8 Analysis of Differences in the Error Matrix

After testing the error matrix for statistical significance, the next step in the accuracy assessment analysis involves exploring why some of the map labels do not match the reference labels. While much attention is commonly placed on overall accuracy percentages (the sum of the major diagonal divided by the total), by far the more interesting analysis concerns discovering why some of the accuracy assessment samples did not fall on the diagonal of the error matrix. To both effectively use the map and to make better maps in the future, we need to know what causes these off-diagonal samples or differences in the matrix to occur.

All off-diagonal samples or differences in the error matrix will be the result of one of four possible sources:

1. Errors in the reference data,
2. Sensitivity of the classification scheme to observer variability,
3. Inappropriateness of the remote sensing data employed for mapping a specific land cover class, and
4. Mapping error.

This chapter reviews each one of these sources and discusses the impacts of each one to the accuracy assessment results.

ERRORS IN THE REFERENCE DATA

A major and required assumption of the error matrix is that the label from the reference data represents the "correct" label of the site and that all differences between the map and the reference label are due to classification and/or delineation error. Although this assumption is necessary, the reference data will never be perfect. As previously discussed, the term "ground truth" should be avoided for exactly this reason. Throughout this book, the authors prefer the use of reference data or reference label to refer to the sample data set being compared to the map that is being assessed. Unfortunately, error matrices can be inadequate indicators of map error because they are often confused by errors in the reference data (Congalton and Green, 1993). Errors in the reference data can be a function of the following:

- *Registration differences.* Registration differences between the reference data and the remotely sensed map classification caused by delineation and/or digitizing errors. For example, if GPS is not used in the field during accuracy assessment, it is possible for field personnel to collect data in the wrong area. Other registration errors can occur when an accuracy assessment site is incorrectly delineated or digitized, or when an existing map used for reference data is not precisely registered to the map being assessed.

- *Data entry errors*. Data entry errors are common in any database project and can be controlled only through rigorous quality control. Developing digital data entry forms that will only allow a certain set of characters for specific fields can catch errors during data entry. One of the best—but expensive—methods for catching data entry errors is to enter all data twice and then compare the two data sets. Differences usually indicate an error.
- *Classification scheme errors*. Every accuracy assessment map and reference site must have a label derived from the classification scheme used to create the map. Classification scheme errors occur when personnel misapply the classification scheme to the map or reference data, a common occurrence with complex classification schemes. If the reference data are in a database, then such errors can be avoided, or at least highlighted, by programming the classification scheme rules and using the program to label accuracy assessment sites. Classification scheme errors also occur when the classification scheme used to label the reference site is different from the one used to create the map, a common occurrence when existing data or maps are used as reference data.
- *Change*. Changes in land cover between the date of the remotely sensed imagery collection and the date of the reference data. As discussed in Chapter 6, land cover change can have a profound effect on accuracy assessment results. Tidal differences, crop or tree harvesting, urban development, fire, and pests all can cause the landscape to change in the time period between capturing the remotely sensed data and the accuracy assessment reference data collection.
- *Mistakes in labeling reference data*. Labeling mistakes usually occur because inexperienced personnel are used to collect reference data. Even with experienced personnel, the more detailed the classification scheme, the more likely it is that an error in labeling the reference data will occur. For example, some conifer and hardwood species are difficult to distinguish on the ground, much less from aerial photography. Young crops of broccoli, brussels sprouts, and cauliflower are easily confused. Thus, an accuracy assessment must also be performed on the reference data. If manual photo interpretation is used to assess a map created through semi-automated methods, then a sample of the photo-interpreted sites must be visited on the ground. If field data are used, then some sample of the sites must be visited by two different personnel and their answers compared. If the answers mostly agree, then the collection is satisfactory. If the answers mostly disagree, then there is a problem with the reference data collection method.

TABLE 8.1
Analysis of Map and Reference Label Differences

Map versus Ref. Difference	Number of Sites Different	Map Error	Reference Label Error	Date Change	Class. Scheme Difference	Variation in Estimation
Barren vs Water	19	0	6	8	0	5
Hardwood vs Water	6	0	0	0	0	6
Herb vs Forested	50	6	17	4	0	23
Wetland vs All Other Types	50	0	0	0	50	0
Total	125	6	23	12	50	34

Table 8.1 summarizes the reference data errors discovered during the quality control process for an actual accuracy assessment. Only 6 out of 125 of the differences between the map and reference labels were caused by actual errors in the map. Over two thirds of the differences (85 sites) were caused by mistakes in the reference data. The most significant error occurred from using different classification schemes (50 sites). In this project, National Wetlands Inventory (NWI) maps were used exclusively to map wetlands (i.e., wetlands were defined in the classification scheme to be those areas identified by NWI data as wetlands). However, when the accuracy assessment was performed, the photo interpreters collecting the reference data used a different definition of wetlands and disagreed with all the NWI labels. The remaining differences were caused by landscape change, reference label error, and observer variation, which is discussed in the next section of this chapter.

SENSITIVITY OF THE CLASSIFICATION SCHEME TO OBSERVER VARIABILITY

Classification scheme rules often impose discrete boundaries on continuous conditions in nature such as vegetation cover, soil type, or land use. In situations where breaks in the classification scheme represent artificial distinctions along a continuum, observer variability is often difficult to control and, although unavoidable, can have profound effects on accuracy assessment results (Congalton, 1991; Congalton and Green, 1993). Analysis of the error matrix must include exploring how many of the matrix differences are the result of observers being unable to precisely distinguish between classes when the accuracy assessment site is on the margin between two or more classes in the classification scheme.

Plato's parable of the shadows in the cave is useful for thinking about observer variability. In the parable, Plato describes prisoners who cannot move:

> "Above and behind them a fire is blazing in the distance, and between the fire and the prisoners there is a … screen which marionette players have in front of them over which they show puppets … (the prisoners) see only their own shadows, or the shadows of one another which the fire throws on the opposite wall of the cave … To them … the truth would be literally nothing but the shadows of the images." (Plato, *The Republic,* Book VII, 515-B, from Benjamin Jowett's translation as published in Vintage Classics, Random House, New York).

Like Plato's prisoners in the cave, we all perceive the world within the context of our experience. The difference between reality and perceptions of reality is often as fuzzy as Plato's shadows. Our observations and perceptions vary day to day and depend on our training, experience, or mood.

The analysis in Table 8.1 shows the impact that variation in interpretation can have on accuracy assessment. In the project, two photo interpreters were asked to label the same accuracy assessment reference sites. Almost 30% (34 of 125) of the differences between the map and reference label were caused by variation in interpretation.

Consider, for example, the assessment of a map of tree crown closure with classification scheme rules defining classes as:

Unvegetated	0–10%
Sparse	11–30%
Light	31–50%
Medium	51–70%
Heavy	71–100%

An accuracy assessment reference site from photo interpretation estimated at 45% tree crown cover would be labeled "Light." However, since it is recognized that crown closure can only be interpreted on aerial photos to ±10% (Spurr, 1960), it is also feasible that the proper label could be "Medium." Either the label of Light or the label of Medium is within the variability of the reference data collection. The map user would be much more concerned with a difference caused by a map label of Unvegetated compared to a reference label of Heavy tree crown cover. Differences on class margins are both inevitable and far less significant to the map user than other types of differences.

Classification schemes that employ estimates of percentage of vegetative cover are particularly susceptible to this type of confusion in the error matrix. Appendix 8.1 of this chapter presents a set of very complex classification scheme rules for a mapping project of Wrangell-St. Elias National Park in Alaska. The classification scheme is highly sensitive to estimates of percent vegetative cover. Sensitivity analysis on 140 accuracy assessment sites revealed that nearly 33% of the sites received new class labels when estimates of vegetative cover were varied by as little as 5%.

Several researchers have noted the impact of the variation in human interpretation on map results and accuracy assessment (Gong and Chen, 1992; Lowell, 1992;

Congalton and Biging, 1992; Congalton and Green, 1993). Woodcock and Gopal (1992) state, "The problem that makes accuracy assessment difficult is that there is ambiguity regarding the appropriate map label for some locations. The situation of one category being exactly right and all other categories being equally and exactly wrong often does not exist." Lowell (1992) calls for "a new model of space which shows transition zones for boundaries, and polygon attributes as indefinite." As Congalton and Biging (1992) conclude in their study of the validation of photo-interpreted stand-type maps, "The differences in how interpreters delineated stand boundaries was most surprising. We were expecting some shifts in position, but nothing to the extent that we witnessed. This result again demonstrates just how variable forests are and the subjectiveness of photo interpretation."

While it is difficult to control observer variation, it is possible to measure the variation, and to use the measurements to compensate for differences between reference and map data that are caused not by map error but by variation in interpretation. One option is to measure each reference site precisely to reduce observer variance in reference site labels. This method can be prohibitively expensive, usually requiring extensive field sampling. The second option incorporates fuzzy logic into the reference data to compensate for nonerror differences between reference and map data, and is discussed in Chapter 9.

INAPPROPRIATENESS OF THE REMOTE SENSING DATA EMPLOYED TO MAKE THE MAP

Early satellite remote sensing projects were primarily concerned with testing the viability of various remote sensing data for mapping certain types of land cover. Researchers tested the hypotheses of whether or not the imagery could be used to detect land use, or crop types, or forest types. Many accuracy assessment techniques were developed primarily to test these hypotheses.

Recently, accuracy assessment has focused more on learning about the reliability of a map for land management or policy analysis. However, some of the differences in the error matrix will be because the map producer was attempting to use remote sensing data or methods that were incapable of distinguishing certain land cover or vegetation class types. Understanding what differences are caused by the technology is useful to the map producer when the next map is being made.

In the Wrangell-St. Elias example cited earlier, Landsat TM data was employed as the primary remotely sensed data, with 1:60,000 aerial photography as ancillary data. The classification scheme included distinguishing between pure and mixed stands of black and white spruce. Accuracy assessment analysis consistently showed success at differentiating pure stands of black versus white spruce. However, consistently differentiating these species in mixed or occasional hybrid stands was found to be unreliable. This phenomenon is not surprising considering the difficulty often associated with differentiating these species in mixed and hybrid stands on the ground. In summary, moderate resolution multispectral remotely sensed data, at the scales employed, cannot be used to reliably and consistently differentiate between mixed classes of these two tree species.

To make the map more reliable, the map user can collapse the classification system across classes. In this example, the nonpure spruce classes of Closed, Open, and Woodland were collapsed into an "Unspecified Interior Spruce" class. In the difference matrix, "Unspecified Interior Spruce," map labels were considered to be mapped correctly if they corresponded to a pure or mixed white spruce or black spruce reference site demonstrating the same density class of Closed, Open, or Woodland. For example, a map label of "Open Unspecified Interior Spruce" was considered to be correctly mapped if its corresponding reference label for the site was "Open Black Spruce," "Open White Spruce," or "Open Black/White Spruce" mix. While less information is displayed on the map, the remaining information is more reliable.

MAPPING ERROR

The final cause of differences in the error matrix is the result of mapping error (i.e., the actual real errors). Often, these are difficult to distinguish from an inappropriate use of remote sensing data. Usually, they are errors that are particularly obvious and unacceptable. For example, it is not uncommon for an inexperienced remote sensing professional to produce a map of land cover from satellite data that misclassifies northeast-facing forests on steep slopes as water. Because water and shadowed wooded slopes both absorb most energy, this type of error is explainable but unacceptable, and must be avoided. Many map users will be appalled at this type of error and are not particularly interested in having the electromagnetic spectrum explained to them as an excuse. However, careful editing and comparison with aerial photography, checking that all water exists in areas without slope, and comparison to existing maps of waterways and lakes will reduce the possibility of this type of map error.

Understanding the causes of true error can point the map producer to additional methods of improving the accuracy of the map. Perhaps other bands or band combinations will improve accuracy. Incorporation of ancillary data such as slope, aspect, or elevation may be useful. In the Wrangell-St. Elias example, confusion existed between the Dwarf Shrub classes and the Graminoid class. The confusion was addressed through the use of unsupervised classifications and parkwide models utilizing digital elevation data, field-based data, and aerial photography. First, an unsupervised classification with 20 classes was run for only those areas of the imagery classified as Dwarf Shrub in the map. A digital elevation coverage was utilized to stratify the study area for subsequent relabeling of unsupervised classes previously mapped as Dwarf Shrub but actually representing areas of Graminoid cover on the ground. From the unsupervised classification, two spectral classes were found to consistently represent Graminoid cover throughout the study area while another spectral class was found to represent Graminoid cover in areas below 3500 ft elevation. These spectral classes were subsequently recoded to the Graminoid class.

SUMMARY

Analysis of the causes of differences in the error matrix can be one of the most important and interesting steps in the creation of a map from remotely sensed data. In the past, too much emphasis was placed on the overall accuracy of the map, without investigating the conditions that give rise to that accuracy. By understanding what caused the reference and map data to differ, we can use the map more reliably, and produce both better maps and better accuracy assessments in the future.

APPENDIX 8.1

Wrangell-St. Elias National Park and Preserve: Land Cover Mapping Classification Key

If tree total ≥ 10% (Forested)	
If Conifer ≥ 75% of tree total	
If (Pigl + Pima) ≥ 67% of conifer total	
If (Pigl/(Pigl+Pima)) ≥ 75%	**PIGL**
If (Pima/(Pigl+Pima)) ≥ 75%	**PIMA**
Else	**Unspecified Spruce**
If Broadleaf ≥ 75% of tree total	**Broadleaf**
Else (mixed conifer/broadleaf)	**Spruce/Broadleaf**
Else if shrub total ≥ 25% (Shrub)	
If tall shrub total ≥ 25%	**Tall Shrub**
If low shrub total ≥ 25%	**Low Shrub**
If dwarf shrub total ≥ 25%	**Dwarf Shrub**
Else (tall, low, or dwarf are not individually > 25%)	
If tall shrub total ≥ 67% of shrub total	**Tall Shrub**
If low shrub total ≥ 67% or shrub total	**Low Shrub**
If dwarf shrub total ≥ 67% of shrub total	**Dwarf Shrub**
Else "pick the largest percent of ":	
tall shrub	**Tall Shrub**
low shrub	**Low Shrub**
dwarf shrub	**Dwarf Shrub**
(ties go to the "tallest")	
Else if herbaceous ≥ 15% (Herbaceous)	
If graminoid ≥ 50% or (graminoid/herb total) ≥ 50%	**Graminoid**
Else if forb ≥ 50% or (forb/herb total) ≥ 50%	**Forb**
Else if moss ≥ 50% or (moss/herb total) ≥ 50%	**Moss/Lichen**
Else if lichen ≥ 50% or (lichen/herb total) ≥ 50%)	**Moss/Lichen**
Else "pick the largest percent of ":	
graminoid	
forb	
moss	
lichen	
(preference for ties go in the order listed)	
Else if total vegetation ≥ 10% and < 30%	**Sparse Vegetation**
Else (nonvegetated)	

Water
Barren
Glacier/Snow
Clouds/Cloud
Shadow

Forested (>10% tree cover)
- **Conifer** (>75% conifer)
  - **Closed** (60–100%)
    - **Pigl**
    - **Pima**
    - **Pigl/Pima**
    - **Pisi**
    - **Tshe**
    - **Tsme**
    - **Pisi/Tsme**
    - **Pisi/Tshe**
    - **Tshe/Tsme**
    - **Spruce**
    - **Mixed Conifer**
  - **Open** (25–59%)
    - **Pigl**
    - **Pima**
    - **Pigl/Pima**
    - **Pisi**
    - **Tshe**
    - **Tsme**
    - **Pisi/Tsme**
    - **Pisi/Tshe**
    - **Tshe/Tsme**
    - **Spruce**
    - **Mixed Conifer**
  - **Woodland** (10–24%)
    - **Pigl**
    - **Pima**
    - **Pigl/Pima**
    - **Pisi**
    - **Tshe**
    - **Tsme**
    - **Pisi/Tsme**
    - **Pisi/Tshe**
    - **Tshe/Tsme**
    - **Spruce**
    - **Mixed Conifer**
- **Broadleaf** (>75% broadleaf)
  - **Closed** (60–100%)
    - **Closed Broadleaf**
  - **Open** (10–59%)
    - **Open Broadleaf**

Mixed

Closed (60–100%)

Pigl/Pima-Broadleaf

Pisi-Broadleaf

Tshe-Broadleaf

Conifer-Broadleaf

Open (10–59%)

Pigl/Pima-Broadleaf

Pisi-Broadleaf

Tshe-Broadleaf

Conifer-Broadleaf

Shrub (>25% shrub)

Tall (tall shrub > 25% or dominant)

Closed (>75%)

Open (25–74%)

Low (low shrub > 25% or dominant)

Closed (>75%)

Open (25–74%)

Dwarf (dwarf shrub > 25% or dominant)

Herbaceous (herbaceous > 15%)

Graminoid

Forb

Moss

Lichen

Sparse vegetation

Sparse vegetation

Nonvegetated

Water

Barren

Glacier/Snow

Clouds/Cloud Shadow