053
2.12	La carretera muere en el parque	0.053	0.474	0.000	0.000	0.000	**0.474**	0.000
2.14	La carretera acaba en el parque	0.105	0.368	0.000	0.000	0.000	**0.474**	0.053
2.09	La carretera se interna en el parque	0.000	0.053	0.053	0.000	0.263	**0.421**	0.210
2.13	La carretera termina en el parque	0.158	0.368	0.000	0.000	0.000	**0.421**	0.053
2.35	La carretera deja el parque	0.158	0.211	0.000	0.000	0.105	**0.421**	0.105
2.37	La carretera empréza en el parque	0.053	0.263	0.000	0.000	0.000	**0.421**	0.263
1.19	The road ends outside the park	0.344	0.094	0.000	0.000	0.063	**0.375**	0.124
2.25	La carretera termina fuera del parque	0.158	0.053	0.000	0.000	0.316	**0.316**	0.157

TABLE 6: Prototype = 13 (Goes to)

								
	Test sentence	11	13	42	44	71	75	all other
2.15	La carretera está conectada al parque	0.053	**0.684**	0.000	0.000	0.158	0.105	0.000
2.29	La carretera va al parque	0.158	**0.632**	0.000	0.000	0.000	0.211	0.000
1.23	The road goes to the park	0.000	**0.625**	0.000	0.000	0.063	0.313	0.000
2.12	La carretera muere en el parque	0.053	**0.474**	0.000	0.000	0.000	0.474	0.000
2.26	La carretera está conectada con el parque	0.053	**0.474**	0.000	0.000	0.316	0.158	0.000
1.55	The road ends at the park	0.063	**0.469**	0.000	0.031	0.000	0.375	0.062
1.60	The road goes up to the park	0.188	**0.469**	0.031	0.000	0.031	0.188	0.093
2.43	La carretera llega hasta el parque	0.158	**0.368**	0.000	0.000	0.000	0.263	0.211
1.62	The road is connected to the park	0.000	**0.344**	0.000	0.031	0.188	0.188	0.249

Summary

As in our previous work, this paper shows that the 9-intersection model provides a good basis for cognitive and linguistic analysis of spatial relations. A small number of line-region spatial relations made up the great majority of drawings produced to exemplify 107 sentences. Over-all frequency of the spatial relations was significantly different for the English and Spanish data. However, similarities are more striking than differences. In quite a few cases, pairs of synonymous Spanish and English sentences appear consecutive, or at least close together, in the summary tables, indicating that their ranges of topological 'meanings' are very similar.

Also of interest is the fact that some natural language road-park sentences show strong agreement in the topology drawn by subjects, and others do not. For 26 of the 107 sentences, more than 90 percent of subjects drew examples with the same topology. It would probably be 'safe' to map these sentences onto particular spatial-relation predicates for database queries or spatial reasoning. On the other hand, there were 20 other sentences for which no single topological pattern was drawn by even 50 percent of subjects. If a user typed or spoke one of these sentences into a natural-language interface for a database or GIS, it might be difficult to satisfy the user's request with a formal query based on the 9-intersection. In some of these sentences, the query might be satisfied by the union of several relations, but others, for example "the road ends outside the park," are ambiguous, since although one end of the road must be outside the park, the sentence does not specify where the other end, or the body of the road, must fall. Also of interest is the fact that the sentences we have tested using an agreement task (Mark and Egenhofer, 1994a, 1994b) all have high consensus in the present study, with 90.6 % of subjects drawing the same topology for 'crosses' and 'goes into', 81.3% for 'goes through', 75 % for 'enters', 68.8 % for 'goes across'. The agreement task might produce results difficult to interpret if applied to sentences with low subject agreement in this drawing task.

In summary, the patterns for responses in Spanish and English show a great deal of similarity, and the results suggest that many spatial relations between roads and parks can be translated between Spanish and English in a fairly direct manner. Since Spanish and English are from quite different branches of the Indo-European languages, this result suggests that natural-language systems for at least the Romance and Germanic languages might be cross-linguistically robust. On the other hand, results suggest that topology based on the 9-intersection model can account for most variation in the meanings of some sentences, but relatively little for others. Whether some of this remaining variation can be accounted for by geometric factors (shapes, lengths, angles) will await further studies.

References

Abler, R. F., 1987. The National Science Foundation National Center for Geographic Information and Analysis. *International Journal of Geographical Information Systems*, 1(4), 303-326.

Claire, R., and Guptill, S., 1982. Spatial operators for selected data structures. *Proceedings, Auto-Carto 5*, Crystal City, Virginia, pp. 189-200.

Egenhofer, M., and Herring, J., 1994. Categorizing Topological Spatial Relations Between Point, Line, and Area Objects. In Egenhofer, M. J., Mark, D. M., and Herring, J. R., 1994. The 9-Intersection: Formalism and its Use For Natural-Language Spatial Predicates. Santa Barbara, CA: National Center for Geographic Information and Analysis, Report 94-1.

Egenhofer, M. J., and Mark, D. M., Modeling Conceptual Neighborhoods of Topological Relations. International Journal of Geographical Information Systems, in press.

Mark, D. M., and Egenhofer, M., 1992. An Evaluation of the 9-Intersection for Region-Line Relations. In: GIS/LIS '92, San Jose, CA, pp. 513-521.

Mark, D. M., and Egenhofer, M. J., 1994a. Calibrating the Meanings of Spatial Predicates From Natural Language: Line-region Relations. Proceedings, Spatial Data Handling 1994, pp. 538-553.

Mark, D. M., and Egenhofer, M. J., 1994b. Modeling Spatial Relations Between Lines and Regions: Combining Formal Mathematical Models and Human Subjects Testing. Cartography and Geographic Information Systems, October 1994, in press.

Peuquet, D. J., 1986. The use of spatial relationships to aid spatial database retrieval. *Proceedings, Second International Symposium on Spatial Data Handling*, Seattle, Washington, 459-471.

Pullar, D. V., and Egenhofer, M. J., 1988. Toward formal definitions of topological relations among spatial objects. *Proceedings, Third International Symposium on Spatial Data Handling*, Sydney, Australia, August 17-19, 1988, 225-241.

Talmy, L., 1983. How language structures space. In H. Pick and L. Acredolo (editors) Spatial Orientation: Theory, Research and Application. Plenum Press.

QUALITY AND RELIABILITY IN REPRESENTATIONS OF SOIL PROPERTY VARIATIONS

Howard Veregin[1], Kate Beard[2] and Bheshem Ramlal[2]

[1] Department of Geography
413 McGilvrey Hall
Kent State University
Kent, OH 44242-0001

[2] Department of Surveying Engineering/NCGIA
5711 Boardman Hall
University of Maine
Orono, ME 04469-5711

ABSTRACT

Soil maps portray variations in the physical and chemical properties of soil using a mapping unit model. According to this model, soil properties are homogeneous within mapping units and change abruptly at mapping unit boundaries. This model masks within-unit variations in soil properties and accentuates discontinuities between adjacent mapping units.

The purpose of this study is to explore the reliability with which spatial variations in soil properties are represented on digital soil maps. The approach involves kriging of soil property data collected at point locations and comparison of the resulting interpolated surfaces against the values predicted by the traditional soil map. The study area is Bear Brook Watershed in eastern Maine. Available data include a detailed, Level I soil survey, and independent estimates of depth to bedrock derived from a seismic survey.

Results suggest that the quality of the interpolated surface can vary significantly with changes in the number of points used in interpolation. Single measures of quality, such as the mean squared error, may alone be insufficient to assess the reliability of an interpolated surface. Excessive smoothing of surfaces can be offset by merging the interpolated surface with information from the soil map. In the present study, however, the resulting hybrid surfaces have relatively low reliability, due to apparent systematic bias in depth estimates derived from the soil map.

INTRODUCTION

In an environment in which federal agencies will be asked to document the quality of their products to be in compliance with the new Content Standards for Digital Geospatial Metadata (FGDC 1994), methods of quality reporting for geographical databases need to be addressed. Currently soil survey reports give little or no information on the quality of soil maps (Heuvelink and Bierkins 1992). The quality of the map is also difficult to assess in hindsight as field information has been massaged into a cartographic product for which quality assessment is nearly impossible.

Soil maps are made by various survey methods but most include variations on the following steps: definition of classes of soil profiles; observation of soil profiles in the study area; delineation and naming of areas corresponding to the named soil classes. Soil classification is based on several definitive soil properties. Once classified, prototype classes form the basis for predicting individual properties. The soil surveyor identifies locations in the field for test profiles and describes the vertical variation of the soil by distinguishing different soil horizons. Lateral variations are delineated as soil unit boundaries on aerial photo-based field sheets.

This approach yields maps which display classes of soils as discrete units with sharp boundaries. This representation has several advantages from a production perspective but several disadvantages for map users. Soil classification is a useful abstraction method for condensing and describing the continuum of the soil surface. However, soil properties can exhibit significant amounts of internal variability and often do not show sharp discontinuities associated with mapping unit boundaries (Burrough 1986).

Users of soil information are often interested in the individual soil properties or suitability ratings of a soil class rather than the class per se. Hence they are more likely to be directly interested in the reliability of the spatial distribution of the property of interest rather than the reliability of the soil classes. Soil scientists have investigated different methods to test the quality of soil maps and recommended survey methods for improving quality (Marsman and de Gruijter 1986). The quality of soil maps has typically been assessed by validating the classification methods. This approach however does not address the reliability of the spatial distribution of any individual property.

This paper looks at an alternate form for representing soil information. The emphasis is on generating spatial representations of individual soil properties as opposed to discrete soil classes. The advantage lies in the ability to derive reliability estimates for the spatial representation of individual soil properties. Improvements in computing power and geostatistical tools make such an approach feasible.

The purpose of this study is to explore the quality with which patterns of spatial variation in soil properties are represented on digital soil maps. The study area is Bear Brook Watershed in eastern Maine. Available data include a detailed, Level I soil survey, and independent estimates of depth to bedrock derived from a seismic survey. Depth to bedrock data at sampled locations are used to derive a smooth statistical surface based on kriging. The resulting surface is then compared against the values predicted by the traditional soil map.

Ordinary kriging assumes that soil properties vary smoothly over space, rather than being constrained by the boundaries of mapping units. Because the distribution of soil properties may also contain a discontinuous component, methods for incorporating information about discontinuous variation into the interpolation procedure are also considered. Visualization techniques are presented that facilitate comparison of the interpolated surfaces and their mapping unit-based counterparts.

BACKGROUND

Empirical studies indicate that even at relatively large map scales it is generally not feasible to delineate mapping units within which soil properties are strictly homogeneous (Beckett and Burrough 1971). The degree of internal variability depends on mapping unit size, which is constrained by such factors as map scale and purpose. Maps of larger scale can depict variation at higher frequencies but this necessarily entails a reduction in map simplicity and interpretability. Internal variation in soil mapping units is thus normally viewed as an inevitable consequence of soil mapping procedures.

One solution to this problem is to allow soil properties to vary continuously over space rather than forcing these properties to honor the locations of mapping unit boundaries. This can be achieved through interpolation of soil property values at a set of regularly-spaced grid locations. Grid estimates are weighted combinations of the values at neighboring sampled points. Interpolation can be achieved using a variety of methods, including trend surface analysis, kriging and cubic splines (Lam 1981, Laslett et al. 1987). Interpolated surfaces tend to be smoother than traditional soil maps and are generally more representative of values at sampled point locations (Voltz and Webster 1990).

The interpolation method that has met with the most success to date is kriging. Kriging yields unbiased, minimum-variance estimates and provides an estimate of the interpolation error variance at each grid location (Burgess and Webster 1980). Kriging models spatial variation as a composite of systematic and random components. This distinction is somewhat scale-dependent, since random variation at one scale may be resolvable at another. Normally, kriging assumes that the variable of interest belongs to a random field with constant mean (stationarity) and a semivariance that depends only on distance between sample locations (isotropy). Under these conditions the interpolation error has a mean of zero and a variance that depends on the location of sample points.

Various authors have demonstrated the potential of kriging for interpolating soil property data (Oliver and Webster 1986, Bregt, Bouma and Jellinek 1987, Bregt and Beemster 1989). In general, properties are represented with higher accuracy on surfaces derived from kriging than on traditional soil maps. Soil property values derived from soil maps are reasonably good predictors of mean soil property values within mapping units. However, due to their reliance on the mapping unit model, soil maps are unable to capture much of the spatial variation in soil properties. This variation often occurs at a relatively high spatial frequency and makes it difficult to predict soil property values at specific locations

One limitation of kriging is that it can cause excessive smoothing, thus eliminating any discontinuities that do exist in soil property distributions. Several studies have experimented with methods for incorporating information about such discontinuities into the kriging model (Heuvelink and Bierkens 1992, Van Meirvenne et al. 1994). The most common method of combining estimates is to use a weighted sum of the kriged value and the value derived from soil map. Weights can be selected to reflect the error variances of the two sets of values, with the contribution of each source inversely proportional to its error variance. Results suggest that the combination of data from kriging and soil maps can increase accuracy relative to estimates derived from either kriging or soil maps alone.

METHODS

Our study area is Bear Brook Watershed in eastern Maine. Data for this area includes a detailed, Level I soil survey, and independent estimates of depth to bedrock data. Depth values were measured for 47 point locations in the watershed by the Center for Earth and Environmental Science, State University of New York at Plattsburgh (Bogucki and Greundling, no date). Depth values were obtained using seismic refraction techniques. The accuracy of these estimates is reported to be on the order of ± 0.3 m.

Only that area of the watershed within the minimum bounding rectangle defined by the 47 point locations is used in the analysis. The mapping units from the soil survey are shown in Figure 1. Point locations are also shown in this figure.

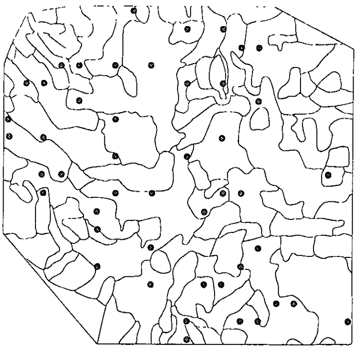


Figure 1. Soil mapping units and point locations for seismic data.

Table 1 contains summary depth statistics for soil mapping units and point locations. Note that for the soil data, depth to bedrock is given as a range of values for each mapping unit.

Component	n	Mean	Std.Dev.	Min.	Max.
Minimum depth (soil map)	104	0.706	0.540	0.0	1.524
Maximum depth (soil map)	104	1.067	0.358	0.0	1.524
Point data	47	2.606	0.975	0.0	5.200

Table 1. Descriptive statistics for depth to bedrock (in meters).

The seismic depth data at the 47 sampled points are used to derive interpolated surfaces using kriging. Since there are too few points to allow for separate calibration and validation subsets, cross-validation techniques based on jack-knifing are used to assess the quality of the interpolated surface. Two measures of quality are used. The first, mean error (ME), is a measure of bias in interpolation,

$$ME = \frac{1}{n} \sum_{i=1}^{n} e_i \tag{1}$$

The second index is the mean squared error (MSE),

$$MSE = \frac{1}{n} \sum_{i=1}^{n} e_i^2 \tag{2}$$

In the equations, n is the number of sampled points and e_i is the difference between the seismic depth value and the kriged value for point i.

Kriging can be performed using various neighborhood sizes. The neighborhood size affects the number of points used to interpolate the value at a given grid point location. It is generally held that as the influence of points decreases with distance, only a small neighborhood is required to obtain an accurate estimate at a given grid location (Burgess and Webster 1980). In this study various neighborhood sizes between 5 and 45 sample points are examined.

Comparison of the soil map depth estimates to the point data is also achieved using the ME and MSE statistics. In this case, e_i is the difference between the seismic depth value for point i and the soil map depth value for the mapping unit within which point i is located. For comparison of the soil map depth estimates to the kriged data, e_i is the difference between the kriged value at point i and the soil map depth value for the mapping unit within which point i is located.

Methods for incorporating information about discontinuities in the distribution of depth to bedrock data are also considered. The approach adopted here is to produce a hybrid surface as a weighted combination of the soil map depth values and the values derived from the kriged surface. Other authors have suggested using weights based on the interpolation error variance (Heuvelink and Bierkens 1992). However, there is no estimate of error variance available for the soil data in the present study. For this reason, our approach is to examine various combinations of weights that favor either the kriged data or the soil map. The quality of these results is assessed with cross-validation procedures based on the ME and MSE indices.

RESULTS

Neighborhood Size

Kriging can be performed using different neighborhood sizes. Our analysis examines neighborhoods of between 5 and 45 points Results are presented in Table 2. In all cases a linear variogram model (with sill) is used, as this provides the closest fit with observed data. The parameters of the model are as follows: a = 390.0 m, c_0 = 0.67 m, c = 0.043 m.

Number of points	ME (m)	MSE (m^2)
5	0.069	1.721
10	0.058	1.569
20	-0.097	1.118
30	0.053	1.006
40	-0.029	0.659
45	-0.139	0.782

Table 2. Cross-validation results for surfaces derived from kriging.

The table indicates that as neighborhood size increases, bias (ME) fluctuates. However, MSE declines consistently as the neighborhood size increases (with the exception of the largest neighborhood of 45 points). On the basis of these results, one might conclude that a larger neighborhood produces a more accurate surface. This is true from the standpoint of the MSE index. However, larger neighborhoods also tend to produce surfaces with unrealistically high degrees of smoothness. This effect is seen in Figure 2, which shows two kriged surfaces, one based on a neighborhood size of 5 points and one based on a neighborhood size of 20 points. While the second of these has a lower MSE, it is also extremely smooth. At the largest neighborhood sizes, the kriged surface is nearly flat.

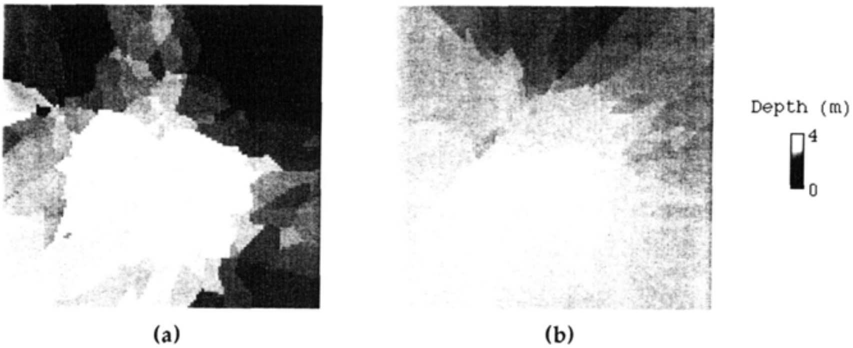


Figure 2. Examples of surfaces derived by kriging.
(a) Linear model, 5 points. **(b)** Linear model, 20 points.

This result suggests that there is a need to base the selection of an appropriate neighborhood size on criteria other than MSE. One suggestion is to select a surface with an acceptable level of accuracy that has the same overall semivariogram as the original data. For the depth to bedrock data, this criterion suggests a neighborhood size of about 5 cells. (This neighborhood size is used in the remainder of the study.)

Soil Data Quality

Figure 3 shows the depth to bedrock values derived from the soil map.

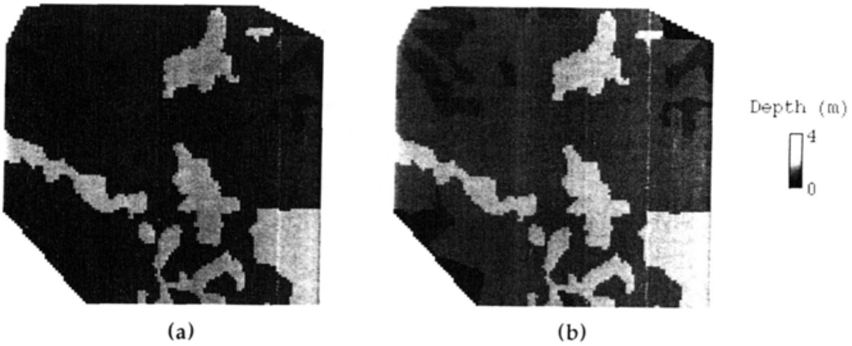


Figure 3. Soil map depth to bedrock values.
(a) Minimum depth. **(b)** Maximum depth.

Table 3 compares depth estimates from the soil map and seismic data. Results indicate a consistent bias in soil map data. If bias were absent, a positive ME value would be observed for minimum depth (i.e., seismic value > minimum) and a negative ME value would be observed for maximum depth (i.e., seismic value < maximum).

Component	ME (m)	MSE (m^2)
Minimum depth	1.886	4.633
Maximum depth	1.486	3.173

Table 3. Deviations between soil map depth estimates and seismic data.

Table 4 compares soil map depth estimates to kriged data. Results are similar to those for soil data and kriged estimates (Table 3).

Component	ME (m)	MSE (m^2)
Minimum depth	1.919	3.992
Maximum depth	1.519	2.505

Table 4. Deviations between soil map depth estimates and kriged surface.

Combined Estimation

Kriged surfaces have low error but are relatively smooth compared to the soil map. Information on discontinuities can be incorporated into the interpolation procedure by producing a hybrid surface as a weighted combination of the soil depth values and the depth value from the kriged surface. Table 5 shows the cross-validation results for various combinations of weights. The table indicates that accuracy decreases as the soil map is weighted more heavily. This is due to the apparent bias in depth estimates from soil maps (Tables 3 and 4).

Relative weights		ME	MSE
Soil map	Kriging	(m)	(m^2)
0.25	0.75	0.637	1.814
0.50	0.50	1.230	2.756
0.75	0.25	1.826	4.429

Table 5. Cross-validation results for hybrid surfaces.

Figure 4 shows the hybrid surface derived using equal weights for the soil map and kriged surface (using a linear semivariogram model with a neighborhood size of 5 points). The surface shares some of the characteristics of the smoother kriged surfaces and the soil map.

Visualization

Figure 5 shows a pair of difference maps computed for the depth data derived from soil maps and the hybrid surface in Figure 4. Figure 5a shows overestimation in soil data (i.e., locations where the soil map minimum depth value is greater than the interpolated value on the kriged surface). Figure 5b shows underestimation in soil data (i.e., locations where the soil map maximum depth value is less than the kriged surface). The degree of over- or under-estimation is shown in shades of gray.

Underestimation appears to be higher in prevalence and degree than overestimation. This results from the general tendency for the soil map to underestimate depth values relative to the seismic data (Tables 3 and 4). Areas of underestimation are associated with the fact that the hybrid surface is smoother than the soil map. Thus errors tends to occur where the soil map deviates away from the smoother interpolated surface.

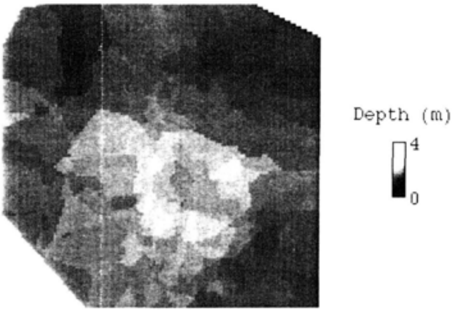
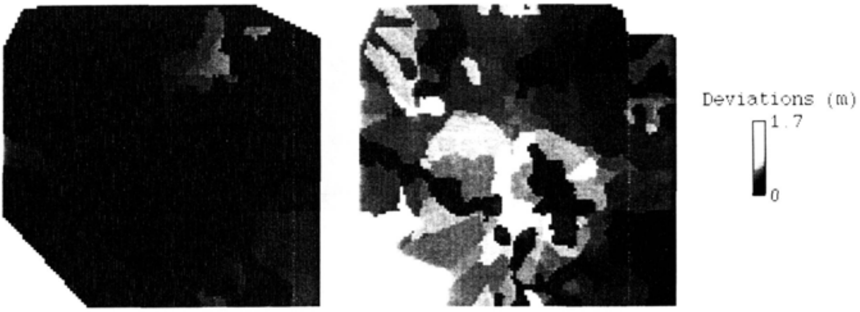


Figure 4. Hybrid surface.



(a) **(b)**

Figure 5 Differences between soil map and hybrid surface.
(a) Overestimation in soil map. **(b)** Underestimation in soil map.

CONCLUSIONS

The quality of the surfaces derived from kriging is dependent on neighborhood size. As neighborhood size increases (i.e., a larger number of points is used in interpolation) the MSE declines fairly consistently, indicating an improvement in interpolation accuracy. The ME index, indicative of bias, fluctuates without a definite pattern. A larger neighborhood is also associated with increased smoothing of the interpolated surface. This suggests that there is a need to look at criteria other than MSE when evaluating the quality of interpolation. One suggestion is to use the interpolated surface that yields approximately the same semivariogram as the original point data values, such that the original spatial pattern is preserved.

Results also show consistent bias (under-estimation) in soil depth values relative to seismic data. This suggests a systematic difference in the way in which depth is measured in the soil map and by seismic techniques. This type of systematic error may not be typical of all soil properties. Merging of soil and kriged data yields hybrid surfaces that are less smooth than the original kriged surfaces. In this study, the quality of these surfaces is lower than that of the kriged surfaces due to systematic differences in soil and seismic data. Quality declines as the relative weight of soil data is increased.

REFERENCES

Beckett, P.H.T. & Burrough, P.A. 1971, The relation between cost and utility in soil survey. *Journal of Soil Science* 22, 466-480.

Bogucki, D.J. & Greundling, G.K. (No date). *Depth to Bedrock at the Bear Brook Watershed, Watershed Manipulation Project.* Unpublished Manuscript, Center for Earth and Environmental Science, State University of New York at Plattsburgh.

Bregt, A.K., & Beemster, J.G.R. 1989, Accuracy in predicting moisture deficits and changes in yield from soil maps. *Geoderma 43,* 301-310.

Bregt, A.K., Bouma, J. & Jellinek, M. 1987, Comparison of thematic maps derived from a soil map and from kriging of point data. *Geoderma 39,* 281-291.

Burgess, T.M. & Webster, R. 1980, Optimal interpolation and isarithmic mapping of soil properties. *Journal of Soil Science 31,* 315-331.

Burrough, P.A. 1986, *Principles of Geographic Information Systems for Land Resources Assessment.* Oxford: Clarendon.

FGDC 1994, *Content Standard for Digital Geospatial Metadata.* Draft document, US Geological Survey.

Heuvelink, G.B.M. & Bierkens, M.F.P. 1992, Combining soil maps with interpolations from point observations to predict quantitative soil properties. *Geoderma 55,* 1-15.

Lam, N. S. 1983, Spatial Interpolation methods: A review. *The American Cartographer 10,* 129-149.

Laslett, G.M., McBratney, A.B., Pahl, P.J., & Hutchinson, M.F. 1987, Comparison of several spatial prediction methods for soil pH. *Journal of Soil Science 38,* 325-2341.

Marsman B. A. & de Gruijter, J.J. 1986, *Quality of soil maps.* Soil Survey Institute, Wageningen, The Netherlands.

Oliver, M.A. & Webster, R. 1986, Combining nested and linear sampling for determining the scale and form of spatial variation of regionalised variables. *Geographical Analysis 18,* 227.

Van Meirvenne, M., Scheldeman, K., Baert, G. & Hofman, G. 1994, Quantification of soil textural fractions of Bas-Zäire using soil map polygons and/or point observations. *Geoderma 62,* 69-82.

Voltz, M. & Webster, R. 1990, A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *Journal of Soil Science 41,* 473-490.

ON UNCERTAINTY IN GEOGRAPHIC DATABASES DRIVING REGIONAL BIOGEOCHEMICAL MODELS[*] (EXTENDED ABSTRACT)

Ferenc Csillag

Geography Department & Institute for Land Information Management
University of Toronto, Erindale College
Mississauga, ONT, L5L 1C6, Canada
<fcs@eratos.erin.utoronto.ca>

ABSTRACT

Large regional geographic databases are becoming crucially important in driving regional environmental models. When these databases contain information from various sources about different geographic phenomena, defining a common reference of "environmental units" may be difficult. A general framework is proposed to evaluate regular and irregular landscape partitioning strategies, as well as numerous spatial predictors to provide input data for complex models. The strategy is based on information available on spatial (and temporal) variability. It is concluded that it is advantageous to adjust the "effective scale" of representation to local variability.

INTRODUCTION

There is increasing demand to link environmental simulation models and geographic information systems (GIS) to predict the status of various environmental phenomena. Typically, these efforts extend a *site-based (process) model* over a region (Figure 1), which may be several orders of magnitude larger than "calibrated sites". Therefore, the extrapolation requires some *spatial model*, which accounts for the spatial variability of the landscape.

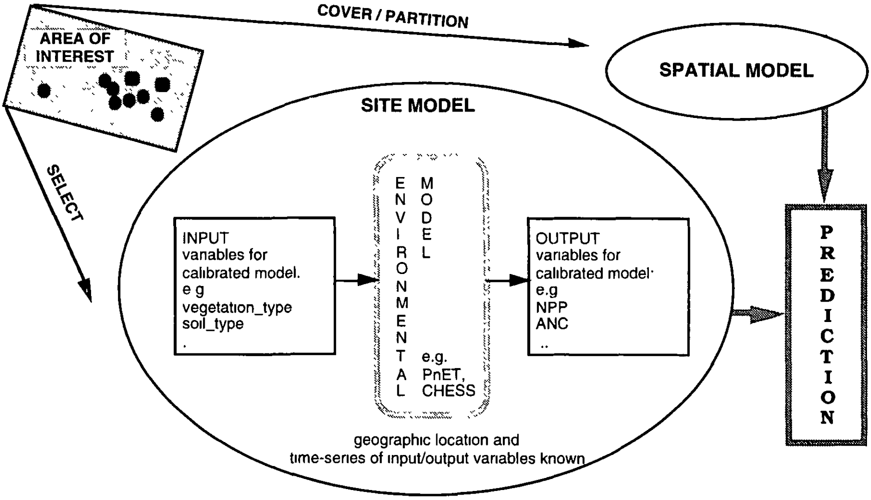


Figure 1. Linking "calibrated" site-based models with GIS for prediction

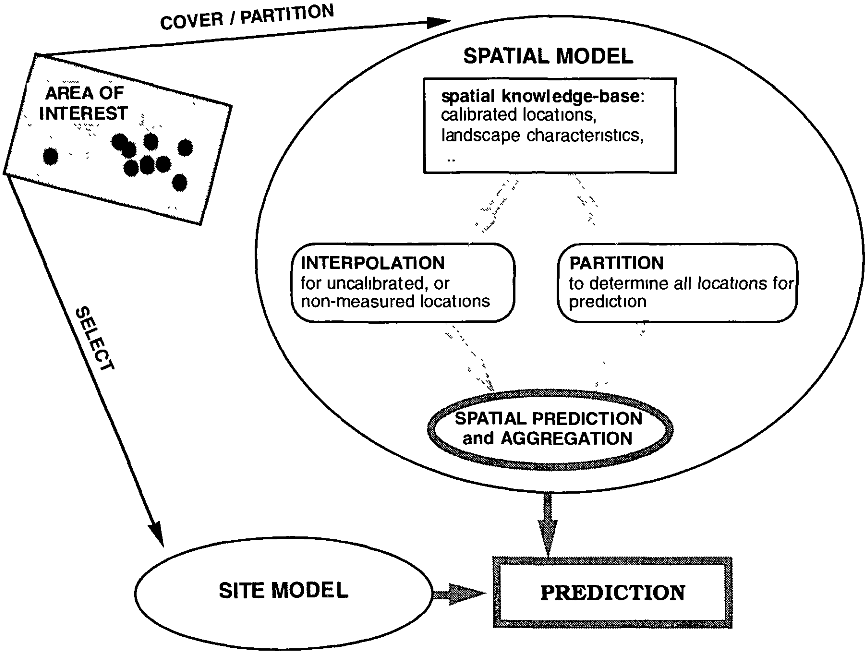
[*] This research has been supported by NSERC and the IBM Environmental Research Program.

Considerations must be given to the nature of spatial models, for example, when net primary productivity of forested ecosystems (Aber et al. 1993), acid neutralizing capacity of aquatic ecosystems (Driscoll and Van Dreason 1993), or grassland soil nutrient availability (Parton et al. 1988) is to be predicted at *regional scales*.

This paper focuses on possible choices of the spatial model for regional application of environmental process models as a function of information available about the modeled landscape. A general framework is presented for partitioning the landscape into "soft objects" (mapping or environmental units), which can be derived within a GIS by analysis of local variability. These units facilitate both loose and close coupling of GIS and the model (across data structures), and serve as common ground for sensitivity analysis of the predictions.

LINKING SITE-LEVEL MODELS AND REGIONAL MODELS

Extending site-level models to large regions frequently relies on the assumption that the region is a collection of comparable sites; thus they imply some form of *partitioning* of the region. This regional partitioning may require different strategies depending on which landscape, or environmental characteristics are predicted. The geographic context, which is quite different for climate variables, lake chemistry, or vegetation composition, is usually described within a GIS, and it will constrain the *interpolation* within the partitions (Figure 2).



COVER / PARTITION
AREA OF INTEREST
SELECT
SPATIAL MODEL
spatial knowledge-base: calibrated locations, landscape characteristics, ..
INTERPOLATION for uncalibrated, or non-measured locations
PARTITION to determine all locations for prediction
SPATIAL PREDICTION and AGGREGATION
SITE MODEL
PREDICTION

Figure 2. Expansion of the spatial model

Partitioning (which in itself may be scale-dependent) these data sets to units, that correspond to input and control requirements of the underlying processes and their models, is a function of spatial (and temporal) variability of the phenomena to be characterized, and the amount and nature of information available. In a general framework a variety of strategies can be identified to derive "mapping units" according to preference given to statistical, geometric and/or biophysical control over partitioning -- such as in geostatistical (Webster and Oliver 1990), tessellation-based (Gold and Cormack 1987), or watershed-oriented (Band et al. 1991) models, respectively.

These strategies can also be compared and classified from a data structure perspective; primarily considering whether they employ regular or irregular spatial units. For example, a grid of biomass (e.g., as derived from satellite images) does not use any "environmental" control in determining the spatial units (i.e., all sites/locations are "equal") and it is completely regular; as opposed to a biophysical and statistical knowledge-based watershed delineation (e.g., derived from a DEM) which is rather irregular; or a set of Voronoi-polygons constructed based on a set of observation sites (e.g., weather stations, or lakes) is irregular and based on geometric distribution.

Once spatial units are defined for the database, the next step is to supply data for the entire region of interest, which usually requires some form of interpolation. The more information is available about spatial structure of the phenomenon, the more reliable spatial prediction can be. The reliability of describing spatial structure, however, partly depends on the arrangement and size (distribution) of spatial units. Furthermore, efficiency and feasibility considerations may conflict with statistical optimality.

The combination of these functionality requirements for linking and interfacing GIS and environmental models results in a fairly complex system (Figure 3). There are two primary objectives for the system currently under development: (1) to keep the complete functionality of both GIS and environmental model, and (2) to provide guidelines for (or, potentially automate) the choice of the spatial model according to uncertainty analysis of the prediction.

DATABASE AND INTERFACE COMPONENTS

One of the key challenges facing the linkage between GIS and site-based models lies in considering the sensitivity of the model to errors in the parametrization derived from the geographic database. This problem is often avoided in site-based calibration, which is sometimes called "validation", because the uncertainty (e.g. variance) in input parameters is set to zero. In regional studies, however, the emphasis is on the "collection of sites", therefore, it is imperative to consider the relationship between the uncertainty of location and attributes (Csillag 1991). For example, an effort to predict the acid/base chemistry in the lakes of the Adirondack Mountains, NY, requires, among others, input of precipitation, min/max temperature, slope and aspect (for solar radiation), vegetation cover and soil characteristics information to a biogeochemical model (PnET, Aber and Federer 1992). Whenever a set of input parameters is available, the model is "ready to run", and will

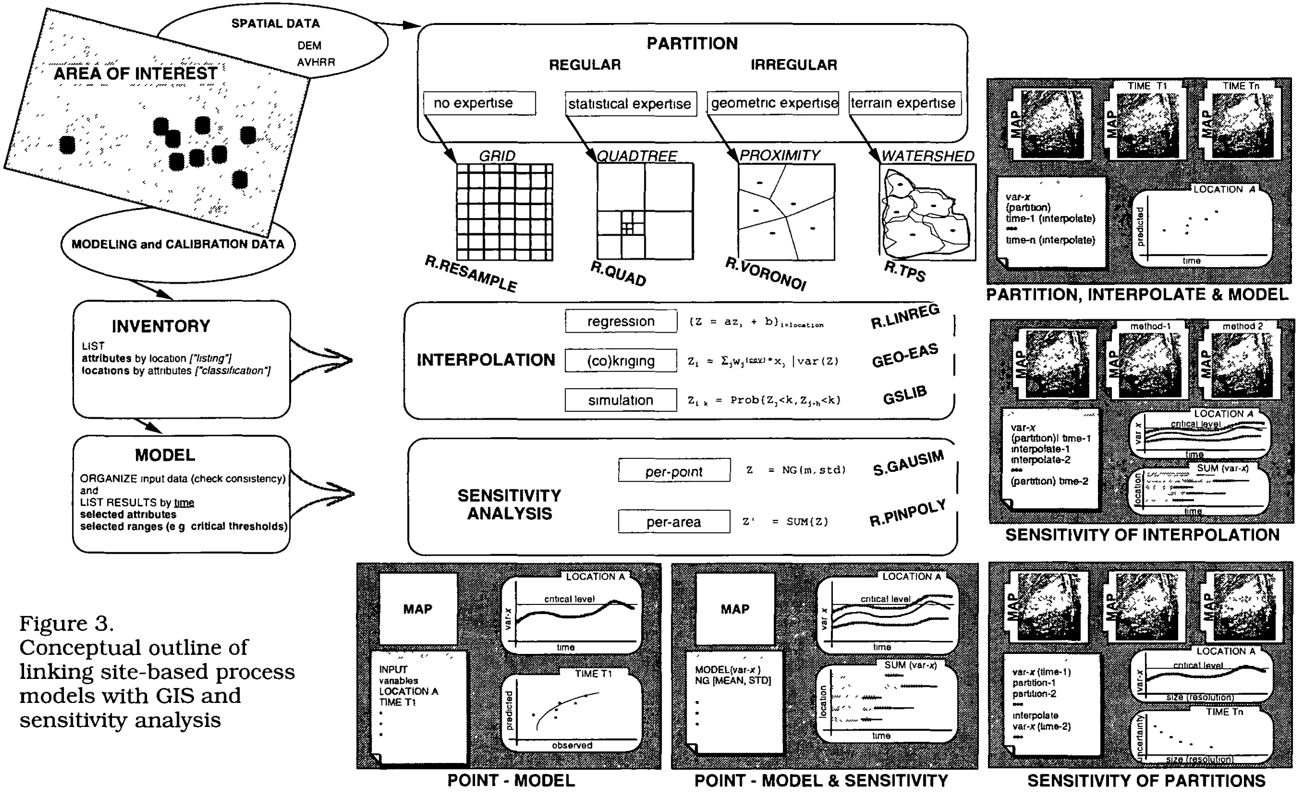


Figure 3.
Conceptual outline of linking site-based process models with GIS and sensitivity analysis

result in a time-series of a variable (e.g., the amount of dissolved N in drainage water).

Whenever collections of sites are used as input to the model, it is worthwhile and necessary to assess the sensitivity of the model to the errors in the input parameters. Since most of the process models are non-linear, their sensitivity to variation in input values is generally assessed by Monte-Carlo simulation, and is reported as "confidence intervals" around calibrated values. In spatial context then, if we can determine, or even limit, the (residual) variance of our partitions, this information can be directly related to the sensitivity of the model output. For example, if we have two (or more sites) close to each other with similar characteristics, we may be better off aggregating them into one "soft unit"; it would result in significant (50% or more) savings in computing requirements while the uncertainty in prediction may be kept below a required threshold. Since the "calibrated sites" are usually very small compared to the regions in question (e.g., a 4 ha lake watershed compared to the 3.5 million ha Adirondack ecological zone), this strategy potentially offers major compensation of attribute versus spatial accuracy. Consequently, during partition and interpolation one should control the tradeoff between the level of spatial detail (resolution) and the accuracy of prediction.

TOWARD USING MULTI-PARTITIONS IN SPATIAL ESTIMATION

As outlined on Figure 3, there are many feasible approaches toward partitioning the geographic landscape and interpolating environmental state-variables across partitions. Depending on the nature of data (e.g., a DEM versus a collection of lakes), and the level of expertise (e.g., is a detailed DEM available, or not), the partitioning strategies are classified into four groups. Since data structures in GIS and environmentally meaningful "units" do not necessarily coincide, the strategies are also grouped from a data structure perspective. Regular partitions (Figure 4) include *grids*, which do not utilize any expertise, and *quadtrees*, which can be constructed by statistical constraints on within-leaf variance (Csillag et al. 1994). Irregular partitions include *Voronoi-polygons*, which rely on geometric expertise, and *watersheds*, which are based on terrain expertise.

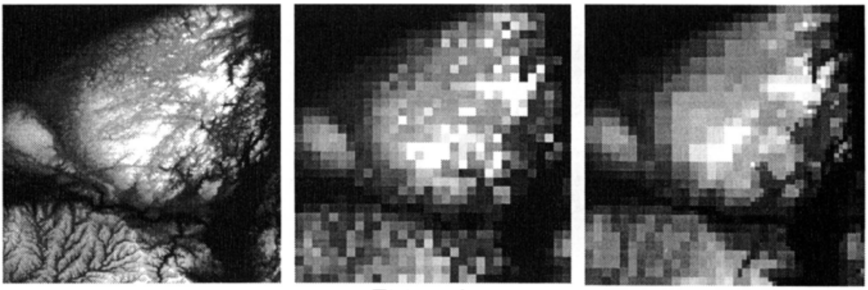


Figure 4.
Original (100m) DEM (left), regular grid aggregation to 1460 units (middle)and quadtree-tessellation with 1460 leaves (right) of the Adirondack Mts. (Note: the total variance of aggregates is reduced by using the quadtree instead of grid-based aggregation.)

Once the GIS is capable of controlling the within-unit variance, it can provide guidance to the user, who sets the accuracy thresholds for the model prediction. At the same time, it is not required that all input variables be partitioned the same way; the "soft objects" can be carried over and can be further used in interpolation within the partitions.

The combination of interpolation with (optionally limited-variance) partitions facilitates further control of uncertainty (Dungan et al. 1994, Mason et al. 1994). Since interpolation is always carried out using as much information about spatial variability as possible, for each partition the partition-mean will be more robust, and the lack of information will not spread from one partition to another. Furthermore, during interpolation the uncertainty can also be determined for each partition. The combined uncertainty associated with partitioning and interpolation can be reassessed before running the model. This is particularly important when information on one variable (e.g. elevation) is used to (co-)estimate another (e.g. acid deposition). Without partitioning the predictor variable simple estimators (e.g., non-spatial regression) lead to enormous residual variance; however, partitions can dramatically reduce it (Figure 5).

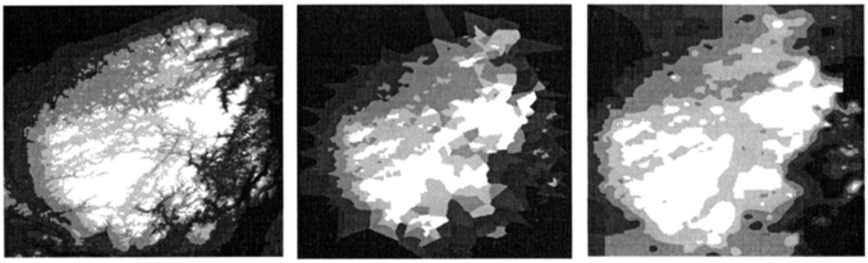


Figure 5.
Original (100m) DEM grid of the Adirondack Park (left); Voronoi-polygon mean elevations based on 1468 lakes (middle); kriged elevation based on 1468 lakes (right).

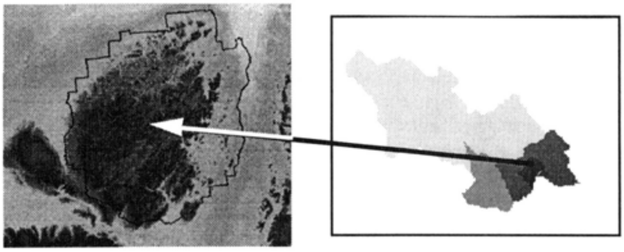


Figure 6.
PnET prediction of the amount of dissolved N on a regular (1 km) grid. (The gray-levels represent 0.4-1.4 mg/liter.) Arbutus Lake (one of the calibration sites) is marked with watersheds derived with various limits on internal heterogeneity.

CONCLUSION AND WORK IN PROGRESS

One of the major challenges in linking environmental models and GIS is to control the uncertainty related to input derived from the geographic database to drive the environmental (process) model. A general framework is proposed to combine the analytical capabilities of GIS with sensitivity analysis of a biogeochemical model (PnET) to control uncertainty in a simulation study, aiming to predict acidification in the northeast US (Figure 6). Most of the elements for interfacing GIS and environmental models by "soft objects" have been implemented in grass (see Figure 3). Current efforts are focused on automating the use of the analytical components.

REFERENCES

Aber, J.D., C. Driscoll, C.A. Federer, R. Lathorp, G. Lovett, J.M. Melilo, P.A. Steudler and J. Vogelman. 1993. A strategy for the regional analysis of the effects of physical and chemical climate change on biogeochemical processes in northeastern US forests. Ecological Modelling 67:37-47.

Aber, J.D. and C.A. Federer. 1992. A generalized, lumped-parameter model of photosynthesis, evapotranspiration and net primary productivity in temperate and boreal ecosystems. Oecologia 92: 463-474.

Band, L.E., D.L. Peterson, S.W. Running, J. Coughlan, R. Lammers, J. Dungan and R. Nemani. 1991. Ecosystem processes at the watershed level: basis for distributed simulation. Ecological Modeling 56: 171-196.

Csillag, F. 1991. Resolution revisited. AutoCarto-10 (ASPRS-ACSM, Bethesda) p. 15-29.

Csillag, F., M. Kertész and Á. Kummert. 1994. Sampling and mapping of two-dimensional lattices by stepwise hierarchical tiling based on a local measure of heterogeneity. International Journal of GIS (in press)

Driscoll, C.T. and R. Van Dreason. 1993. Seasonal and long-term temporal patterns in the chemistry of Adirondack lakes. Water, Air and Soil Pollution 67: 319-344.

Dungan, J. L., D.L. Peterson and P.J. Curran. 1994. Alternative approaches to mapping of vegetation amount. In: Michener, W.K., J.W. Brunt and S.G. Stafford (Eds.) Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor & Francis, London (in press)

Gold, C. and S. Cormack. 1987. Spatially ordered networks and topographic reconstructions. International Journal of GIS 1:137-148.

Mason, D.C., M. O'Conaill and I. McKendrick. 1994. Variable resolution block-kriging using a hierarchical spatial data structure. International Journal of GIS 8:429-450.

Parton, W.J., J.W.B. Stewart, C.V. Cole. 1988. Dynamics of C, N, P and S in grassland soils: a model. Biogeochemistry 5: 109-131.

Beyond Stevens: A revised approach to measurement for geographic information

Nicholas R. Chrisman
CHRISMAN@u.washington.edu
Department of Geography DP 10, University of Washington
Seattle, Washington 98195 USA

ABSTRACT

Measurement is commonly divided into nominal, ordinal, interval and ratio 'scales' in both geography and cartography. These scales have been accepted unquestioned from research in psychology that had a particular scientific agenda. These four scales do not cover all the kinds of measurements common in a geographic information system. The idea of a simple list of measurement scales may not serve the purpose of prescribing appropriate techniques. Informed use of tools does not depend on the nature of the numbers, but of the whole 'measurement framework', the system of objects, relationships and axioms implied by a given system of representation.

Introduction

The approach to measurement in certain social sciences is still strongly influenced by Stevens' (1946) paper in *Science*. His 'scales of measurement' form the basis for geography (Unwin 1981) and for cartography (Muehrcke 1976; Chang 1978). While measurement has continued to develop in social science research (Churchman and Ratoosh 1959; Coombs 1964; Ellis 1966; Krantz et al. 1971; Narens 1985; Suppes et al. 1989; 1990), these continuing developments have not been followed in the cartography and GIS literature.

Development of theories of measurement

The 'classical' school of measurement developed in physics and other sciences by the end of the nineteenth century. In the classical view, measurement discovered the numerical relationship between a standard object and the one measured. The property was seen as inherent in the object. This viewpoint is deeply ingrained in our language and society.

Let us take the attribute 'length'. Every entity in space can be measured by comparing its length to some other length. If we adopt a 'standard' measuring rod, we can obtain numbers (the ratio between the length of the rod and the objects measured) by a physical procedure that mimics addition – laying the rod successively along the edge. Nineteenth century physics was able to build up a rather complex model of the world with remarkably few of these fundamental properties (length, mass, electrical charge, etc.). These properties were termed 'extensive' because they extended in some way as length does in space. Other properties (like density) were built up as ratios of the extensive properties and were thus 'derived'. The laws of physics prescribed the rules for derived measures.

Extensive properties are rather restrictive, and the idea of a universal standard measuring rod in Sèvres, France is not very practical for all the properties that must be measured. Physicists began to move beyond the classical concept that the meter was an intrinsic property of one particular rod. The method of measurement became just as important as the physical standard, thus separating the object and the measurement. A twentieth-century philosophy of measurement called 'representationalism' saw numbers, not as properties inherent in an object, but as the result of relationships between measurement operations and the object.

Exclusive focus on extensive measurement in physics left almost no room for the social sciences to develop a measurement theory. The physicists could not consider phenomena like perceived loudness of sounds as a measurement, since it did not involve extensive properties like addition. Stevens' system arises from this context, a part of the movement to create a quantitative social science.

This paper will begin with a quick review of Stevens' scheme, followed with some examples of measurements in one and more dimension which require another approach. At one end, the 'ratio' level is not the highest level, nor is it so unified. At the other, the nature of categories need to be reexamined. Stevens' hierarchy also fails to treat the circumstances of multidimensional measurement.

Stevens' Scales of Measurement

Table 1: verbatim copy of Stevens' (1946, 678) Table 1

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

Stevens adopted the representationalist philosophy in a 'nominalist' form (Michell 1993), defining measurement as the 'assignment of numbers to objects according to a rule'. Table 1 reproduces Stevens' original table exactly so that his presentation is not clouded by the reinterpretations developed over the past fifty years.

The scales are defined by groups of mathematical operations that were increasingly restrictive. The focus was upon 'groups' of transformations under which the meaning of the scale remains invariant. A nominal scale could be replaced by any other scale that could be mapped one-to-one onto the original one. One subset of those operations are 'isotonic' [meaning monotonic]; a subset of these are linear, and a subset of these are simply multiplicative. From the start, the levels of measurement can be associated with an attempt to bring mathematical order into fields that do not seem to be as rigorous as physics, which had controlled the earlier developments of measurement theory (Campbell 1920; Bridgeman 1927; Michell 1993).

A key element of Table 1 is the connection between the 'scales' and 'permissible statistics'. Many textbooks on statistics for social sciences (beginning with Siegel's (1956) classic on non-parametric methods) adopted this connection between a variable and appropriate techniques. At the root, measurement is seen as a choice to represent an entity by a number, relationships were simplified to those inherent in the number system chosen. The association between numbers and methods may not be as simple as Stevens and Siegel conceived, particularly when dealing with geographic information.

Stevens tried to expunge the distinction that physicists had drawn between 'extensive' and 'derived' measurement. In Stevens' reductionist viewpoint, the properties applied to the number system, not the method by which it was generated [this had been the viewpoint of the operationists like Percy Bridgeman (1927)]. It is ironic that cartographers teach Stevens' system, with a unified 'ratio' level, then must make a distinction between those attributes permissible on choropleth maps (densities and other derived measures) and those permissible on proportional symbol maps ('extensive' measures where addition is the underlying mechanism). Using Stevens for cartography has been established for years (Muehrcke 1976; Chang 1978), and the inadequacies do not seem to be recognized.

Above Ratio

Stevens' four 'scales' are usually presented as a complete set, but they are far from exhaustive. Stevens (1959) himself proposed another scale at the same level as interval for logarithmically scaled measures. The invariance is the exponent, while the zero is fixed. This 'logarithmic-interval' scale is not cited in any of the geographic literature, though it is used for earthquake intensities and similar measurements. Following Stevens' invariance scheme to its conclusion, ratio is not the highest level of measurement. The ratio scale has one fixed point (zero) and the choice of the value of 'one' is essentially arbitrary. A higher level of measurement can be obtained if the value of one is fixed as well. Then the whole scale is predetermined or *'absolute'* (Ellis 1966) and no transformations can be made that preserve the meaning of the measurement. One example of an absolute scale is probability, where the axioms fix the meaning of zero and one simultaneously. Bayes' Law of conditional probability works because the scale is fixed between zero and one. Probability is just one example of a scale not recognized by Stevens.

Another class of geographic measurements consist of counts aggregated over some region in space. Counts are discrete, since there is no half person to count, but a count captures more mathematical structure than the other discrete levels (nominal and ordinal). Since the zero is a fixed value, counts may seem ratios, but, being tied to the discrete unit counted, it cannot be rescaled. Counts have different properties from the absolute scale, as well. Ellis (1966, p. 157) points out the difference between ratio scales and counting with the example that it is acceptable to posit a unit by saying "Let this object be 1 minch long", but it is not possible to say "Let this group contain one apple", since it either has one apple or some other number when you start. As I will demonstrate below, the process of counting depends upon the recognition of objects, a procedure tied to nominal measures.

Cyclical measures
While Stevens' levels deal with an unbounded number line, there are many measures which are bounded within a range and repeat in some cyclical manner. Angles seem to be ratio, in the sense that there is a zero and an arbitrary unit (degrees, grads or radians). However, angles repeat the cycle. The direction 359° is as far from 0° as 1° is. Any general measurement scheme needs to recognize the existence of non-linear systems. Some aspects of time, have repeating or cyclical elements. In environmental studies of all kinds, the seasons play an important role. Stevens' scheme does not allow for measurements that can be ordered spring–summer–fall–winter–spring or fall–winterspring–summer–fall. The seasonal relationships are invariant to the starting point in the cycle.

Spatial measurement raises questions about measurement scales. In the one-dimensional world of Stevens, the open-ended ratio scale seems to provide the most information content. A real number line contains the most promise for mathematical relationships. When representing a two-dimensional space, the normal scheme, attributed to Descartes, uses two orthogonal number lines. Analytical geometry can demonstrate the conversion between coordinates on two orthogonal axes and a radial system (Figure 1). These two representations are equivalent even though the units of measurement do not seem equivalent. The reason is that the two orthogonal distances create a triangle. The radial coordinates specify that same triangle using the hypotenuse and an angle. The theorems of geometry demonstrate that the two triangles are congruent, a finding that would not be apparent from their measurement scales.

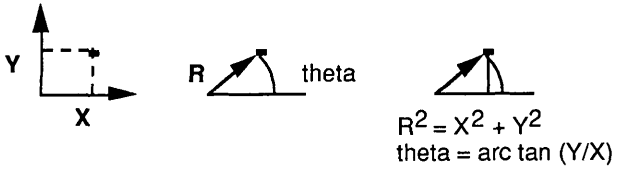


Figure 1: Cartesian axes convert to radial reference without loss

The conversion from two ratio scales to one ratio scale plus an angle is not unique to geometric constructions. Potentially infinite vectors can be simplified into lower dimensional renditions, adding great complexity to the intuitive structure propounded by Stevens. For example, all the gravitational forces from various directions can be resolved into a resultant force measured in three space. Similarly, radiant energy at various wavelengths coalesce into a particular color that can be represented in a simple conical object (Munsell's space or equivalent). The color cone is thus a "fact of nature, not a mathematical trick" (Suppes et al. 1989, p. 226). Multidimensional measurements create interactions not imagined in the simple linear world of Stevens. Since GIS is inherently multidimensional, the linear model limits our understanding concerning the interactions of measurements.

If there is any theory to GIS, it would have to start from the storage of attribute values in their spatial context. Tomlin (1983; 1990) has built a complex range of tools around the raster model of values stored for an array of point/areas. Goodchild (1987) contrasts the object view (isolated objects in a void) and the 'field' view (a z value for all pairs x,y). This commonality of thinking is strongly influenced by the storage systems that we have invented. We must remember that the slope of a surface is characterized by two numbers [gradient and aspect to use the terminology of Burrough (1986)]. We lose much understanding by the reductionism that treats these as arrays of numbers, not the vector space that the two numbers taken together portray. GIS is still stuck with scalar values as the basic conception, while vector fields and tensor fields are necessary to connect the representations to process. Higher numbers of dimensions require more complex spatial data structures (Pigot and Hazelton 1992; Worboys 1992, for example).

Rethinking nominal measurement

While Stevens' top end, the ratio scale, leads off in the direction of multidimensional measures, the bottom end is equally problematical. The nominal scale is not even considered to be a kind of measurement in many theoretical discussions (Ellis 1966; Krantz et al. 1971; Narens 1985). Social scientists had much discussion about identifying numbers, such a 'football numbers' (Lord 1953). A strictly arbitrary string assigned to each object is not really a category that groups together any individuals. Thus, it does not support Stevens' 'equality' operator. Furthermore, most numbering systems (like Lord's football team numbers) provide some kind of ordinal information about the sequence in which they are assigned or some other logic internal to the authority responsible. Identifying numbers are not really the categories that concern this discussion. Basically, a nominal category defies the logic expected of a 'scale'; order, systems of inequalities and some concatenation operations (Krantz, Luce et al. 1971, p. 4). These are the ingredients of ordinal measurement or higher.

Does this mean that nominal measurement must be abandoned? In a careful reading of measurement theory, the tide has changed from Stevens' simplification. For Stevens, the numbers determined the nature of the methods. Even some theorists who ignore nominal measures provide a basic definition that leads in another direction. Volume 1 of *Foundations of Measurement* (Krantz et al. 1971, p. 9) defines a scale as a construct of "homeomorphisms from empirical relational structures of interest into numerical relational structures that are useful". The key issue is not the invariance of some algebraic properties, but the invariance of the underlying *relationships*. Though restricted to numbers, this definition can be broadened to deal with categories. If the measurement preserves the empirical relationships and provides a useful structure for analysis, a nominal categorization fits the general requirements for a scale of measurement.

The trouble has been the oversimplification of nominal distinctions. In most treatment, Aristotle's rules are applied. Each member of a set must share common characteristics. Stevens adopts this rule by requiring that all members of a nominal group are equivalent. Certainly there is plenty of precedent for these kinds of rigid categories, but representations of the world do not always fit the simplicity of this logic. Many scientific categories, and even more of the categories of every-day life, do not live up to the purity of 'shared attribute' categories. Modern category theory (Johnson 1987; Lakoff 1987) describes at least two other alternatives; probabilistic and prototypes. While classical set theory assigns an object either as a full member or not in each category, a probability approach provides for a gradation of membership. The purest application of probability states a likelihood that an object will be discovered to belong in the classical sense. This is the approach taken to interpret soils classes by Fisher (1991). Goodchild (1992) suggests a gradation of membership that moves from the strict interpretation of probability towards a fuzzy set membership interpretation. Taken strictly, fuzzy memberships do not have to sum to one, though this normalization is often implied (Burrough 1989). This is a fracture zone for cartographers. Partial membership is often implied, and the specific model, whether probability or fuzzy sets, is rarely articulated.

On many occasions that cartographers refer to fuzzy sets or probability of membership, they really are using a 'prototype' approach to categories. The prototype refers to a 'central' example that represents the ideal form of the category. Objects are not matched attribute by attribute, but assigned to the prototype that fits most closely. There is some measure of 'distance' involved that may be mistaken for probability. The difference is that probability normalizes the separation so that every object sums to one. Distances from a prototype do not have to sum to any particular value. Some objects are just closer than others. Classification in remote sensing usually uses prototypes and distance based analysis internally before sending a sharp set out for final consumption. Supervised classification establishes the prototypes directly, then assigns each pixel to the 'closest' using the distance in spectral space. An unsupervised classification looks for the smallest set of clusters that will partition the spectral space, but pixels are also seen as more or less central to the cluster. The key trick, as always, is assigning the category names to the clusters; a process that often involves a complex interpretation of the spatial context. Lakoff and Johnson point out that the human mind tends to use prototype logic, rather than the rigid formalism of classical categories. The nature of human cognition is not the issue here, but the question of which relationships must be modeled to make the categories represent the scientific intentions.

Probability and prototype approaches to categories may dominate the real applications of geographic information, though classical categories pervade the explanation. There is a lot of literature talking about the inflexibility of categories, as if all geographic categories involve exact matches to a list of defining characteristics. Actual practice is far different. Categories are conceived in taxonomies, as a comprehensive system. All land is presumed to fit in a category, even if it is 'Not Elsewhere Classified' – a category that certainly does not share attributes amongst its members. The landscape is assigned to the closest fit category, or maybe to the most likely category. To return to measurement theory, geographers should remember that categories are not used to share formal properties along the Aristotelian scheme, but to partition a space into a nearest grouping. Geographic categories are developed to generalize.

Stevens' four scales of measurement are not the end of the story. The concept of a closed list of 'scales' arranged on a progression from simple to more complex does not cover the diversity of geographic measurement. Still, Stevens' terminology is so deeply entrenched that it may remain in use when it applies.

A larger framework for measurement

The largest difficulties with Stevens' scheme come not from the specific 'scales' of measurement, but with the overall model of the process. The levels of measurement presume a rather simple framework; the classical social science 'case' 'has' attributes. Such a model was proposed for most social sciences in the early quantitative period. The version proposed in geography was called the 'Geographical Matrix' (Berry 1964), simply a matrix with 'places' on one axis and attributes on the other. But all 'cases' or places do not have the same attributes. A more fruitful model sees measurement not in terms of properties, but in terms of relationships. Geographic information involves many more kinds of measurement. These distinctions have usually been discussed as 'data models', with an emphasis on representation. Viewed from the perspective of measurement, these old issues take on a new clarity.

This paper proposes a scheme of measurement frameworks developed from the simple taxonomy presented by Sinton (1978). Each model or framework for geographic measurement must account for each of these elements interacting in the roles of fixed, controlled and measured. In Sinton's scheme, in order to measure one component, one of the others had to be 'fixed' and one served as 'control'. At the most basic, Sinton's scheme distinguishes vector from raster because the first controls by object (attribute), while the later controls by space. This rough division provides a starting point, but it does not explain the divisions within these two approaches.

A Taxonomy for measurement: Object as control

When the attribute serves as the control, the spatial location is the subject of the measurement. While this is the common framework for a vector representation, there are large differences between the situation with an isolated category and a connected system of categories. Table 2 summarizes the distinctions.

Table 2: **Object Control Frameworks**

Isolated Objects	
Spatial Object	Single category distinguishes from void
Isoline	Regular slices of continuous variable
Connected Objects	
Network	Spatial objects connect to each other, form topology (one category possible)
Categorical Coverage	Network formed by exhaustive classification (multiple categories, forming an exhaustive set)

The simplest object control framework involves isolated objects, distinguished by a single category. While 'cartographic feature' might be apt, this framework will be termed 'spatial object'. Each point or area object is described as a geometric whole, since it will forcibly occur in isolation. The message of the object framework is: 'Here is an airport'; 'Here is another airport.' and so on. In the pure form of this framework, the only relationship is between the object and a position; there are no relationships between objects. Linear objects depart from this to some extent, creating the need for the network framework discussed below.

Isolines are formed by controlling for a specific value on a surface. Since isolines follow the contours and do not intersect, they have no topological relationships, beyond the ordering of nested contours.

In the creation of advanced GIS software, it was important to recognize that there were relationships between the objects in a database. When a coverage is formed with multiple categories, there will be topological relationships. Similar structure can be created by linear networks. The basic topology is required whether the categories form strict equivalence classes or some form of probabilistic or prototype categories. The distinction between the isolated coverages and connected coverages is not a matter of database design, but a recognition of the underlying measurement structure of the source material.

Spatial Control

Control can also come from a set of predefined spatial objects (Table 3).

Table 3: **Spatial Control Frameworks**

Point-based Control	
Center point	Systematic sampling in regular grid
Systematic unaligned	Random point chosen within cell
Area-based Control	
Extreme value	Maximum (or minimum) of values in cell
Total	Sum of quantities (eg. reflected light) in cell
Predominant type	Most common category in cell
Presence / absence	Binary result for single category
Percent cover	Amount of cell covered by single category
Precedence of types	Highest ranking category present in cell

Control by a set of points has different rules compared to control by areas. While both would be encoded in a raster representation, they must be understood differently. With a point-based control there are not too many rules. Center point provides a regular sampling of a landscape. Digital Elevation Matrices tend to use a point-based sample, though the photogrammetric equipment may actually work on a tiny area to match the photographs. Systematic unaligned is recognized in textbooks, but rarely performed.

Control by area is more common for remote sensing and other applications of grid sampling. In each cell there is some rule that has been applied to all the possible values. Some sensors add up all the reflectance in a certain band width; other gridding takes the highest or lowest value. A system that optimizes each cell by taking the most likely value for the cell may remove all traces of linear features and the minority elements. Unless these rules are known to the analyst, the information can be sorely misconstrued.

Other kinds of control
Control by object and control by space seem to be the only options, but they do not cover all the cases found in existing geographic information. The well-known *choropleth* map is an example of a **composite** framework, in that the base map is created using a categorical coverage for the set of collection units, then these objects serve as a secondary form of spatial control to tabulate the variable in question. Due to these two stages, the spatial measurements of the boundaries have little bearing on the precision of the measurement.

Triangular Irregular Networks (TIN) do not fit the scheme either. While the points may come from an isolated bunch of measurements, the TIN represents a set of relationships that cover space. The ideal TIN is constructed so that the triangles represent zones of uniform slope and aspect, within the resolution available. Thus, a TIN represents a novel class of measurement frameworks where relationships form the control, not the values of the attribute or the location.

Conclusions

The list of measurement scales developed by Stevens do not serve the purpose of providing a structure for geographic measurement. Any scheme to handle geographic measurement must deal with relationships between attribute and location, and eventually with time. A system of 'measurement frameworks' may provide a clearer focus on the design and implementation of geographic information systems. The frameworks proposed here place the measurement in the context of axioms and relationships to preserve.

Acknowledgements

The research reflected in this paper was supported by sabbatical leave from the University of Washington. Portions of this paper are extracts from a book manuscript currently under review by John Wiley.

References Cited

Berry, B. J. L. 1964: Approaches to Regional Analysis: A Synthesis. *Annals of the Association of American Geographers* **54**: 2-11.

Bridgeman, P. 1927: *The logic of modern physics.* Cambridge MA: Harvard Press.

Burrough, P. A. 1986: *Principles of Geographical Information Systems for Land Resource Assessment.* Oxford: Clarendon Press.

Burrough, P. A. 1989: Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Science* **40**: 477-492.

Campbell, N. R. 1920: *Physics: the elements.* Cambridge: Cambridge University Press.

Chang, K.-T. 1978: Measurement scales in cartography. *The American Cartographer* **5**: 57-64.

Churchman, C. W. and Ratoosh, P., eds. 1959: *Measurement: Definition and Theories.* New York: John Wiley.

Coombs, C. H. 1964: *A theory of data.* New York: John Wiley.

Ellis, B. 1966: *Basic Concepts of Measurement.* Cambridge: University Press.

Fisher, P. F. 1991: Modelling soil map-unit inclusions by Monte Carlo simulation. *International Journal of Geographical Information Systems* **5**(2): 193-208.

Goodchild, M. F. 1987: A spatial analytical perspective on geographical information systems. *International Journal of GIS* **1**(4): 327-334.

Goodchild, M. F., Guoqing, S. and Shiren, Y. 1992: Development and test of an error model for categorical data. *International Journal of Geographical Information Systems* **6**(2): 87-104.

Johnson, M. 1987: *The Body in the Mind: The Bodily Basis of Meaning, Imagination and Reasoning.* Chicago: University of Chicago Press.

Krantz, D. H., Luce, R. D., Suppes, P. and Tversky, A. 1971: *Foundations of Measurement: Volume 1: Additive and Polynomial Representations.* New York: Academic Press.

Lakoff, G. 1987: *Women, Fire and Dangerous Things,.* Chicago: University of Chicago Press.

Lord, F. M. 1953: On the statistical treatment of football numbers. *American Psychologist* **8**: 750-751.

Michell, J. 1993: The origins of the representational theory of measurement: Helmholtz, Hölder and Russell. *Studies in the History and Philosophy of Science* **24**: 185-206.

Muehrcke, P. C. 1976: Concepts of scaling from the map reader's point of view. *The American Cartographer* **3**: 123-141.

Narens, L. 1985: *Abstract measurement theory.* Cambridge MA: MIT Press.

Pigot, S. and Hazelton, B. 1992: The Fundamentals of a Topological Model for a Four-Dimensional GIS. *Proceedings of the 5th International Symposium on Spatial Data Handling* **2**: 580-591.

Siegel, S. 1956: *Nonparametric statistics.* New York: McGraw-Hill.

Sinton, D. F. 1978: The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. In *Harvard Papers on Geographic Information Systems*, ed. G. Dutton, p. 1-17. Reading MA: Addison Wesley.

Stevens, S. S. 1946: On the theory of scales of measurement. *Science* **103**: 677-680.

Stevens, S. S. 1959: Measurement, psychophysics and utility. In *Measurement: Definitions and Theories*, eds. C. W. Churchman and P. Ratoosh, p. 18-63. New York: Wiley.

Suppes, P., Krantz, D. H., Luce, R. D. and Tversky, A. 1989: *Foundations of Measurement: Volume 2.* New York: Academic Press.

Suppes, P., Krantz, D. H., Luce, R. D. and Tversky, A. 1990: *Foundations of Measurement: Volume 3.* New York: Academic Press.

Tomlin, C. D. 1983: Digital Cartographic Modeling Techniques in Environmental Planning. unpublished Ph.D., Yale University.

Tomlin, C. D. 1990: *Geographic Information Systems and Cartographic Modeling.* Englewood Cliffs NJ: Prentice Hall.

Unwin, D. 1981: *Introductory Spatial Analysis.* London: Methuen.

Worboys, M. F. 1992: A Model for Spatio-Temporal Information. *Proceedings of the 5th International Symposium on Spatial Data Handling* **2**: 602-611.

Map-Overlay within a Geographic Interaction Language

Vincent Schenkelaars*
Erasmus University / Tinbergen Institute,
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands,
Phone: +31 10-4081416, Fax: +31 10-4526177,
E-mail: Schenkelaars@cs.few.eur.nl.
and
Peter van Oosterom
TNO Physics and Electronics Laboratory,
P.O. Box 96864, 2509 JG The Hague, The Netherlands,
Phone: +31 70-3264221, Fax: +31 70-3280961,
Email: oosterom@fel.tno.nl.

The current generation of spatial SQL languages has still severe problems in specifying queries which contain complex operations. One of these complex operations is map-overlay with topological structured layers. In this paper an attempt is made to model the map-overlay operation into an object-relational query language. This query language is the formal part of a geographic interaction language. An example application of the concepts of this language is given which shows that map-overlay can be specified with relative ease. This paper also deals with the creation of topological structured layers.

1 Introduction

Previous research, by various authors [2, 6, 7], proved that the original relational model is not very suitable for a Geographic Information System (GIS). One of the main problems with the relational data model is that it lacks the geometric data types. A very similar problem occurs in the relational query language SQL. The extended relational data model solves the problem quite well, the extended SQL still has a number of remaining problems. The extended SQL approach has difficulties when dealing with more complex objects and operations, for example operations that apply to topological structured input sets. Examples of this type of operations are: shortest path in road network, network analyses, map-overlay, visibility analyses in 3D terrain, and computations of corridors. Since map-overlay is generally regarded as one of the crucial operations in a GIS, our attention is focused on modelling this operation

*This research was funded by a grant from TNO-FEL under AIO-contract number 93/399/FEL

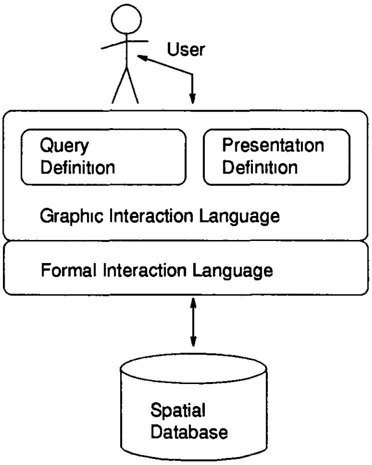


Fig. 1: Structure of the Interaction Language

in the formal part of the interaction language. We propose an *object-relational* formal syntax for this map-overlay operation.

Our final goal is to develop a general geographic interaction language. This interaction language is not meant to be yet another extended SQL. It consists of two layers. The first layer is formed by the object-relational model. The extended relational algebra, which contains all required spatial types and operations, describes the formal side of this geographic interaction language. On top of this formal language a more graphic, user-oriented language will be defined. This graphic interaction language performs two major tasks. First of all, it takes care of easy definition of queries and operations. The second task of this graphic interaction language is the definition of the presentation of query results [7]. Issues which are involved in the presentation are: shape, size, color, shading, projection, transformation, etc. All these factors can be defined interactively with the presentation language in a graphic way. Figure 1 visualizes the structure of the interaction language. More information about our interaction language can be found in [13].

2 Map Layers

An important principle in GIS is the layer concept. In this concept the geographic data is stored into layers. Each layer describes a certain aspect of the modeled real world [4]. Using map layers is a *natural* technique to organize the data from possibly different sources.

Two distinctive types of layer organizations can be identified: thematic and structured layers. The most common type is the *thematic* layer [4]. For each theme on a map, a separate layer is specified without a topological structure. In order to solve user queries containing features of more than one layer, a spatial join operation has to be performed. Elements of both layers are individually combined. The output of such combination is usually a set of object pairs. Each pair consists of an object of both

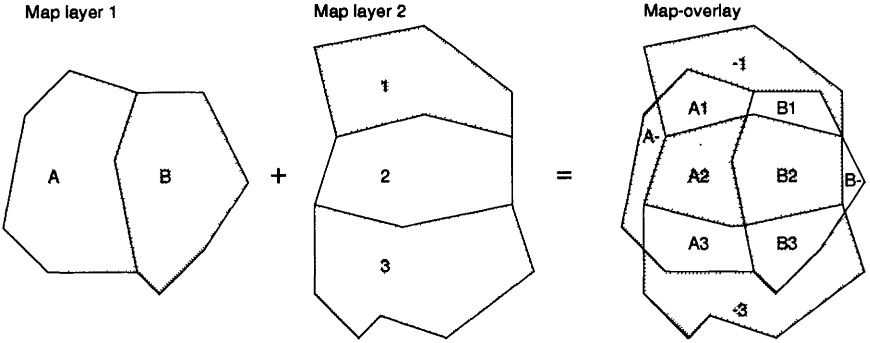


Fig. 2: Map-overlay computation and label propagation

layers.

Multiple thematic layers can be stored into one *structured* layer. These kind of layers are organized with a topological structure. The use of a topological structure removes a large amount of redundant information. Each edge is only stored once and contains references to the polygons on each side of the line. Each polygon only stores a reference to its defining edges. Although this topological structure stores map information more efficient, solving queries which contain features of more than one structured layer becomes more awkward. The combination of two of these layers is not performed separately on each element of the layers but on the layer as a whole. The result of this combination is a new structured layer.

This paper concentrates on the latter type of layers: the topological structured layer. It may be clear that the combination of two ore more of these structured layers is far more complex compared to the spatial join of two relatively simple thematic layers. We focus on the specification of the map-overlay process in a formal query language.

3 Map Overlay

The input of a map-overlay operation consists of two or more topological structured layers of edges only, in case of a linear network, or edges and faces, in case of a polygonal map. The output of the process is a new topological structured layer in which the attribute values of the new edges and faces are based on the attribute values of the elements of the input layers. Figure 2 shows the map-overlay process.

Note that the spatial domains of the different input layers do not have to match exactly. The map-overlay is usually computed in three logical phases [8]. The first step is performed at the metric level and computes all *intersections* between the edges (line segments) from the different layers. Followed by a reconstruction of the *topology* and assignment of labels or *attribute* values in the next two steps [18]. Several algorithms have been developed for computing the map-overlay:

- brute force method;
- plane-sweep method [1] (including several variants);

- uniform grid method [9];
- z-order-based method [11];
- R-tree-based method [15].

A problem related to map-overlay is the introduction of sliver polygons. Solutions for this have been presented in [3, 19]. Certain parameters have to be specified (e.g. minimum face area) in the map-overlay process.

When inserting a new line into a topological layer, several situations can occur: touch, cross, overlap, or disjoint from the lines already present in the structure. Geometric computations are used to determine the actual situation. Because computers have only finite precision floating-point arithmetic [10, 12], *epsilon*-distance computations have to be used. The actual epsilon value has to be given (by default) for every topological layer.

4 Extended Relational Approach

This section describes an attempt to model the map-overlay process directly into a relational query language. Since we use the extensible relational database management system (eDBMS) Postgres [14] in our GIS research, we used its query language Postquel as a framework. It is clear that the modeling can be done in other eDBMS's in a very similar way.

First, at least two layers have to be created. This can be done by creating a relation for each layer. The code fragment below shows the statements. The first line creates a parcel layer with owner and value information. The other layer contains soil information. Note that the layers contain explicit polygons and there is no topological structure.

Code Fragment 1

```
create layer1 (name=text, shape=POLYGON2, owner=text, value=int4)

create layer2 (id=int4, soil=text, location=POLYGON2)
```

With this layer definition, it is possible to specify the map-overlay operation as is shown in the next code fragment. The shape of the objects of the new layer are defined as the intersection between an object from `layer1` and an object from `layer2`. The attribute value `newval` is calculated from attribute value `soil` of `layer2`, `value` of `layer1`, and the area of the new polygon. Note that some functions are used in the calculation of the value of `newval`.

Code Fragment 2

```
retrieve into layer3 (oldname=layer1.name,
    newshape=Intersect(layer1.shape, layer2.location),
    newval=SoilRef(layer2.soil)*layer1.value*AreaPgn2(newshape))
```

We now have specified map-overlay in relational terms. The number of input layers does not have to be limited to two, but can be increased to any number. However, for each number of layers to be combined an `Intersect` function with the appropriate number of parameters has to be defined. Other attributes can be specified at will.

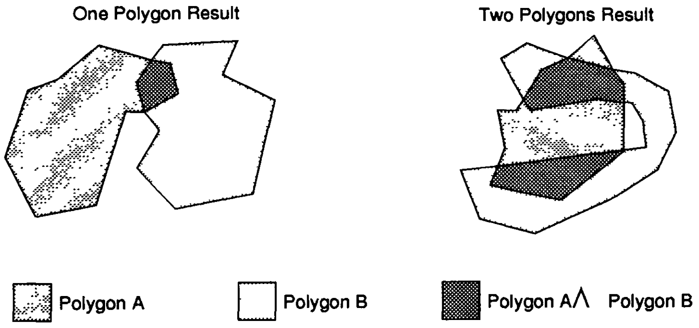


Fig. 3: Intersection of two polygons

They can be copies of old attributes or functions applied to attributes of any layer. This is a very flexible and elegant way to specify the propagated attribute values in the map-overlay operation.

Although the relational approach is simple and straight forward, this 'solution' has several severe drawbacks:

a. It is based on explicit polygons, not on a topological data model. As is stated before, there are clear advantages in topologically structuring the layers. It is possible to create topological structured layers in a relational model, but specifying the map-overlay process becomes impossible. Since one does not longer have explicit polygons, it is necessary to execute a query for each polygon to get its defining edges. This construction does not fit in the relational algebra.

b. It assumes that each pair of intersection polygons result in at most one polygon, in general this is not the case. Any number of polygons can be returned as result of the intersection; see righthand side of figure 3. Therefore, a new type of polygon with disconnected parts must be defined. It is clear that this is not desirable.

c. This method does not work for the map-overlay of two linear network layers. Intersections of lines return in general points and not lines. One could define the intersection function in such way that it returns the resulting line parts, but then again a new polyline type with separated parts must be defined.

d. It is not possible to use an efficient plane-sweep algorithm, because the intersection function operates on pairs of individual polygons. This is not efficient.

The common cause behind these problems is that map-overlay should be applied to complete layers and not to the individual polygon instances making up the whole layer structure. In order to be able to do this, the concept of complex objects is needed. The topological structured layer is considered to be a complex object with its own intersection operation. So there seems to be a need to move away from the pure relational model and adopt some object-oriented concepts. The next section describes what is needed to specify the map-overlay operation in an object-relational model.

5 Object Relational Approach

This section described our new approach to model the map-overlay process in a more object oriented approach. The next subsection describes the way to define a topological layer structure. Section 5.2 describes the actual creation of the topological layer, while section 5.3 deals with the final map-overlay operation specification.

5.1 Topological Layer Definition

To solve the problems associated with the extended relational approach, we need to create a complex object `layers`. A fully topological structured layer contains nodes, edges, and faces. To make the model simple, we assume in our examples that the nodes are stored in the edges, and that only faces have labels (attributes). Before the topological structure of a layer can be created, we need to define some prototypes. This is done in the first two lines of the next code fragment. These prototypes are essentially the same as ordinary relations, but they can not contain any data. They form the framework links between faces and edges.

Code Fragment 3

```
define prototype faces (id=oid, boundary=edges.id[])

define prototype edges (id=oid, line=POLYLINE2, left=faces.id,
    right=faces.id)

create layers (layer_id=unique text, boundaries=prototype edges,
    areas=prototype faces)

define topology _layers_topol on layers using polygonal (boundaries,
    areas)

define index _layers_bdy on layers.boundaries using
    rtree (line polyline_ops)
```

The prototypes `faces` and `edges` define the basic attributes of a face and an edge respectively. Both have an attribute `id` of type `oid`[1], which contains a unique identifier for each edge and face. The `faces` prototype has an attribute `boundary` which is a variable length array (indicated by the square brackets) of `id` values of the prototype edges. This concept of references forms the actual link between the two prototypes. Note that the `faces` prototype has a forward reference to the `edges` prototype. This forward referencing is only allowed in the definition of the prototypes. These have to be defined before the actual `layers` relation can be created.

The creation of the `layers` relation is quite straight forward. The layer has a unique identifying name and *references* to the edges and the faces member relations. These two relations are not directly visible for the user. The only way a user can retrieve individual edges or faces is through the layers relation similar to the object-oriented concept of class and member class. The edges and faces can therefore be in one layer only.

[1] oid stands for object identification

The next line in code fragment 3 registers the topological structure in the DBMS. This line will enhance the `append`, `replace`, and `delete` statements whenever these statements deal with a topological structured layer. This enhancement of these statements is similar to the enhancement of the same statements when defining an index on a relation attribute. The syntax is also similar to the index definition syntax in Postgres as can be seen in the last line of this code fragment. The two parameters between the parenthesis relate to the attributes of the layer which contain the edges and the faces. A optional third parameter could be the epsilon value; see section 3 The keyword `polygonal` denotes that the topological structure contains edges and faces. The possible topological structures are:

- `full_polyhedral`: A three dimensional topological structure with nodes, edges, faces, solids;
- `polyhedral`: A three dimensional topological structure with edges, faces, solids;
- `full_polygonal`: A two dimensional topological structure with nodes, edges, and faces;
- `polygonal`: A two dimensional topological structure with edges, and faces;
- `network`: A two dimensional topological structure with nodes and edges.

5.2 Topological Layer Creation

The definition of the framework structure of the layer is now complete. It is good to realize that the topological structured layers itself have still to be created. The following code fragment shows how this could be specified in the formal database language.

Code Fragment 4

```
define prototype faces2 (name=text, owner=text, value=int4)
    inherit faces

append layers (layer_id="parcels", areas=prototype faces2,
    boundaries=prototype edges)

append l1.boundaries (polyline="(....)"::POLYLINE2)
    from l1 in layers where l1.layer_id="parcels"

replace l1.areas (name="....", value="....", owner="....")
    from l1 in layers
    where PointInPolygon ("(x,y)"::POINT2, current))
          and l1.layer_id="parcels"
```

The first line of code fragment 4 defines the additional attributes of the areas in this layer. The `inherit` keyword allows `faces2` to be used wherever `faces` can be used. In this way the generic topological structured layer can be extended to have arbitrary number of attributes.

The next line creates the new empty layer. Only the `layer_id` attribute value has to be provided. The append statement checks the type of `faces2` against the original definition of `faces`. Note that the original definition of `edges` is used as the `boundaries` attribute. If one would need to have boundaries with additional attributes, a new prototype has to be defined in a similar way as is done for faces2.

Now the layer is defined and ready to receive its defining edges. The third line in code fragment 4 is executed for every edge in the layer. It is an append into the member relation of the layer where the edges are stored. The location of this relation is stored in the edges attribute of the layer. This append has some extra functionality due to the definition of the topology structure. Whenever an edge El intersects an edge E2 already in the layer, both edges are split by the append operation. Edge E2 is removed from the layer, the resulting parts of the splitted edge are appended to the layer one by one. Note that in case of edges with additional attributes, both parts of the splitted edge get the original attribute values. While appending edges, faces are being formed. Those faces are stored in the faces relation of the layer in a similar way. After all the edges are added to the layer we have a layer with all edges and faces, but the faces have no labels yet; the additional attributes have to be given values. This is done in the last line of code fragment 4. Each area is checked whether it contains the location. The keyword current refers to the area which is checked at that time.

Note that the described process above creates a topological layer from scratch. When one has a data set which contains topological references as is for example the case in DIGEST data [5], this process can be simplified by defining the topology after the areas and lines are inserted in layer. When inserting the edges and faces in this case, no extra functionality is needed in the append. When the topology is defined at the end, it provides the additional append functionality at that time. The definition of the topology triggers also a topology check process on the already inserted data. This to ensure that the topology structure of the data is valid. When new edges are added to the layer, the append statement will take care of the topology maintenance. It also contains an object id manager. This manager will keep object id's unique and insures referential integrity of the topological structured layer.

5.3 Map Overlay

Once the layers have been created, they can be manipulated as layer objects; that is as *complex objects*. Since complex objects can be handled in the same way as ordinary objects, one can write an intersection function which is executed on the layer as a single object. The intersection of two layers is exactly what map-overlay is. The next code fragment shows how this can be specified in the query language. We assume that the map-overlay function is registered in the database.

Code Fragment 5

```
append layers (layer_id="combined layer")

retrieve (count=overlay(l1,l2,new_layer,"FaceAttrSpecStr",
    epsilon, sliver))
    from l1, l2, new_layer in layers
    where l1.layer_id="parcels" and l2.layer_id="soil" and
          new_layer.layer_id="combined layer"
```

Before the map-overlay can be executed, we need to provide a structured layer in which the map-overlay result can be stored. This is done in the first line of code fragment 5. Note that we do not need to initialize the areas and boundaries sets. This initialization is done in the `overlay` function. Now we are ready to compute the resulting layer. The map-overlay is now nothing more than a retrieve using a user-defined function. Since a function cannot be called without having a variable to receive its result, some extra information can be returned from the overlay function. In this case the number of areas in the new layer is returned. The `FaceAttrSpecStr` contains for each attribute of the areas in the new layer a specification string. Each part of the specification string contains the name of the attribute and the expression which specifies the value of the attribute. Each expression is similar to the expression in code fragment 2. This total string is parsed by the overlay function.

The result of the map-overlay operation is stored in a new layer. The coordinates of all the edges in the new layer are redundantly stored. However, since some layers may be regarded as temporary layers, a user has two option to remove the redundancy. The user can either remove the new layer after studying the result, or one or more of the original layers is no longer useful and can therefore be removed.

6 Conclusion & Further Research

We have presented a way to model the important map-overlay concept into a formal query language. The following components are added to the Postquel language: `prototype`, `define topology`, and special `append`, `replace`, and `delete` statements. Besides these modifications in the DBMS (backend), also modifications to the geographic user interface (frontend) have to be made in order to visualize the topologically structured data [17].

An implementation of the suggested extensions in Postgres will be non-trivial, but partly comparable to adding a new access method [16]. An alternative is to use a true OODBMS as platform on which this spatial eRDBMS will implemented.

In this paper, we focussed on one specific type of topology, but as indicated, other types of topology structures can be supported as well. This will make the spatial eRDBMS a good generic basis for GIS-application, and also for other spatial applications; e.g. CAD systems.

References

[1] Jon L. Bentley and Thomas A. Ottmann. Algorithms for reporting and counting geometric intersections. *IEEE Transactions on Computers*, C-28(9):643–647, September 1979.

[2] M.S. Bundock. SQL-SX: Spatial extended SQL - becoming a reality. In *Proceedings of EGIS '91*, Brussels, 1991. EGIS Foundation.

[3] Nicholas R. Chrisman. Epsilon filtering – a technique for automated scale changing. In *Technical Papers of the 43rd Annual Meeting of the American Congress on Surveying and Mapping*, pages 322–331, March 1983.

[4] Sylvia de Hoop, Peter van Oosterom, and Martien Molenaar. Topological querying of multiple map layers. In *COSIT'93, Elba Island, Italy*, pages 139–157, Berlin, September 1993. Springer-Verlag.

[5] DGIWG. DIGEST - digital geographic information - exchange standards - edition 1.1. Technical report, Defence Mapping Agency, USA, Digital Geographic Information Working Group, October 1992.

[6] Max J. Egenhofer. An extended SQL syntax to treat spatial objects. In *Proceedings of the 2nd International Seminar on Trends and Conce rns of Spatial Sciences*, New Brunswick, 1987.

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

[8] Andrew U. Frank. Overlay processing in spatial information systems. In *Auto-Carto 8*, pages 12–31, 1987.

[9] Wm. Randolph Franklin, Chandrasekhar Narayanaswami, Mohan Kankanhall ans David Sun, Meng-Chu Zhou, and Peter Y. P. Wu. Uniform grids: A technique for intersection detection on serial and parallel machines. In *Auto-Carto 9*, pages 100–109, April 1989.

[10] David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1):5–48, March 1991.

[11] Jack Orenstein. An algorithm for computing the overlay of k-dimensional spaces. In *Advances in Spatial Databases, 2nd Symposium, SSD'91, Zürich*, pages 381–400, Berlin, August 1991. Springer-Verlag.

[12] David Pullar. Spatial overlay with inexact numerical data. In *Auto-Carto 10*, pages 313–329, March 1991.

[13] Vincent Schenkelaars. Query classification, a first step towards a geographical interaction language. In *Proceedings of Advanced Geographical Data Modeling, AGDM'94*, Delft, September 1994.

[14] Michael Stonebraker and Lawrence A. Rowe. The design of Postgres. *ACM SIGMOD*, 15(2):340–355, 1986.

[15] Peter van Oosterom. An R-tree based map-overlay algorithm. In *Proceedings EGIS/MARI'94: Fifth European Conference on Geographical Information Systems*, pages 318–327. EGIS Foundation, March 1994.

[16] Peter van Oosterom and Vincent Schenkelaars. Design and implementation of a multiscale GIS. In *Proceedings EGIS'93: Fourth European Conference on Geographical Information Systems*, pages 712–722. EGIS Foundation, March 1993.

[17] Peter van Oosterom and Tom Vijlbrief. Integrating complex spatial analysis functions in an extensible GIS. In *Proceedings of the 6th International Symposium on Spatial Data Handling, Edinburgh, Scotland*, pages 277–296, September 1994.

[18] Jan W. van Roessel. Attribute propagation and line segment classification in planesweep overlay. In *4th International Symposium on Spatial Data Handling, Zürich*, pages 127–140, Columbus, OH, July 1990. International Geographical Union IGU.

[19] Guangyu Zhang and John Tulip. An algorithm for the avoidance of sliver polygons and clusters of points in spatial overlays. In *4th International Symposium on Spatial Data Handling, Zürich*, pages 141–150, Columbus, OH, July 1990. International Geographical Union IGU.

VISUALIZATION SUPPORT FOR FUZZY SPATIAL ANALYSIS

Bin Jiang*, Ferjan Ormeling**, and Wolfgang Kainz*

*Department of Geoinformatics
International Institute for Aerospace Survey and Earth Sciences (ITC)
P. O. Box 6, 7500 AA Enschede, The Netherlands
Tel: +31 53 874 449, Fax: +31 53 874 335
Email: {bin, kainz@itc.nl}

**Department of Cartography, Universiteit Utrecht
P. O. Box 80115, 3508 TC Utrecht, The Netherlands
Tel: +31 30 531373, Fax: +3130 540604
Email: F.Ormeling@frw.ruu.nl

ABSTRACT

Visualization techniques benefit fuzzy spatial analysis in at least two aspects. One is in the field of exploratory analysis, and another is in the representation of uncertainty. This paper intends to discuss the first issue.

Fuzzy spatial analysis may be distinguished from conventional analysis in that the former is a form of concept analysis which is closer to natural language and the latter in most cases refers to numerical processing. Due to the fuzzy nature of the analysis approach, suitable visualization techniques to support the analysis process are highly needed. In this paper, the fundamentals of fuzzy spatial analysis are outlined and consecutively, the visualization tools supporting the exploration process are focused on. Finally, an approach towards a complete system framework for exploration is presented using some advanced techniques such as the object-oriented system building approach, Graphical User Interfaces (GUI) and hypertext techniques.

1 INTRODUCTION

There have been quite a lot of discussions on as well as practical contributions to exploratory spatial statistical techniques (Goodchild 1987, Openshaw 1990). It is gradually forming a new field in between Spatial Data Analysis and Visualization, and it is referred to now either as Exploratory Data Analysis (EDA) by Tukey (1977), or Exploratory Geographical Analysis (EGA) by Openshaw (1990). These two terms are based on the same assumption, that is that spatial data analysis (SDA) techniques need to be developed further in order to provide tools that allow us to discover patterns or structures unknown to us. It is only in a process of efficient exploration that potential patterns or anomalies can be recognized by analysts. Much research has been carried out in order to incorporate SDA into GIS (Tang 1992, Xia 1993). However, these EDA techniques are restricted to statistical analysis. We will attempt to develop the issue by emphasizing the potential of fuzzy spatial analysis.

The incorporation of linguistic notions into conventional spatial analysis leads to fuzzy spatial analysis (Leung 1984, Jiang, Kainz and Muller 1995, Jiang and Kainz 1994), which is closer to the way of human thinking. The fuzziness inherent in linguistic notions requires an effective visualization tool to support this exploratory analysis. In this connection, the following features of visualization would be needed extensively: 1) direct manipulation; 2) multiple perspective treatments; 3) real-time operation modes; 4) flexible online help.

"In the context of scientific visualization, 'to visualize' refers specifically to using visual tools (usually computer graphics) to help scientists/analysts explore data and develop insights. Emphasis is on purposeful exploration, search for pattern, and development of questions and hypotheses" (MacEachren and Monmonier 1992).

Visualization benefits fuzzy spatial analysis for purposes of discovery and better understanding of structures and patterns in at least two aspects. One is the dynamic exploratory process in which patterns and anomalies are easily identified. Another is the representation of uncertainty in an efficient way.

The purpose of the research presented in this paper is to develop a methodology for visualization tools to support fuzzy spatial analysis. To accomplish this goal, first the distinction between conventional spatial analyses and fuzzy analysis is discussed in section 2. The following section 3 shows the directions in which fuzzy spatial analysis operations need to be developed. Section 4 deals with exploratory tools. A complete system framework with an object-oriented approach will be discussed in section 5. The paper concludes with possible future research directions. It should be noted that the work is somewhat different from effort contributed to the visualization of data quality with fuzzy set theory (Van der Wel et al. 1994), although some principles and results obtained in this discussion can be equally applied.

2 CONVENTIONAL SPATIAL ANALYSIS VS FUZZY SPATIAL ANALYSIS

Conventional spatial analysis is a form of numerical analysis while fuzzy spatial analysis is a form of higher level analysis related to Artificial Intelligence (AI). What differentiates fuzzy spatial analysis from the conventional kind are the fundamental aspects of its exploration modes. These will be outlined first.

Numerical Analysis versus Concept Analysis

By stating that conventional spatial analysis is a form of numerical analysis while fuzzy spatial analysis is a form of concept analysis, we do by no means imply that concept analysis is the kind of qualitative analysis adhered to before the quantitative revolution in geography. The object of numerical analysis is data, and what is represented with graphics is also data. Due to the fuzzy nature of human thinking, the numerical analysis approach may mislead. A straightforward example is that two objects slightly different in value when on different sides of a crisp boundary value may be divided into two different classes. But instead of finding for instance an area with a slope of less than 30 degrees, in most cases, we would be interested in locating the area with relatively *gentle* slopes. By an efficient color scale, a pattern with the characteristic of '*gentle* slope' can be obtained. Taking a step further, the areas with *very gentle* or with *less gentle* slopes could be obtained as well.

Layer versus Sublayer

A layer is a subset of a multiple-element map, produced for spatial analysis purposes in the context of GIS. A sublayer is derived from a layer for the purpose of fuzzy spatial analysis. With the advent of digital technology, maps have been split up into sets of layers or coverages in GIS databases, and each layer corresponds to a type of thematic elements, like vegetation or transportation, etc. Layers have been playing a great role in spatial analysis, in answering series of queries relevant to reality. But they have limitations in answering and describing the fuzzy side of reality, and for that purpose sublayers are needed, which can be derived from the corresponding layers.

As we have already shown, there is no difference in nature between maps and layers except for their contents. However, sublayers are different from layers in the basic nature of what they represent. What the sublayers represent is uncertainty about one single concept indicated by a linguistic notion. Through intuitive representation, sublayers offer an analyst a more direct perception of the fuzzy side of the real world.

Probability versus (Un)certainty

Fuzzy spatial analysis is concerned with uncertainty or fuzziness and not with probability, although probability could also exist in some fuzzy linguistic notions, like *likely* and *most likely* etc. Conventional analysis is mostly based on probability theory, as what the layer represents still refers to the data or class in question. What exists in concept communication is primarily uncertainty, although it might also entail a probability issue. There is a well-known assertion in the field of mathematics that randomness is not equal to fuzziness. In the statement "show me the *gentle* slope area", what *gentle* implies has something to do with uncertainty about the concept of *gentle,* but rarely with probability.

Statistical Graphs versus Membergrams

Conventional exploratory analysis is based on statistical graphs such as histograms, scatterplots and scatterplot matrices. Fuzzy spatial analysis, as will be shown below, focuses on a series of sublayers that can be represented by color scales. Through the operations applied to a single sublayer or a set of sublayers, the analyst gets deep insight into fuzzy aspects of the real world.

(Un)certainty values ranging from zero to one can be represented graphically in a form similar to the statistical graphs, but because of differences in the nature of the contents, we refer to them as membergrams. Mathematically, they provide an intuitive representation and serve the task of visualizing the distribution of uncertainty. With membergrams, we can easily explore structures like fuzzy patterns, correlation and co-occurrences.

3 EXAMPLES OF EXPLORATORY OPERATIONS OF FUZZY SPATIAL ANALYSIS

Fuzzy spatial analysis changes our perspective from numerical analysis to concept analysis which is closer to human thinking in natural languages. The basic exploratory operations of fuzzy spatial analysis which have been discussed by Jiang and Kainz (1994) can be designed as a basic toolbox.

Operation on Primary Terms

Primary terms like *gentle, moderate,* and *steep,* are fundamental for fuzzy spatial analysis. Each of the primary terms usually constitutes a sublayer which appears in one window. Compared to layers in non-fuzzy spatial analysis, sublayers can perfectly describe these objects which belong not only to this level of the term but also to a higher level term. According to experience of domain analysts, for example, if 40 degrees is the ideal value for moderate slopes, then a value of 50 or 60 is most likely to be a vague one, which can be either regarded as moderate or as a steep slope. For such an operation, two components of a Graphical User Interface (GUI) are used to do the exploratory analysis. One is a dialog to adjust the shape of the membership function, and another is the child window required for a view of the sublayer. Fig. 1 is an example of this operation.

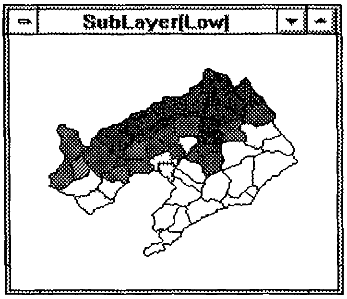


Fig. 1: Operation on primary term *low*

Operation with Negation

Negation is similar to the expression in daily communication, that what we want to know is the opposite of a certain property. For instance, instead of gentle slopes, we may want to get a pattern of *non-gentle* slopes. This would be worked out through an operator assigned by an icon which is arranged in a toolbox or is listed in the margin of the work area. Fig. 2 shows a result of the operation applied to *low* .

Operation with Hedges

Hedge operations are usually applied to the corresponding primary term when the degree of an expression is to be changed (Fig. 3). Instead of being interested in steep slopes, the analyst may wish to obtain *very steep* slopes. It can also be designed as an operator that includes reinforcing and weakening, just like a negation.

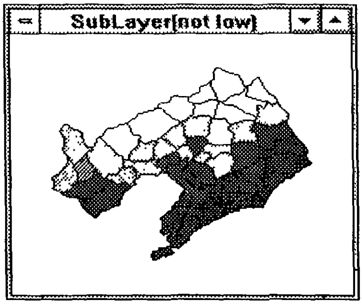


Fig. 2: Operation with negation *not low*

Operation on Connectives

Contrary to operations with hedges or negation, connective operations have at least two primary terms involved simultaneously. It is the most complex one among the set of operations, so its exploratory function is relatively difficult to perform. Sometimes, what we are interested in is neither gentle nor moderate, but *gentle and moderate*, or *gentle or moderate*. Another important application of connectives is fuzzy overlay in which combinations can be performed by a variety of operators (Jiang, Kainz and Muller 1995).

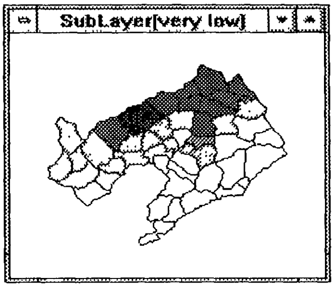


Fig. 3: Operation with hedges *very low*

4 EXPLORATORY TOOLS

Similar to statistical exploratory analysis, fuzzy exploratory analysis also has a set of tools to assist the analytical process. With these tools, the system can provide a better understanding about fuzzy aspects of reality.

Membergrams

Membergrams are similar in form to statistic graphs used for statistic exploratory analysis. However, what the membergrams indicate is the uncertainty (or possibility) distribution about assigning a set of objects to a certain attribute. They provide an intuitive visualization means to show the membership function. The following are a group of potential structures to be explored. They often serve the task of controller in exploratory analysis.

Fuzzy Pattern: this occurs in a single sublayer which has been derived from the layer and is represented with tints of gray. In this scheme, dark grey indicates a relatively full membership, and light grey indicates an intermediate membership or non-membership. It is much like the style of region-pattern masks introduced by Monmonier (1990). Fig. 1, Fig. 2 and Fig. 3 are examples of the fuzzy pattern representing regions in which dark areas have a higher certainty value than light areas.

Correlation: another question the analyst may be interested in is the correlation of two concepts in correspondence with different layers, say low pollution index and high forest coverage percentage. Two juxtaposed sublayers for simultaneous viewing will provide a rough answer about the relationships. The fact is that the visual comparison will only provide a superficial inspection; a deeper insight can be obtained through the precise calculation of the correlation coefficients. The exact correlation values should be shown in another window for inspection. To enhance the potential intuitive analysis, a suitable color scale could be used to represent the magnitude of coefficient values.

Co-occurrence: One of the important applications of overlay is to find the locations which satisfy certain conditions. The best solution is the provision of color mixture schemes in which each color represents one single condition. Intuitive representation can enhance the possibility for memorizing patterns. If there are few sublayers involved in the overlay operation, an analyst always wishes to obtain a general idea at a quick glance. In this connection, different color combinations (Olson 1987) provide alternative solutions.

Color-based Operation

Color has great potential in representing uncertainty just as observed by Zadeh (1973) who stated that "If we regard the color of an object as a variable, then its values, red, blue, yellow, green, etc., may be interpreted as labels of fuzzy subsets of a universe of objects. In this sense, the attribute color is a fuzzy variable, that is, a variable whose values are labels of fuzzy sets. It is important to note that the characterization of a value of the variable color by a natural label such as red is much less precise than the numerical value of the wavelength of a particular color". A fuzzy color system based on the recognition of identical hue as the same property or attribute may be constructed as a hue-layered color solid.

Against this background it can be surmised that, if a primary notion, for example 'low Pollution Index (PI)', is identified by a hue as red, a different degree of red with different lightness or saturation could then be used to represent (un)certainty about the concept 'low PI'. Thus based on user's preferences, certain color hues can be assigned to a primary notion. Once the provisional assignment of a given hue to a primary notion is established, the uncertainty can be represented by color, and further explorations can be seen as operations applied on color. We will refer to them as color-based operations. One of the important advantages is that this provides a real-time analytical tool.

Color is a useful tool to represent uncertainty about linguistic concepts. Fuzzy color specifications have been designed on a PC platform with 16 million colors. With the decreasing costs of personal computers, the cost of color was also reduced greatly. Once the uncertainty is visualized by color, the exploratory analysis can proceed in a real time manner.

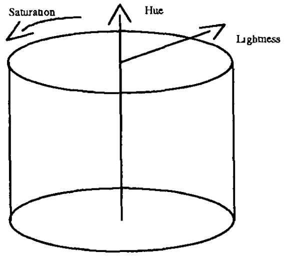


Fig. 4: The fuzzy color system used for representation of uncertainty

Color Solid: A color system used for the representation of uncertainty has been designed, in which, according to three psychological variables, all colors are rearranged to construct a new intuition-based color solid. In this model, full ranges of colors are organized to layers in terms of different hues (Fig. 4). The basic idea of the model originates from the metaphor that unitary hue is usually referred to as a homogeneous feature, as vegetation is represented by green, water by blue and so on. In addition, an uncertainty specification for each color facilitates the representation of uncertainty in fuzzy spatial analysis.

Uncertainty Specification: Along with the color solid, uncertainty is specified in equations (1), (2) and (3).

$$\mu_C(l)|_{Lum} = \frac{(1-\upsilon)^{\lambda-1} l^{\lambda}}{(1-\upsilon)^{\lambda-1} l^{\lambda} + \upsilon^{\lambda-1}(120-l)} \qquad [0,\ 120] \qquad (1)$$

$$\mu_C(l)|_{Lum} = \frac{(1-\upsilon)^{\lambda-1}(240-l)^{\lambda}}{(1-\upsilon)^{\lambda-1}(240-l)^{\lambda} + \upsilon^{\lambda-1}(l-120)} \qquad [120, 240] \qquad (2)$$

$$\mu_C(s)|_{Sat} = \frac{(1-\upsilon)^{\lambda-1}(240-s)^{\lambda}}{(1-\upsilon)^{\lambda-1}(240-s)^{\lambda} + \upsilon^{\lambda-1} s} \qquad [0,\ 240] \qquad (3)$$

The two parameters υ and λ serve for the purpose of visual equality in color scales. That is, the users can adjust the magnitude of the two parameters up to a certain level in which color scales satisfy the request of visual equality. On the other hand, the illustrations of Fig. 5 and Fig. 6 provide a good visualization method of how color could be perceived in conjunction with representation of uncertainty.

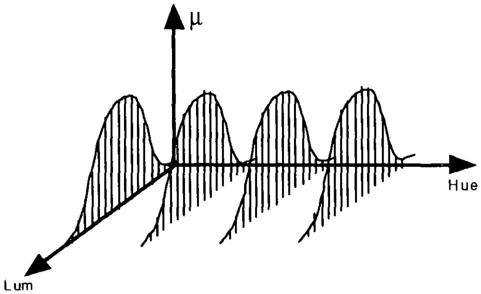


Fig. 5: Uncertainty specification corresponding to equation (1) and (2)

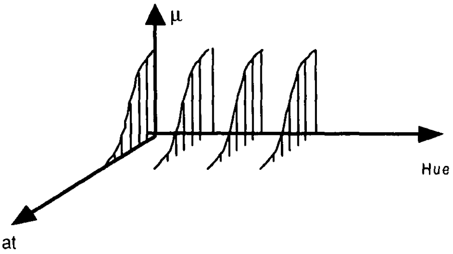


Fig. 6: Uncertainty specification corresponding to equation (3)

5 A SYSTEM FRAMEWORK FOR FUZZY EXPLORATORY ANALYSIS

In this section, we design a framework for exploratory analysis which is mainly oriented to the fuzzy spatial analysis. The framework is also applied to conventional non-fuzzy spatial analysis.

The whole system framework consists of four parts, and there is message communication between parts as the two way arrow indicate in Fig. 7. The framework actually is a object-oriented architecture, which uses the well-known Model-View-Control (MVC) model, first introduced for object-oriented development in the Smalltalk-80 language (Strand 1993).

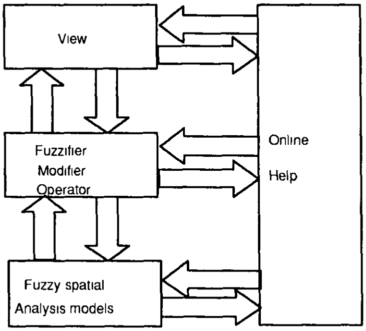


Fig. 7: A system framework for fuzzy exploratory analysis

View

View is the interface of the system for end-users, which is supported by a Graphical User Interface (GUI). Presently, there are a number of GUI standards available like Microsoft Windows initial SDK or OSF/Motif.

A Multiple Document Interface (MDI) offers the multiple visual perspectives simultaneously for deep insight and comparison in fuzzy exploratory analysis. There is one principal window serving as the entry of the system with various child windows to represent different layers and sublayers. These windows could also be regarded as an interface with the database in which any change of data in the database will be reflected in the windows while the analysis is progressing. Basically, there are two kinds of windows, one is the graphics-oriented window in which attribute data are visualized with color or symbols, and which offers intuitive visualization; and the other provides purely attribute data, which are visualized in graphics-oriented windows simultaneously. It is allowed to switch between these different windows when required.

Control

The control serves the exploratory function, by supporting different interface components such as dialog, icon, etc. According to the architecture of fuzzy spatial analysis, it is mainly three kinds of controls that need to be taken into account.

1) Fuzzifier: It is the principal component to do fuzzification. It can be designed as a modeless dialog to make the exploratory results appear on the corresponding views in real-time.
2) Modifier: It mainly indicates the linguistic hedges in fuzzy spatial analysis. In addition to modeless dialog, a slider bar could serve the task of the modifier.
3) Operator: For the operator, there are basically two kinds, one is the unary operator, and another is a binary operator.

Model

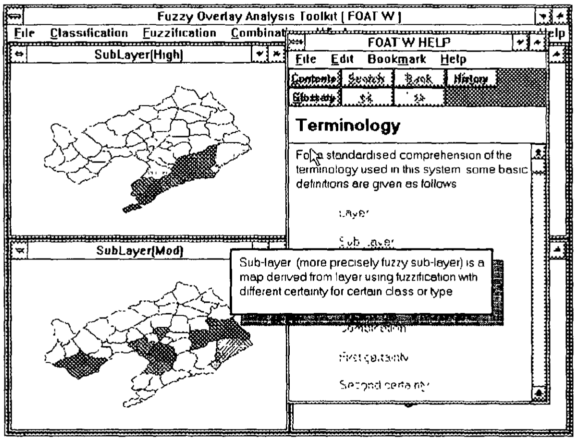
Model is the term for full fuzzy spatial analysis including fuzzy overlay model, fuzzy buffer model, fuzzy search model etc.

Online Help

Online help provides not only the instructions to use the software itself, but to use the help of analysis methods as well. For a successful exploratory analysis system, online help is critically important to facilitate the exploratory process. Preliminary, there are three basic modules to be considered (Fig. 8).

1. Terminology and notations: for a new system, it is essential to offer the analyst some basic concepts about the system. In our prototype system FOAT:W, for example, a set of definitions like fuzzification, sublayer, first certainty, second certainty is presented (Jiang, Kainz and Muller 1995).
2. Commands: It gives details about the operation of the system.
3. Interpretation: It provides detailed explanation about the strategy of visualization in online style. If the above two items are commonly available to other kinds of system, this one is available uniquely for exploratory analysis systems.

The online help of existing software consist mainly of texts or documents, but for exploratory analysis in GIS graphics will be highly incorporated.


Fuzzy Overlay Analysis Toolkit [FOAT W]
File Classification Fuzzification Combinat
elp
SubLayer(High)
SubLayer(Mod)
FOAT W HELP
File Edit Bookmark Help
Contents Search Back History
Glossary << >>

Terminology

For a standardised comprehension of the terminology used in this system some basic definitions are given as follows

Layer

Sub-layer

Sub-layer (more precisely fuzzy sub-layer) is a map derived from layer using fuzzification with different certainty for certain class or type

First certainty

Second certainty

Fig. 8: An online help implemented in FOAT:W

6 CONCLUSION

The framework presented in this paper needs to be expanded further for practical implementation. Parts of it and the prototype system FOAT:W have been implemented, and although the functions are only partly available, they have shown the promise that it is possible to produce a powerful visualization support for fuzzy spatial analysis. The FOAT:W will be extended to a practical visualization tool in the near future in the following aspects:

1. Integration: It will be integrated into existing GISs based on some standardized GUIs as OSF/Motif, X-Windows, especially for MS-Windows and MS-Windows NT.
2. The color-based operation proposed in this paper opens up many new possibilities for exploratory analysis.

3. The option of online help promises substantial potential in the development of exploratory analysis systems. In addition to the nonlinear text, graphics can greatly improve the capacity for exploratory analysis.

REFERENCES

Goodchild, M. F. (1987). A Spatial Analysis Perspective on Geographical Information Systems, *Int. J. Geographical Information System*, Vol. 1, pp. 327-334.

Jiang, B. and Kainz, W., (1994). Fuzzification as a Basis for Fuzzy Spatial Analysis. *Proceedings of IAI'94*, Wuhan, P.R. China, pp. 294-302.

Jiang, B., Kainz, W., and Muller, J. C. (1995). Fuzzy Overlay Model -- Overlay Operation Incorporated with Linguistic Notions, to be published.

Leung, Y. (1984). Fuzzy Sets Approach to Spatial Analysis and Planning -- A Nontechnical Evaluation (Occasional Paper), The Chinese University of Hong Kong.

MacEachren, A. M. and Monmonier, M. (1992). Introduction, Special Issue of *Cartography and Geographic Information Systems*, Vol. 19, No. 4, pp. 197-200.

Monmonier, M. (1990). Summary graphics for Integrated Visualization in Dynamic Cartography, *Cartography and Geographic Information Systems*, Vol. 19, No. 1, pp. 23-26

Olson, J. M. (1987). Color and the Computer in Cartography, In: Durrett H. J. (ed.) *Color and the Computer*, Academic Press Inc., pp. 205-219.

Openshaw, S. (1990). Spatial Analysis and Geographical Information Systems: a Review of Progress and Possibility, In: Scholten H.J., and Stillwell J.C.H. (eds.) *Geographical Information System for Urban and Regional Planning*, Kluwer Academic Publishers. pp. 153-163.

Strand, E. J. (1993). Model-View-Controller Architecture Expedites Embedded Application, *GIS World*, Oct. 1993, pp. 20-21.

Tang, Q. (1992). A Personal Visualization System for Visual Analysis of Area-based Spatial Data, *Proceedings of GIS/LIS'92*, pp. 767-776.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley.

Van der Wel, F. J. M., Hootsmans, R. M. and Ormeling, F. J. (1994). Visualization of Data Quality. In: MacEachren A. M. and Taylor D. R. F. (eds.) *Visualization in Modern Cartography*, Pergamon Press, pp. 313-331.

Xia, F. F. and Fotheringham, A. S. (1993). Exploratory Spatial Data Analysis with GIS -- The Development of the ESDA Module Under ARC/INFO, *Proceedings of GIS/LIS'93*, pp. 801-810.

Zadeh, L. A. (1973). Outline of a New Approach to the Analysis of Complex Systems and Decision Process, *IEEE Trans. System, Man, and Cybernetics*, SMC-3, pp. 28-44.

On the Robustness of Qualitative Distance- and Direction-Reasoning*

Jung-Hong Hong
Department of Surveying Engineering
National Cheng-Kung University
1 University Road
Tainan, Taiwan, Republic of China
junghong@mail.ncku.edu.tw

Max J. Egenhofer
National Center for Geographic Information and Analysis
and
Department of Surveying Engineering
Department of Computer Science
University of Maine
Orono, ME 04469-5711, U.S.A.
max@mecan1.maine.edu

Andrew U. Frank
Department of Geo-Information E127
Technical University Vienna
Gusshausstr. 27-29
A-1040 Vienna, Austria
frank@geoinfo.tuwien.ac.at

Abstract

This paper focuses on spatial information derived from the composition of two pairs of cardinal directions (e.g., North and North-East) and approximate distances (e.g., near and far), i.e., given the approximate distances a1 (A, B) and a2 (B, C) and the cardinal directions c1 (A, B) and c2 (B, C), what are a3 (A, C) and c3 (A, C)? Spatial reasoning about cardinal directions and approximate distances is challenging because distance *and* direction will affect the composition. This paper investigates the dependency between qualitative and quantitative inference methods for reasoning about cardinal directions and approximate distances. Cardinal directions are based on a 4-sector model (North, East, South, West), while approximate distance correspond to a set of ordered intervals that provide a complete partition (non-overlapping and mutually exclusive) such that the following interval is greater than or equal to the previous one (for example, "far" would extend over a distance that is at least as great as "medium.") We ran comprehensive simulations of quantitative reasoning, and compared the results with the ones obtained from quantitative reasoning. The results indicate that the composition is robust if the ratio between two consecutive intervals of quantitative distances is greater than 3.

Introduction

The domain of this paper is the intelligent inference of spatial information in Geographic Information Systems (GISs), which record a variety of geographic data to aid human decision making about the real world. Spatial reasoning denotes the inference of new

* This work was partially supported through the NCGIA by NSF grant No. SES-8810917. Additionally, Max Egenhofer's work is also supported by NSF grant No. IRI-9309230, a grant from Intergraph Corporation, and the Scientific Division of the North Atlantic Treaty Organization. Jung-Hong Hong's work was partially supported by a University of Maine International Student Tuition Waiver. Andrew Frank's work is partially supported by a grant from Intergraph Corporation.

spatial information that is otherwise not available. People do spatial reasoning using knowledge they acquire from environment and learning experiences, so that they can make decisions even if the available spatial information is incomplete, uncertain, or inconsistent (Collins *et al.* 1975). For example, after living in an area for a period of time, people can usually find a path from one place to another, or draw a sketch map about the whole area—even if they would not know the exact relationship between any two objects. People have the flexibility to adapt to the environment and the reasoning processes do not necessarily follow human-designed models like mathematics or logic. Unlike humans, information systems with spatial reasoning capabilities must rely on an appropriately designed spatial models and their formalizations. Of course the result of such a reasoning must make sense to humans (Kieras 1990). From the perspective of data management and query processing, the design of spatial reasoning mechanisms for GISs must consider at least two factors: (1) it must have unambiguous definitions of spatial relations and operations to process queries and (2) it must have the capability to search for appropriate combinations of relations and derive reasonable answers promptly.

The domain of this paper is reasoning about qualitative distances and directions (e.g., near and North), also known as approximate distances and cardinal directions (Frank 1992). Since people use spatial relations to make sense of observations, spatial relations become an essential part of GISs. Qualitative spatial relations—topological relations, cardinal directions, and approximate distances—are important, because they are closer to human cognition in everyday lives than their quantitative counterparts. Qualitative relations are based on a small set of symbols and a set of inference rules of how to manipulate symbols. Although qualitative relations are often vague in their geometric meaning and have less resolution than their quantitative counterpart, people have little difficulty in processing them and using them to communicate with others. Dutta (1989) and Freksa (1992) even argued that most human spatial reasoning is qualitative rather than quantitative. Currently, human spatial behaviors are yet incompletely understood and consequently the design of spatial theory and its formalization for GIS is difficult. Future GISs should not only be mechanisms for the storage and retrieval of geographic information as most current database systems do, but also be intelligent knowledge-based systems capable of incorporating human expertise to mimic human behaviors in decision making (Abler 1987). Since qualitative spatial relations stored in databases are often incomplete, the deduction of new relations has to rely on built-in inference mechanisms (Kidner and Jones 1994; Sharma *et al.* 1994); however, such processing of qualitative spatial relations in computers is currently impeded by the lack of a better understanding of human spatial knowledge.

Most commonly used reasoning mechanisms for qualitative spatial relations are purely quantitative. Examples are coordinate calculations, which derive from a number of quantitative spatial relations a new spatial relation, also in a quantitative format. These reasoning mechanisms, however, cannot be directly applied to qualitative spatial relations (Futch *et al.* 1992). Recent research focused on the individual types of qualitative spatial relations (Egenhofer and Franzosa 1991; Peuquet 1992; Cui *et al.* 1993), but only few researchers investigated spatial reasoning involving different types of spatial relations (Dutta 1991; Freksa and Zimmermann 1992; Hernández 1993). This paper investigates reasoning about the spatial relations of distance *and* direction, called *locational relation.* By considering these two types of relations simultaneously, stronger constraints between two objects can be established. We suggest to build a reasoning model on the basis of composition operators, which describes the behavior of the combination of two locational relations. The goal of this model is to derive approximate, reasonable, and qualitative reasoning results with the defined composition operators.

Although this paper deals with only the reasoning about qualitative locational relations, it does not suggest that qualitative reasoning should replace the widely used quantitative reasoning. Qualitative and quantitative representation are complementary approaches for human abstractions of the spatial relations in the real world, and one or the other should be used whenever it best serves users' needs.

Background

When qualitative spatial relations are explicitly stored in a database, there are two scenarios when processing a query (Sharma *et al.* 1994):

- If the queried relation is already stored in the database, the system retrieves this information and delivers it to the user.
- If the queried relation is not directly available, a reasoning mechanism must be invoked to infer the queried relation from those relations that exist in the database.

The reasoning mechanism must have two basic functions: First, it must be able to analyze if the available information is sufficient to derive the queried relation. If so, the reasoning model takes selected relations to derive the queried relations. The reasoning result is preferred to be conclusive, i.e., the number of possible answers should be as small as possible. The core of this reasoning model is a set of composition that define the composition behavior for the particular types of relations. A composition operator ";" takes two known relations, r1 (A, B) and r2 (B, C), to derive the relation r3 between A and C, i.e.,

$$r1\ (A, B)\ ;\ r2\ (B, C) \Rightarrow r3\ (A, C) \qquad (1)$$

Vector addition is a good analogy to composition. The addition of vector (A, B) and vector (B, C) will yield vector (A, C). For locational relations, vector addition is actually what the quantitative-based reasoning approach is based on. Nevertheless, this concept is not directly applicable to the composition of qualitative locational relations.

Related Work

Most current spatial reasoning systems and GISs only store locational relations in a quantitative format and consequently only deal with spatial reasoning in a quantitative matter, e.g., through numerical calculations for distances and directions as in Kuipers' (1978) TOUR model. The last few years have seen a growing interest in qualitative inferences of spatial relations. The common part among different reasoning models suggested is that they start with the modeling of the spatial domain (geographic space or spatial relations) and then define their respective composition operators. Depending on the way the reasoning is conducted, we divide the different approaches into those that transform qualitative locational relations into a quantitative format and solve the reasoning problem quantitatively; and those that use operators to define the composition behavior of qualitative relations.

In the first class, the result can be either kept in a quantitative format or transformed back to a qualitative format. For example, the SPAM model (McDermott and Davis 1984) treats a qualitative locational relation as a fuzz box and calculates the possible range for the queried relation. On the other hand, Dutta (1989) used fuzzy set theory (Zadeh 1974) to model the approximate and uncertain nature of qualitative locational relations. Both models require a transformation between qualitative and quantitative representations to be established first. After such transformations, the newly inferred relations are calculated with a quantitative method.

Spatial reasoning models in the second class employ a purely qualitative reasoning process. For example, symbolic projections (Chang *et al.* 1987) were originally proposed to store the spatial knowledge in an image. This model records the relationships among objects in a symbolic image separately in horizontal or vertical directions. Several extensions have been suggested in the past years (Jungert 1988; Jungert 1992; Lee and Hsu 1992). Despite of these extensions, the symbolic projection model does not record distance information, that is, it cannot appropriately represent "how far" two separated objects are. Allen's (1983) temporal logic, originally proposed to solve the reasoning in one-dimensional domain, was expanded to reasoning in higher-dimensional spaces (Guesgen 1989). Papadias and Sellis (1992; 1993) suggest the use of symbolic arrays to store the spatial information hierarchically and take advantage of the array structure for reasoning. Freksa (1992) developed a reasoning model for qualitative directions based on an orientation grid and the conceptual neighbors of qualitative spatial relations. Zimmermann (1993) extended this model to include distance constraints, represented by comparing a distance value to a known distance (> di, = di, < di). Hernández (1991) suggested a model that deals with the reasoning of both directions and topological relationships. He separated these two types of reasoning and solve the reasoning problem individually. No model mentioned above investigates the reasoning of qualitative distances (e.g., near). Frank (1992) developed an algebra for the reasoning of qualitative distances and directions. Like Hernández's model, it also separates these two types of relations and solve the reasoning individually. This algebra can achieve satisfactory results under some restricted condition; for some cases, only approximate results can be derived, i.e., the queried relation is not conclusive.

A Model of Qualitative Distances and Directions

A locational relation includes a qualitative distance component and a qualitative direction component. To simplify the problem domain, only locational relations between point-like objects will be considered here. One property of qualitative distances and directions is their imprecise geometric meaning. Take distances for example, near is often interpreted as a range of quantitative distances rather than a specific value ("The church is about 50 meters away.") The same applies to qualitative directions. Although one may intuitively interpret East as a specific quantitative direction (i.e., azimuth of 90 degree), people often consider an object whose azimuth is 85 degrees to be East as well. Peuquet and Zhan (1987) adopted the cone-shaped concept to investigate the cardinal directions between two extended objects, which also treats a cardinal direction as a range of quantitative directions.

Qualitative relations will be represented by *symbolic values* (in comparison with numerical values for quantitative relations). We therefore have symbolic distance values and symbolic direction values; each one of them represents a specific locational constraint. The name of the symbolic values can be chosen arbitrarily as long as its semantic meaning is reasonable and will not cause confusion. For example, North and South are usually understood as two directions in the opposite direction and the name selection should not violate that. Theoretically the number of symbolic distance values is not limited, yet research in psychology and cognitive science has demonstrated that the number of categories humans can handle simultaneously has a limitation. In this paper, the number of symbolic distance values is chosen to be four. The proposed model can be expanded to include less or more symbolic distance values according to the application. The number of symbolic direction values depends on the applications. Two often used direction systems include either four or eight symbolic values. The following lists a model consisting of four symbolic distance values and eight symbolic direction values.

Distance: {very near, near, far, very far}
Direction: {North, Northeast, East, Southeast, South, Southwest, West, Northwest}

Certain order relationships exist among these symbolic values. The order among symbolic distance values describes distances from the nearest to the furthest. The order among symbolic direction values can be either clockwise or counter-clockwise. To simplify the model design, the following two criteria are introduced:

- complete coverage, i.e., the designed symbolic values describe any situation in its respected domain, and
- mutual exclusive, i.e., any situation in the domain can be described by one and only one symbolic value.

Mapping Qualitative onto Quantitative Locational Relations

Qualitative and quantitative representations describe the same domain, only the symbols used are different (symbolic values vs. numerical values). Since a relation between two objects can be represented qualitatively or quantitatively, it should be possible to transform between these two representations. Also, the number of symbolic values is smaller than its quantitative counterpart, but it has to describe the same domain, so it is reasonable to assume that a symbolic value should correspond to a range of quantitative values (an interval on a one-dimensional scale). On the other hand, a quantitative value should correspond to only one symbolic value. Because of the property of mutual exclusiveness, no gap or overlap will be allowed between two neighboring intervals. Therefore, a number of symbolic values correspond to the same number of intervals of the quantitative values; that is, there is an interval-based transformation between the qualitative and quantitative representations.

This interval-based transformation is context dependent. For example, near can be interpreted as an interval from 0 to 500 meters for walking, while also as an interval from 5 km to 10 km for driving. Such an interval-based transformation can be applied to both the domain of distances and directions. For distance systems, every symbolic value corresponds to an interval of quantitative distances. This divides a two-dimensional space into several tracks; each track represent a unique qualitative distance (Figure 1a). If the cone-shaped concept is chosen for direction systems, the direction domain is divided into a number of cones with the same resolution (Figure 1b). The basic property is that the geometric range of each cone increases with the increase of the distance (Peuquet and Zhan 1987).

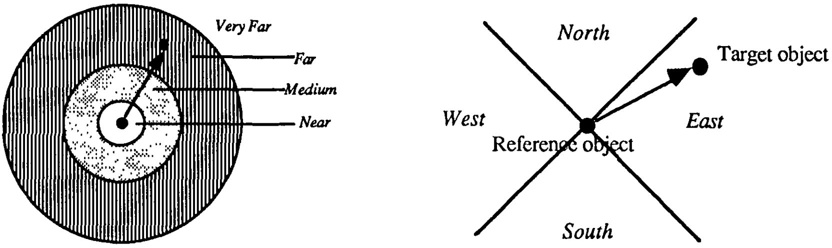


Figure 1: (a) Qualitative distances and (b) qualitative directions.

By considering distances and directions together, the locational relation system becomes a sector-based model (Figure 2), where each sector corresponds to a specific pair of symbolic distance and direction values. Objects in the same sector share the same qualitative locational relationship with respect to the origin of the system.

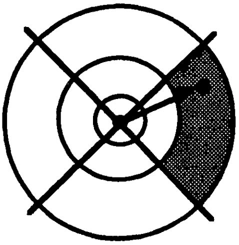


Figure 2: Illustration of the space divided by the distance and direction systems.

With this model, the composition of two qualitative locational relations can be numerically simulated by the compositions of their corresponding quantitative locational relations.

Assuming that every sector can determine N quantitative locational relations, the total number of possible composition between these two sectors is N^2. Every quantitative composition can be mapped onto a qualitative locational relation, which represents a possible answer for the particular composition. The set of possible answers can be used to define the composition operators for qualitative locational relations. Figure 3 illustrates this transformation. Given two qualitative locational relations QL_1 and QL_2, determine the transformations $f(QL_i)$ and $f^{-1}(QN_j)$, and map with $f\,QL_1$ onto a set of quantitative relations QN_1 and QL_2 to a set of QN_2 , respectively. In the quantitative domain, apply quantitative reasoning methods to all the possible combinations between the sets pf QN_1 and QN_2 to derive a set of results QN_3, which is then mapped onto a set of QL_3 using f^1.This process is based on the interval-based numeric simulation and well-developed quantitative inference methods.

$$\begin{tikzcd}
\begin{matrix} QN_1 \\ QN_2 \end{matrix} \arrow[rr, "QN_1;QN_2"] & & QN_3 \arrow[d, "f^{-1}(QN_3)"] \\
\begin{matrix} QL_1 \\ QL_2 \end{matrix} \arrow[u, "f(QL_i)"] \arrow[r, dotted, "QL_1;QL_2"] & QL_3 & QL_3'
\end{tikzcd}$$

Figure 3: Framework of the reasoning model design.

Given any pair of locational relations and an interval-based transformation, their composition operator can thus be defined. For example,

$$\text{(very near, North)} ; \text{(very near, North)} \Rightarrow \text{(very near, North) or (near, North)} \qquad (2)$$

Theoretically there is an infinite number of interval-based transformation depending on the applications. To use composition operators as the basis for reasoning, it is important to investigate if the interval-based transformation will affect the definition of composition operators. If the composition operators are largely dependent on the interval-based transformation, it is not necessary to define composition operators, as the queried relation must be calculated anyway. If the interval-based transformation is insignificant to the definition of composition operators, i.e., composition operators are robust, then it is possible to build a qualitative reasoning mechanism on the basis of the composition operators, in which no calculation will be necessary for the inference.

The All-Answer Model

To ensure the correctness for the queried relation, most qualitative reasoning models define their composition operators in a way that all the possible answers will be found. This concept will be called the *all-answer model.* Since this model is based on numeric simulation, efficient sampling becomes very important. We find that it is only necessary to select samples on the boundary of the two sectors with which the two locational relations are corresponding to. This can largely reduce the number of samples, while still derive all the possible answers. Since the all-answer model only checks if a particular relation is a possible answer, every possible answer will be treated equally. That is, although some answers may turn out to be more probable, the all-answer model will not distinguish them. To simplify this answer selection process, Hong (1994) suggested two other models, the *likely-answer model*—eliminating compositions that have a low probability—and the single-answer model—selecting the composition with the highest probability.

The advantage of the all-answer model is that the actual answer is guaranteed to be one of the possible answers at the end of the reasoning process. However, the disadvantage is that sometimes the number of possible answers becomes too high and the reasoning process becomes very complicated. To solve this problem, more than one combination of locations must be found so that the queried relations can be better constrained. Furthermore, for better efficiency, the reasoning mechanism should have built-in intelligence to select appropriate combinations of relations with better locational constraints and discard those without.

Simulation of Test Data

The interval-based transformation is subjective to applications and individual experiences. To investigate the influence the interval-based transformation has on the definition of composition operators, thirteen interval sets were tested (Table 1). The interval sets are designed in a way that there is a constant ratio relationship between the lengths of two neighboring intervals. Of course these thirteen sets do not make a complete list for the intervals humans may use. The intention here is to observe the results of the defined composition operators to investigate their robustness and distribution. The finding is very important to the design and evaluation of qualitative reasoning mechanism.

Ratio	$dist_0$	$dist_1$	$dist_2$	$dist_3$
1	(0, 1]	(1, 2]	(2,3]	(3, 4]
2	(0, 1]	(1, 3]	(3 , 7]	(7, 15]
3	(0, 1]	(1, 4]	(4, 13]	(13, 40]
4	(0, 1]	(1, 5]	(5, 21]	(21, 85]
5	(0, 1]	(1, 6]	(6, 31]	(31, 156]
6	(0, 1]	(1, 7]	(7, 43]	(43, 259]
7	(0, 1]	(1, 8]	(8, 57]	(57, 400]
8	(0, 1]	(1, 9]	(9, 73]	(73, 585]
9	(0, 1]	(1, 10]	(10, 91]	(91, 820]
10	(0, 1]	(1, 11]	(11, 111]	(111, 1111]
20	(0, 1]	(1, 21]	(21, 421]	(421, 8421]
50	(0, 1]	(1, 51]	(51, 2551]	(2551, 127551]
100	(0, 1]	(1, 101]	(101, 10101]	(10101 , 1010101]

Table 1: Simulated intervals for four symbolic distance values.

Although tests on different numbers of symbolic distance and direction values were conducted, only the group of four symbolic distance values and eight symbolic direction values will be discussed here. Detailed discussions about other groups (e.g., three distances and eight directions) can be found in (Hong 1994).

Robustness

Given two locational relations, if their composition operator remains the same despite the changes of interval-based transformation, the composition operator is *robust*. If so, it is unnecessary to define composition operators for every interval-based transformation and a group of transformation can share the same set of composition operators.

To measure the robustness, a quantitative measure, called *robustness measure* (*RM*), is introduced. It is the ratio between the numbers of answers in the robust set (the intersection of the qualitative and quantitative sector), and the union set (the set union of the two sectors). The domain of *RM* is $\{0 \leq RM \leq 1\}$. Two compositions are compatible if *RM* is equal to 1. If there is no common answer between the two sets of answers, *RM* is equal to 0. Two compositions are therefore incompatible if their robustness measure is less than 1. The advantage of this method is that if most of the selected answers between these two sets are similar, their robustness measure will be close to 1.

Table 2 gives the robustness measures based on the direction differences. The first variable shows the number of incompatible composition in a group and the second variable shows the robustness measure for the group. Between ratio 1 and ratio 2, thirteen compositions do not have the same set of selected answers. The average robustness measure of this group is 0.95, which indicates that the selected answers between ratio 1 and ratio 2 are either identical or very similar. All compositions are robust if the ratio is greater than 2.

	Δdir = 0	Δdir = 1	Δdir = 2	Δdir = 3	Δdir = 4	Total
ratio (1, 2)	(2, 0.92)	(1, 0.98)	(1, 0.99)	(3, 0.94)	(6, 0.92)	(13, 0.95)
ratio (2, 3)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
ratio (2, 100)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)

Table 2: Robustness measures.
ratio (a, b): a = number of incompatible compositions; b = *RM* for the group.
Δdir = direction difference.

The majority of the compositions for the thirteen groups are robust, with the first group (ratio = 1) as the only exception. When we specifically compare the two groups of compositions where ratio is 1 and 2, there are 13 cases (out of 74) between which the possible answers are different. This result indicates that even if the composition operators are not rigorously robust, most of them are robust. From a cognitive and linguistic point of view, such an interval set—ratio = 1, length (near) = length (far)—rarely exists. It is therefore possible to build a model on the basis of the composition operators, such that the context-dependent nature of qualitative distances will not affect the definition of composition operators except for some extreme cases (e.g., ratio = 1).

Distribution

Distribution is the analysis of the selected answers for a particular composition. To make the reasoning process easier, it is important to keep the number of selected answers as few as possible. The following lists some important finding regarding to the distribution of the selected answers using the all-answer model. Through such an analysis, we wish to identify some "preferred" compositions (i.e., fewer possible answers) that can provide better constraints on the queried relation. This can certainly be used as the basis for the design of an intelligent reasoning mechanism.

- The number of possible answers increases with the direction differences; therefore, the further apart the two directions are, the less-constrained their composition is.
- A composition usually extends over 2 or 3 distance values.
- The number of possible directions for each composition largely depends on the direction differences.
- Compositions of relations in the same direction provide the best constraint for both distances and directions.
- Compositions of two relations in opposite directions provide the least constraints, especially for directions.
- Compositions of relations with different distance values usually provide better constraints than compositions of relations with the same distance value.
- If the distance values of the two relations are different, the composition largely depend on the relation with the greater distance value.

Conclusions

This paper demonstrated the first results for investigations into reasoning about qualitative distances and directions. We proposed to build a qualitative reasoning model that takes two qualitative locational relations to derive a new locational relation also in a qualitative format. An important finding is that the composition operator of two qualitative locational relations are robust for the majority of the cases tested. Since the context-dependent nature of the qualitative distances does not seem to be a significant factor, we can build a reasoning mechanism on the basis of the composition operators.

From the above discussion, we can conclude that the distance constraint is usually poor, no matter what kind of combination is used, because it cannot be narrowed down to one single answer. In most situations, two relations with greater distance differences will provide better constraints. Compared with human reasoning, this finding is reasonable. For example, when asked about the relationship between San Francisco and Washington D.C., people are likely to select the relation San Francisco-Baltimore and Baltimore-Washington D.C. for reasoning rather than using the relation San Francisco-New Orleans and Washington D.C.-New Orleans. On the other hand, it is clear that the direction is much better constrained if the two locational relations are in the same direction. This is again not surprising when compared to human reasoning.

Although the all-answer model provides all the possible answers, the number of selected answers is often high and the reasoning process is expected to be tedious and inefficient. For the worst case, there is probably no conclusive result at the end of the reasoning. Further modification on the reasoning model that only track more likely answers is an alternative (Hong 1994).

To simplify the problem domain, we enforced some assumptions (e.g., mutual exclusiveness) on the transformation between qualitative and quantitative representations. It is not clear if humans do possess such a fine line to distinguish between two symbolic distance or direction values. It is therefore of interests to further investigate if the robustness still exists provided some parameters are changed (e.g., if one allows that two neighboring intervals overlap). Also, the discussion in this paper was restricted to point-like objects; to be used in GISs, this model must be further expanded to higher-dimensional domains, or integrated into a hierarchical spatial inference model.

References

R. Abler (1987) The National Science Foundation National Center for Geographic Information and Analysis. *International Journal of Geographical Information Systems* 1(4): 303-326.

J. F. Allen (1983) Maintaining Knowledge about Temporal Intervals. *Communications of the ACM* 26(11): 832-843.

S. K. Chang, Q. Y. Shi, and C. W. Yan (1987) Iconic Indexing by 2-D Strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9(6): 413-428.

A. Collins, E. Warnock, N. Aiello, and M. Miller (1975) Reasoning From Incomplete Knowledge. in: D. Bobrow and A. Collins (Eds.), *Representation and Understanding*. pp. 383-415, Academic Press, New York, NY.

Z. Cui, A. Cohn, and D. Randell (1993) Qualitative and Topological Relationships in Spatial Databases. in: D. Abel and B. Ooi (Eds.), *Third International Symposium on Large Spatial Databases. Lecture Notes in Computer Science* 692, pp. 296-315, Springer-Verlag, New York, NY.

S. Dutta (1989) Qualitative Spatial Reasoning: A Semi-Quantitative Approach Using Fuzzy Logic. in: A. Buchmann, O. Günther, T. Smith, and Y. Wang (Eds.), *Symposium on the Design and Implementation of Large Spatial Databases. Lecture Notes in Computer Science* 409, pp. 345-364, Springer-Verlag, New York, NY.

S. Dutta (1991) Topological Constraints: A Representational Framework for Approximate Spatial and Temporal Reasoning. in: O. Günther and H.-J. Schek (Eds.), *Advances in Spatial Databases—Second Symposium, SSD '91. Lecture Notes in Computer Science* 525, pp. 161-180, Springer-Verlag, New York, NY.

M. Egenhofer and R. Franzosa (1991) Point-Set Topological Spatial Relations. *International Journal of Geographical Information Systems* 5(2): 161-174.

A. Frank (1992) Qualitative Spatial Reasoning about Distances and Directions in Geographic Space. *Journal of Visual Languages and Computing* 3(4): 343-371.

C. Freksa (1992) Using Orientation Information for Qualitative Spatial Reasoning. in: A. Frank, I. Campari, and U. Formentini (Eds.), *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Lecture Notes in Computer Science* 639, pp. 162-178, Springer-Verlag, New York, NY.

C. Freksa and K. Zimmermann (1992) On the Utilization of Spatial Structures for Cognitively Plausible and Efficient Reasoning. in: *IEEE International Conference on Systems, Man, and Cybernetics*, Chicago, IL, pp.

S. Futch, D. Chin, M. McGranaghan, and J.-G. Lay (1992) Spatial-Linguistic Reasoning in LEI (Locality and Elevation Interpreter). in: A. Frank, I. Campari, and U. Formentini (Eds.), *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Lecture Notes in Computer Science* 639, pp. 318-327, Springer-Verlag, New York, NY.

H. W. Guesgen (1989) *Spatial Reasoning Based on Allen's Temporal Logic*. International Computer Science Institute, Technical Report TR-89-049.

D. Hernández (1991) Relative Representation of Spatial Knowledge: The 2-D Case. in: D. Mark and A. Frank (Eds.), *Cognitive and Linguistic Aspects of Geographic Space*. pp. 373-385, Kluwer Academic Publishers, Dordrecht.

D. Hernández (1993) Maintaining Qualitative Spatial Knowledge. in: A. Frank and I. Campari (Eds.), *Spatial Information Theory, European Conference COSIT '93, Marciana Marina, Elba Island, Italy*. 716, pp. 36-53, Springer-Verlag, New York, NY.

J.-H. Hong (1994) *Qualitative Distance and Direction Reasoning in Geographic Space*. Ph.D. Thesis, University of Maine.

E. Jungert (1988) Extended Symbolic Projections as a Knowledge Structure for Spatial Reasoning. in: J. Kittler (Ed.), *4th International Conference on Pattern Recognition. Lecture Notes in Computer Science* 301, pp. 343-351, Springer-Verlag, New York, NY.

E. Jungert (1992) The Observer's Point of View: An Extension of Symbolic Projections. in: A. Frank, I. Campari, and U. Formentini (Eds.), *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Lecture Notes in Computer Science* 639, pp. 179-195, Springer-Verlag, New York, NY.

D. Kidner and C. Jones (1994) A Deductive Object-Oriented GIS for Handling Multiple Representations. in: T. Waugh and R. Healey (Eds.), *Sixth International Symposium on Spatial Data Handling*, Edinburgh, Scotland, pp. 882-900.

D. Kieras (1990) Cognitive Modeling. in: S. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence*. pp. 111-115, Wiley-Interscience, New York, NY.

B. Kuipers (1978) Modeling Spatial Knowledge. *Cognitive Science* 2: 129-153.

S.-Y. Lee and F.-J. Hsu (1992) Spatial Reasoning and Similarity Retrieval of Images Using 2D C-String Knowledge Representation. *Pattern Recognition* 25(3): 305-318.

D. McDermott and E. Davis (1984) Planning Routes through Uncertain Territory. *Artificial Intelligence* 22: 107-156.

D. Papadias and T. Sellis (1992) Spatial Reasoning Using Symbolic Arrays. in: A. Frank, I. Campari, and U. Formentini (Eds.), *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space. Lecture Notes in Computer Science* 639, pp. 153-161, Springer-Verlag, New York, NY.

D. Papadias and T. Sellis (1993) The Semantics of Relations in 2D Space Using Representative Points: Spatial Indexes. in: A. Frank and I. Campari (Eds.), *Spatial Information Theory, European Conference COSIT '93, Marciana Marina, Elba Island, Italy*. 716, pp. 234-247, Springer-Verlag, New York, NY.

D. Peuquet (1992) An Algorithm for Calculating Minimum Euclidean Distance Between Two Geographic Features. *Computers and Geosciences* 18(8): 989-1001.

D. Peuquet and C.-X. Zhan (1987) An Algorithm to Determine the Directional Relationship Between Arbitrarily-Shaped Polygons in the Plane. *Pattern Recognition* 20(1): 65-74.

J. Sharma, D. Flewelling, and M. Egenhofer (1994) A Qualitative Spatial Reasoner. in: T. Waugh and R. Healey (Eds.), *Sixth International Symposium on Spatial Data Handling*, Edinburgh, Scotland, pp. 665-681.

L. Zadeh (1974) *Fuzzy Logic and Its Application to Approximate Reasoning*. North-Holland Publishing Company.

K. Zimmermann (1993) Enhancing Qualitative Spatial Reasoning—Combining Orientation and Distance. in: A. Frank and I. Campari (Eds.), *Spatial Information Theory, European Conference COSIT '93, Marciana Marina, Elba Island, Italy. Lecture Notes in Computer Science* 716, pp. 69-76, Springer-Verlag, New York, NY.

AUTOMATING LINEAR TEXT PLACEMENT WITHIN DENSE FEATURE NETWORKS

David H. Alexander
Carl S. Hantman
Geography Division
U.S. Bureau of the Census
Washington, D.C. 20233

ABSTRACT

In preparation for the 2000 census, the Census Bureau plans to provide maps with address information to local governments for review and update. The algorithms that have been used by the Census Bureau in the past have not been flexible enough to pl'ace this kind of text readably on small-scale maps. This paper describes a new non-interactive algorithm which has been developed at the Census Bureau. This algorithm simultaneously considers all the address range text which must be plotted in a certain area of the map and can make incremental adjustments to the location of each text item in the set in order to position it for maximum clarity before any positions are fixed. The flexibility that this approach provides has already made it possible to produce maps that display address ranges, census block numbers, and linear features and their names at a scale significantly smaller than comparable maps produced for the 1990 Census.

INTRODUCTION

During the 1980's, cartographers and computer programmers at the U.S. Census Bureau developed a fully automated mapping system to produce the thousands of maps that support census field operations and data products. The software developed for the automatic placement of feature names on maps formed an integral part of this system. The text placement algorithms were generally successful in meeting the requirements of the maps for the 1990 decennial census.

Currently, the Census Bureau is developing a Master Address File for use in the 2000 decennial census and other surveys and censuses. In conjunction with this effort, the Census Bureau is updating its TIGER data base to keep it as current and accurate as possible. As part of the TIGER Improvement Program (TIP), the Census Bureau will provide local governments with maps displaying streets and address ranges within their jurisdictions. Participating local officials will update the maps with new and revised feature and address information. These changes will then be incorporated into TIGER to improve the linkage of the Master Address File to the TIGER data base.

THE PROBLEM

The addition of address range text to maps that must also display linear and areal feature names, poses cartographic challenges for automated text placement algorithms. In addition, the large number of unique map sheets that will be generated for this program precludes interactive intervention for text placement refinement. Fully automated placement algorithms must be used and must result in effective maps.

These maps must allow local officials to readily determine what information exists in TIGER and to enter updates. To make the maps as manageable as possible, the number of map sheets must be held to a minimum by using the smallest practical map scales. The requirement to display high volumes of text items within relatively small areas of the map makes demands that the existing Census Bureau text placement algorithms could not meet.

Over the last three decades, as computers have been applied more and more to the job of making maps, one of the last areas of cartographic skill to be addressed has been the automated placement of feature labels. Even in recent years, with the advent of numerous GIS and mapping packages, the full automation of text positioning still lags behind other components of the mapping process.

Researchers working in this field have pointed out the difficulty of the task (Ahn & Freeman, 1987). Names must be easily read, clearly associated with the feature to which they belong, and avoid overlapping other text (Imhof, 1975). The complexity of rules and algorithms necessary to achieve fully automated names placement have caused many otherwise automated systems to ultimately rely on interactive methods for final text placement.

Much of the research on automated names placement has taken place either in academic institutions or research environments. This has included the work of Ahn and Freeman (1987), Basoglu (1984), Zoraster(1986), and Marks and Shieber (1991). Their work has primarily emphasized the problems of point and areal label placement.

The linear address ranges, however, provide the greatest challenge for the Tiger Improvement Program (TIP) maps. Address information is particularly difficult to place because, unlike most linear feature text, the address data cannot be positioned along a wide range of locations following the feature. The address number must be placed in close proximity to the point where the address break (the last potential address number at an end of a segment) occurs and on the correct side of the street. Since these maps will serve as a window into TIGER, the positioning of the address numbers must effectively convey the address range for each side of each street segment stored in the data base.

The production environment in which automated mapping research is conducted at the Census Bureau provides its own set of constraints. The automated map production system consists of many integrated software components that have been continually refined and supplemented to meet changing cartographic requirements. Limited staff resources and tight time constraints do not allow large scale rewriting of the system. New mapping needs must be met by software that can be readily and effectively incorporated into the existing mapping system within very tight schedules.

Automatic name placement algorithms used at the Census Bureau have followed a non-recursive, sequential approach. Categories of text having the greatest importance for the map would be placed first. Each subsequent name processed is positioned at a location that is not already occupied by earlier placed text. Backtracking capability has not been provided. Once a piece of text had been placed, it could not be repositioned. If a free area, which maintained the visual association of name with feature, could not be found for the new text, the name would either be suppressed or allowed to overlap previously placed text. On large scale maps this was generally not a problem. Smaller scale maps, however, had a higher ratio of unlabeled to labeled features, because of increased text conflict in areas of high feature density (Ebinger & Goulette, 1989). For the TIP maps, a more flexible approach was required.

THE SOLUTION

Unit of Analysis

The first and most important decision to be made in the development of this new algorithm was to choose the algorithm's fundamental unit of analysis. Three possibilities were considered: the segment, referred to as a 1-cell in TIGER, the point or 0-cell, or the polygon or 2-cell.

Using the 1-cell as the unit of analysis meant placing addresses on both ends and both sides of a section of a linear feature in one step. An advantage of this is that the TIGER system retrieves address information one 1-cell at a time. This minimizes the need to store many addresses at once and the need to retrieve address information more than once per 1-cell. The chief disadvantage is that the biggest challenge in placing addresses is avoiding collisions between addresses which are neighbors at a street intersection and thus belong to different 1-cells. The 1-cell method would place one of these addresses without considering the placement of the other.

Using the 0-cell as the unit of analysis meant placing all the addresses around an intersection of linear features at one time. This solved the problem of placing addresses which are neighbors at a street intersection. It lost the ability, however, to deal with the next most important challenge in placement, coordinating the placement of addresses on the same side of the same 1-cell. This is especially important when considering short 1-cells.

The third method, which proved to be the best, is to consider an areal unit to be the fundamental unit of analysis. If each address range end applies to one side of one section of a linear feature, then each can be associated unambiguously with the area of which that section of the linear feature forms one side. All that needs to be done then is to restrict each number to being placed strictly within that area and it becomes possible to consider each such area as the unit of analysis. The only possible collisions that any number need guard against will be with the other numbers associated with that area. If a data structure could be devised that would contain everything about an area and the addresses that needed to be arranged within it, then all the information needed to place these addresses, avoiding any overlaps among them, would be present.

Address Suppression

The exact definition of these areas remained to be established. The smallest possible unit was the 2-cell, which is the fundamental unit of area in the TIGER system. The data and the software would be simplest if the 2-cell was adopted as the fundamental unit of analysis, getting all the data for each 2-cell, placing the addresses, and then moving on to the next 2-cell. The limitation of this approach is that it does not allow 2-cells to be merged to consider larger areas as a unit. In some situations, merging 2-cells can provide advantages for enhanced text placement. Certain classes of 1-cells are not plotted on certain map types and others that are plotted are irrelevant for addressing. When a linear feature is broken by such a feature, there is an opportunity to simplify the addressing for those 2-cells by merging 2-cells together and by suppressing addresses at the intersection of certain 1-cells (see Fig. 1).

In order to be able to merge 2-cells together, several 2-cells must be gathered at one time and searched for 1-cells which can be eliminated. The next highest level of geography that the TIGER system provides is the census block. This is a convenient unit to use as the fundamental unit of analysis because it gives some opportunity for merging 2-cells, but is not so large that the merging process would be too time-consuming and also because the TIGER Improvement Program (TIP) maps display geography which is composed of whole blocks.

The first restriction on this suppression is that the neighboring 1-cells must have the same feature identifier (street name). It is critical for anyone using these maps to study the pattern of addressing to see changes in street names and so addresses will always be printed at these points.

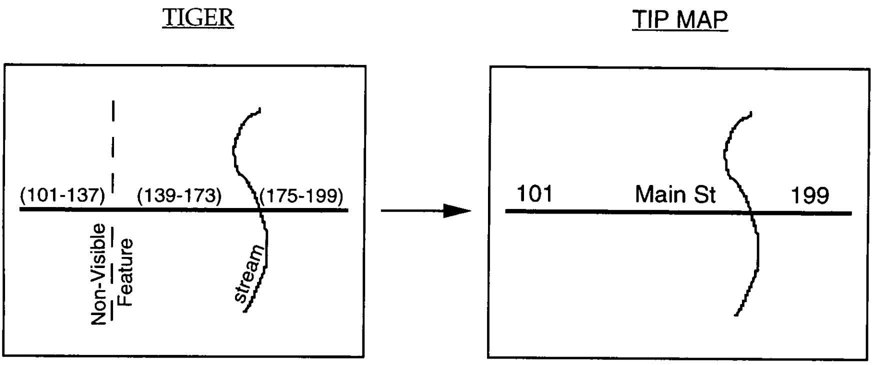


Fig. 1 Address suppression across 2-cells

The second restriction is that there is no addressable feature intersecting at this point on the side under consideration. If a non-addressable linear feature, such as a stream or a railroad track, intersects on the same side as the addresses under consideration, this is not important to the addressing scheme and the address break there may be suppressed. If any linear feature intersects on the opposite side as the addresses under consideration, this is also not a part of the addressing scheme on the other side of the street and it may also be suppressed. A street intersection on the same side, however, will always force the address break to be shown.

Another restriction is that the merged address range must not be misleading. If there is a change in parity (between odd and even address numbers), a change in direction of increase, or a gap of more than two between the adjacent 1-cells, the map user is likely to be misled or confused by looking at the merged range. To let the user see these important shifts in addressing schemes and to not see addresses that are not there, any of these conditions at an address break will force the address break to be shown.

Because the area that the algorithm works on during each pass is a single block, it is impractical to consider suppressing pairs of addresses separated by a block boundary. The final restriction on merging 2-cells and suppressing addresses, therefore, is that addresses are not suppressed across block boundaries.

The restrictions on suppression are therefore:

1) the feature identifier (street name) is unchanged
2) there is no addressable feature intersecting on the side under consideration
3) the address pattern must be consistent between 1-cells with no address gap
4) there is no suppression across block boundaries.

Address Ranges and TIGER

The discussion up to now has assumed that addresses are always simple ranges, with a low number belonging to one end and a high number at the other end. The TIGER system has the capacity to, and often does, store more complicated address information than this. This is because TIGER can store multiple address ranges associated with each 1-cell. This presents another problem. It is already difficult to present simple ranges on a map and yet for the TIGER Improvement Program that these maps support, these variant ranges may be very important. The compromise between displaying all address data as stored in TIGER and simplifying the display to provide map clarity and practical scales, was to merge multiple overlapping ranges into one range for display purposes. Multiple ranges with gaps in the normal two unit number ascending or descending sequence, and those which reverse direction or change parity would not be merged. Instead, a partial range, based on that portion of address information which is contiguous and reflects the same direction and parity, is printed on the map. A special symbol ('+') is appended to each number to show that there is an anomalous address range situation in TIGER that is too complex to display without creating excessive map clutter (see Fig. 2).

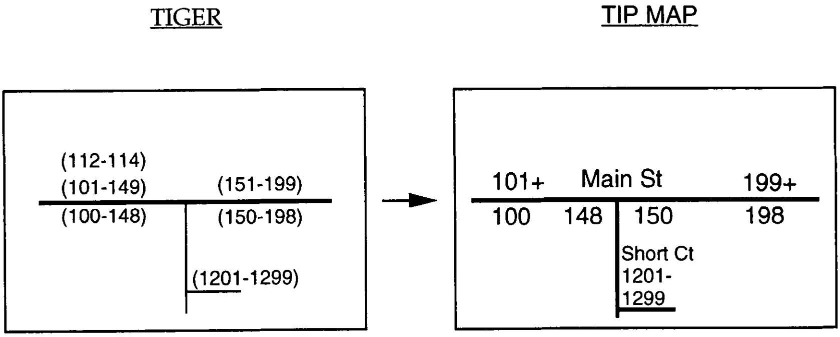


Fig. 2 Display of variant address ranges

Placement

Once all the information about linear features and addresses for a particular sub-area of the map is in place, the algorithm must begin to make decisions about where to place each number. This must be done so that there is a direct, unambiguous visual association between each number and the street side-end that it labels and so that no number overlaps a linear feature or another number (Ahn and Freeman, 1987).

The placement rules that create this visual association are:

1) Place each number as close to the correct end of the street as possible
2) Place each number along a subsegment at least as long as the number
3) Place each number parallel to a subsegment
4) Place each number a standard offset distance from the subsegment
5) Place pairs of numbers adjacent at an intersection in a balanced way
6) When necessary, stack pairs of addresses in a consistent and intuitive manner.

The heart of the algorithm starts with a pair of numbers which are neighbors at a street intersection. Each of the streets is searched starting at the intersection and moving away, searching for a subsegment which is large enough to hold each number. When these are found, a position is calculated for each number along the subsegment so that when the numbers are plotted parallel to and offset from their respective subsegments, the upper left and upper right corners of the boxes that enclose the numbers meet on the line that bisects the angle of intersection.

There are many reasons why the algorithm described above will fail to work for all cases. One is that a street may have no subsegment long enough to hold an entire number. The present implementation of this algorithm will suppress that number. It will be necessary, before TIP mapping begins, to provide a better solution in which the address is placed on the longest subsegment available close to the intersection if its neighbors curve away from it, or to search for a chord of the polygon of sufficient length if the neighboring subsegments curve back toward it.

Another problem may arise if the first subsegments of sufficient length do not allow for placements that do not collide. Currently the algorithm calculates the positions along the infinite lines corresponding to the subsegments where the numbers would go in order to avoid collision. If either of these positions is off the subsegment, the subsegment pair fails and the algorithm will try another pair. The algorithm is set up to try every possible combination of subsegments given the two linear features. If no pair of subsegments will serve, then the present implementation will suppress and the algorithm needs to be improved as described above.

Another problem occurs when the location for a number would position it across the midpoint toward the other endpoint of the 1-cell. Such a placement is likely to result in confusion as to which end of the street an address is associated with, especially if it meets the address from the other end of the street coming its way. The solution for this situation is to stack the addresses in a consistent manner. When stacking, the number which belongs to the left as the addresses are read is placed on top and has a hyphen appended to it (see Fig. 2). The address for the right is placed on the bottom. If the flagging character '+' is required for these addresses, it is appended to the bottom address only.

In many instances there is only one address to be placed at an intersection and this is simpler than the general case described above. In this case, it is still necessary to look for a subsegment long enough to hold the number, but the inside upper corner of the box is offset a small distance from the nearest subsegment, instead of looking for the bisecting line.

Overlap Detection

The operations described above have used geometrical calculations on pairs of rectangles to place address ranges so that they do not overlap. This has the advantage of allowing very precise placements which waste a minimum of space, but cannot be used to detect any collision within a large group of objects because of the excessive time that it would consume. The algorithm also uses a grid-based procedure like the one described by Ebinger & Goulette (1990) for two purposes. One use is as a back-up to ensure that overlaps that the algorithm did not foresee do not happen. After the text has been placed for a block by the procedure described above, the grid is checked to see if any of the text overlaps already placed text. If not, the grid is updated with the new text and the text information is written to a Map Image Metafile (MIM), a text file containing basic graphics commands, from which the mapping program reads it when the complete map is put together. If there is a collision, further adjustment of the text positions will be necessary.

The other use of the grid-based text overlap detection method is to pass information about text positions between different modules of the mapping system. Because of the special challenges of placing address range text, the TIP mapping process has assigned address range placement the highest priority of the map text types and so this algorithm has a clean slate to begin with. Subsequent text placement modules, e.g. block numbering, must have an efficient way to check for overlaps with address range text as they operate. The block numbering module does this by reading the address range MIM, which has already been created, and loading the positions into the grid-based system. The address range algorithm does not do this itself, but it will have to be included whenever a map is created on which address ranges do not have the highest priority of the text types.

FUTURE DEVELOPMENT

One of the most important areas for improvement for the future is to make the algorithm flexible enough to consider the placement of other types of text along with the address ranges. Up to now, the method has assumed that the placement of address ranges has been unquestionably the highest priority. Address ranges could be placed without consideration for other text and then the other text would be placed later to avoid the address ranges. In the application that this has been developed for, in which the placement of address ranges is clearly the most challenging part of text placement, this is satisfactory. In another application, however, it might be important to be able to develop rules more flexibly, with less clear-cut priorities among text types. In this case, it would be necessary to consider other text along with the address range text.

The situation that it would be easiest to adapt to would be if the other text were associated with the same small areas on the map. The notion that each address range number belongs to a definite polygon on the map enables the algorithm to work with blocks and polygons as its fundamental units. If some text type, a block number for example, is also associated with similar areas, the fundamental flow of the algorithm need not be changed and some additional overlap checks and placement steps can be added for each area.

A more difficult case will be to coordinate with the placement of text which adheres to linear features or to larger areas, linear feature names, for example. In such a case, it won't be generally possible to know which of the areas a piece of the linear feature name text will be placed in. All the areas adjacent to the feature will have to be considered first. With linear text labeling features in every direction, it will be necessary to look at every area before deciding on address range and linear feature name placement for any of them.

One possibility would be to attempt to place the linear feature name text in each of the candidate areas, quantify the quality of the resulting placement and then go back and place each where it fit the best. This means that it will not be possible to break text items among segments of linear features that cross areas. Allowing the ability to break up the text and coordinate its placement with the address ranges will add a great deal of flexibility to the placement of linear text, but will also add considerable complexity to the algorithm.

CONCLUSION

The algorithm for the placement of address ranges described in this paper will provide greater success in the display of TIGER data on Census Bureau maps. By increasing the flexibility of the techniques employed for the fully automated positioning of feature labels, these methods have already achieved a more complete display of map information, without resorting to larger scales or increased numbers of sheets. The utility of the maps as a window into TIGER is thus improved and the task of the map user is facilitated.

As the principles involved in the address range algorithms are extended to additional categories of linear, areal, and point labels, even greater success can be expected in producing a well integrated and balanced display of cartographic data.

REFERENCES

Ahn, J., and H. Freeman. 1987. "On the Problem of Placing Names in a Geographic Map." International Journal of Pattern Recognition and Artificial Intelligence, vol. 1, no. 1, pp. 121-140.

Basoglu, U. 1982. "A New Approach to Automated Name Placement." Proceedings Auto Carto V, pp.103-112.

Imhof, E. 1975. "Positioning Names on Maps.", The American Cartographer, vol.2, no.2, pp. 128-144.

Ebinger, L., and A. Goulette. 1990. "Noninteractive Automated Names Placement for the 1990 Decennial Census.", Cartography and Geographic Information Systems, vol. 17, no. 1, pp.69-78.

Marks, J., and Shieber, Stuart. 1991. "The Computational Complexity of Cartographic Label Placement." Harvard University, Center for Research in Computing Technology, Cambridge, Massachusetts.

Zoraster, S. 1986. "Integer Programming Applied to the Map Label Placement Problem." Cartographica, vol. 23, no. 3, pp. 16-27.

An Automated System for Linear Feature Name Placement which Complies with Cartographic Quality Criteria

Mathieu Barrault - François Lecordix
Institut Géographique National - Laboratoire Cogit
2, Avenue Pasteur - BP 68 - 94160 Saint-Mandé France
barrault@cogit.ign.fr

ABSTRACT

Lettering is of prime importance for maps, but positioning is a time-consuming process which may account for up to 50% of the map-editing. This paper adresses the main parameters that lead positioning, and the problems entailed. Then an approach for automated road administrative name-placement is presented which takes into account these problems. We have implemented a research system which follows this approach. The results achieved so far indicate that the problem modeling we used allows to meet high quality requierements.

Keywords : Computational cartography, Optimization, automated name placement.

INTRODUCTION

Lettering is an important medium in cartography. Besides identifying geographic features, text parameters can convey classification (family), hierarchy (police size) and even feature location. But those information depend on the reliability of name text spatial allocation. Actually, this is an expensive task that may represent up to 50% of the map creation process [Yoeli, 72]. Although computer-assisted cartography developments have reduced processing time, only the automation of the label placement could improve on this cost sensibly.

The aim of this paper is both to analyze the main axes which constrain label placement, and to describe a program for automatic road administrative labels placement. The first part presents the three variables which constrain name placement and the problems they raise. After a quick paper review which clearly reveals the preponderance of the point-feature name placement, we'll focus on the problem of labeling linear features, and then bring in a system for automated road labels placement, developed at COGIT laboratory.

THE NAME PLACEMENT PROBLEM

A set of cartographic rules guides name placement [Imhof, 75] [Cuenin, 72]. Some change with the type of feature. Even if they do not provide a rigorous name placement theory, they reflect the three bases the label process leans on.

Three factors

A text attribute does not carry any explicit spatial location. But it cannot be positioned anywhere on a map because of three kinds of conditioners which guide the labeling. The first one is the feature the text designates. The name has to refer to it clearly (distance between the object and its name, text curvature...). Thus, most aesthetics rules depend on the type of feature (cf. figure 1).

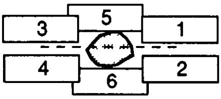


Figure 1 : A model for positional prioritization of point-features. [Cuenin, 72].

The second factor is the cartographic background i.e. every feature already placed and that cannot be moved to help the name to be well positioned. A name must not overlap point-features nor come to close to other significant features. Besides, name placement has to take into account the ambiguities the features of the cartographic background can produce (cf. figure 2).

The last factor is the set of other labels placed on the map. Two names must neither overlap nor be too close. What's more, they must not be evenly spread out nor densely clustered.

Problems of labeling.

Integrating those constraints raises problems we have to face in order to insure a suitable automated name placement under exacting cartographic constraints.

We can group criteria which don't rely on other names positions. They make up the *intrinsic quality* of a position. Cartographic requirements do not establish a strict hierarchy between them. They must be handled simultaneously. To achieve a significant intrinsic quality, we must find an exhaustive list of criteria and define a realistic measure to emulate the cartographer's assessment of each position a name can take. These criteria belong to the first two categories introduced above.

The other problem to face up is the label conflict, a NP-complete problem. What's more, a fine modeling has to manage ambiguities resulting from interactions between names and features. The last difficulty, but not the least, is to devise a system integrating all those without wronging any of them .

Progress in automated name placement

Several researchs have already been undertaken during the last 20 years. Most of them consider point feature names which represent the largest part of labeled features. Hirsch uses spatial search techniques to solve conflicts by computing an overlap vector from which the new position is derived [Hirsch, 82]. In the opposite way, a proposal is to produce a sorted set of possible positions for each feature and to pick up one effective position in each set so that none of them overlaps. This is a discrete approach which reduces combinatorial complexity. The effective position quest has different solutions. Sequential heuristics are proposed [Yoeli, 72] [Ahn & Freeman, 83] [Mower, 86] [Jones & Cook, 91] : each author has his own criteria to sort out features. Then, each feature is labeled : The best allowed position is used and conflicting positions are forbidden. If no position can be selected because none is available (cf. figure 3), there is a backtrack in the feature list to modify a previous selected position so as to free a possible position.

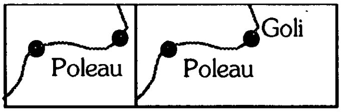


Figure 2 The link between Poleau and its feature depends on the legibility of the other name feature.

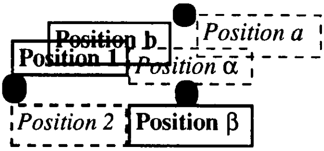


Figure 3 : Two conflictual labels (1,b). Solving this problem may induce another one and start a chain reaction even circular.

Cromley and then Zoraster proposed to consider name placement as a spatial allocation problem [Cromley, 86] [Zoraster, 87] that can be modeled as :

P_{ij} is a quality measure which takes its values in [0,1] where 0 denotes the best position and 1 the worst..
X_{ij}, is a boolean decision variable. $X_{ij}=1$ means the *j*th position is chosen for the *i*th point.
The distribution is modeled as the objective function : $Min\sum_{i}\sum_{j}p_{ij}X_{ij}$

with the constraint set :

Each point symbol demands a label $\forall i, \sum_{j}X_{ij} \geq 1.$

A set *K* of overlapping positions demands at most one position. $\sum_{k\in K}X_{k} \leq 1.$

Chirié and then Lecordix focus on the cartographic quality [Chirié, 92] [Lecordix & al, 94]. Destandau considers the road network [Destandau, 84]. She splits up long roads, needing to be labeled in more than one place, at specific nodes of the network she defines. Jones, who considers road networks too, focuses his works on the named features selection [Jones, 93].
This is a non-exhaustive literature review to introduce the main orientations.

LINEAR FEATURE NAME PLACEMENT PROBLEM

Firstly, the general guidelines listed in the previous section need to be instantiated for two classes of names. Nevertheless, point features name placement is the most investigated domain. In fact, it is the first issue which has been investigated at the IGN (see above), but our current research deals with linear features name placement. [Alinhac, 64] has defined such a set of cartographic rules for linear features.
-The name must be positioned along the linear feature it is labeling.
-It must be placed above the line.
-The label for a line feature should conform to the curvature of the line.
-Strong curvatures must be avoided. On the contrary, straight or almost straight parts must be preferred, preferably horizontal ones, the text being written from the left to the right.
-Small irregularities must be neglected.
-Vertical parts must be avoided. If not, in the left half of the map, the name must be placed on the left side of the line to read upward, and within the right part, it has to be placed on the right side to read downward.
-The name must not be separated from its linear feature by an other feature.
-The test must not cross the feature it is labeling.
-The name must not be spread out , but may be repeated at reasonable intervals along the feature.
-If needed, the name must be repeated along the feature. In this case, it must not be spread out or randomly repeated : The inter-distance should not vary dramatically.

Those rules show up new criteria we'll have to take into account : intrinsic quality criteria such as curvature, word spacing. Also, because of the long size a line can have, which can span across the whole map and because it almost belongs to a network (road network, river network....), the name should better be repeated along the feature. This is a good way to improve map legibility but the name placement process becomes more complex.

The main road network

To support this analysis, here is the description of our first automated name road network placement prototype, elaborated at COGIT laboratory. It is based on three guidelines :

- Road division in sections to label.
- Constitution of a finite set of possible positions (for each section) and computation of the intrinsic quality of those positions.
- Name placement by selecting among possible positions while taking into account intrinsic quality and label conflicts.

Repetition

Some roads are very long. This may cause a lack of legibility. The high density of neighboring features can affect its identification too. If the linear feature belongs to a network, nodes and other edges tend to obscure legibility. To retain the information all along the road, it must be repeated along the line. To process this repetition, a long road will be split into several sections according to specific criteria that must be labeled.

No section must be longer than an explicit size so as to bound the distance between two adjacent names on a road. Density is treated by restricting the number of nodes a section can hold. But this is not enough to comply with cartographic constraints. A crossroad can be embarrassing if sections are not well labeled. Some intersections, defined as a node where two edges of the road go through, are *dubious nodes* (cf. figure 4) because:

1. The two road's edges have different symbolization, inducing that it might be no more the same road. Otherwise :
2. The road crosses another road which is significantly thick. Again, the two edges may appear as belonging to two different roads.
3. Other roads with same legend cross our linear feature in an ambiguous way.

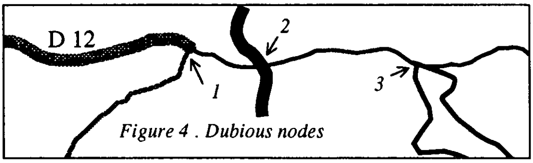


Figure 4 . Dubious nodes

We will call a *road unit*, a part of the linear feature which contains no dubious nodes except its ultimate nodes. Those road units are designed so as to have a reasonable length and to contain a limited set of nodes. In order to alleviate possible ambiguities induced by dubious nodes, every road unit has to be labeled. Labeling each road unit will clarify any ambiguity so that every road is easy to identified. But the map might be too dense. Some dubious crossroads, hence road units, can be neglected :

In fact, if both road units which precede and succeed a given road unit are labeled, then it may be unnecessary to label also this road unit (cf. figure 5).

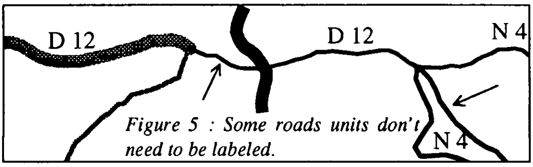


Figure 5 : Some roads units don't need to be labeled.

The main road network on small scale maps yields this kind of situation because of its structure (many tiny roads easily labeled, removing most ambiguities). To keep the message, road units are grouped so as to maximize nodes density and size without stepping over thresholds. These subsets are the *sections* of the roads. Each one of them has to be annotated.

Possible positions

The main road names are administrative labels (N 124, D 12). They are short text associated to artificial linear geographical features hence smooth linear features. Both characteristics allow us to restrict the prospecting of possible positions to straight parts and perform computations by using the bounding box of the label.

Prospecting possible positions provides a set of positions which can be subtituted to a conflicting position. An intrinsic quality [Lecordix & al, 94] must be defined and quantified for each possible position, at first to sort out the best possible positions. This intrinsic quality has to be normalized to permit the comparisons between possible positions of different features.

Restricted to straight parts, possible positions are tested by slipping the bounding box along the main axes of the linear feature. For each tried position, intrinsic quality's criteria are measured.

The main axes are computed with an algorithm issued from [Cromley, 92] based on the least mean square linear regression of the road sections.

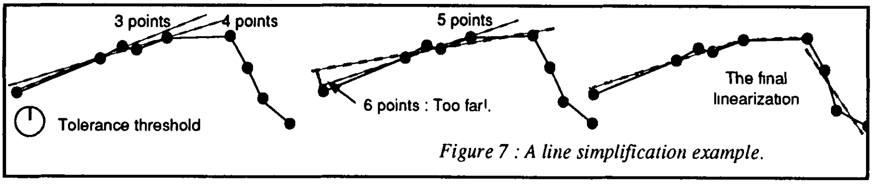


Figure 7 : A line simplification example.

Here are the intrinsic quality criteria for administrative labels.

- The name curvature : We have decided to restrict ourselves to straight parts so that we can use the bounding box of the label. The curvature is summed up to the angle the bounding box makes with the horizontal. This criterion is defined on [-π/2, π/2].
- The axis : Most labels must be written above or below the line. Above is more suitable. Of course, if the name has to be placed on the center-line of the feature, this description does not apply.
- The local center : The more a position is centered between two crossroads, the finer the quality to be. To compute this, the bounding box is assimilated to its center of gravity. This point is projected on the line and its curvilinear abscissa is measured from one of the crossroads. This distance falls in [0, M/2] where M is the curvilinear abscissa of the other crossroad.
- The straight part : Because micro-inflections are neglected, we look for the straigthness of the labeled section. The less chaotic the section, the better the criterion. Assuming that the name and the line falls between S_{min} and S_{max} , this variable is defined in [0, $(l. (S_{max}\text{-}S_{min})$] where l is the width of the bounding box.
- The background concealment : The label must avoid overlapping cartographic background. The more features overlap, the poorest the quality. Some features may even be forbidden. A position overlapping such a feature is called off. If the position does not conceal any prohibited feature, the background removal can be measured. The background map is locally rasterized and each pixel is weighted. The removal is computed by summing the cost of all pixels overlapped by the bounding box of the tried position. This measure is defined on [0, k.(N-1)] where k is the number of pixels in the bounding box and N is the highest possible pixel cost.

Each of those criteria partakes in the intrinsic quality. But, even if the cartographer can privilege some criteria among others, there is no strict theoretical hierarchy between them which could help to model the global intrinsic quality. They must be taken simultaneously and weighted. Another point is that extreme situations must be avoided. The valuation has to enhance these situations, we square each variable for this purpose.
So, for each possible position *j* of the section *i*, we can compute an intrinsic quality P_{ij}, depending on K criteria c_k as: $P_{ij} = \sum_{k=1,..,K} \alpha_k C_k(ij)^2 \text{ with} \sum_k \alpha_k = 1$

$c_k(ij)$ being the estimated value of the *k*th criteria for the *j*th position the *i*th section, normalized on [0,1].

The higher the position quality, the smaller P_{ij}. 0 denotes the best position on a section.

Allocation

We have now a set of sections to label and for each, a sorted set of quantified possible positions. The size of the list must be a compromise between the result's quality and the processing time. We have to find the *effective position* of each section which will be labeled eventually. Actually, the placement of a label must take into account other names to guarantee a suitable legibility. In fact this problem must be considered at two levels. The local level at which we have to consider occurrences of the same name (repetition), and the global level at which all labels may interact.

Local allocation : labels for a specific road must be harmoniously scattered. The distance between any two successive effective positions must be roughly constant. Each label occurrence on the road must have a maximum intrinsic quality while joining in the regular allocation along the linear feature (cf. figure 8).

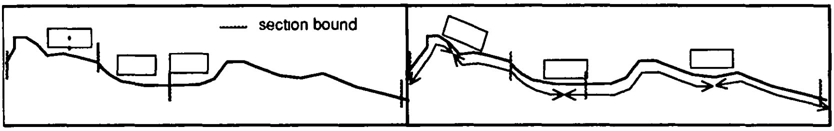


Figure 8 : Contribution of the local allocation.

Distance between two successive positions is calculated by the difference of the two curvilinear abscissae (cf. local center). Extremity labels must be near from the roads end and the others must be positioned so as to be regularly separated along the road while providing the best possible intrinsic quality. For a N-section road, we're looking for a set of effective positions (P_1^e,..,P_N^e) which minimize :

$$\alpha \frac{((d_{ac}(p_1)-\frac{n}{2})^2+\sum_{i=1}^{k-1}(d_{ac}(p_{i+1})-d_{ac}(p_i)-n)^2+d_{ac}(p_k-\frac{n}{2})^2)}{n^2(N+1)} + \beta \frac{\sum_i p_i^2}{N}$$

with n= Length T of the road / number of sections N.
$d_{ac}(p_i)$ the curvilinear abscissa of the *i*th section's tested position .
$\alpha+\beta=1$. A weighted sum to privilege one or the other quality criteria (regularity, intrinsic quality).
A simulated annealing followed by a discrete gradient descent gives a suitable solution.

Global allocation : Cartographic rules compel names neither to overlap each other nor to be too close. The proximity problem is assimilated to an overlap problem through expanding the bounding boxes. The position allocation must reject all overlapping expanded boxes.
Two different approaches have been used to tackle this NP-complete problem. The sequential approach is interesting for linear features because they have usually a lot of

possible positions, and so that backtracking is rarely needed. But on the other hand intrinsic qualities worsen rapidly. Sequential methods are quick but they may yield a non optimal result. Moreover, the backtrack cost is higher with roads because changing an effective position will be passed on other label occurrences of the road, which may start up a chain reaction.

The holistic approach by Cromley & al [Cromley, 86] [Zoraster, 87], is theoretically adequate but the minimization is bound to be solved only through statistical heuristics (such as simulated annealing) because the objective function is neither derivable nor convex. In addition to this problem, taking into account the local allocation may complicate the function and lengthen the processing time needed for reaching a satisfactory solution.

So far, our method is "pseudo-sequential", since roads are processed one by one, while sections of the road are treated simultaneously. In our case, the sequential heuristic is sufficient because conflicts are rare and subsidiary positions handsome.
But in the long run, *global harmony* should also be considered by the name placement process. Currently, it is reduced to a boolean concealment. Actually, to improve name placement, it should involved the same kind of regularity we found in the local allocation, extended to the whole network.

Results
An experiment was performed on an extract of the French road map at 1:1,000,000 scale (Surroundings of Bordeaux - see figure 10). Local allocation is makes use of the simulated annealing method. Roads are processed one by one.
This extract is composed of 632 roads to be named, making 2125 edges.
The segmentation yields 1238 road units, grouped into 865 sections to be labeled.

It takes 2 minutes 30 to run the whole process with a DEC Alpha workstation. Over the whole French territory (26066 edges), it takes 1h40 to label 10187 sections.
The output of the extract was assessed by a cartographer from The Institut Géographique National and 88% of the positions were scored satisfactory. Further experiments will be carried on with other map sheets at other scales.

An automated mileage placement system derived from the method described above, has also been developed at the COGIT laboratory, based on this method [Marrot, 94]. Preprocessing includes the search of the best routes to be labeled. Once again, results are quite excellent with 85% of satisfactory mileages, within 3 minutes 10 of CPU time. Figure 10 illustrates place-names, road labels and mileages as positioned by the different systems developed at the laboratory (Note : Place-names are processed in 2 minutes, 80% of them being well-positioned).

CONCLUSION AND OUTLOOK

Even if lettering is a complex task, these encouraging results show that efficient automation becomes realistic. A common process for positioning different kinds of names shows up through these experimentations. However, the cartographic quality is highly dependent on demanding specifications that vary with the natures of the features.
Research will now focus on longer texts placed along linear features (such as river names...). Complexity increases with the amount of quality criteria and their measurement. Some texts are made up of several words. Besides repetition along the curve, they have to be positioned while taking into account each word's curvature and the space between two words... A modeling of the general problem similar to that described above

(which itself was deduced from [Lecordix & al, 94]) would be welcome. The ultimate aim is to find a global frame for name positioning, with specializations according to the types of names. This would provides with a common processing for names of different natures.

Another issue, besides local positioning, consists in addressing the relative positions between names, which is currently measured only in a boolean way (overlap or not). The limits of this approach are quickly revealed. This layout results into clusters of names which, even if not colliding nor overlapping, are nevertheless unsatisfactory. A finer measure is currently being studied so as to ensure global harmony between name-positions (for a given name category). The global harmony should eventually be extended to all kinds of features over the whole map so that all features may fit harmoniously and aesthetically into the map content.

References

[Alinhac, 64] G. Alinhac, Cartographie théorique et technique (theoratical and technical cartography), fascicule 1, chap. IV, I.G.N. Paris, 1964.

[Ahn, 84] J. Ahn, Automatic Map Name Placement system, Image Processing Laboratory Technical Report 063, Electrical, Computer, and Systems Engineer Department, Rensselaer Polytechnic Institue, Troy, New York, 1984.

[Ahn & Freeman, 83]J. Ahn, H. Freeman, A program for automatic name placement, AUTO-CARTO VI, vol. 2, pp. 444-453, 1983.

[Barrault, 93] Placement automatique des toponymes sur le réseau routier au million, Rapport de stage de DEA, nov. 1993.

[Bonomi, 88] E. Bonomi, J.L. Lutton ,Le recuit simulé, Pour la Science, vol. 129, juillet 1988, pp.68-77, 1988.

[Chirié, 92] F. Chirié, Programme de positionnement automatique des noms de communes, Rapport de stage de DEA, sept. 1992.

[Cook et al, 91] A. Cook, C.B. Jones, J. McBride, Rule-Based control of automated name placement. Proceedings ICA'91 Bournemouth pp. 675-679, 1991.

[Cromley, 86] R.C. Cromley, A spatial allocation anaysis of the point annotation problem, Proceedings, Spatial Data Handling, pp. 39-49, 1986.

[Cromley, 92] R. Cromley, Principal axis line generalization, Computers & Geosciences col. 18, n. 8, p.1003 à 1011, 1992. [Cuenin,72] R. Cuenin, Cartographie générale, Tome 1, pp. 233-245, 1992.

[Destandau, 84] C. Destandau, Essai de positionnement automatique des numéros sur un réseau routier, Rapport de stage de fin d'études, IGN, 1984.

[Fairbain,93] D.J. Fairbain, On the nature of cartographic text, The Cartographic Journal, pp.104-111,Décembre 1993.

[Jones & Cook, 89] C.B. Jones, A.C. Cook, Rule-based name placement with Prolog, AUTO-CARTO 9, pp.231-240, 1989.

[Hirsch, 82]	S.A. Hirsch, An algorithm for automatic name placement around point data, The American Cartographer, vol. 9, No. 1, pp.5-17, 1982.
[Imhof,75]	E. Imhof, Positioning Names on Maps. The American Cartographer, vol. 2, pp.128-144, 1975.
[Jones, 93]	C.B. Jones, Placenames, cartographic generalisation and deductive databases, NCGIA, october 93.
[Lecordix & al, 94]	F. Lecordix, C. Plazanet, F. Chirié, J.P. Lagrange, T. Banel, Y. Cras, Automated name placement on map under high quality cartographic constraints, EGIS'94, vol.1, pp. 22-32, 1994.
[Marrot, 94]	J.M. Marrot, Positionnement automatique des kilométrages, Rapport de stage de DESS, sept. 94.
[Mower, 86]	J.E. Mower, Name placement of point features through constaint propagation, Proceedings Second International Spatial Data Handlings, pp. 65-73, 1986.
[Mower, 93]	J.E. Mower, Automated Feature and Name Placement on Parallel Computers, Cartography and Geographic Information Systems, vol. 20, n. 2, pp. 69-82, 1993.
[Yoeli, 72]	P. Yoeli, The logic of automatic map lettering, The cartographic journal, vol. 9, no. 2, pp. 99-108, 1972.
[Zoraster, 87]	S. Zoraster, Practical experience with a map label placement program. Proceedings, Auto-Carto VIII. pp.701-707, 1987

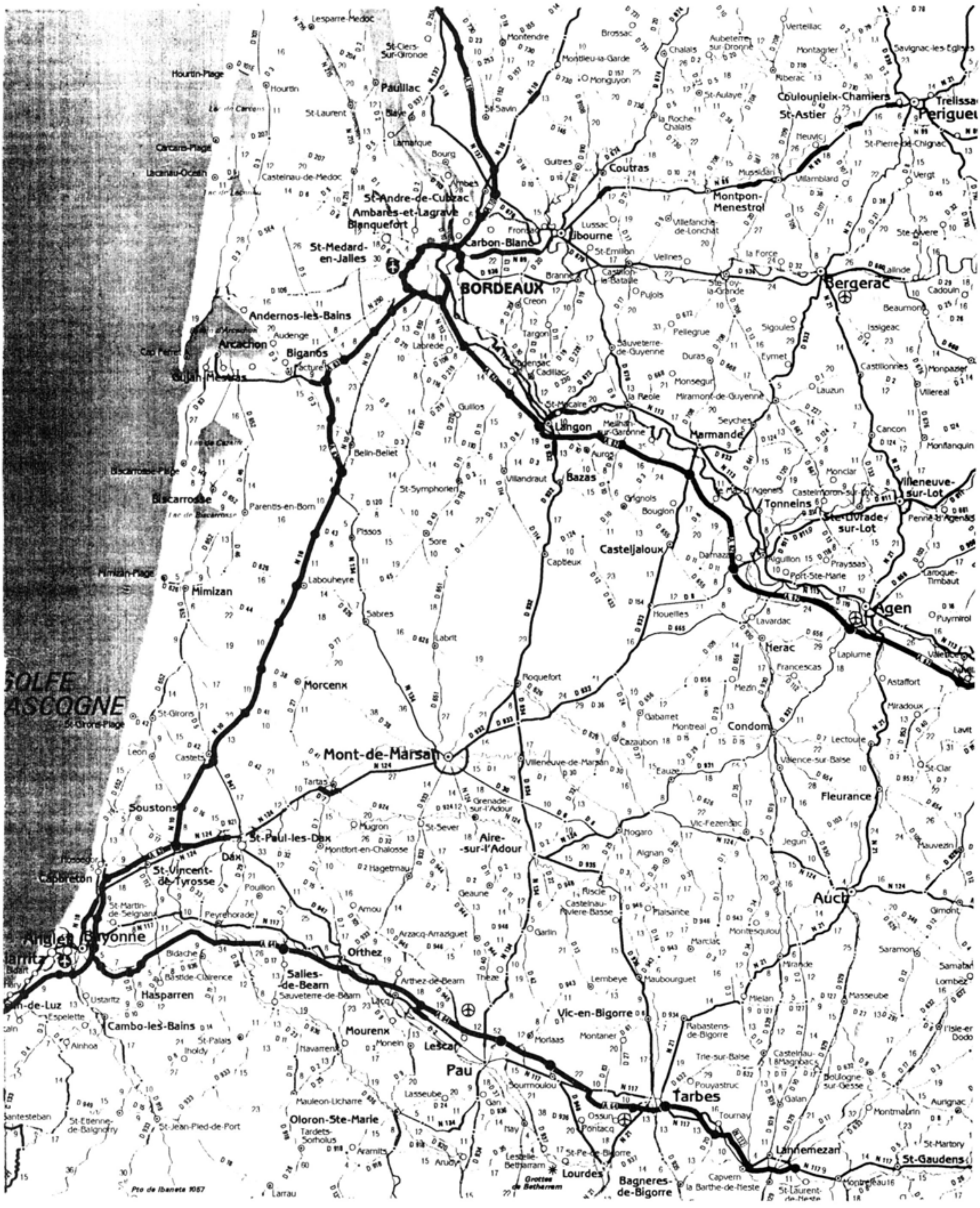


Figure 10 : Automated road label placement on an extract of the French road map at 1:1 000 000.

COMPRESSION OF SPATIAL DATA

Roman M Krzanowski
NYNEX Science & Technology
500 Westchester Ave.
White Plains , NY

ABSTRACT

The objective of this study was to quantify the compressibility of selected spatial data sets: USGS DEM 3 arc-sec, ETAK, and TIGER/Line. The study serves several purposes: it provides the detailed description of the structure of spatial data sets from the perspective of the compression process (using n-gram statistics and entropy); compares the effectiveness of different compression methods (using compression rates), and provides the recommendations on the use of compression methods for the compression of spatial data for both UNIX and DOS operating systems. Three main conclusions are reached in this paper: the compression rates for spatial data sets may be predicted from their entropy; the compression rates for a given type of spatial data remain stable for different instances of those data (exception are DEM data); and currently available compression programs can achieve between 80 percent and 90 percent compression rates on spatial data.

INTRODUCTION

This paper presents a comprehensive study of the compression properties of spatial data. The size of spatial data sets quite frequently exceeds 10 megabytes (Mb). Transfer of such data in large volumes requires either high capacity storage media or high capacity data networks. In either case, the transfer of data is greatly enhanced if the transferred data are efficiently "packed" (that is compressed) (Storer, 1992). Efficient packing of data requires knowledge of data packing properties which are very data dependent * . A review of the literature has revealed that most of the compression studies reported have been concerned with the compression of binary images, sound, voice, or textual files (Held, 1991; Nelson, 1991b; Nelson, 1992; Storer, 1988; Welch, 1984). To the author's knowledge no systematic study of the compression of spatial data has been published. This paper is an attempt to provide such information .

The results of this study may be used in planning data networks, designing of distributed data bases; planning storage space requirements for spatial data bases; and defining of the requirements for secondary storage media. The intended audience includes vendors of spatial data, and any federal, state or local, agencies dealing with spatial data transfer, storage, and distribution.

The next section of this paper introduces fundamental concepts and definitions related to data compression and also reviews modern compression algorithms. Then, findings of a series of experiments establishing the compression characteristics of selected spatial

* "It is clear that the performance of each of ... (compression) ... methods is dependent on the characteristics of the source..." (Lelewer & Hirschberg, 1988, p.288).

data are presented. Final sections review the results of those experiments, and also provide guidelines for the selection of compression methods for spatial data.

FUNDAMENTAL CONCEPTS

Basic definitions

Data compression is the process of encoding a set of data D into a smaller set of data ΔD. It must be possible to decode the set ΔD into the set D -- or to its approximation. Compression methods can be "lossless" and "lossy". "Lossless" methods compress a set of data and thereafter, decompress it into exactly the same set of original data. "Lossy" data compression converts the set of data into a smaller set from which an approximation of the original set may be decoded. "Lossy" data compression is used for the packing of images, speech, or sound, and is appropriate for the compression of data when their accuracy can be compromised. As the accuracy of the spatial data cannot be sacrificed in the compression process, "lossy" compression methods are not suitable for the compression of spatial data as defined in this study. The reminder of this paper will concentrate on "lossless" compression methods.

In this paper a data set processed by the computer is synonymous with a source message. A source message is composed of words from a source alphabet* . Computer processing of a data set involves the process of coding of a source message. Coding is a mapping of a source message (words from a source alphabet) into the code words (words from an code alphabet). A simple example of coding is the replacement of letters from an English alphabet by the 7-bit ASCII code from $\{0,1\}$ alphabet.

The measure of the efficiency of coding (or the message information content) of a message (word) a_i in bits is $-\log_2 p(a_i)$ where $p(a_i)$ is a probability of occurrence of the message a_i in the source message. The average information content of a source message is called entropy (H) and is expressed as:

$$H = \sum_{i=1}^{n} [-p(a_i)\log_2 p(a_i)] \qquad (1)$$

In terms of the coding efficiency, the entropy gives the lower bound on the number of bits necessary to encode the source message (Lelewer & Hirschberg, 1987). Number of bits in the coded message above its entropy is called the redundancy. The amount of redundancy (R) in the message is expressed as:

$$R = \sum p(a_i)l_i - \sum [p(a_i)\log p(a_i)] \qquad (2)$$

where l_i is the length of the code word representing the message a_i .

In most cases, uncompressed coding of data creates redundancy** . The most obvious form of redundancy are repeated patterns of words in the source message. Those patterns are called grams. The existence of repeated patterns (a_i) in the message is determined by

* Source message A is an ensemble of messages (words) $a_i \in A, A = \{a_i,\ldots,a_n\}$.

** Welch (1984) distinguishes four types of redundancy that affect compression: character distribution, character repetition, high-usage patterns, and positional redundancy.

the compilation of "dictionaries" of the patterns with their frequencies (probabilities of occurrence - $p(a_j)$) *. Patterns or grams may be of 1,2 ,3 or higher order representing one, two, three or more character patterns. 1-order (also called 0-order) or 1-grams ignore the dependency of a pattern on preceding or following words. For any meaningful source message this is an unrealistic assumption. Yet, despite their simplicity, frequencies (and entropy) of 1-order patterns provide valuable information about the source message (to be demonstrated later). The grams of order 2 (2-grams),3 (3-grams) or higher provide frequencies (and entropies) of the longer patterns. Longer patterns quantify the dependency between preceding and following words. Figure 1 demonstrates some of redundancies encountered in spatial data files.

The overall efficiency of any compression method is measured by its compression rate. There are several measures of the compression rate (Lelewer & Hirschberg, 1988; Nelson, 1991b) **. In this paper the compression rate is expressed as a ratio of the size (in bytes) of compressed to uncompressed source message:

$$c = \frac{fs}{fsc} * 100\% \qquad (3)$$

where ***fs*** is a size of an uncompressed source message in bytes, ***fsc*** is a size of a compressed source message in bytes, and ***c*** is the compression ratio (Nelson , 1991b).

Compression algorithms and their implementations

Compression is a two step process consisting of modeling and coding (Hirschberg & Lelewer, 1992; Nelson, 1992). Modeling step creates the representation of the source message. The coding step generates the compressed message based on the model. Models may be either statistical or dictionary based.

In "statistical" modeling, the frequency of occurrence of words in the source message is first calculated. Then, using this frequency the new codes that use fewer bits than the original codes are assigned to words. Compression is achieved by the difference between the size of the original code words and the new code words. An early example of the statistical coding is Morse's code optimized for the English language. The important parameter of the compression based on statistical models is the order of the grams for which the frequency of occurrence is calculated. A detailed explanation of statistical modeling methods is offered in Lelewer & Hirschberg (1988). The implementation details are also explained in Nelson (1992).

In dictionary-based methods the coding does not produce the smaller code words but it produces pointers to the patterns encountered in the source message. The source message is scanned using the "window" of predefined size. A dictionary of patterns in the window is created as the source message is scanned and pointers to those patterns are inserted in the coded message. Each pointer contains the index of the pattern in the dictionary and the first character not in the pattern. The dictionary of patterns is updated as the new patterns are scanned till the maximum size of the dictionary or patterns is reached. Compression is achieved when frequent long patterns are substituted with shorter pointers

* Source message with purely random patterns cannot be compresssed (Strorer 1988).

** Compression rate is a compression yielded by a coding scheme. In addition to the measure adopted in this study compression rate may be measured by the ratio of the average message length to the average code word length (Lelewer & Hirschberg, 1988).

to dictionary entries. The important parameters of the dictionary based methods are: size (in bytes) of the scanning window, the size (in bites) of the index, and the size of the largest pattern held in the dictionary. The detailed explanation of algorithms of dictionary compression methods is offered in Lelewer & Hirschberg (1988). The implementation details are explained in Nelson (1992).

Both statistical and dictionary-based compression methods can be either static or adaptive (dynamic). Static methods do not adjust their parameters to the type of the source message; adaptive methods change their parameters in the response to the properties of the processed message.

The statistical compression methods use Huffman, Shannon-Fano, or arithmetic coding (Held, 1991; Lelewer & Hirschberg, 1988; Howard & Vitter, 1992; Nelson, 1991a) and are uncommonly the primary coding algorithms found in the commercial compression software. Rather, dictionary-based compression methods, which use LZ77 or LZ78 algorithms or their derivatives (LZSS or LZW) are employed in the creation of most of the modern compression software (Nelson 1992) * .

Compression methods based on statistical models are limited by the size of the model of the source message that increases with the order of the model. Huffman based coding methods loose efficacy as they use only whole bits for code words (frequently, information may be represented by a fraction of a bit - in a sense of the information content) (Hirschberg & Lelewer, 1992). Arithmetic methods, while very efficient, require large computer resources and are generally very slow (Nelson, 1992). The most efficient compression methods currently used are dictionary-based, an observation which is confirmed by the prevalence of those methods in most commercial implementation (some drawbacks of those methods are explained in Lelewer & Hirschberg (1988)).

Compression rates achieved by compression programs may be as high as 98 percent (Lelewer & Hirschberg, 1988) for the source specific compression programs. On average, the reported compression rates for English texts range from 50-60 percent (Lelewer & Hirschberg, 1987); 40 percent for Huffman based compression (Nelson, 1992); and 40 - 60 percent for dictionary based methods. As noted earlier, no systematic analysis of compression rates was reported for spatial data.

COMPRESSION TESTS

Methodology

Compression algorithms. This study evaluated the commercial compression packages listed in Table 1 as well as generic compression methods listed in Table 2. Compression packages COMPRESS, GZIP, ARJ, PKZIP are available either as the part of OS, from Internet sites (GZIP, 1993; PKZIP, 1993), or through software vendors. Compression software listed in Table 2 is available in Nelson (1992). Compression methods based on statistical modeling (Huffman; Arithmetic coding) have been tested for comparison purposes only. No current production compression software uses either as their primary compression method.

* A different taxonomy of compression algorithms has been proposed by Lelewer & Hirschberg (1988).

COMPRESS utility implements LZW algorithm based on LZ78 which is an extension of LZ77 algorithm (Nelson, 1992; Welch, 1984). GZIP is a variation of LZ77 with the elimination of duplicate strings and use of Huffman coding for compression of indexes (GZIP 1992). ARJ, PKZIP both implement LZ77 based algorithms. Huffman compression, adaptive Huffman compression, and arithmetic compression programs tested in this study are based on 1-order models. LZSS program tested in this study is an implementation of LZSS extension to LZ77 with 4096 -byte window size, 12 bit index to the window, and up to 17 bytes of the coded pattern length (Nelson, 1992). LZW program tested here uses 12 bit fixed code length and is an extension of LZ78 (Nelson, 1992; Welch, 1984).

Spatial data. Spatial data sets tested in this study are listed in Table 3. The selected spatial data usually constitute the fabric of the land information systems. Some of these spatial data (USGS DEM *) are already available on Internet, others (ETAK ** , TIGER***) are distributed on CD-ROMs.

Statistical Measures. In this study the compressibility of the data sets was assessed using the entropy defined by formula (1) , redundancy defined by the formula (2), and the frequency of grams for f 1-, 2-, and 3-, orders . The compression rates were evaluated using the formula (3). The compression rates for Tiger/line files (T1,T2,T3) were averaged for all of the files in the set (records 1 to 8, and a to r).

Results of experiments

The results of the analysis of the structure of spatial data - entropy of 1-,2-,and 3- order, the statistics of three most frequent 1-order grams, and the redundancy - are given in Tables 4, 5, and 6 respectively. Table 7 reports the compression rates achieved with the selected compression methods and calculated using the formula (3) .

CONCLUSIONS

The following observations can be made about the compression properties of the tested spatial data sets:

- Compression rates for spatial data are above those of English text, or program files;

- entropies and redundancies of the same order, for a given type of spatial data, are similar (see Table 4 and Table 6);

- average entropies of order 1, 2, and 3 (Table 8) for ETAK and TIGER/Line data sets are either above (1-gram) or below (2-gram, 3-gram) entropy of DEM data sets. This suggests the similar coding structures and coding efficiencies of ETAK and TIGER /Line;

* Format of USGS DEM 250 data sets is defined in Digital Elevation Models, (1992), Data Users Guide 5. US Department of the Interior, U.S. Geological Survey, Reston, Virginia.

** Format of ETAK data is described in ETAK (1993), MapBase File Definition, File Format version 2.0-2.2, ETAK, The Digital Map Company.

*** Format of TIGER /Line data is defined in TIGER/Line Precensus Files,(1990), Technical Documentation, U.S. Department of Commerce, Washington, D.C.

- the most frequent 1-gram (Table 5) in tested data sets is a white space character (ASCII 32), it constitutes from 44 to 57 percent of characters in the data sets (D3 excluded);

- the statistics of 1-grams are very similar for all of the tested data sets. Most of the grams have a frequency below 4 percent (i.e. the most frequent gram has a frequency above 40 percent, the next one has a frequency above 4 percent, and the rest has a frequency below 4 percent);

- a data format that maps the actual locations in space to its file format (DEM) has a lot of redundancy. This is reflected in the high compression rates for DEM data set.

The following observations can be made about the compression methods evaluated in this project (Table 7):

- Compression methods based on statistical modeling (Huffman, Adaptive Huffman, Arithmetic coding) are inferior to dictionary-based methods (LZ77 and LZ78 and their derivatives). Dictionary-based compression methods demonstrated from 10 percent to 20 percent greater compression rates;

- regardless of the packing method studied, the compression rates for a given spatial data type do not vary significantly (the sole exception being DEM data);

- commercial, dictionary-based compression methods yield the compression rates above the redundancy calculated from 3- grams.

RECOMMENDATIONS AND SUGGESTIONS

The following presents recommendations for the compression of spatial data, utilization of compression programs for the packing of spatial data and concludes with suggestions for further research in this area:

- Knowledge of the n-order entropies and related redundancies may be used for the prediction of the compression rate for a given data type: compression rates with compression methods based on the statistical model are usually close to the 1-order redundancy, compression rates with compression methods based on dictionaries exceed the 3-order redundancy by 10 to 15 percent (Table 9 and Table 10);

- when using commercial dictionary-based compression methods one may expect compression rates of spatial data sets to be from 84 percent to 90 percent. In rare cases (for specific types of formats) it may be as high as 99 percent;

- the best compression rate in this study was demonstrated by GZIP compression software for UNIX, and ARJ and PKZIP compression software for DOS;

- future studies should be carried out into the time aspect of the data packing process as the amount of time taken by the compression program varies significantly from an implementation to another (time was not recorded in this study except when it exceeded an arbitrary limit of 30 minutes - see Table 7);

- future studies into the compressibility of spatial data sets should concentrate on the analysis of longer than 3- grams, and on the specific features of spatial data (use of floating point numbers, absolute coordinate system);

- observations and conclusions reported in this study necessarily reflect the amount and type of data tested, studies should be carried out on the larger amount of spatial data sets before findings of this study could be safely generalized.

REFERENCES

GZIP 1993 , Algorithm documentation, ftp.cso.uiuc.edu.

Held, G. 1991 , Data Compression, New York: John Wiley & Sons, Ltd.

Hirschberg, D.S. & Lelewer, D.A. 1992 , Context Modeling for Text Compression, in Image and Text Compression, J.A. Storer (ed), Kluwer Academic Publishers: London, pp.113-144.

Howard, P.G. & Vitter, J.S. 1992 , Practical Implementations of Arithmetic Coding, in Image and Text Compression, J.A. Storer (ed), Boston: Kluwer Academic Publishers, pp. 85-109.

Lelewer, D.A. & Hirschberg, D.S. 1988 , Data Compression, ACM Computing Surveys, 19(3), pp. 261-296.

Nelson, M.R. 1991a , Arithmetic Coding and Statistical Modeling, Dr.Dobb's Journal, 2, pp. 16-29.

Nelson, M. R. 1991b, Data Compression Contest Results, Dr. Dobb's Journal, 11, pp. 62-64.

PKZIP 1993, Algorithm documentation, ftp.cso.uiuc.edu.

Nelson, M. 1992, The Data Compression Book, New York: M&T Publishing, Inc.

Storer, J.A. 1988, Data Compression, Methods and Theory, Maryland: Computer Science Press, Inc.

Storer, J.A. 1992, Introduction, in Image and Text Compression, J.A. Storer (ed), Boston: Kluwer Academic Publishers, pp.v-viii.

Thomas, K. 1991, Entropy, Dr. Dobb's Journal, 2, pp. 32-34.

Welch, T.A. 1984 , A Technique for High Performance Data Compression, Computer, June, pp. 8-18.

```
NOGALES - W (a)____________AZ          NH12-02W  1    1    0    0   0.0
 0.0             0.0             0.0        0.0             0.0             0.0             0.0
0.0             0 0             0.0             0.0             0.0             0.0
0.0
   3    2    4  -0.403200000000000D+06   0.111600000000000D+06  -0.403200000000000
D+06   0.115200000000000D+06  -0.399600000000000D+06   0 115200000000000D+06  -
0.399600000000000D+06   0.111600000000000D+06   0.570000000000000D+03  (b)
0.235600000000000D+04   0 0    10.300000E+010.300000E+010.100000E+01    1  1201
  1    1  1201    1  -0 403200000000000D+06   0 111600000000000D+06   0 0
 0 570000000000000D+03   0.105300000000000D+04   575(c)   577  576   576 (b)  575   575
 575  (c)574  574  574  573  573  573  573  573  573  573  573  573  573  573
  574  574  575  575  577  579  578  577  577  576  576  575  575  574
 574  573  573  572  572  571  571  571  570  570  570  570  570  570  570
```

Figure 1. Example of redundancies in Spatıal Data file (DEM) - (a) white spaces; (b) use of numbers (limited alphabet); (c) repeated patterns.

Compression Software	Operating System	Compression method
COMPRESS*	UNIX	Dictionary-based (LZW)
GZIP*	UNIX	Dıctionary-based (LZW)
ARC,PKARC	DOS	Dictionary-based (LZ78)
ARJ*	DOS	Dictıonary-based/ (LZ78) Huffman
LHarc	DOS	Dictionary-based
PKZIP*	DOS	Dictionary-based (LZ78)

(*)- software tested in this study.

Table 1. Most common compression programs and related compression algorithms.

COMPRESSION ALGORITHM	COMPRESSION METHOD
Huffman	statistical
Adaptive Huffman	statistical
Arithmetic Codıng 0	statistical
LZSS	dıctionary-based
LZW	dictionary-based

Table 2 . Compression algorithms tested in this study.

CODE	DATA TYPE	DATA FILE	SIZE [byte]
D1	USGS DEM	Nebraska-w	9,840,640
D2	250 3arc-sec	Nogales-w	9,840,640
D3		Noyo-canyon-e	9,840,640
E1	ETAK	dnv_co.mbs	80,253,440
E2		ftc_co.mbs	41,170,688
E3		mhn_cy.mbs	19,004,772
	TIGER/line	state 36	
	census		
T1	county 003	record 1...r	9,197,728
T2	county 071	record 2. .r	22,395,596
T3	county 111	record 3...r	15,296,530

Table 3. Data sets used in the study and their respective sizes in bytes.

DATA SET	1-gram	2-gram	3-gram
E1	2.66	4.31	5.77
E2	2.64	4.31	5 75
E3	2.78	4.49	5.95
D1	2.28	4.05	5 42
D2	2.59	4.89	6.85
D3	0.63	1.22	1.75
T1	2.78(1.94-3.4)*	4.20(1.94-3.40)	5.61(4.15-7 3)
T2	2.77(2.0-3.5)	4.38(2.0-3.51)	5.52(5.0-7.51)
T3	2.74(2.0-3.5)	4.35(4.0-5.7)	5.51(4.9-7.5)

(*)- for TIGER line files the reported entropy of grams is the average entropy of grams for all of the files in the given data set and the minimum and maximum in the set.

Table 4. Entropy of 1- 2- and 3-grams for tested data sets.

DATA SET	1-grams
E1	44%(32) 24%(48) 4%(49)*
E2	42%(32) 26%(48) 4%(49)
E3	40%(32) 25%(48) 5%(49)
D1	54%(32) 13%(52) 4%(53)
D2	47%(32) 12%(49) 4%(50)
D3	84%(32) 15%(48) 7%(41)
T1	58%(32) 6%(48) 4%(51/50)**
T2	55%(32) 6%(48/45) 4%(51)
T3	57%(32) 7%(49) 4%(48)

(*) - percentage of 1-grams(ASCII code of 1-gram); (**) - statistics for record type 1.

Table 5. Statistics of the most frequent 1-grams

DATA SET	1-order	2-order	3-order
E1	66%	73%	76%
E2	67%	73%	76%
E3	65%	72%	75%
D1	71%	74%	77%
D2	67%	69%	71%
D3	92%	92%	92%
T1	65%	73%	76%
T2	65%	72%	77%
T3	65%	73%	77%

Table 6. Redundancy for tested data sets for 1-, 2-, 3-order entropy.

DATA SET	COM PRES S	GZIP	HUF	AHUF	AR-0	LZSS	LZW	ARJ	PKZI P
E1	84%	88%	f*	f	f	78%	72%	88%	88%
E2	84%	88%	f	f	f	78%	83%	88%	88%
E3	84%	89%	f	f	f	80%	74%	88%	88%
D1	89%	91%	70%	71%	73%	81%	78%	91%	91%
D2	81%	82%	66%	67%	70%	71%	67%	85%	87%
D3	99%	99%	86%	86%	92%	88%	99%	99%	99%
T1(**)	85%	90%	65%	66%	65%	79%	80%	90%	90%
T2	84%	90%	64%	64%	64%	72%	77%	90%	90%
T3	85%	90%	64%	64%	64%	79%	79%	90%	90%

(*) f - failed to compress in 30 min; (**) - the average for all files in the set; HUF- Huffman compression order 0; AHUF- adaptive Huffman compression order 0; AR-0 - arithmetic compression order 0; LZSS - LZ77 based; LZW - modified LZ78;

Table 7. Compression rates for tested algorithms and methods.

	E	D*	T
1-gram	2.69	2.43	2.76
2-gram	4.37	4.47	4.26
3-gram	5.82	6.13	5.53

(*) - D3 excluded; E - ETAK ; D - DEM; T - TIGER .

Table 8. Average entropy for tested data sets.

SPATIAL SIMPLIFICATION OF INPUT DATA FOR HYDROLOGIC MODELS: ITS EFFECT ON MAP ACCURACY AND MODEL RESULTS

Casson Stallings
Casson_Stallings@ncsu.edu
Computer Graphics Center, Department of Forestry

Siamak Khorram
Khorram@ncsu.edu
Computer Graphics Center, Departments of Forestry and of
Electrical and Computer Engineering

Rodney L. Huffman
Huffman@eos.ncsu.edu
Department of Biological and Agricultural Engineering
North Carolina State University
Raleigh, NC 27695

ABSTRACT

Statistical generalization can be used on map data to reduce the complexity of the data before analysis. The generalization scheme tested here calculates the dominant cover class (mode) of a theme within specified areas. In this work, the specified areas are agricultural fields, and the themes are soils, slope, and aspect. The purpose of this study was to assess the impact of using a vector-based mode amalgamation scheme on the map polygon count, map accuracy, and hydrologic model results at field and watershed scales.

The Groundwater Loading Effects of Agricultural Management Systems (GLEAMS) model was initially run on fields using inputs based on the original coverages describing soils, slope, aspect, and field boundaries. Polygons representing one or more themes were amalgamated and the simulations were then repeated. The map accuracy of the amalgamated themes were calculated. The original model results were compared to those based on the amalgamated data.

The amalgamation technique decreased the number of polygons, but introduced substantial map errors. The soils amalgamation introduced error into most measures, but more so at the field scale than the watershed scale Amalgamations of the slope and aspect themes had less effect on the model outputs investigated. We concluded that the aggregation method and extent should depend on the model being used and on the modeling objectives.

INTRODUCTION

In using vector-based GIS it is often necessary to intersect multiple themes. Each new intersection typically increases the number of polygons in the combined coverage by many more than the added theme contained. It is likely that most vector-based GIS-model interfaces must deal with this problem as they combine information on soils, slope, watersheds, land use, and other themes. Raster-based interfaces, if they are to handle homogenous regions and use a small cell size, encounter the identical problem

Stallings *et al.* (1992) described the GIS Interface for Ground-Water Models. That GIS-model interface used multiple layers of detailed map data over a small watershed. The interface is designed to be used over larger areas, however. Application in this manner leads to large amounts of data and many polygons to process. For this and other vector-based modeling applications, each polygon must be processed individually, so that the processing (CPU) time is approximately proportional to the number of polygons. Decreasing the number of polygons required to represent the geographic data decreases the general computational effort, but at the cost of reducing the accuracy of the analysis.

Map generalization techniques can be classified according to their purpose (cartographic, statistical), domain (spatial, thematic), and data structures processed (raster, vector). Map generalization is described and reviewed by Monmonier (1983), Brassel and Weibel (1988), McMaster and Monmonier (1989), and McMaster and Shea (1992). The generalization technique investigated in this study is a statistical vector-based polygon amalgamation technique that affects both the spatial and thematic data. It is used to reduce the polygon count of the final coverages, decreasing the resulting processing effort

McMaster and Shea (1992) describe two generalization methods appropriate for reducing the polygon count of a vector-based coverage: classification and amalgamation. Classification (as described by McMaster and Shea) consists of redefining the thematic attributes into fewer classes. This can then be followed by an operation to join adjacent polygons that are no longer distinct. 'Amalgamation' is the general term McMaster and Shea use to describe the merging of adjacent polygons. Typically this is done in one of two ways. The first is to eliminate--by merging with adjacent polygons--all polygons below a set size threshold (e.g., ELIMINATE in Arc/Info, ESRI, 1992). This technique was originally used to reduce the number of slivers in intersected coverages The second is to merge adjacent polygons that share some attribute (e.g., DISSOLVE in Arc/Info, ESRI, 1992).

<u>Map Simplification Experiences</u> The effects of the elimination technique on map error have been investigated by Wang and Donaghy (1992). In general the increase in map error was proportional to the increase in the size threshold used. However, at very small size thresholds, many polygons could be eliminated with very little error. A similar trend was found in spatial simplification described in Stallings *et al.* (1992) In that simplification the size threshold was based on polygon sizes within agricultural fields. A one-percent threshold would result in each polygon that consisted of less than one percent of the field area being merged with an adjacent polygon This assured that despite their small size, the relatively large (percentage wise) polygons within small fields would not be decimated in the simplification process.

Most investigations of map generalizations and their effect on GIS-model results have looked at grid-cell size changes. Whede (1982) quantified the effect of cell size on map errors. He found that map errors, and the variability of the map errors, increased as cell size increased. Brown *et al.* (1993) investigated generalization of input data for the agricultural non-point-source pollution model (AGNPS). They varied grid-cell size and looked at the errors in land-cover distributions, predicted erosion, and deposition summaries and areas; they then related the errors to the fractal dimension of the inputs and to the semi-variograms of the inputs. The authors found significant errors in the model results as the grid-cell size increased beyond 120-180 meters on a side One of their conclusions was that cell size should be no larger than the lag distance of the shortest semi-variogram range. In their hydrologic modeling study, Srinivasan and Arnold (1994) used a mode-based amalgamation technique for the soils data within each subbasin. The subbasins were defined using the r.watershed program in GRASS. Within each subbasin, the mode (or dominant) soil class and land-use attributes were determined and used for simulation of the entire subbasin. The effect of this amalgamation on the model results was not tested.

Purpose and Approach The purpose of this study was to assess the impact of using a vector-based mode amalgamation scheme (Figure 1) on the map polygon count, map accuracy, and the hydrologic model results at field and watershed scales.

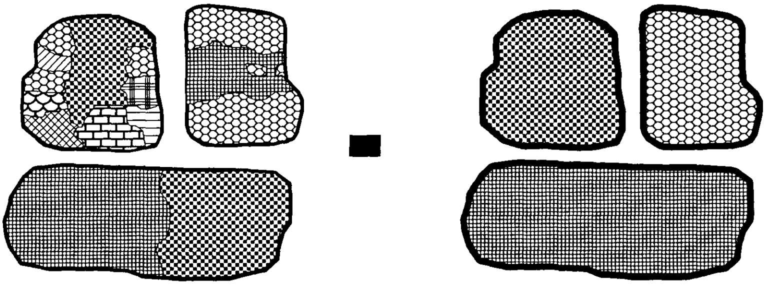


Figure 1 Application of the mode amalgamation technique to three regions (thick lines) Each class is represented by a separate shading pattern After the amalgamation each region contains only one class

Polygon coverages representing agricultural fields, soils, slope, aspect, and other pertinent factors were created and intersected to create a fundamental unit coverage (Figure 2). Each polygon in this coverage is locally unique with respect to all the pertinent attributes (e.g , field, soil, slope, aspect). A set of fields was selected at random from the study area. The coverage was simplified using mode amalgamation within each agricultural field. The amalgamations used were (1) slope, (2) aspect, (3) slope and aspect, and (4) soils, slope, and aspect (Figure 2). An amalgamated theme was reduced to one measure for each field polygon by assigning to all polygons in the field the mode (or dominant) class within the field and then merging these polygons (Figure 1). These coverages were incorporated into the GIS Interface for Ground-Water Models (Stallings *et al.*, 1992). GLEAMS simulations were run on each field using the fundamental unit coverage and four simplified versions of the coverage

The unsimplified fundamental unit coverage was used as a benchmark and is referred to as either the standard coverage or the standard fundamental unit coverage. The standard runs used this coverage to derive simulation inputs. Results based on the amalgamated themes are always compared to the results based on the standard run. Simulations used typical parameters for Duplin County, assuming that corn was grown and alachlor applied. Alachlor was the most commonly used pesticide on the fields simulated. Annual model outputs of water and alachlor percolation and runoff and sediment loss were integrated into the GIS and summarized by field before analysis

METHODS

Study Site The study site in Duplin County, North Carolina, is the 2044 ha Herrings Marsh Run Watershed This watershed is part of the larger Goshen Swamp which serves

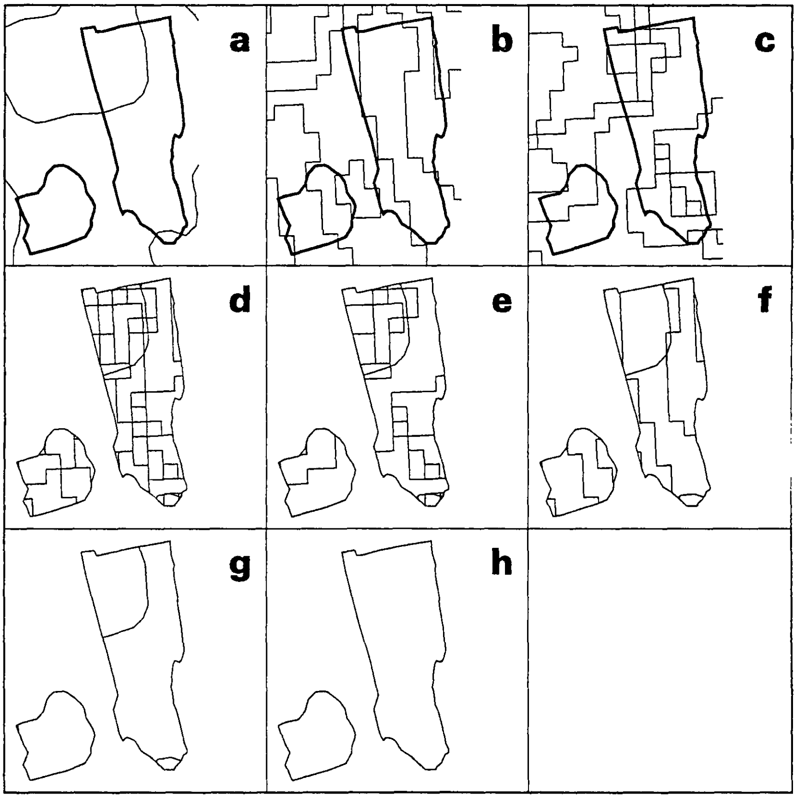


Figure 2 The original, intersected, and amalgamated themes near two of the agricultural fields Original themes agricultural fields (thick lines in a, b, and c), soils (a), slope (b), and aspect (c) Fundamental unit coverage (intersected and clipped themes) (d) Coverages resulting from amalgamations slope (e), aspect (f), slope and aspect (g), soils, slope, and aspect (h)

as the headwaters to the Northeast Cape Fear River. The watershed is in the coastal plain physiographic region. The soils are generally sandy; the major crops are soybeans, corn, cotton, tobacco, and hay.

The agricultural land is composed of 363 fields. The database contained adequate field, crop, pesticide, and soils data for 129 fields. Due to the extensive computer requirements, it was desirable to limit this study to an area represented by approximately 1000 polygons Of the 129 fields, 41 were chosen at random Intersection with the slope, aspect, and soils themes, resulted in 990 fundamental unit polygons within the fields. Of these, 147 were less than 5 square feet and were ignored in the simulation runs The area ignored was less than 0.0003% of the 41 fields. One agricultural field was represented by two distinct polygons, thus the fields are represented by 42 field polygons.

Software The GIS aspects of the work--digitizing, intersecting, mapping--primarily used ESRI's Arc/Info software, although Atlas*GIS was used for the initial digitization of the field coverages. GLEAMS version 1.8.55 (Davis *et al.*, 1990; Leonard *et al.*, 1987) was used to model water and pesticide transport.

The GIS Interface for Ground-Water Models (Stallings *et al.*, 1992) was used to overlay the coverage and run the model simulations for the watershed This software provides a linkage between Arc/Info, tabular databases, and GLEAMS. It allows one to perform GLEAMS simulations for many fields simultaneously (Stallings *et al.*, 1992). Graphical and statistical analyses were done using Splus version 3.2 (Statistical Sciences, Inc., 1993).

Spatial and Tabular Data Details of the spatial and tabular data can be found in Stallings *et al.* (1992) or Stallings (1995), they will be described here briefly. Polygon coverages describing agricultural fields, soil series, slope, aspect, and the watershed boundary were used.

The fields coverage was produced using a Zoom Transfer Scope and aerial photography. Identical cropping and pesticide practices were assumed on all fields studied. The cropping practices for corn (e.g., planting dates, field practices) were based on common usage within the county as determined by a Duplin County extension agent (Curtis Fountain, pers. comm., 1992); the typical planting date was March 15 and the harvest date was September 15 A rooting depth of 30" was used for these scenarios, therefore percolation loss values represent the loss below 30". Alachlor application was simulated on April 1 at a rate of 1.4 kg/ha. This application was based on actual alachlor applications as determined by a Cooperative Extension Service field survey of the watershed. Generalized information about Leaf Area Index (LAI) was taken from the GLEAMS manual (Davis *et al.*, 1990).

The soils coverage was digitized from a 1:24,700 prepublication map provided by the Soil Conservation Service (SCS). The SCS also provided the most recent version of their detailed soils attribute data for Duplin County. Generalized data on soil hydrologic conductivity, structure, and texture were taken from the GLEAMS manual (Davis *et al.*, 1990) and linked to the soils data.

The slope and aspect polygon coverages were based on 1:24,000 USGS digital elevation models. The DEMs were smoothed with a low-pass filter prior to calculation of the slope and aspect

The daily rainfall data for the closest weather station (Warsaw, NC) was taken from the Hydrological Information Storage and Retrieval System (Wiser, 1975). Missing values were filled with the average value for that date determined using the six years of available data (1985-1990). Monthly average high and low temperatures were also derived from the daily temperature data for this weather station. Average monthly solar irradiance was calculated based on data from the Raleigh, NC, weather station.

The corn and alachlor simulations were run on 10 years' data using identical cropping, pesticide, and 1990 weather data. After the first three years of simulation, the outputs stabilized to within three significant figures of the final values. This was also true in the Bleeker *et al.* (In Press) study

Coverage Preparation GLEAMS, at least as it is applied by the GIS-model interface, assumes polygons are homogenous with respect to the input themes. GIS intersection combines all pertinent attributes into a single coverage. The coverage breaks the fields into

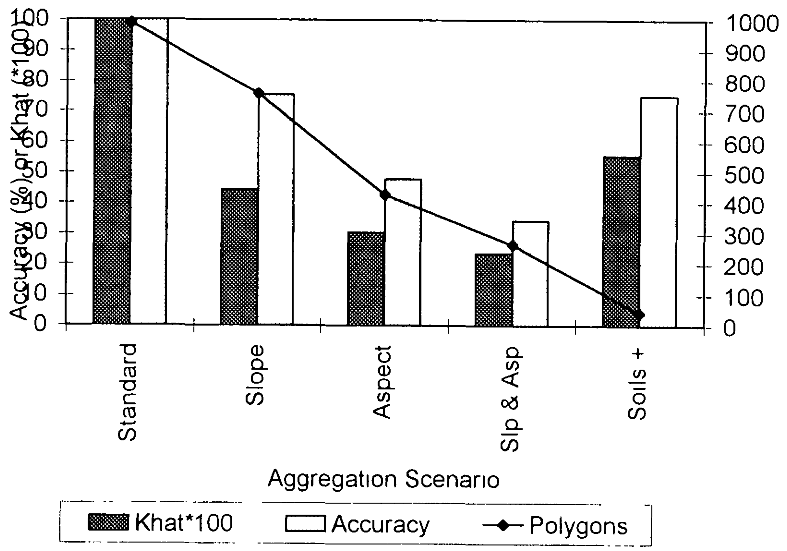


Figure 3 Change in polygon count, overall map accuracy, and Khat with different amalgamation scenarios

fundamental unit polygons that are locally unique with respect to their field, slope, aspect, and soils. The intersection and maximum overland flow calculations were performed as described in Stallings *et al.* (1992), except that the spatial-simplification routine in the GIS-model interface was not used. The coverage resulting from the intersection is the standard fundamental unit coverage; it, and the results based on it, was the benchmark against which changes were compared

Four separate amalgamations were made to the fundamental unit coverage: slope, aspect; slope and aspect; and soils, slope, and aspect (Figure 3). In these four amalgamations, one or more themes within a field were replaced with the dominant category of each theme within that field.

Simulation Runs and Initial Data Analysis The fundamental unit coverage and the four simplified coverages were used as input to parallel simulations using GLEAMS and the GIS-model interface Thus the slope, aspect, soil, number of polygons per field, and parameters based on polygon size and orientation (e.g., maximum overland flow length) varied with each amalgamation scenario. Inputs relating to the weather and field treatments remained constant across all polygons in each scenario.

Annual values of water and alachlor percolation and runoff and sediment loss were output and integrated into the GIS. The annual values associated with each polygon for the tenth year were extracted. Field summaries were then calculated by taking the spatially weighted average of the polygon results within each field.

STATISTICAL TESTS

Descriptive and test statistics were used to identify which amalgamation scenarios caused significant differences in the model results. Differences are always compared to model results based on the standard fundamental unit coverage; these results are considered correct. The absolute values of the results and their paired differences were judged to be non-normal based on quantile-quantile plots, so non-parametric and descriptive statistics were used for most analyses. The statistics examine results by model output parameter and amalgamation scenario at watershed and field scales. Each test produced a table of results. Conclusions are drawn based on the general trends. Watershed means were used to quantify differences in the entire population. Kendall's Tau rank correlation coefficient and the number of fields correctly classified into the top quartile are used to show the differences in field rankings. The distribution of field-specific relative errors are used to describe the differences in outputs at the field scale.

Watershed Means A t-test based on meaningfully paired differences (Steel and Torrie, 1980) was used to assess the deviation of the watershed means for each output parameter caused by each amalgamation. This test reduces the field-to-field variation. This test assumes normally distributed differences, which is not the case, so the results should be viewed with skepticism.

Kendall's Tau Rank Correlation Coefficient A likely use of the GIS-model interface is in ranking fields with respect to pesticide losses This does not require good results in the absolute sense or even a good linear correlation. Kendall's Tau (Kruskal, 1958) was used to measure the rank correlation of field specific outputs based on different map amalgamations The ranks resulting from the standard run are always used as the basis of comparison. The interpretation of Tau is similar to that for the standard Pearson sample correlation coefficient, except of course that Tau does not require a normal distribution

Relative Errors The relative error of field-specific values was calculated. Values from the standard simulation run are used as a reference. The equation is

Relative_Error = (Output_Amal. - Output_Standard) / Output_Standard

Map Accuracy Two measures are used to quantify map accuracy after amalgamation. overall map accuracy and the Khat statistic. Each is derived from an error matrix

The Khat statistic (Congalton and Mead, 1983, Hudson and Ramm, 1987) measures the amount of agreement between two maps accounting for chance agreement. Perfect agreement is indicated by a Khat of 1. One or more attributes of the standard fundamental unit coverage are used as the reference data. An error matrix is created by comparing an attribute of the fundamental unit coverage with the same attribute from a simplified coverage. A complete sample of the agricultural fields was taken The cells of the error matrix created did not represent discrete samples, but rather the area of agreement, in square feet, between the classes

RESULTS

Polygon Counts and Map Accuracy The map simplifications had the desired effect of decreasing the map polygon count (Figure 3) from 990 to as low as 42. The overall accuracy and Khat statistic decreased substantially with the amalgamations (Figure 3) The accuracy measures are based on only one attribute per map, except for the dominant slope and aspect coverage. The accuracy measures for this map consider each unique pair of

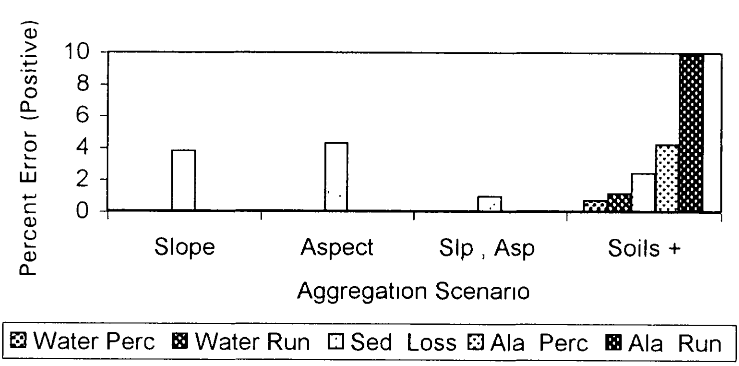


Figure 4 Percentage error in watershed means of model outputs for different amalgamation scenarios

slope and aspect as a class (e.g., slope = 1 and aspect = 2). The map accuracy for the soils simplification represents the accuracy of the soils data and ignores the fact that the slope and aspect were also amalgamated for these fields.

Watershed Means Many of the watershed means based on the amalgamation of slope and aspect were identical to the standard run means to two or three significant figures The exception was sediment loss, which varied a small amount (Figure 4). However, the runs based on the soils simplification show larger differences. None of the means reported here are significantly different from those based on the standard run. The soil simplification did cause a significant difference in soil-borne alachlor runoff loss (see Stallings (1995) for details).

Field Rankings The rank correlation coefficients comparing the standard and amalgamated themes are shown in Figure 5 Most of the Tau values showed significant departures from zero It was generally appropriate to accept the hypothesis that the results were correlated. Two Tau values are similar to zero ($P(H_0) > 0.1$), these result from the soil amalgamation

The low Tau values can be used to identify scenarios differing more from the standard

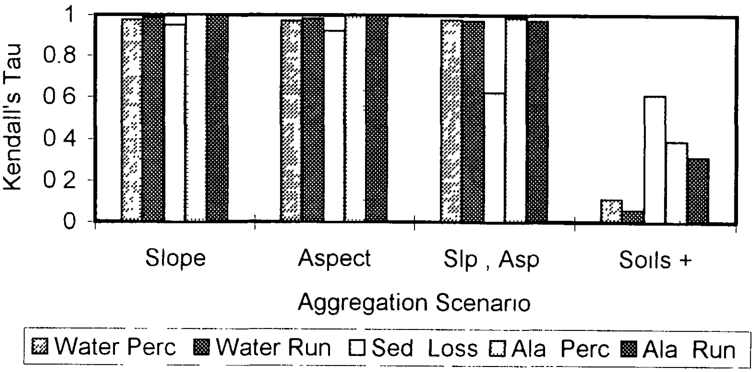


Figure 5 Kendall's Tau (rank correlation coefficient) between field results based on the standard and amalgamated data

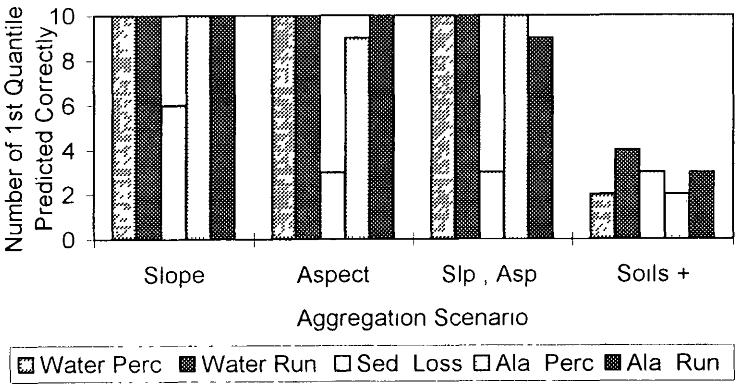


Figure 6 Number of fields correctly placed in the first quartile using results based on amalgamated data

run. The slope and aspect scenario resulted in ranks that were similar to the standard run for most output values (Tau values above 0.9). However, sediment loss shows a bigger difference (Tau values of 0.622). The soils amalgamation caused large differences in model outputs (Tau values ranging from 0.122 to 0.610).

The frequency with which the same fields were classified as being in the top quartile for different scenarios was investigated (Figure 6). The slope and aspect amalgamations caused the fewest errors, matching nine or ten of the ten fields for the majority of output parameters. Sediments loss, however, had more errors, matching only three to six of the ten fields. The soils amalgamation resulted in many errors, matching from two to four of ten fields.

Field Error The field-specific errors are perhaps the largest, since there is less averaging and abstraction to hide errors. This error is pertinent since fields are generally the smallest managed units. The error is presented as absolute relative errors (i.e , 0 10 represents a plus or minus 10% relative error). The field-specific relative errors for each model output and amalgamation scenario are shown in Figure 7. Once again the slope and aspect amalgamations introduced little error into the output values as indicated by the very narrow distributions of relative errors (left three columns). The exception once again is sediment loss (bottom row) The soils amalgamation resulted in much larger relative errors, ranging from approximately -0.8 to 0 9 (right column)

DISCUSSION

A number of factors are considered in this study the theme being amalgamated, field vs. watershed scale, absolute vs. rank results, and output parameter. The statistics used tested hypotheses related to single factors These are used to extrapolate to general trends in the data.

Clearly, the polygon count is greatly reduced by the amalgamation procedures. This results in substantial CPU savings since the CPU use is proportional to the polygon count. The map accuracy is also approximately proportional to polygon count. This relationship is hidden in Figure 3 because the stated map accuracies for the soils amalgamation do not account for the amalgamation errors of slope and aspect.

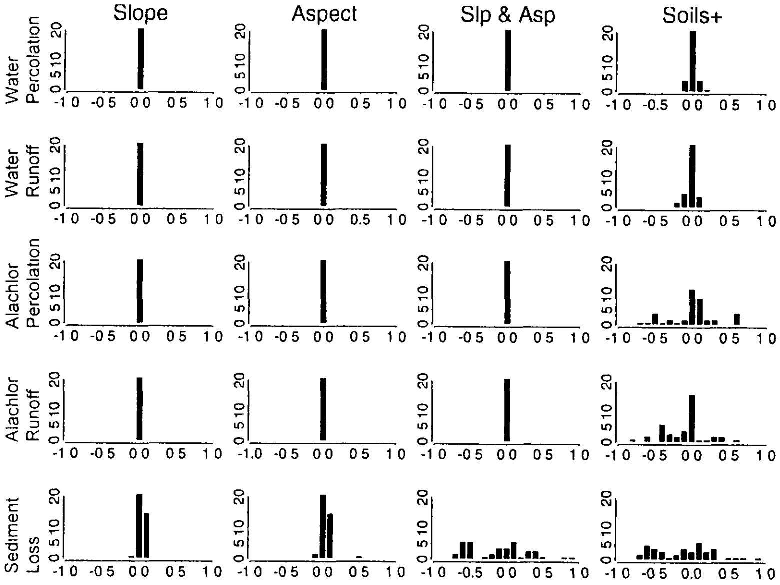


Figure 7 Histograms of field-specific relative errors for each model output and amalgamation scenario

One major advantage of the elimination-based techniques is that they can be applied incrementally. That is, the degree of generalization and map error can be controlled by changing the threshold used to select the polygons to be eliminated. The mode-amalgamation technique cannot be applied at various levels, although it can be applied to a single theme at a time, whereas the elimination techniques are generally applied to all themes at once in a combined coverage.

There does not appear to be a relationship between the accuracy of the input maps and the accuracy of the model outputs. This is due to the differing influences of the input maps. Greenland *et al.* (1985) concluded that, as a rule of thumb, Khat statistics of less than 0.90 to 0.85 indicated a map unfit for use. While the results based on soils simplifications support this, the results based on slope and aspect simplifications indicate that the sensitivity of the model to specific inputs must also be considered.

General Trends Several trends are apparent from the results It is clear that the soils amalgamation result in significantly greater error than the slope and aspect amalgamations However, the slope and aspect amalgamations do result in errors of sediment loss estimation This is due largely to the model's relative sensitivity to these inputs

The field-scale errors were greater than the watershed-scale errors. The soils amalgamation resulted in relative errors of up to 0.09 for watershed means. The same amalgamation resulted in field-specific maximum relative errors ranging from 0.177 to 0.854 (see Stallings (1995) for details) The field-scale errors tend to cancel one another in the watershed summaries, as would be expected as long as the errors caused by the amalgamation are unbiased.

The field-ranking results were similar to the field-specific results. Both the Tau values (Figure 5) and the first quartile predictions (Figure 6) show that the slope and aspect amalgamations affected model results less than the soils simplifications. The exception was sediment loss, which was affected by the slope and aspect amalgamations. The only statistically significant differences were, however, due to the soils amalgamation.

The errors due to slope and aspect amalgamations are probably acceptable in many of the contexts that GLEAMS may be used. Even when used with the best field data, GLEAMS predicts pesticide center of mass, and solute concentration distributions with depth to ± 50% (Leonard *et al.*, 1987; Pennell *et al.*, 1990). Clearly, however, some of the ranking and field-specific results are unacceptable

Limitations This study looked at only a few of GLEAMS many outputs. In particular It did not look at any daily or monthly outputs, or outputs of pesticide concentrations within the soils layers. It should also be emphasized that GLEAMS is a field-scale model. Neither GLEAMS nor the GIS-model interface accounts for routing between fields, from the edge of fields to streams, or from the bottom of the root zone to the saturated ground water. The watershed summaries are therefore indicative of mean loss at the edge of the fields and not of the amount leaving the watershed

Generality of Results The results presented here are specific to the GLEAMS model. The effects of amalgamation will be different for each model and output parameter, depending on their sensitivity to the input themes. However, most hydrologic models rely heavily on soils inputs and will probably be sensitive to changes in these inputs.

The results should be transferable geographically. The model sensitivity to different inputs will not change The relative importance of the inputs may vary geographically, however Amalgamation of the slope and aspect themes might cause more error in terrain with more relief. In areas with more heterogeneous soils, small but important soil series might be suppressed during amalgamation, causing a greater bias in the watershed means.

Future Work This method of amalgamation needs to be compared more rigorously to at least three other methods· (1) elimination of small polygons; (2) elimination of small polygons by field, and (3) amalgamation by spatially weighted means of the attributes. No single method will be best in all cases The elimination techniques have the advantage that they can be applied incrementally. The spatially weighted mean technique will probably result in smaller root-mean-square errors for numeric map data and perhaps in smaller variances in model results due to amalgamation. The mode method has the advantages that it can be applied to individual themes, it can handle class data, and it is simple

CONCLUSIONS

The slope and aspect themes are less important inputs for GLEAMS compared to the soils theme in the area investigated. They are not taken into account by GLEAMS directly when calculating either percolation or runoff volume. Slope and aspect did have an effect on outputs relating to sediment loss. An analysis of the model sensitivity as carried out by Lane and Ferrira (1980) could be used to determine which inputs are not important. Mode amalgamation of soils introduced significant error into the output parameters This will probably be true for most agricultural models predicting percolation and runoff since many use the same or similar equations to represent percolation and runoff process. Aggregating soils for model input may cause unacceptable errors in many circumstances.

In order to determine which results are acceptable, one must first determine what errors are acceptable for a particular study. Knowing this, one can determine which amalgamation scenarios are acceptable for a particular study. The error in large-area summaries due to the mode amalgamation is probably acceptable. The error in ranking and field-specific results are probably acceptable for specific output parameters and amalgamation scenarios (e.g., slope and aspect amalgamations, if percolation or runoff are the outputs of interest). In those cases where the error introduced by the amalgamation is acceptable, the technique clearly reduces the polygon count and therefore the computing effort required.

The most important conclusion of this study is that the amalgamation method and extent should be considered in context with the model and modeling objectives. At least two methods can be used to do this. In either case one must have determined the modeling objectives which model outputs are of interest, the accuracy desired, and whether site-specific values or areal summaries are of interest. The first method consists of identifying the error introduced into the input coverages by generalization and estimating how the errors will propagate through the GIS-model combination based on the model's sensitivity to the inputs derived from the coverages. Where a complete model sensitivity analysis has been done much of this analysis can be carried out with known data. The second method, used in this study, consists of implementing a pilot study. To do this one must implement the model on a representative data set. This has two advantages. It will give more exacting results for specific study locations. It will better represent the simultaneous changes of many soils parameters that occur as the areal representation of the soil series change

ACKNOWLEDGMENTS

Earlier drafts of this article were reviewed by Nancy White, David Adams, Hugh Devine, and three anonymous reviews; the article is much improved as a result. The SCS and the Duplin County office of the Agricultural Stabilization and Conservation Service supplied map and tabular data Mark Rice digitized the agricultural field boundaries. Vernon Cox, Dana Hoag, and Steve Coffey supplied agricultural and pesticide data. Staff at the Computer Graphics Center provided computer and administrative support throughout the project. This work was funded by the Cooperative State Research Service.

REFERENCES

Brassel K.E., Weibel R. (1988) A review and conceptual framework of automated map generalization. Int J Geographical Information Systems 2:229-244

Brown D.G , Bian L., Walsh S.J. (1993) Response of a distributed watershed erosion model to variations in input data aggregation levels. Computers & Geosciences 19(4):499-509

Congalton R.G., Oderwald R.G., Mead R.A. (1983) Assesing Landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogram Engr and Remote Sensing 49:1671-1678

Davis F M., Leonard R.A., Knisel W.G. (1990) GLEAMS user manual version 1.8.55 USDA-ARS Southeast Watershed Research Laboratory, Tifton, Georgia, Lab Note SEWRL-030190FMD

Greenland A., Socher R.M., Thompson M.R. (1985) Statistical evaluation of accuracy for digital cartographic data bases. In: Auto-Carto 7: Digital representations of spatial knowledge. American Society of Photogrammetry, Falls Church, VA (Proceedings)

Hudson W.D., Ramm C.W. (1987) Correct formulation of the Kappa coefficient of agreement. Photogram Engr and Remote Sensing 53:421-422

Kruskal W.H (1958) Ordinal measures of association. J Amer Stat Assoc 53:814-861

Lane L.J., Ferreira V A (1980) Sensitivity analysis. In· Knisel WG (ed) CREAMS a field scale model for chemicals, runoff, and erosion from agricultural management systems. U.S. Dept of Agriculture, Washington, D.C , Conservation Research Report No. 26

Leonard R.A., Knisel W.G., Still D.A. (1987) GLEAMS: groundwater loading effects of agricultural management systems. Trans of the ASAE 30(5)·1403-1418

McMaster R.B., Monmonier M. (1989) A conceptual framework for quantitative and qualitative raster-mode generalization. In GIS/LIS '89, vol Volume 2. American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland (Proceedings)

McMaster R.B., Shea S.K. (1992) Generalization in digital cartography Association of American Geographers, Washington, D.C.

Monmonier M S. (1983) Raster-mode area generalization for land use and land cover maps. Cartographica 20·63-91

Pennell K D., Hornsby A.G., Jessup R E , Rao P.S.C. (1990) Evaluation of five simulation models for predicting Aldicarb and bromide behavior under field conditions Water Resources Res 26.2679-2693

Srinivasan R., Arnold J G. (1994) Integration of a basin-scale water quality model with GIS Water Resources Bulletin 30(3).453-462

Stallings C. (1995) GIS data integration for ground-water quality modeling Ph.D Dissertation, North Carolina State University. In preparation.

Stallings C , Huffman R.L., Khorram S., Guo Z (1992) Linking GLEAMS and GIS ASAE, St Joseph, Michigan (Written for the International Winter Meeting of the ASAE, Nashville, TN, 15-18 December)

Statistical Sciences (1993) S-PLUS User's Manual, Version 3.2 StatSci, a division of MathSoft, Inc., Seattle

Steele R.G D , Torrie, J.H (1980) Principles and procedures of statistics (2nd) McGraw-Hill, Inc., New York.

Wang F., Donaghy P. (1992) Assessing the impact of automated polygon elimination to overlay analysis accuracy. In. ASPRS/ACSM/RT 92 technical papers, vol 4: Remote sensing and data aquisition American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland

Whede M. (1982) Grid cell size in relation to errors in maps and inventories produced by computerized map processing. Photogram Engr and Remote Sensing 48:1289-1298

Wiser E.H. (1975) HISARS -- Hydrologic information storage and retrieval system -- reference manual. National Technical Information Service, Springfield, Virginia

TOWARD IMPLEMENTING A FORMAL APPROACH TO AUTOMATE THEMATIC ACCURACY CHECKING FOR DIGITAL CARTOGRAPHIC DATASETS

Barbara Bicking and Kate Beard
National Center for Geographic Information and Analysis
and Department of Surveying Engineering
University of Maine, Boardman Hall, Orono, ME 04469-5711
207/581-2187 and 2147; Fax 207/581-2206
bickingb@osprey.umesve.maine.edu and beard@grouse.umesve.maine.edu

ABSTRACT

The automation of the map data conversion process is one of the key issues in GIS database construction. The accuracy of the resulting digital cartographic datasets—used to provide base reference maps in the GIS application domain—is directly related to the extend of automation of this process. Currently, the map data conversion process is only partially automated. Automation is limited to the capture and verification of map geometry and topology. At present no conversion software or GIS provides a comparable mechanism for thematic data. Theme attribute coding and verifying remains a manual process. The chosen approach argues that the lack of formal definitions of the content of cartographic data is a fundamental impediment to the automation of the theme attribution and verification. This paper reports on work in progress on implementing the conceptual model developed to capture the map content by way of symbols and symbol relationships. Algebraic specifications of these objects facilitate automated map data conversion, and the assessment and verification of their accuracy and consistency at the time of their capture. A symbol-based cartographic knowledge-base and the formal specifications form the basis of a simple prototype implementation which will demonstrate the automated accuracy checking as an integral part of the data conversion process.

INTRODUCTION

Many geographic information systems (GIS) currently in use have a cartographic subsystem used to provide base reference maps in the application domain (Ramirez and Lee 1991). The accuracy of such a cartographic database is therefore critical to the GIS database construction. It is also critical to the effective use of the data in analysis and becomes especially important when the analysis results are used in decision-making (Beard et al. 1991). In this context the automation of the map data conversion process is a key issue. It is currently limited to the capture and verification of map geometry and topology. No comparable automated mechanism is available for thematic map data. No computer-based tool exists to perform a comprehensive assessment of the fitness for use of cartographic datasets. The research reported on in this paper works toward that goal.

Research on the accuracy of spatial databases abound in the GIS literature and focuses primarily on positional accuracy, topological consistency, and quantitative thematic accuracy. Considerably less research has been done on factual or qualitative thematic accuracy, although correct thematic data are equally important for a useful database (Brusegard and Menger 1989; Veregin 1989). Thematic data give meaning to spatial objects. They distinguish spatial objects represented by the same geometric type.

And thematic data are essential when querying the database. In the application environment the majority of GIS users thus are primarily concerned with the factual accuracy of thematic data (Millsom 1991). The magnitude of thematic errors, though very important, is a lesser concern. One major source of factual thematic errors is the keyboard entry of thematic data. Attempts to automate this process have proven to be no trivial matter. No research reports exist on the automation of the theme attribution and verification. No software is commercially available to carry out thematic accuracy and consistency checking.

We have developed a conceptual model for automated capture and accuracy and consistency checking of thematic map data. The model uses the methods of algebraic specification to formally define the component domains of map symbols, including the thematic data at the class level, and symbol relationships in terms of their behavior (Bicking 1994; Bicking and Beard 1994). This paper reports on the implementation work in progress and its overall pursuit: the development of a thematically accurate and consistent cartographic knowledge-base rich in detail with an object-based accuracy and consistency checking mechanism.

The paper is organized as follows. The next section discusses why map data conversion by way of symbols and symbol relationships is more suitable for increased automation than the conventional method. The third section outlines the implementation strategy and includes examples of algebraic specifications. Section Four concludes with comments on the status of the implementation and future work.

TWO MAP DATA CONVERSION METHODS

The automation of the map data conversion process is one of the key issues in GIS database construction. The accuracy of the resulting digital cartographic datasets—used to provide base reference maps in the GIS application domain—is directly related to the extend of automation of this process.

Current Approach To Map Data Conversion

The standard map data conversion currently in use is basically a four-step process. In Step One the map geometry and topology are captured. In Step Two geometry and topology are verified in a post-conversion procedure (e.g. the BUILD and CLEAN commands in Arc/Info). Both steps are automated based on formal definitions of the domains are implemented in the digital environment. In Step Three the thematic data are added through keyboard entry and are stored in relational tables. This is a manual process due to the lack of formal specifications of the map theme domain. Equally, the post-conversion accuracy verification of the thematic data is carried out through visual inspection of proof-plots. Both processes require high concentration; they are tedious and prone to error.

The USGS's National Mapping Division developed and uses the Attribute Verification Package (AVP), an automated post-conversion quality control tool, to perform rudimentary checks for correct general purpose attributes of Digital Line Graph data elements (USGS-NMD 1990). The program is limited in scope and effectiveness. Extensive manual and visual thematic accuracy checks remain necessary. A fundamental drawback of the AVP is its lack of a formal base.

Formal Symbol-Based Approach To Map Data Conversion

Maps are powerful communication tools. Their power lies in using symbols to portray real world objects and their thematic and locational relationships to each other. Symbols do encode all relevant information about these objects, which is also deemed sufficiently relevant information for the base reference maps needed in the GIS application domain. So, rather than decompose the map content into geometric and thematic information, we capture it simultaneously by way of symbols and symbol relationships. The automation of this process is based on formal specifications of all the information encoded in a symbol, namely in its geometric (type, locational and topological), representational (visual variables), thematic, and relational components. Note that most map symbols encode only object class level information at the nominal scale of measurement, although each symbol clearly represents an unique object instance. The inclusion of symbol relationships is of particular benefit: they add richness of detail, more complete relational data, and improve the accuracy of the knowledge-base.

The second core element of our approach is the construction of an object-oriented cartographic knowledge-base. It is structured such that the standardized and finite set of symbols of the 1:24,000 USGS Topographic map series—the selected source document—are stored as objects with their behavior encapsulated in their definition. The development of an object-oriented cartographic knowledge-base has the added benefit of allowing the user to produce consistent, standardized, digital reference map, much like the source document, when combined with a map design knowledge-based system like the one described by Steiner et al. (1989) or Zhan (1991).

The map conversion is then a single-step process: The symbol is captured and immediately verified as an occurrence of the knowledge-base and as accurate and consistent with its definition. The accuracy and consistency checking is part of each symbol's definition. Binary symbol relationships are checked in like fashion at the time of their capture and verified against a set of consistency matrices, based on the theme class of each symbol. The symbol-based method leads to a greater degree of automation of map data conversion and thus to increased accuracy and consistency of the resulting cartographic datasets.

IMPLEMENTATION STRATEGY

The task this research addresses is the automated capture and verification of the thematic accuracy and consistency of cartographic datasets. Checks this research will be able to perform are: 'Check the theme accuracy of the specified symbols' 'Check the relationships of symbol x and symbol y based on their theme class', or 'Check if the symbols crossing at point (x, y) have the correct thematic attributes'. The implementation is done in the Arc/INFO environment to be able to test the formalism with cartographic data.

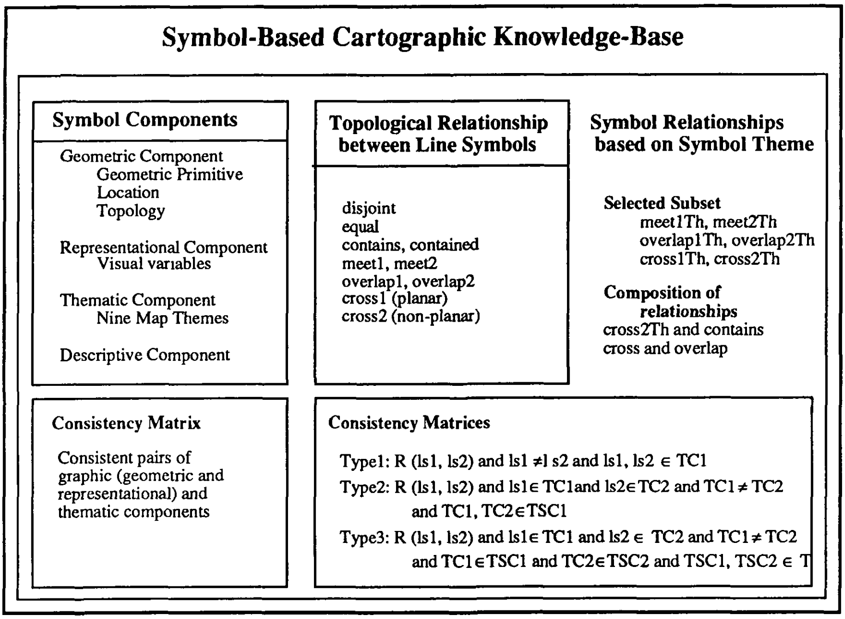
Knowledge-Base Development

Schemata and formal specifications are used to precisely describe the task and the model properties and behavior. The schema in Figure 1 shows the system and its individual modules. (For schemata for the individual modules and their detailed description the reader is referred to Bicking (1994)). The knowledge-base construction began with defining tables in the INFO module into which all consistency matrices were imported. Unique identifier provide the needed links between the tables and the graphic data.

Examples of Algebraic Specifications for the Symbol-based Data Capture and Accuracy Checking

We start with a selected subset of symbols for the prototype implementation from the map theme *Transportation—roads, railroads, and linear hydrographic objects*—and include a subset of topological relationships between them—*meet, overlap,* and the planar and non-planar *cross.*

Figure 1: Schema of the Cartographic Knowledge-base with Symbol-Based Thematic Accuracy Checking



Symbol-Based Cartographic Knowledge-Base

Symbol Components

Geometric Component
Geometric Primitive
Location
Topology

Representational Component
Visual variables

Thematic Component
Nine Map Themes

Descriptive Component

Topological Relationship between Line Symbols

disjoint
equal
contains, contained
meet1, meet2
overlap1, overlap2
cross1 (planar)
cross2 (non-planar)

Symbol Relationships based on Symbol Theme

Selected Subset
meet1Th, meet2Th
overlap1Th, overlap2Th
cross1Th, cross2Th

Composition of relationships
cross2Th and contains
cross and overlap

Consistency Matrix

Consistent pairs of graphic (geometric and representational) and thematic components

Consistency Matrices

Type1: R (ls1, ls2) and ls1 ≠l s2 and ls1, ls2 ∈ TC1

Type2: R (ls1, ls2) and ls1∈ TC1and ls2∈TC2 and TC1 ≠ TC2 and TC1, TC2∈TSC1

Type3: R (ls1, ls2) and ls1∈ TC1 and ls2 ∈ TC2 and TC1 ≠ TC2 and TC1∈TSC1 and TC2∈TSC2 and TSC1, TSC2 ∈ T

Algebraic specifications are written for each module and its components and are combined as needed. Only those properties are specified which are essential to satisfy the task, thus avoiding overspecification (Horebeek and Lewi 1989; Guttag and Horning 1978; Liskov and Zilles 1975). Note that the specifications—given below in abbreviation—use the construct operations *create* and *assign* in simple and modified form and the observe operations *get* and *consistent* to obtain information about the specified object and its accuracy and consistency. Each component is considered a set. In the context of this implementation each set must satisfy certain constraint conditions, some of which are upwardly cumulative (Beard 1991). These will be revised and removed—if needed—as the knowledge-base grows.

Condition 1: All sets must be non-empty.
Condition 2: All facts about sets, i.e., geometry, topology, graphics, map theme must be explicitly stored in the cartographic database or be inferable from the stored data.
Condition 3: A line must not intersect itself, i.e. all interior points have unique values.
Condition 4: A line must not close to form a circle, i.e. its start node and its end node are unique also.
Condition 5: A line can only have one start and one end node. From this follows that rotaries, meanders, and bifurcations are excluded by definition.
Condition 6: All topological relationships are binary relationships.
Condition 7: Topological relationships must be explicitly stored in the cartographic database or be inferable from the stored relationships.

Specification 1 describes the behavior of the sort *symbolType* and imports the sort *geometricPrimitive*, defined elsewhere. It allows to construct a basic graphic symbol object of one of the three primitive geometric types: point, line, and area.

Specification 1: The Symbol Type

```
SORT                    symbolType USES geometricPrimitive, Boolean
OPERATIONS
    create:             —> symbolType
    assignGeomPr:       symbolType x geometricPrimitive —> symbolType
    getGeomPr:          symbolType —> geometricPrimitive
```

The specifications 2 through 4 are derived from the generic specifications for *symbolClass*, *mapDomain*, and *symbol*, respectively. Specification 2 describes the behavior of the cartographic rendering of a line symbol. To this end it USES the sorts symbolType, lTyp, lTex, lSiz, lCol, and lSha—the individual visual variables which have the most expressive power for line symbols. Note that *lTyp* is not a visual variables as defined by Bertin (1983). It is introduced to differentiate between a single line and two parallel lines, called *casing* in cartographic terminology.

Specification 2: The Line Symbol Class

```
SORT                    lineSymbolClass USES symbolType, lTyp, lTex, lSiz,
                        lCol, lSha, INCLUDES visVar with lTex, lSiz, lCol, lSha
                        for visVarVal
OPERATIONS
    createLsc:          symbolType x lTyp x lTex x lSiz x lCol x lSha —>
                            lineSymbolClass
    getVisVarVal:       lineSymbolClass x visualVariable —> visualVariableValue
```

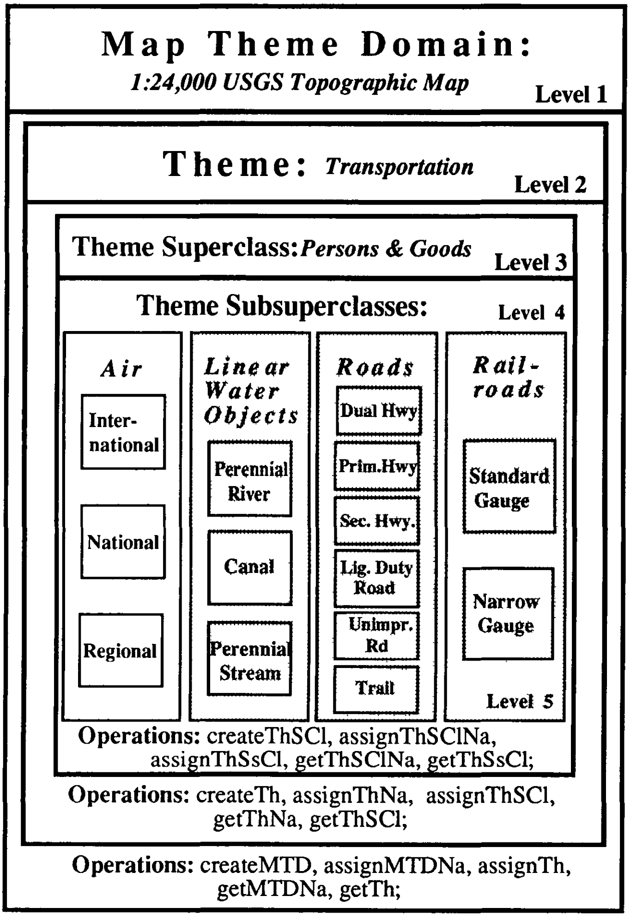
The map theme domain (MTD) of cartographic symbols is hierarchically structured, comprising a finite set of themes (T) followed by a sequence of theme superclasses (TSC) and subsuperclasses (TSsC) and theme classes (TC). Figure 2 shows the specific case of the 1:24,000 USGS Topographic map series and a subset of the theme *Transportation*. The specifications for the *mapThemeDomain*, the *theme*, and the *themeClass* are parallel in structure. Specification 3 is representative for them.

Specification 3: The Theme Class

```
SORT                    themeClass USES themeClassName, themeClassMember
OPERATIONS
    createThCl:         —> themeClass
    assignThClNa:       themeClass x themeClassName —> themeClass
    assignThClMem:      themeClass x themeClassMember —> themeClass
    getThClNa:          themeClass —> themeClassName
    getThClMem:         themeClass —> themeClassMember
```

The linking of the representation and content domains is achieved with the object *symbol*. The symbol is composed of a *lineSymbolClass* component—it comprises the geometric and the graphic domains from Specification 1 and 2—and the *themeClass* component from Specification 3. Its accuracy and consistency is checked against a consistency matrix composed of accurate and consistent pairs of *lineSymbolClass* and *themeClass*.

Figure 2: Schema of the Map Theme Domain and *Transportation* Theme



Specification 4: The Symbol

SORT	symbol USES lineSymbolClass, themeClass, matrix, Boolean
OPERATIONS	
createSym:	lineSymbolClass x themeClass —> symbol
getLsc:	symbol —> lineSymbolClass
getThCl:	symbol —> themeClass
consistent:	symbol x matrix —> Boolean

Specification 5 describes in generic terms the behavior of the thematically constrained symbol relationships. It imports the sort *relationshipName* and the sort *symbol* from the previous specification to create a symbol relationship.

Specification 5: The Symbol Relationship

```
SORT                 symbolRelationship USES symbolRelationshipName,
                        symbol, matrix, Boolean
OPERATIONS
   createSymRel:     symRelName x symbol x symbol —> symbolRelationship
   getSymRelNa:      symbolRelationship —> symRelName
   getFirstSym:      symbolRelationship —> symbol
   getSecSym:        symbolRelationship —> symbol
   consistent:       symbolRelationship x matrix —> Boolean
```

All specifications require slight syntactic adaptations to facilitate their compilation in the Macintosh version of *Gofer*, a functional programming language (Jones 1994). This work is currently in progress. Their integration into the Arc/INFO environment will follow this implementation step.

CONCLUDING REMARKS

The research presented here emphasizes the importance of factually accurate and consistent thematic data in cartographic and GIS databases. It takes an object-oriented approach to data conversion and verification with the accuracy and consistency checking as part of the object, i.e. symbol and symbol relationship, definition.

Based on the insights obtained from Mark and Xia's (1994) implementation of the 9-intersection model for spatial relationships developed by Egenhofer and Franzosa (1991), we will implement the formalism for line-line relationships in Arc/INFO (Egenhofer 1993).

Once the knowledge-base is established the data conversion for the selected subset of line symbols and symbol relationships will be carried out. The implementation of the described approach requires that the map content is captured by way of symbols, either through laser scanning with color and line pattern recognition capability, or line tracing, or digitizing. Digitizing was chosen because such a laser scanner is currently not available to us. Furthermore, by focusing on digitizing we eliminate any errors associated with this technology. These will need to be accommodated in a full-scale program development. The symbol-based data will be stored in thematic layers in Arc/INFO and each layer will be verified prior to their vertical integration into a *Transportation* layer. Since the implementation is work in progress, we expect revisions and adaptations of the what is done to-date.

REFERENCES CITED

Beard, M.K. 1991: *Constraints on Rule Formation.* In: Buttenfield, B., McMasters, R. (eds.) 1991: *Map Generalization. Making Rules for Knowledge Representation.*, 121-135 (Chap. 7), London.

Bicking, B. 1994: *A Formal Approach To Automate Thematic Accuracy Checking For Cartographic Data Sets.* MSc Thesis, NCGIA and Dept. of Surveying Engineering, University of Maine, Orono.

Bicking, B., Beard, M.K. 1994: *A Formal Approach To Automate Thematic Accuracy Checking For Cartographic Data Sets.* In: GIS/LIS '94 Proceedings, Phoenix, AZ, 63-74.

Bertin, J. 1983: *Semiology of Graphics.* Madison, WI.

Brusegard, D., Menger, G. 1989: *Real Data and Real Problems: Dealing with Large Spatial Databases.* In: Goodchild, M.F., Gopal, S. (eds.) 1989: *The Accuracy of Spatial Databases*, London, 55-67.

Egenhofer, M.J., Franzosa, R.D. 1991: *Point-set Topological Spatial Relations.* In: IJGIS, 5, 2, 161-174.

Egenhofer, M.J. 1993: *Definition of Line-Line Relations for Geographic Databases.* In: Bulletin of the Technical Committee on Data Engineering (IEEE), 16, 3, 40-30.

Guttag, J.V., Horning, J.J. 1978: *The Algebraic Specification of Abstract Data Types.* In: Acta Informatica, 10, 1, 27-52.

Horebeek, van I., Lewi, J. 1989: *Algebraic Specifications for Software Engineering.* Berlin, Germany.

Jones, M.P. 1994: *MacGofer, version 2.30.* Glasgow, GB and New Haven, CT.

Liskov, B., Zilles, S.N. 1975: *Specification Techniques for Data Abstractions.* In: IEEE Transactions on Software Engineering, SE-1, 1, 7-19.

Mark, D.M., Xia, F.F. 1994: *Determining Spatial Relations between Lines and Regions in Arc/INFO using the 9-Intersection Model.* In: Proceedings, 14th ESRI User Conference, Palm Springs, CA.

Millsom, R.1991: *Accuracy Moves from Maps to GIS - Has Anything Changed?* In: Hunter, G.J. (ed.) 1991: Symposium of Spatial Database Accuracy (Proceedings). Melbourne, Victoria, Australia, 1-16.

Ramirez, J.R., Lee, D. 1991: *Final Report on the Development of a Cartographic Model for Consistent and Efficient Map Production.* OSU Center for Mapping, Columbus, OH, (Research Report submitted to USGS-NMD).

Steiner, D.R., Egenhofer, M.J., Frank, A.U. 1989: *An Object-oriented Carto-Graphic Output Package.* In: ASPRS/ACSM Ann. Conv., Baltimore, MD, Tech. Papers, 5, Surveying & Cartography, 104-113.

USGS-NMD 1990: *Attribute Verification Package (AVP) - Users Manual.* US Geological Survey: National Mapping Division, Technical Instructions: Software Users Manual. UM1-19-0; (updated 09/90,03/91, 02/93).

Veregin, H. 1989: *A Taxonomy of Error in Spatial Databases.* NCGIA-Santa Barbara, Technical Paper. 89-12.

Zhan, F. 1991: *Structuring the Knowledge of Cartographic Symbolization - An Object-Oriented Approach.* In: ACSM/ASPRS Ann. Conv., Baltimore, MD, Tech. Papers, 6, Auto-Carto 10, 247-260.

IMPROVING REMOTE SENSING-DERIVED LAND USE/LAND COVER CLASSIFICATION WITH THE AID OF SPATIAL INFORMATION

Yingchun Zhou[1], Sunil Narumalani[1], Dennis E. Jelinski[2]

[1]*Department of Geography, University of Nebraska, Lincoln, NE 68588-0135*
[2]*Department of Forestry, Fisheries & Wildlife, University of Nebraska, Lincoln, NE 68583-0814*

ABSTRACT

Most fundamental per-pixel classification techniques group pixels into clusters based on their spectral characteristics. Since various terrestrial objects may exhibit similar spectral responses, the classification accuracies of remote sensing derived image-maps is often reduced. This study focuses on using the shape index of detected ground objects to resolve some of the spectral confusions which occur when pure per-pixel classification algorithms are applied. First, homogeneous areas were identified by using an edge detection algorithm. Second, a stratification procedure divided the image into two strata based on the shape index of patches. One stratum was composed of patches with regular shapes and large sizes, such as agricultural fields and some wet meadows. The other stratum was composed of highly fragmented patches, including urban areas, roads, and riparian vegetation. By stratifying the image, the classes which frequently caused mixed clusters, such as grassy surfaces in urban areas and crop fields, wet fields and riparian forests, were assigned to different strata, thus reducing the possibility of spectral confusion. Third, a spectral classification algorithm was applied to the two strata separately to derive the land cover information for each layer. Finally, the two classifications were merged to produce the final land use/land cover map of the study area.

INTRODUCTION

Multispectral classification techniques have been used for a variety of applications, such as land use/land cover mapping, crop classification, wetland change detection, and landscape diversity measurements. Most of the fundamental classifiers such as ISODATA, sequential clustering, and maximum likelihood classification, group pixels into clusters based on their spectral characteristics. However, due to the spectral sensitivity of remote sensing instruments and the material properties of terrestrial features, pixels belonging to different classes may exhibit inherently similar spectral properties (Gurney and Townshend, 1983). For example, it is not surprising to find that grassy surfaces (e.g., lawns, parks, etc.) within an urban area are often misclassified as agricultural fields. In other cases, bare fields may be confused with concrete surfaces in urban areas, or that riparian woodlands have been mixed with wet agricultural fields when per-pixel spectral classification techniques are used.

Human interpreters can resolve most of these confusions since they possess a comprehensive knowledge of image tone, texture, pattern, association, shape size, position, and other related characteristics of various features (Gurney,1981). Consequently, visual interpretation often achieves a much higher accuracy than automatic digital classifiers though it is laborious and time-consuming process. Innovative models have, therefore, been developed to take into account spatial information in addition to spectral information, to aid in classification (Argialas, 1990). These spatial classifiers often consider such aspects as image texture, pixel proximity, feature size, shape, direction, repetition, and context for improving the classification of an image (Lillesand and Kiefer, 1994).

Spatial information inherent in an image itself can be extracted to assist the spectral classification process. Various kinds of spectral information have been used in pre-classification segmentation, post-classification labelling, or as additional layers input into a statistical classifier. Gurney (1983), used the relative locations of clouds and shadows to successfully separate cloud shadows from spectrally similar water surfaces. Johnsson (1994), improved the spectral classification results by reassigning segments according to a set of decision rules based on size and neighborhood. In another study, Fung and Chan (1994) used the spatial composition of spectral class (SCSC) within a moving window to label pre-classified spectral classes for deriving land use/land cover characteristics. Based on the SCSC ranges, the authors were able to separate water, high density urban land, low density urban land, bare areas, and grassy surfaces. An edge detection segmentation method along with a knowledge based classification that took into account the contextual, textual, and spectral information of segments was developed by Moller-Jensen (1990) to classify an urban area. The author concluded that an expert system-based classification approach produced improved results over traditional classification techniques.

This study focuses on using a stratification process to avoid some of the spectral confusions which occur when pure per-pixel classification is used. A combination of spectral and spatial pattern recognition techniques was used for classifying the land use/land cover of an image. A directional first-differencing algorithm was applied on the original image to highlight edge information. Relatively homogeneous areas were clumped based on the network of edge features, and each homogeneous patch was assigned to either the simple-shape group or the fragmented group, using its shape index. Spectral classification was performed on these two groups separately, and the two classified image-maps were merged to produce a composite classification of the study area.

STUDY AREA

The study area for this project lies in the central Platte River valley. It is comprised of portions of Merrick and Polk counties in Nebraska. A subset of the Landsat Thematic Mapper (TM) image, Path/Row 29/31, acquired on 19 August 1992 was used to illustrate the methodology of the knowledge-based land use/land cover classification of the Platte River flood plain (Figure 1). In the study area, agricultural land dominates the landscape, with corn, sorghum, and soybean being the major crops. Remnants of natural grasslands are found only on the flood plain bluffs, while wet meadows are distributed along the Platte River channel and other small streams. Woody vegetation, which requires more moisture than grasslands is a composite of riparian forests and wetland shrubs and is formed mainly along the stream channels. In sum, natural vegetation cover is extremely fragmented due to the intensive agricultural and other human activities prevalent in this region.

METHODOLOGY

The classification strategy discussed in this paper involved a three-step process. First, homogeneous areas were identified using an edge detection technique on the raw image data, whereby, linear features or edges were detected to isolate these areas. Contiguous non-edge pixels were grouped as a unit. Second, these homogeneous units were stratified into two groups based on their shape index values . One group included agricultural fields and a part of large wet meadow parcels, while the other consisted of highly fragmented features, including urban areas, roads, and riparian vegetation. Finally, a per-pixel spectral classification algorithm was applied to the two groups separately. Pixels were labelled into one of the following eight categories: agricultural, water, forests, wetland shrubs, wet meadows, grassland, urban/roads, and bare.

Identification of Homogeneous Areas

A digital image is a complex of points (i.e., single pixels), and patches (i.e., connected sets of pixels with some uniform property such as grey level or texture) (Argialas, 1990). An edge detection method modified from the directional first differencing algorithm was developed to detect border pixels of homogeneous areas (or patches) and linear features. Every land patch was assigned an unique value to evaluate it as a whole unit in order to facilitate the subsequent measurement of the shape index. This measurement was essential to the stratification of the image.

The first differencing algorithm is designed to emphasize edges in image data (Lillesand and Kiefer, 1994). It is a procedure that systematically compares each pixel in an image to one of its immediate neighbors as follows:

P	H
V	D

P - Primary pixel being processed
H - Horizontal neighbor
V - Vertical neighbor
D - Diagonal neighbor

$$\text{Horizontal first difference} = BV_P - BV_H \quad (1)$$

$$\text{Vertical first difference} = BV_P - BV_V \quad (2)$$

$$\text{Diagonal first difference} = BV_P - BV_D \quad (3)$$

BV = brightness value of pixel

In this study, both horizontal and vertical first differences were computed using TM bands 2, 3, and 4. The differences of the brightness values (BVs) between one pixel and its neighbors can be negative, positive, or zero. However, since the multi-band differencing algorithms (equations 4 and 5) require absolute values, the signs of the differences are removed.

Horizontal first difference = $|DN_{P2}-DN_{H2}| + |DN_{P3}-DN_{H3}| + |DN_{P4}-DN_{H4}|$ (4)
Vertical first difference = $|DN_{P2}-DN_{V2}| + |DN_{P3}-DN_{V3}| + |DN_{P4}-DN_{V4}|$ (5)
where: 2, 3, and 4 represent TM bands 2, 3, and 4 respectively.

If the directional first differencing of a pixel is larger than or equal to the threshold value of 10, in either direction, the pixel is selected as an edge pixel, otherwise it is ignored. The threshold value was determined by experimenting with different values, and examining their effects on the image. Unfortunately, this is not a universally applicable value. An appropriate threshold value should be determined by the users since it may vary in different images.

All edge pixels were assigned a value of 1 in the output image, and all non-edge pixels were assigned a value of 0. Based on this output, contiguous groups of pixels with zero difference values were clumped into patches. Each patch was a relatively homogeneous area and was assigned an unique value so as to be treated as an independent unit in shape measurement.

Essentially, the detected edge pixels make up a boundary network which manifests homogeneous areas (Figure 2). Obviously, some pixels are the borders of fields, but they are also a part of the adjacent patch. For example, if there are two adjacent fields, the boundary pixels between them will be detected as edges (Figure 3). They are basically part of the field on the left side of or above them, and can be extracted and assigned back to the patches they belong to. If the neighboring pixel on the left, or the neighboring pixel above an edge pixel is not an edge pixel, it would indicated that the edge pixel is not significantly different from this neighbor and should therefore, belong to the same patch. These border pixels only served as edges temporally for the identification of homogeneous areas, but they can be reassigned back to the appropriate patches. This edge detection technique has an advantage over other edge detection methods (e.g., high-frequency filtering and texture measurements), which cannot differentiate linear features and land borders, and therefore, often cause "edge errors".

Pre-classification Stratification

Stratification is usually performed before per-pixel classification in order to separate cover types with inherent spectral similarity. The geometric appearance of an object (i.e., its shape), is an important element of pattern recognition. In this project the shape index of each feature was computed using the ratio of perimeter to area algorithm. Implementing this step resulted in the stratification of the image into a stratum of simple, regular shape patches and a stratum of complex-shaped patches.

In many parts of the U.S., where the terrain is smooth, agricultural fields often have regular shapes and very simple perimeters, and therefore, a low perimeter to area ratio. Conversely, urban areas, grasslands, and natural woody vegetation have irregular shapes and complex perimeters, and consequently, yield high shape index values. In the case of some wet meadows, which have large size and smooth texture, medium shape index values were derived. All patches were sorted into two groups. The first group was comprised of patches with regular shape and large size, including agricultural fields with crops, bare fields, and some wet meadows. The second group was comprised of fragmented patches with complex shapes, and included urban areas, roads, streams, upland grasses, wetland shrubs, and forests. Detected edges were the most irregular and complex features, and were accordingly assigned to the fragmented group.

Once this stratification was completed, the classes which were hard to separate based on their spectral characteristics, such as discrimination between grassy surfaces in urban areas and crop fields with low infrared spectral reflectance, or wet fields and riparian forests, or upland grasses and wet meadows, etc., were assigned to different groups, thus reducing the possibility of spectral confusion. Each group of land patches was used to create a mask to extract the corresponding areas from the original image. Consequently, two image-strata were formed. One comprised of a highly fragmented stratum, while the other was a low fragmented stratum.

Spectral Classification

Each image-stratum was classified independently. TM bands 2, 3, 4, and 5 were used to extract 50 clusters from the highly fragmented stratum, using a self-iterative, unsupervised clustering algorithm. The clusters were assigned in feature space using the maximum likelihood rule. Each cluster was grouped into a land use/land cover category by overlaying it on a false-color composite of the image, and delineating its respective location on the red-infrared (TM bands 3, 4) scatterplot. The methodology was also applied to classify the low fragmented stratum. However, only TM bands 3, 4, and 5 were used, since an examination of the TM band 2 histogram for the low fragmented stratum revealed a narrow range of BVs. Such uniformity of BVs makes it difficult to extract reliable clusters and therefore leads to poor classification results.

Once the low fragmented stratum was classified, the two classifications were merged to produce a final land use/land cover image-map with eight categories: water, wetland shrubs, forests, wet meadows, grassland, agricultural fields, urban, and bare fields (Figure 4). It is evident from Table 1 that agriculture occupies a substantial portion of the landscape (nearly 51%), while natural grasslands, wetlands, and forests comprise only 42% of the landscape. The ratio of the natural vegetation land area is deceptive, since much of the grassland included in this image interpreted consists of fields that are used for grazing purposes or those that have left been fallow. This is because the central Platte River Valley has undergone significant transformation over the last century. Most of this has been due to agricultural and development activities. These activities have led to a reduction in the extent of native vegetation and the fragmentation of their remnants (Narumalani et al., 1995). From the perspective of conserving natural resources, it is important to conserve what remains and implement schemes that are compatible with existing land use activities. Future agricultural activities must be carefully monitored to minimize their impact on the remaining natural vegetation of the area.

Table 1. Land Cover Classification of the Study Area.

Class Name	Area (ha)	Percent (%)
Water	532.59	2.2
Wetland shrub	3208.31	13.1
Forest	476.39	1.9
Wet meadow	2336.84	9.6
Agriculture	12410.9	50.8
Grassland	4335.38	17.8
Urban	174.15	0.7
Bare/Fallow	941.64	3.9

SUMMARY AND CONCLUSIONS

The approach discussed in this paper identified homogeneous areas by using the directional first-differencing method. Border pixels could be extracted and assigned back to the patch to which they belonged. The critical step in this edge detection method was thresholding. Determining the threshold was an experimental process affecting the quantum of the edge bias for the analysis. If the threshold is too small, non-edge pixels will be included. Conversely, if it is too large, edges may not be detected, neighboring patches may join, and hence interfere with the shape measurements and the subsequent stratification. Another important aspect related to the effectiveness of the edge detection technique is band selection. In this study, Landsat TM bands 2, 3, and 4 were used due to the following reasons. TM band 4 allows biomass detection and differentiation, while bands 2 and 3 clearly show roads, urban areas, and streams, which are a major component of the border features. A false-color composite (TM bands 2, 3, 4 = RGB) shows the best representation of border features.

In the central Platte River valley study area, crops are usually planted in uniform, distinct fields, often with a single crop to a field. This farming pattern permitted an effective stratification and aided to the spectral classification. However, in many regions of the U.S., and the world, crops are planted in very small fields due to topographic, cultural, or landscape characteristics. Consequently, the geometric differences between natural cover and agricultural fields might be undetectable, and other more effective methodologies may need to be developed.

Stratification involves a division of the study scene into smaller areas or strata based on some criterion or rule so that each stratum can be processed independently (Hutchinson, 1982). The purpose of stratification in this research was to separate different features which cause confusion due to their spectrally similarity. The stratification divided the original image to two strata, but did not alter its original BVs. The stratification results are effective, except for a few small or irregular agricultural fields which were assigned to the high fragmented stratum. However, such fields or patches still can be classified into their appropriate land use/land cover classes if their spectral values do not deviate too far from the class means. A visual comparison of the original image data with the classified image-map showed that the methodology described in this paper was effective in resolving much of the classification confusion, especially between agricultural fields and urban grassy surfaces, or riparian forests.

Spatial information, which is implied in the image, can be a significant ancillary data source for digital classification improvement. Its extraction from digital imagery, especially high resolution images such as those acquired by the Landsat TM and SPOT sensor systems, would greatly improve image classification if proper strategies are used.

REFERENCES

Argialas, D. and C. Harlow, 1990. "Computational Image Interpretation Models: An Overview and a Perspective," *Photogrammetric Engineering and Remote Sensing,* 56(6):871-886.

Gurney, C., 1981. "The Use of Contextual Information to improve Land Cover Classification of Digital Remotely Sensed Data," *International Journal of Remote Sensing,* 2(4):379-388.

Gurney C. and J. Townshend, 1983. "The Use of Contextual Information in the Classification of Remotely Sensed Data," *Photogrammetric Engineering and Remote Sensing,* 49(1):55-64.

Johnsson, K., 1994. "Segment-Based Land-Use Classification from SPOT Satellite Data," *Photogrammetric Engineering and Remote Sensing,* 60(1):47-53.

Lillesand, T. and R. Kiefer, 1994. *Remote Sensing and Image Interpretation,* John Wiley & Sons, Inc., New York.

Moller-Jensen, L., 1990. "Knowledge-Based Classification of an Urban Area Using Texture and Context Information in Landsat-TM Imagery," *Photogrammetric Engineering and Remote Sensing,* 56(6):899-904.

Narumalani, S., D. E. Jelinski, Y. Zhou, and D. Smith, 1995. "Analyzing Landscape Patterns Using Remote Sensing-Derived Land Cover Classification for the Central Platte River Valley in Nebraska," *Technical Papers, ASPRS,* in press.

Tung, F. and K. Chan, 1994. "Spatial Composition of Spectral Classes: A Structural Approach for Image Analysis of Heterogeneous Land-Use and Land-Cover Types," *Photogrammetric Engineering and Remote Sensing,* 60(2):173-180.

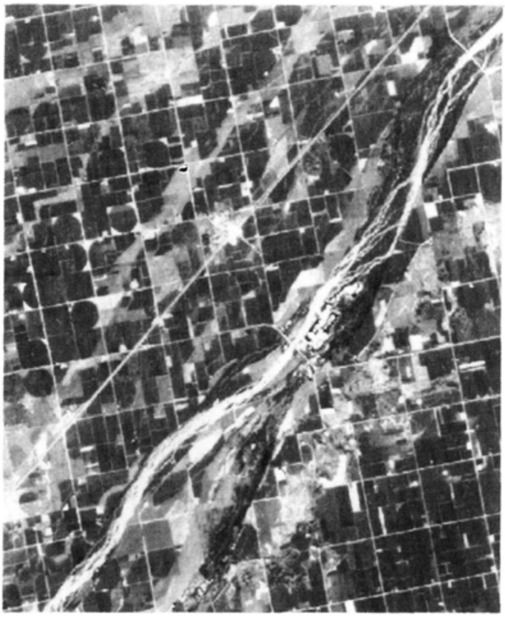


Figure 1. Landsat Thematic Mapper (TM) image subset of the Platte River valley study area, acquired on 19 August 1992, band 3.

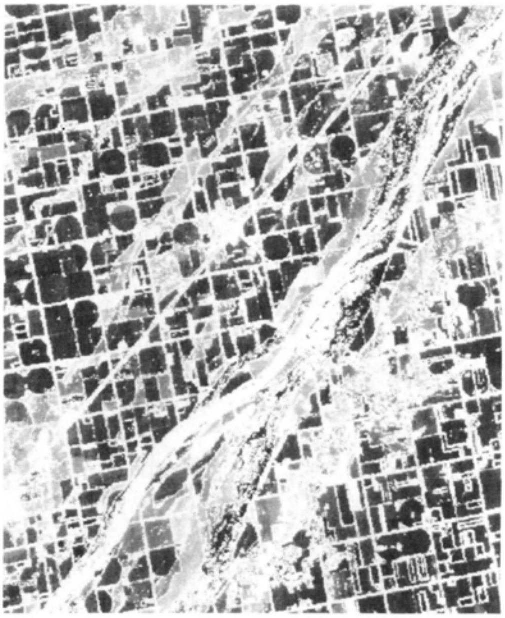
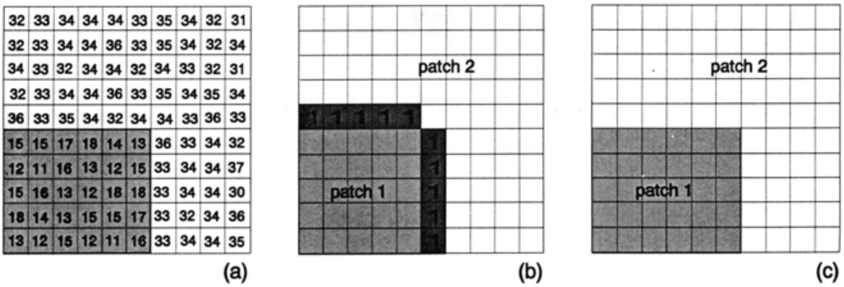


Figure 2. Results of the directional first differencing algorithms overlaid on TM band 3.



32	33	34	34	34	33	35	34	32	31
32	33	34	34	36	33	35	34	32	34
34	33	32	34	34	32	34	33	32	31
32	33	34	34	36	33	35	34	35	34
36	33	35	34	32	34	34	33	36	33
15	15	17	18	14	13	36	33	34	32
12	11	16	13	12	15	33	34	34	37
15	16	13	12	18	18	33	34	34	30
18	14	13	15	15	17	33	32	34	36
13	12	15	12	11	16	33	34	34	35

Figure 3. Edge detection and modification process: (a) a single-band sample of the original digital image data, (b) patches clumped based on edge information, and (c) border pixels assigned to their respective patches.

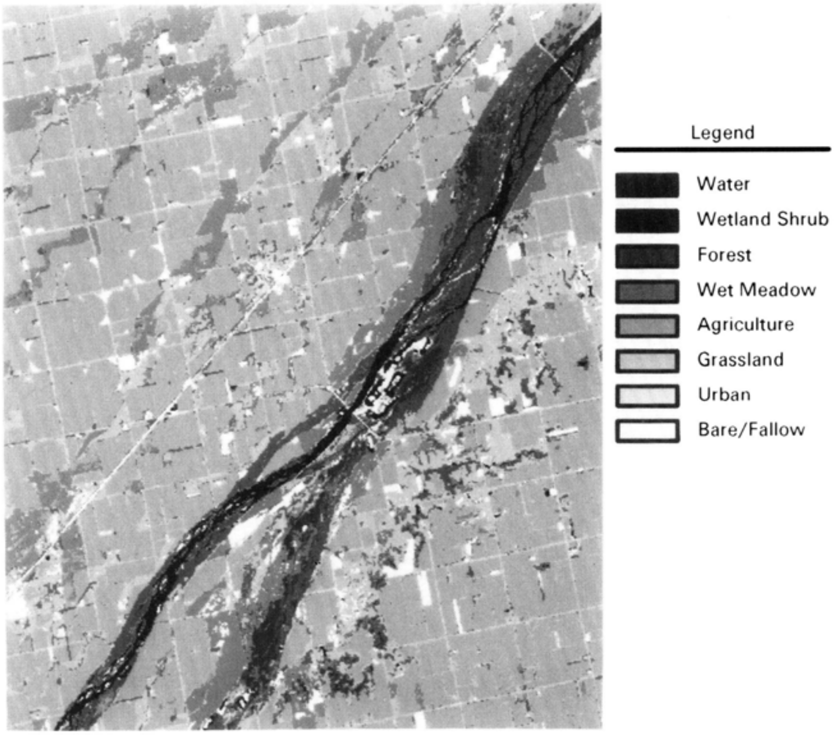


Figure 4. Final land use/land cover classification derived after application of the stratification methodology.

Late Submissions

AUTOMATIC MAP FEATURE EXTRACTION USING ARTIFICIAL NEURAL NETWORKS

Itthi Trisirisatayawong and Mark R. Shortis
Department of Geomatics
University of Melbourne
Parkville VIC 3052 AUSTRALIA
Mark_Shortis@mac.unimelb.edu.au

ABSTRACT

This paper describes the implementation of and experimentation with multi-layer feedforward neural networks to extract particular map features from scanned topographic maps. Commercial scan-conversion systems for automatic map input exist, but their capabilities are limited to vectorisation and other post-conversion processes on clean single-theme map images. This limitation impedes their application on paper maps which are in far more common use than single-theme images. The central issue of this paper is a technique that can be used as generally as possible as a mechanism to extract single-theme data from multi-theme documents. Backpropagation neural networks can establish any complex decision boundaries and therefore can be applied in any classification problem, which makes the technique attractive as a base for the development of such a mechanism. Experiments are carried out on scanned images of two topographic maps at different scales. The results demonstrate that neural networks have the potential to be implemented for automatic map data acquisition for GIS.

INTRODUCTION

The creation of a clean digital databases is a most important and complex task, upon which the usefulness of GIS depends (Burrough, 1988, p.56). Of the number of sources that can be used as input to GIS, secondary data like maps has been the most important source because of the variety, accuracy and effectiveness of maps to convey information about real-world phenomena, and their relationships, to the users. However, although maps are very efficient stores of information, it is surprisingly difficult to obtain certain types of numeric information from them (Goodchild, 1990). To date, the tool normally offered by commercial GISs to capture map data is manual digitisation using a hand-held cursor. It is well known that data capture by this method is slow, inconsistent and error-prone, so spatial database creation is expensive. Screen-based or head-up digitisation may eliminate the inefficiency of looking back and forth between the digitising table and screen, but in order to achieve the accuracy required, it is necessary to magnify feature data. The time required to change view windows in high magnification modes often makes capture of spatial features more time-consuming with screen digitising than similar accuracies achieved by using conventional digitising tablets (Skiles, 1990).

Since the early 1970s, many commercial systems have been offering automated line-tracing, a technique for rapid generation of vector data from line maps. The line-tracing system may be controlled by special hardware devices and software, or purely by software. The core of a scan-conversion system is vectorisation, whose fundamental requirement is that clean single-theme maps such as map separates are available. This assumption leads to a very narrow, well-defined problem domain which allows commercial development. For many reasons, however, clean single-theme maps may not be available (Faust, 1987). This situation is more severe in most developing countries where map separates are strictly controlled, mainly for reasons of national security. Consequently, the application of automatic conversion systems on multi-theme map documents, which in are far more common use than map separates, are impeded by the same assumption that allows their commercial development.

Automatic map-conversion systems can be used more widely if the assumption about the availability of single-theme maps is removed. In other words, a process that can extract particular features from scanned map images and feed them to a vectorisation process is needed. The demand on this missing component will accelerate due to the fall in prices of commercial desktop scanners in recent years, since this means that most organisations now can afford to routinely capture map data in an automatic manner. Unfortunately, the level of automation of feature extraction is far behind vectorisation and other processes thereafter. This creates a situation in which single-theme data may be reproduced by manual tracing on original multi-theme documents before being scanned and vectorised. It can be clearly seen that this practice is in fact equivalent to manual digitisation and thus suffers all the same drawbacks.

There has been little reported research concerned with automatic feature extraction from images produced by scanning multi-theme maps. Fain (1987) addressed pixel classification as part of a solution to automatic data conversion, but no indications about source documents, techniques and results were given and the emphasis was on interactive editing rather than automatic methods. In the paper of Konty et al (1991), some information about test documents was available but the core work is a benchmark test of commercial systems without any attempt to evaluate the underlying techniques. Of the more technical work, Eveleigh and Potter (1989) reported a preliminary study of using the Bayes technique to classify a USGS 7.5 minute quadrangle covering a rural area. The study, however, did not indicate whether the RGB intensity values of the map image had the normal distribution assumed by the Bayes classification technique and neither was any classified result nor quantitative information revealed.

Previous research in automatic map feature extraction concentrates on the implementation of techniques that can be possibly used for particular test maps. The important issue of the generality of the employed techniques has not yet been systematically studied. This paper investigates neural network techniques, which have the potential to be implemented as a general feature extraction mechanism. The property of neural networks that makes it a powerful technique is described first, followed by discussion of the experiments carried out on test maps.

NEURAL NETWORK FEATURE EXTRACTOR

Feature extraction is achievable by classification. This popular scheme in pattern recognition employs a classifier to classify image objects according to their characteristics. A classifiable characteristic is any numeric information used to distinguish one part of image from other parts. The objects may be single pixels or groups of contiguous pixels depending on the level of abstraction at which classification is performed. The derived characteristics are fed into the classifier which produces class labels for objects.

A classification technique may be categorised as supervised or unsupervised. Basically, an unsupervised classification is a data clustering technique whose fundamental idea is that different feature classes should form different clusters in characteristic space. Thus, unsupervised classification is done by grouping objects into clusters based on the criterion that the clusters should be formed as compactly or tightly grouped as possible. Other criteria or additional processing may be employed to enhance the separation between clusters. These techniques are called unsupervised because they make no use of external knowledge or the characteristics of feature classes.

Unsupervised classification is mainly used when there is little information regarding what features the scene contains. This is particularly true for satellite images which may cover inaccessible areas and, in such circumstances, it is impractical or too expensive to obtain sample data. The drawback of unsupervised techniques is that feature classes may only be marginally separable or may not form obvious clusters at all. Unsupervised classification cannot provide correct results in such situations and this problem prevents the technique from being used as a general tool.

On scanned map images, the problem of the unavailability of sample data is certainly not the case and supervised techniques are the obvious choice for the classification task. Instead of relying on compactness, which may not truly reflect the data structure, supervised techniques utilise information contained in the sample data to establish decision boundaries between feature classes.

The classification process can be geometrically interpreted as the establishment of decision boundaries in the characteristic space. The more complex the decision boundaries a classifier can establish, the more general it is. The most important of traditional supervised classifications is the Bayes technique. In the Bayes technique, the sample data is used to estimate the parameters, for instance the mean vector and covariance matrix, of the multivariate normal distribution. The bell shapes of normal probability density functions in two dimensional space mean that the decision boundaries drawn by the Bayes technique are hyper-ellipsoids in *n* dimensional space.

A new and increasingly important technique is backpropagation neural networks. The interest in the application of backpropagation neural networks for classification problems is driven by the analysis shown in Lippmann (1987) that a single hidden-layer backpropagation neural network can form any, possibly unbounded, convex region in *n* dimensional space. A simple extension to a two hidden-layer network can establish arbitrarily complex decision boundaries which can separate even meshed classes. Although the latter case is not explored in this paper, this capability means that, in theory, backpropagation neural networks can be used for virtually any classification problem regardless of the statistical distribution of data. It follows that the backpropagation neural network technique is a more general technique than the Bayes technique, which is based on the assumption that the data is normally distributed.

However, accuracy also needs to be taken into account to assess the validity of the generality. It has long been known that when the underlying assumption about the statistical distribution is met, the Bayes classifier provides an optimum result. This important property has established the Bayes technique as a benchmark against which any newly proposed technique has to measure. The next section compares performances of both techniques when applied to a scanned topographic map.

FEATURE EXTRACTION AT PIXEL LEVEL

The performance analysis of backpropagation neural network classifiers has been undertaken on remotely-sensed data by a number of reported research works, many including a comparison with the Bayes technique (Howald 1989; Hepner et al 1990; Heermann and Khazenie 1992; Bischof et al 1992). However, despite the importance of these two techniques, little attention has been paid by the mapping/GIS community for such studies on map data. The comparison of Bayes and backpropagation neural network techniques on the segmentation of real map images was probably first reported by Trisirisatayawong and Shortis (1993). The study used RGB spectral characteristics to classify a portion of 1:100,000 Australian topographic map which was scanned at a resolution of 150 dots per inch. The advantages and disadvantages of both methods in practical issues were discussed, but a detailed analysis of data and performance of each method on each feature was beyond the scope of the paper.

The analysis below is an extension of the above-mentioned work. Figure 1 illustrates the test map which is shown in grey-scale. The statistics of map features are shown in table 1. Some of the results from Bayes and one hidden-layer neural network classifiers are shown in figure 2.

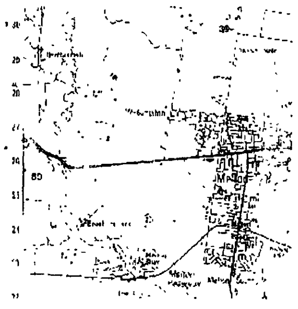


Figure 1: Test map (from the colour original at 1:100,000 scale).

Crown Copyright reproduced with the permission of the General Manager, Australian Surveying and Land Information Group, Department of the Arts and Administrative Services, Canberra, Australia

Feature	Number of Samples	Mean			Standard Deviation			Skew		
		R	G	B	R	G	B	R	G	B
Built areas	39	237	153	155	8	14	14	-0.5	0.0	-0.4
Contours	29	216	172	140	19	25	34	-0.5	0.0	0.6
Forest	50	222	233	155	21	16	36	-0.7	1.6	0.8
Roads	42	213	106	87	45	39	43	-2.2	0.2	0.3
Water	39	161	193	244	35	32	25	-0.8	-1.2	-3.7
Dark	36	80	70	83	35	24	28	1.8	1.1	0.2

Table 1: Statistics of spectral characteristics of features on the test map shown in figure 1.

Skewness, the magnitude of which indicates the degree of deviation of the data from a symmetrical distribution, of RGB characteristics is computed for each feature class. Large skewness values mean that the data is significantly skewed whereas smaller values indicate otherwise. Using skewness values as indicators, it can be seen that for the feature classes such as water, roads and forest, the sample data distributions are substantially skewed and therefore cannot be normal. It is therefore expected that the backpropagation neural network produces more accurate results, and this is evident in figure 1, particularly for the water image of rivers and lakes. On the layers of built areas and contour lines, whose spectral characteristics can be properly assumed to have normal distributions because of small skewness, the results from the two techniques are essentially similar.

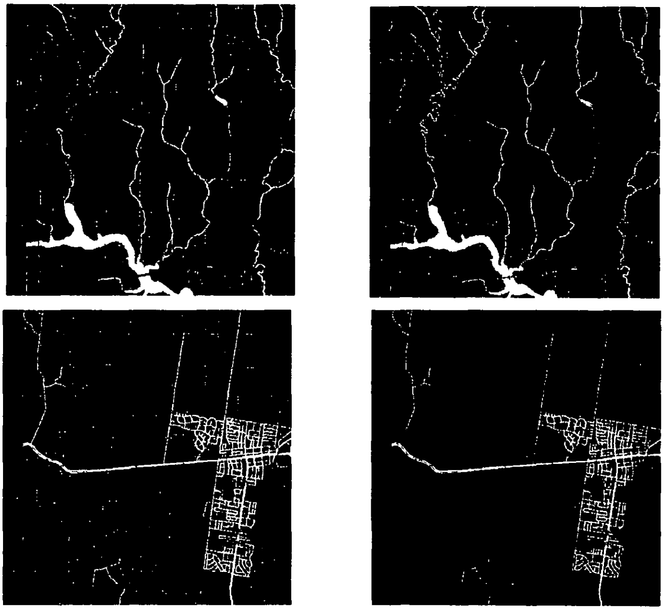


Figure 2a: From top to bottom: Classified images of the features of water bodies and roads respectively. The images on the left and right hand sides are results from the Bayes and backpropagation neural network techniques respectively.

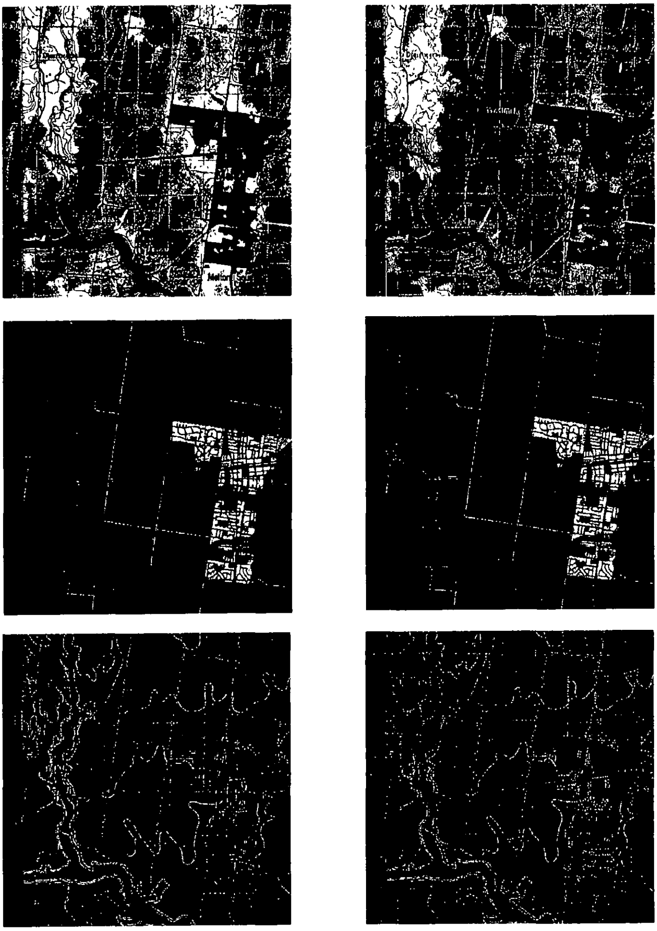


Figure 2b: From top to bottom: Classified images of the features of forest, built areas and contour lines respectively. The images on the left and right hand sides are results from the Bayes and backpropagation neural network techniques respectively.

The results indicate that a backpropagation neural network with one hidden-layer can be implemented for map-feature extraction at the pixel level. In the non-normal cases, the neural network provides better results, but even when the characteristics have a normal distribution, results provided by the Bayes bench mark are satisfactorily approximated by the neural network. Thus, the statistical distribution constraint in the Bayes technique is removed when the alternative neural network technique is applied, solving the same problem with equivalent accuracy. The proposition that neural networks are a general classifier is supported by the experimental results

However, not every feature can be extracted using a multispectral classification. Different features may have the same colour and it is not possible to differentiate one feature from another regardless of the technique used. An example of this problem is shown in figure 3 in which railway, text, house symbols and tracks are incorrectly assigned into the same class. Another example of the same situation is shown figure 4, which is an image resulted from a neural network multispectral classification of another scanned topographic map (original scale 1:50000, scanned at 300 dots per inch). The fact that these features hold the same spectral characteristics means that there is no way at pixel level to avoid the misclassification. Thus, multispectral classification provides only partial solution to the problem of map-feature extraction. Other techniques need to be utilised to resolve ambiguities resulting from the initial spectral classification.

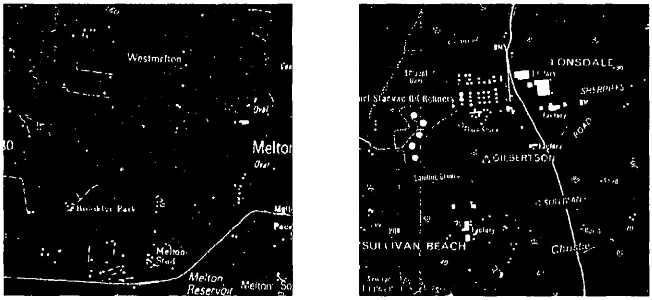


Figure 3 (left) and 4 (right): Images of mixed features resulting from a multispectral classification.

FEATURE EXTRACTION BY SHAPE ANALYSIS

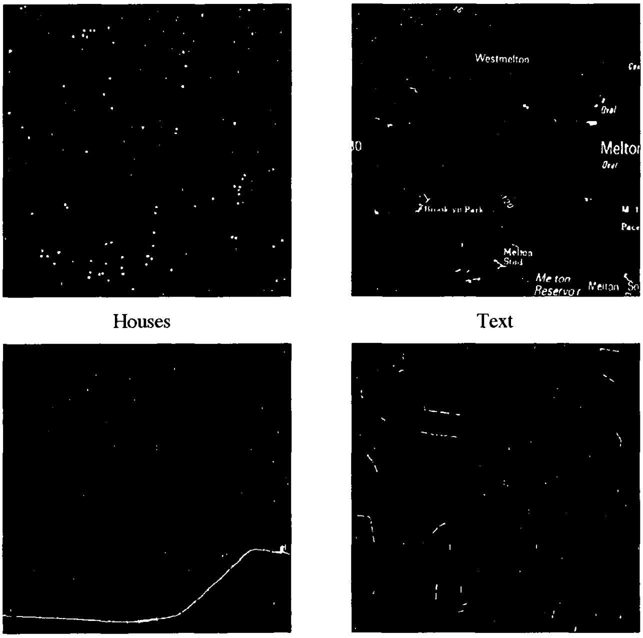
Further extraction on the image of mixed features must be carried out to produce single-theme images. If the appearances of map features are somewhat consistent, templates which define their likely pixel arrangement can be used in a classification by looking at the degree of match or similarity between the image part under consideration and the template being applied. The serious disadvantage of this simple technique is its inability to handle variations in shape and size of each feature type, which is normally the case in most maps. A very large number of templates must be defined to cover all possible occurrences and this may incur extremely heavy computational load. This may lead to an unacceptable situation in which even a high-speed processor will take several hours to locate features within the image. Another situation in which template matching techniques are not suitable is when the appearance of one feature is a part of another larger-sized feature. For example, the letter I also appears in the left portion of the letters B, D, P. Misidentifications (or false alarms) will occur when B, D, P are matched by the template of I. There is no universally effective solution for this problem.

Intuitively, sets of contiguous foreground pixels displayed as regions, identified by a pixel level classifier, can be treated as individual image objects. These objects can be further classified based on the similarities and differences in shapes. Thus, the concept that feature extraction can be formulated as a classification is still applicable, provided that shape information is properly quantified.

The shape characteristics must be tolerant to transformation and uniquely defined by the objects if the problems of the template matching method are to be avoided. Basic shape characteristics are those related to size such as area, perimeter and extent. These characteristics are invariant to translation and rotation but are affected by scale. A possible way to obtain scale-invariant characteristics is by relating the given measurements of objects to some well known geometric figure such as a circle. The result is dimensionless shape measurements which are invariant under magnification or reduction. For example, a compactness ratio could be derived by dividing the area of the object by the area of the circle having the same perimeter as the object. However, although it is possible to produce characteristics which are invariant to translation, rotation and scale in this way, there is no guarantee that two different object types will not produce the same characteristics. The use of shape values which are not uniquely defined by objects prevents the implemented classification technique on a particular test map to be subsequently applied to different map images.

The theory of moment invariants can be applied to produce object characteristics that are invariant under transformation and uniquely defined by objects. This analytical method was first introduced to the classification problem by Hu (1962). Details of moment invariants is omitted here but can be found in Hu (1962). Since its introduction, moment invariants have been used in aircraft identification (Belkasim et al, 1991), detection of ships in remotely sensed images (Smith and Wright 1971), and optical character recognition (El-Dabi et al 1990). All of this research applied moment characteristics in conjunction with conventional supervised or unsupervised classifiers. The performance of the technique combining moment characteristics with neural networks has not yet been explored.

In theory, coupling a neural network classifier with moment characteristics should result in a general classification technique. Of the infinite number of moments that can be chosen, only three, namely m00 (area or number of pixels comprising an object), M1 (spread) and M2 (elongation), are employed. The selection of this subset of moments is based on the consideration that these three values carry substantial shape information and should contain discrimination power adequate for classifying objects within a map image. Statistics of the three moment characteristics of the test image of figure 3 are shown in table 2 and the classified image results from using a one hidden-layer backpropagation neural net are shown in figure 5.



Houses

Text

Railways

Dashed lines

Figure 5: Images of the object classifications of the map image shown in figure 3.

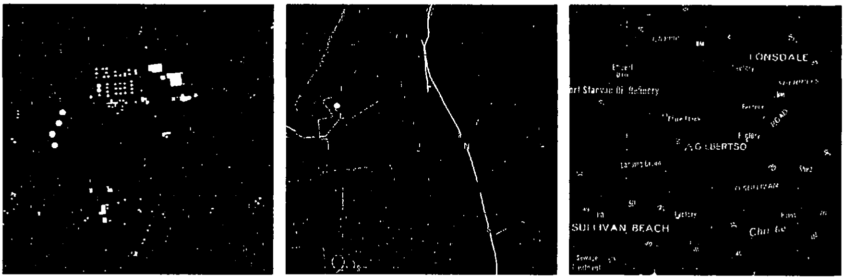
The results clearly illustrate that extraction of features from the test image is achieved with a high degree of accuracy and completeness. However, there are what seem to be misclassifications appearing in each classified image. If the classification of objects into a class is posed as the null hypothesis in statistical test, then it can be seen that most of the misclassifications are type-two errors. The classified images of house symbols and railways are free of type-one errors. In the text image, there are a few type-one errors but all of them occur from characters having similar shapes to dashed lines. This is reasonable since there is no way for the classifier to identify these characters without the help of extra information, such as context.

Object	Mean			Standard Deviation			Skew	
	m00	M1	M2	m00	M1	M2	M1	M2
Dashed lines	49	60	57	7	13	12	-0.38	-0.42
Houses	38	16	3	5	1	2	0.00	0.25
Railway	2719	1526	1414	0	0	0	0.00	0.00
Text	111	226	21	50	16	23	0.34	0.41

Table 2: Statistics of moment characteristics of map objects in figure 3.

Every classified image suffers from a different degree of type-two errors. The most serious case is the text image. However, almost all of the errors occur from compound objects which are an incorrect aggregation of two or more objects. Considering that there is no class representing them and they are not used in the training phase, these type-two errors are not mistakes of the classifier. In fact, a visual inspection reveals that, except in one instance where two characters on the railway image are mistakenly joined by the pixel classification, all compound objects appear on the original document or are a result of the finite sampling size of the scanning process.

A similar process of classification by moment characteristics performed on the test image of figure 4 produces similar results. In this case a slight modification is made. The number of classes of line features is restricted to one only, since each line type has only a few objects. So, there are three classes representing points, lines and text with another class being assigned as a noise channel. The results are illustrated in figure 6 below.



Points Lines Text

Figure 6: Images of the object classifications of the map image shown in figure 4.

Like the previous analysis, it can be seen that most of the features have been correctly classified. Except for two point symbols being misclassified into the layer of text and three elongated characters being incorrectly assigned to the line layer, most of the problems are caused by compound objects. White noise appearing in the classified images can be simply removed by size criterion. The classification accuracy of both text and point symbols are in the high 90 percent range and this is achieved without any extra information, such as contextual information, which certainly will enhance the results.

A number of further processes are required to convert the classified objects into features appropriate for the generation of a spatial database. Aggregated objects must be separated using, for example, mathematical morphology techniques (Trisirisatayawong, 1994). Line objects must be vectorised, text must be recognised and linked to associated map features

(Shortis and Trisirisatayawong, 1994), and a final phase of attribute tagging must be conducted (Trisirisatayawong, 1994).

CONCLUSIONS

Neural networks re-formulate all problems by finding the correct internal weights, so the technique can be viewed as a black-box problem solver in which the weights have no obvious physical meaning in the context of problem. The statistical distribution of data is insignificant compared to other traditional classifiers. This means that neural networks can be universally applied to all classification problems, provided the network is properly trained by appropriate and accurate sampling.

A drawback of the neural network technique is that it is often difficult to determine whether the neural network is correctly trained. Learning error is the only information used by the neural network to indicate the degree of success. There is no guarantee that when the error has converged to a particular value that it is the global, rather than a local, minimum. So, the magnitude of error often does not truly reflect the degree of learning. One widely-used practice is to set an acceptable error threshold and the network is accepted as adequately trained once the learning error has converged to a value less than this threshold. Thus, the amount of training of the network is subjectively determined by the operator, who must specify the threshold based on experience or any other suitable guideline.

Neural networks are extremely flexible in solving a wide variety of problems. The key factor determining the accuracy of a neural network is its structure, which can be constructed as single hidden-layer, multiple hidden-layer, partial inter-layer connection, full connection or other varieties. However, it also means that different neural networks may be constructed to solve the same problem and so, in mathematical sense, the technique of neural networks does not provide a unique solution. There is no general rule to determine whether the chosen structure is optimum. The most serious problem in practice is the determination of the learning rate, the initial weights and especially the structure of neural networks. All of these factors must be pre-determined by the operator who will in general set them from prior knowledge and experience.

Nevertheless, the drawbacks of neural networks occur mostly because the technique is still a relatively young science. The problems will dissipate as the knowledge of neural computing expands. For example, some guidelines about the determination of learning rate can be found in Kung and Hwang (1988), although the final settings must still be determined on a trial and error basis. Also, research on the automation of the determination of structure and the removal of redundant elements in the network to improve efficiency are under way (Wen et al 1992). The efforts in these areas will lead to less time and frustration incurred from training neural networks in the future.

Overall, the advantages of neural networks as a general classifier outweigh the disadvantages. As the experimental results on real map data shown here firmly support the theoretical claims, it is believed that neural networks can be further developed as general map feature extraction mechanism.

REFERENCES

Belkasim S.O., Shridhar M., and Ahmadi M., 1991, "Pattern Recognition with Moment Invariants: A Comparative Study and New Results", *Pattern Recognition*, Vol. 24, No. 12, pp. 1117-1138.

Bischof H., Schneider W., and Pinz A.J., 1992, "Multispectral Classification of Landsat-Images Using Neural Networks", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 30, No. 3, pp. 482-490.

Burrough P.A., 1988, *Principles of Geographical Information Systems for Land Resources Assessment*, Reprinted, Oxford University Press, UK.

El-Dabi S.S., Ramsis R., and Kamel A., 1990, "Arabic Character Recognition System: A Statistical Approach for Recognizing Cursive Typewritten Text", *Pattern Recognition*, Vol. 23, No. 5, pp. 485-495.

Eveleigh T.J., and Potter K.D., 1989, "Spectral/Spatial Exploitation of Digital Raster Graphic Map Products for Improved Data Extraction", *Proceedings of Auto-Carto 9*, Baltimore, Maryland, pp. 348-356.

Fain M.A., 1987, "Optical Scanning and Interactive Color Graphics: An Integrated Approach to Automated Data Capture", *Technical Papers of the 1987 ASPRS-ACSM Annual Convention*, Baltimore, Maryland, Vol. 2, pp. 323-326.

Faust N.L., 1987, "Automated Data Capture for Geographic Information System", *Photogrammetric Engineering and Remote Sensing*, Vol. 53, No. 10, pp. 1389-1390.

Goodchild M.F., 1990, "Geographic Information Systems and Cartography", *Cartography*, Vol. 19, No. 1, pp. 1-13.

Heermann P.D., and Khazenie N., 1992, "Classification of Multispectral Remote Sensing Using a Back-Propagation Neural Network", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 30, No. 1, pp. 81-88.

Hepner G.F., Logan T., Riter N., and Bryant N., 1990, "Artificial Neural Network Classification Using a Minimal Training Set: Comparison to Conventional Supervised Classification", *Photogrammetric Engineering and Remote Sensing*, Vol. 56, No. 4, pp. 469-473.

Howald K.J., 1989, "Neural Network Image Classification", *Technical Papers of ASPRS-ACSM Fall Convention*, Cleveland, pp. 207-215.

Hu M.K., 1962, "Visual Pattern Recognition by Moment Invariants", *IRE Transaction on Information Theory*, Vol. 8, No. 2, pp. 179-187.

Konty L., and Gross G., 1991, "Scanning for GIS Data Conversion: Perils, Pitfalls & Potentials", *URISA Proceedings*, San Francisco, Vol. 2, pp. 192-197.

Kung S.Y., and Hwang J.N., 1988, "An Algebraic Projection Analysis for Optimal Hidden Units Size and Learning Rates in Back-Propagation Learning", Proceedings of IEEE International Conference on Neural Network, San Diego, Vol. 1, pp. 363-370.

Lippmann R.P., 1987, "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, Vol. 4, No.2, pp. 4-22.

Shortis, M. R. and Trisirisatayawong, I., 1994, "Automatic text recognition and the potential for text to feature association on scanned maps", *Australian Journal of Geodesy, Photogrammetry and Surveying*, (in press).

Skiles J.M., 1990, "Using Scanning, Automated Conversion, and Raster Data to Speed GIS Implementation and Lower Cost", *Proceedings of GIS/LIS'90*, Anaheim, California, pp. 476-483.

Smith F.W., and Wright M.H., 1971, "Automatic Ship Photo Interpretation by the Method of Moments", *IEEE Transactions on Computers*, Vol. 20, No. 9, pp.1089-1095.

Trisirisatayawong I., 1994, "Automatic Feature Extraction from Scanned Topographic Maps", unpublished Ph.D. thesis, The University of Melbourne, Australia, 152 pages.

Trisirisatayawong I., and Shortis M.R., 1993, "A Comparison of Two Feature Classification Methods for Scanned Topographic Maps", *Cartography*, Vol. 22, No. 1, pp. 1-14.

Wen W.X., Liu H., and Jennings A., 1992, "Self-Generating Neural Networks", Proceedings of the International Joint Conference on Neural Networks, Baltimore, Maryland, Vol. 2, pp. 111-117.

NOTES

NOTES

NOTES

NOTES

NOTES

NOTES

ISBN 1-57083-019-3
ISBN 1-57083-018-5