QUALITY ASSESSMENT OF GIS PRODUCTS FOR RESOURCE MANAGEMENT

Gary J. Hunter[†] and Michael F. Goodchild[††]

[†] Department of Surveying and Land Information and
Centre for Geographic Information Systems and Modelling
The University of Melbourne, Parkville
Victoria, Australia 3052

[††] National Center for Geographic Information and Analysis
3510 Phelps Hall, University of California
Santa Barbara, California, USA, 93106

ABSTRACT

In this paper, the authors discuss two techniques for assessing the quality of outputs from Geographic Information Systems (GIS) when applied to resource management problems. It is argued that the advent of mandatory data quality reporting requirements, and the need to protect the integrity of decisions made by regulatory agencies using GIS against legal challenges, has bought about a critical need to be able to assess whether the quality of GIS products meets the quality requirements of the tasks for which the information is to be employed. This paper presents two case studies in which techniques for judging the quality of spatial data products are investigated. In the first study, simple probability mapping is employed to communicate the error in elevation arising from the use of Digital Elevation Models (DEMs). In the second case study, a more advanced methodology is presented to assess the effect of DEM error on the terrain corrections applied to remotely sensed imagery in mountainous terrain. It is suggested that techniques such as these can help GIS users to overcome some of the data quality reporting problems now being faced.

INTRODUCTION

An issue which has become of increasing importance to GIS users in recent years is that of assessing the quality of their system products. For resource managers, this difficulty is compounded by the fact that much of the data they routinely deal with applies to natural phenomena in which the definition of physical boundaries and their associated attributes is usually highly subjective. As the applications of GIS become more sophisticated and users become more experienced in this technology's strengths and weaknesses, questions are now being increasingly raised as to whether or not the quality of the outputs necessarily matches the quality requirements of the tasks for which the derived information is to employed.

The reasons for this concern stem from:

(1) the recent introduction, in many countries, of spatial data transfer standards that carry a mandatory requirement for data quality reporting;

(2) the rapidly growing view of spatial data as a commercially saleable commodity; and

(3) the increase in litigation relating to decisions made by agencies in which GIS products have played a role.

The third reason, in particular, is causing anxiety amongst regulatory organisations which see the integrity of their decisions and reputation within the community as being at stake. For example, in USA the Environmental Protection Agency is responsible for the Superfund Project in which toxic waste-affected land is identified and legally binding notices are served on owners of property from which the waste emanates to fund the clean up process. In some cases, restoration costs amount to hundreds of millions of dollars and, not unnaturally, most affected owners are challenging the EPA's decisions in court. Thus far, the amount of litigation pending has become so great that the EPA has set aside over $1 billion to cover expected legal defence costs.

What, then, is meant by data quality? Put simply, it can be defined as the 'fitness for use' of a particular product to meet a user's needs for a given purpose, and it requires a knowledge of the accuracy of the data (that is, the closeness of the observations to the truth or to measurements which are accepted as being true). The dimensions of the accuracy issue are now generally accepted as including positional accuracy, attribute accuracy, logical consistency, completeness and data currency. In the case of positional accuracy, it may be possible to measure the coordinates of a feature on a map and compare them with higher order coordinates gained, for example, from detailed ground survey.

Attribute accuracy assessments may also be made, for example when verifying landuse classifications derived from remotely sensed imagery with the actual landuse on the ground. For each GIS application, identification of the key accuracy components will clearly depend on the use to which the data is to put, and while in some cases it is positional accuracy which is of primary importance to users, in other circumstances attribute accuracy may be the dominant factor in assessing data quality. Alternatively, users may be concerned with a combination of several of the five accuracy factors listed above.

To assist users of spatial data in determining whether a product has the potential to meet their needs, data producers are now beginning to provide detailed data quality statements which report on each of these factors. This concept is known as 'truth in labelling', and it is considered (although not well tested in the courts) that if producers truthfully label products to the best of their ability, then it is reasonable to expect users to take similar care in studying the information provided before using such products. Unfortunately, while this is a step in the right direction towards better quality assessment, many users are striking difficulty in assessing the fitness for use of secondary products which may have been created from a variety of data sets from different sources via the many spatial algorithms and operations now available in commercial software packages.

In such cases, models of error and a knowledge of the way in which error propagates through each stage of a product's creation are required, however at this time there are relatively few widely accepted and tested models available for the hundreds of spatial operations now employed (Goodchild, 1993). Hence, the growing use of the term 'uncertainty' to denote the fact that until we possess the required understanding of spatial data error it is our lack of knowledge which is causing many of the problems associated with the assessment of product quality.

Clearly, new methods and tools are required to assist users in this area and the paper presents two methods recently employed by the authors for this purpose. In the first case study described, an assessment of the elevation error associated with Digital Elevation Models (DEMs) is presented. It is suggested that simple probability theory, when combined with elevation error estimates supplied by data producers, can provide users with more effective information concerning the quality of their GIS products. The method adopted is seen as a useful means of helping to convey the meaning of the Root Mean Square Error (RMSE) statistic currently associated with DEMs.

In the second case study, a model of uncertainty developed by Goodchild, Guoqing and Shiren (1992) has been applied. 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 error propagation resulting from the GIS operations that have been applied.

The model permits a data set (such as a DEM) to be perturbed according to an error descriptor (its RMSE value), and then used for other applications so that different versions of the same map may be produced. It differs from the traditional means of producing random maps in that it has the capability of taking into account spatial autocorrelation between cells so that some degree of intuitive 'truthfulness' can be maintained in the realised maps obtained. The model has been applied to an assessment of the uncertainty of terrain corrections made to remotely sensed images in mountainous country to account for the effects of shadow and large slope gradient variations. Until such time as our knowledge of spatial error considerable improves, the authors suggest that this new model has the potential to greatly assist with quality assessment in certain situations.

CASE STUDY 1

DEMs are commonly used in resource management for purposes ranging from flood plain mapping through to classification of vegetation communities and animal habitats. Errors in elevation can have considerable impact upon the horizontal uncertainty of such boundaries. In the problem presented below, the 350m elevation is to be delineated for the data set described – together with an assessment of its uncertainty.

The data comes from a 448 row by 334 column subset (86%) of the US Geological Survey (USGS) Digital Elevation Model (DEM) for the 7.5-minute 1:24,000 State College (Pennsylvania) mapping quadrangle. The total number of cells in the test file is 149,632 with each one measuring 30m x 30m and covering an area of 900m^2 or 0.09ha. The test site measures approximately 10km by 13km and has considerable variation in elevation, ranging from 255m in the north to 743m in the south-east. The vertical accuracy of the DEM used for the tests is quoted by the USGS as being 7m RMSE, which means that the square root of the average squared difference between the true and observed elevations at approximately thirty test points in the quadrangle is 7m.

In probability mapping, the likelihood of each grid cell value exceeding or being exceeded by a nominated threshold is calculated and used as a means of displaying the variability of a cell's elevation. Assuming a Gaussian or Normal distribution of random error, the concept is illustrated in Figures 1a, b and c where cell (p) is shown with an observed elevation (Z_p) and a vertical error estimate (RMSE). The probability of a cell being greater than the threshold value (Z) can be evaluated by calculating $P(Z_p>Z)$ and writing the value to a separate file for subsequent display. In Figure 1a, where $Z_p-Z = 0$, the characteristics of the Normal probability distribution are such that cell (p) has a likelihood of 0.50 (50%) that its true elevation will be greater than the nominated threshold value, Z. Alternatively, in Figure 1b where $Z_p-Z = 1 \times$ RMSE the probability is about 0.16 (16%), and in Figure 1c for $Z_p-Z = 2 \times$ RMSE the likelihood is approximately 0.025 (2.5%).

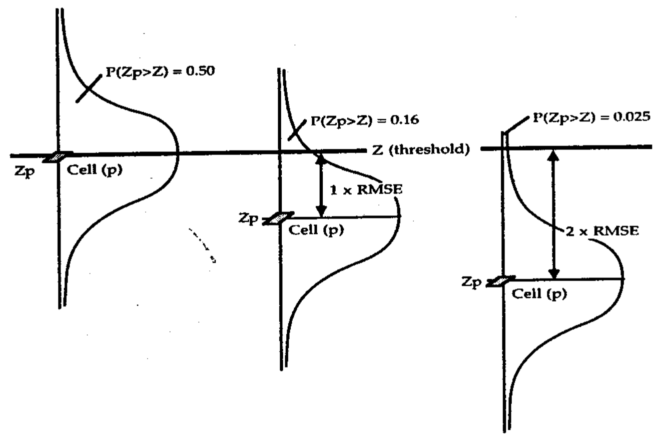


Figure 1a: $Z_p-Z=0$ Figure 1b: $Z_p-Z = 1 \times$ RMSE Figure 1c: $Z_p-Z = 2 \times$ RMSE

While few GIS have a function that will allow cell probabilities to be calculated (IDRISI is an exception), it was found that other means could be just as effectively employed to achieve the same result. Using the Arc/Info GRID software, the procedure adopted was to first select all cells which contained elevation values within nominated error distribution limits centred around the threshold elevation (in this case, 350m ±2 RMSE, or 350m ±14m). From Figure 1c, it can be deduced that cells selected within this range have between 2.5% and 97.5% probability of possessing a true value that exceeds 350m. At this stage, it should be noted that the choice of RMSE limits is context-dependent and best left to users to determine according to their needs. Next, cells are displayed by means of shading which is applied so as to gradually vary between nominated extremes in user defined increments.

In Figure 2a, the 350m contour for the test site is shown as a thick line, superimposed over contours at an interval of 20m. In Figure 2b, cells outside the limits of 350m ±2 RMSE have been masked in neutral gray colour, and the key legend at the right of the image depicts the shade ramp chosen which varies from white (350m – 2 RMSE) through to black (350m + 2 RMSE). The shade ramp was constructed in 1 metre increments from 336m through to 364m. In this case, the shade ramp represents an answer to the query "Show the probability of a cell exceeding the threshold value of 350m". For further details and colour illustrations of the output (which the authors believe enhance the visual representation of elevation error) readers are referred to the paper by Hunter and Goodchild (1994a).

Having visualised the uncertainty in the position of the 350m elevation, the question remains as to how the information can be applied in practice. From a management perspective, users must choose between either reducing the uncertainty in their data or else absorbing (accepting) it.

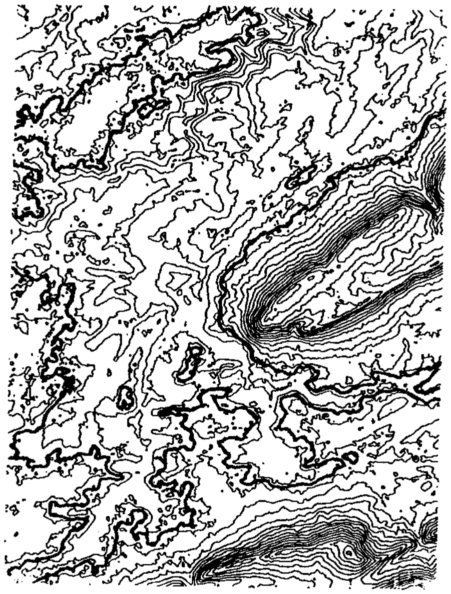


Figure 2a: The 350m contour for the test site is shown as a thick line, superimposed over contours at an interval of 20m.

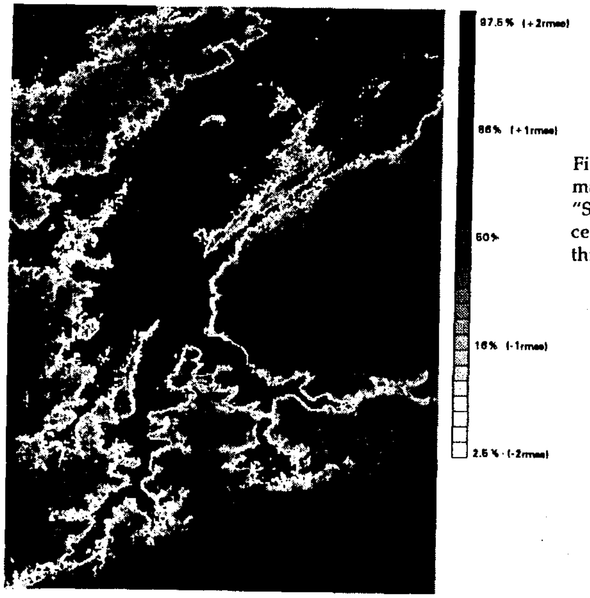


Figure 2b: A probability map in answer to the query "Show the probability of cells exceeding the 350. threshold value".

For instance, if a critical site is located near the western edge of the image where there are large areas of flat terrain with elevations very close to the 350m value, the usual form of error reduction would be to recollect and reprocess elevation data with a higher accuracy for the area of interest. Once the uncertainty in delineating the 350m elevation is satisfactorily reduced, the user must then absorb any remaining uncertainty. Generally, this occurs by simply accepting the fact that the data will never be perfect and that there will always be some likelihood of error because of the means by which DEMs model reality. In other cases, additional factors may be built in to the absorption process. For example, with reservoir construction a buffer zone is applied around the full supply level contour to be certain that no adjoining private or environmentally significant lands are inadvertently inundated.

Alternatively, there may need to be a change in the way elevations are perceived by thinking of them in terms of confidence levels. For instance, a user may want the 350m elevation depicted with a confidence level of 90%, in which case cells with at least a 0.90 probability of exceeding 350m would be selected (that is, 350m + 1.282 RMSE, or 359m). This concept already occurs in flood plain mapping where it is standard practice to compute flood levels which, for instance, have a 1% probability of being met or exceeded in any year (termed a 100-year event). In this case, engineers are able to live with the uncertainty that occurs in natural phenomena, and it is suggested that GIS users need to develop similar attitudes to the data they use and the products they develop.

CASE STUDY 2

The second case study relates to mapping areas burnt by forest fires through the use of remotely sensed Thematic Mapper (TM) imagery. In rugged terrain in particular, many researchers have reported the problem of confusion between shadowed and burnt regions as they both appear the same in most of the bands. In addition, if terrain corrections are not applied for the differences in gradient between cells it is difficult to discriminate between degrees of fire severity, since a given area may seem darker not because the fire was more intense but because it lies on a slope that receives less light by area unit.

Traditionally, DEMs have not been used as part of the assessment for fire severity mapping, but research at the University of California, Santa Barbara, is underway to determine how consideration of topographic effects on the satellite signal may be used to counter this problem, which requires a radiance model to be applied to correct (or normalise) the radiance data for terrain differences. Normalised radiance values are already commonly used in other applications of remote sensing in mountainous areas, and once derived may be used with traditional image analysis techniques such as supervised and unsupervised image classification, and density slicing.

The test site lies in central Portugal near Pampilhosa da Serra, and a DEM with a cell size of 30m x 30m was used as the basis for normalisation of the TM imagery. Figure 3a (left) shows a hill-shaded view of the DEM covering the test site, while the same area is also delineated on the unclassified TM Band 4 scene in Figure 3b (right), in which the effects of fire clearly show in the middle of the image as regions of dark gray/black colour. The portion of the DEM used for this research measures 353 rows by 272 columns (or about 10.6km by 8.2km), with elevations ranging from 287m to 1020m. Unlike DEM data supplied by the U.S. Geological Survey, an estimate of the elevation error for the DEM is not available, and so on the advice of researchers familiar with the Portuguese digital mapping program the standard deviation for elevation error has been estimated to be 10m.

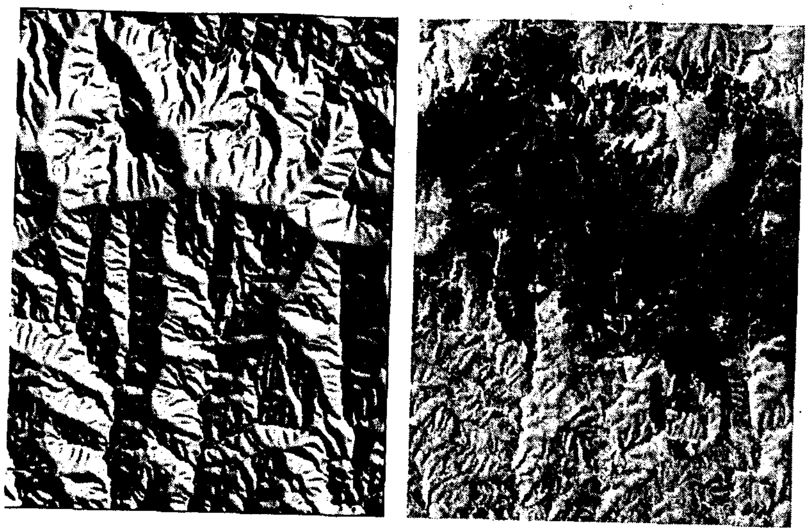


Figure 3a (left) showing a hill-shaded view of the test site DEM, and Figure 3b (right) clearly showing the darker burnt area in the unclassified TM scene.

The traditional procedure used by image analysts is to calculate the slope gradient and aspect for each cell in the DEM and then combine them with the sun's zenith and azimuth angles at the time of image capture (taken from the file header or else calculated for the time of day and the latitude and longitude of the site). This information is used to compute the cosine of the incidence angle, which has values in the range -1.0 to +1.0. Thus, a cell with an aspect equal and opposite to the sun's azimuth (in other words, facing the sun), and a gradient equal to the sun's zenith angle (that is, perpendicular to the sun's rays), will receive the maximum amount of radiance and have an incidence angle of 0° with a cosine of +1.0. The formula for the cosine of the incidence angle (i) is given by equation (1).

$$\cos(i) = \cos(\text{sun zenith}) * \cos(\text{cell gradient}) + \sin(\text{sun zenith}) * \sin(\text{cell gradient}) * \cos(\text{sun azimuth} - \text{cell aspect}) \quad (1)$$

Cells which have an incidence angle cosine equal to or less than zero either lie in a plane parallel to the direction of the sun's rays or else are on reverse hill slopes. These cells are deemed to be in 'self shadow' and are not operated on in the traditional research procedure due to the difficulty of working with diffused light. There is a further process which identifies cells that are in 'cast shadow' from larger features which obscure them from direct sunlight, and these cells also are usually excluded from further calculations.

The incidence angle cosines for all cells in the DEM are then used as a means of normalising the radiance values of pixels in the TM image, given that radiance is affected by the nature of the terrain to which it applies. At this point it should be noted that corrections will have already been made to ensure that both the DEM and the TM image have the same geo-referencing and cell/pixel size. The radiance values (L), being the raw signals from the image in the range 0 to 255, are then normalised by computing the value they would have if each pixel was horizontal (L_H), as in equation (2).

$$L_H = \frac{L}{\cos(i)} \quad (2)$$

Having discussed the analyst's traditional procedures for deriving normalised radiance values for each pixel, it is clear there is considerable potential for applying the model developed by Goodchild et al. (1992) to assess the uncertainty present in the final output, which would include any effects arising from the DEM elevation error, the algorithms used to calculate slope gradient and aspect, and the formulae applied to determine the incidence angle cosines and the normalised radiance value. It should be noted at this time that the process was halted at the point where realised radiance values are calculated, however there is no reason why the realisation process could not continue through to the next stage of analysis of fire severity.

The difference between the traditional approach to calculating the terrain corrected L_H values and the proposed approach which permits uncertainty to be assessed, can be seen in Figure 4. The latter technique applies elevation noise files, with varying levels of spatial autocorrelation (ρ), to the original DEM to establish corresponding sets of slope gradient and aspect files for the test site. Pairs of gradient and aspect realisation files (for each given ρ value) are then taken in turn and used to calculate the corresponding incidence angle cosine file (i), which is applied to the original TM radiance file (L) used for the analysis. The process results in the creation of a family of realised L_H files whose outputs can then be analysed. The entire process was automated by using a macro command script and applied using the Arc/Info GRID software. For further details of this particular case study, readers are referred to the paper by Hunter and Goodchild (1994b).

The adopted procedure resulted in a set of 10 realised L_H files for each of the ρ values 0.0, 0.5, 0.10, 0.15, 0.20, 0.21, 0.22, 0.23, 0.24 and 0.245 (the limit of ρ being 0.25 when taking into account the four neighbouring cells with sides common to the cell under consideration). For the purpose of analysis, the 96,016 grid cells in each realised file were subtracted from their corresponding cells in the original L_H file to provide a 'difference' file. This difference represents the amount by which the final L_H 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 set of 10 difference files was then calculated for the range of ρ values applied. The results were plotted graphically and a gradual increase was observed in the means and standard deviations of the differences as ρ varied from 0 to 0.20, followed by a sudden decrease as ρ approached 0.25. At this stage, analysis of the results shows that the average greatest difference that might be expected in L_H values is about 2.5 units with a standard deviation of approximately 13 units. These extremes occur around $\rho = 0.20$. However while such global statistics are useful in their own right, they say nothing about the spatial variation of the differences and, accordingly, further analysis was made of the realisations made at $\rho = 0.20$.

Taking the 10 realised L_H difference files at $\rho = 0.20$, a composite file was calculated and displayed such that cells with an L_H difference within 2 standard deviations of the overall mean for the file were shaded as mid-gray colour, while cells with an L_H difference below and above 2 standard deviations were shaded as white and black colour respectively. The result is shown in Figure 5a where it can be seen that the white and black cells, representing outlying values or those most susceptible to the spatial operations applied, tend to occur on west-facing slopes of north-south ridgelines – when compared with a hill-shaded view of the test site DEM with contours overlaid at an interval of 100m (Figure 5c).

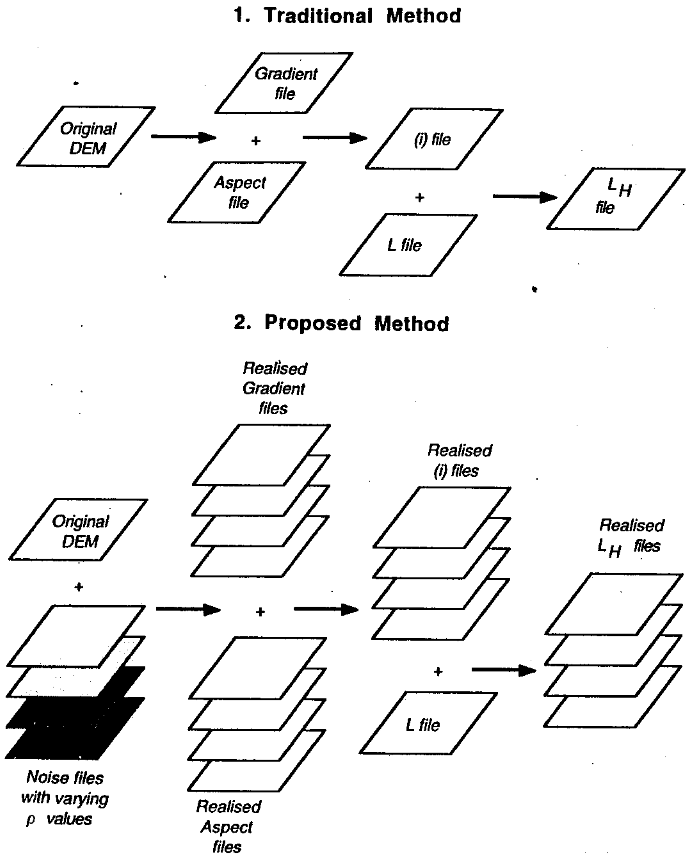
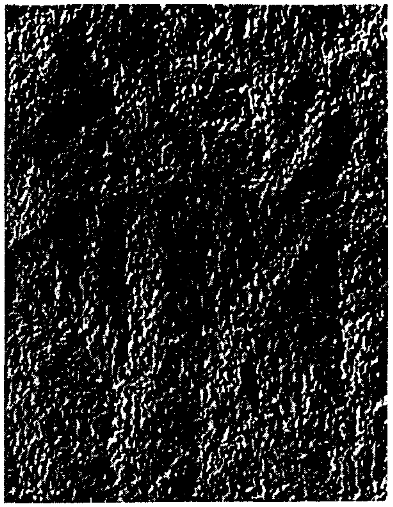
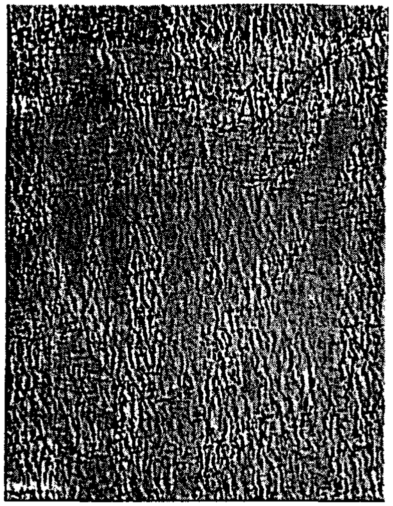


Figure 4: Comparison between the traditional technique of creating the L_H file, and the proposed method which provides for assessment of L_H uncertainty.

The file used in Figure 5a was then hill-shaded from the north-east to communicate both the size and spatial variation of the differences, and cell values beyond the ±2 standard deviation threshold show as a highly disturbed pattern while cells with differences within the threshold display as relatively smooth gray colour (Figure 5b). One site, in particular, in the top north-east corner of the image contains a significant L_H difference witnessed by its long shadow extending to the south-west. Given that this file represents the mean difference value occurring after 10 independent realisations, there is the suggestion of an anomalous TM radiance value present which warrants closer inspection of the original data.





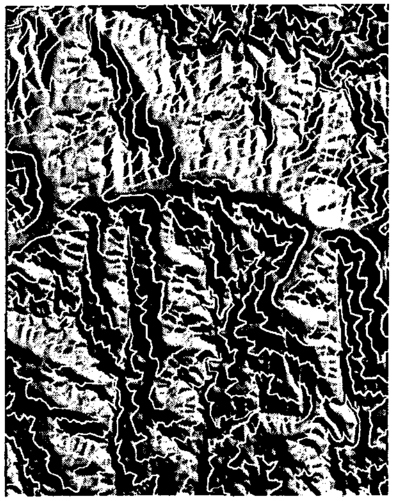


Figure 5a (top left): Showing the mean L_H difference file after 10 realisations at ρ = 0.20, with cells below and above the ±2 standard deviation threshold shown as white and black colour respectively.

Figure 5b (top right): Showing a hill-shaded view of the L_H difference file used in Figure 5a, with cells outside the ±2 standard deviation threshold showing as disturbed areas (note the anomaly in the upper right corner).

Figure 5c (bottom left): Showing a hill-shaded view of the DEM by comparison with 100m contours overlaid.

Having illustrated the spatial variation in the uncertainty of the L_H values, further explanation was sought as to the reason for the apparent correlation between significant differences in L_H and west-facing slopes. This can be explained by the location of the sun at the time of the TM image capture, which was at an azimuth of 117° and zenith angle of 36° (during the middle of the northern hemisphere summer). From this position, the pixels in shadow are clearly affected the most, which confirms the problems encountered when working with pixels in diffused light. It was for this reason that cells found to be in shadow (and thus having an incidence angle cosine < 0) were not removed from the realisation process, but instead deliberately retained to demonstrate any likely susceptibility to variation. Thus, the masking of such pixels during traditional analysis may be considered a valid approach to the problem.

At the same time, it was seen that for the remainder of the image the greatest mean difference in L_H that might be expected is about 2.5 units with a standard deviation of 13 units. It is left to the user of the data to decide whether the tolerances are acceptable for the task at hand, and this assessment of product quality (or fitness for use) forms part of the management approach which needs to be adopted in such cases. If the variation is acceptable then the methodology proposed has confirmed that the particular combination of DEM and TM imagery; the algorithms for gradient, aspect and incidence angle cosine; and the model for terrain correction of L_H value; is suitable for the purpose intended. On the other hand, if these differences are unacceptable then uncertainty reduction methods will need to be employed: such as choosing more accurate DEM data; selecting alternative algorithms and models; or employing TM imagery from other epochs. To this end, the realisation process may be repeated using different combinations of data, algorithms and models to determine which one produces the least uncertainty in the final product.

CONCLUSIONS

Factors such as mandatory data quality reporting requirements, the increasing view of spatial data as a commercially saleable commodity, and the need to protect the integrity of decisions made by regulatory agencies using GIS against legal challenges, have now bought about a critical need to be able to assess whether the quality of GIS products meets the quality requirements of the tasks for which the information is to be employed. This paper has presented two case studies in which techniques for judging the quality of GIS products are investigated. In the first study, simple probability mapping is employed to communicate the error in elevation arising from the use of Digital Elevation Models (DEMs). In the second case study, a more advanced methodology is presented to assess the effect of DEM error on the terrain corrections applied to remotely sensed imagery in mountainous terrain. It is suggested that techniques such as these can help GIS users to overcome some of the data quality reporting problems now being faced.

REFERENCES

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.

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

Hunter, G.J. and Goodchild, M.F., 1994b, "Design and Application of a Methodology for Reporting Uncertainty in Spatial Databases", *Proceedings of the URISA '94 Conference*, Milwaukee, Wisconsin, Vol. 1, pp. 771-784.

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.

ACKNOWLEDGMENTS

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.