Thematic and Positional Accuracy Assessment of Digital Remotely Sensed Data

Russell G. Congalton[1]

Abstract.—Accuracy assessment or validation has become a standard component of any land cover or vegetation map derived from remotely sensed data. Knowing the accuracy of the map is vital to any decisionmaking performed using that map. The process of assessing the map accuracy is time consuming and expensive. It is very important that the procedure be well thought out and carefully planned to be as efficient as possible. This paper presents a brief review of the current methods used in thematic map accuracy assessment. A discussion of positional error is included as it is impossible to assess thematic accuracy without carefully considering positional accuracy.

Introduction

Assessing the accuracy of thematic maps generated from remotely sensed data has become a required component of most mapping projects. Researchers assess their maps because they wish to determine if a newly developed technique or algorithm produces better results than an established method. Government agencies often require a measure of accuracy to meet the standards set up in the contract for the work. Many will use the map as part of a decisionmaking process, while others use map accuracy as a guide throughout the mapping project to evaluate the accuracy of each stage of the mapping process and to improve the map.

Errors come from many sources when generating a thematic map from remotely sensed data. Congalton and Green (1993) provide a good discussion of the errors that can result if the classification scheme is not well understood or if the reference data are poorly collected. Lunetta *et al.* (1991) present a very effective diagram and discussion of the various sources of error that can accumulate from the beginning of a mapping project through to the end. These sources include sensor issues, geometric registration, errors introduced by the classification process, assumptions made in the accuracy assessment, and limitations in the map output, to name just a few. Careful consideration of the entire mapping project before it is begun can go a long way toward reducing these errors.

Accuracy

Assessing the accuracy of maps generated from remotely sensed data requires evaluating both positional accuracy and thematic accuracy. While these two accuracies can be assessed separately, they are very much interrelated and failure to consider both of them is a serious mistake.

Positional Accuracy

Positional accuracy, a measure of how closely the imagery fits the ground, is the most common measure of map accuracy. In other words, positional accuracy is the accuracy of the location of a point in the imagery with reference to its physical location on the ground. It is imperative for any accuracy comparison that the same exact location can be determined both on the image and on the ground. The major factor influencing positional accuracy is topography, while sensor characteristics and viewing angles can also have some affect. It is commonly accepted that a positional accuracy of half a pixel is sufficient for sensors such as Landsat Thematic Mapper and SPOT. As sensors increase in spatial resolution, such as the 4-m multispectral IKONOS data, positional accuracy increases in importance and new standards need to be established. These standards need to be based on current ability to locate the chosen location (sample site) on both the image and the ground.

[1] Professor of Remote Sensing and Geographic Information Systems, Department of Natural Resources, 215 James Hall, 56 College Road, University of New Hampshire, Durham, NH 03824. E-mail: russ.congalton@unh.edu.

Positional accuracy is an integral part of thematic accuracy. If an image is registered to the ground to within half a pixel and a Global Positioning System (GPS) unit is used to locate the place on the ground to within about 15 meters, then it is impossible to use a single pixel as the sampling unit for assessing the thematic accuracy of the map. If positional accuracy is not up to the standard or a GPS is not used to precisely locate the point on the ground, then these factors increase in importance and can significantly affect the thematic accuracy assessment.

Figure 1 shows an example of positional accuracy. In this figure, the digital image is not exactly registered to the Geographic Information System (GIS) road layer. Therefore, the road layer does not line up exactly on top of the roads in the imagery. Positional accuracy has historically been based on National Map Accuracy Standards and measured in terms of root mean square error (RMSE). Most often, the RMSE is computed as the sum of the square of the differences between the position of the point on one data layer as compared to the position of the same point on another data layer (often the ground) using the same data that were used to register the layers together. This measure is, therefore, not an independent measure of positional accuracy. Instead, it would be more useful and more indicative of the true accuracy to collect an independent sample of points from which to compute the RMSE.

Thematic Accuracy

Thematic accuracy refers to the accuracy of a mapped land cover category at a particular time compared to what was actually on the ground at that time. Clearly, to perform a meaningful assessment of accuracy, land cover classifications must be assessed using data that are believed to be correct. Thus, it is vital to have at least some knowledge of the accuracy of the reference data before using it for comparison against the remotely sensed map. Congalton (1991: 42) points out that, "Although no reference data set may be completely accurate, it is important that the reference data have high accuracy or else it is not a fair assessment. Therefore, it is critical that the ground or reference data collection be carefully considered in any accuracy assessment."

Accuracy assessment begins with the generation of an error matrix (fig. 2), a square array of numbers or cells set out in rows and columns, which expresses the number of sample units assigned to each land cover type as compared to the reference data. The columns in the matrix represent the reference data (actual land cover) and the rows represent assigned (mapped) land cover types. The major diagonal of the matrix indicates agreement between the reference data and the interpreted land cover types.

Figure 1.—*Example of positional accuracy.*

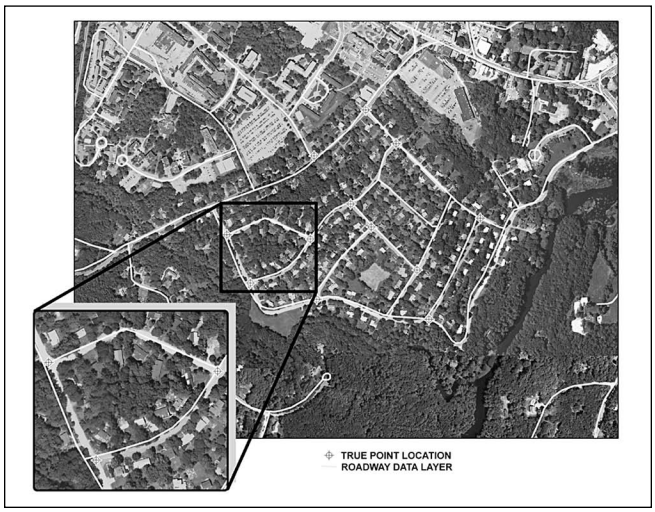

Figure 2.—*Example error matrix showing overall, producer's, and user's accuracies.*

		Reference Data V	W	U	row total
Classified Data	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

V = 43/48 = 90%
W = 23/34 = 68%
U = 30/41 = 73%

USER'S ACCURACY

V = 43/59 = 73%
W = 23/31 = 74%
U = 30/33 = 91%

The error matrix is useful for both visualizing image classification results and for statistically measuring the results. The error matrix is the only way to effectively compare two maps *quantitatively*. A measure of overall accuracy can be calculated by dividing the sum of all the entries in the major diagonal of the matrix by the total number of sample units in the matrix (Story and Congalton 1986). In the ideal situation, all the nonmajor diagonal elements of the error matrix would be zero, indicating that no area had been misclassified and that the map was 100 percent correct (Congalton *et al.* 1983). The error matrix also provides accuracies for each land cover category as well as both errors of exclusion (omission errors) and errors of inclusion (commission errors) present in the classification (Card 1982, Congalton 1991, Congalton and Green 1999).

Omission errors can be calculated by dividing the total number of correctly classified sample units in a category by the total number of sample units in that category from the reference data (the column total) (Congalton 1991, Story and Congalton 1986). This measure is often called the "producer's accuracy," because from this measurement the producer of the classification will know how well a certain area was classified (Congalton 1991). For example, the producer may be interested in knowing how many times vegetation was in fact classified as vegetation (and not, say, urban). To determine this, the 43 correctly classified vegetation samples (fig. 2) would be divided by the total 48 units of vegetation from the reference data, for a producer's accuracy of 90 percent. In other words, vegetation was correctly identified as vegetation 90 percent of the time.

Commission errors, on the other hand, are calculated by dividing the number of correctly classified sample units for a category by the total number of sample units that were classified in that category (Congalton 1991, Congalton and Green 1999, Story and Congalton 1986). This measure is also called "user's accuracy," indicating for the user of the map the probability that a sample unit classified on the map actually represents that category on the ground (Congalton and Green 1999, Story and Congalton 1986). In figure 2, while the producer's accuracy for the vegetation category is 90 percent, the user's accuracy is only 73 percent. That is, only 73 percent of the areas mapped as vegetation are actually vegetation on the ground. However, because each omission from the correct category is a commission to the wrong category, it is critical that both producer's and user's accuracies are considered, since reporting only one value can be misleading.

It is vital that the error matrix generated for the accuracy assessment be valid. An improperly generated error matrix may not be truly representative of the thematic map and, therefore, meaningless. The following factors must be considered to generate a valid error matrix (Congalton 1991):

1. Reference data collection.
2. Classification scheme.
3. Sampling scheme (Congalton 1988b, Hay 1979, Stehman 1992, van Genderen and Lock 1977).
4. Spatial autocorrelation (Campbell 1981, Congalton 1988a).
5. Sample size and sample unit (Congalton 1988b, Congalton and Green 1999, Hay 1979, van Genderen and Lock 1977).

Failure to consider even one of these factors could lead to significant shortcomings in the accuracy assessment process.

Reference Data Collection

Reference data collection is the first step in any assessment procedure, and may be the single most important factor in accuracy assessment, since an assessment will be meaningless if the reference data cannot be trusted. Reference data can be collected in many ways, including photo interpretation, aerial reconnaissance with a helicopter or airplane, video, drive-by surveys, and visiting the area of interest on the ground (Congalton and Biging 1992). Not all of these approaches are valid in every situation and great care needs to be taken to make sure that the reference data are accurate.

A key factor in reference data collection is the separation of training data from accuracy assessment data. In the not-too-distant past, many assessments of remotely sensed maps were conducted using the same data set used to train the classifier (Congalton 1991). This training and testing on the same data set resulted in an improperly generated error matrix that clearly overestimated classification accuracy. For accuracy assessment procedures to be valid and truly representative of the thematic

map, data used to train the image processing system should not be used for accuracy assessment. These data sets must be independent.

Finally, the information used to assess the accuracy of remotely sensed maps should be of the same general vintage as those originally used in map classification. The greater the time period between the imagery used in map classification and the data used in assessing map accuracy, the greater the likelihood that differences are due to change in vegetation (from harvesting, land use changes, etc.) rather than misclassification. Therefore, ground data collection should occur as close as possible to the date of the remotely sensed data.

Classification Scheme

A classification scheme categorizes remotely sensed map information into a meaningful and useful format. The rules used to label the map must be rigorous and well defined. An effective means of ensuring these requirements are met is to define a classification system that is totally exhaustive, mutually exclusive, and hierarchical (Congalton and Green 1999). A totally exhaustive classification scheme guarantees that *everything* in the image falls into a category; i.e., nothing is left unclassified. A mutually exclusive classification scheme means that everything in the image fits into one *and only one* category; i.e., an object in an image can be labeled only one category. Total exhaustion and mutual exclusivity rely on two critical components: (1) a set of labels (e.g., white pine forest, oak forest, nonforest, etc.), and (2) a set of rules (e.g., white pine forest must comprise at least 70 percent of the stand). Without these components, the image classification would be arbitrary and inconsistent. Finally, hierarchical classification schemes—those that can be collapsed from specific categories into more general categories—can be advantageous. For example, if it is discovered that white pine, red pine, and hemlock forest cannot be reliably mapped, these three categories could be collapsed into one general category called coniferous forest.

Sampling Scheme

An accuracy assessment very rarely involves a complete census or total enumeration of the classified image, since this data set is too large to be practical (Hay 1979, Stehman 1996, van Genderen and Lock 1977). Creating an error matrix to evaluate the accuracy of a remotely sensed map therefore requires sampling to determine if the mapped categories agree with the reference data (Rosenfield *et al.* 1982).

To select an appropriate sampling scheme for accuracy assessment, some knowledge of the distribution of the vegetation/land cover classes should be known. Stratified random sampling has historically prevailed for assessing the accuracy of remotely sensed maps. Stratified sampling has been shown to be useful for adequately sampling important minor categories, whereas simple random sampling or systematic sampling tended to oversample categories of high frequency and undersample categories of low frequency (Card 1982, van Genderen *et al.* 1978).

Spatial Autocorrelation

Because of sensor resolution, landscape variability, and other factors, remotely sensed data are often spatially autocorrelated (Congalton 1988a). Spatial autocorrelation involves a dependency between neighboring pixels such that a certain quality or characteristic at one location has an effect on that same quality or characteristic at neighboring locations (Cliff and Ord 1973, Congalton 1988a). Spatial autocorrelation can affect the result of an accuracy assessment if an error in a certain location can be found to positively or negatively influence errors in surrounding locations. The best way to minimize spatial autocorrelation is to impose some minimum distance between sample units.

Sample Size and Sample Unit

An appropriate sample size is essential to derive any meaningful estimates from the error matrix. In particular, small sample sizes can produce misleading results. Sample sizes can be calculated using the equation from the multinomial distribution, ensuring that a sample of appropriate size is obtained (Tortora 1978). Some researchers have suggested using the binomial equation to compute sample size. Given the need to create an error matrix, however, the binomial equation is inappropriate. A general rule of thumb developed from many projects shows that sample sizes of 50 to 100 for each map category are recommended, so that each category can be assessed individually (Congalton and Green 1999).

In addition to determining appropriate sample size, an appropriate sample unit must be chosen. Historically, the sample units chosen have been a single pixel, a cluster of pixels, a polygon, or a cluster of polygons. A single pixel is a poor choice of sample unit (Congalton and Green 1999), since it is an arbitrary delineation of the land cover and may have little relation to the actual land cover delineation. Further, it is nearly impossible to align one pixel in an image to the exact same area in the reference data. In many cases involving single pixel accuracy assessment, the positional accuracy of the data dictates a very low thematic accuracy. A cluster of pixels (e.g., a 3 by 3 pixel square) is always a better choice for the sample unit, since it minimizes registration problems. A good rule of thumb is to choose a sample unit whose area most closely matches the minimum mapping unit of the reference data. For example, if the reference data have been collected in 2-hectare minimum mapping units, then an appropriate sample unit may be a 2-hectare polygon.

Analysis Techniques

Once an error matrix has been properly generated, it can be used as a starting point to calculate various measures of accuracy in addition to overall, producer's, and user's accuracies. Two techniques have been found to be extremely useful. The first is a discrete multivariate technique called Kappa (Bishop *et al.* 1975), which can be used to statistically determine (1) if the remotely sensed classification is better than a random classification, and (2) if two or more error matrices are significantly different from each other. Kappa calculates a KHAT value (Cohen 1960), which is a measure of the *actual* agreement of the cell values minus the *chance* (i.e., random) agreement (Congalton and Mead 1983, Rosenfield and Fitzpatrick-Lins 1986) and can be viewed as a measure of accuracy. The KHAT value can be used to determine whether the results in the error matrix are significantly better than a random result (Congalton 1991). The KHAT accuracy value inherently includes more information than the overall accuracy measure since it indirectly incorporates the error (off-diagonal elements) from the error matrix. In addition, confidence limits can be calculated for the KHAT statistic, which allows for an evaluation of significant differences between KHAT values (Congalton and Green 1999).

Secondly, the analysis of the error matrix can be taken yet another step further by normalizing the cell values. An iterative proportional fitting technique, called *Margfit*, can be used to perform this normalization. Because the cell values in each row and column in the matrix are forced to sum to one, each cell value becomes a proportion of one, which can easily be multiplied by 100 to obtain percentages. Consequently, producer's and user's accuracies are not needed because the cell values along the major diagonal represent the proportions correctly mapped. Congalton *et al.* (1983) argue that the normalized accuracy is a more inclusive measure of accuracy than either KHAT or overall accuracy because it *directly* includes the information in the off-diagonal element of the error matrix. Because each row and column sums to the same value, different cell values (e.g., different forest cover classes) within an error matrix and among different error matrices can be compared despite differences in sample sizes. The software for performing both the Kappa and Margfit analyses is available from the author.

Acknowledgment

This work was funded under University of New Hampshire Agriculture Experiment Station McIntire-Stennis Project MS–33.

Literature Cited

Bishop, Y.; Fienberg, S.; Holland, P. 1975. Discrete multivariate analysis: theory and practice. Cambridge, MA: MIT Press. 575 p.

Campbell, J. 1981. Spatial autocorrelation effects upon the accuracy of supervised classification of land cover. Photogrammetric Engineering and Remote Sensing. 47: 355-363.

Card, D. 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering and Remote Sensing. 48(3): 431-439.

Cliff, A.D.; Ord, J.K. 1973. Spatial autocorrelation. London: Pion Limited. 178 p.

Cohen, J. 1960. A coefficient of agreement for nominal scale. Educational and Psychological Measurement. 20: 37-46.

Congalton, R.; Green, K. 1993. A practical look at the sources of confusion in error matrix generation. Photogrammetric Engineering and Remote Sensing. 59: 641-644.

Congalton, R.G. 1988a. Using spatial autocorrelation analysis to explore errors in maps generated from remotely sensed data. Photogrammetric Engineering and Remote Sensing. 54: 587-592.

Congalton, R.G. 1988b. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. Photogrammetric Engineering and Remote Sensing. 54: 593-600.

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

Congalton, R.G.; Biging, G.S. 1992. A pilot study evaluating ground reference data collection efforts for use in forest inventory. Photogrammetric Engineering and Remote Sensing. 58(12): 1669-1671.

Congalton, R.G.; Green, K. 1999. Assessing the accuracy of remotely sensed data: principles and practices. Boca Raton, FL: Lewis Publishers. 137 p.

Congalton, R.G.; Mead, R.A. 1983. A quantitative method to test for consistency and correctness in photo-interpretation. Photogrammetric Engineering and Remote Sensing. 49: 69-74.

Congalton, R.G.; Oderwald, R.G.; Mead, R.A. 1983. Assessing Landsat classification accuracy using discrete multivariate statistical techniques. Photogrammetric Engineering and Remote Sensing. 49: 1671-1678.

Hay, A.M. 1979. Sampling designs to test land-use map accuracy. Photogrammetric Engineering and Remote Sensing. 45: 529-533.

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.

Rosenfield, G.H. 1982. The analysis of areal data in thematic mapping experiments. Photogrammetric Engineering and Remote Sensing. 48(9): 1455-1462.

Stehman, S. 1992. Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. Photogrammetric Engineering and Remote Sensing. 58: 1343-1350.

Stehman, S. 1996. Estimating the Kappa coefficient and its variance under stratified random sampling. Photogrammetric Engineering and Remote Sensing. 62(4): 401-402.

Story, M.; Congalton, R.G. 1986. Accuracy assessment: a user's perspective. Photogrammetric Engineering and Remote Sensing. 52: 397-399.

Tortora, R. 1978. A note on sample size estimation for multinomial populations. The American Statistician. 32: 100-102.

van Genderen, J.L.; Lock, B.F. 1977. Testing land use map accuracy. Photogrammetric Engineering and Remote Sensing. 43: 1135-1137.

van Genderen, J.L.; Lock, B.F.; Vass, P.A. 1978. Remote sensing: statistical testing of thematic map accuracy. Proceedings of the twelfth international symposium on remote sensing of environment. ERIM. 3-14.