

Statistical Properties of Measures of Association and the Kappa Statistic for Assessing the Accuracy of Remotely Sensed Data Using Double Sampling

Mohammed A. Kalkhan¹, Robin M. Reich², and
Raymond L. Czaplewski³

Abstract.--A Monte Carlo simulation was used to evaluate the statistical properties of measures of association and the Kappa statistic under double sampling with replacement. Three error matrices representing three levels of classification accuracy of Landsat TM Data consisting of four forest cover types in North Carolina. The overall accuracy of the five indices ranged from 0.35% to 82.1% depending on the number of classes, the level of classification accuracy of satellite imagery, and the simulated sample sizes of reference plots. Statistical criteria used in the evaluation included: percent bias, mean squared error, relative error, ratio of the mean variance to the simulation variance, and 95% confidence coverage rates. Results of the simulation indicated that double sampling provided unbiased estimates of the overall accuracy of remotely sensed imagery irrespective of the number of classes in the image being analyzed, or sample size. While no one index was superior for all levels of accuracy, number of classes, or sample size, the Kappa statistic and Pearson's P provided the best estimates of the overall accuracy of the remotely sensed images. Results from previous studies suggest that increasing the sample size, or reducing the number of classification in the remotely sensed image may increase the accuracy and precision of the estimates. However, this was not the case in this study because the low accuracy of the aerial photos to ground data.

¹ Biometrician, Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA.

² Associate Professor, Department of Forest Sciences, Colorado State University, Fort Collins, CO 80523, USA.

³ USDA Forest Service Research, Rocky Mountain Forest and Experiment Station, Fort Collins, CO 80526, USA.

INTRODUCTION

High resolution satellite data (*HRSD*), such as Landsat Thematic Mapper (TM) Data, the Systeme Probatoire D'Observation de la Terre (SPOT-1,2: France) satellite imagery, and JERS-1,OPS (Japan) provide natural resource managers with vital information for mapping, monitoring, and assessing the condition of their resources (Kalkhan 1994). Manager are also concerned with assessing the accuracy of their thematic maps. The high cost of collecting ground data limits the sample sizes used in assessing the accuracy of a thematic map. To reduce this cost, one could use aerial photographs. Aerial photographs are less expensive than ground data, but their degree of misclassification tends to be higher (Kalkhan *et al.* 1995b). As an alternative, some combination of aerial photography and ground data (i.e., double sampling) could be used to improve the accuracy assessment of the remotely sensed data at a lower cost.

Measures of Association and the Kappa Statistic

Indices developed by Turk (1979), Hellden (1980), and Short (1982) estimate the accuracy of individual categories of remotely sensed data based on the errors of commission and omission associated with the estimated error matrix. A major disadvantage of these indices is that they tend to underestimate the accuracy (Rosenfield and Fitzpatrick-Lins 1986). Another disadvantage is that no valid variance formula has been developed for these indices, making it difficult to evaluate the precision of the estimates (Kalkhan 1994). Because of this, these indices are not often used to assess the accuracy of the remotely sensed images. In addition, users of remotely sensed images are more interested in evaluating the overall accuracy of their thematic map not just individual categories (Kalkhan 1994). As an alternative, users of remotely sensed images may prefer to use indices such as Pearson's coefficient of mean square contingency (Q^2), Pearson's P, Tschuprow's T, or Cramer's V, for assessing accuracy. These latter indices are related to the Chi-square statistic and have more desirable statistical properties than the other indices (Kalkhan 1994). In addition, Bishop *et al.* (1975, p 385-386), Kalkhan (1994), and Kalkhan *et al.* (1995a) provide detailed information about the measures of association and evaluations of their statistical properties.

The Kappa statistic (K) was originally developed by Cohen (1960) to measure the observed agreement for categorical data (Landis and Kock 1977). Since then, the Kappa statistic has received considerable attention in remote sensing applications (Congalton and Mead 1983, Rosenfield and Fitzpatrick-Lins 1986, Hudson and Ram 1987, Stehman 1992, Czaplewski 1994, Kalkhan 1994). The Kappa statistic was described in detail by Bishop *et al.* (1975, p. 395-400) and Kalkhan (1994).

Double Sampling and Accuracy Assessments

Kalkhan *et al.* (1995b) point out that double sampling and accuracy assessments involve three phases: (1) unbiased sampling of the landscape; (2) accuracy assessments using a composite estimator; and (3) the integration of data and accuracy assessments into a geographic information systems (GIS). In this paper we discuss phase 2.

Estimating the accuracy of maps derived from remotely sensed data requires statistical sampling of aerial photographs or ground reference plots to insure reliable estimates (Card 1982). However, collecting information to assess the accuracy of a thematic map with respect to reference data is time consuming and costly. In designing a cost-effective sampling method, two feats must be accomplished, both of which affect the quality of information contained in the sample and the precision of the estimate. First, an unbiased sample must be selected from the population. Second, the amount of variation in the sample statistic must be controlled somewhat by the sampling method (Scheaffer *et al.* 1990). Since field observations are expensive, a design that provides a precise estimate of a parameter at minimum cost is desirable (Cochran 1977).

Two phases are required in double sampling. The first phase is to locate a sample point on an aerial photograph. It is preferable to pinpoint a location in the center of the aerial photo because it has less error and distortion (Kalkhan *et al.* 1995b). Next, the sampling point is georeferenced to the Landsat TM imagery through x,y coordinates. In the second phase, the point selected from the aerial photo is located on the ground to verify cover type mapping accuracy.

Accuracy Assessment: The Composite Estimator

Kalkhan *et al.* (1995b) point out that validating the accuracy of coarse- and fine-scale vegetation maps using double sampling requires accurate estimates of bias and variance at multiple spatial scales. For this, a composite estimator developed by Mayback (1979) can be used. The composite estimator is based on data from two phases of sampling, i.e. remotely sensed data and field data. Czaplewski (1992) proposed to use the composite estimator as a statistical method to improve the precision of the estimates. The composite estimator provides considerable flexibility to accommodate complex sampling designs for cross-classified census and sample survey data (Czaplewski 1995). Adding aerial photography as an intermediate phase creates the "double sampling." More details about the composite estimator and its uses with double sampling were provided by Kalkhan *et al.* (1995b).

In assessing the classification accuracy of Landsat TM Data, we used an error matrix "contingency table" which is the most common way to represent the classification accuracy of remotely sensed data or thematic maps. The error matrix consists of a series of rows and columns containing the number of sample units, such as pixels. Each unit is assigned to a particular category relative to its actual type based on a set of reference data. The error matrix provides the user with information on the accuracy of individual categories, and both errors of commission and omission in the classification (Rosenfield and

Fitzpatrick-Lins 1986). Errors of commission relate to user accuracies, while errors of omission represent the accuracy of the remotely sensed data.

The objective of this study is to evaluate the statistical properties of Pearson's coefficient of mean square contingency (Q^2), Pearson's P, Tschuprow's T, Cramer's V, and the Kappa statistic for assessing the classification accuracy of remotely sensed data using double sampling.

METHODS

The data used in this study were from a pilot study designed to evaluate the use of Landsat TM data in identifying forest cover types in the state of North Carolina (Kalkhan 1994). A subsample of data was used to construct three composite error matrices representing three levels of classification accuracy (poor, average, and good) with four forest cover types. Forest cover types were identified using an unsupervised classification procedure available in *ERDAS Software* using bands TM3 (red), TM4 (IR1), and TM5 (IR2). Two error matrices were required for assessing the accuracy of thematic maps using the Kappa statistic and measures of association with double sampling. The first error matrix contained information from Landsat TM data and aerial photos, while the second error matrix contained information from aerial photos and ground data. Each error matrix contained count data (i.e. pixel, spatial information) that were converted to a joint probability matrix for the purpose of sampling and evaluating the accuracy of the maps (Kalkhan 1994). Each error matrix was estimated using Monte Carlo simulation (Rubinstein 1981) using double sampling with replacement. The error matrix relating the aerial photo to ground data was estimated using sample sizes of 50, 100, 150, 300, and 800. The error matrix relating the remotely sensed imagery to the aerial photo was estimated using sample sizes ranged from 200 to 25600 depending on number ground data used in the second phase of sampling. This process was repeated 5000 times using a Monte Carlo simulation based on a *GAUSS* matrix language computer program. The bias for the five indices was computed as the difference between the average estimates of the measures of association and the Kappa statistic and their true population mean which is known without error in the Monte Carlo simulation.

In addition to knowing whether an estimate is unbiased, it is desirable to know something about the precision of the estimates. One method of comparing the precision of estimates with different amounts of bias is to use the mean squared error (MSE). Cochran (1977) defines mean square error as the " $\text{Variance} + \text{Bias}^2$ ". Therefore, an estimator with a smaller MSE is considered more precise than one with a larger MSE, even though the latter may have a smaller variance (Reich and Hussin 1993).

Because of the possible effect of bias on the observed coverage rates, one can estimate the impact of this bias by examining the relative error. The relative error is defined as the ratio of bias to the square root of the mean square error. A large relative error has the effect of distorting confidence probabilities (Cochran 1977). For example, the actual confidence level associated with a nominal 95% confidence interval for an estimate with a

relative error of 0.20 is approximately 0.9454 (Reich and Hussin 1993). Absolute relative errors less than 0.15 have little impact on coverage rates.

In addition to a bias in estimating the overall accuracy, there may also be a bias associated with estimating the variance. This can affect the observed coverage rates. One way of evaluating this bias is to compute the ratio of the mean variance to the simulation variance. The mean variance was computed by averaging the 5000 estimated variances using the classical variance formulae found in the literature (see Kalkhan 1994). For estimating the variance of the Kappa statistic, we used Czaplewski's derived variance approximation formula (Czaplewski 1994). The simulation variance was computed as the variance of the 5000 simulated estimates of overall accuracy. A 95% confidence interval for each index was computed using the appropriate standard error of the estimate, and the proportion of confidence intervals enclosing the true population mean was determined.

RESULTS AND DISCUSSION

The observed (true) values of overall accuracy for the remotely sensed images ranged from 0.33 to 0.87, while the overall accuracy of aerial photo to ground was 0.79 (Table 1). The observed overall accuracy for the Kappa statistic and measures of association ranged from 0.004 to 2.03.

The scenes with four classes had a mean accuracy for the Kappa statistic ranging from -0.8 to 64.3, while Pearson's Q^2 , had an overall accuracy ranging from 0.9 to 140.1 (Table 2). Pearson's P ranged from 9.4 to 75.7, while Tschuprow's T and Cramer's V had an overall accuracy ranging from 5.5 to 68.

Note that the overall accuracy of the Kappa statistic increased for poor classification as the aerial photo to ground sample sizes increased, while for the measures of association their overall accuracy decreased as the sample sizes of remotely sensed images increased. For average and good classifications the overall accuracy decreased for all indices when the aerial photo to ground sample sizes increased. One could expect that the overall accuracy of the aerial photo to ground would be higher compared to the accuracy of remotely sensed images. The Landsat TM imagery was taken on October 8, 1985, while the aerial photos were taken on October 10 and November 9 of 1985. The ground survey to collect ground sampling points was accomplished during November 11, 1982 through November 1986. This might have an effect on assessing the accuracy. Increasing the sample size did not improve the overall accuracy. In fact, the overall accuracy decreased as sample size increased.

Relative bias

A consistent trend was observed in the relative bias with respect to the accuracy of the scene and number of samples at different phases of sampling (Table 3). Pearson's P provided the best estimate of overall accuracy ranging from -0.1 to -0.8 (Table 3) followed by Tschuprow's T, Cramer's V, and Pearson's Q^2 with relative bias ranging from -0.2 to -1.0.

Table 1.--Observed values of five accuracy assessment indices.

Index ^a	Poor	Average	Good	Photo-Ground
O	0.3321	0.5594	0.8744	0.7884
K	0.0035	0.3571	0.8206	0.6967
Q	0.2308	0.9447	2.0320	1.4862
P	0.4330	0.6970	0.8186	0.7732
T	0.2774	0.5611	0.8230	0.7039
V	0.2774	0.5611	0.8230	0.7039

^a Overall accuracy (O), Kappa statistic (K), Pearson's coefficient of mean cohesency (Q), Pearson's (P), Tschuprow's (T), and Cramer's (V).

Table 2.--Estimated¹ values of five accuracy assessment indices.

Index ^a	M	Poor	Average	Good
K 50		-0.7 -0.8 -0.7 -0.8A	30.3 30.3 30.0 29.6	64.1 64.0 64.3 64.2
Q		14.5 13.6 13.3 13.2	44.8 44.8 43.2 42.4	139.0 138.8 140.1 139.9
P		34.3 33.4 33.1 32.9	54.7 54.7 54.1 53.8	75.6 75.5 75.7 75.7
T		21.4 20.8 20.5 20.4	38.4 38.4 37.7 37.4	67.7 67.6 67.9 67.9
V		21.6 20.9 20.7 20.6	38.6 38.6 37.9 37.7	67.8 67.7 68.0 68.0
K 100		-0.4 -0.4 -0.6 -0.5B	28.4 28.1 28.1 28.2	62.2 62.3 62.0 62.3
Q		7.0 6.7 6.7 6.5	34.5 33.9 33.6 33.7	125.4 125.8 125.4 126.1
P		24.8 24.3 24.2 23.9	50.1 49.8 49.7 49.7	74.2 74.3 74.2 74.3
T		14.9 14.6 14.5 14.3	33.7 33.4 33.3 33.3	64.4 64.5 64.4 64.6
V		14.9 14.6 14.5 14.3	33.7 33.4 33.3 33.3	64.4 64.5 64.4 64.6
K 150		-0.3 -0.4 -0.3 -0.4C	27.9 27.7 27.6 27.6	61.5 61.3 61.3 61.4
Q		4.6 4.4 4.4 4.3	31.0 30.5 30.3 30.2	120.4 119.6 119.7 120.2
P		20.4 20.0 19.9 19.7	48.3 48.0 47.9 47.8	73.6 73.5 73.6 73.6
T		12.1 11.8 11.7 11.6	32.0 31.7 31.6 31.5	63.2 63.0 63.0 63.1
V		12.1 11.8 11.7 11.6	32.0 31.7 31.6 31.5	63.2 63.0 63.0 63.1
K 300		-0.2 -0.2 -0.2 -0.2D	27.5 27.3 27.3 27.3	60.3 60.3 60.3 60.1
Q		2.3 2.2 2.2 2.1	27.9 27.5 27.4 27.3	113.1 112.9 113.0 112.3
P		14.6 14.3 14.1 14.1	46.5 46.3 46.2 46.2	72.7 72.7 72.7 72.6
T		8.6 8.4 8.3 8.2	30.4 30.2 30.2 30.1	61.3 61.2 61.3 61.1
V		8.6 8.4 8.2 8.2	30.4 30.2 30.2 30.1	61.3 61.2 61.3 61.1
K 800		-0.1 -0.1 -0.1 -0.1E	27.3 27.3 27.3 27.3	59.6 59.5 59.5 59.4
Q		1.0 1.0 1.0 1.0	26.2 26.0 25.9 26.0	108.6 108.1 108.2 107.8
P		9.9 9.7 9.5 9.4	45.5 45.4 45.3 45.3	72.1 72.0 72.0 72.0
T		5.7 5.6 5.5 5.5	29.5 29.4 29.4 29.4	60.1 60.0 60.0 59.9
V		5.7 5.6 5.5 5.5	29.5 29.4 29.4 29.4	60.1 60.0 60.0 59.9

¹ Estimated values were multiplied by 100.

^a See Table 1 for description of variables.

N Sample size from aerial photo to ground (Second phase of sampling).

A Sample sizes from remotely sensed to aerial photo are 200,400,800,1600.

B Sample sizes from remotely sensed to aerial photo are 400,800,1600,3200.

C Sample sizes from remotely sensed to aerial photo are 600,1200,2400,4800.

D Sample sizes from remotely sensed to aerial photo are 1200,2400,4800,9600.

E Sample sizes from remotely sensed to aerial photo are 3200,6400,12800,25600.

* First phase of sampling from remotely sensed image to aerial photo.

Table 3.--Percent bias* for five accuracy indices.

Index ^a	M	Poor	Average	Good
K 50		-3.1 -3.4 -3.0 -3.3A	-0.2 -0.2 -0.2 -0.2	-0.2 -0.2 -0.2 -0.2
Q		-0.4 -0.4 -0.4 -0.4	-0.5 -0.5 -0.5 -0.6	-0.3 -0.3 -0.3 -0.3
P		-0.2 -0.2 -0.2 -0.2	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T		-0.2 -0.3 -0.3 -0.3	-0.3 -0.3 -0.3 -0.3	-0.2 -0.2 -0.2 -0.2
V		-0.2 -0.3 -0.3 -0.3	-0.3 -0.3 -0.3 -0.3	-0.2 -0.2 -0.2 -0.2
K 100		-2.1 -2.2 -2.8 -2.4B	-0.2 -0.2 -0.2 -0.2	-0.2 -0.2 -0.2 -0.2
Q		-0.7 -0.7 -0.7 -0.7	-0.6 -0.6 -0.6 -0.6	-0.4 -0.4 -0.4 -0.4
P		-0.4 -0.4 -0.4 -0.4	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T		-0.5 -0.5 -0.5 -0.5	-0.4 -0.4 -0.4 -0.4	-0.2 -0.2 -0.2 -0.2
V		-0.5 -0.5 -0.5 -0.5	-0.4 -0.4 -0.4 -0.4	-0.2 -0.2 -0.2 -0.2
K 150		-1.8 -2.0 -2.0 -2.2C	-0.2 -0.2 -0.2 -0.2	-0.3 -0.3 -0.3 -0.3
Q		-0.8 -0.8 -0.8 -0.8	-0.7 -0.7 -0.7 -0.7	-0.4 -0.4 -0.4 -0.4
P		-0.5 -0.5 -0.5 -0.5	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T		-0.4 -0.4 -0.4 -0.4	-0.3 -0.3 -0.3 -0.3	-0.2 -0.2 -0.2 -0.2
V		-0.6 -0.6 -0.6 -0.6	-0.4 -0.4 -0.4 -0.4	-0.2 -0.2 -0.2 -0.2
K 300		-1.5 -1.6 -1.6 -1.5D	-0.2 -0.2 -0.2 -0.2	-0.3 -0.3 -0.3 -0.3
Q		-0.9 -0.9 -0.9 -0.9	-0.7 -0.7 -0.7 -0.7	-0.4 -0.4 -0.4 -0.4
P		-0.7 -0.7 -0.7 -0.7	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T		-0.7 -0.7 -0.7 -0.7	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
V		-0.7 -0.7 -0.7 -0.7	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
K 800		-1.2 -1.3 -1.1 -1.1E	-0.2 -0.2 -0.2 -0.2	-0.3 -0.3 -0.3 -0.3
Q		-1.0 -1.0 -1.0 -1.0	-0.7 -0.7 -0.7 -0.7	-0.5 -0.5 -0.5 -0.5
P		-0.8 -0.8 -0.8 -0.8	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T		-0.8 -0.8 -0.8 -0.8	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
V		-0.8 -0.8 -0.8 -0.8	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1

* Percent bias divided by 100.

^a See Table 1 and 2 for description of variables and sample sizes.

Table 4.--Mean square error* of five accuracy indices.

Index ^a	M	Poor	Average	Good
K 50		954.3 1179.8 894.8 1077.7A	2.3 2.3 2.6 2.9	4.8 4.8 4.7 4.7
Q		13.8 16.7 17.9	18.2 27.7 27.7 29.5	30.3 10.0 10.1 9.6
P		4.3 5.3 5.6	5.8 4.6 4.6 5.0	5.2 0.6 0.6 0.6
T		5.2 6.3 6.7	6.9 10.0 10.0 10.8	11.1 3.2 3.2 3.1
V		5.0 6.1 6.5	6.7 9.7 9.7 10.5	10.8 3.1 3.1 3.0
K 100		441.6 500.0 789.1 561.3B	4.1 4.5 4.5 4.5	5.9 5.8 6.0 5.8
Q		48.3 50.1 50.2	51.7 40.3 41.1 41.5	41.4 14.7 14.5 14.7
P		18.2 19.2 19.4	20.1 7.9 8.1 8.2	8.2 0.9 0.9 0.9
T		21.4 22.6 22.7	23.6 16.0 16.0 16.4	16.6 16.5 4.7 4.7
V		21.4 22.6 22.7	23.6 16.0 16.4 16.6	16.5 4.7 4.7 4.6
K 150		319.2 412.3 392.3 468.9C	4.8 5.1 5.2 5.2	6.3 6.4 6.4 6.4
Q		64.0 65.3 65.5	66.2 45.1 45.9 46.1	46.4 16.6 16.9 16.9
P		28.0 29.1 29.3	29.8 9.4 9.7 9.8	9.9 1.0 1.0 1.0
T		31.9 33.0 33.2	33.8 18.5 18.9 19.0	19.2 5.4 5.5 5.4
V		31.9 33.0 33.2	33.8 18.5 18.9 19.0	19.2 5.4 5.5 5.4
K 300		225.3 272.0 257.1 238.8D	5.3 5.5 5.5 5.5	7.0 7.1 7.1 7.2
Q		80.9 81.6 82.2	82.3 49.6 50.3 50.4	50.5 19.6 19.8 19.7
P		43.8 44.7 45.4	45.6 11.0 11.3 11.3	11.4 1.2 1.3 1.3
T		47.8 48.7 49.4	49.5 21.0 21.3 21.4	21.5 6.5 6.5 6.5
V		47.8 48.7 49.4	49.5 21.0 21.3 21.4	21.5 6.5 6.5 6.5
K 800		144.1 167.7 131.4 130.3E	5.5 5.6 5.7 5.6	7.5 7.5 7.5 7.6
Q		91.3 91.5 91.8	92.0 52.3 52.5 52.7	52.6 21.7 21.9 21.9
P		59.6 60.2 60.9	61.1 12.1 12.2 12.2	12.2 1.4 1.4 1.4
T		63.0 63.5 64.1	64.4 22.5 22.6 22.7	22.7 7.3 7.3 7.4
V		63.0 63.5 64.1	64.4 22.5 22.6 22.7	22.7 7.3 7.3 7.4

* MSE divided by 100.

^a See Table 1 and 2 for description of variables and sample sizes.

Table 5.--Nominal* 95% confidence coverage rates using the classical variance formula for five accuracy indices.

Index ^a	M	Poor	Average	Good
K 50		200 41 80 49 15 73	3 66 4 1 10	15 12 19 4 32
Q		400 26 66 29 8 45	1 50 2 0 4	9 6 11 3 18
P		800 17 50 18 6 29	1 36 1 0 2	7 4 7 2 11
V		1600 11 36 11 3 18	0 25 0 0 1	4 3 4 2 7
K 100		400 2 81 3 0 9	0 45 0 0 0	1 1 2 0 4
Q		800 1 67 1 0 3	0 28 0 0 0	1 0 1 0 2
P		1600 1 51 1 0 2	0 18 0 0 0	0 0 0 0 0
V		3200 0 39 0 0 1	0 13 0 0 0	0 0 0 0 0
K 150		600 0 82 0 0 1	0 30 0 0 0	0 0 0 0 0
Q		1200 0 67 0 0 0	0 16 0 0 0	0 0 0 0 0
P		2400 0 52 0 0 0	0 5 0 0 0	0 0 0 0 0
V		4800 0 39 0 0 0	0 3 0 0 0	0 0 0 0 0
K 300		1200 0 67 0 0 0	0 2 0 0 0	0 0 0 0 0
Q		2400 0 54 0 0 0	0 1 0 0 0	0 0 0 0 0
P		4800 0 38 0 0 0	0 1 0 0 0	0 0 0 0 0
V		9600 0 7 0 0 0	0 0 0 0 0	0 0 0 0 0
K 800		3200 0 67 0 0 0	0 0 0 0 0	0 0 0 0 0
Q		6400 0 53 0 0 0	0 0 0 0 0	0 0 0 0 0
P		12800 0 40 0 0 0	0 0 0 0 0	0 0 0 0 0
V		25600 0 0 0 0 0	0 0 0 0 0	0 0 0 0 0

* 95% Confidence Interval multiply by 100.

^a See Table 1 for description of variables.

^b Sampling from remote sensing image to aerial photo.

Table 6.--Relative error* for five accuracy assessment indices.

Index ^a	M	Poor	Average	Good
K 50		-0.0 -0.0 -0.0 -0.0A	-3.6 -3.6 -3.6 -3.6	-8.2 -8.2 -8.2 -8.2
Q		-2.3 -2.3 -2.3 -2.3	-9.4 -9.4 -9.4 -9.4	-20.3 -20.3 -20.3 -20.3
P		-4.3 -4.3 -4.3 -4.3	-7.0 -7.0 -7.0 -7.0	-8.2 -8.2 -8.2 -8.2
T		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
V		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
K 100		-0.0 -0.0 -0.0 -0.0B	-3.6 -3.6 -3.6 -3.6	-8.2 -8.2 -8.2 -8.2
Q		-2.3 -2.3 -2.3 -2.3	-9.4 -9.4 -9.4 -9.4	-20.3 -20.3 -20.3 -20.3
P		-4.3 -4.3 -4.3 -4.3	-7.0 -7.0 -7.0 -7.0	-8.2 -8.2 -8.2 -8.2
T		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
V		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
K 150		-0.0 -0.0 -0.0 -0.0C	-3.6 -3.6 -3.6 -3.6	-8.2 -8.2 -8.2 -8.2
Q		-2.3 -2.3 -2.3 -2.3	-9.4 -9.4 -9.4 -9.4	-20.3 -20.3 -20.3 -20.3
P		-4.3 -4.3 -4.3 -4.3	-7.0 -7.0 -7.0 -7.0	-8.2 -8.2 -8.2 -8.2
T		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
V		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
K 300		-0.0 -0.0 -0.0 -0.0D	-3.6 -3.6 -3.6 -3.6	-8.2 -8.2 -8.2 -8.2
Q		-2.3 -2.3 -2.3 -2.3	-9.4 -9.4 -9.4 -9.4	-20.3 -20.3 -20.3 -20.3
P		-4.3 -4.3 -4.3 -4.3	-7.0 -7.0 -7.0 -7.0	-8.2 -8.2 -8.2 -8.2
T		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
V		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
K 800		-0.0 -0.0 -0.0 -0.0E	-3.6 -3.6 -3.6 -3.6	-8.2 -8.2 -8.2 -8.2
Q		-2.3 -2.3 -2.3 -2.3	-9.4 -9.4 -9.4 -9.4	-20.3 -20.3 -20.3 -20.3
P		-4.3 -4.3 -4.3 -4.3	-7.0 -7.0 -7.0 -7.0	-8.2 -8.2 -8.2 -8.2
T		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
V		-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2

* Relative error multiply by 1000.

^a See Table 1 and 2 for description of variables and sample sizes.

Table 7.--Ratio of mean variance to simulation variance for five accuracy indices.

Index ^a	M	Poor				Average				Good			
K 50	0.44	0.24	0.12	0.06A	0.34	0.34	0.10	0.05	0.24	0.12	0.06	0.03	
Q	0.38	0.20	0.10	0.04	0.28	0.28	0.08	0.04	0.30	0.10	0.05	0.02	
P	0.46	0.23	0.12	0.06	0.31	0.31	0.08	0.04	0.27	0.10	0.05	0.02	
T	0.05	0.03	0.01	0.01	0.03	0.03	0.01	0.00	0.03	0.01	0.01	0.00	
V	0.97	0.49	0.25	0.12	0.62	0.62	0.17	0.09	0.68	0.23	0.11	0.06	
K 100	0.48	0.23	0.11	0.06B	0.34	0.17	0.09	0.04	0.24	0.12	0.06	0.03	
Q	0.47	0.26	0.13	0.07	0.38	0.19	0.11	0.05	0.26	0.13	0.07	0.03	
P	0.53	0.25	0.12	0.06	0.34	0.18	0.10	0.05	0.25	0.13	0.06	0.03	
T	0.06	0.03	0.01	0.01	0.04	0.02	0.01	0.00	0.03	0.01	0.01	0.00	
V	1.16	0.55	0.26	0.14	0.76	0.38	0.21	0.10	0.55	0.28	0.14	0.07	
K 150	0.48	0.26	0.13	0.07C	0.39	0.21	0.11	0.06	0.28	0.14	0.07	0.04	
Q	0.53	0.25	0.12	0.07	0.37	0.20	0.10	0.05	0.27	0.14	0.07	0.03	
P	0.59	0.28	0.13	0.07	0.40	0.20	0.10	0.05	0.27	0.14	0.07	0.03	
T	0.06	0.03	0.01	0.01	0.04	0.02	0.01	0.01	0.03	0.02	0.01	0.00	
V	1.27	0.62	0.30	0.17	0.83	0.44	0.22	0.11	0.60	0.31	0.16	0.08	
K 300	0.49	0.26	0.14	0.07D	0.42	0.22	0.11	0.06	0.30	0.15	0.08	0.04	
Q	0.49	0.26	0.14	0.07	0.42	0.22	0.11	0.06	0.30	0.15	0.08	0.04	
P	0.62	0.31	0.17	0.08	0.40	0.21	0.13	0.06	0.30	0.15	0.08	0.04	
T	0.07	0.03	0.02	0.01	0.04	0.02	0.02	0.01	0.03	0.02	0.01	0.00	
V	1.39	0.70	0.36	0.19	0.86	0.47	0.24	0.12	0.61	0.31	0.16	0.08	
K 800	0.50	0.27	0.15	0.07E	0.42	0.23	0.12	0.06	0.31	0.16	0.08	0.04	
Q	0.49	0.26	0.15	0.08	0.40	0.22	0.11	0.06	0.33	0.17	0.09	0.04	
P	0.63	0.32	0.16	0.09	0.40	0.22	0.11	0.06	0.33	0.17	0.09	0.04	
T	0.07	0.04	0.02	0.01	0.04	0.02	0.02	0.01	0.04	0.02	0.01	0.00	
V	1.44	0.71	0.36	0.19	0.87	0.47	0.24	0.12	0.61	0.31	0.16	0.08	

The Kappa statistic had high relative bias ranging from -0.2 to -3.4. As the overall accuracy of the scene increased, the relative bias decreased for all indices. The relative bias for the Kappa statistic decreased when the aerial photo to ground sample sizes increased for poor classification. The opposite was found for the measures of association for average and good classification; the bias increased as the sample size increased.

Mean squared error

Pearson's P had the smallest MSE across all sample sizes ranging from 0.6 to 61.1 followed by Cramer's V and Tschuprow's T with the MSE ranging from 3.0 to 64.4 for all levels of classification (Table 4). Pearson's Q² had a MSE ranging from 9.7 to 92, while the Kappa statistic had the largest MSE. In general, the mean square error decreased with increased accuracy of the remotely sensed image. The MSE for the Kappa statistic decreased as the aerial photo to ground sample sizes increased, while the opposite was observed for the measures of association. The MSE increased consistently for all indices as the sample size associated with estimating the error matrix relating that to aerial photo to the ground. For the poor classification, the Kappa statistic decreased as the sample size of the photo-ground increased.

95 Percent coverage rates

None of the five indices provided consistent coverage rates across all sample sizes and levels of accuracy (Table 5). The Kappa statistic provided reasonable coverage rates, but was not consistent for the poor classification. The coverage rate was very poor for the average and good classification. The inconsistency in coverage rates may be due to the bias, which could distort the confidence intervals. The lower than expected coverage rates may also be due to a bias in estimating the variance.

Bias relative to the root mean squared error

In computing the relative errors, the simulation variance was used instead of the classical variance to facilitate comparison between all accuracy indices. Except for a few instances, all the absolute relative errors exceeded 0.20 (Table 6). This suggests that the observed biases are large enough relative to the variability to distort the expected coverage rates. This would help explain why the coverage rates were lower than expected.

Ratio of the mean variance to the simulation variance

As noted, the lower than expected coverage rates also may be due to a bias associated with estimating the variance. One way to evaluate this bias is to calculate the ratio of the mean variance to the simulation variance (Table 7). It is assumed that the simulation variance provides an unbiased estimate of the variance. Thus, a ratio of less than 1 indicates an underestimation of the variance, while a ratio greater than 1 indicates an

overestimation (Cochran 1977). Using this as a guideline, one can see that all indices underestimated the variance, except for a few cases where Cramer's V overestimated the variance. There was also some indication that the bias decreased as the aerial photo to ground sample sizes increased for poor classification. In addition, as the level of the accuracy improved, the bias increased.

The biases associated with estimating the variances may be due to the asymptotic nature of the variance formulae which assume large sample sizes. When used with a small sample size the variance formulae may not provide valid estimates of the variance (see Kalkhan 1994), especially for the Kappa statistic where we used derived variance formula developed by Czaplewski (1994). Also, the variance estimates of Pearson's P, Tschuprow's T, and Cramer's V are based on a nonlinear transformation of the asymptotic variance of Pearson's Q^2 . This transformation may also introduce a bias in the estimation process. In general, increasing the sample size of the remotely sensed image did not seem to decrease the bias of the mean variance estimates. This is because the overall accuracy of aerial photo to ground comparison was less than the Landsat TM imagery.

CONCLUSION

The Kappa statistic and Pearson's P provided the best estimates of overall accuracy for all levels of classification, sample sizes, and numbers of forest cover types. In addition, the results of the Kappa statistic in this study are applicable to Pielou's index of segregation (Pielou 1960, 1977; Kalkhan *et al.* 1995a). Poor performance of the measures of association were due to large biases associated with the statistic. The ratios of the estimated mean variance to the simulation variance suggested that the classical variance formula for the various indices provided biased estimates of variance, irrespective of the number of classes, or the sample size. This bias may be due to the asymptotic nature of the variance formulae. These variance formulae may not provide valid estimates when used with small sample sizes. Some of the bias associated with the variance may be due to the nonlinear transformation used to approximate the variance of Pearson's P, Tschuprow's T, and Cramer's V. To overcome this problem, one could use bootstrapping (Efron 1979) to provide unbiased estimates of the variance (Kalkhan 1994).

Double sampling, as applied in this study, allows one to make inference about the whole image and not a preselected subset of homogenous test sites. It also suggests that, increasing the sample size, or reducing the number of classes on the remotely sensed image may increase the accuracy and precision of the estimates. However, this was not the case in this study because the accuracy of this particular aerial photo to ground was low. No matter what indices one uses to assess the accuracy of a remotely sensed image, bootstrapping should be used to ensure unbiased estimates of the variance (Kalkhan 1994). To improve the findings of this study, we suggest the use of a large number of classes for assessing the accuracy of the thematic map. Prior to assessing the accuracy of any thematic map, the user of remotely sensed must be familiar with the statistical properties of any accuracy assessment index before using it.

REFERENCES

- Bishop, Y. M. M., S. E. Feinberg, and P. W. Hooland. 1975. Discrete multivariate analysis-theory and practice. MIT Press, Cambridge, MA. 575 p.
- Card, D. H. 1982. Using known map categorical marginal frequencies to improve map accuracy. *Photogrammetric Eng. & Remote Sensing*. 48:431-439.
- 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. *Educ. Psychological Meas.* 20:37-46.
- Congalton, R. G., and R. A. Mead. 1983. A quantitative method to test for consistency and correctness in photointerpretation. *Photogramm. Eng. & Remote Sens.* 49:67-74.
- Czaplewski, R. L. 1992. Accuracy assessment of remotely sensed classification with multi-phase sampling and the multivariate composite estimator. P. 22 In: 1992 (XVIth) Intern. Biometric Conference, Hamilton, New Zealand, December 7-11. Vol. 2, 268 p.
- Czaplewski, R. L. 1994. Variance approximation for assessment of classification accuracy. Research Paper, Rocky Mountain For. & Range Exper. Station, Forest Serv. U.S.D.A., 94 p.
- Czaplewski, R. L. 1995. Assessment of classification accuracy and extent estimates for land cover map with double sampling. *Forest Science* (In Review).
- Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Stat.* 7:1-26.
- Helldon, U. 1980. A test of Landsat-2 imagery and digital data for thematic mapping illustrated by an environmental study in northern Kenya. Sweden, Lund University Natural geography Institute, Report No. 7.
- Hudson, W. D., and C. W. Ram. 1987. Correct formulation of the Kappa coefficient of agreement. *Photogramm. Eng. & Remote Sens.* 53:421-422.
- Landis, J. R., and G. G. Kock. 1977. The measurement of observer agreement for categorical data: *Biometrics*. 33:154-174.
- 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.
- Kalkhan, M. A., R. M. Reich, and R. L. Czaplewski. 1995a. Statistical properties of five accuracy indices in assessing the accuracy of remotely sensed data using simple random sampling. In: 1995 ACSM/ASPRS Annual Convention & Exposition, ASPRS Technical Papers, February 27- March, 1995, Charlotte, North Carolina. P. 246-257.
- Kalkhan, M. A., T. J. Stohlgren, and M. Coughneour. 1995b. An investigation of biodiversity and Landscape-scale gap patterns using double sampling: A GIS approach. In: 1995 Ninth International Symposium on Geographic Information Systems for natural resources, environment and land information management, March 27-30, 1995, Vancouver, British Columbia, Canada. P. 708-712.

- Maybeck, P.S. 1979. Stochastic Models, Estimation, and Control, vol. 1. Academic Press, New York.
- Pielou, E. C. 1961. Segregation and symmetry in two species populations as studied by nearest neighbor relations. *J. Ecology*. 49:255-269.
- Pielou, E. C. 1977. Mathematical Ecology. John Wiley and Sons, New York. 385 p.
- Reich, R. M., and Y. A. Hussin. 1993. Estimating average stand biomass for a regional forest inventory using radar backscatter. *ITC Journal*. 1:82-87.
- Rosenfield, G. H. and K. Fitzpatrick-Lins. 1986. A coefficient of agreement as a measure of thematic classification accuracy. *Photogramm. Eng. & Remote Sensing*. 52:223-227.
- Rubinstein, R. Y. 1981. Simulation and the monte carlo method. John Wiley and Sons, New York. 278 p.
- Scheaffer, R. L., W. Mendenhall, and L. Ott. 1990. Elementary Survey Sampling, 4th ed., PWS-Kent Publishing Company, Boston. MA. 390 p.
- Short, N. M. 1982. The Landsat tutorial workbook: Basis of satellite remote sensing. Greenbelt, MD., GSF Center, NASA Reference Publication 1078.
- Stehman, S. V. 1992. Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogramm. Eng. & Remote Sens.* 58:1343-1350.
- Turk, G. 1979. GT index: A measure of the success of prediction. *Remote Sens. Environ.* 8:65-75.

BIOGRAPHICAL SKETCH

Mohammed A. Kalkhan is a Research Scientist/GIS Specialist with the Natural Resource Ecology Laboratory, Colorado State University. He graduated from University of Mosul, Iraq with B.Sc. (Forestry) in 1973 and M.Sc. (Forest Mensuration) in 1980. Mohammed holds a Ph.D. in Forest Biometrics from Colorado State University (1994). Research interest include using remote sensing, GIS, biometrics, and spatial statistics to study natural resources, landscape-gap analysis and ecology, biodiversity, and accuracy assessment of remotely sensed data with sampling techniques.

Robin M. Reich is an Associate Professor with the Department of Forest Sciences, Colorado State University. Research interest spatial statistics, natural resource sampling and forest biometrics. Current research is aimed at integrating of spatial and geostatistical technique with GIS and remote sensing to model the spatial interaction biological populations have with themselves and their environment.

Raymond L. Czaplewski has spent the past 13 years working on statistical design and analysis of regional monitoring systems for forests and other environmental features, primarily the Forest Service's national Forest Inventory and Analysis Program. Other accomplishments include statistical design of the 1990 and 2000 global surveys of tropical forests by the Food and Agricultural Organization of the United Nations, and a design to monitor the effectiveness of the President's 1994 Forest Plan for the States of Oregon, Washington, and California. He has specialized in linking sample surveys to remotely sensed data and geographic information systems.