TOWARDS COMPREHENSION OF DATA QUALITY AND UNCERTAINTY IN DIGITAL CARTOGRAPHY

Trevor J. Davis
Department of Geography
University of British Columbia
1984 West Mall
Vancouver, B.C. Canada, V6T 1Z2
Internet: tdavis@geog.ubc.ca

C. Peter Keller
Department of Geography
University of Victoria
PO Box 3050
Victoria, B.C., Canada, V8W 3P5
Internet: keller@geography.geog.uvic.ca

Abstract

This paper discusses components of a research project where several types of incompatible spatial data uncertainties are modelled and propagated through a complex GIS-based analytical routine [10]. The procedure utilizes two modelling methods: fuzzy logic for spatial uncertainty data, and Monte Carlo simulation for spatial/ non-spatial variance information. The paper reports on uncertainty tracking, commencing with the transition of the spatial data from source paper maps and field reports, through analysis and modelling, to cartographic output required to visualize the uncertainty dimensions. As holds true for traditional cartography, the presentation of information about uncertainty is key to utility. The research [10] and the presentation utilize a variety of proven choropleth techniques in concert with animation capabilities of modern workstations to summarize the uncertainty modelled. This paper focuses on the visualization problem; however, the results cannot be fully presented due to the limitations of a static, monochrome medium.

This paper concludes by arguing that we should abandon some of our traditional techniques of data storage and presentation by moving to data structures and visualization tools that reflect a more accurate version of inherent and introduced uncertainties in cartographic data.

1 Introduction

Geographic Information Systems (GIS) increasingly can be found on the desks of public utility planners, natural resource scientists, and almost anyone else concerned with spatially referenced data. Perhaps the best applications to date have been with the former, for utility management focuses on straight lines and unambiguous locations. Detailed coordinates and precise analytical and modelling routines supported by GIS cater to a utility manager's desire to see the world in black-and-white. Unfortunately, the same black-and-white routines and data structures are also applied in the natural resource sector, where entities in question might be better represented by shades of gray.

This dichotomy in natural resource applications between imprecise reality and its precise digital representation has given rise to a rapidly growing research area in which spatial data experts grapple with the implications of uncertainty and error analysis. As this field has developed, researchers have fanned out across a broad front:

advancing error analysis (e.g., [8, 9]), locational and feature uncertainty analysis [6], error propagation methods [18], the visualization of uncertainty [3], and numerous others.

One particularly difficult problem in this field involves addressing the various *types* of error that exist. Positional error can be represented by an error ellipse (2D) or sphere (3D). However, thematic variables can have many other uncertainty aspects, including:

- uncertainty regarding classification;
- uncertainty regarding classification divisions;
- inherent uncertainty (regarding the resource itself and data gathering techniques);
- model uncertainty (when combining thematic variables);
- uncertainty that varies spatially (e.g. between polygons); and
- error 'envelopes' around non-classified items (e.g. elevation).

When spatial data are combined in a GIS, a common data format is required. Dealing with many types of uncertainty and error will also require some type of common denominator. All of the above, with the exception of the final item, can be manipulated in a common format using fuzzy set theory (e.g. [5, 20, 23]). The partial memberships allowed in fuzzy set theory are particularly well-suited to representing uncertainty in resource data. For example, a fuzzy index of 0.75 at a particular (x,y) location may indicate that the data gathering techniques do not allow absolute certainty regarding the forest type; or, alternately, that the forest at that site demonstrates most of the characteristics of "mature forest" but some of "immature forest".

However, numerical estimates of error (e.g. elevation = 225m ±5m) cannot be incorporated into a fuzzy-based system. In concept they differ substantially from uncertainty information. This difference demands that errors and uncertainties receive separate processing in an integrated analysis.

The research reported in this paper utilizes the concept of spatially variable uncertainty in the form of "fuzzy surfaces" to model the uncertainties in a GIS-based analysis of slope stability. Error elements in the analysis are modelled separately using Monte Carlo simulation. Data from a forested coastal site are utilized to test the uncertainty/error modelling routine. The primary focus of the paper is on the visualization methods required to bring together these uncertainties and errors. This visualization is an attempt to effectively communicate the model's results and their actual utility in light of built-in uncertainty and error.

One important issue raised in this work is the need to address the spatial variability in the uncertainty model. The wide variety of potential errors and uncertainties involved with almost any natural resource data set precludes the use of simple non-spatial meta-data. The recognition of the need for meta-data found in recent initiatives such as the U.S. Spatial Data Transfer Standard should be lauded. However, methods such as these that simply append a meta-data index to each existing polygon (or to each map sheet) ignore the potential spatial variability of uncertainty. In the context of both exploratory spatial data analysis (ESDA) and spatial decision support systems (SDSS), the concept of homogenous objects (i.e. standard polygons) with homogenous attributes is increasingly difficult to defend. Intelligent manipulation at a variety of scales requires knowledge regarding the complex spatially-variable interactions between these objects. Some, such as clear-cut/forest boundaries, have little added uncertainty. Others, such as soil polygons, may have complex, multi-leveled interaction — increasing uncertainty significantly in specific areas. Due to this high degree of complexity, the model presented here will utilize a surficial, rather than object-oriented structure to address this key element of uncertainty.

2 Techniques

A slope stability scenario was chosen to demonstrate this spatially-variable error/uncertainty model. This highly typical type of analysis utilizes data from a variety of sources, each of which demonstrates one or more of the uncertainties listed above. Slope stability analysis depends significantly upon soil data—typically stored as polygons. Deriving the uncertainties while converting such polygons to uncertainty surfaces will serve to demonstrate techniques of manipulating spatially-variable uncertainty. The wide variety of uncertainty data will also allow the exploration of a diverse set of visualization techniques.

The infinite slope stability model is utilized to calculate a factor of safety for potential soil slippage. Predicting slope stability is particularly important on Canada's West Coast, as seasonal peaks in precipitation can often cause mass movement, inflicting damage on buildings or roads located down-slope. The infinite slope stability model requires soil type, slope, and forest cover as source data. From these, values are derived for the model inputs of root depth, soil cohesion, slope, and several other minor items. The uncertainties and errors in these data sources may be classified as follows:

1) uncertainty regarding soil or forest type classification (e.g., soil classed type A, might be B);
2) uncertainty regarding data gathering (e.g., 10% of polygons misclassified);
3) spatial uncertainties (e.g., the uncertainty of item 1 increases near polygon boundaries);
4) error envelopes around derived items (e.g., cohesion of soil type A = Y ±x); and
5) error envelopes around elevation values.

Information regarding the first two uncertainties was gathered from discussions with soil scientists that had personal experience in the study area. Ideally, such information should come from the initial survey or a physical re-sampling of polygon boundaries. However, the Semantic Import Model [26] used to quantify this information is argued to act as a reasonable substitute. In the Semantic Import Model, phrases such as 'close to' or 'nearly' are translated into fuzzy classifiers (see [5]).

Information regarding the third item, spatial uncertainties, was also gathered from expert opinion via Semantic Import and was operationalized using the following 'corridor of transition' algorithm.

A central ridge is defined in each source polygon that is located as far from a boundary as possible without being closer to another boundary: a 'max-min ridge'. The slope of the uncertainty values at the polygon/polygon interface is constrained by the two polygon types as well as the effective size of each polygon, determined from the shortest distance to a 'max-min ridge'. This 'corridor or transition' model acts as a refinement of the epsilon boundary model [4, 7, 25] as well as of Mark and Csillag's [22] parametric function model.

Upon completion of this algorithm, each of the possible soil classifications is assigned an 'uncertainty surface' which defines the likelihood of finding that particular type at a given (x,y) location. A similar procedure is performed for the forest coverage.

The fourth and fifth items in the list, the error envelopes, were gathered from both literature reviews (in the case of soil attributes) and published error values (in the case of elevation data). Given that each of these error terms are distributed normally (or in some cases log-normally) about the mean, a Monte Carlo simulation is utilized to determine the output of the slope stability equation. Values are chosen randomly from within the error distribution and assigned to the equation. The randomization procedure is repeated a sufficient number of times to determine the mean and standard deviation of the outputs (in this case 50 runs was chosen as a conservative figure; see [24] for a discussion). As a variety of soil and forest types are possible at any given location, the entire Monte Carlo procedure is repeated for every possible combination, excluding those highly unlikely (e.g. bedrock with mature forest).

The uncertainties accompanying each surface are combined with a 'fuzzy AND', also known as a 'joint membership function' [6]; in this case a MIN function suffices to produce the output uncertainty surface. In the end, three surfaces are required to describe each combination of forest/soil possibilities: mean, standard deviation, and uncertainty. The large number of resulting surfaces serve to maintain maximum information content for the exploratory visualization to follow.

3 Case Study

An 8500 hectare study site was selected. The area is located on Louise Island, on the east side of Moresby Island in the Queen Charlotte Group, British Columbia, Canada, at 53°N, 132°E (see Figure 1). The area is a forest company test site, and was selected for: 1) the availability of data, and 2) the existence of prior slope stability studies for comparison.

Several examples of the fuzzy surface resulting from the transition corridor algorithm applied to the case study data are illustrated in Figure 2. Note the asymmetric slopes on the boundary between two thematically similar, yet differently sized polygons in Figure 2a. Figure 2b details a bedrock/soil boundary.

The 'non-spatial' items were gathered from an extensive literature review undertaken by the US Forest Service Intermountain Research Station [17] while developing their slope stability modelling system. Soil types found in the Louise Island study site were matched with the classification system used by the USFS, and means and standard deviations of the relevant data were calculated. The USFS study found that, for the most part, the values for each variable are normally distributed, with the exceptions of soil cohesion and root cohesion which are log-normal.

Elevation data consisting of a semi-regular grid of elevation spot heights (British Columbia TRIM data; [28]) produced from stereo-photogrammetry were utilized. Block kriging was used to interpolate from these points to a regular grid. Kriging was chosen as the interpolation method for two reasons: its high accuracy and an ability to produce variance maps of the derived values. The published error values were combined with the variance for each data point to produce a final variance value for each interpolated cell.

The Monte Carlo randomization was repeated 50 times for each forest/soil combination, then summarized in three final surfaces: mean, standard deviation and uncertainty, describing a log-normal result curve. A final tally of 60 surfaces were produced by the procedure.

4 Results

The results of any new data manipulation procedure or model are typically assessed via comparisons with existing techniques, with a focus on the new model's accuracy in predicting phenomena. Unfortunately, increased accuracy is precisely what the techniques of this study attempt to avoid; accurately representing real-world phenomena requires an increase in uncertainty. Most authors dealing with the uncertainty issue have also faced this conceptual road-block. Although they deal with it in a variety of ways, most maintain that the methods they advocate create maps arguably closer to ground truth than the original Boolean versions [6, 13, 21]. This argument is supported by us, with the stipulation that some method must be developed to compare the various

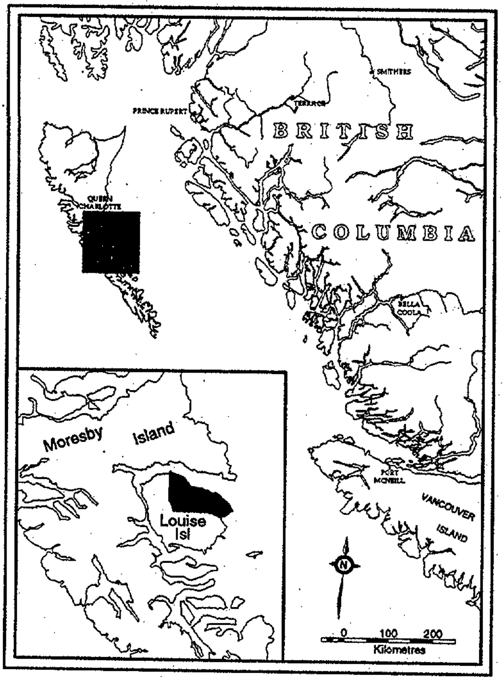

Figure 1: Louise Island Study Area

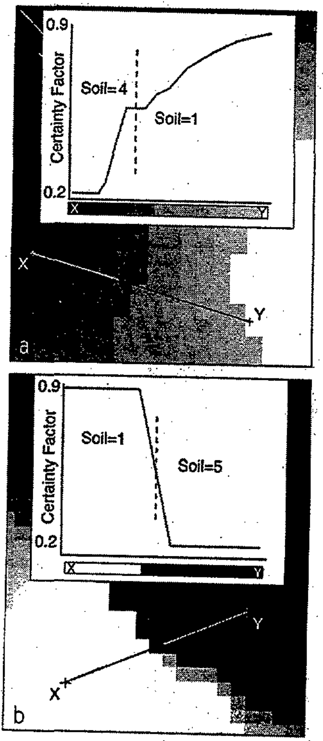

Figure 2: Examples of the boundary model

realizations of uncertainty that research in this area will generate. For now, an understanding of the results must focus on the visualization problem.

The visualization of uncertainty information is a substantial research topic in the geographical literature. As every application of GIS software has different requirements, so too must the visualization question be approached on a broad front. A gathering of specialists at a 1991 NCGIA conference on quality visualization [3] produced a telling discussion under the "research agenda" question. A major point of emphasis was the importance of interactive data exploration as opposed to quality representation for hypothesis testing. The range of disciplines represented led to numerous discrepant views about the nature of an error model and hence the nature of summary displays. One of the only points of agreement was the need to explore, compile examples, and understand the nature of the internal human mechanisms by which spatial or temporal patterns are interpreted.

The major problem encountered in adding uncertainty information to a visual summary is overload — users are not accustomed to dealing with variables that vary in value or certainty. Standard monochrome two-dimensional mapping techniques are not well-suited to incorporating these added dimensions in visual communication. Newer technologies offer several alternatives, such as:

- continuous spatial transition using colour or hue graduations;
- time dependence — animating displays;
- three dimensions — simulating 3-D objects; or
- multi-media — using sound, images and text simultaneously [3].

The raster data structure and continuously variable surfaces in this study eliminate one other commonly stated solution: error graphs. Such a graph might be generated for a particular pixel being queried. However, in our case the entire image is of importance in a summary display.

Fisher [11] uses animation to display the variability in soil maps — the image shifts in real time to show where variability occurs. Others have made use of sound [27] or other dynamic phenomena to display variability. This study will primarily utilize map comparisons, using the assumption that more than four variables (X,Y,Z and theme) in a 2-D display can be confusing for the viewer. Animation offers the opportunity of adding a fifth variable, as does the inclusion of contoured information overlays. All of the above are utilized in the full study (see [10]) and appear in the presentation; only a limited monochrome subset appear in this discussion.

4.1 Comparison with a Boolean Model

Direct comparison with a Boolean model requires the fuzzy results to be fixed, or "de-fuzzed", with some particular threshold value. There are several methods available; however the most appropriate in this situation might be utilizing a maximum-likelihood index of all potential representation of each pixel.

A maximum likelihood map utilizes the highest of the fuzzy joint-membership values and gathers the associated mean slope stability value. The resulting surface shows the most likely model results for every pixel on the map (Figure 3). This is the type of approach used by most non-spatial studies, including Burrough [5] and Fisher [11, 12]. The differences between this type of summary and the Boolean model are highlighted in Figure 4. The maximum-likelihood (M-L) summary map generally displays safer results on slopes while considering the plains more dangerous (probably due to minor variations in the DEM on the plain). However, since this type of summary ignores all variance information, direct comparison with the Boolean model can be misleading. The M-L map gives the appearance of "safer" than the Boolean version in areas of high variance.

The key element of a slope stability study in a "working forest" is the identification of the potentially most unstable zones. An M-L map only focuses on the most likely stability values. However, such a map does not make use of some of the potential of the uncertainty model. Any M-L map ignores the fact that there may be results, almost as likely, that describe more dangerous conditions. For example, a particular pixel might receive a value of 1.0 at a certainty of 0.85 when utilizing soil type 3; however, a value of 0.2 (considerably more dangerous) at a certainty of 0.8 (slightly less likely) might appear with soil type 4. This is one of the most lauded utilities of a fuzzy analysis: the ability to retain information that almost, but not quite, fits into a specified class.

The fuzzy overlay maps were examined for any values above a threshold (arbitrarily set at 0.6). If a particular cell had multiple realizations above this threshold, only the lowest factor of safety was retained. The resulting 'worst-case scenario' map is displayed in Figure 5. The blank areas indicate certainty factors below the 0.6 threshold (i.e., areas too uncertain to trust the model's results). These results confirm those reported by Goodchild et al. [15] in a vegetation map accuracy study: the errors tended to occur more frequently near

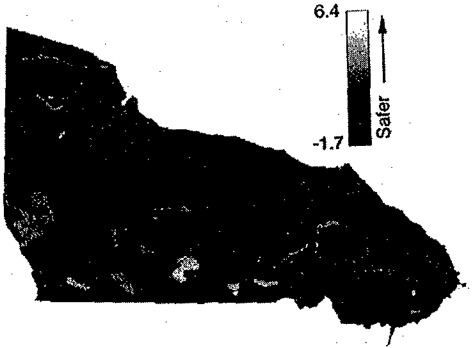

Figure 3: Maximum likelihood factor of safety

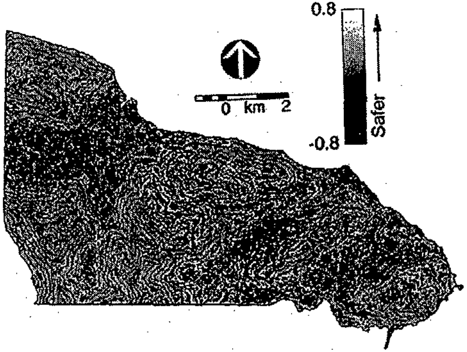

Figure 4: Maximum likelihood - Boolean

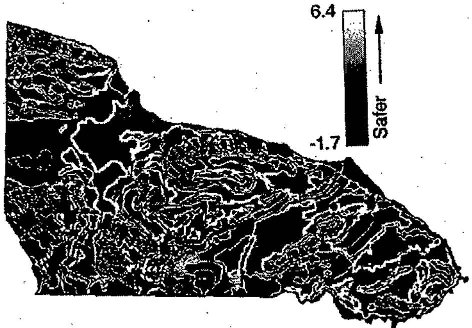

Figure 5: Worst case scenario

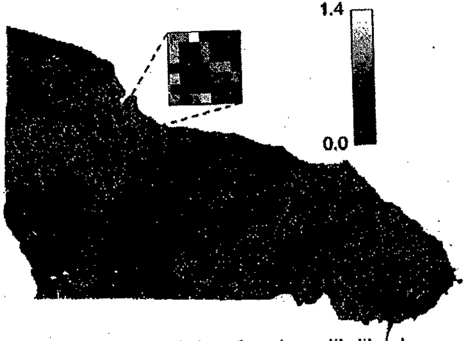

Figure 6: Standard deviation of maximum likelihood

polygon boundaries.

This summary comparison only scratches the surface of the uncertainty model's potential. Unfortunately, humans have a limited ability to perceive multiple criteria simultaneously. The data displays are even more limited: a two-dimensional colour representation can only present a maximum of two or three spatially referenced variables simultaneously with any chance of comprehension.

4.2 *Model Variance*

The log-normally distributed variability in the model's output is summarized with a standard deviation (SD) surface (Figure 6). In order to standardize all output as normally distributed, the log of the original value is stored in the map. The theming in Figure 6 demonstrates that the widest variability is found in the areas with the least slope; smaller SDs occur on the steeper sections. However, a closer examination of the SD themed surface reveals that the areas with the larger SDs are also very spotty — a great deal of local pixel-to-pixel variation is observed. This variation can be seen to be occurring in all of the source maps for the maximum likelihood summary. However, pixel-by-pixel, the sources also vary widely between the different realizations. This implies that the minor slope variations on gentle terrain may influence the model significantly. Such an observation supports a slope stability equation sensitivity study [17] in which slope had the greatest influence over output. On a steep slope such as 50°, a 5° variation introduced by perturbing the DEM does not significantly affect the model. On a relative flat plain, however, this 5° perturbation appears to significantly widen the model output variance.

An animated display increases the amount of comprehensible information that can be presented to the viewer; time becomes a new variable in the visual summary. Variance values are particularly suited to this time-variant method of presentation. A perspective animation using stability numbers modified by standard deviation gives instant feedback as to the type of surface containing greater or lesser uncertainty. Such an animation was produced to display the model's variance.

Exploratory data analysis reveals several interesting data correlations. When the 10% most unsafe areas under the maximum likelihood scenario are overlaid with the 10% widest standard deviations, the results only intersect in 0.04% of their area. When the former is combined with the 10% narrowest SDs, over 50% intersect. This indicates that, generally, one can be most sure of the model's predictions in unsafe areas.

The same test applied to the 10% safest areas demonstrate that, in general, the safest areas have the widest standard deviations (23% coincide). The exceptions — areas that show safe zones with narrow SDs — correspond to bedrock outcrops.

5 Discussion and Conclusions

Natural resource inventory methods are commonly mired in the paper map era. Considerable information is lost in the translation from field data/field experts to analytical data structures. Generally, the restrictions that required such data reduction techniques are no longer in force given current digital capabilities and should, therefore, be eliminated. Altman [1], among others, points out that the traditional conversion to 'hard' data occurs far too early in the modelling process. Retaining a maximum amount of data through the entire analytical process gives analysts flexibility and new windows on data elements such as uncertainty and error. This study indicates one potential method whereby retention of this information could be put to good use. The techniques of exploratory spatial data analysis [19] could yield many other uses for these data.

Boolean maps are simple, clear representations of idealized data structures in which boundaries are sharp and values exact to the nth decimal place. However, adopting the view that our perceptions and knowledge of the world are fraught with uncertainty devalues such simplistic representations. The fuzzy approach allows real-world uncertainty and the innate variability of natural phenoména to be represented. By its very nature, such a model is considerably more difficult to interpret than conventional results; yet the possibilities for database exploration that are opened up by the uncertainty model represent an entirely new spectrum of information.

Acknowledgment

The above research was conducted at the Spatial Sciences Laboratories at the University of Victoria. The research was partially funded by the National Science Engineering Research Council of Canada. Thanks to Dr. O. Niemann for making the Louise Island data available.

References

[1] Altman, D. 1994. Fuzzy set theoretic approaches for handling imprecision in spatial analysis. International Journal of Geographic Information Systems. 8(3):271-289.

[2] Beard, K. 1991. Position statement on visualization of data quality. In M. K. Beard & B. P. Buttenfield (eds.)NCGIA Research Initiative 7: Visualization of Spatial Data Quality. NCGIA Technical Paper 91-26. (NCGIA) pp. C11-C16.

[3] Beard, M.K. & B.P. Buttenfield. (eds) 1991. NCGIA Research Initiative 7: Visualization of Spatial Data Quality. NCGIA Technical Paper 91-26 (NCGIA)

[4] Blakemore, M. 1983. Generalization and error in spatial databases. Cartographica. 21(2/3):131-139.

[5] Burrough, P. A. 1989. Fuzzy mathematical methods for soil survey and land evaluation. Journal of Soil Science. 40:477-492.

[6] Burrough, P. A., R. A. MacMillan & W. VanDeursen. 1992. Fuzzy classification methods for determining land suitability from soil profile observations and topography. Journal of Soil Science. 43:193-210.

[7] Chrisman, N. R. 1982. A theory of cartographic error and its measurement in digital databases. Proceedings, Fifth International Symposium on Computer-Assisted Cartography. AUTO-CARTO 8. Falls Church, Virginia, (ASPRS & ACSM) pp. 159-168.

[8] Chrisman, N. R. 1989. Modelling error in overlaid categorical maps. In Goodchild M. & Gopal S. (eds.)The Accuracy Of Spatial Databases. (Bristol, PA: Taylor & Francis) pp. 21-34.

[9] Chrisman, N. R. 1991. The error component of spatial data. In D. J. Maguire, M. F. Goodchild & D. W. Rhind (eds.) Geographical Information Systems: Principles And Applications. 1. (Harlow, U.K.: Longman Scientific And Technical) pp. 165-174.

[10] Davis, T.J. 1994. Modelling and visualizing spatial uncertainty using fuzzy logic and Monte Carlo simulation. Unpublished M.Sc. Thesis, Department of Geography, University of Victoria, Victoria, British Columbia, Canada.

[11] Fisher, P. F. 1991. Modelling and visualizing uncertainty in GIS. In M. K. Beard & B. P.

Buttenfield (eds.)NCGIA Research Initiative 7: Visualization of Spatial Data Quality. NCGIA Technical Paper 91-26. (NCGIA) pp. C63-C70.

[12] Fisher, P.,F. 1992. Real-time randomization for the visualization of uncertain spatial information. In P. Bresnahan, E. Corwin & D. Cowen (eds.) Proceedings of the Fifth International Symposium on Spatial Data Handling. Charleston, South Carolina, Aug. 3-7. (University of South Carolina) pp. 491-495.

[13] Fisher, P. F. 1993. Visualizing uncertainty in soil maps by animation. Cartographica. 30(2-3):20-29.

[14] Goodchild, M. F. 1991. Issues of quality and uncertainty. In J. C. Muller (ed.) Advances In Cartography. (New York: Elsevier Applied Science Series).

[15] Goodchild, M.F., F.W. Davis, M. Painho & D.M. Stoms. 1991. Use of vegetation maps and geographic information systems for assessing conifer lands in California. NCGIA Technical Paper 91-23 (NCGIA)

[16] Goodchild, M. & S. Gopal. (eds.) 1989. The Accuracy Of Spatial Databases. (New York: Taylor & Francis).

[17] Hammond, C., D. Hall, S. Miller & P. Swetik. 1992. Level I Stability Analysis (LISA) Documentation. Intermountain Research Station, General Technical Report INT-285 (Ogden, Utah: United State Forest Service, Department of Agriculture)

[18] Heuvelink, G. B. & P. A. Burrough. 1993. Error propagation in cartographic modelling using Boolean logic and continuous classification. International Journal of Geographic Information Systems. 7(3):231-246.

[19] Keller, C. P. 1994. Exploratory spatial data analysis (ESDA) - The next revolution in GIS. In So, Now What. Proceedings, 1994 International Symposium on GIS. (Ministry of Supply and Services, Canada) pp. 298-302.

[20] Kollias, V. J. & A. Voliotis. 1991. Fuzzy reasoning in the development of geographical information systems. FRSIS: A prototype soil information system with fuzzy retrieval capabilities. International Journal of Geographical Information Systems. 5(2):209-223.

[21] Lowell, K. E. 1993. Initial studies on the quantification and representation of uncertainty in forestry data. In Proceedings, GIS '93 Symposium. Vancouver, BC, February, 1993. pp. 791-796.

[22] Mark, D. M. & F. Csillag. 1989. The nature of boundaries on 'area-class' maps. Cartographica. 26(1):65-77.

[23] Mendoza, G. A. & W. Sprouse. 1989. Forest planning and decision making under fuzzy environments: an overview and illustration. Forest Science. 35(2):481-502.

[24] Openshaw, S. 1989. Learning to live with errors in spatial databases. In M. Goodchild & S. Gopal (eds.)The Accuracy Of Spatial Databases. (Bristol, PA: Taylor & Francis) pp. 263-276.

[25] Perkal, J. 1966. An attempt at objective generalization. Geodézia es Kartográfia. VII(2):130-142.

[26] Robinson, V. B. 1988. Some implications of fuzzy set theory applied to geographic databases. Computers, Environment and Urban Systems. (12):89-98.

[27] Shiffer, M. J. 1993 . Implementing multimedia collaborative planning technologies. In URISA Proceedings. Atlanta, Georgia.

[28] SRMB - Survey and Resource Mapping Branch, Ministry of Crown Lands & Province of B.C. 1990. British Columbia Specifications and Guidelines for Geomatics. Content Series, Volume 3 (Victoria, B.C.: Survey and Resource Mapping Branch)