AURISA 94
The Annual Conference of the Australasian Urban and Regional Information Systems Association Inc
Sydney, New South Wales
21-25 November 1994

A Toolbox for Assessing Uncertainty in Spatial Databases

G.J. Hunter†, M.F. Goodchild†† & M. Robey†

† Department of Surveying and Land Information, and
Centre for Geographic Information Systems and Modelling
The University of Melbourne, Parkville, Victoria
Australia 3052, Tel. +61-3-344-4626, Fax. +61-3-347-2916

†† National Center for Geographic Information and Analysis
3510 Phelps Hall, University of California
Santa Barbara, California, 93106, USA
Tel. +1-805-893-8224, Fax. +1-805-893-8617

ABSTRACT

As the application of spatial databases becomes increasingly diverse, an issue of growing concern is the inability to determine whether the quality of the products derived from these databases are suitable for their intended purposes. For many users of the technology, understanding the quality with their outputs will help achieve one of the "bottom lines" in using spatial databases, namely the delivery of accurate information. To help meet this demand, a toolbox for communicating the uncertainty associated with certain types of spatial database products has been developed, by which users will be better placed to make data quality assessments. This paper will describe the design and application of the toolbox.

INTRODUCTION

An issue that has become of increasing importance to spatial database users in recent years is that of assessing the quality of their system products. As the applications of these systems become more sophisticated, and as users grow more experienced in the technology's strengths and weaknesses, questions are being increasingly raised as to whether or not the quality of the outputs necessarily matches the requirements of the tasks for which the information is to be employed. For many users, understanding the quality their system outputs will help achieve one of the "bottom lines" in applying spatial databases, namely the delivery of accurate and appropriate information. Thus, the debate about uncertainty has reached the stage where there is a critical need for tools to be developed to assist users in better understanding the nature of the outputs derived from their systems.

Before discussing the options available for dealing with this problem, some explanatory remarks are required regarding the use of the term 'uncertainty'. In general terms, it denotes a lack of sureness or definite knowledge about an outcome or result. In the context of spatial databases, the authors suggest there is a clear distinction to be made between 'error' and 'uncertainty', since the former implies that some degree of knowledge has been attained about differences (and the reasons for their occurrence) between the results or observations and the truth to which they pertain. On the other hand, 'uncertainty' conveys the fact that it is the very lack of such knowledge which is responsible for hesitancy in accepting those same results or observations without caution, and often the term 'error' is used when it would be more appropriate to use 'uncertainty'.

It is well known that there are many potential sources of error in spatial databases (as discussed in Hunter and Beard, 1992), but because so little is understood about the way in which those errors (either singly or in conjunction with each other) affect the outcome of the final products (be they displays, maps, graphs or reports), there is a resultant uncertainty concerning the level of trust which should be placed in them. Indeed, Goodchild (1993) notes there are only some half dozen known and widely accepted error models for the many hundreds of spatial database operations available.

In some ways, the distinction between error and uncertainty is analogous to the legal belief that a person is 'innocent until proven guilty', since in many cases conceptual models of spatial database error simply do not exist and it is suggested that until that situation improves, 'uncertainty' offers a more appropriate means of describing such lack of proof. This does not mean that 'uncertainty' should always be substituted for 'error', as there already exist a small number of well established error models for given spatial operations and they are properly described as such, however in situations where there is little knowledge of the actual errors involved, as in the case study described, it is uncertainty which will be referred to by the authors.

At this time, there are three options available (Goodchild, Chih-Chang and Leung, 1993) for dealing with uncertainty in spatial databases, viz.:

(1) omit all reference to it;
(2) attach some form of descriptor to the output; and
(3) show samples from the range of possible maps.

The first approach (the 'do nothing' option) treats the problem by ignoring it; undoubtedly the easiest solution to adopt, but one which potentially places at risk the reputations of decision-makers

(and their agencies) who have to act on the basis of such information. The second option would see the use of descriptors such as epsilon bands, misclassification matrices, reliability diagrams, and root mean square error estimates. In effect, these are a caveat to users and while they give warnings about product uncertainty they provide little assistance in showing how the resultant output might spatially vary, although with further development they can be more usefully interpreted as Hunter and Goodchild (1994a) have shown in the case of the root mean square error estimate for Digital Elevation Models (DEMs). Finally, different versions of the same map might be presented to users to illustrate the uncertainty their products are subject to due to the particular combination of data, error estimates, algorithms and other models which have been chosen for the task at hand.

This latter approach is the one preferred by the authors, since it would appear to have the greatest potential benefit in both communicating uncertainty and at the same time educating the user community in the significance of the issue. Accordingly, this paper discusses the design and application of a toolbox which permits uncertainty reporting for certain types of spatial database products. By presenting the level of uncertainty which resides in an output, such a toolbox might assist agencies in determining the degree of uncertainty they are willing to tolerate before it either changes the decisions made on the basis of that information, or else (in the worst case) causes the benefits of spatial database usage to be lost. In the reverse role, the toolbox could provide pre-testing of different combinations of data, error estimates, algorithms and models to assess which ones are most likely to suit the user's needs.

At this stage, the toolbox is restricted to the study of grid-cell data and, specifically, the outputs derived from the use of DEMs, however even in this restricted role it has considerable relevance to natural resource and environmental applications where the raster data model has greater suitability for representing inherently continuous variation. In addition, the raster model more easily accommodates simulation techniques such as those used in this research. The paper discusses (1) the underlying model of uncertainty employed, (2) its potential applications, (3) incorporation of the model into a toolbox to handle uncertainty, and (4) a case study illustrating the toolbox's application in determining the uncertainty associated with calculating the 'northness' index derived from DEMs – as used for identifying vegetation communities and studying species diversity.

THE UNDERLYING MODEL OF UNCERTAINTY

The basis for the toolbox is a version of the model developed to represent uncertainty by Goodchild, Guoqing and Shiren (1992). In general terms, the model can be defined as a stochastic process capable of generating a population of distorted versions of the same reality, with each version being a sample from the same population. The traditional Gaussian model (where the mean of the population is an estimate of the true value and the standard deviation is a measure of the variation in the observations) is one attempt at describing error, but it says nothing about the processes by which it has accumulated. In certain cases where the Gaussian model is not applicable, the proposed model offers an alternative solution to assessing uncertainty. Not only does it have the advantage of showing spatial variation in uncertainty, but it includes the effects of spatial autocorrelation (r) and the likely propagation of error arising from the spatial operations that have been applied.

The model permits a data set (such as a DEM) to be perturbed according to an error estimate (its RMSE value), and then used for other applications so that different versions of the same map are

produced. In a previous paper, Hunter and Goodchild (1994b) argued that while it is possible to perturb a data set according to an error descriptor (such as the RMSE value) without consideration of spatial autocorrelation between point sample elevations (that is, $r = 0$), the process is stochastic but lacks 'truthfulness' – since adjacent elevations can be severely distorted by the creation of large pits and peaks which do not intuitively occur in practice. This approach produces what are known as 'random maps'.

On the other hand, assumption of complete spatial dependence between neighbouring points (that is, $r = 0.25$) produces realisations of the DEM which are 'truthful' but not stochastic, since elevations are constrained to maintain their relative differences to each other and the introduction of a noise value only has the effect of moving DEM elevations 'all up' or 'all down' by a constant amount. Hence, there is a need to find the spatial autocorrelation value in the domain $0 < r < 0.25$ at which the dual requirements of being both stochastic and 'truthful' are met. The limit of 0.25 ensures stationarity when the Rook's case is used to test a cell's elevation against its four neighbours sharing a common edge (as discussed in Cliff and Ord, 1981, p. 147).

By producing distorted versions of the DEM for different r values, and by studying the change in differences between the realised data and the original data, it is possible to make reasonable deductions as to what an appropriate r value might be. In the paper already referred to (Hunter and Goodchild 1994b), separate realisations of slope gradient and aspect values were derived from a DEM with the latter, in particular, showing a marked change in response at approximately $r = 0.24$, while slope gradient only started to significantly vary from $r = 0.20$ onwards.

Of course, the realisation process need not stop there, as the different slope gradient and aspect maps can be input to, say, hydrologic models to produce alternative realisations of drainage basin parameters, which in turn can be used to derive realised runoff characteristics. At any stage, the differences between the realised maps and the original (as produced from the source data without any consideration of uncertainty) may be analysed to assess the resultant effects. The attractiveness of this approach is that even though it is not known how error is being propagated, its likely effects are nevertheless displayed.

POTENTIAL APPLICATIONS OF THE MODEL

The potential applications of the model lie in four areas. First, the realisation process may be used to highlight areas of a map which are susceptible to change in parameter values. For instance, Hunter and Goodchild (1994b) demonstrated that the calculation of slope aspect from a DEM was particularly susceptible to variation in terrain elevation in relatively flat regions while large hillside slopes remained relatively stable. While such a conclusion is already fairly well established, this may not always be necessarily so and where complex process models are applied their effects may still be largely unknown. In other applications, the observed differences might be used as input to sensitivity analyses to understand how changes in parameters impact upon the decision making process, as in landuse suitability and capability studies.

Secondly, the technique can be useful in cases where differences *per se*, are not as important as assessing the likelihood of a cell's membership of a particular class. For example, in viewshed computations cells are computed as being either visible or not visible from a viewing point. Sets of

realisations taken at different *r* values can be added to compute a 'score' for each cell (together with a mean and standard deviation), which in turn may be used to calculate the probability of a cell satisfying the criteria associated with the operation – thereby overcoming the 'in or out' Boolean responses normally associated with spatial databases. Users can thus nominate a confidence level to be met when assessing the results of the process (for example, 'cells must have a 90% probability of being seen').

Thirdly, a user might want to display several realisations of a map to understand the degree of variation associated with the processes involved. For example, instead of interpolating contours from a DEM just once, several realisations might be made to assess not only the impact of elevation uncertainty on the process, but also the variation due to the interpolation procedure itself. This approach could also be applied to other raster-to-vector conversion procedures in which class polygons or linear features such as stream patterns are required, thereby producing a family of possible boundaries or linear features.

Finally, simulations can be undertaken to study the effect on map products where competing data sets, error estimates, algorithms and process models are available. This 'reverse engineering' approach might also be applied by users who, having already studied several possible realisations of a desired map, and having identified areas exhibiting levels of uncertainty considered unacceptable, wish to see how different uncertainty reduction options (for example, recollecting data at a higher accuracy) would affect the final outcome – before returning to the field site or purchasing alternative data sets.

DEVELOPMENT OF THE TOOLBOX

The design of the toolbox to handle uncertainty embodies the model previously described and is shown in Figure 1. It consists of four key stages, with the first one requiring the user to combine whatever data, processes and models are needed to generate the desired output – in other words, to apply the spatial database as would normally be done without any consideration of uncertainty. From the beginning of the procedure, a log is kept of the commands used which will later be applied in producing the realisations.

In the second stage, the parameters necessary for the realisation process are determined. By reading system variables associated with the source data file, the number of rows and columns in the data file, the cell size, and the geo-referencing details of the data can be ascertained. These will be required later when the noise files are to be transformed to agree with the actual data set used. Any constraints applied during processing will be embedded in this file, such as in a viewshed computation where cells immediately surrounding the viewing point are usually masked out or held fixed (and therefore assumed to be always seen) so that their elevations are not perturbed and thereby possibly obscuring large areas of the viewshed. An error estimate for the source data will also need to be identified and this would normally be the RMSE estimate supplied with the DEM.

While not a direct step in the realisation procedure as such, the noise files to be employed would usually be previously computed and then permanently stored in the system for future use. The way in which they are generated has already been described in Hunter and Goodchild (1994b). To date, it has been considered sufficient for most applications tested for about ten files to be held for each *r*

value, although users would have the option of creating a greater number of noise files for specific tasks in the final module of the methodology. The default *r* values chosen for the noise files are 0.0, 0.05, 0.10, 0.15, 0.20, 0.21, 0.22, 0.23, 0.24, 0.245, and 0.249. As for the maximum value of *r* offered (0.249), experience to date has shown there is little to be gained from using *r* values higher than this since the realisation process becomes so constrained that there is no discernible difference between the realised maps and the original.

In stage 3 of the toolbox, it is expected that users will want to see a small number of initial trial realisations and the default *r* values listed above are applied. A single realisation for each value is performed by first applying the parameters derived from stage 2 to geo-reference and transform the coordinates of the noise grid. Next, the error estimate is applied to map the noise values from a Normal distribution of N(0,1) to N(0,RMSE) so that it has a distribution similar to the original DEM. This adjusted noise file is added to the source data to produce a realisation to which the commands employed to create the original output are applied. The realised maps and the differences between the realisations and the original outputs can be mapped or graphed. Finally, in stage 4 of the toolbox the user may choose one or more approaches for more detailed investigation of product uncertainty, as discussed in the previous section, and with a greater variety of reporting output products available.

At this time, the toolbox has been developed as a series of macro command files in ARC/INFO for use in a workstation environment. Also included in the toolbox is the executable code needed to produce the sets of realisations for different *r* values, as developed by the U.S. National Center for Geographic Information and Analysis. It is envisaged that the toolbox would be used by analysts within an agency tasked with assessing the uncertainty associated with DEM-based products. To ensure the widest possible use of the toolbox, the software will be made freely available upon request.

A CASE STUDY IN ASSESSING UNCERTAINTY

The case study to be discussed deals with the uncertainty of the derivation of the 'Northness' index commonly used in image analysis to identify vegetation communities and to study species diversity and ecological land classification (Davis & Goetz 1990, Davis & Dozier 1990, Fairbanks 1993). It is applied to grid cells as a measure of how much sunlight is reaching the vegetation within the cell, and is a function of both slope gradient and aspect. It is calculated by equation (1) and takes values in the range ±1.0. Thus, the uncertainty of the index derived for each pixel in an image is a function of the DEM elevation error, the algorithms used to calculate slope gradient and aspect, and the 'Northness' equation itself.

$$\text{'Northness'} = \sin(\text{Slope Gradient}) * \cos(\text{Slope Aspect}) \qquad (1)$$

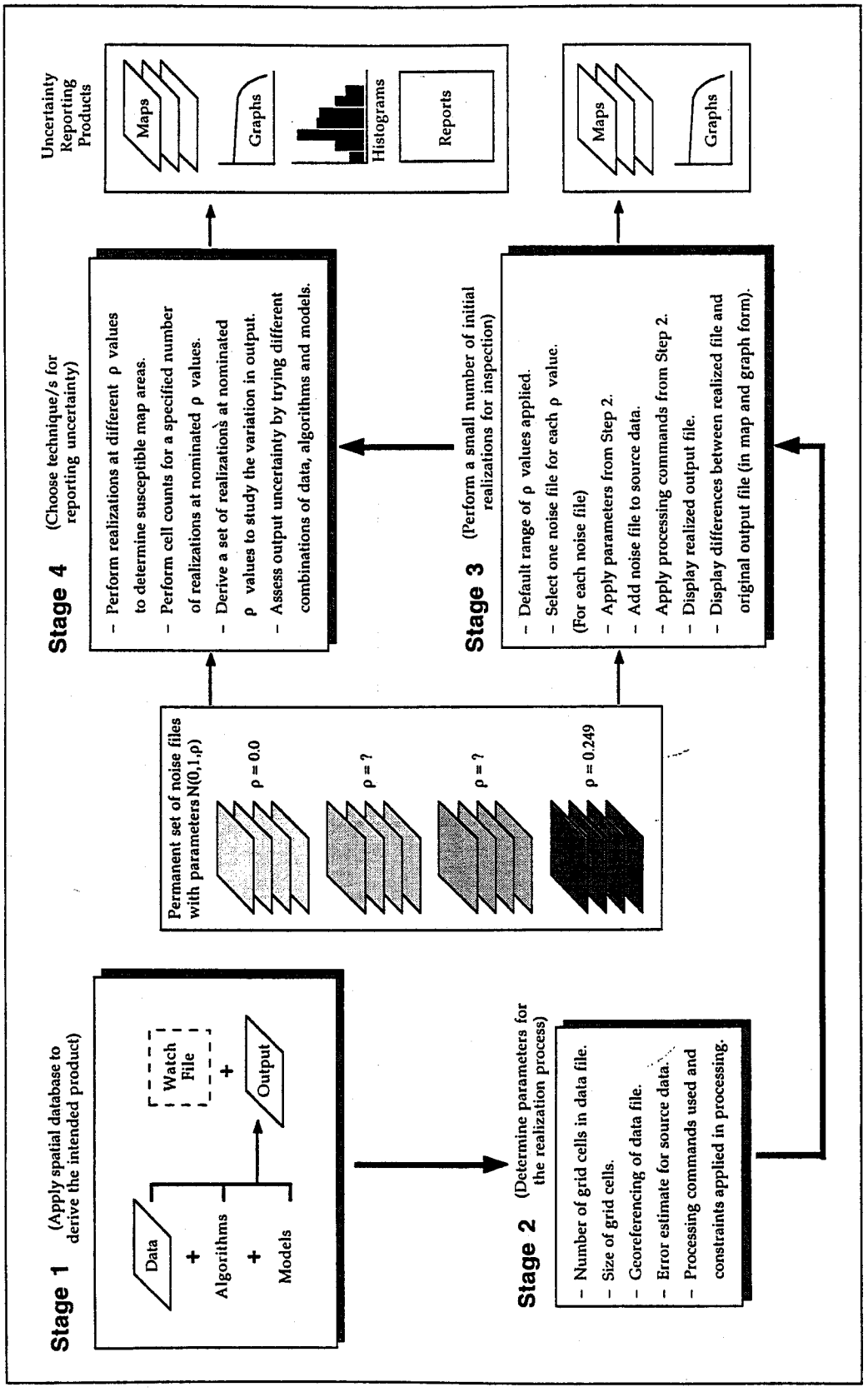
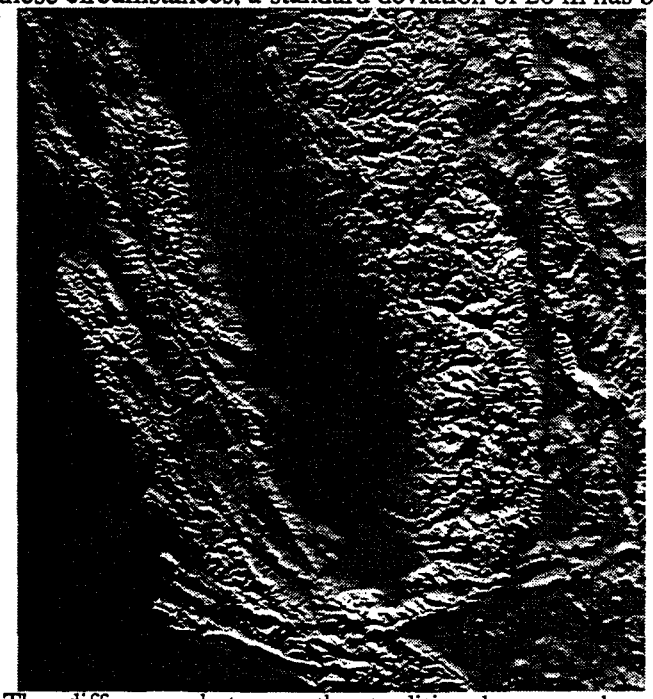


Figure 1: The Design of the Uncertainty Reporting Toolbox.

The topographic data used for the study came from the California-based EROS Data Center and was originally developed for a US continental land cover database. The DEM was first derived by the Defence Mapping Agency (DMA) from 1°x2° 1:250,000 topographic maps with (mostly) 200 foot (60.96 m) contour intervals. The grid nodes were spaced at 3 arc-second intervals in the X and Y directions. The DEM was subsequently interpolated by the National Telecommunications and Information Administration and the elevations rounded to the nearest 20 feet (6.1m) for every 30 arc-seconds of latitude and longitude. The file has since been re-interpolated to a grid size of 1000 m x 1000 m, with point elevations supplied to the nearest metre.

As a result of the considerable reprocessing to which this particular file has been subjected, there is no longer any error estimate attached to it. However, while the specific details of its lineage remain obscure it is suggested that the error estimate which the file had when produced by the DMA serves as a useful (although by now, conservative) basis for this study. From the USGS documentation on DEM accuracy (USGS 1990, pp. 13-14) it is noted that the DMA production objective is to satisfy *"... an absolute vertical accuracy (feature to mean sea level) of ±30 m, linear error at 90-percent probability"*.

While it is not specifically stated, it is assumed by the authors that this relates to a Normal probability distribution which would give a standard deviation of 18.2 m. This is confirmed by the classification of the DMA product as level 3 by the USGS, in which *"... an RMSE of one-third of the contour interval is the maximum permitted"* (USGS 1990, p. 15), and would thus be approximately 20 m given the 60.96 m(200 feet) contour interval on the original 1:250,000 source maps. Under these circumstances, a standard deviation of 20 m has been initially adopted for the research.



The test site measures 480 rows by 435 columns (or 480km x 435km) and covers the central Californian coast ranging from San Francisco Bay in the upper north-west corner and Santa Barbara in the bottom centre. Elevations range from sea level in the west to over 4000 m in the mountain ranges to the east.

Figure 2 (left) shows the 'Northness' index values for the test site shaded from black (-1.0) through to white (+1.0), with mid-gray cells which indicate a value of zero. The mountain ranges either side of the north-south running San Fernando valley can be clearly seen in this illustration.

The difference between the traditional approach to calculating the 'Northness' index and the proposed approach which permits its uncertainty to be assessed, can be seen in Figure 3. The latter technique applies elevation noise files, for different values of spatial auto correlation, to the original DEM to establish corresponding sets of slope gradient and aspect files for the test site. Each pair of

realised gradient and aspect files is then taken in turn and used to calculate a corresponding 'Northness index files.

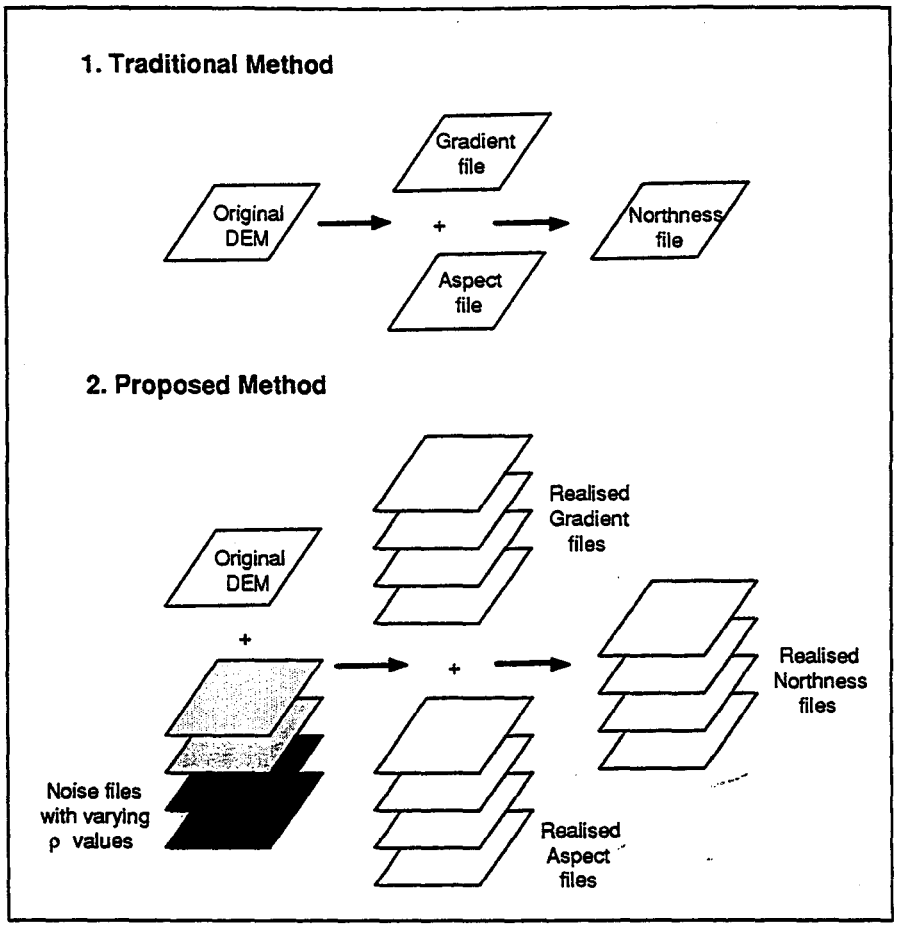


Figure 3: Showing the difference between the traditional approach to calculating the 'Northness' index and the proposed approach which permits its uncertainty to be assessed.

As per stage 3 of the toolbox, an initial set of realised 'Northness' index files was calculated for the r values 0.0, 0.5, 0.10, 0.15, 0.20, 0.21, 0.22, 0.23, 0.24 and 0.245. The grid cells in each realised file were then subtracted from their corresponding cells in the original 'Northness' file to provide a 'difference' file. This difference represents the amount by which the final 'Northness' index value might be expected to vary under terms of uncertainty due to variation in the elevations of the original DEM and the subsequent series of spatial operations that were performed on the data.

The mean and standard deviation of each difference file were then calculated and graphed. Figure 4a shows that the average differences are very small and decrease gradually from $r = 0$, whereas Figure 4b shows a slight initial rise in the standard deviation of the differences, followed by a sharp decrease from about $r = 0.20$. From this information it is clear that the 'Northness' index is only affected in the second or third decimal place of its value.

Stage 4 of the toolbox was then applied and ten realisations were computed for the 'Northness' index at values of r = 0.20, 0.22, 0.24 and 0.245. Each set of ten realisations was added and a mean 'Northness' index was computed for each cell. Visually, there were no significant differences between the four mean 'Northness' grids computed other than a decrease in graininess as r increased (that is, as the DEM perturbations became more constrained).

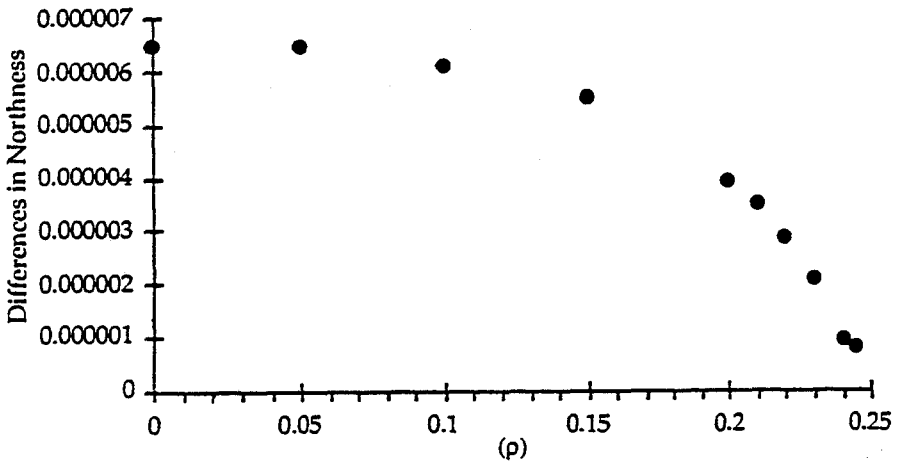


Figure 4a: A Graph showing mean differences versus r value, between the initial realised maps and the 'Northness' map originally derived from the source DEM.

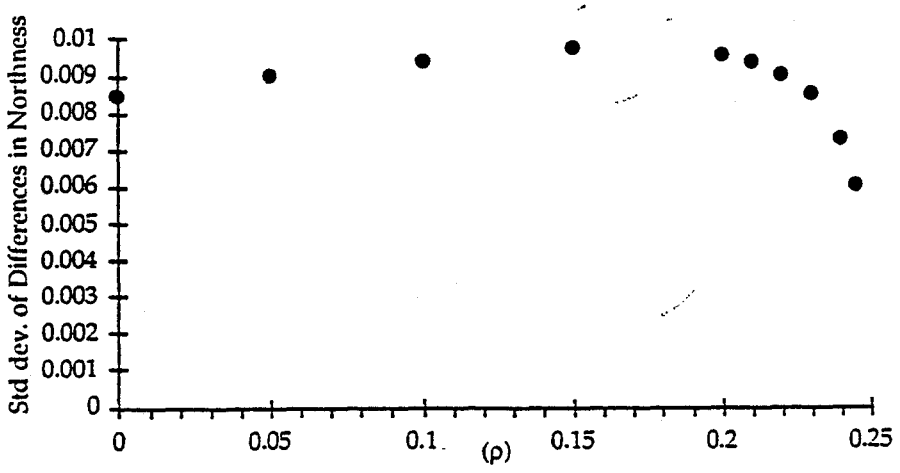
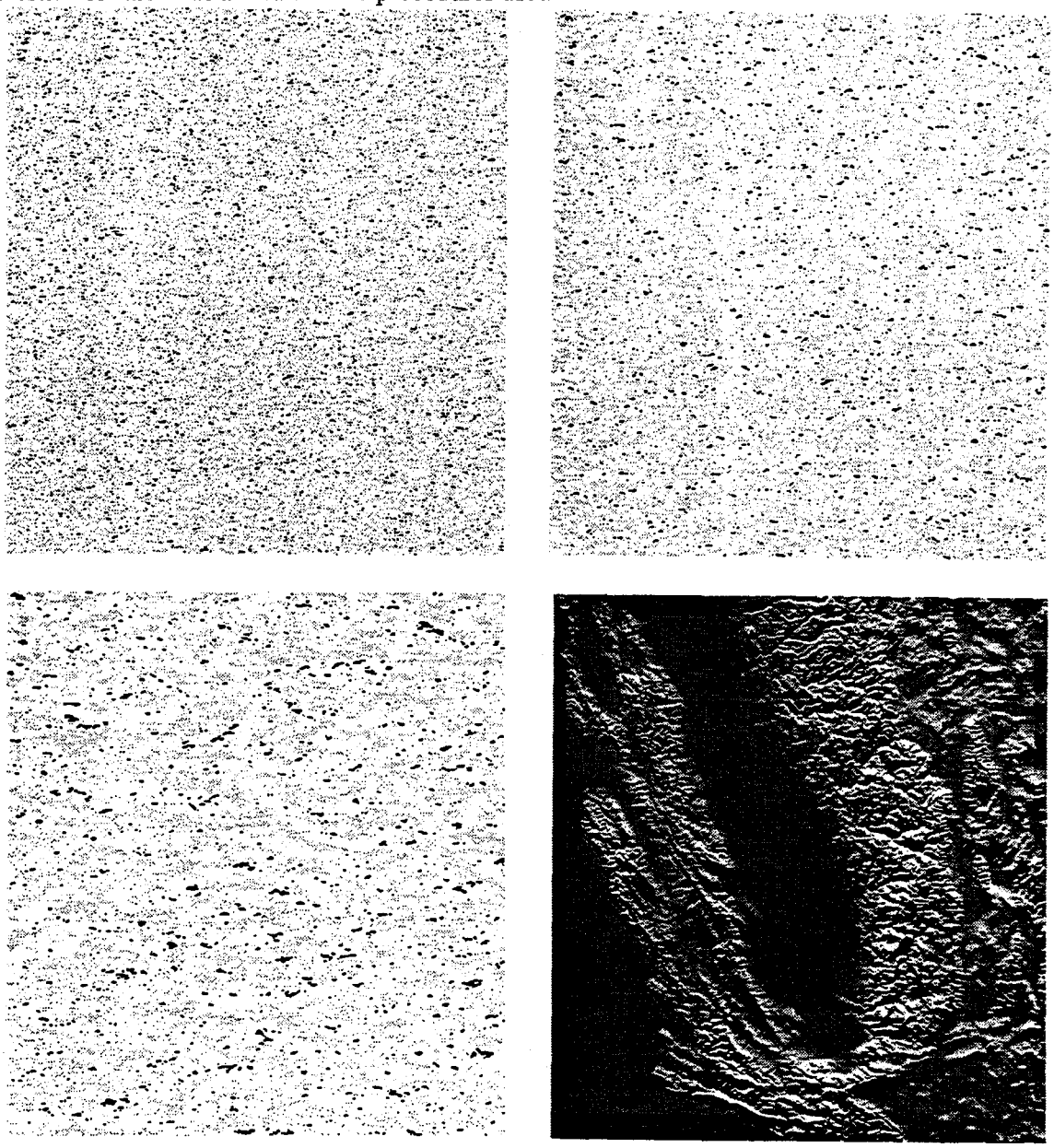


Figure 4b: Graph showing the change in standard deviation of the differences between the initial realised maps and the original 'Northness' map for each value of r applied.

The differences between these mean realised grids and the original 'Northness' grid were also calculated and displayed by masking out differences within ±2 standard deviations of the mean in white, and highlighting those outside these limits in black. As can be seen in Figures 5a, b and c, for realisations at r = 0.20, 0.24 and 0.245, there are no locally significant trends in variations in the 'Northness' index as a result of the procedures used.



Figures 5a (top left), 5b (top right) and 5c (bottom left) show differences between the original 'Northness' index map and the mean realisations for the 'Northness' index at values of r = 0.20, 0.24 and 0.245 respectively. Cells which exhibit a difference beyond the threshold of ±2 standard deviations from the mean are shown in black. In comparison with the original 'Northness' index map shown in Figure 5d (bottom right), there are no obvious areas of terrain that are locally susceptible to uncertainty in the 'Northness' index.

In summary, from these tests it is obvious that derivation of the 'Northness' index for this data set is not unduly affected by uncertainty in either the terrain or the algorithms used in its calculation. Differences were detected only in the second and third decimal places of the index value, which is insignificant considering that only one decimal place is normally used with this index. As for local variation, it is also obvious that there are no areas of the test site which are more susceptible to uncertainty than others, even though there is considerable variation in relief. From the user's perspective, there may well be far greater uncertainty associated with other components of the vegetation classification process (such as the 1000 m grid resolution), but on the basis of this study the means by which the 'Northness' index is derived is not a major contributor to the overall uncertainty of products which may be based upon it.

CONCLUSIONS

In this paper, the authors have presented the design and application of a toolbox that allows uncertainty to be reported for certain types of spatial database products. The work recognises the critical need for tools to be developed to assist users in improving their understanding of the quality of the outputs from their systems. The toolbox has been applied to portray the uncertainty associated with determining the 'Northness' index for grid cells used when classifying vegetation communities from remotely sensed imagery. The results of the study show that given the particular combination of DEM data and algorithms employed, the index is likely to vary only in the second or third decimal place, which is not of significance to users for this particular application. In addition, no particular areas of the test site appear to be locally susceptible to variation in the parameters used to calculate the index.

REFERENCES

Cliff, A.D. & Ord, J.K. 1981, *Spatial Processes: Models and Applications*, Pion, London.

Davis, F. & Dozier, J. 1990, "Information Analysis of a Spatial Database for Ecological Land Classification", *Photogrammetric Engineering & Remote Sensing*, vol. 5, pp. 605-13.

Davis, F. & Goetz, S. 1990, "Modeling Vegetation Pattern Using Digital Terrain Data", *Landscape Ecology*, vol. 4, pp. 69-80.

ESRI, 1992, ARC/INFO User's Guide: GRID Command References (version 6.1), Environmental Systems Research Institute Inc., Redlands, California.

Fairbanks, D. 1983, "Relationship of GIS-based and Remotely Sensed Data to Floristic Gradients within California Vegetation Communities", Unpublished Masters degree Thesis, University of California, Santa Barbara, 137 pp.

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*, Vol. 6, No. 2, pp. 87-104.

Goodchild, M.F.,1993, "Data Models and Data Quality: Problems and Prospects", in M. F. Goodchild, B. O. Parks and L. T. Steyaert (eds), *Environmental Modeling with GIS*, New York: Oxford University Press, pp. 94-103.

Goodchild, M.F., Chih-Chang, L. and Leung, Y., 1993, "Visualizing Fuzzy Maps", in H. Hearnshaw and D. Unwin (eds), *Visualization in Geographic Information Systems*, Association of Geographic Information, London (in press), 12 pp.

Hunter, G.J. and Beard, K., 1992, "Understanding Error in Spatial Databases", *The Australian Surveyor*, Vol. 37, No. 2, pp. 108-119.

Hunter, G.J. and Goodchild, M.F., 1994a, "Dealing with Uncertainty in Spatial Databases: A Simple Case Study", Accepted for publication by *Photogrammetric Engineering and Remote Sensing Journal*, 20 pp.

Hunter, G.J and Goodchild, M.F., 1994b, "Modeling the Uncertainty of Slope Gradient and Aspect Estimates Derived from Spatial Databases', Submitted for publication to the *Photogrammetric Engineering and Remote Sensing Journal*, 25 pp.

USGS, 1990, *Digital Elevation Models: Data Users Guide 5*, Earth Science Information Center, U.S. Geological Survey, Department of the Interior, Reston, Virginia, 51 pp.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support received for this research from the Australian Key Centre for Land Information Studies (AKCLIS), and the National Centre for Geographic Information and Analysis (NCGIA), Santa Barbara, California. The NCGIA is supported by the National Science Foundation, grant SES 88-10917 and this research constitutes part of the Centre's Research Initiative #7 on Visualisation of the Quality of Spatial Data.

Further Information:

Copies of the toolbox software used in this research may be obtained free of charge upon request from Dr. G.J. Hunter at the University of Melbourne.