CSIRO Publishing

International *Journal* of Wildland Fire

Scientific Journal of IAWF

VOLUME 10, 2001

Address manuscripts and editorial enquiries to:
International Journal of Wildland Fire
Editor in Chief
Dr Gwynfor Richards
Department of Mathematics and Computer Science
Brandon University
Brandon, Manitoba, Canada R7A 6A9
Telephone: +1 204 727 7362
Fax: +1 204 728 7346
Email: richards@brandonu.ca

International Association of Wildland Fire

Address subscription enquiries to:
CSIRO PUBLISHING
PO Box 1139 (150 Oxford St)
Collingwood, Vic. 3066
Australia
Telephone: +61 3 9662 7644
Fax: +61 3 9662 7611
Email: ijwf@publish.csiro.au

Published by CSIRO Publishing
for the International Association of Wildland Fire

www.publish.csiro.au/journals/ijwf

International Journal of Wildland Fire, 2001, **10**, 321–328

Accuracy assessment and validation of remotely sensed and other spatial information

Russell G. Congalton

University of New Hampshire, Department of Natural Resources, Durham, NH 03824, USA.
Telephone: +1 603 862 4644; fax +1 603 862 4976; email: russ.congalton@unh.edu

This paper was presented at the conference 'Integrating spatial technologies and ecological principles for a new age in fire management', Boise, Idaho, USA, June 1999

Abstract. Today, validation or accuracy assessment is an integral component of most mapping projects incorporating remotely sensed data. Other spatial information may not be so stringently evaluated, but at least requires meta-data that documents how the information was generated. This emphasis on data quality was not always the case. In the 1970s only a few brave scientists and researchers dared ask the question, 'How good is this map derived from Landsat MSS imagery?' In the 1980s, the use of the error matrix became a common tool for representing the accuracy of individual map categories. By the 1990s, most maps derived from remotely sensed imagery were required to meet some minimum accuracy standard. A similar progression can be outlined for other spatial information. However, this progression is about 5 years behind the validation of remotely sensed data. This paper presents a series of steps moving towards better assessment and validation of spatial information and asks the reader to evaluate where they are in this series and to move forward.

Keywords: Accuracy assessment, validation, remotely sensed data, spatial information

Introduction

Accuracy assessment or validation is a key component of any project employing spatial data. There are a number of reasons why this assessment is so important including:

(1) The need to know how well you are doing and to learn from your mistakes;
(2) The ability to quantitatively compare methods; and
(3) The ability to use the information resulting from your spatial data analysis in some decision-making process.

There are many examples in the literature as well as an overwhelming selection of anecdotal evidence to demonstrate the need for accuracy assessment. Many different groups have mapped or quantified the amount of tropical deforestation in the Amazon Basin (e.g. Skole and Tucker 1993). Estimates have ranged by almost an order of magnitude. Which estimate is correct? Without a valid accuracy assessment we may never know. Several Federal, State, and local agencies have created maps of wetlands in a county on the Eastern Shore of Maryland. Techniques used to make these maps included satellite imagery, aerial photography (at various scales and film types), and ground sampling, all with varying classification schemes and wetlands definitions. Comparing the various maps yielded very little agreement about where wetlands actually existed. Without a valid accuracy assessment we may never know which of these maps to use.

It is no longer sufficient to have the final step in creating a map from remotely sensed or other spatial data simply be printing out the map. Instead, it is absolutely necessary to take some steps towards assessing the accuracy or validity of that map. There are a number of ways to investigate the accuracy/error in spatial data. These steps should be viewed as a progression and are as follows:

(1) Visual inspection;
(2) Non-site specific analysis;
(3) Difference image creation;
(4) Error budgeting; and
(5) Quantitative accuracy assessment

It is the goal of this paper to review these steps and help the reader to move along this progression of steps to increase the information gained from the spatial data analysis. In other words, to motivate everyone to do more accuracy assessment/validation.

10.1071/WF01031 1049-8001/01/030321

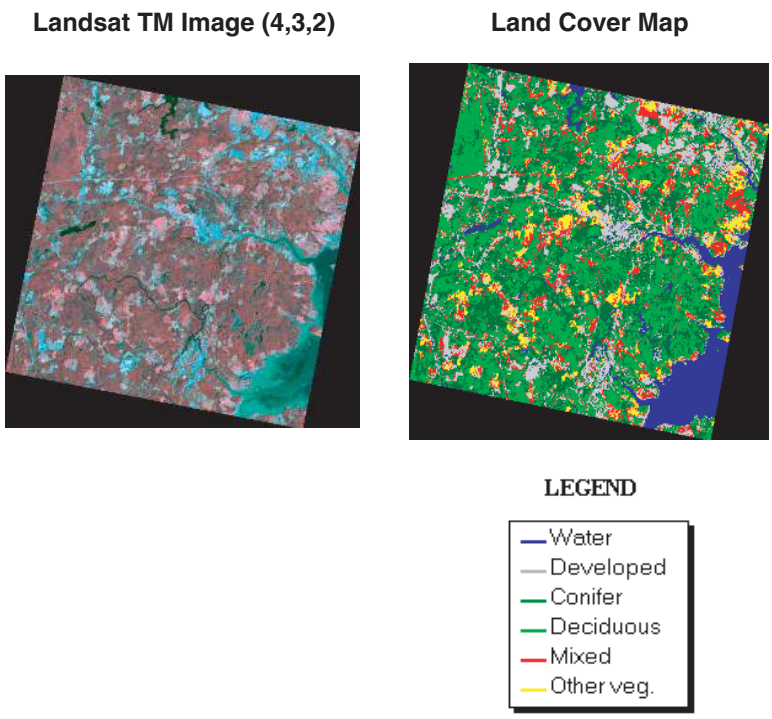

Fig. 1. Example of visual inspection of a map as the first step in accuracy assessment/validation.

Accuracy assessment/validation methods

Visual inspection

The visual inspection of your map derived from spatial data should be the first step in any assessment (Fig. 1). Visual inspection is a necessary first step, but it is not sufficient. In other words, it is very important to perform a visual assessment of your map and to be convinced that it looks right. After all, it would not make sense to further assess a map that does not even look right. However, it is not appropriate to conclude your assessment with a visual inspection. It is simply not sufficient. Many maps that 'looked good' were later found to have serious errors as a result of further accuracy assessment. In the example in Fig. 1, it is important that the water in the image be labeled water in the land cover map. It is also possible to compare other map classes to determine if the visual inspection makes sense. If the map fails the visual inspection then the analysis should be redone before any further accuracy assessment is undertaken.

Non-site-specific analysis

Non-site-specific analysis of a map derived from spatial data involves only the comparison of overall amounts of various areas without regard to any locational component. As shown in Table 1, the total area of forest created by two different image analysts can be compared to some reference amount. The reference amount could come from a variety of sources such as county agricultural statistics or US Forest Service Forest Inventory and Analysis (FIA) data. Reference data are assumed to be correct (Congalton and Green 1999). Whichever analyst's estimate is closer to the reference value will be deemed better. A quick study of Fig. 2 demonstrates why non-site specific analysis may not provide enough information for a valid map assessment. Note that, despite concluding through the non-site-specific analysis that the map created by image analyst #1 is better (error of only 113 ha), there is very little correspondence between where the reference data show the forests are located and where they are on the map. The map generated by image analyst #2 has much better spatial correspondence, although using only non-site-specific analysis one would conclude that this map is inferior.

Table 1. Results of a non-site-specific assessment

	Total area of forest (ha)
Analyst #1	2322
Analyst #2	2635
Analyst #3	2435
Assessment	
Difference #1 = 2435–2322 = 113 ha	
Difference #2 = 2435–2635 = 200 ha	

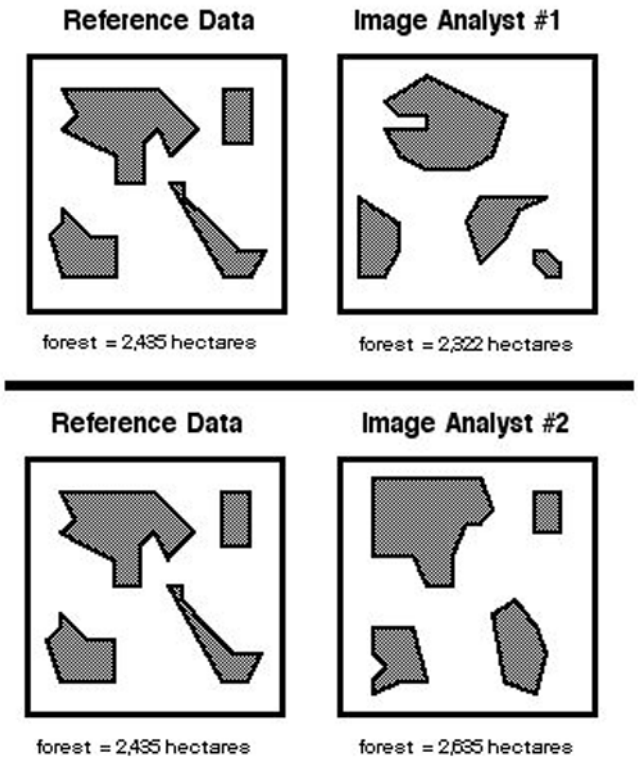

Fig. 2. Pictorial representation of the problems of a non-site specific assessment.

Difference image creation

The creation of a difference image is the first step at evaluating the spatial component of the map error. A difference image is a direct comparison of any two registered images/maps of the same area. It is produced by comparing the two images/maps, pixel by pixel, and representing areas of agreement in black and areas of disagreement in white (Fig. 3). If one of the maps is assumed to be correct, then the difference image represents the exact pattern of error in the map. If the maps are from different dates, then the difference image represents changes over time. If the maps are from different analysts using the same date of imagery, then the difference image represents differences in analyst methodology. In any case, the difference image presents a very graphic and easy to understand method for studying spatial patterns in maps and is a first step towards quantifying error.

Error budgeting

Over the last 20 years, many papers have been written about the quantification of error associated with remotely sensed and other spatial data (Congalton and Green 1999). Our ability to quantify the total error in a spatial data set has developed substantially. However, little has been done to partition this error into its component parts and construct an error budget. Without this division into parts, it is not possible to evaluate the impact of a certain error on the whole. Therefore, it is not possible to determine which components contribute the most error and which are most easily corrected. Some work was begun in this area in a paper by Lunetta *et al.* (1991) and resulted in a diagram like the one shown in Fig. 4. This figure shows the accumulation of error throughout a remote sensing project and is the beginning of understanding that the final map accuracy is a combination of many errors along the way. Each of the major error sources adds to the total error budget separately, and/or through a mixing process. For many applications there is a definite need to identify and understand (1) error sources, and (2) the appropriate mechanisms for controlling, reducing and/or reporting to the end users the magnitude of such errors.

Table 2 presents the results of performing an example error budget analysis for a GIS project. The table is generated column by column, beginning with a listing of the possible sources of error for the project. In order to make effective use of any GIS, it is important to understand the errors associated with the spatial information (Goodchild and Gopal 1989). This knowledge is critical whether you are a user of a GIS or whether you are one of the suppliers of spatial information (i.e. data layers) used in the GIS. Errors associated with spatial information can be divided into three groups as follows: (1) user errors; (2) measurement/data errors; and (3) processing errors (Burrough 1986). User errors are those errors which are probably most obvious and are more directly in the control of the user. Measurement/data error deals with the variability in the spatial information and the corresponding accuracy with which it was acquired. Finally, processing error involves errors inherent in the techniques used to input, access, and manipulate the spatial information.

Once the various components that comprise the total error are listed, then each component is assessed to determine its contribution to the overall error. Next, our ability to deal with this error is evaluated. It should be noted that some errors may be very large but are easy to correct while others may be rather small. In this example, an error index is created directly by multiplying the error contribution potential by the error control difficulty. Combining these two factors allows one to establish priorities in dealing with error.

Quantitative accuracy assessment

The key element of a quantitative accuracy assessment is the creation of an error matrix. An error matrix is a square array of numbers organized in rows and columns which express the number of sample units (i.e. pixels, clusters of pixels, or polygons) assigned to a particular category relative to the actual category as indicated by the reference data (Table 3). The columns usually represent the reference data while the rows indicate the classification generated from the remotely sensed data. Reference data are assumed correct and can be collected from a variety of sources including photo

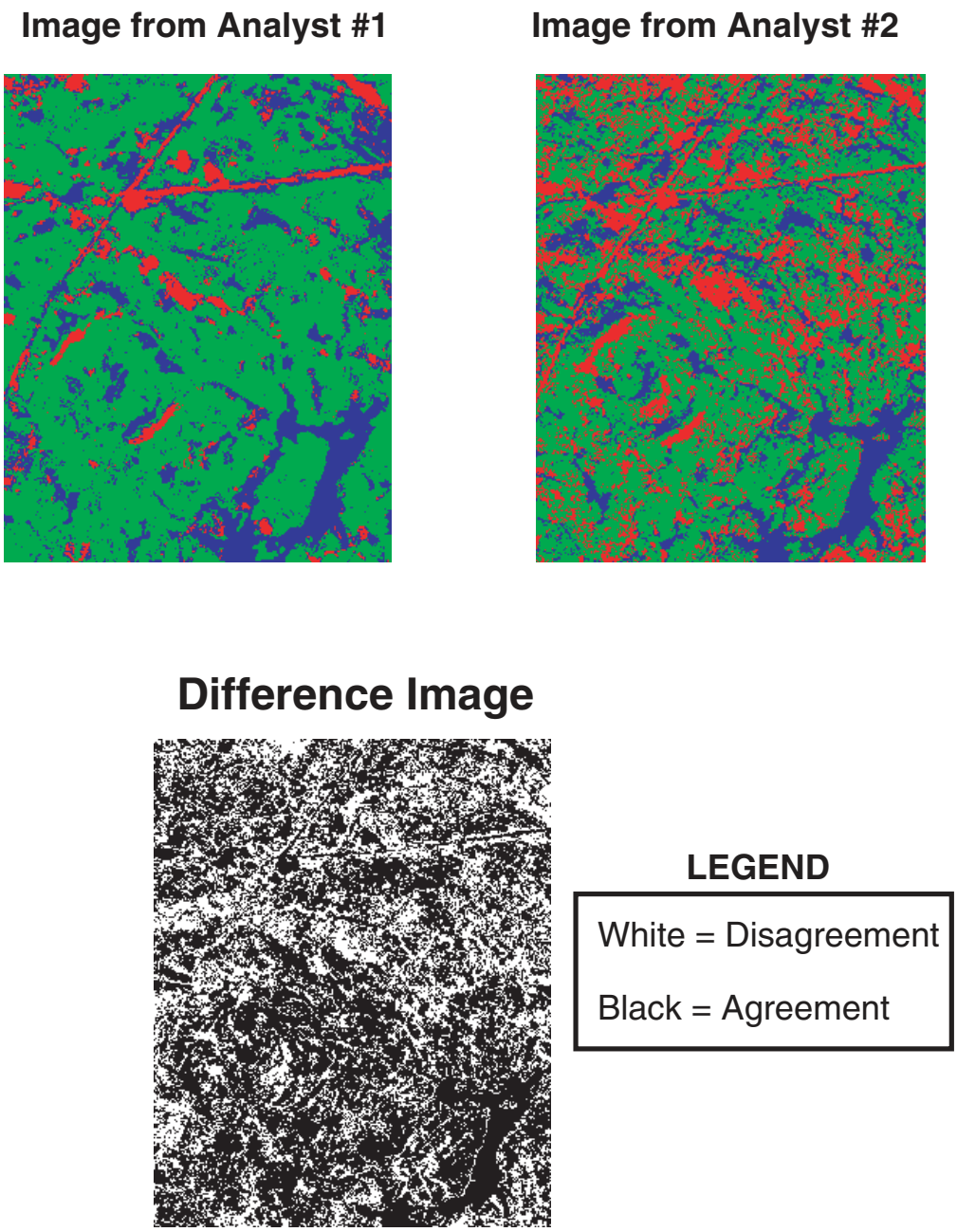

Fig. 3. Example of generating a difference image.

interpretation, videography, ground observation, and ground measurement. An error matrix is a very effective way to represent accuracy in that the accuracies of each category are plainly described along with both the errors of inclusion (commission errors) and errors of exclusion (omission errors) present in the classification.

The error matrix can then be used as a starting point for a series of descriptive and analytical statistical techniques. Perhaps the simplest descriptive statistic is overall accuracy, which is computed by dividing the total correct (i.e. the sum of the major diagonal) by the total number of samples in the error matrix. In addition, accuracies of individual categories can be computed in a similar manner. However, this case is a little more complex in that one has a choice of dividing the number of correct samples in that category by either the total number of samples in the corresponding row or the corresponding column. Traditionally, the total number of correct samples in a category is divided by the total number of samples of that category as derived from the reference data (i.e. the column total). This accuracy measure indicates

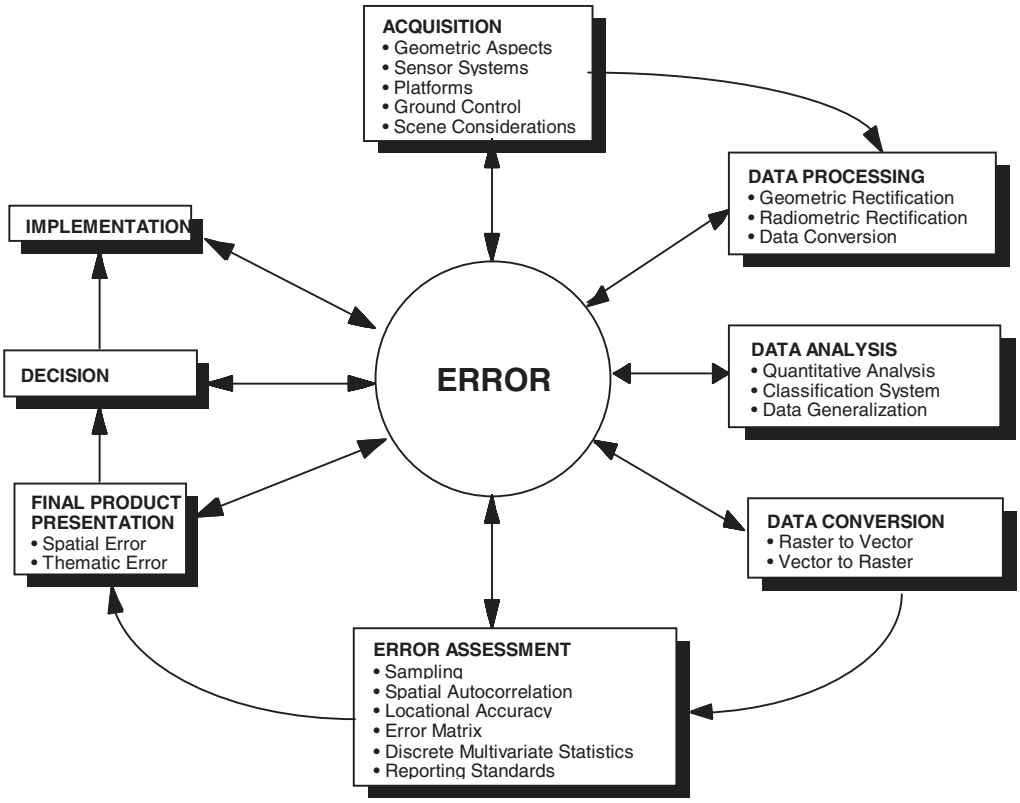

Fig. 4. Error source accumulation process in a remote sensing project.

the probability of a reference sample being correctly classified and is really a measure of omission error. This accuracy measure is often called 'producer's accuracy' because the producer of the classification is interested in how well a certain area can be classified. On the other hand, if the total number of correct samples in a category is divided by the total number of samples that were classified in that category, then this result is a measure of commission error. This measure, called 'user's accuracy' or reliability, is indicative of the probability that a sample classified on the map/image actually represents that category on the ground (Story and Congalton 1986).

In addition to these descriptive techniques, an error matrix is an appropriate beginning for many analytical statistical techniques. This is especially true of the discrete multivariate techniques. Starting with Congalton *et al.* (1983), discrete multivariate techniques have been used for performing statistical tests on the classification accuracy of digital remotely sensed data. Since that time many others have adopted these techniques as the standard accuracy assessment tools (e.g. Rosenfield and Fitzpatrick-Lins 1986; Campbell 1987; Hudson and Ramm 1987; Lillesand and Kiefer 1994). Discrete multivariate techniques are appropriate because remotely sensed data are discrete rather than continuous. The data are also binomially or multinomially distributed rather than normally distributed. Therefore, many common normal theory statistical techniques do not apply.

One analytical step to perform once the error matrix has been built is to 'normalize' or standardize the matrix using a technique known as 'MARGFIT' (Congalton *et al.* 1983). This technique uses an iterative proportional fitting procedure, which forces each row and column in the matrix to sum to 1. The rows and column totals are called marginals, hence the technique name MARGFIT. In this way, differences in sample sizes used to generate the matrices are eliminated and therefore individual cell values within the matrix are directly comparable. Also, because the iterative process totals the rows and columns, the resulting normalized matrix is more indicative of the off-diagonal cell values (i.e. the errors of omission and commission) than is the original matrix. The major diagonal of the normalized matrix can be summed and divided by the total of the entire matrix to compute a normalized overall accuracy.

Another discrete multivariate technique of use in accuracy assessment is called KAPPA (Cohen 1960). The result of performing a KAPPA analysis is a KHAT statistic (an estimate of KAPPA), which is another measure of agreement or accuracy. The values can range from +1 to –1. However, since there should be a positive correlation

Table 2. Matrix showing GIS error sources and priority for dealing with them

Error contribution potential: relative potential for this source as contributing factor to the total error (1 = low; 2 = medium; 3 = high). Error control difficulty: given the current knowledge about this source, how difficult is controlling the error contribution (1 = not very difficult; 5 = very difficult); Error index: an index that represents the combination of error potential and error difficulty; Error priority: order in which method should be implemented to understand, control, reduce, and/or report the error due to this source based on the error index

Error source	Error contribution potential	Error control difficulty	Error index	Error priority
		User		
Age	2	3	6	6
Scale	3	4	12	11
Coverage/extent	1	2	2	1
Indirect/derived layer	2	4	8	7
		Measurement/data		
Instrument error	2	4	8	7
Field error	1	3	3	4
Natural variation	1	5	5	5
		Processing		
Precision	1	2	2	1
Interpolation	2	4	8	7
Generalization	3	3	9	10
Conversion	3	4	12	11
Digitization	1	2	2	1

between the remotely sensed classification and the reference data, positive KHAT values are expected. Landis and Koch (1977) characterized the possible ranges for KHAT into three groupings: a value greater than 0.80 (i.e. 80%) represents strong agreement; a value between 0.40 and 0.80 (i.e. 40–80%) represents moderate agreement; and a value below 0.40 (i.e. 40%) represents poor agreement.

The power of the KAPPA analysis is that it provides two statistical tests of significance. Using this technique, it is possible to test if an individual land cover map generated from remotely sensed data is significantly better than if the map had been generated by randomly assigning labels to areas. The second test allows for the comparison of any two matrices to see if they are statistically significantly different. In this way, it is possible to determine that one algorithm is different than another one and, based on a chosen accuracy measure (e.g. overall accuracy), to conclude which is better.

The above descriptive and analytical techniques are based on the error matrix. An assumption made here is that the matrix was properly generated and is therefore indicative of the map it represents. If the matrix was not properly created then it is useless, or at best anecdotal evidence. Certain statistical considerations are required in order to assure that this assumption is valid. Developing a statistically rigorous accuracy assessment requires choosing an appropriate sampling scheme, sample size, sampling unit, maintaining independence between the training and reference data, and considering the effects of spatial autocorrelation (Congalton 1991).

Table 3. Example error matrix

Classified Data \ Reference Data	V	W	U	row total
V	43	10	6	59
W	3	23	5	31
U	2	1	30	33
column total	48	34	41	123

Land Cover Categories

V = Vegetation

W = Water

U = Urban

OVERALL ACCURACY = 96/123 = 78%

PRODUCER'S ACCURACY		USER'S ACCURACY	
V = 43/48 =	90%	V = 43/59 =	73%
W = 23/34 =	68%	W = 23/31 =	74%
U = 30/41 =	73%	U = 30/33 =	91%

Sampling scheme

There are numerous possible sampling schemes used in collecting accuracy assessment data including: simple random sampling, systematic sampling, stratified random sampling, cluster sampling, and stratified systematic

unaligned sampling. Each scheme has its own advantages and disadvantages. Randomness provides very nice statistical properties that are important for further analysis of the results. Systematic and cluster sampling can provide practical advantages. It is important to understand each scheme and apply the one most appropriate for the situation. The analysis undertaken must then match the sampling scheme chosen. In most cases, stratified random sampling is most appropriate.

Independence

It is critical that the data collected for training in the classification process be independent (separate) from the data used in the accuracy assessment. After the ground reference data are collected, a stratified random sample of the data should be selected for accuracy assessment and put aside and not looked at until after the map has been generated. The remaining data can then be used for training. It is important to stratify the data by map class to insure that sufficient training and accuracy samples exist for each map class.

Sample size

Sample size is dictated by the need to express accuracy in an error matrix. The sample size must be large enough to provide that the error matrix estimates have adequate precision. An error matrix does not fall into the right/wrong binomial scenario but rather a multinomial situation in which there is one correct for each class and n–1 wrongs (where n is the number of map classes). Therefore, experience as well as the multinomial equation show that approximately 50 samples (30 as an absolute minimum) per map class are required to adequately populate an error matrix (Story and Congalton 1986).

Sampling unit

There are three common sampling units proposed in assessing the accuracy of remotely sensed data. They are (1) the pixel; (2) a 3×3 grouping of pixels; or (3) a polygon. It should be noted that the pixel should not be used as the sampling unit because of our inability to accurately locate it on the ground (even using GPS) and on the imagery. Either a grouping of pixels, such as a 3×3 block or a polygon, should be selected as the sample unit depending on the specific needs of the project.

Spatial autocorrelation

Spatial autocorrelation is a measure of the influence, positive or negative, that some characteristic at a certain location has on its surrounding neighbors. Spatial autocorrelation is an important consideration when deciding which sampling scheme to employ. If there is positive correlation between samples, then it is important for precision of the accuracy estimates to space the samples far enough apart to minimize this correlation. This issue is particularly important for certain schemes such as cluster sampling and systematic sampling.

Conclusions

This paper presents a variety of techniques that can be used to assess or validate maps derived from remotely sensed and other spatial data. Although it is important to perform a visual examination of the map, it is not sufficient. Other techniques, such as non-site-specific analysis and difference images, can help. Error budgeting is a very useful exercise in helping to realize error and consider ways to minimize it. Quantitative accuracy assessment provides a very powerful mechanism for both descriptive and analytical evaluation of the spatial data. As our use of spatial data continues to grow, so must our use of these tools for evaluation. If you are a novice spatial data user, please consider the techniques proposed here and implement as many as you can. If you are an advanced spatial data user, there is no excuse for not employing these techniques to better evaluate your analysis. Let us not stay stuck in a mode of 'it looks good,' but rather let us struggle forward to advance the use of spatial data in all aspects of our work. To obtain more details on any of the procedures and techniques described in this paper, see Congalton and Green (1999); Stehman and Czaplewski (1998); Janssen and van der Well (1994); and Congalton (1991).

Acknowledgements

The author acknowledges that funding for this work came from the GLOBE Program and the University of New Hampshire McIntire-Stennis Project MS33. Thanks to Ms Lucie Plourde for her review of this manuscript.

References

Burrough PA (1986) 'Principles of Geographical Information Systems for land resources assessment.' (Oxford University Press: New York) 193 pp.

Campbell J (1987) 'Introduction to remote sensing.' (Guilford Press: New York) 551 pp.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46.

Congalton RG, Oderwald RG, Mead RA (1983) Assessing Landsat classification accuracy using discrete multivariate statistical techniques. *Photogrammetric Engineering and Remote Sensing* **49**, 1671–1678.

Congalton R (1991) A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* **37**, 35–46.

Congalton R, Green K (1999) 'Assessing the accuracy of remotely sensed data: Principles and practices.' (CRC/Lewis Press: Boca Raton) 137 pp.

Goodchild M, Gopal S (Eds) (1989) 'The accuracy of spatial databases.' (Taylor and Francis: New York) 290 pp.

Hudson W, Ramm C (1987) Correct formulation of the kappa coefficient of agreement. *Photogrammetric Engineering and Remote Sensing*. **53**, 421–422.

Janssen L, van der Well F (1994) Accuracy assessment of satellite derived land-cover data: A review. *Photogrammetric Engineering and Remote Sensing* **60**, 419–426.

Landis J, Koch G (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Lillesand T, Kiefer R (1994) 'Remote sensing and image interpretation.' 3rd edn. (John Wiley and Sons: New York) 750 pp.

Lunetta R, Congalton R, Fenstermaker L, Jensen J, McGwire K, Tinney L (1991) Remote sensing and geographic information system data integration: error sources and research issues. *Photogrammetric Engineering and Remote Sensing*. **57**, 677–687.

Rosenfield G, Fitzpatrick-Lins K (1986) A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing*. **52**, 223–227.

Skole D, Tucker C (1993) Tropical deforestation, fragmented habitat, and adversely affected habitat in the Brazilian Amazon: 1978–1988. *Science* **260**, 1905–1910.

Stehman S, Czaplewski R (1998) Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment* **64**, 331–344.

Story M, Congalton R (1986) Accuracy assessment: A user's perspective. *Photogrammetric Engineering and Remote Sensing* **52**, 397–399.