A Practical Look at the Sources of Confusion in Error Matrix Generation

Abstract

The need for assessing the accuracy of spatial data derived from remote sensing techniques and used in geographic information system (GIS) analyses has been recognized as a critical component of many projects. The results of this assessment are typically represented by an error matrix presenting the overall accuracy, the accuracies of each category, and the omission and commission errors. This paper presents a discussion of the various sources of confusion that are represented within an error matrix.

Ideally, the matrix is representative of only classification error. However, other factors can be significant and substantially lower the accuracy as represented by the matrix. An example is presented from mapping old growth forests on the Rogue River National Forest in Oregon. A detailed accuracy assessment was performed and the sources of confusion between the remotely sensed map classification and the reference data set were explored. The results demonstrated that these other non-classification differences can be substantial sources of error. Finally, methods for controlling these differences are discussed.

Introduction

The need for assessing the accuracy of remotely sensed and other spatial data is a continuing one and the issue is receiving more interest every year. The use of remotely sensed data in geographic information systems (GIS) has made such assessments even more important. This has not always been the case. Historically, accuracy assessment was an afterthought and often performed in a haphazard manner. Many times the assessment was only qualitative in nature and achieved by visiting a few areas and determining that the classification "looked good."

Early on some researchers, notably Hord and Brooner (1976), van Genderen and Lock (1977), and Ginevan (1979), proposed some criteria and techniques for testing map accuracy. In the early 1980s more in-depth studies were conducted and new techniques were proposed (Rosenfield *et al.*, 1982; Congalton *et al.*, 1983; Aronoff, 1985). As a result of this work and work by others, the error matrix (Story and Congalton, 1986) became the standard medium for reporting remotely sensed data classification accuracies. In addition, the use of the Kappa statistic (Cohen, 1960) was recommended by many researchers as an acceptable measure of accuracy. Work continued in this area throughout the rest of the 1980s and into the 90s by incorporating other factors influencing the accuracy of spatial data, including sampling scheme and sample size, classification scheme, and spatial autocorrelation. A detailed review of the techniques and considerations for accuracy assessment can be found in Congalton (1991).

As techniques became more established and accuracy assessments were required as a critical component of any mapping project, other important considerations became evident. These issues include ground verification techniques, incorporating assessments into regional and global projects, and evaluating all the sources of error in the spatial data, not just classification accuracy. Lunetta *et al.* (1991) looked at the effect the error associated with remote sensing and GIS data acquisition, processing, analysis, conversion, and final presentation can have on the decision making process. Smith *et al.* (1991) explored the use of different algorithms and their effects on the same data set and Fenstermaker (1991) discussed the issues of global assessments in terms of the EPA nationwide Environmental Monitoring and Assessment Program (EMAP).

As part of an effort to identify old growth forests in the Pacific Northwest from remotely sensed and other spatial data (Congalton *et al.*, 1993), an extensive accuracy assessment was conducted. In this study, classifications derived from remotely sensed data were used as inputs into a GIS along with other layers of spatial information such as slope, aspect, elevation, hydrology, and other factors used to determine areas of old growth forest. Old growth forests were defined by a variety of characteristics including tree canopy structure (i.e., multistoried), tree diameter, crown closure, and size of the stand.

Instead of using a traditional remote sensing approach and identifying old growth areas on the satellite data and building old growth training areas, the factors mentioned above were identified in a GIS and used to determine where old growth forests exist. In other words, the remotely sensed data were used to identify tree species, size class (diameter), crown closure, and structure instead of just identifying a category as old growth. This way, if the definition of old growth was modified, the rules for selecting old growth in the GIS database could be changed and a new map based on the revised definition of old growth could be produced. For example, if the size class definition of big old growth were changed from 32 inches in diameter to 36 inches, a map of old growth using this new definition could quickly be generated from the GIS. If the traditional approach was used, the entire image classification process would need to be repeated and old growth areas meeting the new definition identified.

Much was learned about the practical and technical aspects of accuracy assessment during this project. Specifically, observations were made that directly affect generation of the error matrix, a critical component for representing the accuracy determined by any assessment. It is the knowledge

Photogrammetric Engineering & Remote Sensing,
Vol. 59, No. 5, May 1993, pp. 641–644.

0099-1112/93/5905-641$03.00/0

Russell G. Congalton
Department of Natural Resources, University of New Hampshire, Durham, NH 03824.
Kass Green
Pacific Meridian Resources, 5915 Hollis St., Building B, Emeryville, CA 94608

gained from this extensive effort that we would like to report on in this paper. Hence, the objectives addressed in this paper were:

- To identify and enumerate sources of confusion between the remotely sensed classification and the reference data used to assess it.
- To estimate the impact of non-error differences on the error matrix. The term non-error differences is used here to differentiate classification error from the other non-error differences that can occur in the error matrix.
- To suggest methods of controlling these non-error differences.

Methods

The study area used to perform an in-depth analysis of the components of error matrix generation was the Rogue River National Forest in southern Oregon (Figure 1). This forest is part of the Douglas-fir old growth forests of the Pacific Northwest. The categories of interest in this study were big old growth, little old growth, and other. Big old growth was defined as a coniferous forested area with more than 70 percent total crown closure and greater than 10 percent crown closure in trees 32 inches in diameter at breast height (DBH) or more. Little old growth was defined as a coniferous forested area with more than 70 percent total crown closure and greater than 10 percent crown closure in trees 21 to 31 inches in diameter at breast height (DBH). The other category encompassed everything else but old growth.

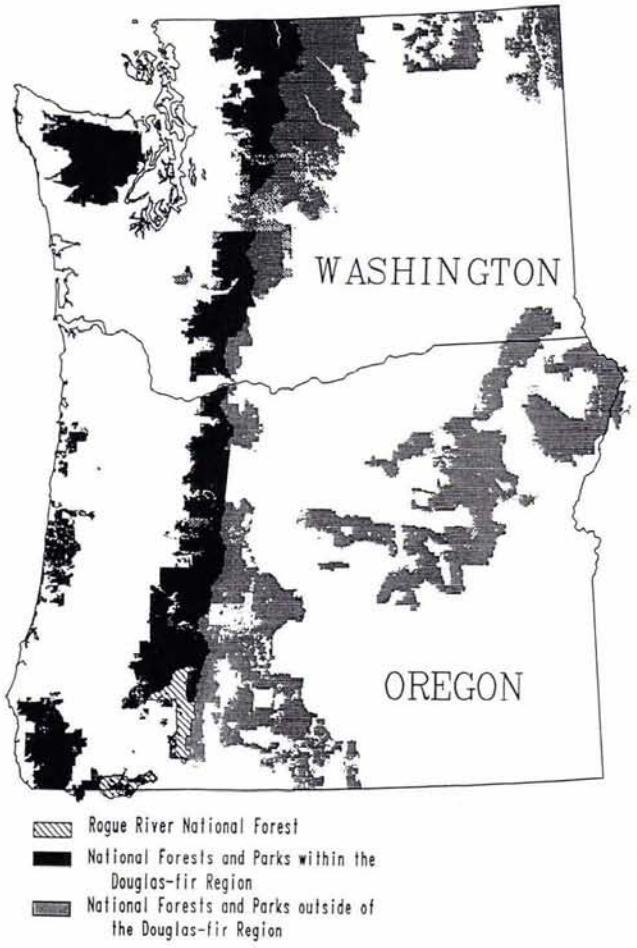

Figure 1. The Rogue River National Forest study area is part of the Douglas-fir Old Growth region of the Pacific Northwest.

Error Matrix Generation

An error matrix is a square array of numbers set out in rows and columns which expresses the number of sample units (i.e., pixels, clusters of pixels, or polygons) assigned to a particular category relative to the actual category as verified by some reference data. The columns usually represent the reference data while the rows indicate the classification generated from the remotely sensed data. In other words, an error matrix is a comparison between sampled areas on the map generated from the remotely sensed data and those same areas as determined by some reference data. The reference data are typically ground visits or large scale (1:12,000 or larger) color aerial photography. The object then of the error matrix is to represent the accuracy of the remotely sensed classification (i.e., the errors in the map). An assumption made here is that all differences between the remotely sensed map classification and the reference data are due to classification and/or delineation error. However, there are many other sources of confusion between the remotely sensed classification and the reference data that must also be considered. They include

- Registration differences between the reference data and the remotely sensed map classification;
- Delineation error encountered when the sites chosen for accuracy assessment are digitized;
- Data entry error when the reference data is entered into the accuracy assessment database;
- Error in interpretation and delineation of the reference data (e.g., photointerpretation error);
- Changes in land cover between the date of the remotely sensed data and the date of the reference data (temporal error), for example, changes due to fires or urban development or harvesting;
- Variation in classification and delineation of the reference data due to inconsistencies in human interpretation of heterogeneous vegetation;
- Errors in the remotely sensed map classification; and
- Errors in the remotely sensed map delineation.

Each of these possible sources of confusion were evaluated during the error matrix generation for the Rogue River National Forest. Methods were devised for controlling the differences represented by the first six factors in the above list (i.e., the non-error differences). It is these six factors that can significantly lower the accuracy in an error matrix and make the classification (i.e, factors 7 and 8) look far worse than it actually is. Error matrices were generated that compared the remotely sensed map classification only to ground visited reference data and to a combination of ground visited plus photo interpreted reference data. Finally, an error matrix was generated that compared the ground visited reference data to the photo interpreted reference data.

Results

Table 1 presents the error matrix for the remotely sensed map classification as compared to the photo interpreted and ground visited reference data combined. Table 2 shows the error matrix where only the ground visited reference data was used and Table 3 is a comparison between the ground

visited and photo interpreted reference data sets. These matrices are provided for comparison purposes. Clearly, it is more expensive and time consuming to ground visit reference data sites rather than to photo interpret them. Minimizing ground visits is therefore very desirable.

However, it is important to check that the photointerpretation is correct (i.e., that it agrees with what a ground visit would indicate). Table 3 presents this comparison between assessment sites that were both photo interpreted and ground visited. Given the desire to collect as few ground visits as possible, this matrix is going to have fewer samples. This limits the statistical inferences that can be made from the matrix, but it can still be indicative of whether problems (confusion) exist. Taking sufficient samples to provide a statistically valid matrix would defeat the purpose of substituting the photo interpreted sites for the ground visits.

The detailed analysis of the sources of confusion for the Rogue River National Forest accuracy assessment revealed the following:

- No registration differences;
- No delineation errors;
- No data entry errors;
- Assuming the ground reference data to be the truth, eight of the 40 photo sites that were visited on the ground were incorrectly photo interpreted;
- Nine of the accuracy assessment sites had been harvested after the date of the photography used for reference data but before the data of the remotely sensed data acquisition; and
- Of the 27 sites that were photo interpreted by two different interpreters, 11 of them or 41 percent were given a different class by each interpreter.

As a result of this detailed analysis, methods for controlling these differences between the remotely sensed map classification and the reference data were devised. These actions included

- Registration differences were controlled by transferring the reference data to the remotely sensed map classification before generating the matrix. This procedure allowed the analyst to check that both data sets coincided by visual inspection.

TABLE 1. ERROR MATRIX COMPARING THE REMOTELY SENSED MAP CLASSIFICATION TO THE COMBINATION OF BOTH PHOTOINTERPRETED AND GROUND VISITED REFERENCE DATA SITES.

Combination of photo interpreted and ground visited reference data

Remotely Sensed Map Classification	BOG	LOG	O	row total
BOG	40	16	15	71
LOG	7	5	1	13
O	9	10	92	111
column total	56	31	108	195

Land Cover Categories
BOG = big old growth
LOG = little old growth
O = other

OVERALL ACCURACY = 137/195 = 70%

PRODUCER'S ACCURACY
BOG = 40/56 = 71%
LOG = 5/31 = 16%
O = 92/108 = 85%

USER'S ACCURACY
BOG = 40/71 = 56%
LOG = 5/13 = 38%
O = 92/111 = 83%

TABLE 2 . ERROR MATRIX COMPARING THE REMOTELY SENSED MAP CLASSIFICATION TO JUST THE GROUND VISITED REFERENCE DATA SITES.

Ground visited Reference Data

Remotely Sensed Map Classification	BOG	LOG	O	row total
BOG	8	1	5	14
LOG	1	1	0	2
O	1	0	27	28
column total	10	2	32	44

Land Cover Categories
BOG = big old growth
LOG = little old growth
O = other

OVERALL ACCURACY = 36/44 = 82%

PRODUCER'S ACCURACY
BOG = 8/10 = 80%
LOG = 1/2 = 50%
O = 27/32 = 84%

USER'S ACCURACY
BOG = 8/14 = 57%
LOG = 1/2 = 50%
O = 27/28 = 96%

- Delineation in the digitizing process was controlled by digitizing all data twice.
- Data entry differences were handled by implementing a strict set of procedures in which all personnel were well trained and a series of checks instituted to ensure consistency (i.e., quality control).
- In order to minimize photo interpretation differences, ground visits were substituted for interpretation wherever practical. In places where interpretation was used, a subsample of results for both the photointerpretation and the ground visits were compared as in Table 3.
- Temporal problems can be quite significant and were controlled by requiring the reference data be as close in date to the remotely sensed map classification as possible. An evaluation should be made of the temporal problems in each mapping project. If found to be significant, then newer, more appropriate reference data must be obtained. In the Rogue River National Forest example, temporal problems were caused by timber harvesting between the time the photographs used as reference data were obtained and the time of the remotely sensed data acquisition. Fortunately, cut areas can be easily detected on the raw remotely sensed data and so these temporal problems can be easily detected. Other

TABLE 3 . ERROR MATRIX COMPARING THE GROUND VISITED REFERENCE DATA SITES TO THE PHOTO INTERPRETED REFERENCE DATA SITES.

Ground Visited Reference Data

Photo intepreted Reference Data	BOG	LOG	O	row total
BOG	5	1	0	6
LOG	2	1	3	6
O	2	0	26	28
column total	9	2	29	40

Land Cover Categories
BOG = big old growth
LOG = little old growth
O = other

OVERALL ACCURACY = 32/40 = 80%

PRODUCER'S ACCURACY
BOG = 5/9 = 56%
LOG = 1/2 = 50%
O = 26/29 = 90%

USER'S ACCURACY
BOG = 5/6 = 83%
LOG = 1/6 = 17%
O = 26/28 = 93%

problems are not so easy to control. For example, the change in the size class (i.e., diameter of the trees) can change between the time of the photo and remotely sensed data acquisitions, especially in a fast growing area such as the Olympic Peninsula in Washington.

- Inconsistencies in human interpretation, especially for heterogeneous areas, can be a very difficult factor to control. Measures of variation in interpretation need to be further developed that can test the validity of class boundaries while at the same time provide for allowable variances in the accuracy assessment.

Conclusions

There are many sources of confusion between the remotely sensed map classification and the reference data as represented by an error matrix. The non-classification error differences can significantly lower the accuracy as determined from the matrix. Therefore, the error matrix is really a "difference" matrix and is only representative of error if these other differences (sources of confusion) have been accounted for. It is critical that all these differences be considered in any accuracy assessment.

The importance of assessing the accuracy of remotely sensed data classifications is evident. However, assessing other spatial data layers are equally important. Decisions made as a result of the data in a GIS are only as good as the input data. Many of the techniques and problems described in this paper apply to other spatial data besides just remotely sensed data. As more and more emphasis is placed on computer databases and GIS, it is increasingly important to consider techniques for evaluating the accuracy of these layers.

Acknowledgments

This old growth project has been a monumental team effort involving numerous subcontractors, Forest Service personnel, and, at one time or another, every one of Pacific Meridian's employees. We would like to acknowledge each of these individuals for their essential involvement and contribution.

References

Aronoff, Stan, 1985. The minimum accuracy value as an index of classification accuracy. *Photogrammetric Engineering & Remote Sensing*, Vol. 51, No. 1 pp. 99-111.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37-46.

Congalton, R., 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, Vol. 37, pp. 35-46.

Congalton, R., K. Green, and J. Teply, 1993. Mapping old growth forests on National Forest and Park lands in the Pacific Northwest from remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, Vol. 59, No. 4, pp. 529-535.

Congalton, R. G., R. G. Oderwald, and R. A. Mead, 1983. Assessing Landsat classification accuracy using discrete multivariate statistical techniques. *Photogrammetric Engineering & Remote Sensing*, Vol. 49, No. 12, pp. 1671-1678.

Fenstermaker, L., 1991. A proposed approach for national to global scale error assessment. *Proceedings of GIS/LIS 91*, Atlanta, Georgia, October, pp. 293-300.

Ginevan, M. E., 1979. Testing land-use map accuracy: another look. *Photogrammetric Engineering & Remote Sensing*, Vol. 45, No. 10, pp. 1371-1377.

Hord, R. M., and W. Brooner, 1976. Land use map accuracy criteria. *Photogrammetric Engineering & Remote Sensing*, Vol. 42, No. 5, pp. 671-677.

Lunetta, R., R. Congalton, L. Fenstermaker, J. Jensen, K. McGwire, and L. Tinney, 1991. Remote sensing and geographic information system data integration: error sources and research issues. *Photogrammetric Engineering & Remote Sensing*, Vol. 57, No. 6, pp. 677-687.

Rosenfield, G. H., K. Fitzpatrick-Lins, and H. Ling, 1982. Sampling for thematic map accuracy testing. *Photogrammetric Engineering & Remote Sensing*, Vol. 48, No. 1, pp. 131-137.

Smith, J., S. Prisley, and R. Weih, 1991. Considering the effect of spatial variability on the outcomes of forest management decisions. *Proceedings of GIS/LIS 91*, Atlanta, Georgia, October, pp. 286-292.

Story, M., and R. Congalton, 1986. Accuracy assessment: A user's perspective. *Photogrammetric Engineering & Remote Sensing*, Vol. 52, No. 3, pp. 397-399.

Van Genderen, J. L., and B. F. Lock, 1977. Testing land use map accuracy, *Photogrammetric Engineering & Remote Sensing*, Vol. 43, No. 9, pp. 1135-1137.

(Received 10 July 1992; accepted 2 September 1992; revised 30 September 1992)

LIST OF "LOST" CERTIFIED PHOTOGRAMMETRISTS

We no longer have valid addresses for the following Certified Photogrammetrists. If you know the whereabouts of any of the persons on this list, please contact ASPRS headquarters so we can update their records and keep them informed of all the changes in the Certification Program. Thank you.

Jack R. Anthony
Dewayne Blackburn
Albert Brown
Eugene Caudell
Robert Denny
Leo Ferran
Robert Fuoco
Franek Gajdeczka
George Glaser
William Grehn, Jr.
Louis T. Harrod
Elwood Haynes
F.A. Hildebrand, Jr.
James Hogan
William Janssen
Lawrence Johnson
Spero Kapelas
Andre J. Langevin
Harry J. Miller
Marinus Moojen
Gene A. Pearl
Sherman Rosen
Lane Schultz
Keith Syrett
William Thomasset
Conrad Toledo
Robert Tracy
Lawrence Watson
Tad Wojenka