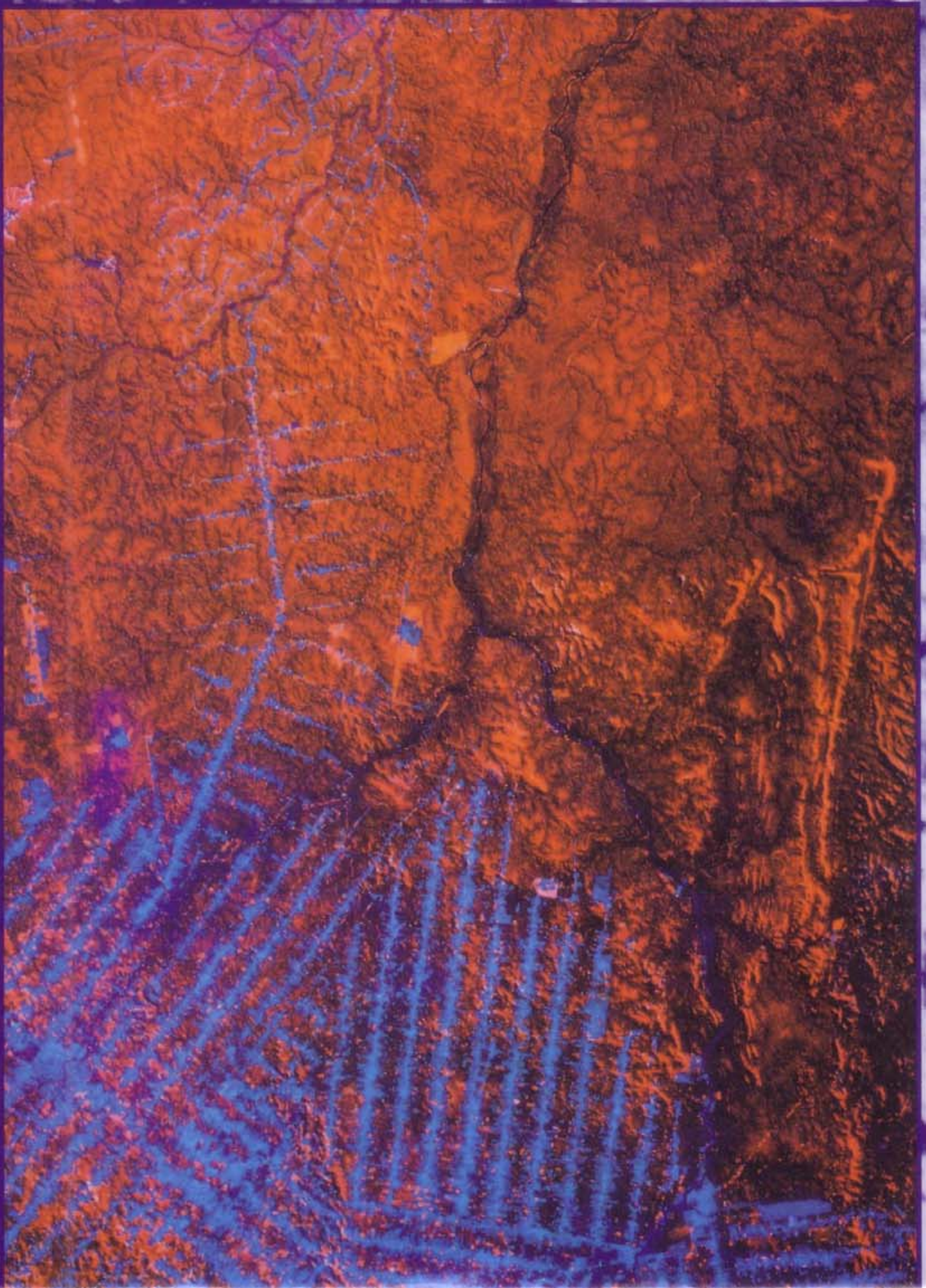
LILLESAND • KIEFER

REMOTE SENSING AND IMAGE INTERPRETATION

THIRD EDITION



desirable to "smooth" the classified output to show only the dominant (presumably correct) classification (Figure 7.48*b*). Initially, one might consider the application of the previously described low-pass spatial filters for this purpose. The problem with this approach is that the output from an image classification is an array of pixel locations containing numbers serving the function of *labels*, not *quantities*. That is, a pixel containing land cover 1 may be coded with a 1. A pixel containing land cover 2 may be coded with a 2, and so on. A moving low pass filter will not properly smooth such data because, for example, the averaging of class 3 and class 5 to arrive at class 4 makes no sense. In short, postclassification smoothing algorithms must operate on the basis of logical operations, rather than simple arithmetic computations.

One means of classification smoothing involves the application of a *majority filter*. In such operations a moving window is passed through the classified data set and the majority class within the window is determined. If the center pixel in the window is not the majority class, its identity is changed to the majority class. If there is no majority class in the window, the identity of the center pixel is not changed. As the window progresses through the data set, the original class codes are continually used, not the labels as modified from the previous window positions. (Figure 7.48*b* was prepared in this manner, applying a 3 × 3 pixel majority filter to the data shown in Figure 7.48*a*.)

Majority filters can also incorporate some form of class and/or spatial weighting function. Data may also be smoothed more than once. Certain algorithms can preserve the boundaries between land cover regions and also involve a user-specified minimum area of any given land cover type that will be maintained in the smoothed output.

One way of obtaining smoother classifications is to integrate the types of logical operations described above directly into the classification process. This involves the use of spatial pattern recognition techniques that are sensitive to such factors as image texture and pixel context. Compared to purely spectrally based procedures, these types of classifiers have received only limited attention in remote sensing in the past. However, with the continued improvement in the spatial resolution of remote sensing systems and the increasing computational power of image processing systems, such procedures will likely become more common.

7.14 CLASSIFICATION ACCURACY ASSESSMENT

Another area that is continuing to receive increased attention by remote sensing specialists is that of classification accuracy assessment. Unfortunately, to date the ability to produce digital land cover classifications far exceeds the ability to meaningfully quantify their accuracy. In fact, this problem sometimes precludes the application of automated land cover classification techniques even when their cost compares favorably with more traditional means of data collection. The lesson to be learned here is embodied in the expression "A classification is not complete until its accuracy is assessed."

Classification Error Matrix

One of the most common means of expressing classification accuracy is the preparation of a classification *error matrix* (sometimes called a *confusion matrix* or a *contingency table*). Error matrices compare, on a category-by-category basis, the relationship between known reference data (ground truth) and the corresponding results of an automated classification. Such matrices are square, with the number of rows and columns equal to the number of categories whose classification accuracy is being assessed.

Table 7.3 is an error matrix that an image analyst has prepared to determine how well a classification has categorized a representative subset of pixels used in the training process of a supervised classification. This matrix stems from classifying the sampled training set pixels and listing the known cover

TABLE 7.3 Error Matrix Resulting from Classifying Training Set Pixels

		Training Set Data (Known Cover Types)[a]						
		W	S	F	U	C	H	Row Total
Classification Data	W	480	0	5	0	0	0	485
	S	0	52	0	20	0	0	72
	F	0	0	313	40	0	0	353
	U	0	16	0	126	0	0	142
	C	0	0	0	38	342	79	459
	H	0	0	38	24	60	359	481
	Column total	480	68	356	248	402	438	1992

Producer's Accuracy	**User's Accuracy**
W = 480/480 = 100%	W = 480/485 = 99%
S = 052/068 = 76%	S = 052/072 = 72%
F = 313/356 = 88%	F = 313/353 = 87%
U = 126/248 = 51%	U = 126/142 = 89%
C = 342/402 = 85%	C = 342/459 = 74%
H = 359/438 = 82%	H = 359/481 = 75%

Overall accuracy = (480 + 52 + 313 + 126 + 342 + 359)/1992 = 84%

[a] W, water; S, sand; F, forest; U, urban; C, corn; H, hay.

types used for training (columns) versus the pixels actually classified into each land cover category by the classifier (rows).

Several characteristics about classification performance are expressed by an error matrix. For example, one can study the various classification errors of omission (exclusion) and commission (inclusion). Note in Table 7.3 that the training set pixels that are classified into the proper land cover categories are located along the major diagonal of the error matrix (running from upper left to lower right). All nondiagonal elements of the matrix represent errors of omission or commission. Omission errors correspond to nondiagonal column elements (e.g., 16 pixels that should have been classified as "sand" were omitted from that category). Commission errors are represented by nondiagonal row elements (e.g., 38 "urban" pixels plus 79 "hay" pixels were improperly included in the "corn" category).

Several other descriptive measures can be obtained from the error matrix. For example, the *overall accuracy* is computed by dividing the total number of correctly classified pixels (i.e., the sum of the elements along the major diagonal) by the total number of reference pixels. Likewise, the accuracies of individual categories can be calculated by dividing the number of correctly classified pixels in each category by either the total number of pixels in the corresponding row or column. What are often termed *producer's accuracies* result from dividing the number of correctly classified pixels in each category (on the major diagonal) by the number of training set pixels used for that category (the column total). This figure indicates how well training set pixels of the given cover type are classified.

User's accuracies are computed by dividing the number of correctly classified pixels in each category by the total number of pixels that were classified in that category (the row total). This figure is a measure of commission error and indicates the probability that a pixel classified into a given category actually represents that category on the ground [182].

Note that the error matrix in Table 7.3 indicates an overall accuracy of 84 percent. However, producer's accuracies range from just 51 percent ("urban") to 100 percent ("water") and user's accuracies vary from 72 percent ("sand") to 99 percent ("water"). Furthermore, this error matrix is based on training data. *It should be remembered that such procedures only indicate how well the statistics extracted from these areas can be used to categorize the same areas! If the results are good, it means nothing more than the training areas are homogeneous, the training classes are spectrally separable, and the classification strategy being employed works well in the training areas. This aids in the training set refinement process, but it indicates little about how the classifier performs elsewhere in a scene. One should expect training area accuracies to be overly optimistic, especially if they are derived from limited data sets.* (Nevertheless, training area accuracies are sometimes used in the literature as an indication of overall accuracy. They should not be!)

Sampling Considerations

Test areas are areas of representative, uniform land cover that are different from, and considerably more extensive than, training areas. They are often located during the training stage of supervised classification by intentionally designating more candidate training areas than are actually needed to develop the classification statistics. A subset of these may then be withheld for the postclassification accuracy assessment. The accuracies obtained in these areas represent at least a first approximation to classification performance throughout the scene. However, being homogeneous, test areas might not provide a valid indication of classification accuracy at the individual pixel level of land cover variability.

One way that would appear to ensure adequate accuracy assessment at the pixel level of specificity would be to compare the land cover classification at every pixel in an image with a reference source. While such "wall-to-wall" comparisons may have value in research situations, assembling reference land cover information for an entire project area is expensive and defeats the whole purpose of performing a remote sensing-based classification in the first place.

Random sampling of pixels circumvents the above problems, but it is plagued with its own set of limitations. First, collection of reference data for a large sample of randomly distributed points is often very difficult and costly. For example, travel distance and access to random sites might be prohibitive. Second, the validity of random sampling depends on the ability to precisely register the reference data to the image data. This is often difficult to do. One way to overcome this problem is to sample only pixels whose identity is not influenced by potential registration errors (for example, points at least several pixels away from field boundaries).

Another consideration is making certain that the randomly selected test pixels or areas are geographically representative of the data set under analysis. Simple random sampling tends to undersample small but potentially important areas. Stratified random sampling, where each land cover category may be considered a stratum, is frequently used in such cases. Clearly, the sampling approach appropriate for an agricultural inventory would differ from that of a wetlands mapping activity. Each sample design must account for the area being studied and the cover type being classified.

One common means of accomplishing random sampling is to overlay classified output data with a grid. Test cells within the grid are then selected randomly and groups of pixels within the test cells are evaluated. The cover types present are determined through ground verification (or other reference data) and compared to the classification data.

Several papers have been written about the proper sampling scheme to be used for accuracy assessment under various conditions, and opinions vary among researchers [40, 60, 113]. One suggestion has been the concept of combining both random and systematic sampling [42]. Such a technique may use systematically sampled areas to collect some accuracy assessment data early in a project (perhaps as part of the training area selection process) and random sampling within strata after the classification is complete.

Consideration must also be given to the *sample unit* employed in accuracy assessment. Depending upon the application, the appropriate sample unit might be individual pixels, clusters of pixels, or polygons.

Sample size must also weigh heavily in the development and interpretation of classification accuracy figures. Again, several researchers have published recommendations for choosing the appropriate sample size [40, 77, 162, 196]. However, these techniques primarily produce the sample size of test areas or pixels needed to compute the overall accuracy of a classification or of a single category. In general, they are not appropriate for filling in a classification error matrix wherein errors of omission and commission are of interest.

As a broad guideline, it has been suggested that a minimum of 50 samples of each vegetation or land use category be included in the error matrix. Further, "if the area is especially large (i.e., more than a million acres) or the classification has a large number of vegetation or land use categories (i.e., more than 12 categories), the minimum number of samples should be increased to 75 or 100 samples per category" [42]. Similarly, the number of samples for each category might be adjusted based on the relative importance of that category for a particular application (i.e., more samples taken in more important categories). Also, sampling might be allocated with respect to the variability within each category (i.e., more samples taken in more variable categories such as wetlands and fewer in less variable categories such as open water).

Evaluating Classification Error Matrices

Once accuracy data are collected (either in the form of pixels, clusters of pixels, or polygons) and summarized in an error matrix, they are normally subject to detailed interpretation and further statistical analysis. For example, a number of features are readily apparent from inspection of the error matrix included in Table 7.4 (resulting from randomly sampled test pixels). First, we can begin to appreciate the need for considering overall, producer's, and user's accuracies simultaneously. In this example, the overall accuracy of the classification is 65%. However, if the primary purpose of the classification is to map the locations of the "forest" category, we might note that the producer's accuracy of this class is quite good (84 percent). This would potentially lead one to the conclusion that although the overall accuracy of the classification was poor (65 percent), it is adequate for the purpose of mapping the forest class. The problem with this conclusion is the fact that the user's accuracy for this class is only 60 percent. That is, even though 84 percent of the forested areas have been correctly identified as "forest," only 60 percent of the areas identified as "forest" within the classification are truly of that category. A more careful inspection of the error matrix shows that there is significant confusion between the "forest" and "urban" classes. Accordingly, although the producer of the classification can reasonably claim that 84 percent of the time an area that was forested was identified as such, a user of this classification would find that only 60 percent of the time will an area visited on the ground that the classification says is "forest" actually be "forest." In fact, the only highly reliable category associated with this classification from both a producer's and a user's perspective is "water."

A further point to be made about interpreting classification accuracies is the fact that even a completely random assignment of pixels to classes will produce percentage correct values in the error matrix. In fact, such a random assignment could result in a surprisingly good apparent classification result. The $\hat{k}$ ("KHAT") statistic is a measure of the difference between the actual agreement between reference data and an automated classifier and the chance agreement between the reference data and a random classifier.

Conceptually, $\hat{k}$ can be defined as

$$\hat{k} = \frac{\textit{observed accuracy} - \textit{chance agreement}}{1 - \textit{chance agreement}} \tag{7.10}$$

This statistic serves as an indicator of the extent to which the percentage correct values of an error matrix are due to "true" agreement versus "chance" agreement. As true agreement (observed) approaches 1 and chance agreement approaches 0, $\hat{k}$ approaches 1. This is the ideal case. In reality, $\hat{k}$ usually ranges between 0 and 1. For example, a $\hat{k}$ value of 0.67 can be thought of as an indication that an observed classification is 67 percent better than one resulting from chance. A $\hat{k}$ of 0 suggests that a given classification is no better than a random assignment of pixels. In cases where chance agreement is large enough, $\hat{k}$ can take on negative values—an indication of very poor classification performance. (Because the possible range of negative values depends on the specific matrix, the magnitude of negative values should not be interpreted as an indication of relative classification performance).

The KHAT statistic is computed as

$$\hat{k} = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+} \cdot x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+} \cdot x_{+i})} \tag{7.11}$$

where

r = number of rows in the error matrix

x_{ii} = the number of observations in row i and column i (on the major diagonal)

x_{i+} = total of observations in row i (shown as marginal total to right of the matrix)
x_{+i} = total of observations in column i (shown as marginal total at bottom of the matrix)
N = total number of observations included in matrix

To illustrate the computation of KHAT for the error matrix included in Table 7.4

$$\sum_{i=1}^{r} x_{ii} = 226 + 216 + 360 + 397 + 190 + 219 = 1608$$

$$\begin{aligned}\sum_{i=1}^{r} (x_{i+} \cdot x_{+i}) &= (239 \cdot 233) + (309 \cdot 328) + (599 \cdot 429) \\ &\quad + (521 \cdot 945) + (453 \cdot 238) + (359 \cdot 307) = 1{,}124{,}382\end{aligned}$$

$$\hat{K} = \frac{2480(1608) - 1{,}124{,}382}{(2480)^2 - 1{,}124{,}382} = 0.57$$

Note that the KHAT value (0.57) obtained in the above example is somewhat lower than the overall accuracy (0.65) computed earlier. Differences in these two measures are to be expected in that each incorporates different forms of information from the error matrix. The overall accuracy only includes the data along the major diagonal and excludes the errors of omission and commission. On the other hand, KHAT incorporates the nondiagonal elements of the error matrix as a product of the row and column marginal. Accordingly, it is not possible to give definitive advice as to when each measure should be used in any given application. Normally, it is desirable to compute and analyze both of these values.

One of the principal advantages of computing KHAT is the ability to use this value as a basis for determining the statistical significance of any given matrix or the differences among matrices. For example, one might wish to compare the error matrices resulting from different dates of images, classification techniques, or individuals performing the classification. Such tests are based on computing an estimate of the variance of $\hat{k}$ and then using a Z test to determine if an individual matrix is significantly different from a random result and if $\hat{k}$ values from two separate matrices are significantly different from one another. Readers interested in performing such analyses and learning more about accuracy assessment in general are urged to consult the various references on this subject in the Selected Bibliography [7, 25, 40–43, 54, 160–162, 177, 182].

There are three other facets of classification accuracy assessment that we wish to emphasize before leaving the subject. The first relates to the fact that the quality of any accuracy estimate is only as good as the information used

TABLE 7.4 Error Matrix Resulting from Classifying Randomly Sampled Test Pixels

		Reference Data[a]						
		W	S	F	U	C	H	Row Total
Classification Data	W	226	0	0	12	0	1	239
	S	0	216	0	92	1	0	309
	F	3	0	360	228	3	5	599
	U	2	108	2	397	8	4	521
	C	1	4	48	132	190	78	453
	H	1	0	19	84	36	219	359
	Column total	233	328	429	945	238	307	2480

Producer's Accuracy

W = 226/233 = 97%

S = 216/328 = 66%

F = 360/429 = 84%

U = 397/945 = 42%

C = 190/238 = 80%

H = 219/307 = 71%

User's Accuracy

W = 226/239 = 94%

S = 216/309 = 70%

F = 360/599 = 60%

U = 397/521 = 76%

C = 190/453 = 42%

H = 219/359 = 61%

Overall accuracy = (226 + 216 + 360 + 397 + 190 + 219)/2480 = 65%

[a] W, water; S, sand; F, forest; U, urban; C, corn; H, hay.

to establish the "true" land cover types present in the test sites. To the extent possible, some estimate of the errors present in the reference data should be incorporated into the accuracy assessment process. It is not uncommon to have the accuracy of the reference data influenced by such factors as spatial misregistration, photo interpretation errors, data entry errors, and changes in land cover between the date of the classified image and the date of the reference data. The second point to be made is that the accuracy assessment procedure must be designed to reflect the intended use of the classification. For example, a single pixel misclassified as "wetland" in the midst of a "corn" field might be of little significance in the development of a regional land use plan. However, this same error might be intolerable if the classification forms the basis for land taxation or for enforcement of wetland preservation legislation. Finally, it should be noted that remotely sensed data are normally just a small subset of many possible forms of data resident in a GIS. How errors accumulate through the multiple layers of information in a GIS is the subject of ongoing research [177].