Comparison of Systematic and Random Sampling for Estimating the Accuracy of Maps Generated from Remotely Sensed Data

Stephen V. Stehman
SUNY College of Environmental Science and Foresty, 211 Marshall Hall, 1 Forestry Drive, Syracuse, NY 13210

ABSTRACT: Properties of statistical analyses of error matrices generated for accuracy assessment of remote sensing classifications were evaluated for three sampling designs: systematic, stratified systematic unaligned, and simple random sampling (SRS). The population parameters investigated were the proportion of misclassified pixels, P, and the Kappa coefficient of agreement, κ. Systematic designs were generally more precise than SRS for the populations studied, except when sampling in phase with periodicity in a population. Bias of the estimated proportion of misclassified pixels, $\hat{P}$, was negligible for the systematic designs. The common practice of estimating the variance of $\hat{P}$ for systematic designs by using an SRS variance estimator resulted in over- or underestimation of variance, depending on whether the systematic design was more or less precise than SRS. A small simulation study showed that the usual standard error formula for the estimated Kappa coefficient of agreement can perform poorly for systematic designs.

INTRODUCTION

LAND-USE AND LAND-COVER CLASSIFICATION MAPS generated from remote sensing data are valuable management and planning tools, and the importance of assessing the accuracy of land-use and land-cover classifications from remotely sensed data has long been recognized. Reference data are needed to properly assess the accuracy of classifications obtained by remote sensing, and obtaining these data by a statistically valid sampling design provides the mathematical foundation for scientifically rigorous inferences. Reference data are typically used for estimating the overall classification accuracy of the map or the accuracy of individual map categories. Reference data are also used to construct error matrices, and these matrices may be subjected to further analyses, such as comparison of different algorithms used to create classification maps.

Congalton (1988) and Maling (1989) review recommendations on choice of sampling design for accuracy assessment. Many papers focus on hypothesis testing to decide whether or not the map is of acceptable accuracy (Hord and Brooner, 1976; Hay, 1979; Van Genderen *et al.*, 1978; Aronoff, 1982a; Aronoff, 1982b). The hypothesis tests may be based on overall accuracy, or accuracy for individual map categories. Review papers by Iachan (1982) and Bellhouse (1988) provide references to the large body of statistical literature comparing sampling designs when the variable of interest is a quantitative variable. For binary variables, Yates (1948) provides some guidance based on his investigation of sampling in one dimension.

In the context of sampling for accuracy applications, Berry and Baker (1968) stated, "For land-use data, where geographic autocorrelation is known to decline monotonically with increased distance, experiments show that greatest relative efficiency is obtained by systematic sampling." Berry and Baker further cautioned, however, that if the shape of the autocorrelation function is unknown, "a stratified systematic unaligned sample appears to yield both greatest relative efficiency and safety to estimation procedures." Stratified systematic unaligned sampling (SSUS) has received support from several other authors. Ayeni (1982) recommended SSUS, based on efficiency as it "relates to interpolation accuracy," for sampling from Digital Terrain Models. Further support for this design was expressed by Maling (1989, p. 173), "There is increasing awareness that only the unrestricted random sample, or the unaligned stratified systematic sample offer satisfactory statistical possibilities." Rosenfield and Melley (1980) and Rosenfield *et al.* (1982) recommended SSUS, with augmentation of the sample by addition of randomly selected pixels in rare map categories to bring the sample sizes in these categories up to some minimum number. Finally, Campbell (1987, p. 359) stated that, "If the analyst knows enough about the region to make a good choice of grid size, the stratified systematic nonaligned sample is likely to be among the most effective."

Congalton's (1988) comparison of sampling designs for accuracy assessment is one of the few empirical studies specifically addressing sampling in remote sensing. His simulation study of three populations compared five sampling schemes: simple random, stratified random (with geographic strata, not stratification by map class), cluster, systematic, and SSUS. The three populations studied consisted of pixels arranged in a grid pattern, wherein each pixel was assigned the value 0 or 1 depending on whether the classification obtained from the remote sensing data at that pixel was correct or incorrect. The populations studied differed in spatial complexity of the pattern of misclassifications. As stated in his abstract, Congalton summarized his results as follows:

> The results indicate that simple random sampling always provides adequate estimates of the population parameters, provided the sample size is sufficient. For the less spatially complex agriculture and range areas, systematic sampling and stratified systematic unaligned sampling greatly overestimated the population parameters and, therefore, should be used only with extreme caution. Cluster sampling worked reasonably well.

Congalton's concern with bias of systematic designs appears contradictory to Maling's (1989) and Berry and Baker's (1968) statements, as well as to Fitzpatrick-Lins' (1981, p. 345) interpretation, "This technique [SSUS] has been found to be the most bias-free sampling design (Berry and Baker, 1968)."

This study examines some of the issues pertinent to statistical inference for accuracy assessment. The objectives are to describe the important criteria for rigorous statistical comparison of sampling designs, and to clarify some of the confusion surrounding systematic designs. The scope will be limited to investigation of simple random sampling (SRS), and of two systematic designs, stratified systematic unaligned sampling (SSUS) and systematic sampling (SS), focusing on inferences concerning two population parameters: the overall proportion of misclassifications and the Kappa coefficient of agreement.

ESTIMATING THE OVERALL MISCLASSIFICATION PROPORTION

Let the parameter P denote the population proportion of incorrect classifications in an image of N pixels, with each pixel

0099-1112/92/5809–1341$03.00/0

assigned the value of 0 or 1 depending on whether the pixel is classified correctly or incorrectly. Because it is not practical to verify the accuracy of every single pixel in the population, a sampling procedure must be used. Estimation of P is a classical example of a finite population sampling problem. Statistical inferences in finite population sampling are based on the randomization distribution generated by repeated application of the sampling design. This approach to inference is the topic of standard sampling texts such as Kish (1965), Cochran (1977), and Stuart (1984). Familiar designs such as simple random, stratified random, cluster, and systematic sampling are commonly used for inference in surveys.

Let y be the number of pixels misclassified in the sample, and let n be the sample size. Then $\hat{P}=y/n$, the sample proportion of pixels misclassified, is an estimator of the parameter P. Two statistical criteria for comparing sampling designs are that $\hat{P}$ should be unbiased and have small sampling variance, $V(\hat{P})$. $V(\hat{P})$ measures the variability of $\hat{P}$ over the set of all possible samples that could be selected; that is, $V(\hat{P})$ measures the "spread" of the sampling distribution of $\hat{P}$. $V(\hat{P})$ is a parameter and depends on the sampling design. The sampling design with the smallest $V(\hat{P})$ for a given population would be preferred, other considerations such as cost or practical convenience being equal. Formulas for $V(\hat{P})$ are found in most sampling texts. For example, the formula for SRS is (Cochran, 1977, p. 51)

$$V(\hat{P}) = \frac{P(1-P)(N-n)}{n(N-1)}, \tag{1}$$

while the formula for SS is

$$V(\hat{P}) = \frac{P(1-P)}{n}[1 + (n-1)\rho_w], \tag{2}$$

where ρ_w is the correlation coefficient between pairs of units that are in the same systematic sample (Cochran, 1977, p. 209). The variance of the estimator $\hat{P}$ should not be confused with the finite population variance of y, which is $S^2=NP(1\text{-}P)/(N\text{-}1)$ for a binary response variable (Cochran, 1977, p. 51, Equation 3.4). $V(\hat{P})$, not S^2, is the relevant variance for inference concerning P.

In practice, estimation of the unknown parameter $V(\hat{P})$ is often part of the descriptive use of accuracy data. $V(\hat{P})$ must be estimated from the sample data. For SRS, the estimated variance of $V(\hat{P})$ is (Cochran, 1977, p. 52)

$$\hat{v}(\hat{P}) = \frac{(N-n)\hat{P}(1-\hat{P})}{(n-1)N}, \tag{3}$$

The crux of the problem with systematic designs is that an unbiased estimator of $V(\hat{P})$ is unavailable, so this variance has to be approximated. A common strategy is to treat the systematic sample as a simple random sample and use Equation 3 as an approximation of the variance given by Equation 2. In general, $\hat{v}(\hat{P})$ will overestimate the true variance if the systematic design results in a gain in precision over SRS, and underestimate the variance if the systematic design results in a loss of precision relative to SRS.

Comparison of sampling designs on the basis of precision, $V(\hat{P})$, differs from Congalton's (1988) criteria. His comparisons were based on bias of the estimator $\hat{P}$, and ability of each sampling design to provide an unbiased estimate of $P(1\text{-}P)$. For all five sampling designs Congalton investigated, an unbiased estimator of P is available, so bias is not a useful criterion for distinguishing among these designs. For large N, $P(1\text{-}P)=S^2$. Because S^2 is a parameter of the population, it does not change for different sampling designs. Therefore, S^2 is not relevant to comparison of sampling designs on the precision criterion.

SYSTEMATIC SAMPLING

The simplicity and convenience of systematic sampling strongly appeal to practitioners. Confusion about properties of systematic sampling, however, has led to concerns not supported by statistical evidence about its use in practice. The usual source of confusion arises because an unbiased estimator of variance for systematic sampling is unavailable. But lack of an unbiased variance estimator does not imply bias of the estimator of P. Unbiased estimation of $V(\hat{P})$, not of P, is the problem.

Systematic samples have also been characterized as not being equal probability samples. Berry and Baker (1968, p. 93) stated that a systematic selection procedure "implies that all parts of the study area do not have an equal chance of being included in the sample." This may be the source of Congalton's (1988, p. 595) remark, "The major disadvantage of systematic sampling is that the selection procedure implies that each unit in the population does not have an equal chance of being included in the sample." These statements are not true in the remote sensing application in which pixels are the sampling units, if the systematic design has been properly applied with a randomized start.

Systematic samples are equal probability samples because every unit in the population has the same chance of being included in the sample. For example, consider a simple case of systematic sampling of a discrete universe of seven units, $y_1, y_2, \ldots, y_7$. If the systematic sampling interval is $k=3$, one of three possible samples,

Sample 1: y_1, y_4, y_7
Sample 2: y_2, y_5
Sample 3: y_3, y_6

is selected depending on whether the random starting value is 1, 2, or 3, respectively. The probability that a given unit is included in the sample is simply the probability that the sample containing that unit is selected. Because all three samples have probability 1/3 of being selected, all seven units have the same probability, 1/3, of being included in the sample. Extension to two-dimensional systematic sampling of pixels in a square-grid alignment follows the same reasoning.

Another source of confusion about systematic sampling arises because, for some applications, the sample size is not fixed for certain values of the sampling interval, k. In the example just presented, the sample size may be 2 or 3. For systematic designs with fixed sample size, $\hat{P}$ is unbiased for $\hat{P}$. If the sample size is not fixed, $\hat{P}$ is biased, but this bias is trivial as long as the sample size and population size are not both small (Cochran, 1977, p. 206). An unbiased estimator of P for the variable sample size case is simply ky/N. Alternative systematic selection procedures, such as circular systematic sampling (Cochran, 1977) or use of a fractional sampling interval (Murthy, 1967, p. 141), result in fixed sample size and unbiasedness of $\hat{P}$. Two-dimensional versions of these systematic selection procedures exist.

"Representativeness" of systematic samples is another issue that must be considered carefully. Campbell (1987, p. 358) stated, "Because selection of the starting point predetermines positions of all subsequent observations, data derived from systematic samples will not meet requirements of inferential statistics for randomly selected (and therefore representative) observations." Systematic samples certainly meet requirements of descriptive "inferential statistics" and therefore provide "representative observations" for the objective of estimating classification accuracy. Campbell's statement, therefore, should be interpreted to mean that systematic samples do not satisfy the sampling models required for some statistical procedures, such as contingency table analyses.

Investigators sometimes claim that systematic samples are "biased" or not "representative" when the sampling interval is in phase with periodicity in the population. A particular sample from this population may indeed be "unrepresentative," but this assertion can also be applied to any other sampling design. Some individual samples from any design will poorly "represent" the population because of sampling error. Matern (1986, p. 66) provides an informative discussion of this issue. Systematic sampling in phase with a periodic signal results in large $V(\hat{P})$, so the precision of the design is unfavorable in this circumstance. But systematic sampling still permits an unbiased estimator of P even if periodicity is present in the population.

EMPIRICAL COMPARISON OF SAMPLING DESIGNS

Properties of two systematic designs, SS and SSUS, and SRS were evaluated empirically by a simulation study of eight populations (Figures 1 and 2, Table 1). The eight populations were selected to represent a variety of circumstances, but did not exhaust all possible spatial patterns of misclassification. Rather, the empirical study was intended to illustrate some general features of the sampling designs and the proper approach for statistically evaluating designs.

Population DIAGONAL was constructed to create a strongly periodic spatial pattern, while BLOCK was constructed to generate a pattern similar to Congalton's (1988) range population. The other six populations represented subregions of actual land-use images. Boundaries between land-use categories were labeled as misclassification errors to generate spatial patterns of errors corresponding to increased likelihood of misclassification at boundaries of polygons. Several sample sizes, approximately 0.2 percent, 0.5 percent, 1 percent, 3 percent, and 5 percent, were evaluated for each design. While the larger sample sizes may not be realistic for most remote sensing applications, they were included to illustrate a broad range of properties of the sampling designs.

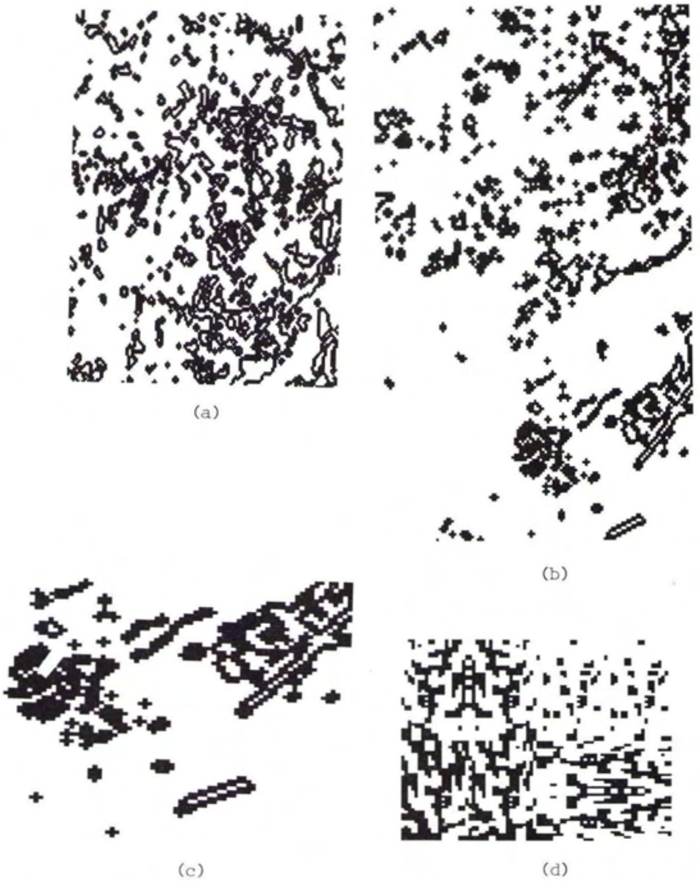

FIG. 1. Difference Images for (a) MASSLAND, (b) AIRPORT2, (c) AIRPORT1, and (d) BLOCK (misclassified pixels are shown in black, correctly classified pixels are shown in white).

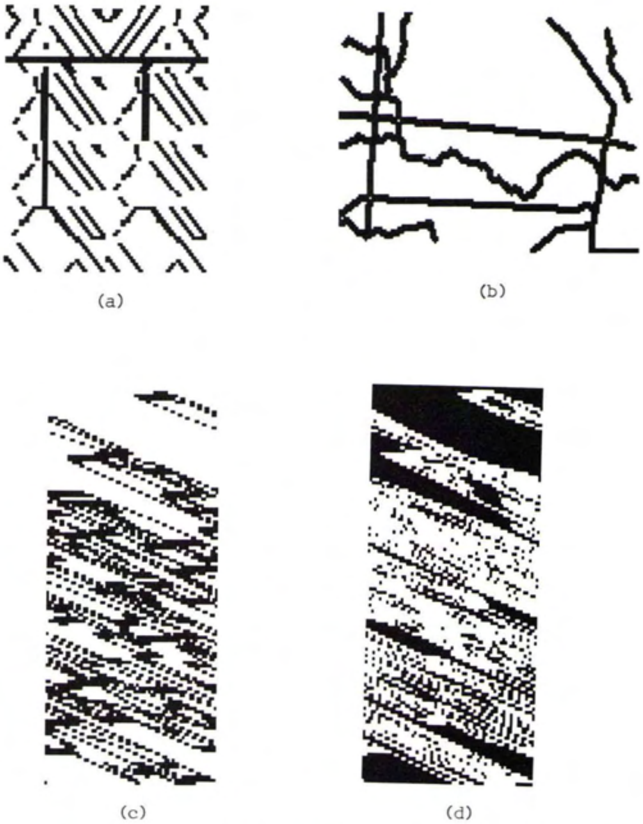

FIG. 2. Difference Images for (a) DIAGONAL, (b) RD&STRM, (c) SOIL, and (d) COMPART (misclassified pixels are shown in black, correctly classified pixels are shown in white).

TABLE 1. DESCRIPTION OF POPULATIONS (ASCII FILES OF ALL DIFFERENCE IMAGES MAY BE OBTAINED BY WRITING TO THE AUTHOR)

Population	Dimensions	Number of Pixels	Proportion Misclassified
AIRPORT1[1]	75 × 75	5,625	0.2091
AIRPORT2[1]	125 × 200	25,000	0.1993
BLOCK	80 × 80	6,400	0.3481
COMPART[2]	123 × 70	8,610	0.3590
DIAGONAL	80 × 80	6,400	0.2300
MASSLAND[3]	175 × 150	26,250	0.3463
RD & STRM[4]	96 × 138	13,248	0.1946
SOIL[5]	123 × 70	8,610	0.4494

[1] Land-use map of a subregion of Worcester Airport Band 4 Thematic Mapper image (portion of AIRPORT image in IDRISI, Version 3.1, Graduate School of Geography, Clark University, Worcester, MA 01610)
[2] Land-use and ownership compartments of Heiburg Forest, Tully, New York
[3] Massachusetts Land-Cover Map (portion of MASSLAND image in IDRISI)
[4] Roads and streams from the west side of Houston, Texas
[5] Soil map of Heiburg Forest, Tully, New York

The primary objective of the simulation study was to compare the precision, $V(\hat{P})$, of the three sampling designs. Secondary

TABLE 2. DESIGN EFFECT* OF SYSTEMATIC SAMPLING AND STRATIFIED SYSTEMATIC UNALIGNED SAMPLING

% of Population Sampled	Design Effect							
	AIRPORT1		AIRPORT2		BLOCK		COMPART	
	SS	SSUS	SS	SSUS	SS	SSUS	SS	SSUS
0.2	1.024	0.946	0.603	0.952	0.858	1.187	0.935	0.972
0.5	0.960	0.869	0.609	0.832	0.898	1.044	0.792	0.916
1.0	0.449	0.800	0.550	0.861	1.691	0.895	1.291	0.938
3.0	0.259	0.559	0.385	0.692	1.322	0.869	1.927	0.934
5.0	0.275	0.524	0.271	0.743	1.572	1.099	1.118	0.868

% of Population Sampled	Design Effect							
	DIAGONAL		MASSLAND		RD & STRM		SOIL	
	SS	SSUS	SS	SSUS	SS	SSUS	SS	SSUS
0.2	1.162	1.430	1.107	0.949	0.685	1.139	0.753	0.784
0.5	1.276	1.434	0.872	0.954	0.899	1.047	0.654	0.706
1.0	3.688	1.714	0.606	0.874	1.239	0.928	0.597	0.700
3.0	7.654	2.083	0.549	0.841	0.263	1.028	0.601	0.558
5.0	3.957	2.744	0.363	0.774	0.124	0.931	0.126	0.566

* The design effect multiplied by 1,000 is the number of observations required in a systematic design to provide the same precision as SRS of 1,000 observations.

objectives were to evaluate the standard practice of using $\hat{v}(\hat{P})$ (Equation 3) to approximate the variance of $\hat{P}$ for a systematic design, and to verify the theoretically known result that bias of $\hat{P}$ is negligible for systematic designs. The latter objective was necessary in light of statements presented by Congalton (1988), which imply that the bias of $\hat{P}$ is significant for systematic designs.

The GAUSS programming language* was used for all simulations. For each population and sample size, 1,500 samples were simulated, and $\hat{P}$ and $\hat{v}(\hat{P})$ were calculated for each sample. The expected values of $\hat{P}$ and $\hat{v}(\hat{P})$ were calculated by averaging the values of each estimate over the 1,500 samples. These estimated expected values were then compared to the parameters P and $V(\hat{P})$ for each design. For SRS, $V(\hat{P})$ was calculated directly from Equation (1). Direct calculation of $V(\hat{P})$ for the systematic designs requires a time-consuming enumeration of all possible samples, so $V(\hat{P})$ was estimated by simulation using the formula

$$V(\hat{P}) = \sum_{i=1}^{1,500} (\hat{P}_i - P)^2/1,500. \tag{4}$$

Precision of the two systematic designs was assessed relative to SRS by calculating the "design effect" (Kish, 1965, p. 258), the ratio of $V(\hat{P})$ for a systematic design to $V(\hat{P})$ for SRS. The design effect multiplied by 1,000 is the number of observations required of a systematic design to provide the same precision as SRS of 1,000 observations. Conversely, the reciprocal of the design effect is the number of observations required of SRS to achieve the same precision as a systematic design of 1,000 observations. If the design effect exceeds 1, SRS has better precision than the systematic design.

Results were difficult to generalize because precision of the systematic designs depended on the particular spatial configuration of misclassifications (Table 2). Even a ranking of the three designs was not always possible within a specific population because the ordering of precision varied for different sample sizes. For example, for population BLOCK, the design effect for SS was below 1 for the 0.2 percent and 0.5 percent samples, but increased to well above 1 for the 1 percent, 3 percent, and 5 percent samples. For SSUS of the same population, the design effect was greater than 1 at 0.2 percent, 0.5 percent, and 5 percent sampling, but decreased below 1 for the 1 percent and 3 percent samples. COMPART was another population in which the design effect of both SS and SSUS varied markedly for different sampling fractions.

Based on the populations studied, the two systematic designs were generally as precise as or more precise than SRS. For populations COMPART, SOIL, MASSLAND, AIRPORT1, and AIRPORT2, SSUS was more precise than SRS at all sample sizes examined. SS was more precise than SRS for all sample sizes in populations SOIL and AIRPORT2, and more precise than SRS for all but one sample size in each of populations MASSLAND, RD&STRM, and AIRPORT1. Consistent with Berry and Baker's (1968) empirical results, SS was generally more precise than SSUS, but the design effect of SS showed greater variation than SSUS. Neither systematic design performed well for the strongly periodic DIAGONAL population, and the systematic design effects generally increased with sample size for this population. DIAGONAL is clearly an example of a population for which systematic designs can have poor precision.

Bias of $\hat{P}$ for both SS and SSUS was negligible for all populations, including the periodic populations DIAGONAL and BLOCK (Table 3). These empirical results confirm the theoretical result that bias of $\hat{P}$ is trivial for the sample and population sizes likely in remote sensing applications, and this result holds whether or not periodicity is present in the population.

When the sampling design was SS or SSUS, the ratio of the expected (average) value of $\hat{v}(\hat{P})$ to $V(\hat{P})$ for SRS was nearly 1 for all populations (Table 4). In other words, $\hat{v}(\hat{P})$ estimated $V(\hat{P})$ for SRS even if the actual sampling design was SS or SSUS. Translating this result to practice, if SS or SSUS has better precision than SRS, $\hat{v}(\hat{P})$ overestimates the true variance of the systematic design. If SS or SSUS results in poorer precision than SRS, $\hat{v}(\hat{P})$ underestimates the actual systematic design variance. The magnitude of the over- or underestimation is proportional to the design effect. Murthy (1967, p. 157) reported similar behavior when a variance estimator for SRS was applied to systematic sampling of a quantitative variable.

OTHER ANALYSES OF ERROR MATRICES

Accuracy data are also used to construct error matrices (Story and Congalton, 1986). Contingency table analyses, such as es-

TABLE 3. ESTIMATED BIAS OF $\hat{P}$ FOR SYSTEMATIC DESIGNS AND SELECTED POPULATIONS

% of Population Sampled	Population							
	AIRPORT1		BLOCK		DIAGONAL		MASSLAND	
	SS	SSUS	SS	SSUS	SS	SSUS	SS	SSUS
0.2	−0.0099	0.0002	−0.0013	−0.0030	−0.0014	−0.0036	0.0014	0.0011
0.5	−0.0066	0.0000	0.0006	0.0024	−0.0005	−0.0002	−0.0025	0.0004
1.0	0.0002	−0.0010	−0.0038	−0.0008	0.0024	−0.0007	0.0016	0.0003
3.0	0.0000	0.0002	−0.0008	0.0003	0.0022	0.0004	−0.0005	−0.0004
5.0	0.0000	−0.0005	−0.0012	−0.0007	−0.0003	0.0009	0.0002	−0.0002

*Version 2.0, Aptech Systems Inc., 26250 196th Place Southeast, Kent, WA 98042

TABLE 4. RATIO OF AVERAGE (EXPECTED VALUE) OF $\hat{v}(\hat{P})$ FROM SS AND SSUS TO $V(\hat{P})$ OF SRS

% of Population Sampled	Population							
	AIRPORT1		BLOCK		DIAGONAL		MASSLAND	
	SS	SSUS	SS	SSUS	SS	SSUS	SS	SSUS
0.2	1.16	1.10	1.11	1.10	1.11	1.09	1.02	1.17
0.5	1.07	1.04	1.03	1.02	1.01	1.01	1.01	1.01
1.0	1.04	1.03	1.01	1.02	0.97	1.01	1.01	1.00
3.0	1.00	1.00	0.99	1.00	0.97	1.00	1.00	1.00
5.0	1.02	1.00	1.00	1.00	1.00	0.99	1.01	0.99

timation of the Kappa coefficient of agreement, κ, estimation of parameters of log-linear models (Congalton *et al.*, 1983), and estimation of the conditional Kappa coefficient (Rosenfield and Fitzpatrick-Lins, 1986), are often performed on the error matrix. We should heed Maling's (1989) advice, "No analyses of an error matrix can be done unless the method for collecting the data is known," and Congalton's (1988) suggestion that we must be sure that "the proper sampling approach was used in generating the error matrix on which all future analysis will be based."

Statistical inferences from contingency table analyses of error matrices are based on specific sampling models, rather than the randomization distribution of descriptive surveys. In general, these analyses require independent Poisson, multinomial, or product multinomial sampling (Bishop *et al.*, 1975; Agresti, 1989). For example, the standard error of the estimated Kappa coefficient is derived under the assumption of multinomial sampling (Bishop *et al.*, 1975, p. 396: Agresti, 1989, p. 366). The log-linear model analyses described by Congalton *et al.* (1983) also require one of the three sampling models listed above (Bishop *et al.*, 1975). Multinomial and product multinomial sampling are common designs in accuracy assessment. The multinomial sampling model applies when SRS is employed. The product multinomial sampling model applies when the pixels are stratified by map category, and SRS of pixels is employed within each stratum.

When the reference data are not obtained by a sampling design on which the contingency table analyses are modeled, the assumptions of the statistical analyses are violated. For cluster and systematic designs, the observations may display spatial autocorrelation so that observations are not independent (Upton and Fingleton, 1989, p. 81). Further, if the observations are correlated, the assumptions of the usual Chi-square tests are violated and the Chi-square approximation is not valid. While no direct study of the effects of correlated data on the estimation of the Kappa coefficient and its standard error have been reported, positive correlation of observations has been found to inflate Chi-square statistics for cluster samples (Cohen, 1976) and systematic samples (Fingleton, 1983a). Holt *et al.* (1980) and Skinner *et al.* (1989) described performance of Chi-square tests for other complex sampling designs, while Fingleton (1983b) investigated log-linear model analyses for systematic samples.

Kappa coefficients and standard errors have been calculated from data obtained by systematic sampling (Agbu and Nizeyimana, 1991) and stratified systematic unaligned sampling (Stenback and Congalton, 1990). Gong and Howarth (1990) computed Kappa coefficients and standard errors for a design they called stratified systematic unaligned. The effect of not satisfying the sampling model on these analyses of Kappa coefficients has not been studied in the remote sensing literature. The following empirical investigation explores what consequences may arise in an analysis of the Kappa coefficient for SS or SSUS.

EMPIRICAL ASSESSMENT OF KAPPA COEFFICIENT

The spatial patterns of misclassifications for the difference images of AIRPORT1, BLOCK, DIAGONAL, and MASSLAND were selected for study. For each difference image, correctly classified pixels were randomly assigned to one of the categories represented by the diagonal cells of the population error matrix, and misclassified pixels were randomly assigned to one of the categories represented by the off-diagonal cells of the population error matrix (Table 5). The spatial patterns of misclassifications of the original difference images were retained. Map classes A through D are arbitrary and not related to actual categories of landuse.

The two systematic designs were again investigated, using sampling percentages of 0.2 percent, 0.5 percent, 1 percent, 3 percent, and 5 percent. In place of SRS, an independent random sample (IRS), sampling with equal probability and with replacement, was used because it satisfies the exact multinomial sampling model required for calculating $\hat{v}(\hat{\kappa})$, the estimated variance of $\hat{\kappa}$. The formula for $\hat{v}(\hat{\kappa})$ was obtained from Hudson and Ramm (1987). The same computing formulas were used to calculate $\hat{\kappa}$ and $\hat{v}(\hat{\kappa})$ for all three designs.

For each sampling design and sample size, 5,000 independent samples were simulated. For each sample, $\hat{\kappa}$ and $\hat{v}(\hat{\kappa})$ were cal-

TABLE 5. POPULATION ERROR MATRICES USED IN ANALYSES OF KAPPA COEFFICIENT

		AIRPORT1 (κ=0.6845)					BLOCK (κ=0.4544)			
		Reference Class					Reference Class			
		A	B	C			A	B	C	
Map	A	1750	218	140	Map	A	2165	565	309	
Class	B	330	1331	152	Class	B	678	944	108	
	C	136	200	1368		C	136	432	1063	
		DIAGONAL (κ=0.6539)					MASSLAND (κ=0.4785)			
		Reference Class					Reference Class			
		A	B	C			A	B	C	D
Map	A	1950	351	104		A	5999	2169	1764	152
Class	B	343	1230	107	Map	B	637	1877	486	27
	C	110	457	1748	Class	C	1753	752	8429	271
						D	109	220	751	854

TABLE 6. DESIGN EFFECT OF SYSTEMATIC SAMPLING AND STRATIFIED SYSTEMATIC UNALIGNED SAMPLING FOR ESTIMATING κ

% of Population Sampled	Population							
	AIRPORT1		BLOCK		DIAGONAL		MASSLAND	
	SS	SSUS	SS	SSUS	SS	SSUS	SS	SSUS
0.2	1.03	0.95	0.78	1.00	1.00	1.17	0.97	0.97
0.5	0.90	0.83	0.88	0.97	1.18	1.43	0.82	0.89
1.0	0.43	0.73	1.70	0.89	3.43	1.62	0.57	0.89
3.0	0.20	0.44	0.86	0.57	5.86	1.46	0.36	0.56
5.0	0.21	0.39	1.15	0.81	2.97	1.96	0.42	0.57

culated, and an 80 percent confidence interval for κ was calculated employing the formula $\hat{\kappa} \pm 1.28^{*}\sqrt{\hat{v}(\hat{\kappa})}$. The estimated expected values of $\hat{\kappa}$ and $\hat{v}(\hat{\kappa})$ were obtained by averaging the 5,000 values of $\hat{\kappa}$ and $\hat{v}(\hat{\kappa})$, respectively. For each design, the sampling variance of $\hat{\kappa}$, denoted by $V(\hat{\kappa})$, was estimated by simulation using the formula

$$V(\hat{\kappa}) = \sum_{i=1}^{5,000} (\hat{\kappa}_i - \kappa)^2/5,000, \qquad (5)$$

where $\hat{\kappa}_i$ is the estimated Kappa for the i^{th} replication. Observed confidence interval coverage was the percentage of the 5,000 samples in which the 80 percent confidence interval contained the parameter κ.

The systematic design effects for estimating κ (Table 6) reflected a pattern similar to the design effects for estimating P. The systematic designs provided approximately the same or better precision than IRS, except for population DIAGONAL and the 1 percent and 5 percent systematic samples of population BLOCK. For all three designs, $\hat{\kappa}$ was nearly unbiased for κ, although at the smallest sample sizes (n=14), biases between −0.015 and −0.025 were observed (Table 7).

The practical consequences of applying $\hat{v}(\hat{\kappa})$ to a systematic design are effectively illustrated by examining properties of confidence intervals constructed with $\hat{v}(\hat{\kappa})$. The results for IRS are presented first to verify that the simulation algorithm operated correctly. For IRS, observed confidence interval coverage was close to the expected 80 percent except when sample size was small (Table 8). Sample sizes less than approximately 60 were apparently too small to satisfy the asymptotic (large sample) assumption used in the derivation of $\hat{v}(\hat{\kappa})$, but otherwise the simulation results for IRS were as predicted by theory.

Observed confidence interval coverage for the two systematic designs depended on the design effect for estimating κ. For example, for population DIAGONAL, in which precision of the systematic designs was worse than IRS, observed coverage of the confidence intervals for κ from SS and SSUS was less than the nominal 80 percent. Conversely, for populations AIRPORT1 and MASSLAND, in which SS and SSUS had better precision than IRS, observed confidence interval coverage of SS and SSUS was higher than the nominal 80 percent except for the 0.2 percent samples. The confidence interval coverage properties reflected the general result that $\hat{v}(\hat{\kappa})$ underestimated $V(\hat{\kappa})$ when the systematic designs were less precise than IRS, and overestimated $V(\hat{\kappa})$ when the systematic designs were more precise than IRS.

CONCLUSIONS

Recommendation for use of a systematic design or SRS depends on the spatial pattern of misclassification and the objectives of accuracy assessment in a given application. If the primary objective is to estimate P, sampling designs should be compared on the criterion of precision, $V(\hat{P})$. Based on the populations studied, and consistent with the results reviewed in the Introduction, systematic designs are generally more precise, and therefore use sampling resources more efficiently, than SRS. SS offers the greatest potential gains and losses in precision relative to SRS. If strong periodicity in the spatial pattern of misclassifications is suspected, SS and SSUS should be avoided, unless sufficient information is available to avoid an unfavorable sampling interval (Matern, 1986, p. 66). It is important to recognize that even if such an unfavorable sampling interval were selected, biases of $\hat{P}$ and $\hat{\kappa}$ are still negligible for systematic designs.

If a systematic design is selected, estimation of variance requires special consideration. The common procedure of using $\hat{v}(\hat{P})$ to estimate the systematic design variance does not work well if the design effect is not close to 1. However, because $\hat{v}(\hat{P})$ estimates $V(\hat{P})$ of SRS even when the actual sampling design is SS or SSUS, $\hat{v}(\hat{P})$ provides a conservative estimate (i.e., overestimate) of the systematic design variance if the design effect is less than 1. In this circumstance, the precision of the systematic design is better than that of SRS, but the *estimate* of the systematic design variance will not reflect the improvement in precision. Conversely, if the design effect is greater than 1, the more

TABLE 7. ESTIMATED BIAS OF $\hat{\kappa}$

	AIRPORT1				BLOCK		
	Sampling Design				Sampling Design		
Sample Size*	IRS	SS	SSUS	Sample Size	IRS	SS	SSUS
12	−0.0186	−0.0172	−0.0207	14	−0.0209	−0.0151	−0.0233
29	−0.0068	−0.0135	−0.0078	33	−0.0092	−0.0153	−0.0107
57	−0.0034	−0.0030	−0.0050	64	−0.0030	−0.0056	−0.0032
226	−0.0007	−0.0003	−0.0006	256	−0.0002	0.0016	−0.0003
352	0.0005	−0.0012	−0.0011	400	−0.0006	−0.0013	−0.0016
	DIAGONAL				MASSLAND		
	Sampling Design				Sampling Design		
Sample Size	IRS	SS	SSUS	Sample Size	IRS	SS	SSUS
14	−0.0208	−0.0193	−0.0144	55	−0.0032	−0.0019	−0.0029
33	−0.0072	−0.0039	−0.0068	135	−0.0034	−0.0014	−0.0030
65	−0.0004	−0.0023	−0.0035	263	−0.0024	−0.0003	−0.0014
257	−0.0008	−0.0012	−0.0008	1051	−0.0002	−0.0007	−0.0001
401	−0.0013	−0.0013	0.0001	1642	0.0004	−0.0003	0.0001

* Sample size has been reported in this table to record the actual sample sizes represented by each sampling percentage.

TABLE 8. PERCENT OF OBSERVED CONFIDENCE INTERVALS, $\hat{\kappa} \pm 1.28^{*}\sqrt{\hat{v}(\hat{\kappa})}$, THAT CONTAIN κ (80 PERCENT NOMINAL COVERAGE)

% of Population	Population											
	AIRPORT1			BLOCK			DIAGONAL			MASSLAND		
Sampled	IRS	SS	SSUS	IRS	SS	SSUS	IRS	SS	SSUS	IRS	SS	SSUS
0.2	75	76	74	74	80	74	69	73	67	79	78	80
0.5	76	84	83	78	79	80	79	73	67	79	84	82
1.0	76	95	84	79	67	81	80	46	69	80	90	83
3.0	81	92	91	79	68	85	80	16	61	79	92	86
5.0	80	100	94	78	81	79	80	45	57	80	94	87

serious problem of underestimation of variance will occur if $\hat{v}(\hat{P})$ is used to estimate the systematic design variance. Alternative estimators of $V(\hat{P})$ for systematic designs are available (Wolter, 1985), but these estimators have not been evaluated for use in accuracy assessment.

The empirical study of the Kappa coefficient demonstrated that bias of $\hat{\kappa}$ is negligible for systematic designs. Violation of the multinomial sampling model may result in poor estimation of the standard error of $\hat{\kappa}$, which translates into erroneous reporting of confidence interval coverage percentages. Results for the populations investigated suggest that the spatial patterns of misclassification that result in better precision of systematic designs relative to SRS for estimating P also result in enhanced precision of the systematic designs over IRS for estimating κ. Estimation of κ would then parallel estimation of P, in that systematic designs provide a more precise estimate of the parameter of interest for some populations, but estimation of variance would be approximate rather than unbiased. If unbiased estimation of $V(\hat{\kappa})$ is crucial, SRS is the only design that approximately assures the required sampling model for $\hat{v}(\hat{\kappa})$ is satisfied. Empirical investigation of other population error matrices and spatial patterns are needed to better generalize the practical effect of systematic designs on inference for κ.

Because different analyses of error matrices are based on different sampling models, selection of a sampling design for accuracy assessment is a complicated task. Study objectives may require several different analyses of the data, and selection of a sampling design will depend on the priority assigned to different objectives. The researcher must understand the statistical properties and assumptions of various sampling designs and analyses to choose an effective design adequate for the objectives of the investigation.

ACKNOWLEDGMENTS

Suggestions and insights from W. S. Overton significantly improved the manuscript. R. Czaplewski, D. White, and M. B. Ritter also provided helpful comments. W. E. Bahls assisted with obtaining and constructing populations, and G. Weaver assisted with plotting. Populations AIRPORT1, AIRPORT2, and MASSLAND were used with permission from the IDRISI Project. This research has been supported by cooperative agreements CR815422 and CR816721 between the Environmental Protection Agency and Oregon State University.

REFERENCES

Agbu, P. A., and E. Nizeyimana, 1991. Comparisons between spectral mapping units derived from SPOT image texture and field soil map units. *Photogrammetric Engineering & Remote Sensing* 57: 397-405.

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

Aronoff, S., 1982a. Classification accuracy: a user approach. *Photogrammetric Engineering & Remote Sensing* 48: 1299-1307.

———, 1982b. The map accuracy report: a user's view. *Photogrammetric Engineering & Remote Sensing* 48: 1309-1312.

Ayeni, O. O., 1982. Optimum sampling for Digital Terrain Models: A trend towards automation. *Photogrammetric Engineering & Remote Sensing* 48: 1687-1694.

Bellhouse, D. R., 1988. Systematic Sampling. *Handbook of Statistics, Vol. 6*, (P. R. Krishnaiah and C. R. Rao, eds.), Elsevier Science Publishers: Amsterdam, pp. 125-145.

Berry, B. J. L., and A. M. Baker, 1968. Geographic sampling. *Spatial Analysis: A Reader in Statistical Geography*, (B. J. L. Berry and D. F. Marble, eds.) Prentice-Hall, Inc.: Englewood Cliffs, N.J., pp. 91-100.

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

Campbell, J. B., 1987. *Introduction to Remote Sensing*. Guilford Press: New York, 551 p.

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

Cohen, J. E., 1976. The distribution of the Chi-squared statistic under clustered sampling from contingency tables. *Journal of the American Statistical Association* 71: 665-670.

Congalton, R. G., 1988. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering & Remote Sensing* 54: 593-600.

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.

Fingleton, B., 1983a. Independence, stationarity, categorical spatial data and the chi-squared test. *Environment and Planning A* 15: 483-499.

———, 1983b. Log-linear models with dependent spatial data. *Environment and Planning A* 15: 801-813.

Fitzpatrick-Lins, K., 1978. An evaluation of errors in mapping land use changes for the Central Atlantic Regional Ecological Test Site. *Journal of Research U.S. Geological Survey* 6: 339-346.

———, 1981. Comparison of sampling procedures and data analysis for a land-use and land-cover map. *Photogrammetric Engineering & Remote Sensing* 47: 343-351.

Gong, P., and P. J. Howarth, 1990. An assessment of some factors influencing multispectral land-cover classification. *Photogrammetric Engineering & Remote Sensing* 56: 597-603.

Hay, A. M., 1979. Sampling designs to test land-use map accuracy. *Photogrammetric Engineering & Remote Sensing* 45: 529-533.

Hord, R. M., and W. Brooner, 1976. Land-use map accuracy criteria. *Photogrammetric Engineering & Remote Sensing* 42: 671-677.

Holt, D., A. J. Scott, and P. D. Ewings, 1980. Chi-squared tests with survey data. *Journal of the Royal Statistical Society A* 143: 303-320.

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

Iachan, R., 1982. Systematic sampling: A critical review. *International Statistical Review* 50: 293-303.

Kish, L., 1965. *Survey Sampling*. John Wiley & Sons: New York, 643 p.

Maling, D. H., 1989. *Measurements for Maps: Principles and Methods of Cartometry*. Pergamon Press: New York, 577 p.

Matern, B., 1986. *Spatial Variation* (2nd Edition). Springer-Verlag: New York, 151 p.

Murthy, M. N., 1967. *Sampling Theory and Methods*. Statistical Publishing Society: Calcutta, 706 p.

Quenouille, M. H., 1949. Problems in plane sampling. *Annals of Mathematical Statistics* 20: 355-375.

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.

Rosenfield, G. H., K. Fitzpatrick-Lins, and H. S. Ling, 1982. Sampling for thematic map accuracy testing. *Photogrammetric Engineering & Remote Sensing* 48: 131-137.

Rosenfield, G. H., and M. L. Melley, 1980. Applications of statistics to thematic mapping. *Photogrammetric Engineering & Remote Sensing* 46: 1287-1294.

Skinner, C. J., D. Holt, and T. M. F. Smith, 1989. *Analysis of Complex Surveys*. John Wiley & Sons: New York, 309 p.

Stenback, J. M., and R. G. Congalton, 1990. Using thematic mapper imagery to examine forest understory. *Photogrammetric Engineering & Remote Sensing* 56: 1285-1290.

Story, M., and R. G. Congalton, 1986. Accuracy assessment: a user's perspective. *Photogrammetric Engineering & Remote Sensing* 52: 397-399.

Stuart, A., 1984. *The Ideas of Sampling* (3rd edition). Oxford University Press: New York, 91 p.

Upton, G. J. G., and B. Fingleton, 1989. *Spatial Data Analysis by Example: Volume* 2. John Wiley & Sons: New York, 416 p.

Van Genderen, J. L., B. F. Lock, and P. A. Vass, 1978. Remote sensing: statistical testing of thematic map accuracy. *Remote Sensing of Environment* 7: 3-14.

Wolter, K., 1985. *Introduction to Variance Estimation*. Springer-Verlag: New York, 427 p.

Yates, F., 1948. Systematic sampling. *Philosophical Transactions of the Royal Society A* 241: 345-377.

(Received 26 September 1990; revised and accepted 22 January 1992)

SOFTWARE REVIEW

VGA-ERDAS Image Processing and GIS Software

Product Information

Software Name: VGA-ERDAS, Version 7.5
Release Date: 26 June 1991
Vendor: ERDAS, Inc., 2801 Buford Highway, Suite 300, Atlanta, GA 30329; Phone: (404) 248-9000; Fax: (404) 248-9400

Price:	List	Educational
Individual Modules		
Core	$ 1,500	$ 1,125
GISMO	$ 1,500	$ 1,125
Image Processing	$ 1,500	$ 1,125
Tapes Handling	$ 2,000	$ 1,500
Hard Copy (Ink Jet)	$ 2,000	$ 1,500
Hard Copy (Thermal)	$ 2,000	$ 1,500
Terrain (Topo & 3D)	$ 2,500	$ 1,875
pcARC/Info Live Link	$ 1,000	$ 750
Tablet Digitizing	$ 1,000	$ 750
Data Conversion	$ 2,000	$ 1,500
Bundled Software Option I (includes Core, GISMO, Image Processing	$ 3,500	$ 2,625
Bundled Software Option II (includes Core, GISMO, Image Processing, Tablet Digitizing and Live Link)	$ 5,000	$ 3,750
Educational 5-key Lab Kit (includes Core, GISMO, Image Processing, Tablet Digitizing, Data Conversion and LiveLink and first year software subscription)		$ 9,000

Other Bundles and Pricing Available from ERDAS, Inc.

Distribution Medium: 3½- and 5¼-inch floppy diskettes

Hardware Requirements

Computer Platform: IBM-AT and Compatibles (minimum 80286, recommended 80386 or 80486)
Operating System: DOS 3.1 or later, Expanded Memory Manager Required
Minimum RAM Required: 640 Kbytes plus expanded memory beyond 1 Mb (amount depends on display configuration: 0 for 320 by 200 resolution; 1 Mb for 640 by 400; 1.2 Mb for 640 by 480; 1.9 Mb for 800 by 600; and 3.1 Mb for 1024 by 768)
Hard Disk Space Required: Depends on which modules are installed. Also, extra disk space is required during the installation process for file decompression. A final installation of the ERDAS Root files, Core, GISMO, and Image Processing Modules will consume more than 33 Mb of hard disk space. Other individual modules require as much as 4.5 Mb.