Estimating the Kappa Coefficient and its Variance under Stratified Random Sampling

Stephen V. Stehman

Abstract

The kappa coefficient of agreement is frequently used to summarize the results of an accuracy assessment used to evaluate land-use or land-cover classifications obtained by remote sensing. The standard estimator of the kappa coefficient along with the standard error of this estimator require a sampling model that is approximated by simple random sampling. Formulas are presented for estimating the kappa coefficient and its variance for stratified random sampling. Empirical results demonstrate that these estimators have little bias, and confidence intervals perform well, often even at relatively small sample sizes.

Introduction

Accuracy assessment of land-use or land-cover classifications obtained by satellite remote sensing is necessary to evaluate the quality of maps developed from remotely sensed data. A typical strategy for accuracy assessment is to use a statistically sound sampling design to select a sample of locations (pixels) in the study region, and to determine if the land-use or land-cover classification assigned to that pixel matches the true classification of the ground location represented by that pixel. The reference classification, whether obtained on the basis of ground visit or photointerpretation, is assumed to be correct. The sample data are often summarized in an error matrix, which is then subjected to various statistical analyses (Congalton *et al.*, 1983). The kappa coefficient of agreement (KAPPA) (Cohen, 1960) is one parameter frequently used in these analyses of error matrices (here "parameter" is used in the statistical sense of a number describing a characteristic of the population). Congalton *et al.* (1983) and Rosenfield and Fitzpatrick-Lins (1986) provide additional details on applications of KAPPA in remote sensing.

To estimate KAPPA from the sample data, Bishop *et al.* (1975, p. 396) and Agresti (1989, p. 366) present formulas for the maximum-likelihood estimator, KHAT, of KAPPA, and the standard error of KHAT. These formulas were derived under the assumption of multinomial sampling, which is approximately satisfied by simple random sampling. The effect of using these formulas when the sampling design is not simple random has received little attention (Congalton, 1991). Stehman (1992) found that the usual formula for estimating KAPPA had negligible bias when the sampling design was systematic or systematic unaligned, but the estimator of the variance of KHAT was biased. Other sampling designs have not been studied.

Stratified sampling (Cochran, 1977, Chap. 5) is a potentially useful design for accuracy assessment. In particular, if strata are constructed on the basis of the categories of the remotely sensed image, stratified sampling permits control over the number of sample observations in each map category. This guarantees that a minimum sample size can be selected in each stratum or category. Because the stratified design does not satisfy the multinomial sampling model, neither the standard formula for KHAT nor the formula for the standard error of KHAT is appropriate.

In this article, an estimator (denoted KS) of KAPPA is derived for use with stratified sampling. In addition, the variance and a sample-based variance estimator of KS are derived under stratified random sampling. KS is appropriate for systematic sampling within strata, a design which might be used if stratification is by geographic region rather than by map category. However, the variance and variance estimator derived in this article are not appropriate for a stratified systematic design. The variance formulas depend on a large sample approximation, so empirical results are presented to confirm the validity of the estimators proposed, and to evaluate properties of these estimators at small sample sizes.

Description of Estimators

Suppose a remote sensing image of N pixels is classified into q categories. Given a census of all N pixels and the true classification of each pixel, the population error matrix is

		Reference				Row
		1	2	. . .	q	Total
Image (Stratum)	1	N_{11}	N_{12}	. . .	N_{1q}	N_1
	2	N_{21}	N_{22}	. . .	N_{2q}	N_2
	⋮	⋮	⋮	. . .	⋮	⋮
	q	N_{q1}	N_{q2}	. . .	N_{qq}	N_q
Col. Tot.		M_1	M_2	. . .	M_q	N

A census is, of course, impractical, so sampling is necessary to obtain an estimate of KAPPA. For a stratified sampling design, in which the strata are the map classes of the image, the row totals, N_h, are known, but the column totals, M_j, are unknown. All entries, N_{hj}, within the error matrix are also unknown.

From the population error matrix, the parameter of interest is (Bishop *et al.*, 1975, p. 395)

$$\text{KAPPA} = \frac{\sum_{i=1}^{q} p_{ii} - \sum_{i=1}^{q} p_{i+}p_{+i}}{1 - \sum_{i=1}^{q} p_{i+}p_{+i}} \qquad (1)$$

where $p_{ii} = N_{ii}/N$, $p_{i+} = N_i/N$, and $p_{+i} = M_i/N$. Because it is a parameter of the population, KAPPA is unchanged by the choice of sampling design. If the q rows of the population er-

Photogrammetric Engineering & Remote Sensing,
Vol. 62, No. 4, April 1996, pp. 401–402.

0099-1112/96/6204-401$3.00/0

SUNY-ESF, 320 Bray Hall, 1 Forestry Drive, Syracuse, NY 13210.

ror matrix are used as strata, KAPPA may be written in the following forms, algebraically equivalent to Equation 1, corresponding to the population stratification:

$$\text{KAPPA} = \frac{\sum_{h=1}^{q} (N_{hh}/N) - \sum_{h=1}^{q} \left(\frac{N_h}{N}\right)\left(\frac{M_h}{N}\right)}{1 - \sum_{h=1}^{q} \left(\frac{N_h}{N}\right)\left(\frac{M_h}{N}\right)} \quad (2)$$

$$= \frac{N \sum_{h=1}^{q} N_{hh} - \sum_{h=1}^{q} N_h M_h}{N^2 - \sum_{h=1}^{q} N_h M_h}. \quad (3)$$

In stratified random sampling, a simple random sample of pixels is selected in each stratum. The stratum sample sizes, n_h, are specified in advance of the sample collection by the investigator, and each stratum is sampled independently of the other strata. Ground visits or photointerpretation are used to determine if the sampled pixel is classified correctly or not. The results for a sample of $n = \sum_{h=1}^{q} n_h$ pixels are organized into a sample error matrix,

		Reference				Row
		1	2	. . .	q	Total
Image	1	n_{11}	n_{12}	. . .	n_{1q}	n_1
	2	n_{21}	n_{22}	. . .	n_{2q}	n_2
	⋮	⋮	⋮	. . .	⋮	⋮
	q	n_{q1}	n_{q2}	. . .	n_{qq}	n_q
						n

The row totals (n_h) are fixed by the stratified design, but the n_{hj}'s depend on the observed sample.

The details of the derivation of the stratified sampling estimator of KAPPA and the variance of this estimator are deferred to the Appendix. The general approach follows from Result 5.5.1 of Särndal *et al.*, (1992) in which the population quantities N_{hh} and M_h in Equation 3 are replaced by standard, unbiased, stratified random sampling estimators of these population totals. The estimator of KAPPA for stratified sampling is

$$\text{KS} = \frac{N \sum_{h=1}^{q} \hat{N}_{hh} - \sum_{h=1}^{q} N_h \hat{M}_h}{N^2 - \sum_{h=1}^{q} N_h \hat{M}_h}, \quad (4)$$

where $\hat{N}_{hh} = \frac{N_h}{n_h} n_{hh}$ is an unbiased estimator of N_{hh}. The notation is somewhat awkward here in that $\hat{M}_h$ estimates a column total, not a row total. For any column $j = 1,...,q$, an unbiased estimator of M_j is $\hat{M}_j = \sum_{h=1}^{q} \frac{N_h}{n_h} n_{hj}$ (the estimated column total requires summation over the q strata). KS is not an unbiased estimator of KAPPA, but it is a consistent estimator (Särndal *et al.*, 1992, p. 168), and the bias of KS is shown to be small in the populations examined in the subsequent empirical study.

The variance of KS, denoted **V**(KS), requires knowing the population error matrix, and a large-sample, approximate formula, denoted **AV**(KS), is presented in the Appendix. Note that **V**(KS) is the parameter of interest, and that **AV**(KS) is a population quantity that will approximate **V**(KS) for large samples. The important practical problem of estimating **V**(KS) from the sample data is addressed as follows. For convenience in writing the formula for the estimated variance, define $\hat{D} = \sum_{h=1}^{q} \hat{N}_{hh}$, and $\hat{C} = \sum_{h=1}^{q} N_h \hat{M}_h$. For each stratum h of the sample error matrix, calculate

$$\overline{u}_h = \frac{1}{n_h} \left\{ n_{hh} \left[\frac{N}{N^2 - \hat{C}} \right] + \frac{N(\hat{D} - N)}{(N^2 - \hat{C})^2} \sum_{\substack{j=1 \\ j \neq h}}^{q} n_{hj} N_j \right\}, \quad (5)$$

$$u_h^2 = n_{hh} \left[\frac{N}{N^2 - \hat{C}} + \frac{N_h N(\hat{D} - N)}{(N^2 - \hat{C})^2} \right]^2 + N^2 \frac{(\hat{D} - N)^2}{(N^2 - \hat{C})^4} \sum_{\substack{j=1 \\ j \neq h}}^{q} n_{hj} N_j^2 \quad (6)$$

and

$$\hat{\mathbf{V}}_h = (u_h^2 - n_h \overline{u}_h^2)/(n_h - 1). \quad (7)$$

Then the estimated variance of KS is

$$\hat{\mathbf{V}}(\text{KS}) = \sum_{h=1}^{q} N_h^2 (1 - f_h) \hat{\mathbf{V}}_h / n_h, \quad (8)$$

where $f_h = n_h/N_h$ is the sampling fraction in stratum h. For strata in which n_h is small relative to N_h, the finite population correction factor $(1 - f_h)$ may be ignored. A confidence interval for KAPPA is constructed using KS ± $z_\alpha \sqrt{\hat{\mathbf{V}}(\text{KS})}$, where z_α is the percentile from a standard normal distribution appropriate for the desired confidence probability.

Empirical Results

A simulation study was conducted to evaluate how closely **AV**(KS) approximates **V**(KS), to evaluate bias of KS and $\hat{\mathbf{V}}$(KS), and to explore the properties of these estimators for small sample sizes. The 11 populations investigated were selected to represent a variety of population error matrices (Table 1). Four of these matrices (AIRPORT1, BLOCK, DIAGONAL, and MASSLAND) were studied by Stehman (1992), three were artificial populations constructed for this study (STRAT3, STRAT4, and STRAT8), and four were constructed by expanding published sample error matrices (BLIGHT, Table 2 in Rosenfield and Fitzpatrick-Lins (1986); OLDGROWTH, Table 3 in Congalton *et al.* (1993); STANDCON, Table 3 in Fiorella and Ripple (1993); and GREEN, Table 1 in Green *et al.* (1993)). The latter four population error matrices were created by multiplying every entry of the sample error matrix by a constant. The constructed population error matrix thus has the same KAPPA, the same proportions for each entry within the error matrix (i.e., n_{hj}/n of the original sample error matrix is equal to N_{hj}/N of the constructed population matrix), and the same row and column marginal proportions as the sample error matrix from which it was constructed. Population error matrices created in this manner should resemble those encountered in real applications. Several sample sizes were examined for each population. Under equal allocation of samples to strata, the smallest sample size evaluated was 10 per stratum, and the largest was 75 per stratum.

The simulation results were based on 10,000 replications of the stratified random sampling design for each sample size and population error matrix. For each sample, KS and $\hat{\mathbf{V}}$ (KS) were calculated. The simulated expected value of KS was computed by

$$\text{E(KS)} = \sum_{i=1}^{10{,}000} \text{KS}_i/10{,}000, \quad (9)$$

where KS$_i$ is KS for sample i. Bias of KS was estimated by E(KS)-KAPPA. The simulated expected value of $\hat{\mathbf{V}}$ (KS) was

TABLE 1. POPULATION ERROR MATRICES (ORDERED BY NUMBER OF CLASSIFICATION CATEGORIES AND KAPPA). THE COLUMNS REPRESENT THE REFERENCE CLASSIFICATION, AND THE ROWS REPRESENT THE IMAGE CLASSIFICATION.

OLDGROWTH (KAPPA=0.6389)

	1	2	N_h	N_h/N
1	4560	705	5265	0.392
2	1685	6495	8180	0.608
M_h	6245	7200	13445	

BLOCK (KAPPA=0.4544)

	1	2	3	N_h	N_h/N
1	2165	565	309	3039	0.475
2	678	944	108	1730	0.270
3	136	432	1063	1631	0.255
M_h	2979	1941	1480	6400	

DIAGONAL (KAPPA=0.6539)

	1	2	3	N_h	N_h/N
1	1950	351	104	2405	0.376
2	343	1230	107	1680	0.263
3	110	457	1748	2315	0.362
M_h	2403	2038	1959	6400	

AIRPORT1 (KAPPA=0.6845)

	1	2	3	N_h	N_h/N
1	1750	218	140	2108	0.375
2	330	1331	152	1813	0.322
3	136	200	1368	1704	0.303
M_h	2216	1749	1660	5625	

STRAT3 (KAPPA=0.8053)

	1	2	3	N_h	N_h/N
1	6840	180	180	7200	0.950
2	270	3060	270	3600	0.850
3	180	180	840	1200	0.700
M_h	7290	3420	1290	12000	

MASSLAND (KAPPA=0.4785)

	1	2	3	4	N_h	N_h/N
1	5999	2169	1764	152	10084	0.384
2	637	1877	486	27	3027	0.115
3	1753	752	8429	271	11205	0.427
4	109	220	751	854	1934	0.074
M_h	8498	5018	11430	1304	26250	

GREEN (KAPPA=0.6533)

	1	2	3	4	N_h	N_h/N
1	2000	200	300	0	2500	0.250
2	100	2100	200	100	2500	0.250
3	700	800	1000	0	2500	0.250
4	0	200	0	2300	2500	0.250
M_h	2800	3300	1500	2400	10000	

STRAT4 (KAPPA=0.7729)

	1	2	3	4	N_h	N_h/N
1	4135	258	167	82	4642	0.542
2	220	2127	88	43	2478	0.290
3	27	125	953	89	1194	0.140
4	12	27	55	149	243	0.028
M_h	4394	2537	1263	363	8557	

$$\mathrm{E}[\hat{\mathbf{V}}(\text{KS})] = \sum_{i=1}^{10,000} \hat{\mathbf{V}}(\text{KS}_i)/10,000, \qquad (10)$$

and the simulated variance of KS was

TABLE 1. (CONTINUED)

BLIGHT (KAPPA=0.7544)

	1	2	3	4	5	N_h	N_h/N
1	4440	0	30	30	30	4530	0.469
2	30	1500	180	0	0	1710	0.177
3	240	450	1170	180	0	2040	0.211
4	60	90	210	750	30	1140	0.118
5	0	0	30	30	180	240	0.025
M_h	4770	2040	1620	990	240	9660	

STANDCON (KAPPA=0.7184)

	1	2	3	4	5	N_h	N_h/N
1	1700	200	100	0	0	2000	0.167
2	300	1300	400	0	0	2000	0.167
3	0	100	1900	0	0	2000	0.167
4	0	0	100	900	1000	2000	0.167
5	0	0	0	400	3600	4000	0.333
M_h	2000	1600	2500	1300	4600	12000	

STRAT8 (KAPPA=0.8529)

	1	2	3	4	5	6	7	8	N_h	N_h/N
1	4000	300	200	100	50	25	10	30	4715	0.339
2	5	3000	50	10	10	5	3	6	3089	0.222
3	20	20	1800	30	10	10	5	5	1900	0.136
4	5	10	20	500	30	25	20	10	620	0.045
5	10	25	35	45	750	58	20	15	958	0.069
6	30	3	8	8	39	1021	40	20	1169	0.084
7	5	10	15	20	20	30	700	25	825	0.059
8	5	10	15	20	25	30	40	500	645	0.046
M_h	4080	3378	2143	733	934	1204	838	611	13921	

$$\mathbf{V}(\text{KS}) = \sum_{i=1}^{10,000} (\text{KS}_i - \text{KAPPA})^2/10,000. \qquad (11)$$

AV(KS) was obtained directly from the population error matrix using Appendix Equation A17. The relative error of **AV**(KS) was calculated as [**AV**(KS) − **V**(KS)]/**V**(KS), and the relative bias of $\hat{\mathbf{V}}$(KS) was calculated as {E[$\hat{\mathbf{V}}$(KS)] − **V**(KS)}/**V**(KS).

For the 11 population error matrices investigated in the simulation study, the bias of KS was negligible even at the smallest sample sizes (Table 2). The maximum absolute bias observed was 0.002. The relative error of **AV**(KS) was generally between 1 percent and 2 percent for most sample sizes and populations, and the maximum relative error was 3.8 percent. Relative bias of $\hat{\mathbf{V}}$(KS) was also usually small, generally falling between 1 percent and 2 percent. The maximum relative bias was 3.4 percent. Because the relative error of **AV**(KS) and the relative bias of $\hat{\mathbf{V}}$(KS) were small even when n_h was as small as 10 or 15, the large-sample requirement of the derivation of the asymptotic variance and variance estimator appeared to be approximately satisfied for sample sizes as low as n_h = 10.

Confidence interval properties were related to sample size, reflecting the additional requirement of confidence intervals for an approximate normal distribution of KS. As expected, observed coverage usually improved as the sample size increased. But even at the smallest sample sizes, observed coverage was not poor for all populations. Coverage was poorest for STRAT4 and STRAT8, but observed coverage was 91.5 percent or better for the smallest sample sizes evaluated for all other populations (nominal coverage of 95 percent). For n_h = 25, observed coverage was 90.0 percent in STRAT3, 92.9 percent in OLDGROWTH, 91.5 percent in STRAT4, and 92.3 percent in STRAT8, but it was between 94 percent and 95 percent in the remaining populations. For n_h = 50 or

Table 2. Properties of KS and $\hat{\mathbf{V}}$(KS) in Simulation Study. Results are Based on 10,000 Replications of Stratified Random Sampling Using Equal Allocation with n_h Observations Per Stratum. Relative Error of **AV**(KS) is [**AV**(KS)−**V**(KS)]/**V**(KS), and Relative Bias of $\hat{\mathbf{V}}$(KS) is {E[$\hat{\mathbf{V}}$(KS)] − **V**(KS)}/**V**(KS).

Population	KAPPA	n_h	Estimated Bias of KS	$\sqrt{\mathbf{V}(\text{KS})}$	Relative Error of **AV**(KS)	Relative Bias of $\hat{\mathbf{V}}$(KS)	Observed Coverage (%) (Nominal 95%)
OLDGROWTH	0.6389	15	−0.001	0.1401	0.009	−0.005	91.5
		25	0.001	0.1095	−0.011	−0.019	92.9
		50	0.002	0.0772	−0.009	−0.016	93.8
		75	0.000	0.0631	−0.013	−0.016	94.2
BLOCK	0.4544	10	0.001	0.1336	0.012	0.004	93.3
		25	0.001	0.0850	−0.007	−0.011	94.2
		50	0.000	0.0601	−0.019	−0.021	94.3
		75	0.000	0.0482	0.004	0.002	94.8
DIAGONAL	0.6539	10	0.001	0.1144	−0.003	−0.009	91.6
		25	0.001	0.0722	−0.007	−0.009	94.1
		50	−0.000	0.0504	0.009	0.008	94.7
		75	−0.001	0.0409	0.007	0.009	95.0
AIRPORT1	0.6845	15	0.000	0.0912	−0.010	−0.012	92.1
		25	0.001	0.0694	0.018	0.017	94.2
		50	−0.000	0.0499	−0.027	−0.027	94.3
		75	0.000	0.0398	0.007	0.007	94.6
STRAT3	0.8053	25	0.000	0.0655	−0.022	−0.027	90.0
		50	0.000	0.0456	0.005	−0.000	92.6
		75	0.000	0.0370	0.011	0.007	93.7
MASSLAND	0.4785	10	0.000	0.1229	−0.007	−0.007	92.9
		25	0.002	0.0774	0.000	−0.000	94.3
		50	−0.000	0.0543	0.013	0.013	94.8
		75	0.000	0.0444	0.005	0.006	94.9
GREEN	0.6533	10	0.002	0.0805	0.038	0.034	94.1
		25	0.000	0.0518	−0.002	−0.001	94.7
		50	−0.000	0.0366	−0.011	−0.011	94.6
		75	0.001	0.0295	0.003	0.002	94.8
STRAT4	0.7729	10	0.002	0.0991	0.037	−0.014	87.1
		25	0.000	0.0641	−0.015	−0.030	91.5
		50	0.000	0.0448	0.001	−0.008	93.6
		75	0.000	0.0365	−0.003	−0.008	94.1
BLIGHT	0.7544	15	0.001	0.0559	−0.017	−0.024	93.1
		25	0.001	0.0426	0.009	0.001	94.1
		50	0.000	0.0300	0.005	0.004	94.7
		75	0.000	0.0244	−0.001	−0.003	94.8
STANDCON	0.7184	15	−0.001	0.0557	0.002	−0.004	93.9
		25	−0.000	0.0436	−0.024	−0.026	94.0
		50	0.000	0.0302	0.007	0.006	94.8
		75	−0.000	0.0246	0.003	0.002	94.7
STRAT8	0.8530	10	0.001	0.0533	0.023	0.006	87.0
		25	−0.000	0.0342	−0.008	−0.012	92.3
		50	−0.000	0.0239	0.008	0.005	94.0
		75	−0.000	0.0196	−0.016	−0.017	94.0

n_h = 75, observed coverage was between 93.6 percent and 95 percent for all populations except STRAT3 (92.6 percent at n_h = 50).

To illustrate the effect of ignoring the stratified design, simulation results (Table 3) are also presented for KHAT and $\hat{\mathbf{V}}$(KHAT), the estimated variance of KHAT (Hudson and Ramm 1987). Recall that stratified random sampling does not satisfy the sampling model under which KHAT and $\hat{\mathbf{V}}$(KHAT) were derived. For the 11 populations studied, bias of KHAT was generally within what might be considered tolerable limits in practice, as the absolute bias exceeded 0.04 in only three populations (STRAT3, STRAT4, and BLIGHT). The observed coverage of confidence intervals constructed using KHAT and $\hat{\mathbf{V}}$(KHAT) was close to the nominal 95 percent in seven populations, but for the other four populations coverage was poor, and became poorer as sample size increased. In these four populations, the bias of KHAT remained the same as sample size increased, but $\hat{\mathbf{V}}$(KHAT) decreased with increasing sample size. Therefore, observed coverage decreased because the intervals were often too narrow to compensate for the bias of KHAT so that the interval did not cover the true KAPPA. From just these 11 populations, it is difficult to ascertain when it is safe to use the simpler formulas KHAT and $\hat{\mathbf{V}}$(KHAT). If proportional allocation is employed, or equal allocation is used and the strata all have approximately the same N_h, KHAT is likely to be nearly unbiased, but it is unclear if $\hat{\mathbf{V}}$(KHAT) will result in adequate confidence intervals for KAPPA. KHAT is not a consistent estimator of KAPPA for the stratified design, so using KHAT when the design is stratified must also be discouraged on this theoretical basis.

The results shown in Table 2 provide some guidance on sample size selection if the objective of accuracy assessment is to estimate KAPPA. A sample size of at least 25 pixels per

TABLE 3. PROPERTIES OF SIMPLE RANDOM SAMPLING ESTIMATORS WHEN APPLIED TO A STRATIFIED RANDOM DESIGN. RESULTS ARE BASED ON 10,000 REPLICATIONS OF STRATIFIED RANDOM SAMPLING USING EQUAL ALLOCATION WITH n_h OBSERVATIONS PER STRATUM. VARIANCE OF KHAT USED IN CONSTRUCTING CONFIDENCE INTERVALS WAS ESTIMATED USING THE FORMULA PRESENTED IN HUDSON AND RAMM (1987).

		Estimated Bias of KHAT			Observed Coverage (Nominal 95%)		
Population	KAPPA	25	50	75	25	50	75
OLDGROWTH	0.6389	0.020	0.020	0.021	93.9	92.8	91.0
BLOCK	0.4544	0.000	0.000	0.000	95.2	95.6	95.6
DIAGONAL	0.6539	−0.005	−0.004	−0.005	93.8	94.9	95.7
AIRPORT1	0.6845	−0.001	−0.001	−0.002	92.3	95.7	95.3
STRAT3	0.8053	−0.056	−0.055	−0.055	89.5	79.5	71.6
MASSLAND	0.4785	−0.008	−0.009	−0.008	94.4	94.6	94.3
GREEN	0.6533	0.000	−0.000	0.003	96.5	97.0	96.8
STRAT4	0.7729	−0.052	−0.053	−0.053	88.1	74.7	62.8
BLIGHT	0.7544	−0.044	−0.045	−0.045	89.9	77.1	64.1
STANDCON	0.7184	−0.018	−0.018	−0.018	95.5	93.9	92.7
STRAT8	0.8530	−0.017	−0.017	−0.017	94.0	88.9	85.5

stratum is needed to guarantee good coverage properties of confidence intervals, although observed coverage of confidence intervals was adequate for smaller sample sizes in some populations. Excellent coverage properties may be expected for sample sizes of 50 or higher per stratum. An increase in sample size from 25 to 50 pixels per stratum resulted in an average reduction in standard deviation of 0.019 for the ten populations studied, while an increase in sample size from 50 to 75 pixels per stratum resulted in an average standard deviation decrease of only 0.009. Because confidence interval properties were not much better at $n_h = 75$ compared to $n_h = 50$, increasing the sample size from 50 to 75 does not appear to provide any meaningful advantages for the objective of estimating KAPPA. Other objectives of a sampling design for accuracy assessment may require additional sample size considerations.

Summary

An estimator and variance estimator applicable for estimating KAPPA under stratified random sampling have been derived. These formulas allow practitioners the flexibility to apply a stratified sampling design for accuracy assessment while still being able to estimate KAPPA and the variance of the estimator of KAPPA. Previously, estimators were only available for simple random sampling. Based on the empirical investigation, bias of KS and $\hat{\mathbf{V}}$(KS) is negligible, and **AV**(KS) provides a good approximation to **V**(KS). Confidence intervals constructed using KS and $\hat{\mathbf{V}}$(KS) generally possess the specified nominal coverage, with the exceptions to this good behavior occurring at small sample sizes in some populations. A sample size of 25 pixels per stratum is the recommended minimum to assure adequate confidence interval coverage, and increasing the sample size to 50 pixels per stratum produces even better coverage and a meaningful reduction in the standard error of the estimator of KAPPA. A further increase in sample size from 50 to 75 pixels per stratum does not appear warranted on the basis of this investigation.

Acknowledgments

This research has been supported by cooperative agreement CR821782 between the Environmental Protection Agency and SUNY-ESF. This manuscript has not been subjected to EPA's peer and policy review, and does not necessarily reflect the views of the Agency. Ray Czaplewski and an anonymous reviewer provided helpful suggestions, and Mary Beth Ritter provided technical editing assistance.

References

Agresti, A., 1989. *Categorical Data Analysis*, John Wiley and Sons, New York, 558 p.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland, 1975. *Discrete Multivariate Analysis Theory and Practice*, MIT Press, Cambridge, Massachusetts, 557 p.

Cochran, W.G., 1977. *Sampling Techniques* (3rd Ed.), John Wiley and Sons, New York, 428 p.

Cohen, J., 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20:37–46.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data, *Remote Sensing of Environment*, 37:35–46.

Congalton, R.G., R.G. Oderwald, and R.A. Mead, 1983. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques, *Photogrammetric Engineering & Remote Sensing*, 49:1671–1678.

Congalton, R.G., K. Green, and J. Teply, 1993. Mapping old growth forests on national forest and park lands in the Pacific Northwest from remotely sensed data, *Photogrammetric Engineering & Remote Sensing*, 59:529–535.

Fiorella, M., and W.J. Ripple, 1993. Determining successional stage of temperate coniferous forests with Landsat satellite data, *Photogrammetric Engineering & Remote Sensing*, 59:239–246.

Green, E.J., W.E. Strawderman, and T.M. Airola, 1993. Assessing classification probabilities for thematic maps, *Photogrammetric Engineering & Remote Sensing*, 59:635–639.

Hudson, W.D., and C.W. Ramm, 1987. Correct formulation of the Kappa coefficient of agreement, *Photogrammetric Engineering & Remote Sensing*, 53:421–422.

Rosenfield, G.H., and K. Fitzpatrick-Lins, 1986. A coefficient of agreement as a measure of thematic classification accuracy, *Photogrammetric Engineering & Remote Sensing*, 52:223–227.

Särndal, C.E., B. Swensson, and J. Wretman, 1992. *Model-Assisted Survey Sampling*, Springer-Verlag, New York, 694 p.

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

(Received 1 February 1994; accepted 4 August 1994; revised 10 October 1994)

Appendix

The necessary general results for deriving the asymptotic variance and estimated variance of KS are summarized by Särndal *et al.* (1992, Result 5.5.1). To aid the reader, the derivation presented is designed to follow the formulas and notation of Särndal *et al.* (1992) as closely as possible. This sometimes requires writing equations in the general forms used by Särndal *et al.* (1992), then proceeding to the special case and more familiar formulas of stratified sampling.

The first step is to write KAPPA as a function of population totals, $t_0, t_1, \ldots, t_q$. In particular, parameters of the population error matrix such as the N_{hj}'s and M_j's may be written as totals of specific indicator or "dummy" variables defined on each of the N pixels (see for example, Equations A4 and A7). We then have

$$\text{KAPPA} = f(t_0, t_1, \ldots, t_q) = \frac{Nt_0 - \sum_{j=1}^{q} N_j t_j}{N^2 - \sum_{j=1}^{q} N_j t_j}, \tag{A1}$$

where $t_0 = D = \sum_{h=1}^{q} N_{hh}$, and $t_j = M_j$ for $j = 1, 2, \ldots, q$. Alternatively, in terms of notation used by Särndal *et al.* (1992), we can define the sum of the diagonal elements of the population error matrix as

$$t_0 = \sum_{h=1}^{q} \sum_{i=1}^{Nh} y_{0,hi} = \sum_{k=1}^{N} y_{0,k}, \qquad \text{(A2)}$$

where

$$y_{0,hi} = \begin{cases} 1, \text{ if pixel } i \text{ of stratum } h \text{ is in column } h \text{ (on the diagonal)} \\ 0, \text{ otherwise (off the diagonal of the error matrix)} \end{cases} \qquad \text{(A3)}$$

and

$$y_{0,k} = \begin{cases} 1, \text{ if pixel } k \text{ is on the diagonal of the error matrix} \\ 0, \text{ otherwise} \end{cases} \qquad \text{(A4)}$$

Similarly, define the column totals of the population error matrix as

$$t_j = M_j = \sum_{h=1}^{q} \sum_{i=1}^{Nh} y_{j,hi} = \sum_{k=1}^{N} y_{j,k} \text{ for } j = 1,\dots, q, \qquad \text{(A5)}$$

where

$$y_{j,hi} = \begin{cases} 1, \text{ if pixel } i \text{ of stratum } h \text{ is in column } j \\ 0, \text{ otherwise} \end{cases} \qquad \text{(A6)}$$

$$y_{j,k} = \begin{cases} 1, \text{ if pixel } k \text{ is in column } j \\ 0, \text{ otherwise} \end{cases}. \qquad \text{(A7)}$$

The general results of Särndal *et al.* (1992), which do not require the stratified population structure, use indicator variables (Equations A4 and A7), while the special case stratified random sampling formulas are more conveniently written using the indicator variables defined by Equations A3 and A6.

KAPPA is estimated by replacing the totals $t_0, t_1,\dots, t_q$ in Equation A1 by unbiased estimators of these totals. The general form of Särndal *et al.* (1992) for unbiased estimation of a total is $\hat{t} = \sum_{k=1}^{n} y_k/\pi_k$, where π_k is the inclusion probability of sample element k, and summation is over the n elements of the sample. Result 3.7.2 (Särndal *et al.*, 1992, p. 103) establishes the special case form for unbiased estimation of totals under stratified sampling. This requires rewriting the response variable y_k to reflect the stratified structure, and noting that the inclusion probability for any element in stratum h is n_h/N_h. We then have

$$\hat{t}_0 = \hat{D} = \sum_{k=1}^{n} y_{0,k}/\pi_k \text{ (Särndal } et\ al. \text{ (1992) general form)} \qquad \text{(A8)}$$

$$= \sum_{h=1}^{q} (N_h/n_h) \sum_{i=1}^{nh} y_{0,hi} = \sum_{h=1}^{q} (N_h/n_h)n_{hh}, \qquad \text{(A9)}$$

and

$$\hat{t}_j = \sum_{k=1}^{n} y_{j,k}/\pi_k \text{ (Särndal } et\ al. \text{ (1992) general form)} \qquad \text{(A10)}$$

$$= \sum_{h=1}^{q} (N_h/n_h) \sum_{i=1}^{nh} y_{j,hi} = \sum_{h=1}^{q} (N_h/n_h)n_{hj}. \qquad \text{(A11)}$$

The subscripts in Equations A8 through A11 index the elements of the sample. Substituting $\hat{t}_0$ and $\hat{t}_j$ into Equation A1 leads to the estimator KS (Equation 4).

From Equation 5.5.10 (Särndal *et al.*, 1992), the asymptotic variance of KS, based on the population error matrix, is

$$\mathbf{AV}(\text{KS}) = \sum_{k=1}^{N} \sum_{l=1}^{N} (\pi_{kl} - \pi_k\pi_l) \frac{u_k u_l}{\pi_k \pi_l}, \qquad \text{(A12)}$$

where π_{kl} is the second-order inclusion probability for the design (Särndal *et al.*, 1992, p. 31),

$$u_k = \sum_{j=0}^{q} a_j y_{j,k} \text{ (for } k=1,\dots, N), \qquad \text{(A13)}$$

and

$$a_j = \frac{\partial f}{\partial t_j} \mid t_0, t_1,\dots, t_q \text{ for } j = 0, 1,\dots, q. \qquad \text{(A14)}$$

Then, after taking the appropriate partial derivatives of Equation A1, we obtain

$$a_0 = \frac{N}{N^2 - \sum_{h=1}^{q} N_h M_h} = \frac{N}{N^2 - C}, \qquad \text{(A15)}$$

and

$$a_j = N_j \frac{N(t_0 - N)}{(N^2 - C)^2} \text{ for } j = 1,\dots, q, \qquad \text{(A16)}$$

where $C = \sum_{h=1}^{q} N_h M_h$. For stratified random sampling (Särndal *et al.* 1992, Result 3.7.2, p. 103), the general formula (Equation A12) for $\mathbf{AV}$(KS) reduces to

$$\mathbf{AV}(\text{KS}) = \sum_{h=1}^{q} N_h^2 (1 - f_h)\mathbf{V}_h/n_h, \qquad \text{(A17)}$$

where $f_h = n_h/N_h$, $\mathbf{V}_h = \sum_{i=1}^{Nh} (u_{hi} - \overline{U}_h)^2/(N_h - 1)$, $\overline{U}_h = \sum_{i=1}^{Nh} u_{hi}/N_h$, and $u_{hi} = \sum_{j=0}^{q} a_j y_{j,hi}$. u_{hi} is the stratified sampling equivalent of u_k (Equation A13). $\mathbf{AV}$(KS) is the standard formula for the variance of an estimated total under stratified random sampling (Cochran, 1977, Equation 5.10), with the variable u_{hi} serving as the response variable.

An estimator of $\mathbf{AV}$(KS) is obtained from the sample error matrix by replacing u_{hi} with $\hat{u}_{hi}$ (Särndal *et al.*, 1992, p. 174). That is, u_{hi} is estimated by

$$\hat{u}_{hi} = \sum_{j=0}^{q} \hat{a}_j y_{j,hi}, \qquad \text{(A18)}$$

where

$$\hat{a}_0 = \frac{N}{N^2 - \hat{C}}, \qquad \text{(A19)}$$

$$\hat{a}_j = N_j \frac{N(\hat{t}_0 - N)}{(N^2 - \hat{C})^2} \text{ for } j = 1,\dots, q, \qquad \text{(A20)}$$

and $\hat{C} = \sum_{h=1}^{q} N_h \hat{M}_h$. Then from Equation A17, the estimated asymptotic variance of KS is

$$\hat{\mathbf{V}}(\text{KS}) = \sum_{h=1}^{q} N_h^2 (1 - f_h)\hat{\mathbf{V}}_h/n_h, \qquad \text{(A21)}$$

where

$$\hat{\mathbf{V}}_h = \sum_{i=1}^{nh} (\hat{u}_{h,i} - \overline{u}_h)^2/(n_h - 1)$$

$$= (\sum_{i=1}^{nh} \hat{u}_{hi}^2 - n_h\overline{u}_h^2)/(n_h - 1), \qquad \text{(A22)}$$

and

$$\overline{u}_h = \sum_{i=1}^{nh} \hat{u}_{hi}/n_h. \qquad \text{(A23)}$$

Again, in Equations A22 and A23, the subscripts index elements of the sample.

Equations 5 and 6 provide simpler computational formulas for the estimated variance by eliminating the need to calculate $\hat{u}_{hi}$ for every sample pixel. To obtain Equations 5 and 6, note that n_{hh} pixels have $y_{0,hi} = 1$, so that in stratum h,

$$\sum_{i=1}^{nh} \hat{a}_0 y_{0,hi} = n_{hh} \frac{N}{N^2 - \hat{C}}. \qquad \text{(A24)}$$

Also, n_{hh} pixels have $y_{j,hi} = 1$ when $j = h$, so that in stratum h,

$$\sum_{i=1}^{nh} \hat{a}_j y_{j,hi} = n_{hh}\hat{a}_h = n_{hh}N_h \frac{N(\hat{t}_0 - N)}{(N^2 - \hat{C})^2}. \qquad \text{(A25)}$$

Equations A24 and A25 represent the contribution from the diagonal entry of the sample error matrix in stratum h; the contribution to $\sum_{i=1}^{nh} \hat{u}_{hi}$ is $n_{hh} \left[\frac{N}{N^2 - \hat{C}} + N_h \frac{N(\hat{t}_0 - N)}{(N^2 - \hat{C})^2}\right]$, and the contribution to $\sum_{i=1}^{nh} \hat{u}_{hi}^2$ is $n_{hh} \left[\frac{N}{N^2 - \hat{C}} + N_h \frac{N(\hat{t}_0 - N)}{(N^2 - \hat{C})^2}\right]^2$, which is the first term in Equation 6. For any off-diagonal column in stratum h, $y_{j,hi} = 1$ for the n_{hj} pixels in column j, so that each off-diagonal column contributes $n_{hj}\, N_j \frac{N(\hat{t}_0 - N)}{(N^2 - \hat{C})^2}$ to $\sum_{i=1}^{nh} \hat{u}_{hi}$ and $n_{hj} \left[N_j \frac{N(\hat{t}_0 - N)}{(N^2 - \hat{C})^2}\right]^2$ to $\sum_{i=1}^{nh} \hat{u}_{hi}^2$. In stratum h, the contribution of all off-diagonal columns to $\sum_{i=1}^{nh} \hat{u}_{hi}$ is then

$$\frac{N(\hat{t}_0 - N)}{(N^2 - \hat{C})^2} \sum_{\substack{j=1 \\ j\neq h}}^{q} n_{hj}N_j. \qquad \text{(A26)}$$

The contribution of all off-diagonal columns in stratum h to $\sum_{i=1}^{nh} \hat{u}_{hi}^2$ is

$$N^2 \frac{(\hat{t}_0 - N)^2}{(N^2 - \hat{C})^4} \sum_{\substack{j=1 \\ j\neq h}}^{q} n_{hj}N_j^2, \qquad \text{(A27)}$$

which is the second major term of Equation 6. Thus, $\overline{u}_h$ (Equation 5) is $\sum_{i=1}^{nh} \hat{u}_{hi}/n_h$, and u_h^2 (Equation 6) is $\sum_{i=1}^{nh} \hat{u}_{hi}^2$.