7 Basic Analysis Techniques

This chapter presents the basic analysis techniques needed to perform an accuracy assessment. Once an error matrix has been properly generated, any or all of the following analysis techniques can be performed. These techniques clearly demonstrate why the error matrix is such a powerful tool and should be included in any published accuracy assessment. Without having the error matrix as a starting point, none of these analysis techniques would be possible.

KAPPA

The Kappa analysis is a discrete multivariate technique used in accuracy assessment to statistically determine if one error matrix is significantly different from another (Bishop et al., 1975). The result of performing a Kappa analysis is a KHAT statistic (actually $\hat{K}$, an estimate of Kappa), which is another measure of agreement or accuracy (Cohen, 1960). This measure of agreement is based on the difference between the actual agreement in the error matrix (i.e., the agreement between the remotely sensed classification and the reference data as indicated by the major diagonal) and the chance agreement that is indicated by the row and column totals (i.e., marginals). In this way, the KHAT statistic is similar to the more familiar *chi*-square analysis.

Although this analysis technique has been in the sociology and psychology literature for many years, the method was not introduced to the remote sensing community until 1981 (Congalton, 1981) and not published in a remote sensing journal before Congalton et al. (1983). Since then numerous papers have been published recommending this technique. Consequently, the Kappa analysis has become a standard component of most every accuracy assessment (Congalton et al., 1983; Rosenfield and Fitzpatrick-Lins, 1986; Hudson and Ramm, 1987; Congalton, 1991) and is considered a required component of most image analysis software packages that include accuracy assessment procedures.

The following equations are used for computing the KHAT statistic and its variance.

Let

$$p_o = \sum_{i=1}^{k} p_{ii}$$

be the actual agreement, and

$$p_c = \sum_{i=1}^{k} p_{i+} p_{+j}$$

with p_{i+} and p_{+j} as previously defined, the "chance agreement."

Assuming a *multinomial sampling model,* the maximum likelihood estimate of Kappa is given by

$$\hat{K} = \frac{p_o - p_c}{1 - p_c}.$$

For computational purposes

$$\hat{K} = \frac{n\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i+}n_{+i}}{n^2 - \sum_{i=1}^{k} n_{i+}n_{+i}}$$

with n_{ii}, n_{i+}, and n_{+i} as previously defined.

The approximate large sample variance of Kappa is computed using the Delta method as follows:

$$\hat{\text{var}}(\hat{K}) = \frac{1}{n}\left\{\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2 - \theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4 - 4\theta_2^2)}{(1-\theta_2)^4}\right\}$$

where

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{k} n_{ii},$$

$$\theta_2 = \frac{1}{n^2}\sum_{i=1}^{k} n_{i+}n_{+i},$$

$$\theta_3 = \frac{1}{n^2}\sum_{i=1}^{k} n_{ii}(n_{i+} + n_{+i}),$$

and

$$\theta_4 = \frac{1}{n^3}\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}(n_{j+} + n_{+i})^2.$$

A KHAT value is computed for each error matrix and is a measure of how well the remotely sensed classification agrees with the reference data. Confidence intervals around the KHAT value can be computed using the approximate large sample variance and the fact that the KHAT statistic is asymptotically normally distributed. This fact also provides a means for testing the significance of the KHAT statistic for a single error matrix to determine if the agreement between the remotely sensed classification and the reference data is significantly greater than 0 (i.e., better than a random classification).

It is always satisfying to perform this test on a single matrix and confirm that your classification is meaningful and significantly better than a random classification. If it is not, you know that something has gone terribly wrong during the classification process.

Finally, there is a test to determine if two independent KHAT values, and therefore two error matrices, are significantly different. With this test, it is possible to statistically compare two analysts, the same analyst over time, two algorithms, two types of imagery, or even two dates of imagery and see which produces the higher accuracy. Both the single error matrix and paired error matrix tests of significance rely on the standard normal deviate as follows:

Let $\hat{K}_1$ and $\hat{K}_2$ denote the estimates of the Kappa statistic for error matrix #1 and #2, respectively. Also, let $\hat{\text{var}}(\hat{K}_1)$ and $\hat{\text{var}}(\hat{K}_2)$ be the corresponding estimates of the variance as computed from the appropriate equations. The test statistic for testing the significance of a single error matrix is expressed by:

$$Z = \frac{\hat{K}_1}{\sqrt{\hat{\text{var}}(\hat{K}_1)}}.$$

Z is standardized and normally distributed (i.e., standard normal deviate). Given the null hypothesis $H_O\!:\!K_1 = 0$, and the alternative $H_1 : K_1 \neq 0$, H_0 is rejected if $Z \geq Z_{\alpha/2}$, where $\alpha/2$ is the confidence level of the two-tailed Z test and the degrees of freedom are assumed to be ∞ (infinity).

The test statistic for testing if two independent error matrices are significantly different is expressed by:

$$Z = \frac{\left|\hat{K}_1 - \hat{K}_2\right|}{\sqrt{\hat{\text{var}}(\hat{K}_1) + \hat{\text{var}}(\hat{K}_2)}}.$$

Z is standardized and normally distributed. Given the null hypothesis $H_O\!:\!(K_1 - K_2) = 0$, and the alternative $H_1\!:\!(K_1 - K_2) \neq 0$, H_0 is rejected if $Z \geq Z_{\alpha/2}$.

It is prudent at this point to provide an actual example so that the equations and theory can come alive for the reader. The error matrix presented as an example in Table 7.1 was generated from Landsat Thematic Mapper data using an unsupervised classification approach by analyst #1. A second error matrix was generated using precisely the same imagery and same classification approach; however, the clusters were labeled by analyst #2 (Table 7.2). It is important to note that analyst #2 was not as ambitious as analyst #1, and did not collect as much accuracy assessment data.

Table 7.3 presents the results of the Kappa analysis on the individual error matrices. The KHAT values are a measure of agreement or accuracy. The values can range from +1 to −1. However, since there should be a positive correlation between

TABLE 7.1
Error Matrix Produced Using Landsat Thematic Mapper Imagery and an Unsupervised Classification Approach by Analyst #1

		Reference Data				Row
		D	**C**	**AG**	**SB**	**Total**
Classified Data	**D**	65	4	22	24	115
	C	6	81	5	8	100
	AG	0	11	85	19	115
	SB	4	7	3	90	104
	Column Total	75	103	115	141	434

Land Cover Categories

D = deciduous
C = conifer
AG = agriculture
SB = shrub

OVERALL ACCURACY =
(65+81+85+90)/434 =
321/434 = 74%

PRODUCER'S ACCURACY	USER'S ACCURACY
D = 65/75 = 87%	D = 65/115 = 57%
C = 81/103 = 79%	C = 81/100 = 81%
AG = 85/115 = 74%	AG = 85/115 = 74%
SB = 90/141 = 64%	SB = 90/104 = 87%

the remotely sensed classification and the reference data, positive KHAT values are expected. Landis and Koch (1977) characterized the possible ranges for KHAT into three groupings: a value greater than 0.80 (i.e., >80%) represents strong agreement; a value between 0.40 and 0.80 (i.e., 40–80%) represents moderate agreement; and a value below 0.40 (i.e., <40%) represents poor agreement.

Table 7.3 also presents the variance of the KHAT statistic and the *Z* statistic used to determine if the classification is significantly better than a random result. At the 95% confidence level, the critical value would be 1.96. Therefore, if the absolute value of the test *Z* statistic is greater than 1.96, the result is significant and you would conclude that the classification is better than random. The *Z* statistic values for the two error matrices in Table 7.3 are both 20 or more, and so both classifications are significantly better than random.

Table 7.4 presents the results of the Kappa analysis that compares the error matrices, two at a time, to determine if they are significantly different. This test is based on the standard normal deviate and the fact that although remotely sensed data are discrete, the KHAT statistic is asymptotically normally distributed. The results of this pairwise test for significance between two error matrices reveals that these

TABLE 7.2
Error Matrix Using the Same Imagery and Classification Algorithm as in Table 7.1 Except that the Work Was Performed by a Different Analyst

		Reference Data				Row
		D	**C**	**AG**	**SB**	**Total**
Classified Data	**D**	45	4	12	24	85
	C	6	91	5	8	110
	AG	0	8	55	9	72
	SB	4	7	3	55	69
	Column Total	55	110	75	96	336

Land Cover Categories

D = deciduous
C = conifer
AG = agriculture
SB = shrub

OVERALL ACCURACY =
(45+91+55+55)/336 =
246/336 = 73%

PRODUCER'S ACCURACY

D = 45/55 = 82%
C = 91/110 = 83%
AG= 55/75 = 73%
SB = 55/96 = 57%

USER'S ACCURACY

D = 45/85 = 53%
C = 91/110 = 83%
AG = 55/72 = 76%
SB = 55/69 = 80%

TABLE 7.3
Individual Error Matrix Kappa Analysis Results

Error Matrix	KHAT	Variance	Z Statistic
Analyst #1	0.65	0.0007778	23.4
Analyst #2	0.64	0.0010233	20.0

TABLE 7.4
Kappa Analysis Results for the Pairwise Comparison of the Error Matrices

Pairwise Comparison	Z Statistic
Analyst #1 vs. Analyst #2	0.3087

two matrices are not significantly different. This is not surprising since the overall accuracies were 74 and 73% and the KHAT values were 0.65 and 0.64, respectively. Therefore, it could be concluded that these two analysts may work together because they produce approximately equal classifications. If two different techniques or algorithms were being tested and if they were shown to be not significantly different, then it would be best to use the cheaper, quicker, or more efficient approach.

MARGFIT

In addition to the Kappa analysis, a second technique called Margfit can be applied to "normalize" or standardize the error matrices for comparison purposes. Margfit uses an iterative proportional fitting procedure that forces each row and column (i.e., marginal) in the matrix to sum to a predetermined value; hence the name Margfit (marginal fitting). If the predetermined value is 1, then each cell value is a proportion of 1 and can easily be multiplied by 100 to represent percentages or accuracies. The predetermined value could also be set to 100 to obtain percentages directly, or to any other value the analyst chooses.

In this normalization process, differences in sample sizes used to generate the matrices are eliminated and, therefore, individual cell values within the matrix are directly comparable. In addition, because, as part of the iterative process, the rows and columns are totaled (i.e., marginals), the resulting normalized matrix is more indicative of the off-diagonal cell values (i.e., the errors of omission and commission). In other words, all the values in the matrix are iteratively balanced by row and column, thereby incorporating information from that row and column into each individual cell value. This process then changes the cell values along the major diagonal of the matrix (correct classifications) and, therefore, a normalized overall accuracy can be computed for each matrix by summing the major diagonal and dividing by the total of the entire matrix.

Consequently, one could argue that the normalized accuracy is a better representation of accuracy than is the overall accuracy computed from the original matrix, because it contains information about the off-diagonal cell values. Table 7.5 presents the normalized matrix generated from the original error matrix presented in Table 7.1 (an unsupervised classification of Landsat TM data by analyst #1) using the Margfit procedure. Table 7.6 presents the normalized matrix generated from the original error matrix presented in Table 7.3, which used the same imagery and classifier, but was performed by analyst #2.

In addition to computing a normalized accuracy, the normalized matrix can also be used to directly compare cell values between matrices. For example, we may be interested in comparing the accuracy each analyst obtained for the conifer category. From the original matrices we can see that analyst #1 classified 81 sample units correctly, while analyst #2 classified 91 correctly. Neither of these numbers means much because they are not directly comparable due to the differences in the number of samples used to generate the error matrix by each analyst. Instead, these numbers would need to be converted into percentages or user's and producer's accuracies so that a comparison could be made.

TABLE 7.5
Normalized Error Matrix from Analyst #1

		Reference Data			
		D	C	AG	SB
Classified Data	D	0.7537	0.0261	0.1300	0.0909
	C	0.1226	0.7735	0.0521	0.0517
	AG	0.0090	0.1042	0.7731	0.1133
	SB	0.1147	0.0962	0.0448	0.7440

3.0443

Land Cover Categories

D = deciduous
C = conifer
AG = agriculture
SB = shrub

NORMALIZED ACCURACY =
0.7537+0.7735+0.7731+0.7440 =
3.0443/4.0 = 76%

TABLE 7.6
Normalized Error Matrix from Analyst #2

		Reference Data			
		D	C	AG	SB
Classified Data	D	0.7181	0.0312	0.1025	0.1488
	C	0.1230	0.7607	0.0541	0.0619
	AG	0.0136	0.1017	0.7848	0.0995
	SB	0.1453	0.1064	0.0587	0.6898

2.9534

Land Cover Categories

D = deciduous
C = conifer
AG = agriculture
SB = shrub

NORMALIZED ACCURACY =
0.7181+0.7607+0.7848+0.6898 =
2.9534/4.0 = 74%

TABLE 7.7
Comparison of the Accuracy Values for an Individual Category

Error Matrix	Original Cell Value	Producer's Accuracy	User's Accuracy	Normalized Value
Analyst #1	81	79%	81%	77%
Analyst #2	91	83%	83%	76%

Here, another problem arises: do we divide the total correct by the row total (user's accuracy) or by the column total (producer's accuracy)? We could calculate both and compare the results or we could use the cell value in the normalized matrix. Because of the iterative proportional fitting routine, each cell value in the matrix has been balanced by the other values in its corresponding row and column. This balancing has the effect of incorporating producer's and user's accuracies together. Also, since each row and column adds to one, an individual cell value can quickly be converted to a percentage by multiplying by 100. Therefore, the normalization process provides a convenient way of comparing individual cell values between error matrices regardless of the number of samples used to derive the matrix (Table 7.7).

Table 7.8 provides a comparison of the overall accuracy, the normalized accuracy, and the KHAT statistic for the two analysts. In this particular example, there is agreement among all three measures of accuracy about the relative ranking of the results. However, it is possible for these rankings to disagree simply because each measure incorporates various levels of information from the error matrix into its computations. Overall accuracy only incorporates the major diagonal and excludes the omission and commission errors. As already described, normalized accuracy directly includes the off-diagonal elements (omission and commission errors) because of the iterative proportional fitting procedure. As shown in the KHAT equation, KHAT accuracy indirectly incorporates the off-diagonal elements as a product of the row and column marginals. Therefore, depending on the amount of error included in the matrix, these three measures may not agree.

It is not possible to give clear-cut rules as to when each measure should be used. Each accuracy measure incorporates different information about the error matrix and therefore must be thought of as different computations attempting to explain the error. Our experience has shown that if the error matrix tends to have a great many

TABLE 7.8
Summary of the Three Accuracy Measures for Analyst #1 and #2

Error Matrix	Overall Accuracy	KHAT	Normalized Accuracy
Analyst #1	74%	65%	76%
Analyst #2	73%	64%	74%

off-diagonal cell values with zeros in them, then the normalized results tend to disagree with the overall and Kappa results.

Many zeros occur in a matrix when an insufficient sample has been taken or when the classification is exceptionally good. Because of the iterative proportional fitting routine, these zeros tend to take on positive values in the normalization process, showing that some error could be expected. The normalization process then tends to reduce the accuracy because of these positive values in the off-diagonal cells. If a large number of off-diagonal cells do not contain zeros, then the results of the three measures tend to agree. There are also times when the Kappa measure will disagree with the other two measures. Because of the ease of computing all three measures and because each measure reflects different information contained within the error matrix, we recommend an analysis such as the one performed here to glean as much information from the error matrix as possible.

CONDITIONAL KAPPA

In addition to computing the Kappa coefficient for an entire error matrix, it may be useful to look at the agreement for an individual category within the matrix. Individual category agreement can be tested using the conditional Kappa coefficient. The maximum likelihood estimate of the Kappa coefficient for conditional agreement for the *i*th category is given by

$$\hat{K}_i = \frac{nn_{ii} - n_{i+}n_{+i}}{nn_{i+} - n_{i+}n_{+i}},$$

where n_{i+} and n_{+i} are as previously defined and the approximate large sample variance for the *i*th category is estimated by

$$\hat{\mathrm{var}}(\hat{K}_i) = \frac{n(n_{i+} - n_{ii})}{[n_{i+}(n - n_{+i})]^3}[(n_{i+} - n_{ii})(n_{i+}n_{+i} - nn_{ii}) + nn_{ii}(n - n_{i+} - n_{+i} + n_{ii})].$$

The same comparison tests available for the Kappa coefficient apply to this conditional Kappa for an individual category.

WEIGHTED KAPPA

The Kappa analysis is appropriate when all the errors in the matrix can be considered of equal importance. However, it is easy to imagine a classification scheme in which errors may vary in their importance. In fact, this latter situation is really the more realistic approach. For example, it may be far worse to classify a forested area as water than to classify it as shrub. In this case, the ability to weight the Kappa analysis would be very powerful (Cohen, 1968). The following section describes the procedure to conduct a weighted Kappa analysis.

Let w_{ij} be the weight assigned to the *i, j*th cell in the matrix. This means that the proportion p_{ij} in the *i, j*th cell is to be weighted by w_{ij}. The weights should be restricted to the interval $0 \le w_{ij} \le 1$ for $i \ne j$, and the weights representing the maximum agreement are equal to 1; that is, $w_{ij} = 1$ (Fleiss et al., 1969).

Therefore, let

$$p_o^* = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{ij}$$

be the weighted agreement, and

$$p_c^* = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{i+} p_{+j}$$

where p_{ij}, p_{i+}, and p_{+j} are, as previously defined, the weighted "chance agreement."

Then the weighted Kappa is defined by

$$\hat{K}_w = \frac{p_o^* - p_c^*}{1 - p_c^*}.$$

To compute the large sample variance of the weighted Kappa, define the weighted average of the weights in the ith category of the remotely sensed classification by

$$\bar{w}_{i+} = \sum_{j=1}^{k} w_{ij} p_{+j},$$

where p_{+j} is as previously defined and the weighted average of the weights in the jth category of the reference data set by

$$\bar{w}_{+j} = \sum_{i=1}^{k} w_{ij} p_{i+},$$

where p_{i+} is as previously defined.

The variance may be estimated by

$$\hat{\text{var}}(\hat{K}_w) = \frac{1}{n(1-p_c^*)^4}\left\{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left[w_{ij}(1-p_c^*) - (\bar{w}_{i+} + \bar{w}_{+j})(1-p_o^*)\right]^2 - (p_o^* p_c^* - 2p_c^* + p_o^*)^2\right\}$$

The same tests of significant difference described previously for the Kappa analysis apply to the weighted Kappa. An individual weighted Kappa value can be evaluated to see if the classification is significantly better than random. Two independent weighted Kappas can also be tested to see if they are significantly different.

Although the weighted Kappa has been in the literature since the 1960s and was even suggested to the remote sensing community by Rosenfield and Fitzpatrick-Lins (1986), it has not received widespread attention. The reason for this lack of use is undoubtedly the need to select appropriate weights. Manipulating the weighting scheme can significantly change the results. Therefore, comparisons between different projects using different weighting schemes would be very difficult. The subjectivity of choosing the weights is always hard to justify. Using the unweighted Kappa analysis avoids these problems.

COMPENSATION FOR CHANCE AGREEMENT

Some researchers and scientists have objected to the use of the Kappa coefficient for assessing the accuracy of remotely sensed classifications because the degree of chance agreement may be overestimated (Foody, 1992). Remember from the equation for computing the Kappa coefficient,

$$\hat{K} = \frac{p_o - p_c}{1 - p_c},$$

that p_O is the observed proportion of agreement (i.e., the actual agreement) and p_C is the proportion of agreement that is expected to occur by chance (i.e., the chance agreement). However, in addition to the chance agreement, p_C also includes some actual agreement (Brennan and Prediger, 1981) or agreement for cause (Aickin, 1990). Therefore, since the chance agreement term does not consist solely of chance agreement, the Kappa coefficient may underestimate the classification agreement.

This problem is known to occur when the marginals are free (not fixed *a priori*), which is most often the case with remotely sensed classifications. Foody (1992) presents a number of possible solutions to this problem, including two Kappa-like coefficients that compensate for chance agreement in different ways. Others have suggested additional measures. However, given the very powerful properties of the Kappa coefficient, including the ability to test for significant differences between two independent coefficients, it must still be considered a vital accuracy assessment measure.

CONFIDENCE LIMITS

Confidence intervals are extremely common and are an expected component of any statistical estimate. However, computing confidence intervals for values in an error matrix are significantly more complex than simply computing a confidence interval for a traditional statistical analysis. The following example illustrates the calculations derived from the error matrix (Card, 1982). This example is designed assuming simple random sampling. If another sampling scheme is used, the variance equations change slightly.

TABLE 7.9
Error Matrix Showing Map Marginal Proportions

		True (i) Reference Data					
		D	C	AG	SB	Row Total	Map Marginal Proportions, π_j
Map (j) Classified Data	D	65	4	22	24	115	0.3
	C	6	81	5	8	100	0.4
	AG	0	11	85	19	115	0.1
	SB	4	7	3	90	104	0.2
	Column Total	75	103	115	141	434	OVERALL ACCURACY = (65+81+85+90)/434 = 321/434 = 74%

The same error matrix as in Table 7.1 will be used to compute the confidence intervals. However, the map marginal proportions, π_j, computed as the proportion of the map falling into each map category, are also required (Table 7.9). The map marginal proportions are not derived from the error matrix but are simply the proportion of the total map area falling into each category. These proportions can quickly be obtained by dividing the area of each category by the total map area.

Given this matrix, the first step is to compute the individual cell probabilities using the following equation:

$$\hat{p}_{ij} = \pi_j \; n_{ij}/n_{.j}$$

The individual cell probabilities are simply the map marginal proportion multiplied by the individual cell value all divided by the row marginal. The results of these computations are shown in Table 7.10.

The true marginal proportions, $\hat{p}_i$, can then be computed using the equation:

$$\hat{p}_i = \sum_{j=1}^{r} \pi_j \; n_{ij}/n_{.j}$$

The true marginal proportions can also be computed simply by summing the individual cell probabilities in each column. For example, $\hat{p}_1 = 0.170 + 0.024 + 0.000 + 0.008 = 0.202$, $\hat{p}_2 = 0.357$, $\hat{p}_3 = 0.157$, and $\hat{p}_4 = 0.285$.

The third step is to compute the probability correct given the true class *i;* in other words, the producer's accuracy. It should be noted that the values here differ somewhat from those computed in the error matrix discussion because these values have been corrected for bias by incorporating the true marginal proportions as shown in

TABLE 7.10
Error Matrix of Individual Cell Probabilities, $\hat{p}_{ij}$

		True (i) Reference Data			
		D	**C**	**AG**	**SB**
Map (j) Classified Data	**D**	0.170	0.101	0.057	0.063
	C	0.024	0.324	0.020	0.032
	AG	0.000	0.010	0.074	0.017
	SB	0.008	0.013	0.006	0.0173

the following equation:

$$\hat{\theta}_{ii} = (\pi_i/\hat{p}_i)(n_{ii}/n_{.i}) \quad \text{or} \quad \hat{p}_{ii}/\hat{p}_i$$

As expected, the producer's accuracy is computed taking the diagonal cell value from the cell probability matrix (Table 7.10) and dividing by the true marginal proportion. For example, θ_{11} = 0.170/0.202 = 0.841, or 84%; θ_{22} = 0.908; θ_{33} = 0.471; and θ_{44} = 0.607.

The next step is to compute the probability correct given map class *j;* in other words, the user's accuracy. This computation is made exactly as described in the error matrix discussion by taking the diagonal cell value and dividing by the row (*j*) marginal. The equation for this calculation is as follows:

$$\hat{l}_{jj} = n_{jj}/n_{.j}$$

Therefore, $\hat{l}_{11}$ = 65/115 = 0.565, or 57%; $\hat{l}_{22}$ = 0.810; $\hat{l}_{33}$ = 0.739; and $\hat{l}_{44}$ = 0.865.

Step five is to compute the overall probability correct by summing the major diagonal of the cell probabilities or using the equation:

$$\hat{P}_c = \sum_{j=1}^{r} \pi_j \, n_{jj}/n_{.j}$$

Therefore, in this example, $\hat{P}_c$ = 0.170 + 0.324 + 0.074 + 0.173 = 0.741, or 74%.

We have now made essentially the same calculations as described in the error matrix discussion except that we have corrected for bias by using the true marginal proportions. The next step is to compute the variances for those terms (overall, producer's, and user's accuracies) that we wish to calculate confidence intervals.

Variance for overall accuracy, $\hat{P}_c$

$$V(\hat{P}_c) = \sum_{i=1}^{r} p_{ii}(\pi_i - p_{ii})/(\pi_i n)$$

Therefore, in this example, $\hat{P}_c$ $= [0.170(0.3 - 0.170)/(0.3)(434)$
$+ 0.324(0.4 - 0.324)/(0.4)(434)$
$+ 0.074(0.1 - 0.074)/(0.1)(434)$
$+ 0.173(0.2 - 0.173)/(0.2)(434)]$
$= 0.00040$

Confidence interval for overall accuracy, $\hat{P}_c$

$$\hat{P}_c = 2[V(\hat{P}_c)]^{1/2}$$

Therefore, in this example, the confidence interval for $\hat{P}_c$ $= 0.741 \pm 2(0.0004)^{1/2}$
$= 0.741 \pm 2(0.02)$
$= 0.741 \pm 0.04$
$= (0.701, 0.781)$ or 70% to 78%

Variance for producer's accuracy, $\hat{\theta}_{ii}$

$$V(\hat{\theta}_{ii}) = p_{ii} p_i^{-4} \left[p_{ii} \sum_{j \neq 1}^{r} p_{ij}(\pi_j - p_{ij})/\pi_j n + (\pi_i - p_{ii})(p_i - p_{ii})^2/\pi_i n \right]$$

Therefore, in this example, $V(\hat{\theta}_{11})$ $= 0.170\ (0.202)^{-4}\ \{0.170[0.024(0.4 - 0.024)$
$/(0.4)(434) + 0.008(0.2 - 0.008)/(0.2)(434)]$
$+ (0.3 - 0.170)(0.202 - 0.170)^2\ /(0.3)(434)\}$
$= 0.00132$

Confidence interval for producer's accuracy, $\hat{\theta}_{ii}$

$$\hat{\theta}_{ii} \pm 2[V(\hat{\theta}_{ii})]^{1/2}$$

Therefore, in this example, the confidence interval for $\hat{\theta}_{11}$ $= 0.841 \pm 2(0.00132)^{1/2}$
$= 0.841 \pm 2(0.036)$
$= 0.841 \pm 0.072$
$= (0.768, 0.914)$ or 77% to 91%

Variance for user's accuracy, $\hat{l}_{ii}$

$$V(\hat{l}_{ii}) = p_{ii}(\pi_i - p_{ii})/\pi_i^2 n$$

Therefore, in this example, $V(\hat{l}_{11}) = 0.170(0.3 - 0.170)/(0.3)^2(434)$
$= 0.00057$

Confidence interval for

$$\hat{l}_{ii} \pm 2[V(\hat{l}_{ii})]^{1/2}$$

Therefore, in this example, the confidence interval for $\hat{l}_{11} = 0.565 \pm 2(0.00057)^{1/2}$
$= 0.565 \pm 2(0.024)$
$= 0.741 \pm 0.048$
$= (0.517, 0.613)$ or 52% to 61%

It must be remembered that these confidence intervals are computed from asymptotic variances. If the normality assumption is valid, then these are 95% confidence intervals. If not, then by Chebyshev's inequality, they are at least 75% confidence intervals.

AREA ESTIMATION/CORRECTION

In addition to all the uses of an error matrix already presented, it can also be used to update the areal estimates of the map categories. The map derived from the remotely sensed data is a complete enumeration of the ground. However, the error matrix is an indicator of where misclassification occurred between what the map said is on the ground and what is actually on the ground. Therefore, it is possible to use the information from the error matrix to revise the estimates of total area for each map category. It is not possible to update the map itself or to revise a specific location on the map, but it is possible to revise total area estimates. Updating in this way may be especially important for small, rare categories whose estimates of total area could vary greatly depending on even small misclassification errors.

Czaplewski and Catts (1990) and Czaplewski (1992) have reviewed the use of the error matrix to update the areal estimates of map categories. They propose an informal method, both numerically and graphically, to determine the magnitude of bias introduced in the areal estimates by the misclassification. They also review two methods of statistically calibrating the misclassification bias. The first method is called the classical estimator and was proposed to the statistical community by Grassia and Sundberg (1982) and used in a remotely sensed application by Prisley and Smith (1987) and Hay (1988). The classical estimator uses the probabilities from the omission errors for calibration.

The second method is the inverse estimator, and it uses the probabilities from the commission errors to calibrate the areal estimates. Tenenbein (1972) introduced this technique in the statistical literature, and Chrisman (1982) and Card (1982) have used it for remote sensing applications. The confidence calculations derived in the previous section are from Card's (1982) work using the inverse estimator for calibration. More recently, Woodcock (1996) proposed a modification of the Card approach incorporating fuzzy set theory into the calibration process.

Despite all this work, not many users have picked up on these calibration techniques or the need to perform the calibration. From a practical standpoint, overall total areas are not that important. We have already discussed this in terms of non-site-specific accuracy assessment. However, as more and more work is done with looking at change, and especially changes of small, rare categories, the use of these calibration techniques may gain in importance.