Accuracy Assessment Guidance

(Summarized from Congalton and Green 1999)

Maps are rarely 100 percent correct and yet they are widely used with unknown accuracy to inform decision-making processes. Every mapping project requires trade-offs, and some level of error is accepted as a trade-off for the cost savings inherent in using remotely sensed data. Knowing the extent of this error however, is critical for appropriate application of the derived maps.

The purpose of quantitative accuracy assessment is the identification and measurement of map errors. It involves comparison of a site on a map ('classified data') against field or photo information ('reference data') for the same site. All accuracy assessments include four fundamental steps: 1) design the sample, 2) collect data, 3) build and test the error matrix, and 4) analyze the results.

Step 1: Sample Design

The sample design used to collect valid reference data is one of the most important components of any accuracy assessment because it determines both the cost and statistical validity of the assessment. Several aspects need to be addressed:

- *Distribution of map information* - Reference data must be collected and labeled using the same classification scheme (including both the labels and decision rules) used to generate the map.

- *Appropriate sample unit* - Reference data should be collected at the same minimum mapping unit as was applied to the map. The sample unit (i.e., single pixel, cluster of pixels, polygons) dictates the level of detail for the accuracy assessment. For several reasons, a single pixel is rarely used and not recommended.

- *Number of samples* - A minimum of 50 samples for each land cover category is a good 'rule of thumb' as a starting point. If the area is excessively large greater than 1 million acres) or the classification has a large number of categories greater than 12) then the sample size may need to increased (e.g., 75 to 100 per category). Sample sizes can also be adjusted depending on the relative importance of each category or the inherent variability within each category.

- *Sampling scheme* - A stratified random sampling design where the samples are selected from each strata (i.e., map category) is recommended. The major advantage of this approach is that all map categories are sampled. However, this design requires known strata and therefore must take place after the map has been produced. This can be more expensive and time consuming than if the reference data were collected at the same time as the training data (i.e., at the start of the project). If data are collected at the start of the project, then a simple random sampling is recommended with subsequent subsetting of training and reference data so as to maintain independence.

Step 2: Data Collection

There are many methods for collecting reference data with a wide range in validity, reliability, and expense. In all cases, data collection requires three steps: 1) locate the sample site on both the reference data and the map, 2) delineate the sample unit, and 3) collect information based on the map classification scheme. Although each of these steps is fairly obvious and straightforward, serious problems can arise and, if the resulting reference data are inaccurate, the entire assessment becomes meaningless. To ensure objectivity and consistency, reference data must be independent of any training data, collected consistently from each sample site, and quality controlled. Data collection considerations include:

- *Source data* - Reference data must be collected from source data that is assumed to be more reliable than the remotely sensed data used to create the map. In general, a combination of existing aerial photography and field data are used. Existing data can be used if 1) the classification scheme (labels and definitions) are identical to the map, 2) the time frame is similar or any landscape changes have been addressed, and 3) errors inherent in the reference data (i.e., observation or classification errors) are addressed. Aerial photography is generally reliable with simple classification schemes but, as the level of detail in the classification scheme increases, the reliability of aerial photography for reference data decreases. When using aerial photography as reference data, it is necessary for a subsample of these data to also be collected from the field to verify the reliability of the photo interpretation.

- *Site location and delineation* - Spatial coordinates (e.g., latitude/longitude, UTM) and associated projection information (e.g., projection, datum) are necessary to navigate to the appropriate sampling site in the field and must be recorded at each site. The sampling unit (e.g., cluster of pixels or polygon) delineated around the sample site needs to be homogenous and match the minimum mapping unit of the map.

- *Observation versus measurement* - Simple observation is generally sufficient for labeling a reference sample. However, some classification schemes are dependent on measurements to differentiate between classes. In these instances, it must be decided if the variation and possible inaccuracies inherent in observed estimations is acceptable or whether the extra time and expense is necessary for taking measurements.

- *Timing of data collection* - Reference data should be collected as close as possible to the date of the remotely sensed imagery used to make the map.

- *Information collected* - Collected data need to include all information necessary to correctly classify the sample into one of the categories in the classification scheme. In addition, the following information should be included:

    1. Site identification – unique name including coding for field (F) or aerial photo (A), regional or management description, and sample number. For example: F_LAID_23 = the 23rd field reference site in Laidlaw Park
    2. Date
    3. Observer
    4. Photo – photo or map name or number used to delineate the sampling unit

5. Observation level – used as an indicator of the potential accuracy of the observations a '1' being most accurate and '4' being the least accurate
   1 = Walk through stand or polygon
   2 = Viewed from road or trail adjacent to stand
   3 = Viewed from afar (i.e., road or ridge opposite of stand)
   4 = Photo interpreted in office

Step 3: The Error Matrix

An error matrix is an effective way of communicating the accuracy of individual classes (user's and producer's accuracy) as well as overall map accuracy. It is a square array of numbers set in rows and columns that express the number of sample units (i.e., clusters or polygons) assigned to a particular category in one classification relative the number of sample units assigned in another classification (Figure 1). The columns usually represent the reference data, and are assumed to be correct, while the rows indicate the classification generated from the remotely sensed data (i.e., the map). Overall accuracy is the sum of correctly classified samples (i.e., sum of the major diagonal) divided by the total number of samples in the matrix. Producers and users accuracies are ways of representing individual class accuracies based on commission errors (including an area into a category when it does not belong to that category) and omission errors (exclusion of an area from the category to which it belongs).

For example, Figure 1 shows an error matrix for a simple, 6-class, classification scheme. In this case, the overall map accuracy is 72 percent. To assess the ability of the map to classify shrublands, we first calculate a "producer's accuracy" by dividing the total number of correct sample units in the shrubland category (i.e., 50) by the total number of shrubland sample units in the reference data (i.e., 68). This results in a producer's accuracy of 74 percent, which is a reasonable and acceptable value. However, we need to also calculate a "user's accuracy" to make an informed decision. Dividing the total number of correct pixels in the shrubland category (i.e., 50) by the total number of pixels classified as shrubland (i.e., 77) reveals a user's accuracy of 65 percent. In other words, although 74 percent of the known (sampled) shrubland areas have been identified as shrubland, only 65 percent of the areas called shrubland on the map are actually shrubland on the ground. Carefully looking at the error matrix reveals that there is confusion in discriminating shrubland from semi-desert, and to a lesser extent grassland, in this example.

Reference Data

		F	W	SH	GR	SD	AG	Row Total
	F	44	9	0	0	0	0	53
	W	7	41	5	0	1	0	54
	SH	0	6	50	8	13	0	77
Classified Data	GR	0	0	4	42	9	3	58
	SD	1	5	9	9	38	1	63
	AG	0	0	0	3	0	29	32
	Column Total	52	61	68	62	61	33	337

Land Cover Classes

F =Forest	GR = Grassland
W = Woodland	SD = Semi-desert
SH = Shrubland	AG = Agriculture

Overall Accuracy
(44 + 41 + 50 + 42 + 38 + 29) / 337 =
244 / 337 = 72%

Producer's Accuracy
F = 44 / 52 = 85%
W = 41 / 61 = 67%
SH = 50 / 68 = 74%
GR = 42 / 62 = 68%
SD = 38 / 61 = 62%
AG = 29 / 33 = 88%

User's Accuracy
F = 44 / 53 = 83%
W = 41 / 54 = 76%
SH = 50 / 77 = 65%
GR = 42 / 58 = 72%
SD = 38 / 65 = 58%
AG = 29 / 32 = 91%

Figure 1. Example error matrix with producers, users, and overall accuracy calculations.

Several other useful analyses of the error matrix are possible for the advanced user including Kappa analysis (for statistically determining if one error matrix is significantly different than another), Weighted Kappa analysis (for instances when not all errors have equal importance), Margfit (for normalizing the matrix if differences in sample sizes in each of the classes are great), and fuzzy logic (for addressing non-error differences – see step 4).

Step 4: Analysis of Results

Understanding what causes the reference and map data to differ provides for more informed and appropriate uses of the map as well as the creation of better maps in the future. All differences between map and reference data will be the result of one of the four following sources: 1) errors in the reference data, 2) sensitivity of the classification scheme to observer variability, 3) inappropriateness of the remote sensing technology for mapping a specific land cover class, or 4) mapping error. The first two are most easily addressed, the latter two more difficult to differentiate.

- *Errors in reference data* - A major assumption of the error matrix is that the reference data represents the “true” category of the site. If not extremely careful, reference data can include errors in registration, data entry, classification scheme, date changes, and labeling. For example, typical errors in data collection include misreading or mis-entering of the GPS location, mis-delineation of the sampling unit, misidentification of species, and duplicate entries. Any, and all, of these can lead to misinterpretations of the error matrix and resulting accuracy of the map.

- *Errors due to observer variability* - Classification schemes which rely on breaks in continuous conditions (i.e., vegetation cover) can be extremely sensitive to observer variability. For example, it is generally recognized that mapping crown closure typically varies by plus or minus 10 percent. One option for addressing this variability is to incorporate fuzzy logic into the reference data and error matrix to compensate for non-error differences. The key concept behind fuzzy logic is that membership in a category is a matter of degree and, when on the margins of classes dividing a continuum, an item may belong to both classes. For example, while a 100 percent shrubland can be classified as shrubland, and 100 percent grassland can be classified as grassland, an area with 49 percent shrub and 51 percent grass may be acceptably classified as either. Rules for fuzzy logic often rely on experts and can be driven by mapping and management objectives. Incorporating fuzzy logic, off-diagonal cells in the error matrix have two values, the first represents the number that although not absolutely correct are still acceptable within the fuzzy rules and the second value indicates those that are unacceptable. Accuracies are then calculated by combining the major diagonal values and those deemed acceptable (i.e., the first value) in the off-diagonals.

To make maps more reliable, the classification system can be collapsed across classes. While less information is then displayed on the map, the remaining information is more trustworthy.

References

Congalton, R. G. and K. Green. 1999. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis Publishers, Boca Raton, FL.