

# Calibration of Remotely Sensed Proportion or Area Estimates for Misclassification Error

Raymond L. Czaplewski and Glenn P. Catts

United States Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins

*Classifications of remotely sensed data contain misclassification errors that bias areal estimates. Monte Carlo techniques were used to compare two statistical methods that correct or calibrate remotely sensed areal estimates for misclassification bias using reference data from an error matrix. The inverse calibration estimator was consistently superior to the classical estimator using a simple random sample of reference plots. The effects of sample size of reference plots, detail of the classification system, and classification accuracy on the precision of the inverse estimator are discussed. If reference plots are a simple random sample of the study area, then a total sample size of 500–1000 independent reference plots is recommended for calibration.*

## INTRODUCTION

Remote sensing is an efficient means of mapping vegetation, land use, or other characteristics of local sites. Summary statistics from these thematic maps estimate the proportion or area of the geographic area in each cover category. Areal estimates are often needed for strategic planning, land management, and resource assessments. Re-

motely sensed areal estimates are often unacceptable unless they are consistent with the definitions and measurement protocol used for reference data (Thomas, 1986; Burk et al., 1988; Poso, 1988). Frequently, reference and remotely sensed classifications are available for a small sample of reference plots. Such reference data can be used to construct a contingency table or cross-tabulation table, called an "error" or "confusion" matrix in the remote sensing literature, to assess the classification accuracy of remotely sensed data (Card, 1982). In addition, the error matrix is an empirical estimate of the probabilistic association between remotely sensed data and reference data, and an error matrix can be used to correct or calibrate for misclassification bias in remotely sensed statistical estimates of cover proportions (Hay, 1988). Failure to correct for even low probabilities of misclassification error can bias areal estimates from remotely sensed data (Card, 1982; Chrisman, 1982; Czaplewski, 1991).

True misclassification probabilities are unknown if they are estimated with a finite sample of reference plots. Thus, estimates of misclassification probabilities contain sampling errors. These sampling errors are propagated into errors in calibrated areal estimates. As the sample size of reference plots increases, propagated errors will decrease. Merits of alternative calibration estimators can be affected by the sample size used to estimate misclassification probabilities.

Brown (1982), in a key review of the multivari-

Address correspondence to Dr. R. L. Czaplewski, Rocky Mountain Forest and Range Expt. Station, 240 West Prospect Street, Fort Collins, CO 80526.

Received 15 October 1990; revised 9 July 1991.

ate calibration literature, identifies two classes of statistical calibration estimators that treat measurement error: 1) classical models that predict the known but imperfect measurements using the unknown true state; and 2) inverse models that predict the true but unknown state using known but imperfect measurements. Based on many simulation studies, neither the classical nor inverse estimator has been shown universally superior (Brown, 1982; Heldal and Spjotvoll, 1988). Much depends upon the specific application and the evaluation criteria. There have been no direct comparisons of alternative probabilistic estimators that calibrate for measurement errors caused by misclassification.

A classical estimator for misclassification bias was introduced into the statistical literature by Grassia and Sundberg (1982), with remote sensing applications by Bauer et al. (1978), Maxim et al. (1981), Prisley and Smith (1987), and Hay (1988). Selen (1986), Mak (1988), and Li et al. (1991) use the classical calibration estimator for misclassification error in a double sampling scheme, where the reference data are considered a large subsample of the population (Heldal and Spjotvoll, 1988). An inverse calibration estimator for classification error was introduced by Tenenbein (1972), with remote sensing applications by Card (1982) and Chrisman (1982). Czaplewski and Catts (1990) give examples that show how these two methods are applied in remote sensing.

## OBJECTIVES

The first objective of our study was to evaluate two classical and inverse calibration estimators that use misclassification probabilities estimated from reference data that are typical of many remote sensing studies. Evaluation criteria included: 1) infeasibility, which gauges the frequency of numerical problems (i.e., estimates do not exist because of singular matrices) and inadmissible estimates (i.e., negative estimates of proportions); 2) bias, which is the consistent difference between estimated proportions and their true values; and 3) dispersion (i.e., converse of precision), which can be described by the size of the covariance matrix for each multivariate estimate. The second objective was to evaluate the effects of classification accuracy, sample size

of reference plots, and detail of the classification system on the calibration results.

## MONTE CARLO SIMULATION STUDY

It is difficult to compare performance of different calibration estimators in a remote sensing study because exact determination of the true proportions is too expensive. Given that the true proportions are not known, one does not know which estimator tends to produce the most accurate results. Also, infeasibility, bias, and dispersion are expectations over a large number of replicate estimates, not one estimate, and numerous replications of remote sensing studies under controlled conditions are not practical. Therefore, two calibration estimators were compared with a hypothetical yet realistic population, in which the true misclassification probabilities were known exactly by definition and numerous replications were inexpensive. The replications were obtained by Monte Carlo simulation, where a pseudorandom number generator was used to simulate replicate samples of reference sites from the same population.

To be informative, the hypothetical population must have realistic properties that are often encountered in practice. Our simulation was based upon detailed remotely sensed classifications from photointerpretation of 1:12,000 23 cm × 23 cm color infrared stereo transparencies from the Piedmont and coastal plain of North Carolina (Catts et al., 1987; Czaplewski et al., 1987). In this geographic area, land use practices and structure of the temperate forest vegetation are complex, fine-grained, and spatially diverse. Field and photointerpreted classifications were available for a systematic sample of 282 Forest Service Forest Inventory and Analysis (FIA) ground plots, each of which was 0.4 ha in size. Field data were taken from the 1982 USDA FIA survey of North Carolina (Sheffield and Knight, 1986). The location of each 0.4 ha field plot was accurately registered to the aerial photography using the field notes and aerial photography.

The land use/land cover classification system included 21 categories defined by FIA. This classification system is hierarchical, and contains categories for general land uses (agriculture, pasture, shrubland, urban, and forest), and 16 combina-

tions of broad forest management classes (planted pine, natural pine, oak-pine, bottomland hardwood, upland hardwood) and forest size classes (sawtimber, poletimber, seedling/saplings, non-stocked). The number of categories in the classification system is designated by  $k$  (i.e.,  $k = 21$ ).

Data from these 282 field plots were used to construct a  $21 \times 21$  matrix of joint classification probabilities ( $P$ ), which contains all of the data necessary to construct an error matrix [see Eqs. (A7) and (A8)]. However, many elements of this matrix equaled zero because certain types of misclassification errors did not occur with the 282 sample plots. In reality, these unobserved errors might have a nonzero probability of occurrence; these types of errors might not be observed because the sample size is too small. Therefore, the matrix of misclassification probabilities was smoothed with a Bayesian method (Fienberg and Holland, 1973). This produced a matrix of hypothetical misclassification probabilities that contained no probabilities equal to zero, although many probabilities were nearly zero. This smoothed probability matrix is a concise but sufficient characterization of a realistic hypothetical population, from which an infinite number of randomized samples can be drawn to simulate reference data for a real population. Smoothing decreased the overall classification accuracy in the original source, but yielded more realistic definitions of hypothetical true misclassification probabilities.

These smoothed probabilities were treated as the true probabilities  $P$  in our simulation study. A pseudorandom number generator and the true joint probabilities ( $P$ ) were used to generate a large number ( $s$ ) of simulated joint probability matrices  $\hat{P}_j$  ( $1 \leq j \leq s$ ), each of which was estimated from a simulated sample of  $m$  reference sites, where  $m$  could be greater than or less than the 282 field plots used to build the unsmoothed joint probability matrix. Because of sampling error, any one simulated probability matrix had some zero probabilities for specific types of misclassification that did not occur in the Monte Carlo sample of  $m$  reference sites. The sample size ( $m$ ) used to compute each  $\hat{P}_j$  was varied between 50 and 2000 reference plots.

Each simulated joint probability matrix  $\hat{P}_j$  was used to estimate the vector of true proportions of each category with both the classical ( $\hat{t}_{c_j}$ ) and inverse ( $\hat{t}_{i_j}$ ) calibration estimators [Eqs. (A10) and

(A11)]. Each vector estimate was compared to the true vector of proportions ( $t$ ), which is known without error in the simulation [Eq. (A4)]. The difference between estimated and true proportions was caused by chance differences (i.e., sampling error) between  $\hat{P}_j$  and  $P$  in each iteration, and the difference in structure between the two estimators [Eqs. (A10) and (A11)].

Since there were  $s$  simulated matrices, there were  $s$  Monte Carlo estimates of the true vector of proportions ( $t$ ) in the population. Bias is defined as the expected difference between an estimate and its true value. Unbiased estimates have an expected difference of zero. A vector of biases ( $b$ ) for each estimator was readily estimated using the true vector of proportions:

$$b_c = \sum_{j=1}^s [t - (\hat{t}_{c_j})] / s, \quad (1)$$

$$b_i = \sum_{j=1}^s [t - (\hat{t}_{i_j})] / s. \quad (2)$$

Likewise, a sample covariance matrix for each estimator was readily computed from the Monte Carlo simulations:

$$Q_c = \sum_{j=1}^s \{[\hat{t}_{c_j} - (\hat{t}_{c_j})][\hat{t}_{c_j} - (\hat{t}_{c_j})]\} / (s - 1), \quad (3)$$

$$Q_i = \sum_{j=1}^s \{[\hat{t}_{i_j} - (\hat{t}_{i_j})][\hat{t}_{i_j} - (\hat{t}_{i_j})]\} / (s - 1), \quad (4)$$

where

$$\hat{t}_c = \sum_{j=1}^s (\hat{t}_{c_j}) / s,$$

$$\hat{t}_i = \sum_{j=1}^s (\hat{t}_{i_j}) / s.$$

## EVALUATION CRITERIA

We compared the two calibration estimators based on their infeasibility, bias, and dispersion. Infeasible estimates occur when  $\hat{R}$  is singular in the inverse estimator [Eq. (A10)], or  $\hat{T}$  or  $\hat{P}'\hat{T}^{-1}$  is singular in the classical estimator [Eq. (A11)], or the estimate contains negative proportions. The percentage of infeasible estimates in the Monte Carlo simulations is used as an index of infeasibility. A scalar index of bias [Eqs. (5) and (6)] was taken as the sum of the absolute value of the bias vector from Eqs. (1) and (2):

$$b_c = \mathbf{1}'\mathbf{b}_c, \quad (5)$$

$$b_i = \mathbf{1}'\mathbf{b}_i, \quad (6)$$

where  $\mathbf{1}$  is a  $k \times 1$  vector of 1's. Dispersion can be partially described by the size of the covariance matrix for estimation errors, which is estimated by  $\mathbf{Q}_c$  and  $\mathbf{Q}_i$  in Eqs. (3) and (4). The best estimator will have the smallest dispersion. For comparative purposes, the  $k^2$  elements of the covariance matrices in Eqs. (3) and (4) were summarized to yield a scalar statistic: the sum of the variance terms on the diagonal of the covariance matrix, that is, the trace of the covariance matrix.

Evaluations based on scalar descriptors of bias vectors and covariance matrices assume that the Monte Carlo estimates are close to their true values, which requires an adequate number of Monte Carlo simulations. Therefore, simulations were continued until the evaluation indices stabilized near consistent values, and the standard deviations of the bias indices in Eqs. (5) and (6) were small. The number of Monte Carlo simulations ( $s$ ) varied from 20,000 to 80,000, depending on the number of simulated reference sites ( $m$ ).

## CLASSIFICATION DETAIL

The complexity of the classification system can affect sample size of reference sites used to estimate any one misclassification probability in  $\mathbf{P}$ . As the number of categories ( $k$ ) in the classification system increases, the proportion of most categories will approach zero; this can increase the probability of infeasible estimates, bias the estimates of joint classification probabilities, and in-

crease the dispersion of the estimation errors caused by a small sample size of reference sites. Therefore, the effects of different numbers of categories ( $k = 4, 10, 14, 21$ ) in the hierarchical classification system were explored for each estimator. The less detailed levels were formed by collapsing the basic classification system in various ways. The resulting eight classification systems vary widely in detail and classification accuracy (Table 1).

## RESULTS

The inverse estimator was consistently superior to the classical estimator based on all evaluation criteria (infeasibility, bias, and dispersion). The classical estimator had a much higher percentage of infeasible solutions than the inverse estimator for all eight levels of classification detail and accuracy (Table 1), as shown in Figure 1. For the simplest classification systems (A and B), which have only four categories, 25–40% of the classical estimates were infeasible based on 50–100 reference plots, whereas almost none of the inverse estimates were infeasible. For the more detailed classification systems C–E (10 categories), 30–80% of the classical estimates were infeasible based on 500 reference sites, whereas almost none of the inverse estimates were infeasible. For the most detailed classification system H (21 categories), all of the classical estimates were infeasible, whereas none of the inverse estimates were infeasible based on 1000 or more reference sites.

Both the classical and inverse calibration estimators were substantially less biased than the uncalibrated remotely sensed estimates (dashed lines in Fig. 2). However the feasible classical estimates were consistently more biased than the inverse estimates (Fig. 2). The inverse estimator was virtually unbiased for all levels of detail in the classification system and all sample sizes of reference sites. Bias from the classical estimator was greater for smaller sample sizes of plots and for the more detailed levels of the classification system. Even with sample sizes of 2000 reference sites, the classical estimator was biased for classification systems C–G. (None of the estimates were feasible for classification system H, the most detailed system.) Even for the simplest classification

Table 1. Descriptive Statistics for Each Level of Classification Detail and Accuracy

System	Number of Classes ( $k$ )	Kappa <sup>a</sup>	Percent Correct
A	4	0.90	94.6
B	4	0.65	74.9
C	10	0.68	72.5
D	10	0.65	69.1
E	10	0.57	62.3
F	14	0.59	63.1
G	14	0.54	58.0
H	21	0.52	55.9

<sup>a</sup> Kappa statistic (Cohan, 1960) equals 0 for accuracy no greater than expected by chance and 1 for perfect accuracy.

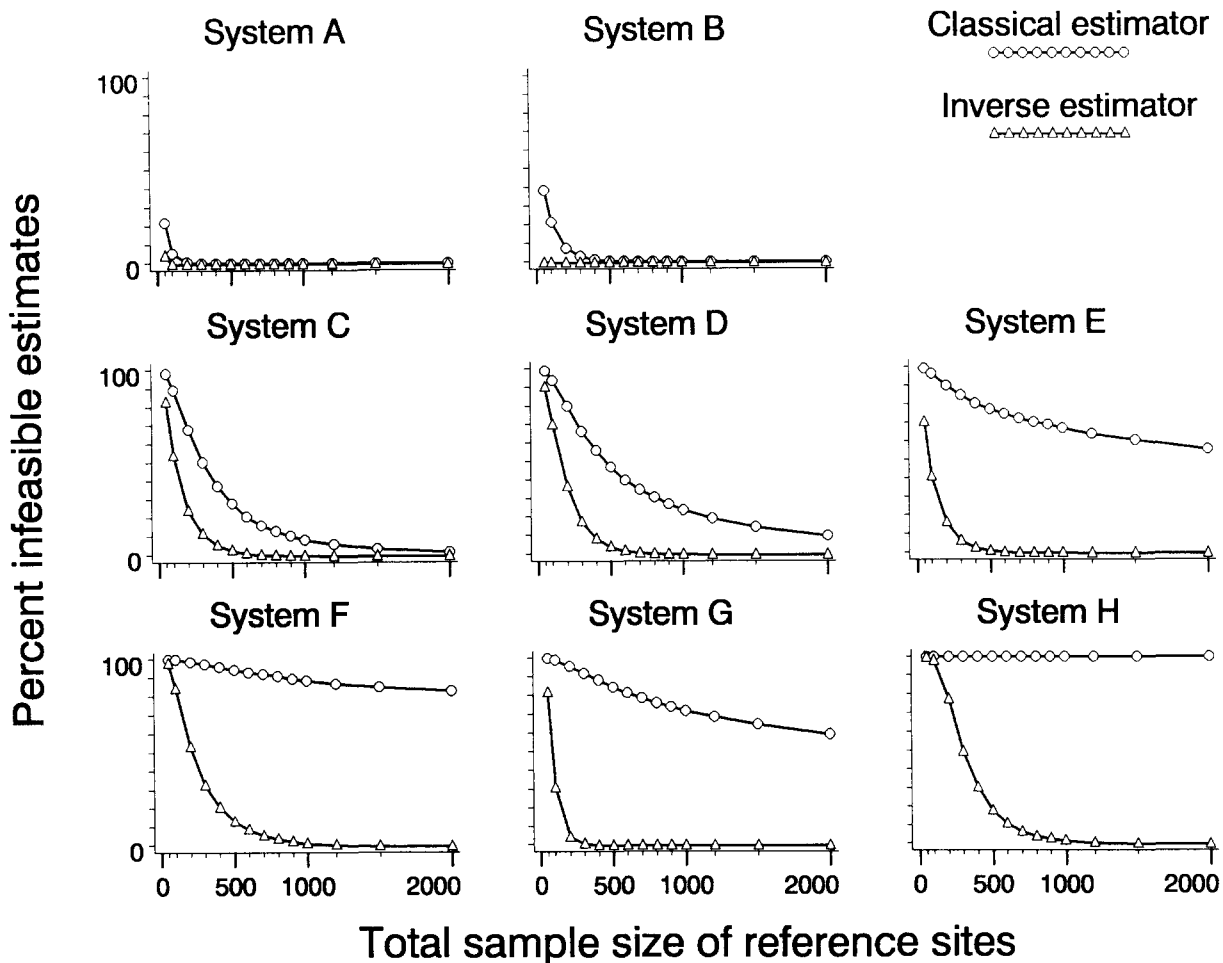


Figure 1. Comparison of classical and inverse estimators based on proportion of infeasible estimates for different levels of classification detail (Table 1) and sample sizes of reference plots. Classification systems A and B have four categories, systems C–E have 10 categories, systems F and G have 14 categories, and system H has 21 categories. The inverse estimator had a smaller percentage of infeasible solutions; in most cases, the inverse estimator was superior to the classical estimator, especially for smaller sample sizes of reference plots and the more detailed levels of classification.

systems (A and B), which had only four categories, the classical estimator required sample sizes of 200–600 reference sites to achieve unbiased estimates (Fig. 2).

The classical estimator had consistently less precision and higher dispersion than the inverse estimator, as shown by the traces for the covariance matrices in Figure 3. For example, the classical estimator required a sample size of 2000 reference sites with classification systems B and C to obtain the same precision as the inverse estimator that used only 200–400 reference sites. This disparity was even greater for the more complex classification systems E–G. The two estimators had comparable precision only with classification

system A, which has high accuracy and contains only four categories.

### Sample Size of Reference Plots

The remote sensing practitioner must determine the value of calibration for misclassification bias in areal estimates. If the expected bias is small relative to user needs, then calibration might not be worth the cost. Czaplewski (1991) provides guidance to help in this decision. If the magnitude of expected misclassification bias is unacceptable, then the practitioner must select a reasonable sample size of reference plots to estimate the

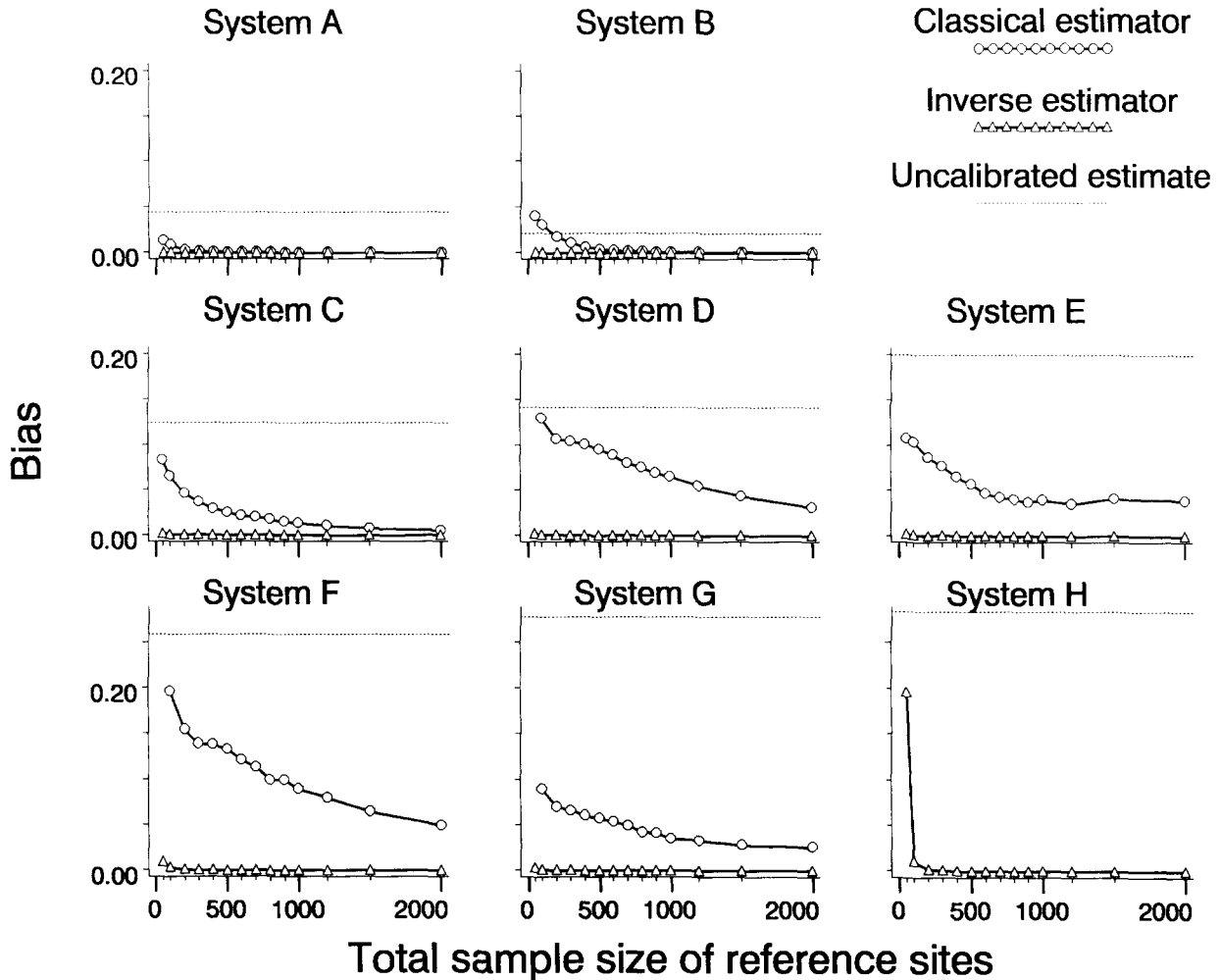


Figure 2. Comparison of classical and inverse calibration estimators based on an index of bias [Eqs. (5) and (6)]. Bias with either calibration estimator was consistently less than bias with uncalibrated estimates (dashed lines). The inverse estimator was unbiased for all levels of classification detail and sample sizes of reference plots. The classical estimator was biased except for very large sample sizes or the very simple classification systems. Bias for the classical estimator in system H is not given since there were no feasible solutions.

misclassification probabilities, which are used for calibration.

Reference plots are expensive, but an inadequate sample size will produce imprecise calibrated estimates. Precision increases with an increase in the sample size of reference plots because there is less sampling error in estimated misclassification probabilities, and less propagation of these sampling errors into estimation errors for the cover proportions. This is apparent in Figure 3, and the same pattern occurs with individual categories, as shown in Figure 4; as sample size of reference plots increases, the coefficient of variation decreases nonlinearly for each category. The coefficient of variation is a

proportion equal to the standard deviation of an estimate divided by the estimate. The relationship between coefficient of variation ( $C_i$ ) for category  $i$  and the total sample size of reference plots ( $m$ ) is approximately linear (Fig. 4), given the following logarithmic transformation

$$\ln(C_i) = a_i + b_i \ln(m). \quad (7)$$

where  $\ln(m)$  is the natural logarithm of the number of reference plots. The slopes of these logarithmic relationships are very similar for all categories (Fig. 5); the principal difference among categories is the intercept of the logarithmic relationship. Figure 5 suggests the intercept for any one category is related to the classification accu-

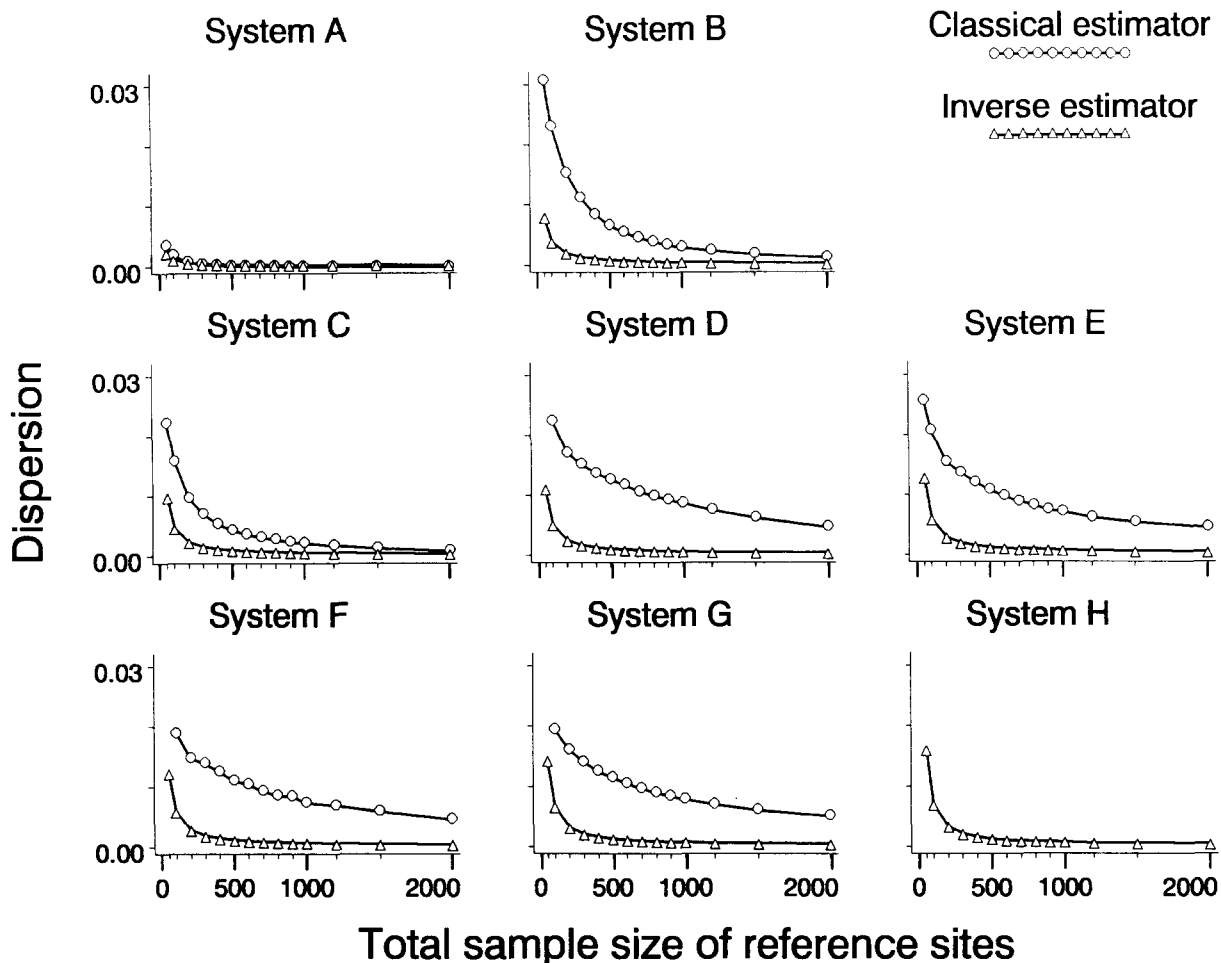


Figure 3. Comparison between classical and inverse estimators based on dispersion (trace of the Monte Carlo covariance matrix); as dispersion decreases, precision increases. The inverse estimator has less dispersion compared to the classical estimator. Dispersion of the classical estimator in System H is not given since there were no feasible solutions.

racy for that category, where accuracy is described by Light's conditional  $\kappa_i$  coefficient of agreement (Light, 1971; Rosenfield and Fitzpatrick-Lins, 1986). This coefficient equals zero when accuracy for a category equals that expected by chance alone and 1 when there is perfect agreement. Categories that are more accurately classified have less propagated error from the calibration model and require fewer reference plots. The intercepts are also related to the remotely sensed proportion ( $p_i$ ) of category  $i$  in the population (Fig. 5). For simple random samples, inaccurately classified or rare categories require relatively larger total sample sizes of reference sites than the more accurately classified or common categories.

Our Monte Carlo results (Fig. 5) were used to fit regression models [Eq. (8)] that predict the

intercepts of the logarithmic relationships in Eq. (7) as functions of conditional accuracy ( $\kappa_i$ ) and proportion ( $p_i$ ) for each category  $i$ :

$$a_i = c_0 + c_1\kappa_i + c_2p_i. \quad (8)$$

Parameter estimates for several versions of this model are given in Table 2, depending on availability of estimates for  $\kappa_i$  or  $p_i$ . Since there was little variation in slopes of these logarithmic relationships (Figs. 4 and 5), the mean overall slope ( $\bar{b} = -0.53$ ) was used to predict the slope for any one category  $i$ . Therefore, the coefficient of variation ( $C_i$ ) for category  $i$  in the Monte Carlo simulations was predicted by

$$\ln(C_i) = c_0 + c_1\kappa_i + c_2p_i - 0.53 \ln(m), \quad (9)$$

where parameter estimates are given in Table 2.

When the inverse estimator is used with a

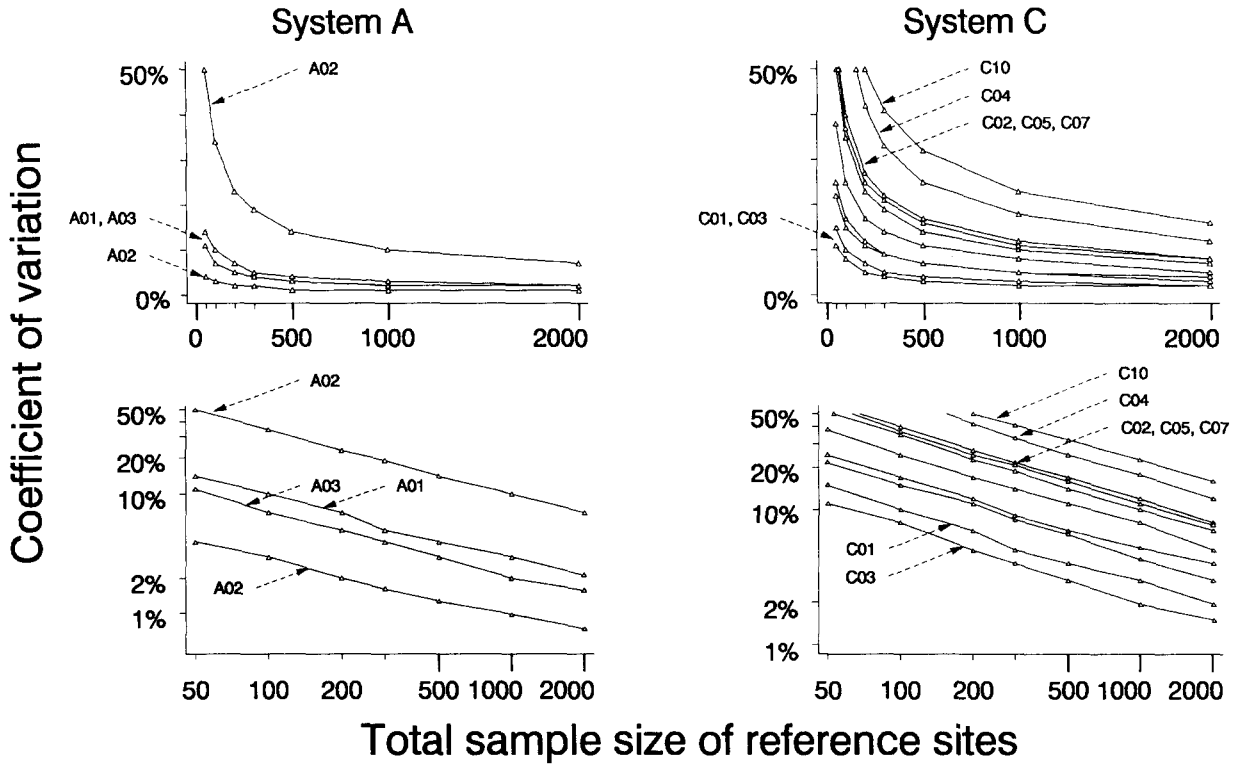


Figure 4. Effect of total sample size of reference sites on the coefficient of variation, which is an index of precision (i.e., converse of dispersion), for individual categories using the inverse calibration estimator. Results are given for two levels of classification detail (classification systems A and C), but are typical of the other levels of detail (Table 1). For a given sample size, the coefficient of variation is higher for categories that are classified with less accuracy (e.g., cover type A02) or rare categories (e.g., cover type C10). There is a logarithmic relationship between coefficient of variation and total sample size of reference sites. The slope of the logarithmic relationship is similar among categories, but the intercept varies considerably depending on classification accuracy and prevalence of each category.

simple random sample of reference sites in other studies, this predictor of the coefficient of variation could be useful to anticipate the total sample size of reference sites required for calibration, assuming preliminary estimates of  $\kappa_i$  or  $p_i$  are available. However, it is not known how well the

results from our Monte Carlo study will extrapolate to other studies; the relationship between precision and sample size might differ for other error matrices. By considering the marginal improvement in the coefficient of variation by incrementally increasing sample size, it is possible to use the slope of the logarithmic relationship ( $\bar{b}$ ) and ignore the less predictable intercept ( $a_i$ ). If  $C_{ij}$  is the coefficient of variation for category  $i$  for a sample size of reference plots  $m_j$ , then the ratio of coefficients of variation for category  $i$  at two alternative sample sizes ( $m_j, j = 1,2$ ) is

Table 2. Coefficients in Regression Models [Eq. (9)] Used to Predict Coefficient of Variation for Any Single Category Given Its Prevalence (Its Proportion in the Population) and the Accuracy Classified Based on Remotely Sensed Data<sup>a</sup>

$c_0$	$c_1$	$c_2$	$R^2$
1.83	-6.74	0.0	0.69
2.35	0.0	-2.04	0.46
2.45	-5.47	-1.32	0.86

<sup>a</sup> These models should only be extrapolated to other remote sensing studies to plan for total sample size of reference sites; they are not to be used to describe variance of estimation errors after reference data are available. If a coefficient equals zero, then the corresponding predictor variable is not used.

$$\frac{C_{i1}}{C_{i2}} = \frac{\exp(a_i) \exp[\bar{b} \ln(m_1)]}{\exp(a_i) \exp[\bar{b} \ln(m_2)]} = \exp[\bar{b}[\ln(m_1) - \ln(m_2)]] \quad (10)$$

This ratio is independent of the intercept  $a_i$  and any parameters unique to category  $i$ ; it depends only on the slope of the logarithmic relationship



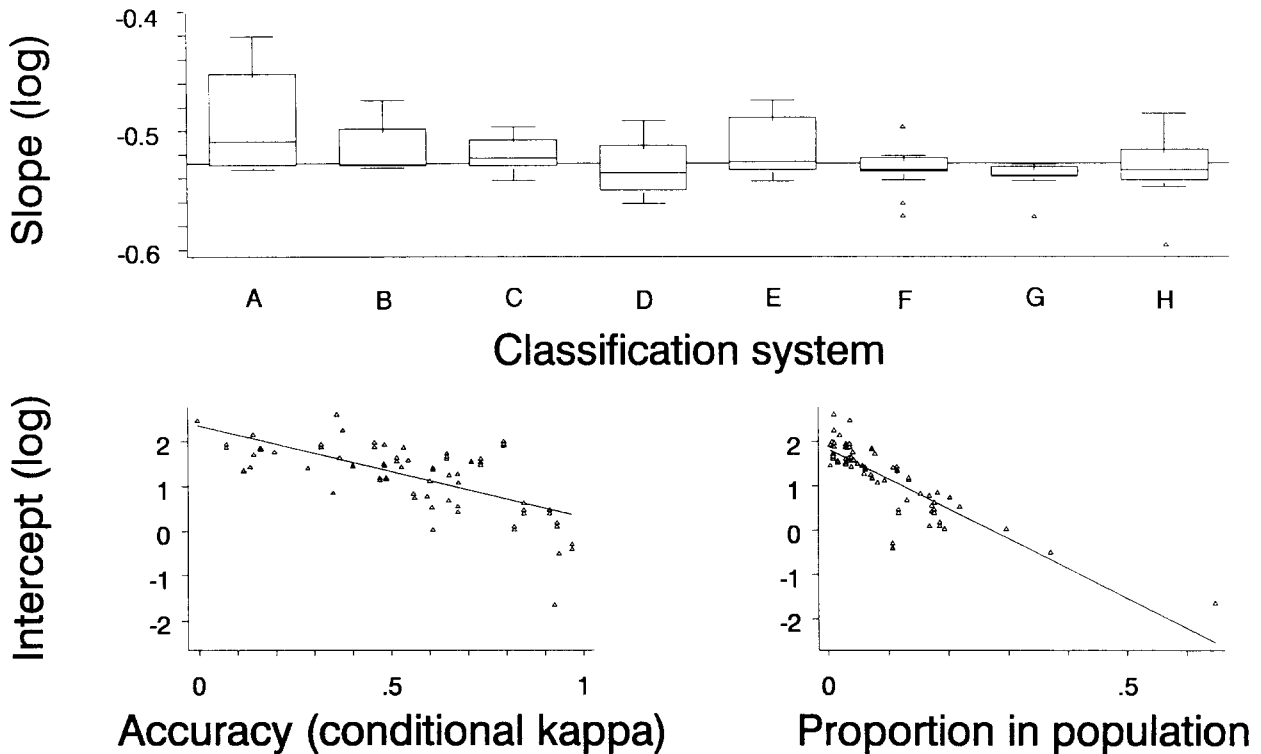


Figure 5. Logarithmic relationships from Figure 4 for all categories in all eight levels of classification detail using the inverse estimator (Table 1). The slope of the logarithmic relationship is similar for all categories within all classification systems; boxplots (Emerson and Strenio, 1983) display the distribution of slopes for each category within a classification system using medians and quartiles. However, the intercept varies considerably. The intercept decreases with increased classification accuracy and increased prevalence (i.e., higher proportion of a category in the population). Classification accuracy is quantified by Light's conditional kappa statistic (Light, 1971); this statistic equals zero if accuracy is no greater than that expected by chance and 1 if there is perfect agreement with reference data. For a given sample size of reference sites, the coefficient of variation is less for smaller intercepts (Fig. 4). It is not known how well these quantitative functions extrapolate to other remote sensing studies. Models for these relationships are given in Eq. (9) and Table 2.

(b), which varies little among different categories within different classification systems (Fig. 5). For a simple random or systematic sample of reference plots, the relative change in the coefficient of variation given a change in the total sample size of reference plots is approximately the same for all categories.

This relationship is illustrated in Figure 6, which gives the percent improvement in coefficient of variation, that is,  $100\%[(C_{i1}/C_{i2}) - 1]$ , when the total sample size of reference plots ( $m_1$ ) is increased by 100 sites ( $m_2 = m_1 + 100$ ). For example, the coefficient of variation for any category is decreased approximately 45% when the total number of reference sites is increased from 100 to 200, 16% when increased from 300 to 400, 12% when increased from 400 to 500, and 10% when increased from 500 to 600 (Fig. 6). The improvement in coefficient of variation for all

categories starts to reach a point of diminishing returns when the total sample sizes of reference sites becomes larger than 500 (Fig. 6), and gains in improving the coefficient of variation by increasing total sample sizes above 1000 are minimal (i.e., less than 5% for every additional 100 reference sites). Therefore, we recommend that a total of 500–1000 independent reference sites be used in other studies, if a simple random sample of reference plots is used.

## DISCUSSION

These results are strictly valid only for the error matrices used in our particular Monte Carlo simulation. It is possible that the classical estimator is superior to the inverse estimator for other error matrices, and further simulations are encouraged.

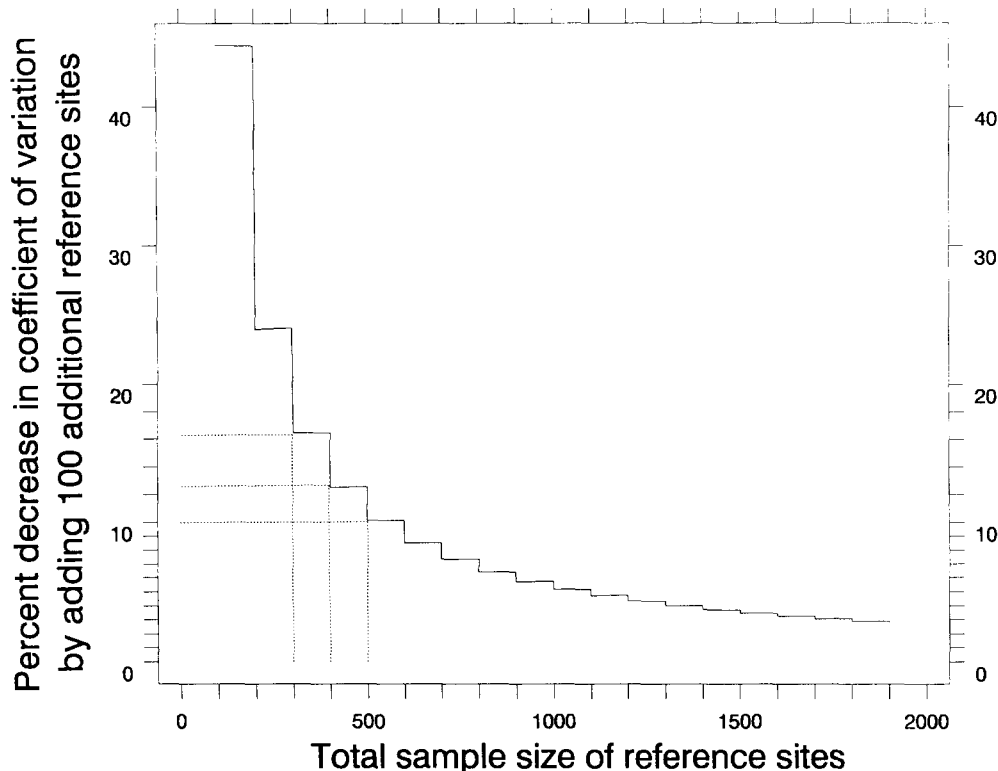


Figure 6. The approximate effect on the coefficient of variation of incrementally increasing total sample size of reference plots by using simple random or systematic sampling [Eq. (10)]. The coefficient of variation describes the estimation precision for each individual category in any of the classification systems when the inverse calibration estimator is used. The dashed lines are examples of this effect. Coefficients of variation for any category decrease approximately 16% when total sample size is increased from 300 to 400 sites; an additional 12% when increased from 400 to 500 sites, and an additional 10% when increased from 500 to 600 sites. Based on this function, we recommend the total sample size be between 500 and 1000 independent reference plots; the improvement is minimal for larger total sample sizes. This function is expected to be more dependably extrapolated to other remote sensing studies than the functions in Figure 5.

However, the error matrices used in this paper are typical for many remote sensing studies, and vary widely in classification detail and accuracy. Therefore, future simulation studies will likely agree with our results.

Emphasis has been placed on estimating proportions of a geographic area in each category of cover. However, estimates of area (e.g., hectares or acres) are more commonly needed. Estimated area in each category can be easily computed from proportion estimates by multiplying  $\hat{t}_c$  or  $\hat{t}_i$  total area in the study area.

The recommended sample sizes of 500–1000 reference plots assumes that the true probabilities of misclassification are constant across the entire study area. If there are subregions that differ in probability of misclassification error (e.g., differ-

ent physiographic regions), then calibrated estimates should be stratified by these subregions, and there should be 500–1000 reference sites within each subregion. If misclassification probabilities vary between different classification processes (i.e., different photointerpreters, dates of image acquisition, or satellite scenes), then an error matrix should be developed for each classification process based on 500–1000 reference sites. Well-known statistical methods are available to test for different probabilities of misclassification (e.g., Cohen, 1960; Congalton et al., 1983), but these require a large number of reference sites to detect differences in classification accuracy. In many cases, prudent judgment of an experienced remote sensing practitioner will be required to assess whether or not misclassification

error probabilities are reasonably constant over a given geographic area, or between different classification processes.

More precise methods for estimating total sample sizes might be possible using variance estimates for the inverse calibration model (Tenenbein, 1972). However, this requires a prior estimate of the error matrix, which may not be available. It also assumes that Tenenbein's approximation for estimating the covariance matrix is adequate for small sample sizes. This assumption needs to be tested in a future study. A total of 500–1000 reference sites is proposed until this recommendation is refined in future studies.

The classical and inverse estimators are derived under the assumption that the reference sites are a simple random sample of the population, although a simple systematic sample might also be used (e.g., Norton-Griffiths, 1988). Suppose, however, that the reference plots are a stratified sample. In this case, an unbiased estimate of the joint classification probabilities cannot be made unless auxiliary data for the marginal probabilities are available. Reference data might be a stratified sample based on the remotely interpreted classifications. This might be the case if the reference data are gathered after the entire geographic area is classified with remote sensing. Then the conditional probabilities [ $\hat{\mathbf{P}}\mathbf{R}^{-1}$  in Eq. (A10)] can be estimated directly, and the inverse estimator can be applied without a direct estimate of  $\hat{\mathbf{P}}$ . Conversely, reference data might be a stratified sample based on the true classifications. This might be the case when existing survey plots are used as reference data, and their true classifications are available independently of their remote interpretation. Then, direct estimates of the conditional probabilities [ $\mathbf{P}'\mathbf{T}^{-1}$  in Eq. (A11)] are available for the classical estimator, and the classical estimator can be applied without a direct estimate of  $\mathbf{P}$ . If unbiased estimates of the marginal probabilities are not available, then unbiased estimates of  $\mathbf{P}$  are not available, and the calibration estimators might be biased if an inappropriate calibration estimator is applied to a stratified sample of reference plots.

It has been assumed that reference sites are a randomized sample of the population. However, reference sites are often purposefully selected in remote sensing studies, without randomization techniques. Such plots are frequently convenient

for training or labeling digital classifiers. Calibration using these types of reference sites might be acceptable if nearly unbiased estimates of the misclassification probabilities ( $\hat{\mathbf{P}}$ ) are produced. Brown (1982) shows that such purposeful sampling (i.e., controlled calibration) can be justified on Bayesian grounds. Unfortunately, it is difficult to verify that purposefully selected reference plots produce approximately unbiased estimates of  $\mathbf{P}$  (Card, 1982), especially when the same plots are used both to build a digital classifier and estimate a calibration model.

Sometimes the number of categories or their definition in the true classification system differs from those in the remotely classified system. Unsupervised digital classification of satellite data is an example, where the number of spectral categories is often greater than the number of categories in the target classification system. In this case, the matrix of misclassification probabilities ( $\mathbf{P}$ ) will be rectangular. It is necessary to label each spectral category to produce a thematic map, but labeling is not necessary to produce statistical estimates. For a rectangular error matrix, the inverse estimator  $\hat{t}_i$  will retain the form given in Eq. (A10), as discussed by Tenenbein (1972), Hochberg (1977), and Card (1982); however, the classical estimator in Eq. (A11) will not exist because the matrix inverse does not exist for a rectangular error matrix.

One of our evaluation criteria was error dispersion of the estimators, which was estimated with sample covariance matrices from Monte Carlo simulations. The objective of our study was to evaluate two calibration estimators, not estimate proportions in a specific study area. In practice, only one realization of reference data is available, and methods other than Monte Carlo simulations would be needed to estimate the covariance matrix for the estimation errors. Estimators for variance terms on the diagonal of the covariance matrix are given by Card (1982) for the inverse estimator, and by Maxim et al. (1981) for the classical estimator. Tenenbein (1972) gives an asymptotically unbiased estimator for the entire covariance matrix for the inverse estimator, while Grassia and Sundberg (1982) and Heldal and Spjøtvoll (1988) present linear approximations of the covariance matrix for the classical estimator. These approximations are unbiased for large sample sizes, but their bias under a small sample size

of reference data is not well known. Therefore, the objectives of our study were met with sample covariance matrices from the Monte Carlo simulations, which are considered more accurate for small sample sizes of simulated reference sites when the true classification probabilities are known. Future studies are needed to evaluate estimators of the covariance matrices cited above.

It has been assumed that the reference plots are a small, independent sample of the population. If the reference data are a larger subsample of the remotely interpreted sites, there might be more efficient estimators that treat reference data as a double sample (e.g., Li et al., 1991). However, the sampling design in a double sample might effect the estimator. If the reference data are a stratified sample based on true classifications, the classical methods of Selén (1986) or Mak and Li (1988) would directly apply; if reference data are stratified on the remotely interpreted classification, then the inverse estimator of Tenenbein (1972) would apply. If the reference data are a simple random or systematic sample, then either the classical or inverse estimator could be used with double sampling.

## CONCLUSIONS

The inverse calibration estimator was more precise and less biased for areal estimates than the classical estimator given the conditions of our simulation study. These conditions are typical of many remote sensing studies in which a simple random sample of homogeneous and accurately registered reference plots are available. However, other types of reference data are also used in remote sensing, such as heterogeneous reference sites, stratified sampling, and purposefully selected reference sites. Future studies are needed to evaluate estimators using these other types of reference data.

It is recommended that sample sizes of 500–1000 reference sites be used to calibrate areal estimates, if the reference sites are homogeneous and a simple random sample of the study area. More precise methods for determining the necessary sample size might be possible using approximate estimators of the covariance matrix for errors propagated from the calibration process, as given by Tenenbein (1972) and Grassia and Sund-

berg (1982). However, this assumes these estimators are reliable for small sample sizes. Future studies are needed to test this assumption. Also, additional work is required to recommend sample sizes for other types of reference data used in remote sensing, such as heterogeneous clusters of pixels.

If areal estimates are an important product of a remote sensing project, then the expense of 500–1000 unstratified, independent reference data plots will often be justified. However, this is more reference data than typical for most remote sensing studies. Efficiency of statistical areal calibration can be improved with a stratified sample of reference plots, and certain issues are discussed regarding the choice of the appropriate statistical estimators for a given stratification scheme, but this subject is beyond the scope of the present study. Efficiency might also be improved with larger, heterogeneous reference plots to estimate the misclassification error matrix. In this case, the inverse and classical estimator evaluated in this paper can be used to calibrate areal estimates; however, the estimators for the error covariance matrix given by Tenenbein (1972) and Grassia and Sundberg (1982) do not apply. However, this too is beyond the scope of the present study.

---

*We thank two anonymous reviewers, and especially Dr. Greg S. Biging from the Remote Sensing Laboratory at University of California Berkeley, for their critical comments and helpful suggestions. We also thank the late Dr. James S. Williams, Statistical Laboratory, Colorado State University, for the suggestion to formulate the calibration model with the underlining joint probability transition matrix.*

## APPENDIX

The classical and inverse calibration estimators for misclassification bias are derived in this section. Part of the following uses notation that was introduced into the remote sensing literature by Bauer et al. (1978) and Hay (1988), but there is no universally accepted notational convention.

First, consider the deterministic situation in which a population of  $N$  sites make up a geographic area. Each site might be a pixel, a homogeneous forest stand or agricultural field, or forest inventory plot. Each site can be inexpensively but imperfectly classified into one of  $k$  mutually exclusive and exhaustive categories using remote

sensing, and independently classified without error into its true category using expensive field measurements or high-resolution aerial photography.

Let the  $(k \times 1)$  indicator vector  $\mathbf{t}_l$  represent the true classification of site  $l$ , where  $i$ th element of  $\mathbf{t}_l$ , that is,  $(t_l)_i$ , has the value of 1 if the true classification of site  $l$  is category  $i$ ; otherwise,  $(t_l)_i$  is 0. Likewise, let the  $(k \times 1)$  indicator vector  $\mathbf{r}_l$  represent the remote classification of site  $l$ . Let  $\mathbf{t}$  be a  $(k \times 1)$  vector of proportions of the population in each category, and  $\mathbf{r}$  be the vector of population proportions from remote interpretations; therefore, these deterministic relationships produce by definition

$$\mathbf{t} = \left( \sum_{l=1}^N \mathbf{t}_l \right) / N, \quad (\text{A1})$$

$$\mathbf{r} = \left( \sum_{l=1}^N \mathbf{r}_l \right) / N. \quad (\text{A2})$$

Let  $\mathbf{P}$  be a  $(k \times k)$  matrix of true joint misclassification probabilities, where the  $ij$ th element  $(P)_{ij}$  equals the number of sites in the population that are truly type  $i$  and remotely classified as type  $j$  divided by the total number of sites in the population ( $N$ ):

$$\mathbf{P} = \sum_{l=1}^N \mathbf{t}_l \mathbf{r}_l' / N. \quad (\text{A3})$$

Since  $\mathbf{r}_l' \mathbf{1} = 1$  by definition, Eq. (A1) can be rewritten as

$$\mathbf{t} = \sum_{l=1}^N \mathbf{t}_l (\mathbf{r}_l' \mathbf{1}) / N.$$

From Eq. (A3), this is equivalent to the deterministic relationship:

$$\mathbf{t} = \mathbf{P} \mathbf{1}, \quad (\text{A4})$$

that is,  $\mathbf{t}$  is a vector of marginal probabilities of  $\mathbf{P}$ .

For reasons that will become clear when we move to the stochastic case, it is necessary to factor the vector of remotely sensed proportions ( $\mathbf{r}$ ) from Eq. (A4). If  $\mathbf{P}$  is solely estimated from a small sample of reference plots, an estimator based on Eq. (A4) would not use the bulk of the remotely sensed information in  $\mathbf{r}$ .

$\mathbf{1}$  can be rewritten as the  $(k \times 1)$  vector

$$\mathbf{1} = \begin{bmatrix} (r)_1 & \cdots & (r)_k \\ (r)_1 & & (r)_k \end{bmatrix}', \quad (\text{A5})$$

where  $(r)_j$  is the  $j$ th element of  $\mathbf{r}$  (i.e., the proportion of sites remotely interpreted as type  $j$ ). Equation (A5) is equivalent to

$$\mathbf{1} = \mathbf{R}^{-1} \mathbf{r}, \quad (\text{A6})$$

where  $\mathbf{R}$  is a  $(k \times k)$  diagonal matrix with vector  $\mathbf{r}$  on the diagonal, that is,  $\mathbf{R} = \text{diag}(\mathbf{r})$ , and  $\mathbf{R}^{-1}$  is diagonal with its  $j$ th diagonal element equal to  $1 / (r)_j$ . Combining Eqs. (A4) and (A6),

$$\mathbf{t} = (\mathbf{P} \mathbf{R}^{-1}) \mathbf{r}. \quad (\text{A7})$$

Using the Bayes theorem,  $(\mathbf{P} \mathbf{R}^{-1})$  can be considered a matrix of conditional probabilities where the  $ij$ th element is the probability that a site is truly type  $i$  given its remote classification is type  $j$ , and each column vector of  $(\mathbf{P} \mathbf{R}^{-1})$  sums to 1. Heldal and Spjøtvoll (1988) term this a transition matrix in a measurement error model. In Eq. (A3), the true proportions ( $\mathbf{t}$ ) of categories in the entire population are a multivariate linear function of proportions of the remotely interpreted categories in the entire population ( $\mathbf{r}$ ) and the conditional probabilities of true classifications given the remotely sensed classifications.

Using  $\mathbf{r} = \mathbf{P}' \mathbf{1}$  similar to Eq. (A4), it can be shown that the proportions of remotely interpreted categories can be expressed as functions of the true proportions in the population and different conditional classification probabilities:

$$\mathbf{r} = (\mathbf{P}' \mathbf{T}^{-1}) \mathbf{t}, \quad (\text{A8})$$

where  $\mathbf{T} = \text{diag}(\mathbf{t})$  and  $(T^{-1})_{ii} = 1 / (t)_i$ .  $(\mathbf{P}' \mathbf{T}^{-1})$  is a matrix of conditional probabilities (Grassia and Sundberg, 1982), where the  $ij$ th element is the probability that the remote classification is type  $i$  given its true classification is type  $j$ , and the column vectors in  $\mathbf{P}' \mathbf{T}^{-1}$  each sum to 1.

If  $(\mathbf{P}' \mathbf{T}^{-1})$  is nonsingular, then Eq. (A8) can be inverted so that the unknown true proportions ( $\mathbf{t}$ ) in the population are a function of the known remotely sensed proportions ( $\mathbf{r}$ ):

$$\mathbf{t} = (\mathbf{P}' \mathbf{T}^{-1})^{-1} \mathbf{r}, \quad (\text{A9})$$

which is structurally different than the deterministic equality in Eq. (A7). The conditional probability matrix  $(\mathbf{P}' \mathbf{T}^{-1})$  is commonly labeled an error matrix in the remote sensing literature, and denoted as  $\mathbf{E}$  by Bauer et al. (1978) and Hay (1988).

In the stochastic case, known estimates  $\hat{\mathbf{P}}$  of the true but unknown joint probabilities  $\mathbf{P}$  are

available using a simple random or systematic sample of  $m$  reference sites:

$$\hat{\mathbf{P}} = \sum_{l=1}^m \mathbf{t}_l \mathbf{r}_l / m,$$

where the randomized sample is considered as the first  $m$  sites for notational convenience. As noted by Maxim and Harrington (1983),  $\hat{\mathbf{P}}$  is the maximum likelihood estimate under certain assumptions (Kendall and Stuart, 1967) and, more generally, the minimum variance estimate (Bishop et al. 1975). Equation (A7) suggests the following inverse estimator of true proportions in the population, where estimates of classification probabilities ( $\hat{\mathbf{P}}$ ) are used in place of the unknown true probabilities ( $\mathbf{P}$ ):

$$\hat{\mathbf{t}}_i = (\hat{\mathbf{P}}\hat{\mathbf{R}}^{-1})\mathbf{r} \quad (\text{A10})$$

with  $\hat{\mathbf{R}} = \text{diag}(\hat{\mathbf{P}}\mathbf{1})$ , that is, the marginal probabilities of each remotely sensed classification estimated solely from the reference data. This estimator was proposed for industrial sampling inspection by Tenenbein (1972), and for remote sensing by Card (1982) and Chrisman (1982). Tenenbein shows that this is an unbiased maximum likelihood estimator; the sum of proportions in  $\hat{\mathbf{t}}_i$  is 1, and each element of  $\hat{\mathbf{t}}_i$  is positive because all elements in  $\hat{\mathbf{P}}\hat{\mathbf{R}}^{-1}$  and  $\mathbf{r}$  are positive. If there are no sampled reference sites that are truly category  $i$  and remotely interpreted as category  $j$ , by convention  $\hat{P}_{ij} = 0$ , even if  $P_{ij} \neq 0$ . In this case,  $\hat{\mathbf{t}}_i$  is unbiased unless  $m$  is small (Tenenbein, 1972). It is possible this estimator is infeasible if estimate  $\hat{\mathbf{R}}$  is singular, even if the true  $\mathbf{R}$  is nonsingular. This infeasibility can occur when  $\hat{\mathbf{R}}$  has a zero element on its diagonal because a remotely sensed category did not occur in a particular sample of reference sites.

Similarly, Eq. (A9) suggests the alternative classical estimator using the estimate  $\hat{\mathbf{P}}$  rather than its true, but unknown, value ( $\mathbf{P}$ ):

$$\hat{\mathbf{t}}_c = (\hat{\mathbf{P}}'\hat{\mathbf{T}}^{-1})^{-1}\mathbf{r}, \quad (\text{A11})$$

where  $\hat{\mathbf{T}} = \text{diag}(\hat{\mathbf{P}}\mathbf{1})$ , that is, the marginal probabilities of true classifications estimated solely from the reference data. This is the estimator proposed by Grassia and Sundberg (1982) for mechanical sorting machines, and by Bauer et al. (1978), Maxim et al. (1981), Prisley and Smith (1987), and Hay (1988) for remote sensing. It has also been employed by Barron (1977) and Greenland (1988) for epidemiology. This estimator is approxi-

mately or asymptotically unbiased (Grassia and Sundberg, 1982). It is also a method of moments estimator (Maxim et al., 1981). Column vectors of estimated conditional probabilities in  $\hat{\mathbf{P}}'\hat{\mathbf{T}}^{-1}$  are statistically independent (Grassia and Sundberg, 1982). Furthermore, the vector of estimated proportions sums to 1. However, it is possible that some proportions in  $\hat{\mathbf{t}}_c$  will be negative because  $(\hat{\mathbf{P}}'\hat{\mathbf{T}}^{-1})^{-1}$  can contain negative elements (Maxim et al., 1981). It is also possible that  $(\hat{\mathbf{P}}'\hat{\mathbf{T}}^{-1})^{-1}$  does not exist, (i.e.,  $\hat{\mathbf{P}}'\hat{\mathbf{T}}^{-1}$  or  $\hat{\mathbf{T}}^{-1}$  is singular) even if  $(\mathbf{P}'\mathbf{T}^{-1})^{-1}$  does exist. For example, a diagonal element of  $\hat{\mathbf{P}}$  could be zero, which causes  $\hat{\mathbf{P}}'\hat{\mathbf{T}}^{-1}$  to be singular, because none of the sampled reference sites happened to be correctly classified. This problem will occur most frequently when there are few reference sites for a particular category, as might be expected for a rare category in a simple random sample of reference sites.

## REFERENCES

- Barron, B. A. (1977), The effects of misclassification on the estimation of relative risk, *Biometrics* 33:414-418.
- Bauer, M. E., Hixson, M. M., Davis, B. J., and Etheridge, J. B. (1978), Area estimation of crops by digital analysis of Landsat data, *Photogramm. Eng. Remote Sens.* 44:1033-1043.
- Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Massachusetts Institute of Technology Press, Cambridge.
- Brown, P. J. (1982), Multivariate calibration, *J. Roy. Stat. Soc. B* 44:287-321.
- Burk, T. E., Bauer, M. E., Ek, A. R., and Ahearn, S. C. (1988), Satellite inventory of Minnesota's forest resource, in *Proceedings of the IUFRO S4.02.05 Meeting*, Hyytiälä, Finland, 29 August-2 September, pp. 43-52.
- Card, D. H. (1982), Using known map categorical marginal frequencies to improve estimates of thematic map accuracy, *Photogramm. Eng. Remote Sens.* 48:431-439.
- Catts, G. P., Cost, N. D., Czaplewski, R. L., and Snook, P. W. (1987), Preliminary results from a method to update timber resource statistics in North Carolina, in *Proceedings of the 11th Biennial Workshop on Color Aerial Photography in the Plant Sciences*, Weslaco, TX, pp. 39-52.
- Chrisman, N. R. (1982), Beyond accuracy assessment: correction of misclassification, in *Proceedings 5th International Symposium on Computer-Assisted Cartography*, Crystal City, VA, 22-28 August, pp. 123-132.
- Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educ. Psychol. Meas.* 20:37-46.
- Congalton, R. G., Oderwald, R. G., and Mead, R. A. (1983),

- Assessing Landsat classification accuracy using discrete multivariate statistical techniques, *Photogramm. Eng. Remote Sens.* 49:1671–1678.
- Czaplewski, R. L. (1991), Misclassification bias in areal estimates, *Photogramm. Eng. Remote Sens.*, forthcoming.
- Czaplewski, R. L., and Catts, G. P. (1990), Calibrating area estimates for classification error using confusion matrices, in *Proceedings ACSM/ASPRS Convention*, Denver, CO, Vol. 4, pp. 431–440.
- Czaplewski, R. L., Catts, G. P., and Snok, P. W. (1987), National land cover monitoring using large, permanent photoplots, in *Proceedings Land and Resource Evaluation for National Planning in the Tropics*, Chetumal, Mexico, pp. 187–202.
- Emerson, J. D., and Strenio, J. (1983), Boxplots and batch comparisons, in *Understanding Robust and Exploratory Data Analysis* (D. C. Hoaglin, F. Mosteller, and J. W. Tukey, Eds.), Wiley, New York, pp. 58–93.
- Fienberg, S. E., and Holland, P. W. (1973), Simultaneous estimation of multinomial cell probabilities, *J. Am. Stat. Assoc.* 68:683–691.
- Grassia, A., and Sundberg, R. (1982), Statistical precision in the calibration and use of sorting machines and other classifiers, *Technometrics* 24:117–121.
- Greenland, S. (1988), Variance estimation for epidemiological effect estimates under misclassification, *Stat. Med.* 7: 745–757.
- Hay, A. M. (1988), The derivation of global estimates from a confusion matrix, *Int. J. Remote Sens.* 9:1395–1398.
- Heldal, J., and Spjotvoll, E. (1988), Combination of surveys and registers: a calibration approach with categorical variables, *Int. Stat. Rev.* 56:153–164.
- Hochberg, Y. (1977), On the use of double sampling schemes in analyzing categorical data with misclassification error, *J. Am. Stat. Assoc.* 72:914–921.
- Kendall, M. G., and Stuart, A. (1967), *The Advanced Theory of Statistics*, Hafner, New York.
- Li, H. G., Schreuder, H. T., Van Hooser, D. D., and Brink, G. E. (1991), Estimating strata means in double sampling with corrections based on second-phase sampling, *Biometrics*, forthcoming.
- Light, R. J. (1971), Measure of response agreement for qualitative data: some generalizations and alternatives, *Psychol. Bull.* 76:365–377.
- Mak, T. K. (1988), Estimating subgroups means with misclassification, *J. Roy. Stat. Soc. B* 50:83–92.
- Mak, T., and Li, W. K. (1988), A new method for estimating subgroup means under misclassification, *Biometrika* 75: 105–111.
- Maxim, L. D., and Harrington, L. (1983), The application of pseudo-Bayesian estimators to remote sensing data: ideas and examples, *Photogramm. Eng. Remote Sens.* 49:649–658.
- Maxim, L. D., Harrington, L., and Kennedy, M. (1981), Alternative “scale-up” estimates for aerial surveys where both detection and classification errors exist, *Photogramm. Eng. Remote Sens.* 47:1227–1239.
- Norton-Griffiths, M. (1988), Aerial point sampling for land use surveys, *J. Biogeogr.* 15:149–156.
- Poso, S. (1988), Seeking for an optimal path for using satellite imageries for forest inventory and monitoring, in *Proceedings of IUFRO S4.02.05 Meeting*, Hyytiälä, Finland, 29 August–2 September.
- Prisley, S. P., and Smith, J. L. (1987), Using classification error matrices to improve the accuracy of weighted land-cover models, *Photogramm. Eng. Remote Sens.* 53:1259–1263.
- Rosenfield, G. H., and Fitzpatrick-Lins, K. (1986), A coefficient of agreement as a measure of thematic classification accuracy, *Photogramm. Eng. Remote Sens.* 52:223–227.
- Selén, J. (1986), Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data, *J. Am. Stat. Assoc.* 81:75–81.
- Sheffield, R. M., and Knight, H. A. (1986), *North Carolina's Forests*, USDA Forest Service Resource Bulletin SE-88, 97 pp.
- Tenenbein, A. (1972), A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection, *Technometrics* 14:187–202.
- Thomas, R. W. (1986), Utility of remote sensing data in renewable resource sample survey, in *Proceedings of the 1986 International Geoscience and Remote Sensing Symposium*, Zürich, 8–11 September, pp. 755–758.