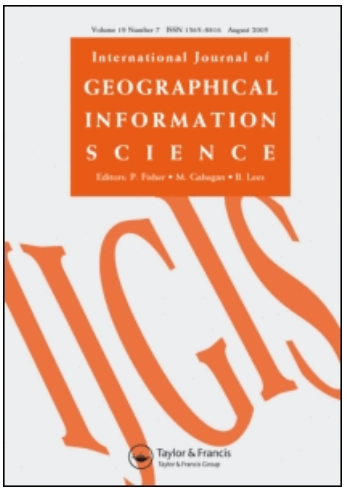

International Journal of Geographical Information Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713599799

Sensitivity analysis of a decision tree classification to input data errors using a general Monte Carlo error sensitivity model

Zhi Huang [§] [a]; Shawn W. Laffan [b]
[a] The Fenner School of Environment and Society, Australian National University, Canberra, ACT, Australia [b] School of Biological, Earth & Environmental Science, University of New South Wales, Sydney, NSW, Australia

Online Publication Date: 01 November 2009

To cite this Article Huang [§], Zhi and Laffan, Shawn W.(2009)'Sensitivity analysis of a decision tree classification to input data errors using a general Monte Carlo error sensitivity model',International Journal of Geographical Information Science,23:11,1433 — 1452
To link to this Article: DOI: 10.1080/13658810802634949
URL: http://dx.doi.org/10.1080/13658810802634949

International Journal of Geographical Information Science
Vol. 23, No. 11, November 2009, 1433–1452

Research Article

Sensitivity analysis of a decision tree classification to input data errors using a general Monte Carlo error sensitivity model

ZHI HUANG[§]*† and SHAWN W. LAFFAN‡

†The Fenner School of Environment and Society, Australian National University, Canberra, ACT, Australia

‡School of Biological, Earth & Environmental Science, University of New South Wales, Sydney, NSW, Australia

(*Received 15 October 2007; final version received 20 October 2008*)

We analysed the sensitivity of a decision tree derived forest type mapping to simulated data errors in input digital elevation model (DEM), geology and remotely sensed (Landsat Thematic Mapper) variables. We used a stochastic Monte Carlo simulation model coupled with a one-at-a-time approach. The DEM error was assumed to be spatially autocorrelated with its magnitude being a percentage of the elevation value. The error of categorical geology data was assumed to be positional and limited to boundary areas. The Landsat data error was assumed to be spatially random following a Gaussian distribution. Each layer was perturbed using its error model with increasing levels of error, and the effect on the forest type mapping was assessed. The results of the three sensitivity analyses were markedly different, with the classification being most sensitive to the DEM error, than to the Landsat data errors, but with only a limited sensitivity to the geology data error used. A linear increase in error resulted in non-linear increases in effect for the DEM and Landsat errors, while it was linear for geology. As an example, a DEM error of as small as $\pm 2\%$ reduced the overall test accuracy by more than 2%. More importantly, the same uncertainty level has caused nearly 10% of the study area to change its initial class assignment at each perturbation, on average. A spatial assessment of the sensitivities indicates that most of the pixel changes occurred within those forest classes expected to be more sensitive to data error. In addition to characterising the effect of errors on forest type mapping using decision trees, this study has demonstrated the generality of employing Monte Carlo analysis for the sensitivity and uncertainty analysis of categorical outputs that have distinctive characteristics from that of numerical outputs.

Keywords: error modelling; vegetation mapping and modelling; land use and land cover; terrain analysis

1. Introduction

Land cover maps play a very important role in natural resource management, both as a final product and as an input to other modelling processes. Land cover classification is used to develop these maps and is often conducted using GIS layers as predictor variables (e.g., topographic, soil, climatic and remotely sensed). All of

[§]Current address: Marine and Coastal Environment Group, Petroleum Marine Division, Geoscience Australia, Canberra, ACT, Australia
*Corresponding author. Email: Zhi.Huang@ga.gov.au

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2009 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810802634949

the variables used in the classification process contain some level of error (Goodchild 1989, Unwin 1995, Van Niel *et al.* 2004, Richards and Jia 2006). This error will be propagated through the classification process, with its effect dependent on the sensitivity of the analysis to that error. There is therefore a constant need to assess its potential impact on land cover classifications.

In terms of modelling error in GIS operations, Veregin (1989) considered a 'hierarchy of needs' as error source identification, error detection and measurement, error propagation modelling, strategies for error management and strategies for error reduction. There has been a great deal of research on these topics since then (Foody 2003). Consequently, sensitivity and uncertainty analyses have been the subject of much attention in spatial and environmental sciences (Crosetto *et al.* 2000, Crosetto and Tarantola 2001, Jager and King 2004). Sensitivity and uncertainty analyses have been used to assess model parameters (Hamby 1994, Hwang *et al.* 1998, McKenney 1999), and input data, both continuous (Davis and Keller 1997, Wang *et al.* 2000, Goovaerts 2001, Canters *et al.* 2002, Gertner *et al.* 2004) and categorical (Goovaerts 1996, Finke *et al.* 1999, Canters *et al.* 2002, Hines *et al.* 2005).

The objective of sensitivity analysis is to help the strategies of error management and reduction. We follow the definition of Jager and King (2004), where the sensitivity analysis is the assessment of which spatially distributed input variables the model is most sensitive to. Therefore, it is not concerned with the actual properties of input error or uncertainty, for example, the error magnitude and its distribution (Jager and King 2004). It does need, however, to characterize uncertainty using reasonable assumptions (Hines *et al.* 2005). There are three decisions to make in terms of conducting a sensitivity analysis. They include a reasonable uncertainty assumption, a suitable error model and an appropriate sensitivity analysis method. Sensitivity analyses can be divided into two groups. One deals with model (parameter) uncertainty and the other deals with input data uncertainty. This research is concerned only with input data uncertainty.

Error models are generally needed to investigate how data error is propagated through a modelling process and can be divided into formal mathematical models and simulation models. Formal mathematical models, such as those described by MacDougall (1975), Taylor (1982), Geman and Geman (1984), Newcomer and Szajgin (1984), Veregin (1989, 1995), and Goodchild *et al.* (1992), have been used to model error propagation through simple GIS overlay functions (e.g., RESELECT, AND, OR, XOR, addition, ratios, univariate overlay, logic functions and area measurement) (Drummond 1987, Walsh *et al.* 1987, Heuvelink *et al.* 1989, Lanter and Veregin 1992, Haining and Arbia 1993, Arbia *et al.* 1998). Simulation models such as Monte Carlo analysis have been strongly recommended for error propagation analysis (Lodwick 1989, Openshaw 1989). The advantage of simulation models over formal mathematical models is that their applications are not limited to simple GIS functions. Instead they are theoretically applicable to any function. For example, simulation models have been used for the buffer function (Veregin 1994, 1996, De Genst *et al.* 2001), digital elevation model (DEM) derivation (Lee *et al.* 1992), logical models and continuous classification (Heuvelink and Burrough 1993) for Bayes theorem (Aspinall 1992) as well as for GIS overlay functions (Openshaw *et al.* 1991). It is also not possible to build a mathematical error model for a classification process that results in discrete classes as it is not continuously linearly differentiable.

Through sensitivity analysis methods, we can rank the sensitivity of the modelling process to individual inputs and even partition individual error contributions to the

model output. Many sensitivity analysis methods have been proposed (see Hamby 1994 for a good review), some of which have been used by GIS researchers. They include the one-at-a-time method (McKenney 1999), variance-based method (Finke *et al.* 1999, Crosetto *et al.* 2001), automatic differentiation (Hwang *et al.* 1998) and the regression method (Gertner *et al.* 2004). Most of the existing sensitivity methods, however, are designed for outputs that are interval or ratio data types and are not suitable for categorical (nominal) data. Multi-source classification often involves using both categorical data and continuous data. The natures of these data clearly differ, and so different strategies have been developed for each.

According to Goodchild *et al.* (1992), an exact probability distribution of any pixel belonging to different classes is required to corrupt (rasterized) categorical data. If the data are a product of classification process, one often-used approach to fulfil this requirement is to use a classification confusion matrix (e.g., Hines *et al.* 2005), although class memberships derived from soft classification can also be used (Canters *et al.* 2002). The Semantic Import Model of expert knowledge has been proposed for when categorical data are not the result of a classification process (Davis and Keller 1997). Geostatistical approaches such as joint sequential simulation and sequential indicator simulation from sample data are also used (e.g., Goovaerts 1996, Finke *et al.* 1999).

One approach to corrupting continuous data is to generate a random error surface based on an assumed error distribution model and then add it back to the original data. The random error surface can be generated using pseudo-random number generators (Van Niel and Laffan 2003). The other common approach is to use geostatistical methods to directly generate an effective input data surface from samples (e.g., Davis and Keller 1997, Atkinson 1999, Goovaerts 2001, Canters *et al.* 2002, Gertner *et al.* 2002, 2004). The advantage of the geostatistical approach is that it can take into account the spatial autocorrelation of a single attribute as well as spatial cross-correlation between different attributes.

As noted above, the subjects of most previous uncertainty studies are continuous modelling outputs, for which it is relatively simple to partition the uncertainty contributions due to the inputs. Comparatively little effort has been spent in assessing the impact of errors for non-continuous data. One possible reason for this is that the outcome of land cover classification is a categorical (nominal) data type, and this does not naturally fit into traditional methods of sensitivity and uncertainty analyses. The nature of classification into land cover classes may also make these systems less sensitive to input data errors than for continuous outputs.

This study is an attempt to fill this gap in our knowledge of the effects of error on land cover classifications, addressing one obvious question among many others: How sensitive is land cover classification, in this case forest type mapping, to the error and uncertainty of input data?

To address this question, we present a case study of sensitivity analysis of a decision tree derived forest type mapping to errors in both categorical and continuous data. In Section 2, we describe the study area, input data and baseline classification. We then present three separate sensitivity analyses for the three input data sources, which include a new approach of corrupting categorical data. The results of the sensitivity analyses and the comparison analyses are summarized in Section 3, followed by discussion of the causes of the observed results.

2. Methods

The purpose of a sensitivity analysis is to investigate and evaluate the sensitivity of model output to model input(s). The need for a sensitivity analysis only arises when we know that a certain level of uncertainty is present in the model input(s). A sensitivity analysis is only feasible when we have sufficient knowledge and thus can make a reasonable assumption about the nature of the input uncertainty. After that, we can select an appropriate error model to simulate the uncertainty assumption. When there are multiple model inputs involved in a sensitivity analysis, methods must be used to allow the evaluation and ranking of the model sensitivities to individual inputs.

2.1 *Study area and data set*

The study area is a 15.75 km by 15.75 km region at Kioloa, NSW, Australia (Figure 1). It has been extensively used for land cover classification research (e.g., Lees and Ritman 1991, Moore *et al.* 1991, Fitzgerald and Lees 1996, Gahegan *et al.* 1999, Huang and Lees 2004, 2005).

The study area has a large variety of vegetation types ranging from eucalypt-dominated scherophyll forest to warm temperate rain forest (Moore *et al.* 1991). It is extremely complex in both physiography and parent material, resulting in complex distributions of the vegetation types. There are approximately 450 forest species that have been classified into 30 forest communities on the basis of dominant species and the composition of understory species (Moore *et al.* 1991). These were subsequently aggregated into seven forest types with additional ocean and cleared classes (Lees and Ritman 1991) (Table 1, Figure 1). The boundaries of the seven forest types are not clear-cut, with considerable spatial overlap.

A base classification (Figure 1, see also Supplementary Material) was developed using the C4.5 decision tree algorithm (Quinlan 1993). This decision tree was used for all the sensitivity analyses in this study (the input data were perturbed, the tree was not). The input in independent variables used to develop the tree are all in raster format with a 30 m pixel size (275,625 pixels), which can be grouped into three sets. The first set includes a DEM (elevation range 0–280 m) and its derivatives slope and aspect. The second set is the geology variable (seven categories: Quaternary Alluvium, Tertiary Essexite, Snapper Point Permian, Pebbly Beach Permian, Wasp Head Permian, Ordovician metasediments and Ocean). The Permian categories are all marine sediments. The third set is three Landsat Thematic Mapper (TM) bands: band2 (digital numbers (DN) range 14–101), band4 (5–89 DN) and band7 (0–86 DN). The classification scheme consists of seven forest types and two other land cover classes (cleared land and water) (Table 1). The dependent variable of this study is a layer of 1708 ground samples. The samples were collected through several stages over four years by several authors (Lees and Ritman 1991). A randomly selected 80% of the ground samples were used for training the decision tree. The remaining 20% of samples were used for testing classification accuracy of the base data set and the results of sensitivity analyses. The overall accuracy of the base classification is 65.1%. The relationships between the classification rules and the input data set are summarised in Tables 2 and 3, from which it is evident that DEM and slope were used most often.

2.2 *Sensitivity analyses*

We used a Monte Carlo simulation method, as it is more readily applied to the C4.5 decision tree classifier than mathematical methods like Taylor series (e.g., Davis and

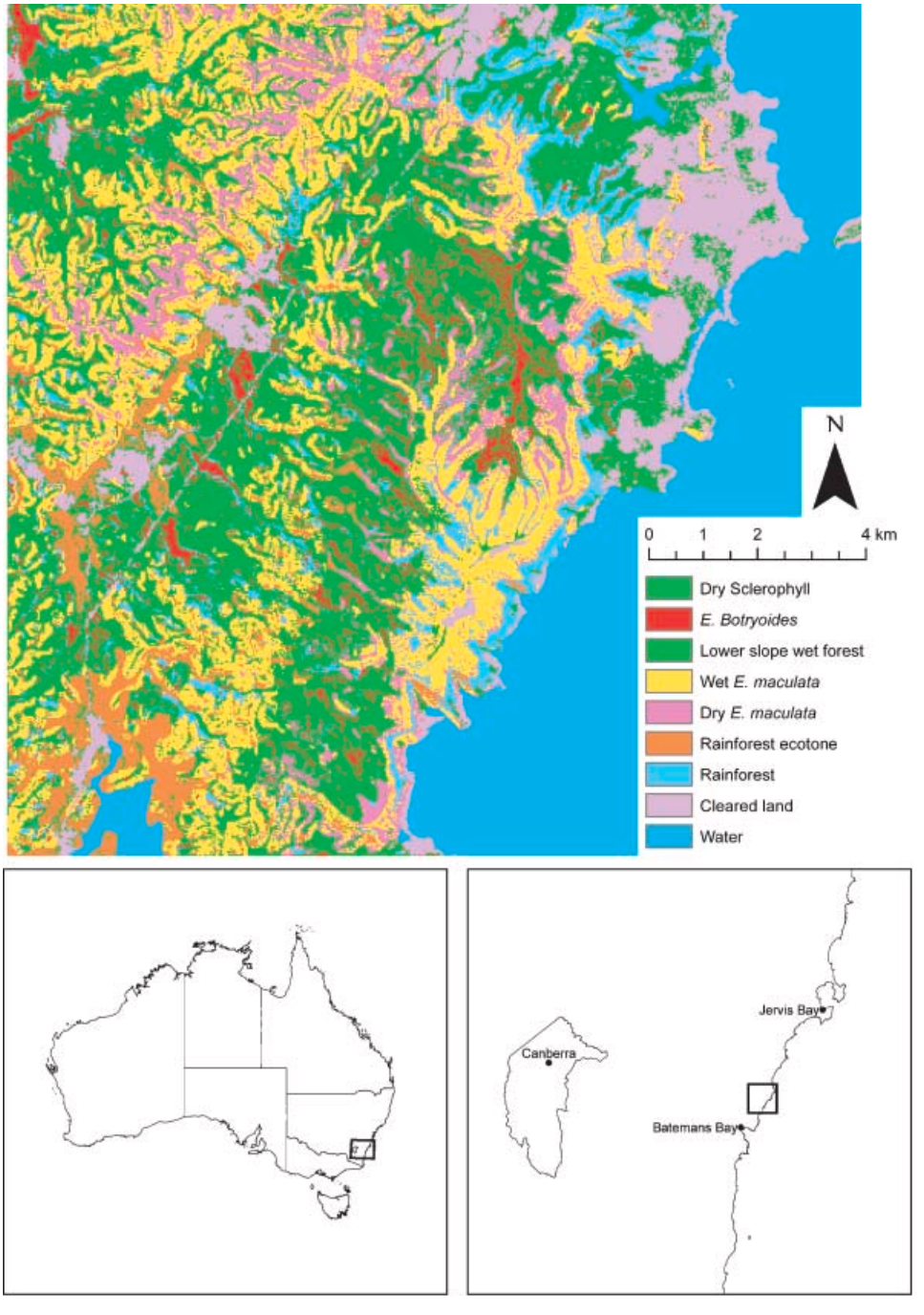

Figure 1. Benchmark forest type map and study area location.

Keller 1997, Crosetto *et al.* 2000, Crosetto and Tarantola 2001, Canters *et al.* 2002). The approach used in a Monte Carlo analysis is to perturb one or more of the input datasets at some chosen level of random error to generate a different realisation of the original data set. This process is repeated for some number of iterations until a

Table 1. Land cover class types and frequencies in the base forest type data set.

Class	Forest type	Cell count
1	Dry sclerophyll	76,315
2	*Eucalyptus botryoides*	7375
3	Lower slope wet forest	6588
4	Wet *E. maculata*	46,364
5	Dry *E. maculata*	15,501
6	Rainforest ecotone	11,568
7	Rainforest	14,063
8	Cleared land	20,213
9	Water	77,638

Table 2. Number of classification rules that use each data set, summarised by class.

Class	Band2	Band4	Band7	DTM	Slope	Aspect	Geology	Number of rules
1	2	2	1	3	3	1	0	3
2	2	4	5	7	4	3	3	7
3	0	0	3	3	3	2	0	3
4	2	2	0	2	4	1	3	4
5	5	3	5	8	10	2	9	10
6	1	2	2	4	5	1	4	5
7	3	3	2	6	5	4	5	8
8	1	0	2	2	0	1	0	3
9	0	0	1	1	0	0	0	1
Total	16	16	21	36	34	15	24	44

Table 3. Number of conditions in which each data set was used in classification rules, summarised by class.

Class	Band2	Band4	Band7	DEM	Slope	Aspect	Geology	Number of conditions
1	3	2	2	4	4	1	0	15
2	2	5	6	11	5	3	3	36
3	0	0	4	5	4	2	0	15
4	3	2	0	2	4	1	3	15
5	5	4	6	9	11	2	9	46
6	1	2	2	5	6	2	4	22
7	3	3	2	9	5	5	5	32
8	1	0	2	3	0	1	0	7
9	0	0	1	1	0	0	0	2
Total	18	18	25	49	39	17	24	190

Note that each data set can be used more than once in each rule.

stable solution is achieved. In this case we used 1000 iterations to ensure convergence, as recommended by Heuvelink (1998). All random values were generated using the Mersenne Twister pseudo-random number generator, as recommended by Van Niel and Laffan (2003).

We used the one-at-a-time sensitivity analysis approach where each variable is perturbed separately (Hamby 1994, McKenney 1999), thus simplifying the interpretation of results. In this case the variables are the data sources (DEM, Landsat image and geology). The DEM and its derivatives were assessed together, as

were the three bands of the satellite image. For example, when the DEM is perturbed, slope and aspect are calculated from the perturbed DEM. The original decision tree classifier was used for each perturbation. Each data source has different errors and each needs to be modelled differently to obtain a fuller understanding of the sensitivity of the classification to error.

2.2.1 The DEM and its derivatives. The DEM is a continuous variable. It was interpolated from elevation contours digitised from the 1:25,000 scale topographic map of the study area, and therefore includes errors in map production, digitisation and interpolation. In this study, we assumed the DEM error is stochastic in nature but with a spatially autocorrelated distribution (Van Niel *et al.* 2004). Also, the DEM error was assumed to be directly associated with, and a small percentage of, the elevation value at individual pixels. This is an appropriate error model for this landscape, as error magnitudes will be larger in the steeper slopes of the upland parts of the study area due to spatial offsets and lower in the comparatively flatter coastal parts. Other landscapes will require a different error model, but this will not require changes to the sensitivity analysis approach.

The sensitivity analysis of the DEM and its derivatives was conducted by perturbing the DEM using five levels of spatially autocorrelated random error, respectively, $\pm 2\%$, $\pm 4\%$, $\pm 6\%$, $\pm 8\%$ and $\pm 10\%$ of individual elevation values. The general Monte Carlo procedure used involves the following steps:

(1) Randomly select 1000 pixels from the DEM (this is an average spacing of approximately 17 cells for this data set).
(2) Randomly generate error values for the 1000 selected pixels, calculated as a percentage of each elevation value. The percentage is from a uniform distribution within the chosen error level (e.g., $\pm 2\%$).
(3) Generate a spatially autocorrelated error surface from the 1000 pixels using ordinary kriging.
(4) Generate a perturbed realisation of the DEM by adding the error surface to the original DEM.
(5) Generate derivative data sets (slope and aspect) from the perturbed DEM.
(6) Generate a perturbed forest type map using the previously trained decision tree with the perturbed DEM, slope and aspect, and the original unperturbed geology and Landsat data sets.

The Monte Carlo analysis was run 1000 times to generate 1000 forest type maps for each of the five error levels.

2.2.2 The geology data. The geology variable is a categorical data type with the seven categories mapped as choropleths. The layer was digitised from Gostin's (1969) 1:25,000 scale geological map and, like the DEM, also includes mapping and digitising errors. We do not have classification outcomes, expert knowledge or sample data to derive probability distributions of individual pixels for the geology data. However, we can assume that some of the most important errors occur near the boundaries of the geology types due to the definition of the boundaries and their subsequent digitisation. This part of the attribute error can be treated as a direct function of the positional (boundary) error. Approaches such as error bands and epsilon bands (Perkal 1956, Goodchild and Hunter 1997), corridor of transition model (Davis and Keller 1997), confidence regions (Shi 1998) and rough sets (Ahlqvist *et al.* 2000, Fisher 2001) have been used to represent positional uncertainty

in vector and raster formats. We propose another way of representing boundary uncertainty for raster data. This is achieved by using a moving window approach and randomly assigning the centre pixel the class of one of its neighbours. By varying the window sizes, the width of boundary error varies. One could also employ a distance decay function to assign different probabilities for neighbours conditioned on their distances from the centre pixel, but this was not done here since we are using relatively small window sizes. The idea and approach are essentially the same as that described in Huang and Lees (2007) for the representation of fuzziness of location.

We applied five moving window sizes to represent five error levels in the geology data: 3 by 3, 5 by 5, 7 by 7, 9 by 9 and 11 by 11. The general Monte Carlo procedure used involves the following steps:

(1) Assign a new geology category to the processing (centre) pixel from a neighbouring pixel randomly selected from within the moving window (including the centre pixel).
(2) Generate a perturbed realisation of the geology data by repeating the above process for all pixels.
(3) Generate a perturbed forest type map using the previously trained decision tree with the perturbed geology data, and the original unperturbed DEM, slope, aspect and Landsat datasets.

As with the DEM sensitivity analyses, the Monte Carlo analysis was run 1000 times to generate 1000 forest type maps for each of the five error levels.

2.2.3 The Landsat data. The Landsat TM data was acquired in April 1988, close in time to the survey of vegetation sites. Signal to noise ratio is usually used to indicate the quality of the remotely sensed data. Noise (i.e., the combination of errors) in remotely sensed data often follows a Gaussian distribution and is independent from the signal (Richards and Jia 2006), so this was used as the random error distribution model for the three Landsat TM bands. We also assumed that the random error had mean of zero and used standard deviations (SD) of 0.25, 0.5, 0.75, 1 and 1.25 to represent five error levels. With the increase of the error levels, the signal-to-noise ratios of the Landsat data were effectively decreased. It should be noted that the quality of the remotely sensed data could also have been affected by such factors as atmospheric effects, geometric aspects, sensor errors and data pre-processing (Lunetta *et al.* 1991). These can be incorporated into further sensitivity assessments where they are known.

The general Monte Carlo procedure used involves the following steps:

(1) Generate a random error surface using a Gaussian distribution with a mean of zero and the specified standard deviation.
(2) Generate a perturbed realisation of the Landsat TM band2 by adding the error surface to the original Landsat TM band2.
(3) Repeat steps 1 and 2 to generate perturbed realisations of Landsat TM bands 4 and 7.
(4) Generate a perturbed forest type map using the previously trained C4.5 decision tree with the perturbed Landsat TM bands, and the original unperturbed DEM, slope, aspect and geology data.

As with the DEM and geology analyses, the Monte Carlo analysis was run 1000 times to generate 1000 forest type maps for each of the five error levels.

2.3 *Assessment of sensitivity analysis results*

Three criteria were used to evaluate the results of the three individual sensitivity analyses. We first assessed the overall test accuracy difference between each of the 1000 forest type maps and the base forest type map (the 'accuracy criterion'). Second, we assessed the number of pixels in the perturbed classification that changed their classes when compared to the base forest type map (the 'pixels changed criterion'). Third, a spatial assessment of the change in pixels was generated by calculating the frequency each pixel changed across all iterations for each error level. To assess the convergence on a stable solution, the mean and standard deviation of the criteria were assessed as the number of iterations increased.

3. Results

All of the results show an increase in sensitivity as the error increases, as is to be expected, but the rates differ between the different data sources (Figure 2). A small proportion of the perturbations for geology and Landsat show a higher accuracy than the original, but this difference is typically small. Less than 2% of the Landsat perturbations, and none of the geology perturbations, are more than 1% more accurate. All of the analyses converged on a stable solution after approximately 500 iterations.

3.1 *The DEM and its derivatives*

Figure 2 and Table 4 indicate the effect of the DEM error on the accuracy assessment of the forest type mapping. As is to be expected, the magnitude of the mean overall accuracy reduction and its variance increases with increasing error levels. However, the effect is not linear and becomes more prominent when the uncertainty level increases. However, one should also note that the range of the median accuracy differences over all error levels is less than 1%.

The pixels changed criterion shows a similar pattern to the accuracy criterion, although the mean change for each error level is close to 10% of all data set pixels. The uncertainty levels of $\pm 2\%$ and $\pm 4\%$ are very similar, but there is an increasing difference as the uncertainty level increases to $\pm 10\%$. These differences are also much more pronounced than for the accuracy criterion.

Spatial assessment of the impact of random DEM error on the results of forest type mapping shows some interesting results (Figure 3). While approximately 10% of all pixels changed their initial class assignment in each perturbed classification, there was considerable variation in which pixels changed. For the $\pm 6\%$ level of error, 26.5% of all pixels changed at least once across all perturbations, 13.1% changed 100 or more times, and 11.6% changed 200 or more times (Figure 3, Table 5). The *Eucalyptus botryoides* forest class was most affected by the $\pm 6\%$ DEM uncertainty level (Table 5). Cells assigned to this class in the base data set changed class assignments 558 times out of 1000 iterations on average, with 64% of cells changing at least 200 times. It is followed by lower slope wet forest and dry sclerophyll forest. Wet *Eucalyptus maculata* forest is the most tolerant to this DEM uncertainty level.

3.2 *The geology data*

Figure 2 and Table 6 indicate that employing some level of random error on the geology data does reduce the overall test accuracy. However, the magnitudes are

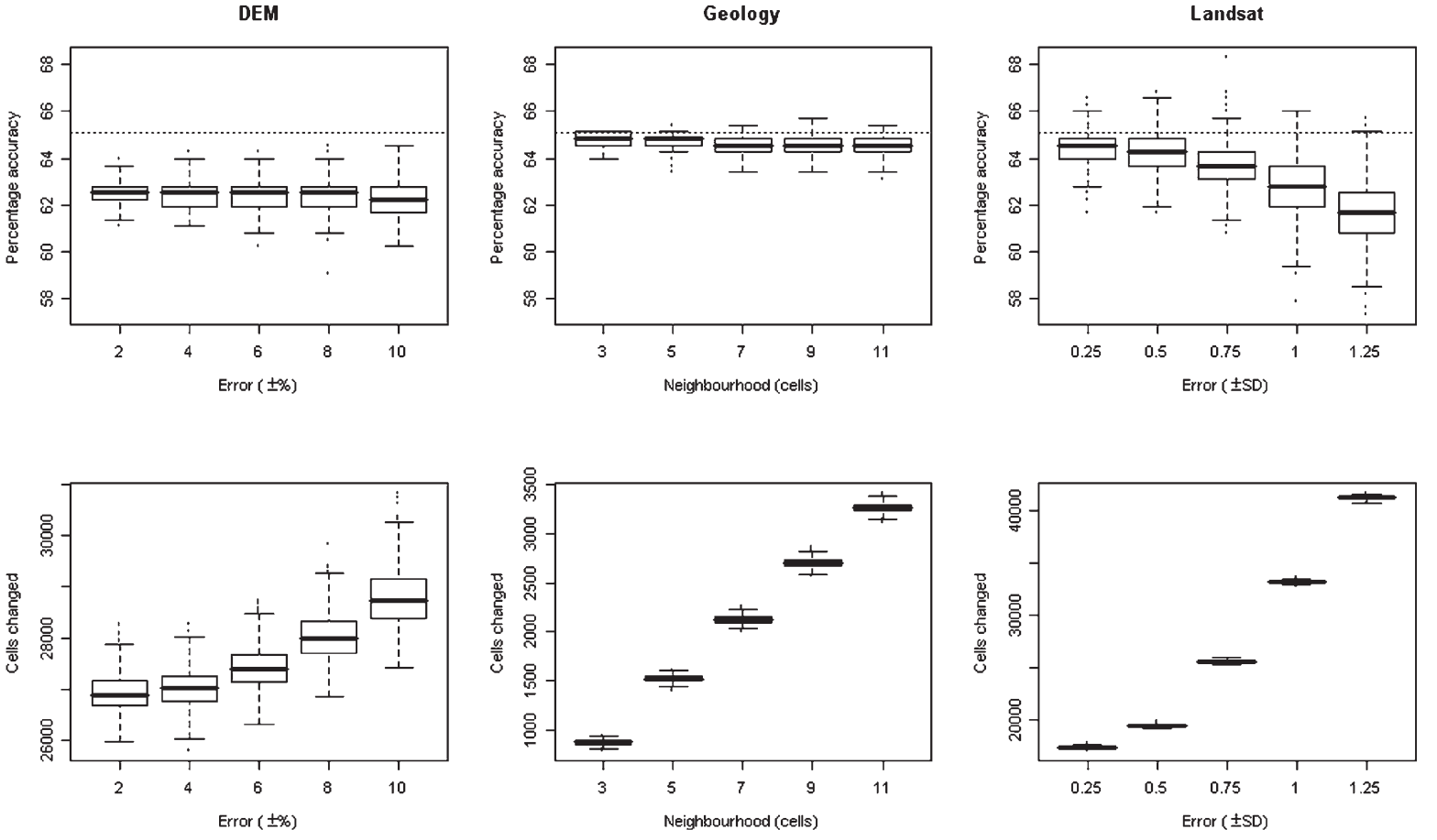

Figure 2. Boxplots of the analysis results for the accuracy (upper) and cells changed (lower) criteria. The accuracy of the benchmark classification is denoted by the dotted line in the accuracy plots.

Table 4. Summarised results of the sensitivity analysis of DEM and its derivatives. Accuracy values are in the interval [0,1].

	±2%	±4%	±6%	±8%	±10%
Difference from overall test accuracy					
Mean	−0.0263	−0.0265	−0.0268	−0.027	−0.0281
SD	0.0046	0.0053	0.0061	0.0067	0.0071
Number of pixels changed					
Mean	26,925	27,016	27,414	28,038	28,807
SD	371	368	403	461	556

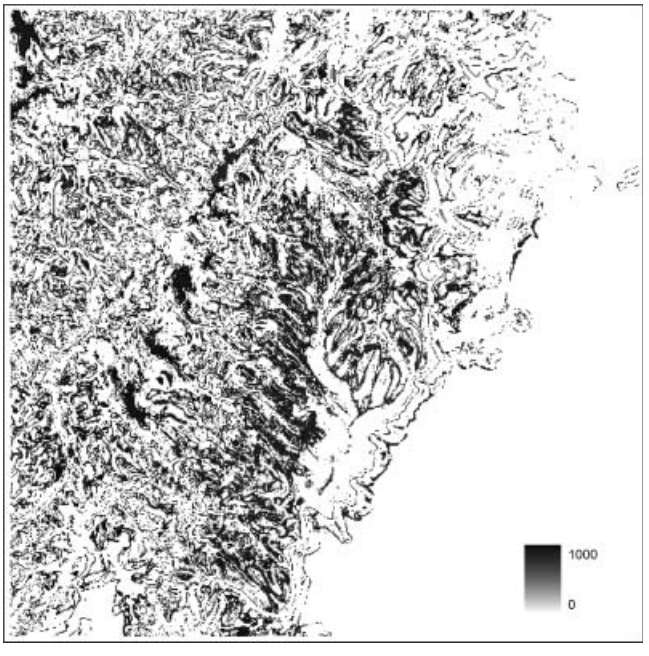

Figure 3. Spatial distribution of pixels changed with the ±6% DEM uncertainty level.

Table 5. Number of class changes for cells in each land cover class when using the ±6% DEM uncertainty level, summarised by class.

Class	Cell count	Min	Max	Mean	SD	Cells changed once (%)	Cells changed >100 times (%)	Cells changed >200 times (%)
1	76,315	0	1000	196	361	50.8	25.0	22.2
2	7375	0	1000	558	430	78.9	65.2	63.6
3	6588	0	1000	343	450	53.2	39.7	38.2
4	46,364	0	1000	32	153	18.6	4.8	3.6
5	15,501	0	1000	98	266	38.1	14.5	10.7
6	11,568	0	1000	155	324	39.2	20.4	18.5
7	14,063	0	1000	58	161	30.0	12.0	10.4
8	20,213	0	913	7	60	2.7	1.4	1.3
9	77,638	0	547	5	49	1.3	1.0	1.0
Overall	275,625					26.5	13.1	11.6

Table 6. Summarised results of the sensitivity analysis of geology data.

	3 by 3	5 by 5	7 by 7	9 by 9	11 by 11
Difference from overall test accuracy					
Mean	−0.00290	−0.00412	−0.00559	−0.00600	−0.00603
SD	0.002452	0.003074	0.003680	0.003956	0.004154
Number of pixels changed					
Mean	882	1524	2125	2701	3257
SD	24	31	38	43	45

very small at less than 1%. As is expected, the magnitude of the mean overall accuracy reduction and its standard deviation increases with increasing error levels. The rate of change is also close to linear. The mean overall accuracy reduction appears to be saturated by the 9 by 9 error level.

The pixels changed criterion is consistent with the accuracy criterion for geology. Figure 2 shows the assessment of the mean and standard deviation of the pixels changed criterion as the number of iterations increases. Figure 2 and Table 6 indicate that applying some level of random boundary error on the geology data makes a small proportion of pixels change their initial class assignments. As is expected, the magnitude of the mean number of pixels and its standard deviation increases with an increasing error level. However, the effect is near linear, and becomes slightly less prominent when the uncertainty level increases. There is also no overlap between the impacts of the different error levels.

When considering the spatial distribution, the impact of random geology (boundary) error on the results of forest type mapping is very different from the case of random DEM error (Figure 4). For example, for the 7 by 7 level geology data error, only 3.4% of all pixels changed at least once across all perturbations, 2.5%

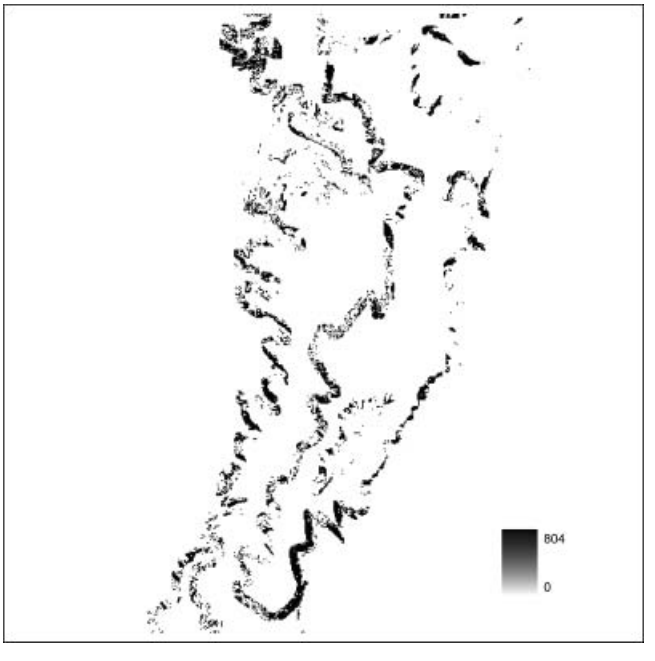

Figure 4. Spatial assessment of pixels changed with the 7 by 7 geology uncertainty level.

Table 7. Number of class changes for cells in each land cover class when using the 7 by 7 geology uncertainty level, summarised by class.

Class	Cell Count	Min	Max	Mean	SD	Cells changed once (%)	Cells changed >100 times (%)	Cells changed >200 times (%)
1	74,852	0	804	7	46.5	3.0	2.1	1.5
2	7215	0	604	9	52.6	4.0	2.8	2.0
3	6588	0	0	0	0.0	0.0	0.0	0.0
4	45,642	0	740	11	60.3	4.4	3.2	2.4
5	15,298	0	671	36	102.0	16.1	11.6	8.1
6	11,246	0	770	7	50.4	2.9	2.2	1.8
7	13,897	0	679	32	94.5	14.9	10.5	7.2
8	20,213	0	0	0	0.0	0.0	0.0	0.0
9	77,638	0	0	0	0.0	0.0	0.0	0.0
Overall	272,589					3.4	2.5	1.7

changed 100 or more times, and 1.7% changed 200 or more times (Table 7). As is expected, the changes of classification are confined to the areas of geology boundaries and also primarily within the forest classes. Dry *E. maculata* forest and rainforest were most affected by the 7 by 7 window (Table 7). Cells assigned to these classes in the base data set changed class assignments close to 30 times out of 1000 iterations on average. Lower slope wet forest is the most tolerant to the geology data uncertainty level. Geology is not used in the rules to define it, and so any effects are due to perturbations of neighbouring classes.

3.3 *The Landsat data*

The effect of the Landsat data error on the overall accuracy varies from 0.7% to 3.4%. (Figure 2 and Table 8). As with the DEM data errors, the rate of change of the error increases with increasing magnitude. The impacts from uncertainty levels of 0.25 SD and 0.5 SD are quite similar, but there is a large difference starting from the uncertainty level of 0.75 SD.

The proportion of pixels that change their initial class assignments is between 6.3% and 15% on average (Figure 2 and Table 8). The effect on the accuracy criterion is not linear, and the rate of change increases as the uncertainty level increases. The impacts from uncertainty levels of 0.25 SD and 0.5 SD are quite similar, but there is a large difference beginning at the 0.75 SD uncertainty level. In addition, the impacts of the five uncertainty levels of Landsat data on the number of pixels changed do not overlap (Figure 2).

The spatial distribution of the impact of random Landsat data error on the forest type mapping differs from that for the DEM and geology (Figure 5). For example,

Table 8. Summarised results of the sensitivity analysis of Landsat data.

	±0.25 SD	±0.50 SD	±0.75 SD	±1.00 SD	±1.25 SD
Difference from overall test accuracy					
Mean	−0.0069	−0.0085	−0.0141	−0.0232	−0.0337
SD	0.0078	0.0086	0.0102	0.0115	0.0129
Number of pixels changed					
Mean	17,441	19,412	25,616	33,201	41,165
SD	93	97	124	134	160

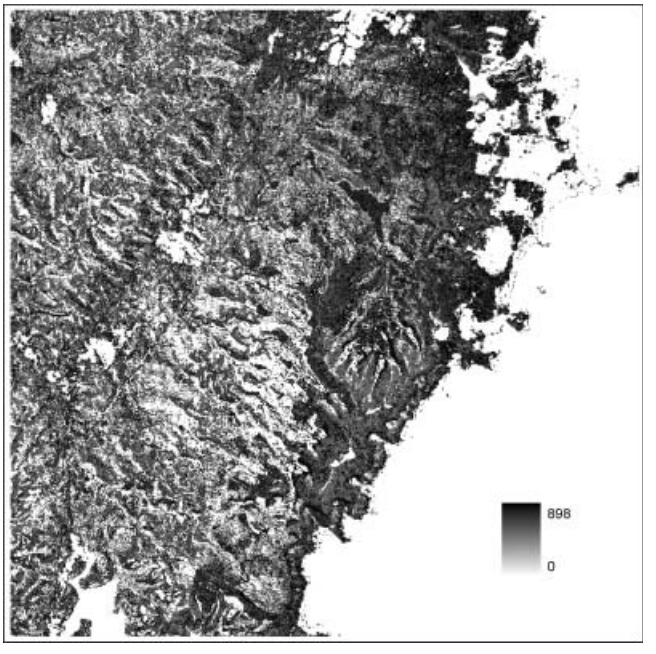

Figure 5. Spatial assessment of pixels changed with the Landsat data uncertainty level of mean zero and standard deviation of 0.75.

for the error level of 0.75 SD, 56.3% of all pixels changed at least once across all perturbations, 21.5% changed 100 or more times and 12.8% changed 200 or more times (Figure 5, Table 9). All seven forest types were noticeably or moderately affected by the 0.75 SD Landsat data uncertainty level (Table 9). Among them, lower slope wet forest was most affected. Cells assigned to this class in the base data set changed class assignments 282 times out of 1000 iterations on average, with 43% of cells changing at least 200 times. It is followed by *E. botryoides* forest. Dry sclerophyll forest is the forest type most tolerant to the Landsat data uncertainty level.

Table 9. Number of class changes for cells in each land cover class when using the Landsat data uncertainty level of mean zero and standard deviation of 0.75, summarised by class.

Class	Cells Count	Min	Max	Mean	SD	Cells changed once (%)	Cells changed >100 times (%)	Cells changed >200 times (%)
1	76,315	0	898	117	194	74.1	25.2	17.8
2	7375	0	820	214	203	95.5	60.3	30.5
3	6588	0	555	282	189	90.9	83.0	42.8
4	46,364	0	858	123	162	95.7	28.9	12.7
5	15,501	0	893	150	201	81.9	38.4	17.9
6	11,568	0	784	189	224	86.8	38.5	32.6
7	14,063	0	778	178	209	89.0	37.8	27.4
8	20,213	0	584	25	76	28.0	4.5	1.7
9	77,638	0	528	0	13	0.4	0.1	0.1
Overall	275,625					56.3	21.5	12.8

4. Discussion and conclusion

This study has demonstrated the feasibility of the sensitivity analysis of categorical model output in general and to land cover classification in particular, to the input data error. However, the nature of categorical model outputs means that many sensitivity analysis techniques cannot be used. In this study the one-at-a-time approach was used with a Monte Carlo simulation model for the sensitivity analysis of land cover classification. The Monte Carlo analysis is computationally demanding but effective for cases where formal mathematical models are not feasible. At least 500 iterations were required before the sensitivity analyses converged on a stable solution. This is less than that proposed by Heuvelink (1998), but is still larger than that used for previous research into the effects of error (e.g., Van Niel *et al.* 2004).

The detailed results of the three sensitivity analyses indicate that the impact patterns and extents of the three error sources to the C4.5 decision tree derived forest type mapping are quite different. These are now discussed.

The geology error model we applied had little effect, in terms of both overall test accuracy and the number of pixels changed. The rate of change between each increasing error level was also approximately linear. That the effect was limited to the geology boundary areas is to be expected, as we assumed only positional (boundary) type error. The number of pixels that actually occur within the spatial window used is a small proportion of those in the entire landscape, and so the relative number of pixels affected will always be less than that for the DEM and Landsat error models. Additionally, the rules involving geology only required categories of Pebbly Beach Permian, Wasp Head Permian and Ocean, albeit this was across 24 of the 47 rules (Table 2). Classification schemes involving more complex spatial arrangements of choropleths, and for which such boundary error is appropriate, will exhibit a greater effect. There are certainly attribute (thematic) error uncertainties in the geology data that cannot be represented as boundary error. If we had information to simulate such attribute error and uncertainty, we would expect to see a much greater effect on the forest type mapping.

Conversely, the decision tree derived forest type mapping was quite sensitive to DEM and Landsat data errors. This is despite the fact that decision tree classifiers have long been claimed to be more error tolerant than traditional statistical models in classification (Quinlan 1986). For example, with as small as a $\pm 2\%$ DEM error level, the overall test accuracy could be reduced by more than 2%. More importantly, the same uncertainty level has on average caused nearly 10% of the study area to change its initial class assignment at each perturbation. The forest type mapping was affected more profoundly with the increase of the DEM and Landsat data uncertainty level than with the geology error. This effect also has a non-linear rate of increase with respect to the error level.

While the impact of random DEM error and Landsat data error on the results of forest type mapping is not entirely surprising, we do now have an assessment of the extent to which this occurs. The percentages of pixels that change their class assignments due to random DEM error (e.g., 13.1% changed 100 or more times with the error level of $\pm 6\%$) or Landsat data error (e.g., 21.5% changed 100 or more times with the error level of 0.75 SD) are considerable when one considers that approximately 28% of the study area is ocean (with DEM values being 0 and band7 value less than 10). Because the classification rule for water (ocean plus lakes) uses pixels with Landsat band7 values less than 15 and DEM values less than 3, the assignment of DEM or Landsat data errors means that all ocean cells away from the

coast will have an error assigned to them that will not change their class assignment. In addition, a spatial assessment of the pixel changes due to either DEM error or Landsat data error indicates that their distribution is primarily within the forest classes (Table 5, Table 9, Figure 3, Figure 5), for which non-parametric classifiers like decision trees are most needed.

The pixels changed criterion and the accuracy criterion have different implications for critical uncertainty levels, albeit the accuracy criterion is not as reliable as the pixels changed criterion as it is based on only a limited number of test samples dispersed across the study area. When mapped, the assessment of changing class assignment can give us insights into which classes of and where the forest type mapping has been affected. In particular, *E. botryoides* forest and lower slope wet forest were most affected by the DEM and Landsat data error. This is to be expected if we look at the classification rules of the base classification (Tables 2 and 3). For example, the DEM was used 11 times in all 7 classification rules for *E. botryoides* forest. It was also used five times in all three classification rules of lower slope wet forest. For band7, it was used four times in all three classification rules of lower slope wet forest and six times across five out of seven classification rules of *E. botryoides* forest. This study thus confirms the findings of Huang and Lees (2004) that *E. botryoides* forest and lower slope wet forest are more difficult to classify for this study site. In comparison, geology was used in nine of the ten classification rules for the Dry *E. maculata* forest and five of the eight rules for rainforest. However, because these form large polygons and the effect of geology error was limited to the boundary areas, we could only see a minimal distribution disturbance for the two forest types due to geology.

We can derive the general conclusion that the decision tree derived forest type mapping is sensitive to topographic variables and remotely sensed data. The sensitivity of the forest type mapping to input data error increases with increasing error, but the rate of change is non-linear and that sensitivity differs considerably for the input variables.

The degree to which the same results will be obtained using other types of classifier remains to be assessed, for example, the Maximum Likelihood and Artificial Neural Network methods. These use very different operational principles (Gahegan 2000), and thus the error will propagate through the classification in different ways (Huang and Lees 2004, 2005). The effect of the errors on decision tree classifiers is restricted to the decision boundaries, whereas continuous or distance-based classifiers are expected to have a more continuous effect, modified by any final cut to a hard classification. It is therefore possible that the impact of the errors modelled here might be reduced. However, any such assessment can use the same general Monte Carlo error approach we have used here. This can be summarised as follows:

(1) Generate a benchmark classification model using the original input data.
(2) Select an input data set, identify its error source(s) and determine an appropriate error model.
(3) Randomly perturb the input data set to generate a perturbed input data set, calculating any necessary derivative data sets.
(4) Apply the benchmark model using the perturbed data set and other original input data to generate a perturbed classification, assessing the change in accuracy or other characteristics of the results (assessment criteria).
(5) Repeat steps 3 and 4 some number of times until the assessment criteria converge on a stable distribution (e.g., the mean and variance of the overall accuracy).

(6) Select the next input data source and repeat steps 2–5, until all required sensitivity analyses have been conducted.

While model sensitivity to errors is often used to determine which variables should be used, knowledge of the sensitivity of classifications to input data errors can also be treated as an opportunity. If the error distribution is known then it can be incorporated into subsequent analysis, most likely as a weighting factor such that the relative contribution of each input layer is adjusted appropriately. The pixels changed criterion means this can be done for each layer as either a constant value or spatially distributed, as appropriate. Such weighting approaches are routinely used for geographical analyses (e.g., Bickford and Laffan 2006). Additionally, ensemble approaches that combine the results of multiple models, for example, using Dempster-Schafer evidential theory (Huang and Lees 2005), could be more explicitly weighted to account for the relative sensitivity of the different models to the input data errors.

The overall implication of our analyses is that, even though numerous studies have demonstrated that multi-source data can be used for complex land cover classifications with good results, the impact of increasing error in the input data on the final accuracy is not necessarily proportional to their magnitudes. This is especially true for the continuous data, being the DEM-driven topographic data and remotely sensed data in this study. An obvious solution is to use more accurate ground elevation measures such as from Radar and LiDAR. However, until such data are available over large areas at appropriate resolutions, most analyses will use data such as those used here with their associated data errors.

Acknowledgements

This paper was improved by the contribution of two anonymous reviewers.

References

Ahlqvist, O., Keukelaar, J. and Oukbir, K., 2000, Rough classification and accuracy assessment. *International Journal of Geographical Information Science*, **14**, pp. 475–496.

Arbia, G., Griffith, D. and Haining, R., 1998, Error propagation modelling in raster GIS: overlay operations. *International Journal of Geographical Information Science*, **12**, pp. 145–167.

Aspinall, R., 1992, An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data. *International Journal of Geographical Information Science*, **6**, pp. 105–121.

Atkinson, P.M., 1999, Geographical information science: geostatistics and uncertainty. *Progress in Physical Geography*, **23**, pp. 134–142.

Bickford, S.A. and Laffan, S.W., 2006, Multi-extent analysis of the relationship between pteridophyte species richness and climate. *Global Ecology and Biogeography*, **15**, pp. 588–601.

Canters, F., De Genst, W. and Dufourmont, H., 2002, Assessing effects of input uncertainty in structural landscape classification. *International Journal of Geographical Information Science*, **16**, pp. 129–149.

Crosetto, M., Ruiz, J.A.M. and Crippa, B., 2001, Uncertainty propagation in models driven by remotely sensed data. *Remote Sensing of Environment*, **76**, pp. 373–385.

Crosetto, M. and Tarantola, S., 2001, Uncertainty and sensitivity analysis: tools for GIS-based model implementation. *International Journal of Geographical Information Science*, **15**, pp. 415–437.

Crosetto, M., Tarantola, S. and Saltelli, A., 2000, Sensitivity and uncertainty analysis in spatial modelling based on GIS. *Agriculture, Ecosystems, and Environment*, **81**, pp. 71–79.

Davis, T.J. and Keller, C.P., 1997, Modelling uncertainty in natural resource analysis using fuzzy sets and Monte Carlo simulation: slope stability prediction. *International Journal of Geographical Information Science*, **11**, pp. 409–434.

De Genst, W., Canters, F. and Gulinck, H., 2001, Uncertainty modeling in buffer operations applied to connectivity analysis. *Transactions in GIS*, **5**, pp. 305–326.

Drummond, J., 1987, A framework for handling error in geographic data manipulation. *ITC Journal*, **1**, pp. 73–82.

Finke, P.A., Wladis, D., Kros, J., Pebesma, E.J. and Reinds, G.J., 1999, Quantification and simulation of errors in categorical data for uncertainty analysis of soil acidification modeling. *Geoderma*, **93**, pp. 177–194.

Fisher, P., 2001, Alternative set theories for uncertainty in spatial information. In C.T. Hunsaker, M.F. Goodchild, M.A. Friedl and T.G. Case (Eds). *Spatial uncertainty in ecology*, pp. 351–362 (Berlin: Springer).

Fitzgerald, R.W. and Lees, B.G., 1996, Temporal context in floristic classification. *Computers and Geosciences*, **22**, pp. 981–994.

Foody, G.M., 2003, Uncertainty, knowledge discovery and data mining in GIS. *Progress in Physical Geography*, **27**, pp. 113–121.

Gahegan, M., 2000, On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, **32**, pp. 113–139.

Gahegan, M., German, G. and West, G., 1999, Improving neural network performance on the classification of complex geographic datasets. *Journal of Geographical Systems*, **1**, pp. 3–22.

Geman, S. and Geman, D., 1984, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, **6**, pp. 721–741.

Gertner, G., Wang, G., Fang, S. and Anderson, B., 2002, Effect and uncertainty of digital elevation model spatial resolutions on predicting the topographical factor for soil loss estimation. *Journal of Soil and Water Conservation*, **57**, pp. 164–174.

Gertner, G.Z., Fang, S., Wang, G. and Anderson, A., 2004, Partitioning spatial model uncertainty based on joint spatial simulation. *Transactions in GIS*, **8**, pp. 441–458.

Goodchild, M.F., 1989, Modeling error in objects and fields. In M.F. Goodchild and S. Gopal (Eds). *Accuracy of spatial databases*, pp. 107–113 (New York: Taylor & Francis).

Goodchild, M.F., Guoqing, S. and Shiren, Y., 1992, Development and test of an error model for categorical data. *International Journal of Geographical Information Science*, **6**, pp. 87–103.

Goodchild, M.F. and Hunter, G.J., 1997, A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, **11**, pp. 299–306.

Goovaerts, P., 1996, Stochastic simulation of categorical variables using a classification algorithm and simulated annealing. *Mathematical Geology*, **28**, pp. 909–921.

Goovaerts, P., 2001, Geostatistical modelling of uncertainty in soil science. *Geoderma*, **103**, pp. 3–26.

Gostin, V.A., 1969, Stratigraphy and sedimentology of the lower Permian sequence in the Durras-Ulladulla area Sydney Basin, *N.S.W.* Thesis (PhD). Australian National University.

Haining, R. and Arbia, G., 1993, Error propagation through map operations. *Technometrics*, **35**, pp. 293–305.

Hamby, D.M., 1994, A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, **32**, pp. 135–154.

Heuvelink, G.B.M., 1998, *Error propagation in environmental modelling with GIS* (London: Taylor & Francis).

HEUVELINK, G.B.M. and BURROUGH, P.A., 1993, Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Science*, **7**, pp. 231–246.

HEUVELINK, G.B.M., BURROUGH, P.A. and STEIN, A., 1989, Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Science*, **3**, pp. 303–322.

HINES, E.M., FRANKLIN, J. and STEPHENSON, J.R., 2005, Estimating the effects of map error on habitat delineation for the California Spotted Owl in Southern California. *Transactions in GIS*, **9**, pp. 541–559.

HUANG, Z. and LEES, B., 2005, Representing and reducing error in natural-resource classification using model combination. *International Journal of Geographical Information Science*, **19**, pp. 603–621.

HUANG, Z. and LEES, B.G., 2004, Combining non-parametric models for multisource predictive forest mapping. *Photogrammetric Engineering and Remote Sensing*, **70**, pp. 415–426.

HUANG, Z. and LEES, B.G., 2007, A classification accuracy measurement to deal with fuzziness of location and fuzziness of class. *Journal of Spatial Science*, **52**, pp. 1–12.

HWANG, D., KARIMI, H.A. and BYUN, D.W., 1998, Uncertainty analysis of environmental models within GIS environments. *Computers and Geosciences*, **24**, pp. 119–130.

JAGER, H.I. and KING, A.W., 2004, Spatial uncertainty and ecological models. *Ecosystems*, **7**, pp. 841–847.

LANTER, D.P. and VEREGIN, H., 1992, A research paradigm for propagating error in layer-based GIS. *Photogrammetric Engineering and Remote Sensing*, **58**, pp. 825–833.

LEE, J., SNYDER, P.K. and FISHER, P.F., 1992, Modeling the effect of data errors on feature extraction from digital elevation models. *Photogrammetric Engineering and Remote Sensing*, **58**, pp. 1461–1467.

LEES, B. and RITMAN, K., 1991, Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. *Environmental Management*, **15**, pp. 823–831.

LODWICK, W.A., 1989, Developing confidence limits on errors of suitability analyses in geographical information systems. In M.F. Goodchild and S. Gopal (Eds). *Accuracy of spatial databases* (New York: Taylor & Francis).

LUNETTA, R.S., CONGALTON, R.G., FENSTERMAKER, L.K., JENSEN, J.R., MCGWIRE, K.C. and TINNEY, L.R., 1991, Remote sensing and geographic information system data integration: error source and research issues. *Photogrammetric Engineering and Remote Sensing*, **57**, pp. 677–687.

MACDOUGALL, E.B., 1975, The accuracy of map overlays. *Landscape and Planning*, **2**, pp. 23–30.

MCKENNEY, D.W., 1999, Calibration and sensitivity analysis of a spatially-distributed solar radiation model. *International Journal of Geographical Information Science*, **13**, pp. 49–65.

MOORE, D., LEES, B. and DAVEY, S., 1991, A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *Environmental Management*, **15**, pp. 59–71.

NEWCOMER, J.A. and SZAJGIN, J., 1984, Accumulation of thematic map errors in digital overlay analysis. *American Cartographer*, **11**, pp. 58–62.

OPENSHAW, S., 1989, Learning to live with errors in spatial databases. In M.F. Goodchild and S. Gopal (Eds). *Accuracy of spatial databases*, pp. 263–276 (New York: Taylor & Francis).

OPENSHAW, S., CHARLTON, M. and CARVER, S., 1991, Error propagation: a Monte Carlo simulation. In I. Masser and M. Blakemore (Eds). *Handling geographical information*, pp. 78–101 (Harlow: Longman).

PERKAL, J., 1956, On epsilon length. *Bulletin de l'Academie Polonaise des Sciences*, **4**, pp. 399–403.

QUINLAN, J.R., 1986, Induction of decision trees. *Machine Learning*, **1**, pp. 81–106.

QUINLAN, J.R., 1993, *C4.5: programs for machine learning* (San Francisco, CA: Morgan Kauffman).

RICHARDS, J.A. and JIA, X., 2006, *Remote sensing digital image analysis*, 4th edn (Berlin: Springer).

SHI, W., 1998, A generic statistical approach for modelling error of geometric features in GIS. *International Journal of Geographical Information Science*, **12**, pp. 131–143.

TAYLOR, J.R., 1982, *An introduction to error analysis: the study of uncertainties in physical measurements* (Mill Valley, CA: University Science Books).

UNWIN, D.J., 1995, Geographical information systems and the problem of 'error and uncertainty'. *Progress in Human Geography*, **19**, pp. 549–558.

VAN NIEL, K. and LAFFAN, S.W., 2003, Gambling with randomness: the use of pseudo-random number generators in GIS. *International Journal of Geographical Information Science*, **17**, pp. 49–68.

VAN NIEL, K.P., LAFFAN, S.W. and LEES, B.G., 2004, Effect of error in the DEM on environmental variables for predictive vegetation modelling. *Journal of Vegetation Science*, **15**, pp. 747–756.

VEREGIN, H., 1989, Error modelling for the map overlay operation. In M.F. Goodchild and S. Gopal (Eds). *Accuracy of spatial databases*, pp. 3–18 (New York: Taylor & Francis).

VEREGIN, H., 1994, Integration of simulation modelling and error propagation for the buffer operation in GIS. *Photogrammetric Engineering and Remote Sensing*, **60**, pp. 427–435.

VEREGIN, H., 1995, Developing and testing of an error propagation model for GIS overlay operations. *International Journal of Geographical Information Science*, **9**, pp. 595–619.

VEREGIN, H., 1996, Error propagation through the buffer operation for probability surfaces. *Photogrammetric Engineering and Remote Sensing*, **62**, pp. 419–428.

WALSH, S.J., LIGHTFOOT, D.R. and BUTLER, D.R., 1987, Recognition and assessment of error in Geographic Information Systems. *Photogrammetric Engineering and Remote Sensing*, **53**, pp. 1423–1430.

WANG, G., et al. 2000, Spatial prediction and uncertainty analysis of topographic factors for the Revised Universal Soil Loss Equation (RUSLE). *Journal of Soil and Water Conservation*, **55**, pp. 374–384.