Vol. 23, No. 3, September/septembre 1997

Variance Estimates and Confidence Intervals for the Kappa Measure of Classification Accuracy

by M.A. Kalkhan • R.M. Reich • R.L. Czaplewski

RÉSUMÉ

On utilise souvent l'analyse statistique Kappa pour caractériser les résultats d'une évaluation d'exactitude destinée à évaluer les classifications d'occupation des sols et de couverture terrestre obtenues à partir de données de télédétection. Cette analyse statistique permet de comparer différentes conceptions d'échantillons, les algorithmes de classification, les photointerprétateurs, etc. Pour pouvoir faire ces comparaisons, il est important de savoir quelle peut être l'étendue de l'erreur d'estimation. On y arrive en construisant des intervalles de confiance autour des points d'estimation. La décision d'utiliser des formules de variance asymptotique ou des estimations de variance d'amorçage pour construire des intervalles de confiance pour l'analyse statistique Kappa n'est pas facile. Cette étude avait pour objet d'aider à trouver une réponse à cette question. Neuf matrices d'erreur représentant trois degrés d'exactitude (mauvaise, moyenne et bonne) des données TM du satellite Landsat consistant en 4, 8 et 16 catégories de couverture terrestre en Caroline du Nord ont été utilisées dans cette étude. Chaque matrice d'erreur a été échantillonnée, avec remplacement, à l'aide d'échantillon de 50, 100, 150, 300 et 800 pixels pour obtenir les estimations de l'analyse statistique Kappa et les estimations de variance d'amorçage. Chaque matrice d'erreur d'échantillon a été ré-échantillonnée 500 fois pour obtenir des estimations de variance d'amorçage. La formule de variance asymptotique pour l'analyse statistique Kappa et la variance d'amorçage ont fourni des estimations sans biais de la variance de l'échantillon. En général, les estimations de variance asymptotique étaient plus grandes que celles obtenues à l'aide de l'amorçage, mais la différence était peu importante. Les intervalles de confiance basés sur les centiles de la distribution d'amorçage ont fourni les meilleurs 95 pour cent de taux de couverture (92 à 96 % avec une médiane de 95 %). Les 95 % les plus bas de taux de couverture ont été obtenu à l'aide de l'estimation de variance d'amorçage (médiane de 83 %).

SUMMARY

The Kappa statistic is frequently used to characterize the results of an accuracy assessment used to evaluate land use and land cover classifications obtained by remotely sensed data. This statistic allows comparisons of alternative sampling designs, classification algorithms, photo-interpreters, and so forth. In order to make these comparisons, it is important to know how far in error the estimate might reasonably be. This is accomplished by constructing confidence intervals around the point estimate. The decision to use either the asymptotic variance formulae or bootstrapping variance estimates in constructing confidence intervals for the Kappa statistic is not a simple task. This study was designed to help answer this question. Nine error matrices representing three levels of accuracy (poor, average, and good) of Landsat TM Data consisting of 4, 8 and 16 land cover types in North Carolina were used in this study. Each error matrix was sampled, with replacement, using sample sizes of 50, 100, 150, 300 and 800 pixels to obtain estimates of the Kappa statistic and sample variance. Each of the sample error matrices were resampled 500 times to obtain bootstrap estimates of the variance. The asymptotic variance formula for the Kappa statistic and bootstrap variance provided unbiased estimates of the sample variance. In general, the asymptotic variance estimates were larger than those obtained using bootstrapping, even though the differences were not significant. Confidence intervals based on percentiles of the bootstrap distribution provided the best 95 percent coverage rates (92 to 96 percent with a median of 95 percent). The lowest 95 percent coverage rates were obtained using the bootstrap variance estimate (median of 83 percent).

• M.A. Kalkhan, Ph.D is with the Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA

• Robin M. Reich, Ph.D is with the Department of Forest Sciences Colorado State University ,Fort Collins, CO 80523, USA;

• Raymond L. Ph.D Czaplewski is with the USDA Forest Service Rocky Mountain Forest and Range Experiment Station, 240 W Prospect Road, Fort Collins, CO 80526, USA

INTRODUCTION

The most common index used to assess the accuracy of remotely sensed data is the Kappa statistic. The Kappa statistic, which was originally developed to measure observer agreement for categorical data (Cohen 1960) has received considerable attention in remote sensing applications (Congalton and Mead 1983; Rosenfield and Fitzpatrick-Lins 1986; Aickin 1990; Congalton 1991; Stehman 1992; Fitzgerald and Lees 1994; and Kalkhan 1994). The Kappa statistic is defined as (Bishop *et al.* 1975, p. 395-400):

$$K = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} x_{i+}x_{+i}}{N^2 - \sum_{i=1}^{r} x_{i+}x_{+i}} \quad (1)$$

where r is the number of rows in the error matrix, x_{ii} is the number of observations in row i and column i (i.e. the diagonal elements), x_{i+} and x_{+i} are marginal totals of row i and column i, respectively, and N is the total number of observations.

If there is perfect agreement between categories, $K = 1$, while a value of $K = 0$ indicates that the observed agreement equals chance agreement (Cohen 1960). Skidmore and Turner (1989) point out that positive values of Kappa occur from greater than chance agreement, while negative values indicates a less than chance agreement. The lower limit of the Kappa statistic depends on the marginal distributions and is not likely to have practical interest (Rosenfield and Fitzpatrick-Lins 1986). Using criteria developed by Landis and Kock (1977), Monserud and Leemans (1992) suggest that a value of Kappa greater than 0.75 indicates very good to excellent agreement, while a value between 0.4 and 0.75 indicates fair to good agreement. A value less than or equal to 0.4 indicates poor agreement between classification categories.

The asymptotic variance of the Kappa statistic is given by (Bishop *et al.* 1975, p.396):

$$\sigma_K^2 = \frac{1}{N}\left(\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2 - \theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4 - 4\theta_2^2)}{(1-\theta_2)^4}\right) \quad (2)$$

where,

$$\theta_1 = \sum_{i=1}^{r} P_{ii}, \qquad \theta_2 = \sum_{i=1}^{r} P_{i+}P_{+i},$$

$$\theta_3 = \sum_{i=1}^{r} P_{ii}(P_{i+} + P_{+i}), \qquad \theta_4 = \sum_{i=1}^{r}\sum_{j=1}^{r} P_{ij}(P_{i+} + P_{+j})^2,$$

$$P_{ii} = \frac{x_{ii}}{N}, \qquad P_{i+} = \frac{x_{i+}}{N},$$

$$P_{+i} = \frac{x_{+i}}{N}, \qquad P_{ij} = \frac{x_{ij+}}{N},$$

In applying the Kappa statistic, it is assumed that one is sampling from a multinomial distribution in which each sampling unit is classified into a mutually exclusive category (Bishop *et al.* 1975; Congalton 1991; Stehman 1992; Kalkhan 1994). With a large sample size, the normal approximation holds reasonably well in terms of constructing confidence intervals and hypothesis testing. With small sample sizes this may not be true.

To overcome this potential problem associated with small sample sizes, one can use bootstrapping (Efron 1979) to obtain unbiased estimates of the sample variance. This is accomplished by generating B bootstrap error matrices, each consisting of N data values drawn with replacement from an error matrix consisting of N objects. For each bootstrap replication of the error matrix, the Kappa statistic is computed using Equation 1. The variance of the Kappa statistic is obtained using the empirical variance of the B bootstrap replication (Efron and Tibshirani 1993).

In general, bootstrap variance estimates may provide better estimates of the variance than those obtained from using the asymptotic variance formula (Equation 2). The estimated variances are often used to assign approximate confidence intervals to a parameter of interest. If the distribution of the Kappa statistic is non-normal, the use of the standard normal distribution in constructing confidence intervals may not adjust the confidence interval to account for skewness in the underlying distribution, or other errors that can result when estimating the Kappa statistic (Efron and Tibshirani 1993). As an alternative, one could use the percentiles of the bootstrap histogram to define confidence intervals without having to make normal theory assumption. If the bootstrap distribution of the Kappa statistic is roughly normal, then the standard normal and percentile interval will generally agree. If the bootstrap distribution is non-normal, then the percentile interval should achieve a better balance in the left and right tails, since one is using more of the information in the bootstrap histogram than just its standard deviation (Efron and Tibshirani 1993).

Previous studies aimed at evaluating the statistical properties of the Kappa statistic in assessing the accuracy of remotely sensed imagery have focused primarily on the bias of the statistic (Stehman 1992). Little or no attention has been given to the statistical properties of the variance estimator and accompanying confidence interval. Thus, the objective of this study is to evaluate the use of bootstrapping to obtain confidence intervals and variance estimates of the Kappa statistic in assessing the accuracy of remotely sensed data.

METHODS

Study Area

The study area is located in the State of North Carolina. North Carolina was selected because of its diversity in physiographic regions, representing land cover conditions commonly found in the eastern and southern United States. Elevations range from sea level to over 2,000 meters. Kuchler (1985) stated that potential climax vegetation for most of the entire state is

Appalachian oak forest. Mountainous areas include forest species commonly found in more northern latitudes. The flat coastal plain includes sand ridges, bays pocosins, and maritime forests. Relative to other temperate forests, vegetation in North Carolina is very diverse. Czaplewski *et al.* (1987) provides a more complete description of the study area.

Data

The data used in this study were from a pilot study designed to evaluate the use of large permanent sample plots and Landsat TM data to monitor short-term land cover and land use changes at the state and regional level (Schreuder *et al.* 1986). Sample plots were located on a systematic grid where each node of the grid corresponded to the approximate center of alternate rows of a 7.5 minute USGS topographic quadrangle, for a total 411 permanent sample plots each with an area of 405 ha (4500 pixels). Sample plots were classified with respect to their land cover and land use using basic land cover categories (Level I) adopted by the Common Terminology Work Group (Powell 1981). Forest categories which represent Level II classifications were developed in association with the Southern Forest Inventory Analysis Unit to ensure the availability of pertinent forest wide information.

Image Analysis and Classification

A Landsat TM scene (October 8, 1985, Identification number Y505061522XO) near Raleigh, North Carolina was selected for this study. The scene contained 35 sample plots. Sixteen cover types were identified using an unsupervised classification procedure available in *ERDAS Software* (version 7.5, 1992) using bands TM3 (red), TM4 (IR1), and TM5 (IR2) (**Table 1**). Of the 35 plots, four were selected to represent images with a poor classification and four were selected to represent an average classification. Criteria used in selecting the sample plots were based on Cramer's V (Bishop *et al.* 1985, p. 386) coefficient of agreement. Plots selected to represent an image with a poor classification had a Cramer's V ranging from 0.059 to 0.178, while the plots representing an average classification had a Cramer's V ranging from 0.355 to 0.460. The interpretation of Cramer's V is similar to that of the Kappa statistic, but Cramer's V is not as widely used in the remote sensing literature.

The sample plots representing the poor and average classifications were combined to form a 16x16 composite error matrix representing the overall accuracy of each group of plots. Each error matrix consisted of a series of rows and columns representing the cover types identified on remotely sensed imagery (columns) and the ground (rows). The error matrix provides the users with information on the accuracy of identifying individual cover types and both errors of commission and omission in the classification (Rosenfield and Fitzpatrick-Lins 1986). Errors of commission relate to the accuracy of the aerial photographs while errors of omission represent the accuracy of the remotely sensed data. The error matrices were then collapsed to form a set of 8x8 and 4x4 error matrices. The collapsed error matrices were created by combining classes with similar class signatures (**Table 1**).

Because of the low accuracy associated with the sample data, we developed a third set of error matrices representing a good level of agreement in the cover type maps. This was accomplished by building up the diagonal of the individual error matrices. In building up the diagonals we reduced the cell counts of the off-diagonal cells associated with obvious misclassification errors such as grassland and pine sawtimber, while cell counts associated with

Table 1.
Initial and reduced classification scheme used to describe the land cover and forest types in North Carolina.

Initial 16 classes	Reduced 8 classes	Reduced 4 classes
1. Idle agriculture, crop land, and nonstocked forest land	Idle agriculture, crop land, nonstocked forest and seedlings/saplings (1,13,14,15,16)	Nonforested, seedling/Saplings (1, 2, 13, 14, 15, 16)
2. Grassland		Urban (3)
3. Urban	Grassland (2)	Pine and mixed sawtimber and pole timber (4,5,9)
4. Pine sawtimber	Urban (3)	
5. Mixed sawtimber	Pine sawtimber (4)	Oak-hardwood sawtimber and pole timber (6,7,8,10,11,12)
6. Pine sawtimber	Mixed sawtimber (5)	
7. Bottom land sawtimber	Oak-pine sawtimber (6)	
8. Upland sawtimber	Hardwood sawtimber (7,8)	
9. Pine and mixed pole timber	Pole timber (9,10,11,12)	
10. Oak-pine pole timber		
11. Bottom land pole timber		
12. Upland pole timber		
13. Pine and mixed seedling/saplings		
14. Oak seedling/saplings		
15. Bottom land seedlings/saplings		
16. Upland seedlings/saplings		

less obvious misclassification errors such as pine sawtimber and pine pole timber were left unchanged. Finally, all error matrices were converted to joint probability matrices for the purpose of sampling. This was accomplished by dividing the individual cell counts by the total number of sample plots.

Data Analysis

Each of the nine composite error matrices (3 levels of accuracy × 3 levels of classes) were sampled n = 50, 100, 150, 300, and 800 times using a two-step process to approximate an equal probability sample. First, a cover type on the remote sensing image was selected with probability proportional to their marginal probabilities (columns). The corresponding ground classification was then selected with probability proportional to the conditional probability of observing a particular cover type given the remote sensing classification obtained in the first step. This was repeated n times to obtain an estimate of the error matrix which was then used to obtain estimates of the Kappa statistic (Equation 1), its variance (Equation 2) and 95% confidence interval assuming a normal distribution. This process was repated 500 times for a total of 22,500 estimated error matrices.

Each of the estimated 22,500 error matrices were resampled B = 500 times using a sample of size n to obtain 500 estimates of the Kappa statistic. The variance of the 500 estimates of the Kappa statistic became the bootstrap variance estimate of the Kappa statistic. The bootstrap variance estimates were used to construct 95% confidence intervals around the estimated Kappa statistic, assuming a normal distribution. Bootstrap percentile intervals (0.025, 0.975) were also computed from the bootstrap histogram of the Kappa statistics.

The bias for the Kappa statistic was computed as the difference between the average of the 500 estimates of the Kappa statistic and their true values. A t-test was used to test the null hypothesis of no significant bias at the 0.05 level of significance. In testing this hypothesis the variance of the 500 estimates of the Kappa statistic (i.e. simulation variance) was used as an estimate of the sample variance.

In addition to knowing whether an estimate is unbiased, it is desirable to know something about the distribution of the sample statistic. The Shapiro and Wilk's (1965) W test was used to test the null hypothesis that the Kappa statistic is normally distributed at the 0.05 level of significance. Randomly selected bootstrap distributions were also tested for normality. The selected bootstrap distributions were compared to the sample distribution of the Kappa statistic using the two sample Kolmogorov-Smirnov goodness-of-fit statistic.

The bias associated with estimating the sample variance was evaluated by computing the ratio of the mean variance to the simulation variance. The mean variance of the asymptotic variance (Equation 2) and bootstrap variance were computed by averaging the 500 estimated variances. The ratio of the asymptotic variance to the bootstrap variance was also computed. An F-test was used to test for significant differences in the variance estimates at the 0.05 level of significance. Finally, we determined the proportion of confidence intervals computed using the classical variance, bootstrap variance, and bootstrap percentile intervals that enclosed the true Kappa.

Results and Discussion

The nine error matrices had Kappa statistics ranging from a low of 0.35 percent to a high of 82.06 percent (**Table 2**). In general,

Table 2.

Influence of sample size, accuracy of the remotely sensed image, and number of cover types in estimating the Kappa statistic. The numbers in parenthases are the standard deviations associated with individual estimates of Kappa. Estimates of the Kappa statistic and standard deviations were based on 500 Monte Carlo simulations.

			Average Estimated Kappa (%) (standard deviation (%))									
No. Classes	Level of Agreement	True Kappa	Sample Size									
		(%)	50		100		150		300		800	
4	Poor	0.35	0.42	(9.53)	0.20*	(6.77)	0.34	(5.08)	0.16*	(3.66)	0.41*	(2.30)
	Average	3571%	35.38*	(9.28)	35.03*	(6.39)	35.23*	(5.29)	35.76	(3.72)	35.76*	(2.23)
	Good	82.06	81.56*	(6.90)	82.02	(4.68)	82.03	(3.83)	82.25*	(2.78)	81.97*	(1.60)
8	Poor	5.17	5.42*	(7.69)	5.19	(5.34)	5.17	(4.10)	5.35*	(2.42)	5.18	(1.83)
	Average	30.84	30.62*	(8.00)	30.65*	(5.48)	31.01*	(4.40)	30.89	(3.08)	30.81	(2.08)
	Good	77.83	77.75	(6.76)	77.77	(4.67)	77.70*	(3.85)	77.89*	(2.77)	77.75*	(1.57)
16	Poor	1.54	1.60	(4.24)	1.53	(2.47)	1.39*	(2.38)	1.42*	(1.82)	1.51*	(1.02)
	Average	20.39	20.14*	(6.18)	20.53*	(4.45)	20.15*	(3.56)	20.31*	(2.73)	20.27*	(1.57)
	Good	74.28	74.01*	(5.99)	74.43*	(5.03)	74.09*	(3.98)	74.35*	(2.83)	74.10*	(1.68)

* The estimated Kappa differed significantly from the true Kappa based on a two-tailed t-test with 499 degrees of freedom at the 0.05 level.

the sample estimates of the Kappa statistic were biased, though there was no consistency in the direction of this bias. There was also no relationship between the sample size and number of classes and whether an estimate was biased or not. There was however, a tendency for the error matrices representing a remote sensing image with an average and good level of accuracy to have a higher proportion of biased estimates than the error matrices representing a poor level of accuracy (**Table 2**). While these biases may not seem large from a practical point of view, they can have a direct impact on the observed coverage rates by distorting the confidence probabilities (Cochran 1977, p. 12-15). In a similar study, Stehman (1992) observed little to no bias associated with the Kappa statistic in estimating the accuracy of remote sensing images. However, Stehman (1992) only considered error matrices with three classes and Kappa statistics ranging from 45 to 68 percent. Because of the differences in the number of classes and the ranges in the level of accuracy, the results of these two studies may not be directly comparable.

The asymptotic variance formulae (Equation 2), for the most part, provided unbiased estimates of the sample variance (**Table 3**). Exceptions to this occurred when the level of accuracy associated with the population error matrix had a Kappa statistic less than 1.6 percent. In these instances, the asymptotic variance formulae significantly overestimated the sample variance, irrespective of the sample size or number of classes associated with the error matrix. Tests of normality indicated that 38 out of 45 sample distributions of the Kappa statistic were normally distributed (**Table 3**). In the few cases in which the hypothesis of normality was rejected, the sample distributions were skewed to the right, and the asymptotic variance formulae tended to slightly underestimate the sample variance.

The bootstrap variance estimate, in contrast, provided unbiased estimates of the sample variance, irrespective of the sample size, level of accuracy, or the number of classes associated with the population error matrix (**Table 4**). There was a general tendency of the bootstrap variance estimate to underestimate the sample variance. A few of the bootstrap distributions did not follow a normal distribution (**Table 4**). In these cases, the bootstrap distributions were skewed to the left, just the opposite of what was observed for the sample distributions of the Kappa statistic.

The asymptotic variance estimates were significantly larger than the bootstrap variance estimate for population error matrices with a Kappa statistic less than 1.6 percent. In general though, the asymptotic variance estimates tended to be larger than those obtained using bootstrapping, even though these differences were not significant. This suggests that the bootstrap distributions have less variability associated with them compared to the original error matrices. When we compared the cumulative density function of the sample estimates of the Kappa statistic with those obtained from the bootstrap procedure, 41 out of 45 were significantly different at the 0.05 level. Comparing the individual distributions, we noticed that the sample distributions were slightly skewed to

Table 3.
Ratio of the asymptotic variance to the simulation variance of the Kappa statistic.

No. Classes	Level of Agreement	Sample Size 50	100	150	300	800
4	Poor	1.26*	1.18	1.39*	1.32	1.24
	Average	1.02	1.17	1.04	1.05	1.09
	Good	0.93#	1.00#	0.99	0.93	1.07
8	Poor	1.08	1.06	1.18	1.16	1.10
	Average	1.00	1.06#	1.09	1.12	0.91
	Good	0.95#	1.00#	0.98	0.95	1.11
16	Poor	1.58*#	1.61*	1.53	1.27*	1.52
	Average	1.17	1.13	1.16	0.99	1.12
	Good	1.07	0.88	1.00	0.94	1.00#

* The asymptotic variance differed significantly from the simulation variance using a two-tailed F-test with 499 degrees of freedom in the numerator and denominator at the 0.05 level.
# The sample distribution of the Kappa statistic differed significantly from a normal distribution using the Shapiro-Wilk's (1965) *W* test at the 0.05 level.

Table 4.
Ratio of the mean bootstrap variance to the simulation variance of the Kappa statistic.

No. Classes	Level of Agreement	Sample Size 50	100	150	300	800
4	Poor	0.85	0.87	1.03	0.99	0.96
	Average	0.88	0.97	0.94	0.94	0.99
	Good	0.95	1.00#	0.99	0.93	1.07
8	Poor	0.85	0.92	1.06	1.05	1.02
	Average	0.92	1.00	1.04	1.07	0.88
	Good	0.92	1.01	0.99	0.96	1.11
16	Poor	0.92#	1.05#	1.03	0.88	1.08
	Average	0.98#	0.99	1.02	0.89	1.00
	Good	1.09	0.89	1.00	0.94	1.00

* The bootstrap variance differed significantly from the simulation variance using a two-tailed F-test with 499 degrees of freedom in the numerator and denominator at the 0.05 level.
# The bootstrap distribution of the Kappa statistic differed significantly from a normal distribution using the Shapiro-Wilk's (1965) *W* test at the 0.05 level.

Table 5.
Ninety -five percent confidence coverage rates for the Kappa statistic using asymptotic variance under the assumption of normality.

No. Classes	Level of Agreement	Sample Size 50	100	150	300	800
4	Poor	0.97	0.96	0.98	0.97	0.96
	Average	0.96	0.95	0.95	0.96	0.96
	Good	0.88	0.89	0.90	0.85	0.85
8	Poor	0.95	0.95	0.98	0.96	0.95
	Average	0.94	0.96	0.96	0.97	0.94
	Good	0.94	0.95	0.94	0.94	0.96
16	Poor	0.98	0.98	0.97	0.97	0.98
	Average	0.94	0.95	0.97	0.97	0.96
	Good	0.94	0.93	0.95	0.93	0.95

the right, while the bootstrap distributions were skewed somewhat to the left. Also, the bootstrap distributions were flatter than the sample distributions which may account for the slightly smaller variance estimates.

Ninety-five percent confidence coverage rates for the estimated Kappa statistic using the asymptotic variance formulae ranged from 88 to 96 percent (**Table 6**). The median coverage rate was 95 percent. **Figure 1A** depicts the frequency distribution of the coverage rates obtained using the asymptotic variance. Error matrices with a low level of accuracy generally had coverage rates larger than expected, while error matrices with a high level

Table 6.
Ninety-five percent confidence coverage rates for the Kappa statistic using bootstrap variance estimates under the assumption of normality.

No. Classes	Level of Agreement	Sample Size 50	100	150	300	800
4	Poor	0.81	0.80	0.84	0.83	0.83
	Average	0.82	0.83	0.82	0.83	0.85
	Good	0.82	0.79	0.84	0.86	0.84
8	Poor	0.84	0.83	0.84	0.86	0.83
	Average	0.85	0.84	0.86	0.83	0.85
	Good	0.82	0.83	0.83	0.87	0.85
16	Poor	0.82	0.86	0.83	0.84	0.86
	Average	0.82	0.84	0.84	0.85	0.83
	Good	0.83	0.80	0.84	0.80	0.83

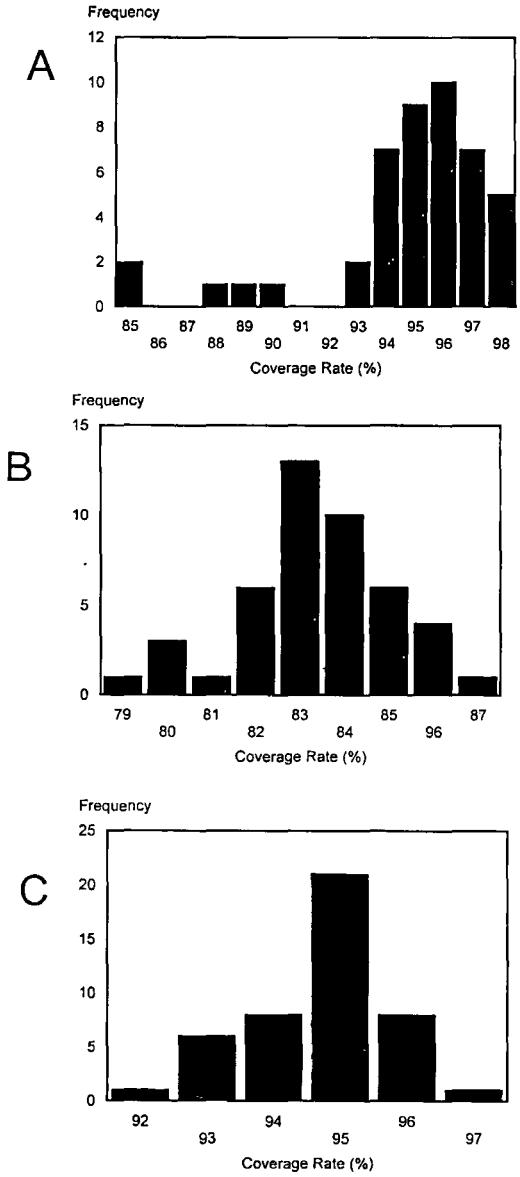

Figure 1.
Frequency distribution of the 95 percent coverage rates obtained using (A) the asymptotic variance formulae, (B) the bootstrap variance estimate, and (C) the percentiles of the bootstrap distribution.

of accuracy had coverage rates lower than expected. The larger than expected coverage rates observed for error matrices with a low level of accuracy was directly attributable to the over estimation of the sample variance (**Table 3**). Similarly, the lower than expected coverage rates associated with error matrices with a high level of accuracy was due to the tendency of the asymptotic variance formulae to underestimate the sample variance and the bias associated with the sample statistic. Both of these factors can result in a distortion in the observed confidence probabilities.

In contrast, the coverage rates using the bootstrap variance estimates ranged from 80 to 87 percent, with a median coverage rate of 83 percent (**Table 6, Figure 1B**). The lower than

expected coverage rates may be due in part to the general tendency of the bootstrapping procedure to underestimate the sample variance (**Table 4**). There is some anecdotal evidence to suggest that the underestimation of the sample variance may be due to the sampling procedure used in obtaining the bootstrap sample (Schreuder 1996, personal communications). Alternative sampling procedures may improve the variance estimates. The biases associated with estimating the Kappa statistic also may be a contributing factor to the lower than expected coverage rates.

Confidence intervals based on the percentiles of the bootstrap distribution provided the best coverage rates (92 to 96 percent with a median of 95 percent) of the three procedures evaluated in this study (**Table 7, Figure 1C**). The percentiles of the bootstrap distribution obtained a better balance in the left and right tails since they use more information than just an estimate of the variability associated with the distribution.

Table 7.
Ninety-five percent confidence coverage rates for the Kappa statistic based on the percentiles of the bootstrap distribution.

No. Classes	Level of Agreement	Sample Size 50	100	150	300	800
4	Poor	0.93	0.95	0.95	0.95	0.95
	Average	0.95	0.95	0.95	0.95	0.95
	Good	0.95	0.94	0.94	0.95	0.96
8	Poor	0.95	0.94	0.93	0.96	0.95
	Average	0.94	0.96	0.93	0.95	0.95
	Good	0.93	0.96	0.93	0.96	0.95
16	Poor	0.94	0.95	0.95	0.96	0.95
	Average	0.92	0.94	0.93	0.96	0.95
	Good	0.94	0.94	0.97	0.96	0.95

Recommendations

Based on the results of this study, we recommend the following for estimating the sample variance and constructing confidence intervals for the Kappa statistic:

1. Use the bootstrap variance estimate to obtain unbiased estimates of the sample variance for purposes of hypothesis testing. We do not recommend using the bootstrap variance estimates for constructing confidence intervals.
2. Use the asymptotic variance formulae for constructing confidence intervals assuming a normal distribution.
3. Percentiles of the bootstrap distribution should be used in constructing confidence intervals for the estimated Kappa statistic.

References

Aickin, M. 1990. "Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's Kappa", Biometrics, 46:293-302.

Bishop, Y. M. M., S. E. Feinberg, and P. W. Hooland. 1975. " Discrete multivariate analysis theory and practice", MIT Press, Cambridge, MA, 575 p.

Cochran, W. G. 1977. "Sampling techniques", 3rd ed. John Wiley and Sons, New York, 428 p.

Cohen, J. 1960. "A coefficient of agreement of nominal scales", Education Psychological Measurements, 20:37 46.

Congalton, R. G., and R. A. Mead. 1983. "A quantitative method to test for consistency and correctness in photointerpretation", Photogrammtric Engineering and Remote Sensing, 49:69 74.

Congalton, R. G. 1991. "A review of assessing the accuracy of classification of remotely sensed data" Remote Sensing of Environment, 37:35-46.

Czaplewski, R. L., G. L. Catts, and P. W. Snook. 1987. "National land cover monitoring using large, permanent photo plots", In: Land and Resource Evaluation for National Planning in the Tropics. An International Conference and Workshop, Chetumal Mexico, January 25-31, 1987, 524 p.

Efron, B. 1979. "Bootstrap methods: Another look at the jackknife", Annals of Statistic, 7:1-26.

Efron, B. And R. J. Tibshirani. 1993. "An introduction to the bootstrap", Chapman & Hall, 431 p.

ERDAS, Inc. 1992. 2081 Buford Highway, Suite 300, Atlanat, Georgia 30329, USA.

Fitzgerald, R. W. and B. G. Lees. 1994. "Assessing the classification accuracy of multisource remote sensing data", Remote Sensing of Enviroment, 47:362-368.

Kalkhan, M. A. 1994. "Statistical properties of six accuracy indices using simple random and stratified random sampling: An application in remote sensing", Ph.D. Dissertation. Colorado State University, 134 p.

Kuchler, A. W. 1985. "Potential natural vegetation", National Atlas of the United States. U.S. Department of Interior, U.S. Geological Survey, (1 map sheet).

Landis, J. R., and G. G. Kock. 1977. "The measurement of observer agreement for categorical data", Biometrics, 33:159-174.

Monserud, R. A. and R. Leemans. 1992. "Comparing global vegetation maps with the Kappa statistic", Ecological Modeling, 62:275-293.

Powell, D. S. 1981. "A solution for solving conflicts between land use and land cover inventories", P. 665-668. In: Place Resource Inventories, Principles and Practices, National SAF Workshop, August 9-14, 1981.

Rosenfield, G. H. and K. Fitzpatrick Lins. 1986. "A coefficient of agreement as a measure of thematic classification accuracy", Photogrammtric Engineering and Remote Sensing, 52:223 227.

Schreuder, H. T. 1996. Rocky Mountain Forest and Range Experiment Station, USDA Forest Service, 240 West Prospect Street, Fort Collins, CO 80526.

Schreuder, H. T., P. W. Snook, R. L. Czaplewski, and G. P. Catts. 1986. "A proposed periodic national inventory of land-use land cover", National American Society of Photogrammetry Conference, September 1986, Anchorage, AK.

Shapiro, S. S. And M. B. Wilk. 1965. "Analysis of variance test for normality (complete test)", Biometrica, 52:591-611.

Skidmore, A., and B. Turner. 1989. "Assessing the accuracy of resource inventory maps", p. 524-535 In: Proc. of Global Natural Resource Monitoring and Assessment: Preparing for the 21st Century, Venice, Italy.

Stehman, S. V. 1992. "Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data", Photogrammtric Engineering and Remote Sensing, 58:1343-1350.