5 Sample Design Considerations

Now that we understand thematic accuracy is typically represented using an error matrix, it is important to know how to correctly generate and populate the matrix. Assessing the thematic accuracy of maps or other spatial data requires sampling because it is not economically feasible to visit every place on the ground. Sampling design requires knowledge of the distribution of thematic classes across the landscape, determination of the types and number of samples to be taken, and choice of a sampling scheme for selecting the samples. Design of an effective and efficient sample to collect valid reference and map accuracy data is one of the most challenging and important components of any accuracy assessment, because the design will determine both the cost and the statistical rigor of the assessment.

Accuracy assessment assumes that the information displayed in the error matrix is a true characterization of the map being assessed. Thus, an improperly designed sample will produce misleading accuracy results. Several considerations are critical to designing an accuracy assessment sample that is truly representative of the map:

1. What are the thematic map classes to be assessed and how are they distributed across the landscape?
2. What is the appropriate sample unit?
3. How many samples should be taken?
4. How should the samples be chosen?

While seemingly straightforward, each of these steps has many potential pitfalls. Failure to consider even one of them can lead to serious shortcomings in the assessment process. This chapter considers each one of these factors.

WHAT ARE THE THEMATIC MAP CLASSES TO BE ASSESSED?

How we sample the map for accuracy will partially be driven by how the thematic classes of the map are distributed across the landscape. This distribution will, in turn, be a function of how we have chosen to categorize the features of the earth being mapped; referred to as the *classification scheme*. Once we know the classification scheme, we can learn more about how the map classes are distributed. Important considerations are the discrete nature of map information, and the spatial interrelationship or autocorrelation of that information. Assumptions made about the

distribution of map categories will affect both how we select accuracy assessment samples and the outcome of the analysis.

The Classification Scheme

Maps categorize the earth's surface. For example, road maps tell us the type of road, its name, and location. Land cover maps typically enumerate the types, mix, and density of vegetation covering the earth (e.g., trees, shrubs, and grass). Land use maps characterize how land is utilized by humans (e.g., urban, agriculture, and forest management).

Thematic map categories are specified by the project's classification scheme. Classification schemes are a means of organizing spatial information in an orderly and logical way (e.g., Cowardin et al., 1979). Classification schemes are fundamental to any mapping project because they create order out of chaos and reduce the total number of items considered to some reasonable number. The classification scheme makes it possible for the map producer to characterize landscape features and for the user to readily recognize them. Without a classification scheme, no mapping is truly possible. The detail of the scheme is driven by (1) the anticipated uses of the map information, and (2) the features of the earth that can be discerned with the remotely sensed data (e.g., aerial or satellite imagery) being used to create the map. If a rigorous classification scheme is not developed before mapping begins, then any subsequent accuracy assessment of the map will be meaningless because it will be impossible to definitively label the accuracy assessment samples.

A classification scheme has two critical components: (1) a set of *labels* (e.g., urban residential, deciduous forest, palustrine emergent wetland, etc.); and (2) a set of *rules* or definitions such as a dichotomous key for assigning labels (e.g., a "deciduous forest must have at least 75% crown closure in deciduous trees"). Without a clear set of rules, the assignment of labels to classes can be arbitrary and lack consistency. For example, everyone has their own idea about what constitutes a forest and yet there are many definitions that could result in very different maps of forest distribution. Consider a situation in which one agency defines forest as an area where 10% of the ground area is covered by trees, and another agency uses a slightly different definition according to which forest exists only if 25% of the ground area is covered by trees. If analysts from each of these agencies were together in a specific plot of land, they could label the area differently based on their agency's definitions of a forest and both of the labels would be correct. Without class definitions expressed as quantifiable rules, there can be little agreement on what area on the ground or the image should be labeled.

The level of detail (i.e., number and complexity of the categories) in the scheme strongly influences the time and effort needed to make the map and to conduct the accuracy assessment. The more detailed the scheme, the more expensive the map and its assessment. Because the classification scheme is so important, no work should begin on a mapping project until the scheme has been thoroughly reviewed and as many problems as possible identified and solved.

In addition to being composed of labels and a set of rules, a classification scheme should be (1) *mutually exclusive* and (2) *totally exhaustive*. Mutual exclusivity requires that each mapped area fall into one and only one category or class. For example,

classification scheme rules would need to clearly distinguish between forest and water (seemingly simple), so that a mangrove swamp cannot receive both a forest and a water label. A totally exhaustive classification scheme results in every area on the mapped landscape receiving a map label; no area can be left unlabeled. One way to ensure that the scheme is totally exhaustive is to have a category labeled as other or unclassified.

If possible, it is also advantageous to use a classification scheme that is *hierarchical*. In hierarchical systems, specific categories within the classification scheme can be collapsed to form more general categories. This ability is especially important when it is discovered that certain map categories cannot be reliably mapped. For example, it may be impossible to separate interior live oak from canyon live oak in California's oak woodlands (these two oak types are almost indistinguishable on the ground). Therefore, these two categories may have to be collapsed to form a live oak category that can be reliably mapped.

Finally, the classification scheme must specify the minimum mapping unit (mmu) for each class being mapped. The mmu is the smallest area of the class to be delineated on the map. Figure 5.1 illustrates this concept. In this example, the rule for mapping a forest is specified as follows:

> An area of 1 acre or more where more than 30% of the ground, as seen from above the tree canopy, is covered by the foliage of hardwood or conifer trees.

The minimum mapping unit for forests is one acre. Areas covered with 30% tree foliage, but smaller than the 1 acre minimum mapping unit will not be labeled as forests. Additionally, areas larger than 1 acre but containing less than 30% tree foliage cover will also not be labeled as forests. Reference data must be collected at the

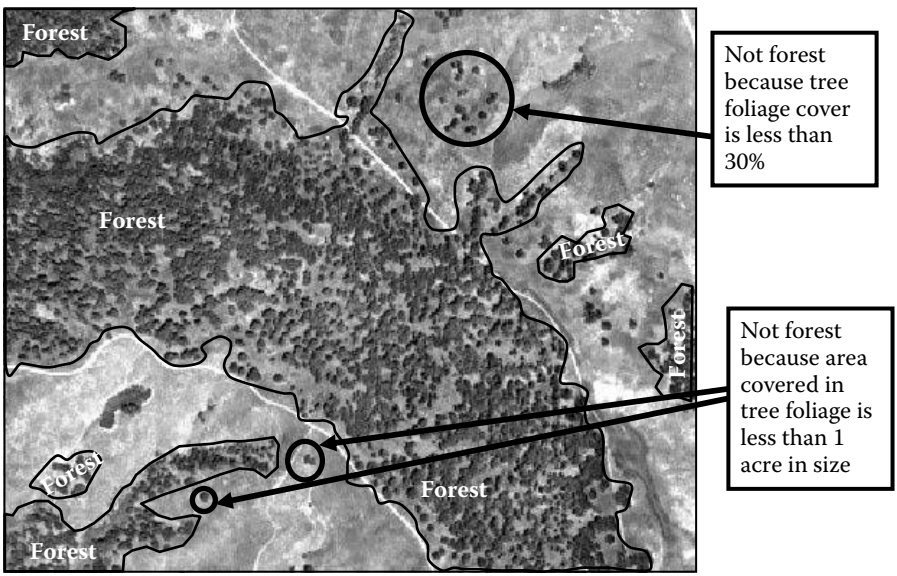


FIGURE 5.1 Example of the impact of a minimum mapping unit.

same minimum mapping unit as was applied to the map generated from the remotely sensed data. For example, it is not possible to assess the accuracy of a Landsat 30 m × 30 m pixel with a single 1/20 hectare ground inventory plot, nor is it possible to assess the accuracy of an AVHRR 1.1 km × 1.1 km pixel using a 30 m × 30 m pixel.

Figure 5.2 provides a dichotomous key for a simple, yet robust, classification scheme for a fire/fuel mapping project. Note how the scheme specifies a minimum mapping unit and is:

- Totally exhaustive—every piece of the landscape will be labeled,
- Mutually exclusive—no one piece of the landscape can receive more than one label, and
- Hierarchical—detailed fuel classes can be lumped into the more general groups of nonfuel, grass, shrub, timber slash, and timber litter.

It is critical that accuracy assessment reference data be collected and labeled using the same classification scheme as that used to generate the map. This may seem obvious until you are tempted to use an existing map to assess the accuracy of a new map. Rarely will any two maps be created using the same classification scheme. Any differences between the classification scheme of the map and the classification scheme of the reference data may result in discrepancies between map and reference accuracy assessment site labels. The result will be an assessment of classification scheme differences, and not of map accuracy.

Other Data Considerations

Continuous versus Noncontinuous Data

Most statistical analysis assumes that the population to be sampled is continuous and normally distributed, and that samples will be independent. Yet we know that classification systems, for all their power in organizing chaos, also take a continuous landscape and divide it into often arbitrarily discrete categories. For example, tree crown closure rarely exists in discrete classes. Yet when we make a map of crown closure, we impose discrete crown closure classes across the landscape. For example, we may create a crown closure map with 4 classes; class 1 from 0 to 10% crown closure, class 2 from 11 to 50% crown closure, class 3 from 51 to 75% crown closure, and class 4 from 76 to 100% crown closure. Given this boundary between two crown closure classes at 75%, one can expect to find confusion between a forest stand with a crown closure of 73% that belongs in class 3 and a stand of 77% that belongs in class 4 (see Chapter 9 for a discussion on fuzzy accuracy assessment). In addition, categories tend to be related spatially, resulting in autocorrelation (discussed next in this chapter). In most situations, some balance between what is statistically valid and what is practically obtainable is desired. Therefore, knowledge of these statistical considerations is a must.

Most students who have completed a beginning statistics course are familiar with sampling and analysis techniques for continuous, normally distributed data. It is these techniques such as analysis of variance (ANOVA) and linear regression that are most familiar to the reader.

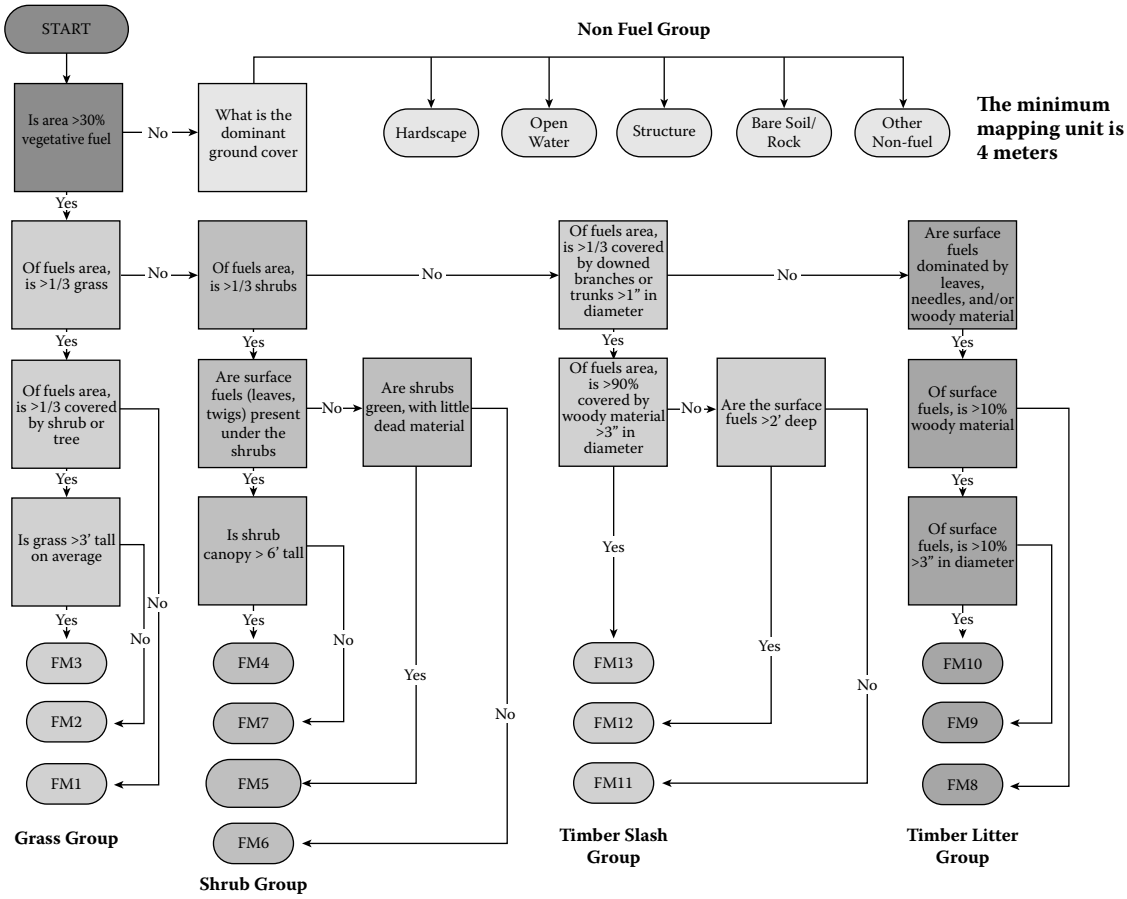


FIGURE 5.2 Example wildland fuel classification scheme.

However, thematic map information is discrete, not continuous, and frequently not normally distributed. Therefore, normal theoretical statistical techniques that assume a continuous normal distribution may be inappropriate for map accuracy assessment. It is important to consider how the data are distributed and what assumptions are being made before performing any statistical analysis. Sometimes there is little that can be done about the artificial delineations in the classification scheme; other times the scheme can be modified to better represent natural breaks. Care and thought must go into this process to achieve the best analysis possible.

Spatial Autocorrelation

Spatial autocorrelation occurs when the presence, absence, or degree of a certain characteristic affects the presence, absence, or degree of that same characteristic in neighboring units (Cliff and Ord, 1973), thereby violating the assumption of sample independence. This condition is particularly important in accuracy assessment if an error in a certain location can be found to positively or negatively influence errors in surrounding locations (Campbell, 1981). Clearly, if spatial autocorrelation exists, the sampling must ensure that the samples are separated by enough distance to minimize this effect, or else the sampling will not adequately represent the entire map.

The existence of spatial autocorrelation is clearly illustrated in work by Congalton (1988a) on Landsat MSS data from three areas of varying spatial diversity/complexity (i.e., an agriculture, a rangeland, and a forested site), which showed a positive influence over 1 mile away. Figure 5.3 presents the results of this analysis. Each image, called a difference image, is a comparison between the remotely sensed classification (i.e., the map) and the reference data. The black areas represent the error, those places where the map and the reference label disagree. The white areas represent the agreement.

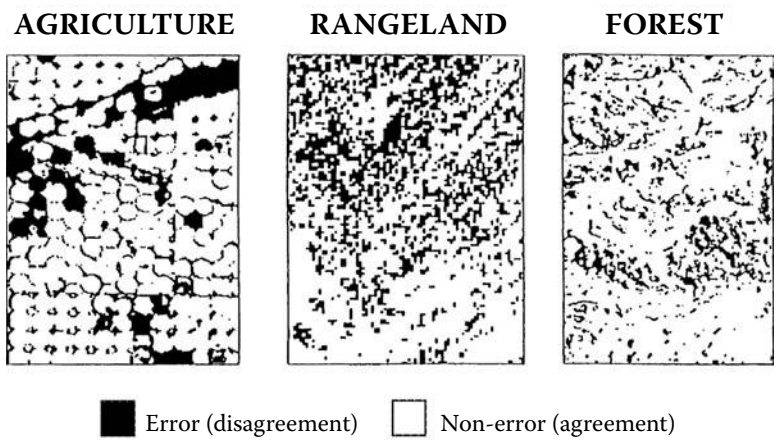


FIGURE 5.3 Difference images (7.5-minute quadrangles) showing the pattern of error for three ecosystems of varying complexity: agriculture, rangeland, and forest. (Reproduced with permission from the American Society for Photogrammetry and Remote Sensing, from Congalton, R. 1988. Using spatial autocorrelation analysis to explore errors in maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*. 54(5): 587–592.)

The pattern of differences between the map and reference labels are readily explainable in an agricultural environment in which field sizes are large and typical misclassification would result in an error in labeling the entire field. In the agricultural difference image in Figure 5.3, the fields are circular fields employing center-pivot irrigation, and examples can be seen of misclassifying entire fields. For example, a field that is mapped as corn when it is actually wheat will result in an entire field (center pivot area in Figure 5.3) being mislabeled. Therefore, it is not surprising that the errors occur in large areas and that there is a positive autocorrelation over a large distance.

However, the results are more surprising for the rangeland and forested classes. Both classes are more spatially complex (i.e., have more fragmentation, edges, and mixtures of land cover) than the agriculture class, and therefore one would expect them to be less spatially autocorrelated. Primarily because of rangeland fencing, the rangeland class does have some of the fields similar to agriculture, but it also reflects some of the edge effects more common to the complex forest class.

The forest class is the most spatially complex, and most map error would be expected to occur along the edges or transition zones between forest types. Although viewing the forest difference image does tend to confirm these edge problems, the results of the analysis still indicate that there is strong positive autocorrelation between errors up to 30 pixels away. In other words, if an error occurs at a given location, it is more likely that another error will be found, even up to this rather large distance away (i.e., 30 MSS pixels or about 240 m), than a correct classification.

The existence of spatial autocorrelation can violate the assumption of sample independence which, in turn, can affect the sample size and especially the sampling scheme used in accuracy assessment. Spatial autocorrelation may indicate the existence of periodicity in the presence of a class across the landscape that could affect the results of any type of systematic sample if the systematic sample design repeats the same periodicity. For example, maple trees need ample water and, in arid landscapes, are usually located along streams. A systematic sampling scheme based on choosing samples near streams would repeat the periodicity of the maple forest class and would result in a biased choice of samples that would oversample maple forests and undersample other map classes.

In addition, autocorrelation may affect the size and number of samples used in cluster sampling because each sample unit may not be contributing new, independent information, but rather, redundant information. Therefore, it would not be effective to collect information in a large cluster sample since very quickly each new sample site in the cluster would be adding very little new information. However, cluster sampling is a very cost-effective method, especially in the field, when the cost of traveling from one sample to another can be very high. Even when the accuracy assessment samples are taken in the office from aerial imagery, cluster sampling can create savings in setup time for each image. Therefore, it is important to consider spatial autocorrelation and balance the impact of having spatially autocorrelated samples against the efficiencies of cluster sampling. This can be done by limiting the number of samples taken in the cluster to 2–4, making sure that each sample unit in the cluster is taken in a different thematic class, and spreading the samples as far apart as possible.

WHAT IS THE APPROPRIATE SAMPLE UNIT?

Sample units are the portions of the map that will be selected for accuracy assessment. There are four possible choices for the sampling unit: (1) a single pixel, (2) a cluster of pixels (often a 3×3 pixel square), (3) a polygon (or object), and (4) a cluster of polygons.

Single Pixel

Historically, a large number of accuracy assessments have been conducted using a single pixel as the sampling unit. However, a single pixel is a very poor choice for the sampling unit for many reasons:

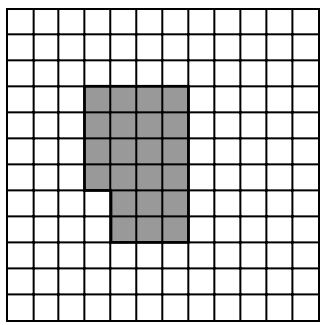
- First, a pixel is an arbitrary rectangular delineation of the landscape that may have little relation to the actual delineation of land cover or land use type. It can be a single land cover or vegetation category (i.e., a pure pixel) or more often than not, it can be a mixture of land cover or vegetation classes.
- Second, before the relatively new geocoding and terrain correction procedures were adopted, it was almost impossible to exactly align one pixel on a map to the exact same area in the reference data. Therefore, there was no way to guarantee that the location of the reference pixel was identical to the location of the map pixel. Even with terrain correction and georeferencing, it is still not possible to get an exact alignment of the boundaries of a pixel. Similarly, until global positioning system (GPS) came along, there was no practical way to ensure that ground-collected reference data was being collected for the exact map pixel being assessed. Even with GPS, this correspondence is not guaranteed to exactly match. Therefore, positional accuracy becomes a large issue, and the thematic accuracy of the map is affected because of positional error.
- Finally, few classification schemes specify a unit as small as a pixel as the minimum mapping unit. If the mmu is larger than a single pixel, then a single pixel is inappropriate as the sample unit.

Even with all the recent technological advances in GPS, terrain correction, and geocoding, accuracy assessment sample units will still have some positional inaccuracies. It is commonly accepted that a positional accuracy of one-half pixel is sufficient for sensors such as Landsat Thematic Mapper and SPOT Multispectral imagery. As sensors increase in spatial resolution, such as that collected from digital airborne cameras and high-resolution satellites, positional accuracy becomes more important and new standards need to be established. If an image with a pixel size of 10–30 m is registered to the ground to within half a pixel (i.e., 5–15 m) and a GPS unit is used to locate the unit on the ground to within 10–15 m, then it is impossible to use a single pixel as the sampling unit for assessing the thematic accuracy of the map. There would simply be no guarantee that the map and the reference data would be collected from the identical area. If the positional accuracy is not up to the standard or if GPS is not used to precisely locate the sample on the ground, then these factors

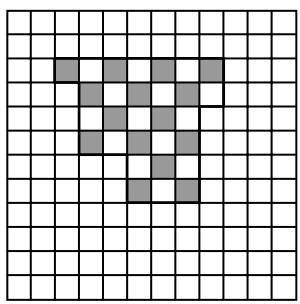
become more important and can significantly affect the thematic accuracy assessment. This is all the more true for higher spatial resolution imagery, in which the pixels may be smaller.

Cluster of Pixels

Given the need to balance thematic accuracy with positional accuracy, a cluster of pixels, typically a 3×3 square for moderate resolution imagery, has recently been the most common choice for the sample unit. A cluster minimizes registration problems because it is easier to locate on the reference data or in the field. However, a cluster of pixels (especially a 3×3 window) may still be an arbitrary delineation of the landscape resulting in the sample unit encompassing more than one map category. To avoid this problem, many analysts require that only homogeneous clusters of pixels be sampled. However, such restrictions may result in a biased sample that avoids heterogeneous areas which are a function of a mix of pixels (e.g., a mixed hardwood-conifer stand of trees) as depicted in Figure 5.4.



Homogeneous polygon



Heterogeneous polygon

FIGURE 5.4 Comparison of accuracy assessment polygons comprising homogeneous versus heterogeneous pixels.

It is important to remember that the sample unit dictates the level of detail for the accuracy assessment. If the assessment is performed on a 3×3 cluster of pixels, then nothing can be said about an individual pixel, nor can anything be said about polygons (i.e., management areas, forest stands, agricultural fields, etc.). Additionally, each sample unit must be considered a single sample. If, for example, a 3×3 cluster of pixels is used as the sample unit, then it must be counted as one sample, and not as nine samples. There are numerous examples in the literature of authors mistakenly counting each pixel in a cluster as a separate accuracy assessment unit. Also, the presence of spatial autocorrelation in most thematic maps dictates that samples should be spaced adequately apart from one another.

Extending the concept of a cluster of pixels to higher-resolution imagery requires knowledge about the positional accuracy of the imagery. As previously stated, common registration (positional) accuracies for Landsat Thematic Mapper (30 m pixels) and SPOT (10 m pixels) satellite imagery are about half a pixel. Therefore, selecting a homogeneous cluster of 3×3 pixels ensures that the center of the sample will definitely fall within the 3×3 cluster. Higher spatial resolution imagery such as that from Ikonos or Digital Globe have pixel sizes of 4 m to below 1 m. However, because of the off-nadir acquisition and other issues, the positional accuracy of these data are often in the range of 10–20 m and can even be much larger. Therefore, a 3×3 pixel cluster as the sampling unit would not be appropriate in this case. If the registration accuracy was 10 m and the pixel size was 4 m, then the cluster would need to be at least 5×5 pixels to account for this positional error. It is imperative that the positional accuracy be considered in the selection of the sample unit cluster size or else the thematic assessment will be flawed.

Polygons

Most large-scale thematic maps delineate the landscape into polygons of homogeneous map classes. Polygons are delineated on edges of classes, where more "between" than "within" class polygon variation exists. While the pixels inside the polygons may vary dramatically (as in a sparse stand of trees), the class label across the pixels is constant. Usually the polygon map is created either through manual interpretation or through the use of image segmentation and object-oriented classification algorithms. If the map to be assessed is a polygon map, then the accuracy assessment sample units should also be polygons. The resulting accuracy values inform the map's user and producer about the level of detail in which they are interested: the polygons. More and more mapping projects using remotely sensed data are generating polygon rather than pixel products as a result of developments in image segmentation and object-based image analysis. As a result, the polygon is replacing the cluster of pixels as the sample unit of choice.

However, using polygons as sample units can cause confusion if the accuracy assessment polygons are collected during the initial training data/calibration fieldwork, which occurs before the map polygons are created. The result can often be manually delineated accuracy assessment polygons with dramatically different delineations than the final map polygons, as illustrated in Figure 5.5. When this occurs, some

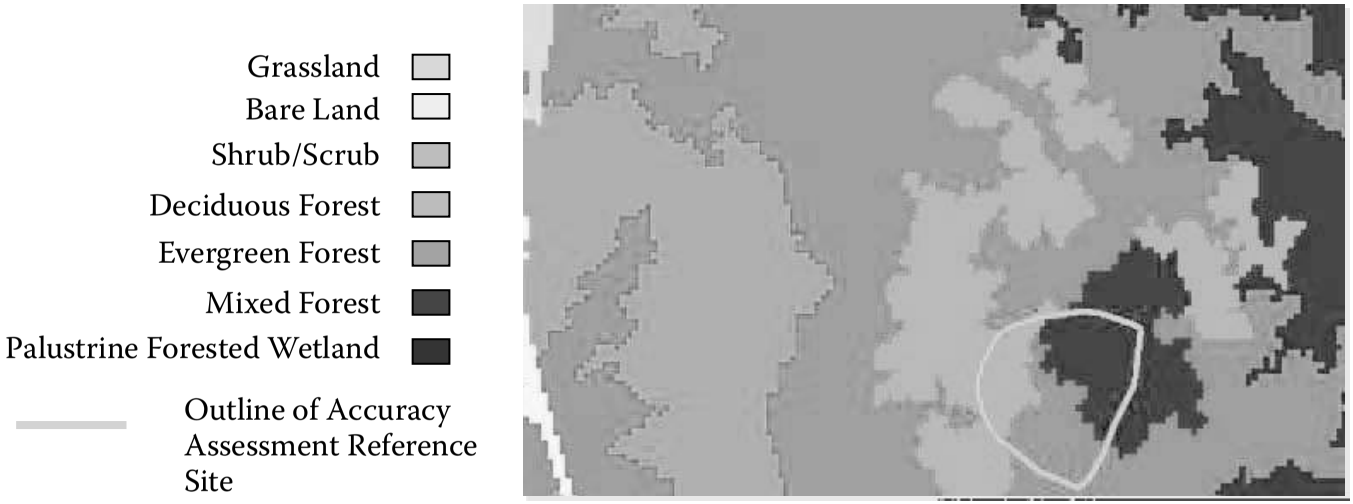


FIGURE 5.5 ***(Color version follows page 112)*** Mixed forest accuracy assessment reference polygon over the map polygons of evergreen, mixed, and deciduous forest. Determining the map label of the accuracy assessment polygon, when the polygon intersects with multiple map classes, can be problematic.

way of creating the map label for the accuracy assessment polygon must be developed. The simplest approach is to use the majority class of the polygon to create the map label. However, this may not work well in heterogeneous conditions, in which the label is more a function of the mix of the ground cover (e.g., patchy seagrass or mixed hardwood conifer forests) rather than the majority of the ground cover.

Another approach is to run the segmentation algorithms and finalize the delineation of the polygons prior to the initial field trip. The resulting polygons will probably vary only slightly from the final polygons and therefore can be reliably used as accuracy-sampling units.

Clusters of Polygons

Sampling clusters of polygons (or a grouping of polygons together), rather than single polygons, can reduce accuracy assessment costs dramatically because travel time and/or setup time is decreased. Unlike clusters of pixels, each polygon within a cluster of polygons can represent a single sampling unit because polygons are by definition separate map class types that have more between than within variation. However, care must be taken to provide some separation between polygons and to limit the number in the cluster. In other words, the impact of spatial autocorrelation still must be considered.

HOW MANY SAMPLES SHOULD BE TAKEN?

Accuracy assessment requires that an adequate number of samples per map class be gathered so that the assessment is a statistically valid representation of the accuracy of the map. However, the collection of reference data at each sample unit is very expensive, requiring that sample size be kept to a minimum to be affordable.

Of all the considerations discussed in this chapter, the most has probably been written about sample size. Many researchers, notably Hord and Brooner (1976), van Genderen and Lock (1977), Hay (1979), Ginevan (1979), Rosenfield et al. (1982), and Congalton (1988b), have published equations and guidelines for choosing the appropriate sample size.

The majority of work performed by early researchers used an equation based on the binomial distribution or the normal approximation to the binomial distribution to compute the required sample size. These techniques are statistically sound for calculating the sample size needed to compute the overall accuracy of a classification or even the overall accuracy of a single category. The equations are based on the proportion of correctly classified sample units and on some allowable error. However, these techniques were not designed to choose a sample size for generating an error matrix.

In the case of creating an error matrix, it is not simply a question of correct or incorrect (the binomial case). Instead, it is a matter of which error or which categories are being confused. Given an error matrix with n land cover categories, for a given category there is one correct answer and $(n - 1)$ incorrect answers. Sufficient samples must be acquired to be able to adequately represent this confusion (i.e., build a statistically valid error matrix). Therefore, the use of the binomial distribution for determining the sample size for an error matrix is not appropriate. Instead, the use of the multinomial distribution is recommended (Tortora, 1978).

The appropriate sample size can and should be computed for each project using the multinomial distribution. However, in our experience, a general guideline or good "rule of thumb" suggests planning to collect a minimum of 50 samples for each map class for maps of less than 1 million acres in size and fewer than 12 classes (Congalton, 1988b). Larger area maps or more complex maps should receive 75 to 100 accuracy assessment sites per class. These guidelines were empirically derived over many projects, and the use of the multinomial equation has confirmed that they are a good balance between statistical validity and practicality.

Because of the large number of potential samples (i.e., pixels, clusters of pixels, polygons, and clusters of polygons) in a remotely sensed image, traditional thinking about sampling in which a 2%, or even 5%, sample is not uncommon often does not apply. To illustrate this point, even as small a sample as 0.5% of a single Landsat Thematic Mapper scene is over 300,000 pixels. As we have previously concluded, accuracy assessment should not be performed on a per-pixel basis, but the same relative argument holds true for the other sample units. Therefore, practical considerations are often a key component of the sample size selection process. For example, the number of samples for each category may be adjusted on the basis of the relative importance of that category within the objectives of the mapping project or by the inherent variability within each of the categories. Sometimes, because of budget constraints or other factors, it is better to concentrate the sampling on the categories of interest and increase their number of samples while reducing the number of samples taken in the less important categories. Also, it may be useful to take fewer samples in categories that show little variability, such as water or forest plantations, and increase the sampling in the categories that are more variable, such as uneven-aged forests or riparian areas. However, in most instances, some minimum number of samples (e.g., 50 samples as per the guidelines or the result of the multinomial equation calculation) should be taken in each land cover category contained in the matrix. Perhaps most importantly, the entire accuracy assessment process should be documented so that others can know exactly what procedures were followed.

Finally, it may be tempting to design a sample that selects many samples in categories which are most accurate and a few in the confused categories. This strategy would guarantee a high accuracy value, but would not be representative of the map accuracy. Care should be taken to ensure that the sampling effort is carefully planned and implemented. It should also be noted that exactly how the sample is selected can affect the analysis performed on the sampled data. Again, the object here is to balance the statistical recommendations in order to get an adequate sample to generate an appropriate error matrix within the time, cost, and practical limitations associated with any viable mapping project. However this balance is achieved, it is critical to document the exact process so that future users of the map can know how the assessment was conducted.

Binomial Distribution

As mentioned earlier, the binomial distribution or the normal approximation to the binomial distribution is appropriate for computing the sample size for determining overall accuracy or the accuracy of an individual category. Later in this book, the

binomial distribution will be used to assess a change/no change map (see Chapter 11). It is appropriate for the two-case situation in which only right and wrong are important. Choosing the appropriate sample size from the binomial or normal approximation is dependent on (1) the level of acceptable error one is willing to tolerate and (2) the desired level of confidence that the actual accuracy is within some minimum range. Numerous publications present look-up tables of the required sample size for a given acceptable error and desired level of confidence (e.g., Cochran, 1977 and Ginevan, 1979).

For example, suppose it is decided that a map is unacceptable if the overall accuracy is 90% or less. Also, let us say that we are willing to accept a 1 in 20 chance that we will make a mistake based on our sample and accept a map that actually has an accuracy of less than 90%. Finally, let us decide that we will accept the same risk, a 1 in 20 chance, of rejecting a map that is actually correct. The appropriate look-up table would then indicate that we must take 298 samples of which only 21 can be misclassified. If more than 21 samples were misclassified, we would conclude that the map is not acceptable.

Multinomial Distribution

As discussed earlier in this chapter, the multinomial distribution provides the appropriate equations for determining the sample size required to generate an error matrix. The procedure for generating the appropriate sample size from the multinomial distribution is summarized here and was originally presented by Tortora (1978).

Consider a population of units divided into k mutually exclusive and exhaustive categories. Let Π_i, $i = 1, \ldots, k$, be the proportion of the population in the ith category, and let n_i, $i = 1, \ldots, k$, be the frequency observed in the ith category in a simple random sample of size n from the population.

For a specified value of α, we wish to obtain a set of intervals S_i, $i = 1, \ldots, k$, such that

$$\Pr\left\{\bigcap_{i=1}^{k}(\Pi_i \in S_i)\right\} \geq 1-\alpha;$$

that is, we require the probability that every interval S_i contains Π_i to be at least $1-\alpha$. Goodman (1965) determined the approximate large-sample confidence interval bounds (when $n \to \infty$) as

$$\Pi_i^- \leq \Pi_i \geq \Pi_i^+,$$

where

$$\Pi_i^- = \Pi_i - \left[B\Pi_i(1-\Pi_i)/n\right]^{1/2} \tag{5.1}$$

$$\Pi_i^+ = \Pi_i + \left[B\Pi_i(1-\Pi_i)/n\right]^{1/2} \tag{5.2}$$

and B is the upper $(\alpha/k) \times 100$th percentile of the χ^2 distribution with 1 degree of freedom. These equations are based on Goodman's (1965) procedure for simultaneous confidence interval estimation.

Examining these equations (Equations 5.1 and 5.2), we see that $[\Pi_i(1-\Pi_i)/n]^{1/2}$ is the standard deviation for the ith cell of the multinomial population. Also, it is important to realize that each marginal probability mass function is binomially distributed. If N is the total population size, then using the finite population correction (fpc) factor and the variance for each Π_i (from Cochran, 1977), the approximate confidence bounds are:

$$\Pi_i^- = \Pi_i - \left[B(N-n)\Pi_i\left(1-\Pi_i\right)/(N-1)n\right]^{1/2} \tag{5.3}$$

$$\Pi_i^+ = \Pi_i + \left[B(N-n)\Pi_i\left(1-\Pi_i\right)/(N-1)n\right]^{1/2} \tag{5.4}$$

Note as $N \to \infty$, Equations 5.3 and 5.4 converge to Equations 5.1 and 5.2, respectively.

Next, in order to determine the required sample size, the precision for each parameter in the multinomial population must be specified. If the absolute precision for each cell is set to b_i, then Equations 5.1 and 5.2 become

$$\Pi_i - b_i = \Pi_i - \left[B\Pi_i\left(1-\Pi_i\right)/n\right]^{1/2} \tag{5.5}$$

$$\Pi_i + b_i = \Pi_i + \left[B\Pi_i\left(1-\Pi_i\right)/n\right]^{1/2} \tag{5.6}$$

respectively. Similar results are obtained when the fpc is included. Equations 5.5 and 5.6 can be rearranged to solve for b_i (the absolute precision of the sample)

$$b_i = \left[B\Pi_i\left(1-\Pi_i\right)/n\right]^{1/2} \tag{5.7}$$

Then by squaring Equation 5.7 and solving for n, the result is:

$$n = B\Pi_i\left(1-\Pi_i\right)/b_i^2 \tag{5.8}$$

or, using the fpc,

$$n = BN\Pi_i\left(1-\Pi_i\right)/\left[b_i^2(N-1) + B\Pi_i\left(1-\Pi_i\right)\right] \tag{5.9}$$

Therefore, one should make k calculations to determine the sample size, one for each pair (b_i, Π_i), $i = 1, \ldots, k$, and select the largest n as the desired sample size. As functions of Π_i and b_i, Equations 5.8 and 5.9 show that n increases as $\Pi_i \to 1/2$ or $b_i \to 0$.

In rare cases, a relative precision b_i' could be specified for each cell in the error matrix and not just each category. Here $b_i = b_i' \ \Pi_i$. Substituting this into Equation 5.8 gives

$$n = B\left(1-\prod_i\right)/\prod_i b_i'^2 \tag{5.10}$$

A similar sample size calculation including the fpc can be computed as before.

Here again, one should make k calculations, one for each pair (b_i', Π_i), $i = 1, \ldots, k$. The largest n computed is selected as the desired sample size. As $\Pi_i \to 1/2$ or $b_i' \to 0$ the sample size increases according to Equation 5.10. If $b_i' = b'$ for all i, then the largest sample size is $n = B(1-\Pi)/\Pi b'^2$, where $\Pi = \min(\Pi_1, \ldots, \Pi_k)$.

In the majority of cases for assessing the accuracy of remotely sensed data, an absolute precision is set for the entire classification and not each category or each cell. Therefore, $b_i = b$ and the only sample size calculation required is for the Π_i closest to 1/2. If there is no prior knowledge about the values of the Π_i's, a "worst-case" calculation of sample size can be made assuming some $\Pi_i = 1/2$ and $b_i = b$ for $i = 1, \ldots, k$. In this worst-case scenario, the sample size required to generate a valid error matrix can be obtained from this simple equation as follows:

$$n = B/4b^2.$$

This approach can be made much clearer with a numerical example. First, let us look at an example using the full equation (Equation 5.8) and then at the corresponding sample size using the worst-case or conservative sample size equation. Assume that there are eight categories in our classification scheme ($k = 8$), that the desired confidence level is 95%, the desired precision is 5%, and that this particular class makes up 30% of the map area ($\Pi_i = 30\%$). The value for B must be determined from a *chi*-square table with 1 degree of freedom and $1 - \alpha/k$. In this case, the appropriate value for B is $\chi^2_{(1,0.99375)} = 7.568$. Therefore, the calculation of the sample size is as follows:

$$n = B\prod_i\left(1-\prod_i\right)/b_i^2$$

$$n = 7.568(0.30)(1-0.30)/(0.05)^2$$

$$n = 1.58928/0.0025$$

$$n = 636$$

A total of 636 samples should be taken to adequately fill an error matrix or approximately 80 samples per class given that there were 8 classes in this map.

If the simplified worst-case scenario equation is used, then the class proportion is assumed to be 50% and the calculation is as follows:

$$n = B/4b^2$$

$$n = 7.568/4(0.05)^2$$

$$n = 7.568/0.01 = 757$$

In this worst-case scenario, approximately 95 samples per class, or 757 total samples would be required.

If the confidence interval is relaxed from 95 to 85%, the required sample sizes decrease. In the earlier example, the new appropriate value for B would be $\chi^2_{(1,0.98125)} = 5.695$ and the total samples required would be 478 and 570 for the complete equation and the worst case scenario, respectively.

HOW SHOULD THE SAMPLES BE CHOSEN?

In addition to the considerations already discussed, the choice and distribution of samples, or sampling scheme, is an important part of any accuracy assessment. Selection of the proper scheme is critical to generating an error matrix that is representative of the entire map. First, to arrive at valid conclusions about a map's accuracy, the sample must be selected without bias. Failure to meet this important criterion affects the validity of any further analysis performed because the resulting error matrix may over- or underestimate the true accuracy. Second, further data analysis will depend on which sampling scheme is selected. Different sampling schemes assume different sampling models, and consequently, different variance equations to compute the required accuracy methods. Finally, the sampling scheme will determine the distribution of samples across the landscape, which will significantly affect accuracy assessment costs.

Sampling Schemes

Many researchers have expressed opinions about the proper sampling scheme to use (e.g., Hord and Brooner, 1976; Rhode, 1978; Ginevan, 1979; Fitzpatrick-Lins, 1981; and Stehman, 1992). These opinions vary greatly among researchers and include everything from simple random sampling to a scheme called stratified, systematic, unaligned sampling.

There are five common sampling schemes that have been applied for collecting reference data: (1) simple random sampling, (2) systematic sampling, (3) stratified random sampling, (4) cluster sampling, and (5) stratified, systematic, unaligned sampling. In a simple random sample, each sample unit in the study area has an equal chance of being selected. In most cases, a random number generator is used to pick random x, y coordinates to identify samples to be collected. The main advantage of simple random sampling is the good statistical properties that result from the random selection of samples (i.e., it results in the unbiased selection of samples).

Systematic sampling is a method in which the sample units are selected at some specified and regular interval over the study area. In most cases, the first sample is randomly selected and each successive sample is taken at some specified interval thereafter. The major advantage of systematic sampling is the ease in sampling somewhat uniformly over the entire study area.

Stratified random sampling is similar to simple random sampling; however, some prior knowledge about the study area is used to divide the area into groups or strata and then each stratum is randomly sampled. In the case of accuracy assessment, the map has been stratified into map classes. The major advantage of stratified random sampling is that all strata (i.e., map classes), no matter how small, will be included in

the sample. This factor is especially important in making sure that sufficient samples are taken in rare but important map classes.

In addition to the sampling schemes already discussed, cluster sampling has also been frequently used in assessing the accuracy of maps from remotely sensed data, especially to collect information on many samples quickly. There are clear advantages to collecting a number of sample units in close proximity to one another. However, cluster sampling must be used intelligently and with great care. Simply taking a large number of sample units (whether they be pixels, a cluster of pixels, or polygons) together is not a valid method of collecting data, because each sample unit is not independent of the other and adds very little additional information. Congalton (1988b), looking at single pixels, recommended that no clusters larger than 10 pixels and certainly not larger than 25 pixels be used because each pixel beyond these cluster sizes did not add further information.

Finally, stratified, systematic, unaligned sampling attempts to combine the advantages of randomness and stratification with the ease of a systematic sample, without falling into the pitfalls of periodicity common to systematic sampling. This method is a combined approach that introduces more randomness than just a random start within each stratum.

Sampling Scheme Considerations

Congalton (1988b) performed sampling simulations on three spatially diverse areas (see Figure 5.5) using all five of these sampling schemes and concluded that in all cases, simple random and stratified random sampling provided satisfactory results.

Simple random sampling allows reference data to be collected simultaneously for both training and assessment. However, it is not always appropriate, because it tends to undersample rarely occurring, but possibly very important, map categories unless the sample size is significantly increased. For this reason, stratified random sampling, in which a minimum number of samples are selected from each stratum (i.e., map category), is often recommended. However, stratified random sampling can be impractical because stratified random samples can only be selected after the map has been completed (i.e., when the location of the strata is known). This limits the accuracy assessment data to being collected late in the project instead of in conjunction with the training data collection, thereby increasing the costs of the project. In addition, in some projects the time between the project beginning and the accuracy assessment may be so long as to cause temporal problems in collecting ground reference data. In other words, the ground may change (e.g., the crop may have been harvested) between the time the project is started and the accuracy assessment is begun.

The concept of randomness is a central issue when performing almost any statistical analysis because a random sample is one in which each member of the population has an equal and independent chance of being selected. Therefore, a random sample ensures that the samples will be chosen without bias. If in-office manual interpretation is used to label reference samples, then random sampling is feasible because

access to the samples will not be a problem. However, a subset of the sample should be visited on the ground to verify the accuracy of the interpretation.

Despite the nice statistical properties of random sampling, access in the field to random sample units can often be problematic because many of the samples will be difficult to locate. Locked gates, fences, travel distances, and rugged terrain all combine to make random field sampling extremely costly and difficult. In forested and other wildland environments, randomly selected samples may be totally inaccessible except by helicopter. The cost of getting to each of the randomly located samples can be more than the cost of the rest of the entire mapping effort.

Obviously, one cannot spend the majority of project resources collecting accuracy assessment reference data. Instead, some balance must be struck. Often, some combination of random and systematic sampling provides the best balance between statistical validity and practical application. Such a system may employ systematic or simple random sampling to collect some assessment data early in a project, and stratified random sampling within strata after the classification is completed to ensure that enough samples were collected for each category and to minimize any periodicity in the data. However, results of Congalton (1988a) showed that periodicity in the errors, as measured by the autocorrelation analysis, could make the use of systematic sampling risky for accuracy assessment.

An example of a combined approach could include a systematic sample tied to existing aerial photography with sample selection based on the center of every *n*th photo. Sample choices based on flight lines should not be highly correlated with a factor determining land cover unless the flight lines were aligned with a landscape feature. Choice of the number of samples per photo and the sampling interval between photos would depend on the size of the area to map and the number of samples to collect. This systematic sample would ensure that the entire mapped area gets covered.

However, rarely occurring map classes will probably be undersampled. It may be necessary to combine this approach with stratified random sample when the map is completed to augment the underrepresented map categories. It may be practical to limit the stratified random field sample selection within some realistic distance of the roads. However, care must be taken because roads tend to occur on flatter areas and in valleys along streams, which will bias sample selection to land cover likely to exist there; so steps must be taken to mitigate these factors so that the most representative sample can be achieved. This type of combined approach minimizes the resources used and obtains the maximum information possible. Still, the statistical complexities of such a combination cannot be neglected. Again, a balance is desirable.

Finally, some analytic techniques assume that certain sampling schemes were used to obtain the data. For example, use of the Kappa analysis for comparing error matrices (see Chapter 7 for details of this analysis technique) assumes a multinomial sampling model. Only simple random sampling completely satisfies this assumption. If another sampling scheme or combination of sampling schemes is used, then it may be necessary to compute the appropriate variance equations for performing the Kappa analysis or other similar technique. The effects (i.e., bias) of using another of

the sampling schemes discussed here and not computing the appropriate variances are unknown.

An interesting project would be to test the effect on the Kappa analysis of using a sampling scheme other than simple random sampling. If the effect is found to be small, then the scheme may be appropriate to use, subject to the conditions discussed earlier. If the effect is found to be large, then that sampling scheme should not be used to perform the Kappa analysis. If that scheme is to be used, then the appropriate correction to the variance equation must be applied. Stehman (1992) has done such an analysis for two sampling schemes (simple random sampling and systematic sampling). His analysis shows that the effect on the Kappa analysis of using systematic sampling is negligible. This result adds further credence to the idea of using a combined systematic initial sample followed by a random sample to fill in the gaps.

Table 5.1 presents a summary of the pros and cons of the different possible accuracy assessment sampling schemes.

TABLE 5.1
A Summary of the Pros and Cons of Various Accuracy Assessment Sampling Schemes

Sampling Scheme	Pros	Cons
Random	Unbiased sample selection. Excellent statistical properties.	Expensive, especially for fieldwork. Does not ensure that enough samples will be taken in each class. Does not ensure good distribution of samples across the landscape.
Stratified Random	Unbiased sample selection. Ensures adequate sample in each class because a minimum number of samples is selected from each stratum (class).	Requires prior knowledge about the distribution of map classes so that strata can be developed. Expensive, especially for fieldwork. Often difficult to find enough samples in rare map classes. Does not ensure good distribution of samples across the landscape.
Systematic	Easy to implement. Less expensive than random sampling. Ensures good distribution of samples across the landscape.	Can be biased if sampling pattern is correlated with a landscape pattern (periodicity). Weaker statistically, as each sample unit does not have equal probability of selection.
Cluster	Least expensive as samples are close to one another, reducing travel time in the field and/or set up time in the office.	Can be impacted by spatial autocorrelation, which results in the samples not being independent. If the samples are not independent from one another, then they are not distinct samples, and more independent samples must be taken.

FINAL CONSIDERATIONS

Because of the many assumptions required for statistical analysis, a few researchers have concluded that some sampling designs can be used for descriptive techniques and others for analytical techniques. However, this conclusion seems quite impractical. Accuracy assessment is expensive, and no one is going to collect data for only descriptive use. Eventually, someone will use that matrix for some analytical technique. It is best to pay close attention to both the practical limitations and the statistical requirements when performing any accuracy assessment.