Chapter 7

Accuracy Assessments and Areal Estimates Using Two-Phase Stratified Random Sampling, Cluster Plots, and the Multivariate Composite Estimator

Raymond L. Czaplewski

INTRODUCTION

Consider the following example of an accuracy assessment. Landsat data are used to build a thematic map of land cover for a multicounty region. The map classifier (e.g., a supervised classification algorithm) assigns each pixel into one category of land cover. The classification system includes 12 different types of forest and land cover: black spruce, balsam fir, white-cedar, other softwoods, aspen, birch, other hardwoods, urban, wetland, water, pasture, and agriculture. The accuracy of the map must be known by a user of the map to conduct credible analyses.

In concept, each and every pixel can be classified with two separate classifiers: (1) the map classifier, and (2) a field crew. I consider the classification by the field crew to be "error-free" because it is treated as the "true" classification by the user of the map. An accuracy assessment compares the agreement of these two classifiers for the entire thematic map. First, I will discuss the "true" error-matrix, in which each and every pixel in the map is classified by classifiers (1) and (2). Then, I will discuss an estimate of the true error-matrix based on a sample of pixels.

The True Error-Matrix

The accuracy assessment in this example uses a 12 by 12 contingency table, called an "error-matrix" in remote sensing terminology (Congalton, 1991). By convention, each column of the error-matrix represents one if the 12 categories on the map, and each row of the error-matrix represents one of the 12 categories determined by the field crew. Each of the 144 cells of the error-matrix represents the proportion of all pixels in the map that are cross-classified into the corresponding column (i.e., map categories) and row (i.e., true categories). The diagonal of the error-matrix represents the proportion of all pixels for which there is agreement between the map and error-free classifiers, i.e., correct classifications. Scalar statistics, such as kappa and the total proportion of correctly classified pixels, con-

cisely describe the degree of overall agreement between the two classifiers as represented by the 144 cells in this error-matrix.

The total proportion of pixels classified as a certain category in the map equals the sum of the 12 cells within the corresponding column of this error-matrix. The column margin of the error-matrix represents the total proportion of pixels in each category within the thematic map, which corresponds to area statistics that would be produced from this map by a geographic information system. The row margin represents the total proportion of pixels in each category as determined by the error-free classifier (i.e., classifications made by field crew). The difference between the proportions in the row and column margins is called misclassification bias (Czaplewski, 1992a). Thematic maps often overestimate the areal extent of rare categories.

The Estimated Error-Matrix

Although it is practical for the map classifier to assign each and every pixel in the entire map into one of the 12 thematic categories, it is generally impractical for a field crew to classify every pixel. Therefore, the accuracy assessment requires a sample of pixels that are cross-classified by both the map classifier and the error-free classifier. A proper probability sample is sufficient to estimate the error-matrix and make inferences about the true error-matrix. As the sample size increases, the estimated error-matrix will more closely match the true error-matrix.

Simple random sampling of individual pixels is the simplest statistical design for estimation of the error-matrix, and most commercial statistical software can produce valid estimates with this design. However, simple random sampling is usually among the most expensive approaches. More complex sampling designs can reduce costs.

Travel costs for the field crew can be reduced by selecting clusters of pixels that are near each other. Assuming spatial autocorrelation of classification errors, pixels within the same cluster can no longer be considered independent. Most commercial software systems are not designed to correctly estimate an error-matrix with cluster plots.

The cost of reference data can be further reduced by photo-interpretation of reference sites, and classifying a small subset of those sites by field crews. However, photo-interpretation is subject to classification errors, which confounds the assessments of map accuracy (Congalton and Green, 1993). For many years, forest inventories have used double-sampling methods to improve efficiencies of field surveys with photo-interpretation. However, these univariate methods are insufficient to estimate many of the accuracy assessment statistics that are derived from transformations of cells in a multivariate error-matrix.

The following section derives a multivariate estimator for the error-matrix that can accommodate double sampling with cluster plots. Poststratification treats all those pixels classified as a single mapped category as a stratum. The independence among strata improves efficiency. If statistical estimators designed for simple random sampling are applied to reference data from this complex sampling design, then the assessment will be biased and can easily lead to false conclusions (Stehman and Czaplewski, 1998). The objective is to present rigorous derivation of statistical methods for implementation by statisticians who serve users of spatial data. Also, we have implemented these same methods in user-friendly software that we call ACAS (ACcuracy ASsessment software).

METHODS

This section will describe the estimation problem using formal statistical terminology and notation. Next, subpopulations and parameter matrices are defined. Then, I present multivariate sample-survey estimators that use homogeneous sample units, such as independent pixels. These estimators are generalized to deal with clusters of sample units. Next, I introduce the multivariate composite estimator, which I use to combine two independent vector estimates: (1) the estimate from the Phase-1 sample, which includes the imperfect (photo-interpreted) reference classifications, but does not include the error-free classifications (field crew); and (2) the estimate from Phase-2, which includes both field classifications and photo-interpreted classifications. The combination of these two independent estimates produces an estimated error-matrix that uses field classifications as the definition of truth, while using photo-interpreted data to improve efficiency. Then, I present methods to estimate the margin of this error-matrix, which provides unbiased and consistent estimates of areal extent for each thematic category. Finally, I present multivariate methods that transform this error-matrix into accuracy assessment statistics, such as kappa statistics and total proportion of correct classifications. I also derive variance estimators for these statistics.

Description of the Statistical Problem

Consider a thematic map that is comprised of many map units (e.g., pixels or polygons), each of which is imperfectly classified into one of k mutually exclusive categories of land cover by an inexpensive, but fallible, classifier (e.g., remotely sensed data). An "error-free" reference classifier determines the true category for a probability sample of map units or control points (Arbia, 1993). This sample serves as the basis of statistical inference to assess the entire map.

A $k \times k$ contingency table is the basis for most accuracy assessments. The rows represent "error-free" reference classifications, and the columns will represent the imperfect classifications on the map. The ijth element of the contingency table is the estimated joint probability that any map unit is labeled as category j on the map and is truly category I.

An assessment of classification accuracy utilizes various scalar statistics computed from this contingency table, as reviewed by Bishop et al. (1975), Fleiss (1981), and Congalton (1991). The total probability of correct classification is the sum of the diagonal elements of the contingency table. Conditional probabilities of correct classification (e.g., accuracy given that the mapped or error-free classification is a certain category) are the joint probabilities on the diagonal divided by their corresponding row or column marginal probability (Agresti, 1990, p. 9; Green et al., 1993). Weighted, unweighted, and conditional kappa statistics are additional assessment statistics that help (Czaplewski, 1994). One margin of this contingency table provides unbiased estimates of areal extent (Van Deusen, 1994), expressed as proportions of the population. The other margin corresponds to the census, or complete enumeration, of map units, all of which are categorized with the map classifier.

Classifications in the field are typically considered "error-free" reference data. However, field observations are expensive. Photo-interpretation is a less expensive source of reference data, but photo-interpretation errors confound the assessment (Congalton and Green, 1993). The combination of photo-interpretations with field classifications can im-

prove precision of the assessment statistics and areal estimates using multivariate two-phase sampling.

A thematic map represents a census of all map units, each of which is classified with the map classifier. The $k \times 1$ column margin of the $k \times k$ contingency table can be fixed as a vector of known constants from the census. This reduces the uncertainty in estimates of the k^2 elements within the contingency table and the k elements of the row margin, which correspond to the estimated true proportion of each category.

Czaplewski (1992b) suggested the multivariate composite estimator to produce the desired $k \times k$ contingency table. This uses methods from stochastic processes (Maybeck, 1979, p. 26). The composite estimator combines vector sample estimates from both the Phase-1 photointerpreted sample and the Phase-2 field samples, and fixes one margin of the contingency table through the census of map units and their classifications on the map. Williamson and Haber (1994) review many closely related problems with cross-classified data, but none of the multiple-sample approaches consider margins fixed through a census. The effect of such complex designs can be substantial on tests of hypothesis and estimated confidence intervals (Holt et al., 1980; Rao and Thomas, 1989).

An imperfect but inexpensive map classifier (e.g., digital classification of remotely sensed imagery) assigns a map unit (e.g., a 0.5-ha plot or a 0.1-ha map pixel) to one and only one of k mutually exclusive categories, where k typically ranges between 2 and 30 categories. The population consists of all N map units on the thematic map, where N is known exactly. The true category for any map unit could be determined with an expensive, error-free reference classifier (e.g., field data) that uses the same k categories as the map classifier. An assessment of the accuracy of the map classifier requires the $k \times k$ contingency table $\mathbf{Z}$. The ijth element of $\mathbf{Z}$ is the joint probability that any map unit in the population is classified as category j on the map and is truly category I.

An imperfect classifier (e.g., manual interpretation of aerial photographs) is also available. This imperfect reference classifier will not always agree with the error-free reference classifier. This imperfect reference classifier uses k_y different categories, which can differ from the k categories used by the map and error-free reference classifiers.

The inexpensive map classifier is applied to all N map units in the population or thematic map (e.g., N equals 10^6 to 10^8 units). The expensive, but imperfect, reference classifier can be applied to a small sample of map units (e.g., n_y equals 10^3 to 10^4 units). The more expensive, error-free reference classifier is applied to an even smaller sample of map units (e.g., n_x equals 10^2 to 10^3 units).

The reference data consist of two independent probability samples of map units within the same subpopulation (defined in the next section). The first sample uses map units that are categorized with both the imperfect reference classifier and the map classifier, which is analogous to a first-phase sample. The second sample, which uses the same type of map units, is categorized with all three classifiers: the map, the imperfect reference, and the error-free reference classifiers. This is analogous to a second-phase sample. Each primary sample unit can be an individual map unit, or a cluster of map units. Assume there is negligible locational error (registration error) in location of a map unit in the field (Arbia, 1993, p. 343). The multivariate composite estimator, which is presented in a following section, combines these two samples into a single, more efficient multivariate estimate of the contingency table.

Definitions: Subpopulations and Parameter Matrices

Since the map classification is known for each map unit, the entire population of N map units can be segregated into k subpopulations. The map classification, denoted as category m, is the same for all map units in subpopulation $M = m$. The number of map units in each subpopulation ($N_{M=m}$) is known exactly. Later, independent estimates of k conditional classification probabilities will be made separately for each of the k subpopulations, then transformed and merged into the required $k \times k$ contingency table.

Let $\mathbf{z}_{M=m}$ be a $k \times 1$ parameter vector for the subpopulation of map units that are assigned to category m by the map classifier. The ith element of $\mathbf{z}_{M=m}$ equals the proportion of map units in this subpopulation that the error-free reference classifier would assign to category I. The sum of all elements in $\mathbf{z}_{M=m}$ is exactly 1, and all elements occur in the interval between 0 and 1. The mth column of the $k \times k$ contingency table $\mathbf{Z}$ equals $\mathbf{z}_{M=m}$ times the prevalence of the mth subpopulation in the total population:

$$\mathbf{Z} = \left[\mathbf{z}_{M=1}\left(\frac{N_{M=1}}{N}\right) \cdots \mathbf{z}_{M=m}\left(\frac{N_{M=m}}{N}\right) \cdots \mathbf{z}_{M=k}\left(\frac{N_{M=k}}{N}\right) \right] \qquad (1)$$

Next, let $\mathbf{y}_{M=m}$ be *a* $k_y \times 1$ parameter vector for the subpopulation that the map classifier assigned to category m. The ith element of $\mathbf{y}_{M=m}$ equals the proportion of all $N_{M=m}$ map units in subpopulation $M = m$ that are assigned to category I, $I \in \{1, ,k_y\}$, by the imperfect reference classifier. The sum of all elements of $\mathbf{y}_{M=m}$ equals exactly 1. Furthermore, let the $k_y \times 1$ measurement vector $\mathbf{y}_p$ represent the classification outcome for an individual map unit, denoted by subscript p; if the imperfect reference classifier assigns the pth map unit to category I, then the ith element of $\mathbf{y}_p$ equals 1 and all other elements equal 0.

Finally, let $\mathbf{X}_{M=m}$ be a $k \times k_y$ parameter matrix for subpopulation $M = m$. The ijth element of $\mathbf{X}_{M=m}$ is the proportion of map units in this subpopulation that the error-free reference classifier would assign to category I and the imperfect reference classifier would assign to category j. Using the notation of Molina (1989) and Christensen (1991, p. 3), define the *Vec* as the vectorization operator that stacks columns of a matrix. Let $\mathbf{x}_{M=m} = Vec(\mathbf{X}_{M=m})$, where $\mathbf{x}_{M=m}$ is the equivalent $(kk_y) \times 1$ vector that contains the same parameters as $\mathbf{X}_{M=m}$. Rearrangement of the $k \times k_y$ matrix ($\mathbf{X}_{M=m}$) into a $(kk_y) \times 1$ stacked vector ($\mathbf{x}_{M=m}$) facilitates computation of the $(kk_y) \times (kk_y)$ covariance matrix $\hat{\mathbf{V}}\mathbf{x}_{M=m}$ for the estimate of $\mathbf{x}_{M=m}$. In addition, let the $(kk_y) \times 1$ measurement vector $\mathbf{x}_p$ represent the joint classification of a single map unit p. If the error-free reference classifier assigns the pth map unit to category I and the imperfect reference classifier assigns it to category j, then the $[(I\text{-}1)k+j]th$ element of $\mathbf{x}_p$ equals 1, and all other elements equal 0.

Subpopulation vectors $\mathbf{z}_{M=m}$ and $\mathbf{y}_{M=m}$ are linear transformations of $\mathbf{x}_{M=m}$:

$$\mathbf{y}_{M=m} = \mathbf{H}_y \, \mathbf{x}_{M=m} \qquad (2)$$

$$\mathbf{z}_{M=m} = \mathbf{H}_z \, \mathbf{x}_{M=m} \qquad (3)$$

$\mathbf{H}_y$ and $\mathbf{H}_z$ are appropriately structured matrices of zeros and ones, examples of which follow. $\mathbf{H}_y$ has dimensions $k_y \times (kk_y)$, and $\mathbf{H}_z$ has dimensions $k \times (kk_y)$. These transformations will be exploited by the multivariate composite estimator in a following section.

The following is a simple example. The map and error-free reference classifiers assign map units into $k = 3$ categories: forest (F), nonforest (N), and water (W). The imperfect reference classifier uses k_y = two categories: vegetated (V) and barren (B). Let $(X_{M=m})_{ij}$ represent the ijth element of matrix $\mathbf{X}_{M=m}$; for example, $(X_{M=N})_{FV}$ is the proportion of map units in the subpopulation assigned to the nonforest (N) category by the map classifier that would be classified as forest (F) by the error-free reference classifier and vegetated (V) by the imperfect reference classifier. In this example,

$$\mathbf{X}_{M=N} = \begin{bmatrix} (X_{M=m})_{FV} & (X_{M=m})_{FB} \\ (X_{M=m})_{NV} & (X_{M=m})_{NB} \\ (X_{M=m})_{WV} & (X_{M=m})_{WB} \end{bmatrix} \qquad \mathbf{x}_{M=N} = Vec(\mathbf{X}_{M=N}) = \begin{bmatrix} (X_{M=m})_{FV} \\ (X_{M=m})_{NV} \\ (X_{M=m})_{WV} \\ (X_{M=m})_{FB} \\ (X_{M=m})_{NB} \\ (X_{M=m})_{WB} \end{bmatrix} \tag{4}$$

$$\mathbf{y}_{M=N} = \begin{bmatrix} (X_{M=m})_{FV} + (X_{M=m})_{NV} + (X_{M=m})_{WV} \\ (X_{M=m})_{FB} + (X_{M=m})_{NB} + (X_{M=m})_{WB} \end{bmatrix}$$

$$= \mathbf{H}_y \, \mathbf{x}_{M=N} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} (X_{M=m})_{FV} \\ (X_{M=m})_{NV} \\ (X_{M=m})_{WV} \\ (X_{M=m})_{FB} \\ (X_{M=m})_{NB} \\ (X_{M=m})_{WB} \end{bmatrix} \tag{5}$$

$$\mathbf{Z}_{M=N} = \begin{bmatrix} (X_{M=m})_{FV} + (X_{M=m})_{FB} \\ (X_{M=m})_{NV} + (X_{M=m})_{NB} \\ (X_{M=m})_{WV} + (X_{M=m})_{WB} \end{bmatrix}$$

$$= \mathbf{H}_z \, \mathbf{x}_{M=N} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} (X_{M=m})_{FV} \\ (X_{M=m})_{NV} \\ (X_{M=m})_{WV} \\ (X_{M=m})_{FB} \\ (X_{M=m})_{NB} \\ (X_{M=m})_{WB} \end{bmatrix} \tag{6}$$

Assume that the *p*th map unit is classified as nonforest (N) by the map classifier, barren by the imperfect reference classifier (B), and water by the error-free classifier (W). Then this map unit is a member of subpopulation $M = \mathrm{N}$, and its measurement vectors are:

$$\mathbf{y}_p = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \mathbf{x}_p = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{7}$$

Estimates from Homogeneous Sample Units

First, consider a simple random sample that has a fixed sample size of $n_{y|M=m}$ map units from subpopulation $M = m$. Let $S_{y|M=m}$ represent the set of $n_{y|M=m}$ subscripts (p) for the map units in this sample. Each map unit in this sample was classified with the map classifier to determine its subpopulation, and with the imperfect reference classifier (e.g., photo-interpretation). From the previous section, vector $\mathbf{y}_{M=m}$ contains the proportions of map units in this subpopulation that would be classified into each of k_y categories by the imperfect reference classifier, and measurement vector $\mathbf{y}_p$ contains the results of the imperfect reference classifier for the *p*th map unit (e.g., Equation 7). The vector sample mean provides an efficient and asymptotically unbiased estimate of $\mathbf{y}_{M=m}$ in Equation 2:

$$\hat{\mathbf{y}}_{M=m} = \frac{1}{\mathbf{n}_{y|M=m}} \left(\sum_{p \in S_{y|M=m}} \mathbf{y}_p \right) = \mathbf{y}_{M=m} + \mathbf{e}_{y|M=m} \tag{8}$$

where $\mathbf{e}_{y|M=m}$ is the $k_y \times 1$ vector of random sampling errors, and $\mathrm{E}[\mathbf{e}_{y|M=m}] = \mathbf{0}$ for a large $n_{y|M=m}$. Equation 8 can be biased for small $n_{y|M=m}$ because zero elements are treated as structural zeros even though they might be sampling zeros (Bishop et al., 1975, Chapters 5 and 12); a structural zero has a probability exactly equal to zero, while a sampling zero element represents a rare event that was not observed in the sample (Bishop et al., 1975, p. 177). Assuming sampling with replacement, the multinomial distribution provides an estimated covariance matrix for these sampling errors ($\mathrm{E}[\mathbf{e}_{y|M=m}\mathbf{e}'_{y|M=m}] = \mathbf{V}_{y|M=m}$):

$$\hat{\mathbf{V}}_{y|M=m} = \frac{Diag(\hat{\mathbf{y}}_{M=m}) - \hat{\mathbf{y}}_{M=m}\hat{\mathbf{y}}'_{M=m}}{\mathbf{n}_{y|M=m}} \tag{9}$$

where $Diag(\hat{\mathbf{y}}_{M=m})$ is the $k_y \times k_y$ matrix with vector $\hat{\mathbf{y}}_{M=m}$ on the diagonal and all other elements equal zero (Agresti 1990, p. 423). Since the sampling fraction is usually small for

thematic maps ($n_{y\,|\,M=m}/N_{M=m} < 0.01$), the multinomial distribution will often be a reasonable approximation for sampling without replacement.

A second, independent sample of $n_{x\,|\,M=m}$ map units is taken from the same subpopulation ($M = m$). $S_{x\,|\,M=m}$ is the set of subscripts (p) for the map units in this second sample. In addition to being classified by the map classifier and the imperfect reference classifier, each map unit in this sample is also classified with the error-free classifier, and the results are represented by measurement vector $\mathbf{x}_p$ (e.g., Equation 7). This provides a sample estimate of $\mathbf{x}_{M=m}$ and its multinomial covariance matrix for sampling errors:

$$\hat{\mathbf{x}}_{M=m} = \sum_{p \in S_{x|M=m}} \frac{\mathbf{x}_p}{\mathbf{n}_{x|M=m}} = \mathbf{x}_{M=m} + \mathbf{e}_{x|M=m} \tag{10}$$

$$\hat{\mathbf{V}}_{x|M=m} = \frac{Diag(\hat{\mathbf{x}}_{M=m}) - \hat{\mathbf{x}}_{M=m}\hat{\mathbf{x}}'_{M=m}}{\mathbf{n}_{y|M=m}} \tag{11}$$

It is assumed with the multivariate composite estimator that the sampling errors $\mathbf{e}_{y\,|\,M=m}$ and $\mathbf{e}_{x\,|\,M=m}$ are independent.

Estimates from Clusters of Sample Units

Now consider a cluster of map units as the primary sample unit. Let subscript c denote the cth cluster, and $S_{c\,|\,M=m}$ represent the set of $N_{c\,|\,M=m}$ subscripts for those map units (p) that make up cluster plot c and are members of subpopulation $M = m$. A map unit can be a member of only one cluster, but more than one subpopulation may occur in a single cluster. The measurement vectors for the cth cluster are:

$$\mathbf{y}_c = \sum_{p \in S_{c|M=m}} \frac{\mathbf{y}_p}{N_{c|M=m}} \tag{12}$$

$$\mathbf{x}_c = \sum_{p \in S_{c|M=m}} \frac{\mathbf{x}_p}{N_{c|M=m}} \tag{13}$$

The ith element of $\mathbf{y}_c$ equals the proportion of category I in the cth cluster, and the sum of all elements in $\mathbf{y}_c$ equals 1. The $[(I\text{-}1)k+j]th$ element of $\mathbf{x}_c$ equals the proportion of the cth cluster that is assigned to category I by the error-free reference classifier and category j by the imperfect reference classifier.

The sample mean vectors remain efficient and asymptotically unbiased estimates of $\mathbf{y}$ and $\mathbf{x}$:

$$\hat{\mathbf{y}}_{M=m} = \sum_{c \in S_{y|M=m}} \frac{\mathbf{y}_c}{n_{y|M=m}} = \mathbf{y}_{M=m} + \mathbf{e}_{y|M=m} \tag{14}$$

$$\hat{\mathbf{x}}_{M=m} = \sum_{c \in S_{x|M=m}} \frac{\mathbf{x}_c}{n_{x|M=m}} = \mathbf{x}_{M=m} + \mathbf{e}_{x|M=m} \tag{15}$$

where $S_{y|M=m}$ represents the set of subscripts for the sample of $n_{y|M=m}$ clusters in subpopulation $M = m$ that are measured with the imperfect reference classifier alone, and $S_{x|M=m}$ represents the set of subscripts for sample of the $n_{x|M=m}$ clusters in the same subpopulation that are measured with both the imperfect and error-free reference classifiers. However, the multinomial distribution (Equations 9 and 11) should not be used for the sampling error covariance matrices $\mathbf{V}_y$ and $\mathbf{V}_x$ because sampling errors for map units within the same cluster are not likely to be independent. The sample covariance matrix, which is an asymptotically unbiased moment estimator, is an alternative:

$$\hat{\mathbf{V}}_{y|M=m} = \sum_{c \in S_{y|M=m}} \left[\frac{(\mathbf{y}_c - \hat{\mathbf{y}}_{M=m})(\mathbf{y}_c - \hat{\mathbf{y}}_{M=m})'}{n_{y|M=m} - 1} \right] \tag{16}$$

$$\hat{\mathbf{V}}_{x|M=m} = \sum_{c \in S_{x|M=m}} \left[\frac{(\mathbf{x}_c - \hat{\mathbf{x}})(\mathbf{x}_c - \hat{\mathbf{x}})'}{n_{x|M=m} - 1} \right] \tag{17}$$

Multivariate Composite Estimator

The multivariate composite estimator (Maybeck, 1979, p. 217) combines independent vector estimates $\hat{\mathbf{y}}_{M=m}$ (Equations 8 or 14) and $\hat{\mathbf{x}}_{M=m}$ (Equations 10 or 15) into a more efficient $(kk_y)\times 1$ vector estimate $\hat{\mathbf{x}}_{M=m|y,x}$ for subpopulation $M = m$:

$$\hat{\mathbf{x}}_{M=m|y,x} = (\mathbf{K}_{M=m})\hat{\mathbf{y}}_{M=m} + (\mathbf{I} - \mathbf{K}_{M=m}\mathbf{H}_y)\hat{\mathbf{x}}_{M=m} \tag{18}$$

where $\mathbf{I}$ is the $(kk_y)\times(kk_y)$ identity matrix, $\mathbf{H}_y$ is given in Equations 2 and 5, and $\mathbf{K}_{M=m}$ is defined below. The covariance matrix for this composite estimate ($\hat{\mathbf{V}}_{x|M=m,y,x}$) is:

$$\hat{\mathbf{V}}_{x|M=m,y,x} = \mathbf{K}_{M=m}\hat{\mathbf{V}}_{y|M=m}\mathbf{K}'_{M=m} + (\mathbf{I} - \mathbf{K}_{M=m}\mathbf{H}_y)\hat{\mathbf{V}}_{x|M=m}(\mathbf{I} - \mathbf{K}_{M=m}\mathbf{H}_y)' \tag{19}$$

$$\hat{\mathbf{V}}_{x|M=m,y,x} = (\mathbf{I} - \mathbf{K}_{M=m}\mathbf{H}_y)\hat{\mathbf{V}}_{x|M=m} \tag{20}$$

Covariance matrices $\hat{\mathbf{V}}_{x|M=m}$ and $\hat{\mathbf{V}}_{y|M=m}$ are given in Equations 11 and 17, and 9 and 16, respectively. Equation 19, which is called the "Joseph form" in the Kalman filter, is more numerically reliable than the equivalent expression in Equation 20 (Maybeck, 1979, p. 237).

$\mathbf{K}_{M=m}$ in Equations 18, 19, and 20 is a $(kk_y)\times k_y$ matrix that places optimal weight on each element of the two vector estimates in Equation 18 using the minimum variance criterion. $\mathbf{K}_{M=m}$ is termed the gain matrix in the Kalman filter. It is analogous to the weight in the univariate composite estimator (e.g., see Green and Strawderman, 1986; Gregoire and Walters, 1988), which is inversely proportional to the variances of two a priori scalar estimates. The gain matrix $\mathbf{K}_{M=m}$ for this subpopulation is:

$$\mathbf{K}_{M=m} = \hat{\mathbf{V}}_{x|M=m}\mathbf{H}_y'\left[\mathbf{H}_y\hat{\mathbf{V}}_{x|M=m}\mathbf{H}_y' + \hat{\mathbf{V}}_{y|M=m}\right]^- \tag{21}$$

The bracketed term in Equation 21 is a singular covariance matrix. Therefore, Equation 21 uses the generalized inverse for a symmetric matrix (Graybill, 1969, pp. 113–115).

The $(kk_y)\times 1$ composite estimate $\hat{\mathbf{x}}_{M=m|y,x}$ for subpopulation $M = m$ (Equation 18) is the basis for the mth column in the estimated $k\times k$ contingency table $\hat{\mathbf{Z}}$, and is related to the $k\times 1$ vector $\mathbf{z}_{M=m}(N_{M=m}/N)$ in Equation 1. Estimate $\hat{\mathbf{z}}_{M=m}$ is a vector of proportions that sums to exactly 1, and it represents the estimated conditional probabilities of a map unit being assigned to any one of the k categories by the error-free reference classifier, given that the map unit is assigned to category m by the map classifier. $\hat{\mathbf{z}}_{M=m}$ is a linear transformation of $\hat{\mathbf{x}}_{M=m|y,x}$ (Equations 3 and 6), with its corresponding $k\times k$ covariance matrix $\hat{\mathbf{V}}_{z|M=m}$:

$$\hat{\mathbf{z}}_{M=m} = \mathbf{H}_z\,\hat{\mathbf{x}}_{M=m|y,x} \tag{22}$$

$$\hat{\mathbf{V}}_{z|M=m} = \mathbf{H}_z\,\hat{\mathbf{V}}_{x|M=m,y,x}\,\mathbf{H}_z' \tag{23}$$

The multivariate composite estimator (Equations 18 to 23) is applied k different times, once for each map category m; then, vector estimates $\hat{\mathbf{z}}_{M=m}$ are concatenated to form the estimated $k\times k$ contingency table $\hat{\mathbf{Z}}$ (Equation 1).

Let $Vec(\hat{\mathbf{Z}})$ be the $k^2\times 1$ vectorized version of $\hat{\mathbf{Z}}$:

$$Vec(\hat{\mathbf{Z}}) = \begin{bmatrix} \hat{\mathbf{z}}_{M=1}(N_{M=1}/N) \\ \vdots \\ \hat{\mathbf{z}}_{M=m}(N_{M=m}/N) \\ \vdots \\ \hat{\mathbf{z}}_{M=k}(N_{M=k}/N) \end{bmatrix} \tag{24}$$

Assume that sampling errors for each subpopulation $M = m$ are independent of the sampling errors for all other subpopulations by design. In this case, the estimation error covariance matrix for $Vec(\hat{\mathbf{Z}})$ is the $k^2\times k^2$ matrix with submatrices $\hat{\mathbf{V}}_{z|M=m}$ (Equation 23) on the diagonal, and all other elements equal 0:

$$\hat{\mathbf{V}}_{Vec(\hat{\mathbf{Z}})} = \begin{bmatrix} \hat{\mathbf{V}}_{z|M=1} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \cdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \hat{\mathbf{V}}_{z|M=m}\left(\frac{N_{M=m}}{N}\right)^2 & \cdots & \mathbf{0} \\ \vdots & \ddots & \cdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \hat{\mathbf{V}}_{z|M=k}\left(\frac{N_{M=k}}{N}\right)^2 \end{bmatrix} \quad (25)$$

This covariance matrix will be used to estimate standard errors of statistics that are functions of the estimated contingency table $\hat{\mathbf{Z}}$.

The multivariate composite estimator cannot always be used to estimate $\hat{\mathbf{z}}_{M=m}$ for subpopulation $M = m$ (Equations 18, 22, and 24). If the error-free classifier assigns all sample units in this subpopulation to only one category, say I, then the ith element of $\hat{\mathbf{z}}_{M=m}$ equals exactly 1, and $\hat{\mathbf{V}}_{x|M=m} = \mathbf{0}$. Alternatively, the sample estimate $\hat{\mathbf{y}}_{M=m}$ might not exist for this subpopulation; the association between the imperfect and error-free classifications might be very poor; or the imperfect classifier might assign all sample units in this subpopulation to only one category. If these situations occur, and the sample estimate with the error-free reference classifications (Equations 10 or 15) exists for this subpopulation, then Equations 22 and 23 for the composite estimator can be replaced by $\hat{\mathbf{z}}_{M=m} = \mathbf{H}_z\ \hat{\mathbf{x}}_{M=m}$ and $\hat{\mathbf{V}}_{z|M=m} = \mathbf{H}_z\ \hat{\mathbf{V}}_{x|M=m}\ \mathbf{H}'_z$ (see Equations 10, 11, 15, and 17), and used in Equations 24 and 25 to estimate the mth column of the contingency table. (This approach also satisfies situations in which $\hat{\mathbf{y}}_{M=m}$ does not exist by design because the imperfect reference classifier is omitted. This permits estimation of the contingency table with heterogeneous cluster plots rather than the typical homogeneous plots, which are classified into one and only one category by each classifier.) If the sample estimate $\hat{\mathbf{x}}_{M=m}$ does not exist, then no estimate for subpopulation $M = m$ is possible, and the mth column of the contingency table will be missing.

The multivariate composite estimator is vulnerable to numerical errors. This problem is common whenever random errors for one vector estimate are much greater than random errors for another, e.g., $det(\mathbf{H}_y\ \hat{\mathbf{V}}_{x|M=m}\ \mathbf{H}'_y) \gg det(\hat{\mathbf{V}}_{y|M=m})$, which can occur when sample sizes $n_{x|M=m}$ and $n_{x|M=m}$ are very different in Equations 8 to 17. Results should always be scrutinized for symptoms of numerical problems, such as: vectors of estimated proportions that do not sum to 1; corresponding covariance matrices that do not sum to 0; negative elements on the diagonal of any covariance matrix; or asymmetric covariance matrices. Solutions include the following:

- All computational routines (e.g., generalized inverse) should use maximal numerical precision and be robust to numerical problems.
- The "Joseph form" for $\hat{\mathbf{V}}_{x|M=m,y,x}$ in Equation 19 should be used rather than Equation 20 because the former is better conditioned and more numerically reliable (Maybeck, 1979, p. 237).
- Matrix dimensions can be reduced by eliminating rows of zeros and their corresponding columns in $(\mathbf{H}_y\hat{\mathbf{V}}_{x|M=m}\mathbf{H}'_y + \hat{\mathbf{V}}_{y|M=m})$ in Equation 21, although this requires careful

bookkeeping in order to expand the resulting composite estimate ($\hat{\mathbf{x}}_{M=m|y,x}$ in Equation 18) back to the original structure of $\hat{\mathbf{x}}_{M=m}$.

- Composite estimation algorithms that use square roots of the covariance matrices often solve stubborn numerical problems (Maybeck, 1979, pp. 377–391).

Validation methods are especially important in composite estimation with remotely sensed data, where logistic and technological difficulties often breed subtle procedural aberrations. Maybeck (1979, p. 229) shows that the vector of residual differences ($\mathbf{H}_y\hat{\mathbf{x}}_{M=m} - \hat{\mathbf{y}}_{M=m}$) has an expected covariance matrix ($\mathbf{H}_y\hat{\mathbf{V}}_{x|M=m}\mathbf{H}'_y + \hat{\mathbf{V}}_{y|M=m}$). Each element of this residual vector for each subpopulation m should be tested for bias, assuming the estimation errors are normally distributed with zero mean and variance on the diagonal of the estimated covariance matrix. Several suspiciously large residuals indicate potential procedural errors. However, off-diagonal elements of the covariance matrix are not necessarily zero, and these separate tests are not mutually independent for the same subpopulation. Therefore, multiple residuals for a subpopulation can be standardized so that they are expected to be independent and identically distributed. First, the dimensions of the residual vector and its covariance matrix are reduced to achieve full rank, then standardized residuals ($\mathbf{r}_{M=m}$) are computed with the Cholesky square root of its covariance matrix:

$$\mathbf{r}_{M=m} = \left[\left(\mathbf{H}_y\hat{\mathbf{V}}_{x|M=m}\mathbf{H}'_y + \hat{\mathbf{V}}_{y|M=m}\right)^{1/2}\right]^{-1}\left(\mathbf{H}_y\hat{\mathbf{x}}_{M=m} - \hat{\mathbf{y}}_{M=m}\right) \tag{26}$$

where $E[\mathbf{r}_{M=m}] = \mathbf{0}$ and $E[\mathbf{r}_{M=m}\mathbf{r}'_{M=m}] = \mathbf{I}$. If all sampling and estimation assumptions are valid, then the standardized residuals from all subpopulations can be pooled, and their pooled mean and variance have expected values of 0 and 1, respectively. These expectations are validated with a t-test for 0 mean with variance 1, and a χ^2 test for variance equal to 1 (see Hoel, 1984, pp. 140–143, 281–284, 298–300). If the validation tests cast doubt upon these expectations, then possible procedural problems should be investigated.

Estimates of Areal Extent

Environmental evaluations and forecasting models often require statistical tabulation of the area occupied by each cover category. These areal estimates are available through enumeration of map units that are classified into the k categories by the imperfect map classifier. However, misclassification can make this enumeration a biased estimate of the true areal extent, especially for rare cover types (Czaplewski, 1992a). Czaplewski and Catts (1992) and Walsh and Burk (1993) reviewed the literature that considers calibration for misclassification bias. However, they did not discuss unbiased areal estimates that use the row margin of the estimated contingency table $\hat{\mathbf{Z}}$ (Equation 24). The ijth element of $\mathbf{Z}$, denoted by Z_{ij}, is the proportion of map units in the population that are classified as category j by the map classifier and truly are category I; therefore, the proportion of map units that are truly category I (p_i) equals:

$$p_{i\cdot} = \sum_{j=1}^{k} Z_{ij} \tag{27}$$

To estimate the true areal extent of each category in the absence of misclassification, the estimated contingency table ($\hat{\mathbf{Z}}$) is substituted for the unknown true matrix in Equation 27, which provides an asymptotically unbiased estimator (Molina C., 1989, p. 122). Since estimation errors for elements of $\hat{\mathbf{Z}}$ are independent between sub populations, which are defined by the imperfect map classifier and denoted by subscript j in Equation 27, the variance for $p_{i\cdot}$, denoted $\hat{V}\hat{p}_{i\cdot}$, is simply the sum of the variances for each Z_{ij} on the diagonal of covariance matrix $\hat{\mathbf{V}}_{Vec(\mathbf{z})}$ in Equation 25.

Matrix algebra provides a concise formulation for areal estimates and their covariance matrix. Let the $k\times 1$ vector $\hat{\mathbf{p}}_I$ represent the estimated true proportions of each of the k categories, *i.e.*, the row margin of $\hat{\mathbf{Z}}$; $\hat{\mathbf{p}}_I$ is a linear transformation of $\hat{\mathbf{Z}}$:

$$\hat{\mathbf{p}}_{i\cdot} = \mathbf{D}'_{\mathbf{p}_{i\cdot}} Vec(\hat{\mathbf{Z}}) \tag{28}$$

where $\mathbf{Dp}_r$ is the following $k^2\times k$ matrix of zeros and ones:

$$\mathbf{D}_{\mathbf{p}_{i\cdot}} = \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \\ \vdots \\ \mathbf{I} \end{bmatrix} \tag{29}$$

$\mathbf{I}$ is the $k\times k$ identity matrix. The covariance matrix for random estimation errors in $\hat{\mathbf{p}}_I$ is the corresponding linear transformation of the $k^2\times k^2$ covariance matrix for $Vec(\hat{\mathbf{Z}})$ in Equation 25:

$$\hat{\mathbf{V}}_{\mathbf{p}_{i\cdot}} = \mathbf{D}'_{\mathbf{p}_{i\cdot}} \hat{\mathbf{V}}_{Vec(\hat{\mathbf{Z}})} \mathbf{D}_{\mathbf{p}_{i\cdot}} \tag{30}$$

Environmental evaluations can require confidence intervals for areal estimates. Confidence intervals often use the normal distribution with an estimated mean and variance. However, the normal distribution is unrealistic for proportions that are near 0 or 1, where the binomial distribution is a more reasonable assumption. The parameters of the binomial distribution are the number of independent trials (n) and the estimated probability of success ($\hat{p}$), with estimated variance $\hat{V}_{\hat{p}} = \hat{p}(1-\hat{p})/\mathrm{n}$. The number of trials ($n$), in the context of a parameter of the binomial distribution, does not pertain to composite estimates; however, the composite estimator does provide estimates $\hat{p}$ and $\hat{V}_{\hat{p}}$. Based on the method of moments approach that Brier (1980) used with cluster sampling for the Dirichlet-multinomial distribution, $\hat{p}$ and $\hat{V}_{\hat{p}}$ provide an estimate of parameter n for the binomial distribution:

$$\hat{\mathrm{n}} = \frac{\hat{\mathrm{p}}(1-\hat{\mathrm{p}})}{\hat{\mathrm{V}}_{\hat{\mathrm{p}}}} \tag{31}$$

The approximate confidence bounds ($\hat{p}_{LO}$ and $\hat{p}_{UP}$) for $\hat{p}$ may be estimated as follows (Rothman, 1986, p. 167). Let a represent an arbitrary confidence level, e.g., $\alpha = 0.95$, and $P(K \geq \hat{p}\hat{n})$ represent the probability that $(\hat{p}\hat{n})$ or more successes are observed out of $\hat{n}$ Bernoulli trials, where $(\hat{p}\hat{n})$ and $\hat{n}$ are rounded to the nearest integer. The lower confidence bound is the probability of success $(\hat{p}_{LO})$ with the binomial distribution for which $P(K \geq \hat{p}\hat{n}) = (1-\alpha)/2$, and the upper bound is the $\hat{p}_{UP}$ for which $P(K \leq \hat{p}\hat{n}) = (1-a)/2$. Solution is by bisection. Confidence intervals so computed are always bounded by 0 and 1. Except when $\hat{p}$ is near 0 or 1, these confidence intervals are approximately equal to intervals that assume a normal distribution for a sufficiently large sample size.

Accuracy Assessment Statistics and Variances

The most common statistic used to assess accuracy is the total proportion of correctly classified map units (p_o), called overall accuracy by Congalton (1991). Once the contingency table is estimated (Equations 22 and 24), the estimated overall accuracy ($\hat{p}_o$) simply equals the sum of the diagonal elements of $\hat{\mathbf{Z}}$. The variance for the overall accuracy statistic $\hat{V}\hat{p}$© is the sum of the diagonal elements of $\hat{\mathbf{V}}_{Vec(Z)}$ (Equations 23 and 25) that correspond to diagonal elements of $\hat{\mathbf{Z}}$. This is expressed in matrix algebra as a prelude to formulae for more complex accuracy statistics:

$$\hat{p}_o = \mathbf{d}'_{p_o} Vec(\hat{\mathbf{Z}}) \tag{32}$$

$$\hat{V}_{\hat{p}_o} = \mathbf{d}'_{p_o} \hat{\mathbf{V}}_{Vec(\hat{Z})} \mathbf{d}_{p_o} \tag{33}$$

where $\mathbf{d}p_c$ is the $k^2 \times 1$ vector in which the $[(I-1)k+i]$th elements equal 1 ($I = 1,2,...,k$), and all other elements equal 0. Confidence intervals are approximated with $\hat{p}_o$, $\hat{V}\hat{p}_c$, and the binomial assumption (Equation 31).

The expected probabilities of correct classification for a category through chance agreement equals the product of the row and column margins for that category in the contingency table. This hypothesis is the basis for Cohen's kappa test, which is commonly applied in remote sensing studies (Congalton, 1991). The column margin for $\hat{\mathbf{Z}}$ is known exactly through the census of map units, each of which is categorized by the map classifier. Let $k \times 1$ vector $\mathbf{p}_{\cdot j}$ represent the column margin of $\hat{\mathbf{Z}}$, in which the ith element equals the constant $N_{M=I}/N$ (Equation 1). The estimated expected accuracy under chance agreement $\hat{p}$© is the product of the fixed column ($\mathbf{p}_{\cdot j}$) and the estimated row margin ($\hat{\mathbf{p}}_{I\cdot}$ in Equation 28), with covariance matrix $\hat{V}_{\hat{p}_c}$:

$$\hat{p}_c = \mathbf{p}'_{\cdot j} \hat{\mathbf{p}}_{i\cdot} = \mathbf{p}'_{\cdot j} \mathbf{D}'_{p_{i\cdot}} Vec(\hat{\mathbf{Z}}) = \mathbf{d}'_{p_c} Vec(\hat{\mathbf{Z}}) \tag{34}$$

$$\hat{V}_{\hat{p}_c} = \mathbf{d}'_{p_c} \hat{\mathbf{V}}_{Vec(\hat{Z})} \mathbf{d}_{p_c} \tag{35}$$

$$\mathbf{d}'_{p_c} = \mathbf{p}'_{\cdot j} \mathbf{D}'_{p_{i\cdot}} = \left[\mathbf{p}'_{\cdot j} | \mathbf{p}'_{\cdot j} | \cdots | \mathbf{p}'_{\cdot j}\right] \tag{36}$$

This suggests a test of hypothesis that the overall classification accuracy is no greater than that expected by chance, i.e., the difference between the observed overall accuracy ($\hat{p}_o$ in Equation 32) and chance accuracy ($\hat{p}_c$ in Equation 34) equals zero:

$$\hat{\mathbf{p}}_o - \hat{\mathbf{p}}_c = [1-1]\begin{bmatrix}\hat{\mathbf{p}}_o \\ \hat{\mathbf{p}}_c\end{bmatrix} = [1-1]\begin{bmatrix}\mathbf{d}'_{\mathbf{p}_o} \\ \mathbf{d}'_{\mathbf{p}_c}\end{bmatrix} Vec(\hat{\mathbf{Z}}) \tag{37}$$

$$\hat{\mathbf{V}}_{\hat{\mathbf{p}}_o-\hat{\mathbf{p}}_c} = [1-1]\begin{bmatrix}\mathbf{d}'_{\mathbf{p}_o} \\ \mathbf{d}'_{\mathbf{p}_c}\end{bmatrix} \hat{\mathbf{V}}_{Vec(\hat{\mathbf{Z}})} \begin{bmatrix}\mathbf{d}'_{\mathbf{p}_o} \\ \mathbf{d}'_{\mathbf{p}_c}\end{bmatrix} [1-1] \tag{38}$$

The normal distribution with mean 0 and variance $\hat{\mathrm{V}}\hat{p}_o - \hat{p}_c$ provides an approximate probability that the null hypothesis is true, although the limiting distribution of this statistic has not been established. Czaplewski (1994) provides a Taylor series approximation for covariance matrix $\hat{\mathbf{V}}_{Vec(Z)}$ in Equation 38 that is more compatible with the null hypothesis in this test.

Some land cover types are very accurately categorized by a map classifier, while others are less successfully classified. The overall accuracy statistic does not isolate differences among individual categories. However, conditional probabilities of correct classification concisely and intuitively describe these differences (e.g., Fleiss, 1981, p. 214). Green et al. (1993) and Stehman (1996) developed variance estimators for these statistics when prestratification is based on the remotely sensed classification, there is simple random sampling within strata, and sample units are homogeneous so that the multinomial distribution applies. The following two paragraphs provide more general results for estimates of conditional probabilities from double sampling.

Consider the conditional probability of correct classification given that the map classifier assigns a map unit to category $M = m$. This is also called "user's accuracy" in the remote sensing and quality control literatures (Congalton, 1991). The vector $\hat{\mathbf{z}}_{M=m}$ in Equation 22 is the transformed composite estimate of the conditional probabilities given that the map classifier assigned the map unit to the mth category (i.e., subpopulation $M = m$). Therefore, user's accuracy for category m is simply the mth element of $\hat{\mathbf{z}}_{M=m}$, with an estimated variance equal to the mth diagonal element of $\hat{\mathbf{V}}_{z|M=m}$ in Equation 23. The confidence interval for user's accuracy for the mth category is approximated with the binomial distribution (Equation 31).

The conditional probability of correct classification, given that the error-free reference classification is category m, is called "producer's accuracy" (Congalton, 1991). The $k\times1$ vector of producer's accuracies, $\hat{\mathbf{p}}_{(I|I\cdot)}$, equals the $k\times1$ diagonal of the $k\times k$ contingency table, $diag(\hat{\mathbf{Z}})$, divided by its row margin (Equation 28). The covariance matrix for producer's accuracies is more difficult to estimate than the covariance matrix for user's accuracies because estimation errors among rows of $\hat{\mathbf{z}}_{M=m}$ are not independent (Equation 25). Czaplewski (1994) derives a Taylor series approximation for this situation. First, define the $k\times k$ matrix $\mathbf{H}_{p(I\cdot)}$, in which all off-diagonal elements are zero, and all diagonal elements equal the iith element of $\hat{\mathbf{Z}}$ divided by the squared inverse of the ith element of $\hat{\mathbf{p}}_{I\cdot}$, $I \in \{1,\cdots,k\}$, in Equation 28. Next, define the $k\times k$ matrix $\mathbf{G}_i$, $I \in \{1,\cdots,k\}$, in which all elements

are zero except the *ii*th element, which equals the inverse of the *i*th element of $(\hat{\mathbf{p}}_I)$. Finally, define the $k^2 \times k$ matrix $\mathbf{D}_{(I|I\cdot)}$ as:

$$\mathbf{D}_{(i|i\cdot)} = \begin{bmatrix} \mathbf{G}_1 \\ \vdots \\ \mathbf{G}_m \\ \vdots \\ \mathbf{G}_k \end{bmatrix} - \begin{bmatrix} \mathbf{H}_{p(i\cdot)} \\ \vdots \\ \mathbf{H}_{p(i\cdot)} \\ \vdots \\ \mathbf{H}_{p(i\cdot)} \end{bmatrix} \tag{39}$$

The approximate $k \times k$ covariance matrix for the $k \times 1$ vector of estimated producer's accuracies $(\hat{\mathbf{p}}_{(I|I\cdot)})$ equals:

$$\hat{\mathbf{V}}_{\hat{\mathbf{p}}_{(i|i\cdot)}} = \mathbf{D}'_{(i|i\cdot)} \hat{\mathbf{V}}_{Vec(\hat{\mathbf{Z}})} \mathbf{D}_{(i|i\cdot)} \tag{40}$$

Czaplewski (1994) provides examples of the matrix in Equation 40. The confidence intervals for elements of $\hat{\mathbf{p}}_{(I|I\cdot)}$ (producer's accuracies) are approximated with their corresponding diagonal elements of covariance matrix $\hat{\mathbf{V}}\hat{\mathbf{p}}_{(I|I\cdot)}$ and the binomial distribution (Equation 31). In addition, Czaplewski (1994) formulates a test of the hypothesis that individual user's and producer's accuracies do not differ from what is expected by chance.

The variances in Equations 35 and 38, and the covariance matrix for user's accuracies, assume that the column margin $\mathbf{p}_{\cdot j}$ is a vector of known constants. This is true when the map classifier is applied to all members of the population, i.e., the column margin of $\hat{\mathbf{Z}}$ is fixed through a census of all pixels. However, these equations do not apply to more general situations, which are considered by Czaplewski (1994).

The coefficient of agreement (kappa) also quantifies overall accuracy in a contingency table $\hat{\mathbf{Z}}$ relative to that expected by chance (Cohen, 1960, 1968). This statistic is commonly used in remote sensing (Congalton, 1991). It is sometimes weighted to consider partial agreement. Let $\mathbf{W}$ represent the $k \times k$ matrix of constants in which the *ij*th element (w_{ij}) is the weight or "partial credit" chosen by the user for the agreement when a map unit is classified as category j by the imperfect map classifier and category I by the error-free reference classifier. The unweighted kappa statistic is merely a special case of the weighted kappa, in which $\mathbf{W} = \mathbf{I}$, the identity matrix. Let $\hat{p}_{o|W} = \mathbf{1}'[\mathbf{W}^*\hat{\mathbf{Z}}]\mathbf{1}$ represent the weighted overall accuracy observed in the sample, and $\hat{p}_{c|W} = \mathbf{1}'[\mathbf{W}^*(\hat{\mathbf{p}}_{I\cdot}\mathbf{p}'_{\cdot j})]\mathbf{1}$ (see Equations 28 and 34) represent the corresponding accuracy expected by chance, where $\mathbf{1}$ is the $k \times 1$ vector of 1's and * represents element-by-element multiplication (i.e., the *ij*th element of $\mathbf{A}^*\mathbf{B}$ is $a_{ij}b_{ij}$). The estimated weighted kappa statistic $(\hat{\kappa}_w)$ is:

$$\kappa_w = \frac{\hat{p}_{o|W} - \hat{p}_{c|W}}{1 - \hat{p}_{c|W}} \tag{41}$$

Cohen derived a variance approximation for kappa in the special case of simple random sampling, where each unit is classified into one and only one category, and Stehman (1996, 1997) has developed variance approximations for stratified random sampling. However, these do not apply to the multivariate composite estimator. Czaplewski (1994) derives a Taylor-series variance approximation for the weighted kappa statistic (k_w) in the more general case, where the contingency table $\hat{\mathbf{Z}}$ is estimated with any appropriate method that provides covariance matrix $\hat{\mathbf{V}}_{Vec(Z)}$ (e.g., Equation 25). First, define the $k\times1$ vectors $\mathbf{w}_{i\cdot} = (\mathbf{W}\,\mathbf{p}_{\cdot j})$ from Equation 34, and $\hat{\mathbf{w}}_{\cdot j} = (\mathbf{W}'\hat{\mathbf{p}}_{I\cdot})$ from Equation 28, and $k^2\times1$ vector $\mathbf{d}_\kappa$:

$$\mathbf{d}_\kappa = \frac{Vec(\mathbf{W})}{(1-\hat{p}_{c|W})} - \frac{(1-\hat{p}_{o|W})}{(1-\hat{p}_{c|W})^2}\left\{\begin{bmatrix}\mathbf{w}_{1\cdot}\\ \vdots \\ \mathbf{w}_{m\cdot}\\ \vdots \\ \mathbf{w}_{k\cdot}\end{bmatrix} + Vec\begin{bmatrix}\hat{\mathbf{w}}'_{\cdot 1}\\ \vdots \\ \hat{\mathbf{w}}'_{\cdot m}\\ \vdots \\ \hat{\mathbf{w}}'_{\cdot k}\end{bmatrix}\right\} \tag{42}$$

where $Vec(\mathbf{W})$ is the $k^2\times1$ vector version of the $k\times k$ weighting matrix $\mathbf{W}$, and the Vec operator is defined for $\mathbf{x}_{M=m}$ in Equation 2; Czaplewski (1994) provides examples. Next, the approximate variance of $\hat{\kappa}_w$ is estimated as:

$$\hat{V}_{\hat{\kappa}_w} = \mathbf{d}'_\kappa\,\hat{\mathbf{V}}_{Vec(\hat{\mathbf{Z}})}\,\mathbf{d}_\kappa \tag{43}$$

The estimated kappa and its variance (Equations 41 and 43) are used with the normal distribution to estimate confidence intervals and test hypotheses. Czaplewski (1994) derives other variance approximations for the weighted and unweighted kappa statistics under the null hypothesis of chance agreement, and the conditional kappa statistics (Light, 1971) for the rows and columns of the contingency table.

Equations 30, 33, 35, 40, and 43 represent variance estimators for areal extent and statistics that assess accuracy of classifications on the map. They are formulated as vector operators on the covariance matrix $\hat{\mathbf{V}}_{Vec(z)}$, which is highly partitioned (Equation 25) because of the assumed independence of sampling errors among subpopulations. This partitioned structure can reduce numerical errors and matrix dimensions. Let the following represent this general structure:

$$\begin{bmatrix}\mathbf{D}'_1 \cdots \mathbf{D}'_m \cdots \mathbf{D}'_k\end{bmatrix}\begin{bmatrix}\mathbf{V}_1 & \cdots & \mathbf{0} & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & \mathbf{V}_m & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{V}_k\end{bmatrix}\begin{bmatrix}\mathbf{D}_1\\ \vdots\\ \mathbf{D}_m\\ \vdots\\ \mathbf{D}_k\end{bmatrix} = \sum_{m=1}^{k}\mathbf{D}'_m\mathbf{V}_m\mathbf{D}_m \tag{44}$$

The left-hand side of Equation 44 contains a $k^2 \times k^2$ covariance matrix; however, the right-hand side contains $k \times k$ covariance matrices. This is very important for detailed classification systems. For example, if the number of categories is $k = 20$, then the covariance matrix on the left-hand side has dimensions 8000×8000, while those on the right-hand side have dimensions 400×400.

DISCUSSION

Classification systems frequently have detailed categories to meet the needs of particular analyses, but accuracy of map classifiers is typically poor for detailed classification systems. The classifier often confuses certain similar categories, and the number of similar categories increases as the classification system becomes more detailed. While increased thematic resolution provides more information about spatial patterns, increased resolution makes modeling of those patterns more difficult (Costanza and Maxwell, 1994). Therefore, the map analyst must often simplify the classification system to attain maps that have reasonable classification accuracy. Statistical assessment of classification accuracy naturally leads to more informed decisions regarding these simplifications. Conditional probabilities of correct classification (user's and producer's accuracies for each category) help strike a compromise between analysts' applications and the classification errors that are endemic to thematic mapping. These compromises are facilitated with user-friendly software, copies of which are available from the author. Other software systems are capable of generating the necessary estimates, especially systems with a matrix language and a library of linear algebra routines.

Occasionally, random errors severely distort estimates of accuracy. Gains in efficiency from the imperfect reference classifier (e.g., photo-interpretation) reduce risk of incorrect evaluations of map reliability. For this reason, reasonable confidence intervals and variance estimates for accuracy statistics are crucial for prudent use of thematic maps. Estimates of classification accuracy improve as registration accuracy increases, as sample sizes for estimates $\hat{\mathbf{y}}_{M=m}$ and $\hat{\mathbf{x}}_{M=m}$ increase (Equations 8 to 17), and as accuracy of the imperfect reference classifier increases. The latter might improve with reductions in classification detail (k_y in Equation 2). All these components affect the cost and reliability of estimating areal extent and accuracy assessments of thematic maps.

Spatial analyses with geographic information systems often generate areal estimates from a small portion of a thematic map, but these estimates can include substantial misclassification. Multivariate calibration methods can correct for this bias (Tenenbein, 1972; Czaplewski and Catts, 1992; Walsh and Burk, 1993) with a transformation of the contingency table ($\hat{\mathbf{Z}}$ from Equations 1 and 24). However, the calibrated areal estimates do not identify the spatial location of classification errors. Also, the calibration model assumes that misclassification probabilities are the same for all portions of the thematic map, which might not be reasonable for small-area estimates. For example, mountain shadows often increase classification errors with multispectral satellite data. The contingency table incorporates this source of error in proportion to the amount of shadow within the entire map. However, a portion of the map can have a very different proportion of shadow, and thus, very different misclassification probabilities. Bauer et al. (1994) propose a solution to this problem, in which a univariate composite method combines predictions from local and global calibration models.

The multivariate composite approach can be expanded to multiway contingency tables. This permits hierarchical loglinear models, and related logit models, and associated methods for systematic testing of hypotheses, similar to analysis of variance for continuous data (Rao and Thomas, 1989; Molina C., 1989). Such methods permit testing hypotheses related to causes of classification errors, and possibly other hypotheses. However, the methods used in this chapter avoid logarithmic transformations for estimation because retransformation bias is very problematic, especially for proportions near zero.

The methods presented here include cluster sampling without the imperfect reference classifier as a special case. This efficient estimator is an alternative to the univariate methods of Stehman (1997). In addition, the difference between the realized overall accuracy and the overall accuracy expected by chance alone (Equation 37) is a new indicator of classification accuracy; Equation 38 provides the variance estimator for this statistic under prestratification. Czaplewski (1994) derives their variance estimators. These statistics provide a more reasonable null hypothesis (Stehman, in press), and are more easily interpreted than Cohen's kappa and partial kappa statistics.

The methods proposed above also apply to situations where the imperfect reference classifier is not photo-interpretation. For example, the imperfect reference data might represent sample units from a different monitoring program, where inconsistencies in protocol and definitions cause "imperfect" classifications, or the imperfect reference data might come from an old survey, where recent changes in land cover cause previous classifications to be imperfect. Lastly, the reference classifications might be considered imperfect if they come from plots that are not well registered between the satellite images and their field locations. The cost of accurate registration could be restricted to a subsample of "error-free" plots.

The methods in this chapter assume each sample unit has the same configuration for the map classifier, the imperfect reference classifier, and error-free reference classifier, which is similar to sample units in a two-phase sampling design. A cluster of pixels in a 1-ha field plot is an example. However, photo-interpretation of larger sample units often adds little marginal cost (Czaplewski and Catts, 1988; Bauer et al., 1994), as in a two-staging sample design. The multivariate composite estimator can accommodate two-stage designs, but requires a different formulation from that considered here.

The composite estimator assumes independence of sampling errors among subpopulations. This assumption is suspect for poststratification by subpopulation of a simple random sample because sample sizes for each subpopulation are not fixed by design. This assumption is also suspect if a cluster plot contains map units from more than one subpopulation. Practical considerations make simple random sampling and heterogeneous cluster plots important options in assessing accuracy. Although the independence assumption is not strictly required by the composite estimator (Czaplewski, 1992a), the independence assumption does reduce numerical problems (e.g., Equation 44). The bootstrap estimator exploits the numerical advantages of this independence assumption, and accounts for dependent sampling errors through resampling. However, bootstrap variance estimates might require hours of computation time, compared with seconds for composite variance estimates. Therefore, the composite estimator can provide reasonable preliminary estimates, especially when assessments of accuracy are used to help simplify the classification system, while final estimates are made with bootstrap methods.

The composite estimator ($\hat{\mathbf{Z}}$ from Equations 18 to 24) is the unbiased minimum variance estimator if $\hat{\mathbf{y}}_{M=m}$ and $\hat{\mathbf{x}}_{M=m}$ are unbiased and the covariance matrices $\mathbf{V}_x$ and $\mathbf{V}_y$ are known. However, only the estimates $\hat{\mathbf{V}}_x$ and $\hat{\mathbf{V}}_y$ are usually known; the composite estimator remains unbiased, but will be suboptimal. Since $\hat{\mathbf{Z}}$ is asymptotically unbiased, and the statistics used to estimate areal extent and assess accuracy are functions of the k^2 estimates in $\hat{\mathbf{Z}}$, these estimates will also be asymptotically unbiased (Molina C., 1989, p. 122; Särndal et al., 1992, p. 168). However, vector estimates $\hat{\mathbf{y}}_{M=m}$ and $\hat{\mathbf{x}}_{M=m}$ in Equations 8 to 17 can be biased for small sample sizes because each zero element in $\hat{\mathbf{y}}_{M=m}$ and $\hat{\mathbf{x}}_{M=m}$ is treated as though the true probability is exactly zero (Bishop et al., 1975, p. 177). A zero element might represent a rare event that was not observed in the sample (sampling zero), even though the true probability of the event exceeds zero. This can bias the composite estimator. Deficiencies in sampling design and measurement protocols can introduce additional biases. The verification techniques described earlier in the Methods section should always be used to detect subtle inconsistencies in the data collection, processing, and estimation processes.

A reference classifier is rarely free of all classification error. Errors are introduced into reference data by errors in locating sample plots on the ground and on the remotely sensed imagery, changes in plot condition over time, measurement and recording errors, and within-plot sampling errors. However, accuracy assessments treat the highest resolution reference data as though they were perfect. The term "error-free reference classifier" emphasizes this assumption to the user of an accuracy assessment. The user should thoroughly understand how the reference data were gathered, and condition their interpretation of the accuracy assessment based on this understanding.

CONCLUSIONS

The contingency table is sufficient for assessing accuracy and estimating the area of different categories of land cover, while maintaining compatibility between these two objectives. The multivariate composite estimator provides efficient estimates of the contingency table because it combines multiple sources of data with an asymptotically unbiased, minimum variance approach. The multivariate composite estimator is a relatively simple method, which permits use of more complex sampling designs and sample units. With statistical methods presented in this chapter, the complexity of the sampling design and estimation method might no longer be a major problem. Rather, the remaining problems might be more pragmatic, such as designing sample units that can be accurately registered to different types of imagery, confidently located in the field, and classified with objective and repeatable measurements in the field.

Classification systems for remote sensing are often very detailed to meet analytical needs, but classification accuracy can be low. Accuracy increases if similar categories of land cover are combined (i.e., collapsing classification system). A statistically reliable assessment of accuracy can help the analyst decide how to best compromise classification detail without sacrificing the objectives of a specific analysis. This interaction between analysts and their geographic data is extremely important (Goodchild and Gopal, 1989). The composite estimator quickly provides the analyst with an approximate description of the consequences of potential compromises.

ACKNOWLEDGMENTS

I appreciate the assistance and helpful suggestions of Steve Stehman, Oliver Schabenberger, Jim Alegria, Geoff Wood, Russell Congalton, Mike Goodchild, C.Y. Ueng, Hans Schreuder, David Evans, Rudy King, Hans Bodmer, Mohammed Kalkhan, and an anonymous reviewer. Any errors remain my responsibility. This work was supported by the Forest Inventory and Analysis Program, the National Forest System's Pacific Northwest Region, and the Forest Health Monitoring Program of the USDA Forest Service.

REFERENCES

Agresti, A. Categorical Data Analysis. John Wiley & Sons, New York, 1990.

Arbia, G. The use of GIS in spatial statistical surveys. *Int. Stat. Rev.* 61, pp. 339–359, 1993.

Bauer, M.E., T.E. Burk, A.R. Ek, P.R. Coppin, S.D. Lime, T. A. Walsh, D.K. Walters, W. Befort, and D.F. Heinzen. Satellite inventory of Minnesota forest resources. *Photogramm. Eng. Remote Sens.* 60, pp. 287–298, 1994.

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

Brier, S.S. Analysis of contingency tables under cluster sampling. *Biometrika* 67, pp. 591–596, 1980.

Christensen, R. *Linear Models for Multivariate, Time Series, and Spatial Data.* Springer-Verlag, New York. 1991.

Cohen, J.A. Coefficient of agreement for nominal scales. *Edu. Psychol. Meas.* 20, pp. 37–46, 1960.

Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* 70, pp. 213–220, 1968.

Congalton, R.G. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.* 37, pp. 35–46, 1991.

Congalton, R.G. and K. Green. A practical look at the sources of confusion in error matrix generation. *Photogramm. Eng. Remote Sens.* 59, pp. 641–644, 1993.

Costanza, R. and T. Maxwell. Resolution and predictability: An approach to the scaling problem. *Landscape Ecol.* 9, pp. 47–57, 1994.

Czaplewski, R.L. and G.P. Catts. Calibration of remotely sensed proportion or area estimates for misclassification error. *Remote Sens. Environ.* 39, pp. 29–43, 1992.

Czaplewski, R.L. Misclassification bias in areal estimates. *Photogramm. Eng. Remote Sens.* 58, pp. 189–192, 1992a.

Czaplewski, R.L. Accuracy Assessment of Remotely Sensed Classifications with Multi-phase Sampling and the Multivariate Composite Estimator, in Vol. 2, *Proceedings of the 14th International Biometric Conference.* International Biometrics Society, Ruakura Agricultural Centre, Hamilton, New Zealand, 1992b, p. 22.

Czaplewski, R.L. *Variance Approximations for Assessments of Classification Accuracy.* USDA For. Serv. Res. Pap. RM-316, USDA Forest Service, Rocky Mountain Research Station, Fort Collins, CO, 1994.

Fleiss, J.L. *Statistical Methods for Rates and Proportions,* 2nd ed. John Wiley & Sons, Inc. New York, 1981.

Goodchild, M.F. and S. Gopal. *Accuracy of Spatial Databases.* Taylor and Francis, London, 1989.

Graybill, F.A. *Introduction to Matrices with Applications in Statistics.* Wadsworth Publishing Co., Belmont, CA, 1969.

Green, E.J. and W.E. Strawderman. Reducing sample size through the use of a composite estimator: An application to timber volume estimation. *Can. J. For. Res.* 16, pp. 1116–1118, 1986.

Green, E.J., W.E. Strawderman, and T.M. Airola. Assessing classification probabilities for thematic maps. *Photogram. Eng. Remote Sens.* 59, pp. 635–639, 1993.

Gregoire, T.G. and D.K. Walters. Composite vector estimates derived by weighting inversely proportional to variance. *Can. J. For. Res.* 18, pp. 282–284, 1988.

Hoel, P.G. *Introduction to Mathematical Statistics,* 5th ed. John Wiley & Sons, New York, 1984.

Holt, D., T.M.F. Smith, and P.D. Winter. Regression analysis of data from complex surveys. *J. R. Statist. Soc. Ser. A* 143, pp. 303–20, 1980.

Light, R.J. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychol. Bull.* 76, pp. 363–377, 1971.

Maybeck, P.S. *Stochastic Models, Estimation, and Control, Volume 1.* Academic Press, New York, 1979.

Molina C., E.A, Measures of Association for Contingency Tables, in *Analysis of Complex Surveys,* C.J. Skinner, D. Holt, and T.M.F. Smith, Eds., John Wiley & Sons, Inc., New York, 1989.

Rao, J.N.K. and D.R. Thomas. Chi-Squared Tests for Contingency Tables, in *Analysis of Complex Surveys,* C.J. Skinner, D. Holt, and T.M.F. Smith, Eds., John Wiley & Sons, Inc., New York, 1989.

Rothman, K.J. *Modern Epidemiology,* Little, Brown and Co., Boston, 1986.

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

Schreuder, H.T., J. Sedransk, and K.D. Ware. 3P Sampling and Some Alternatives, I. *For. Sci.* 14, pp. 429–454, 1968.

Stehman, S.V. Estimating the Kappa coefficient and its variance under stratified random sampling. *Photogram. Eng. Remote Sens.* 62, pp. 401–402, 1996.

Stehman, S.V. Estimating standard errors of accuracy assessment statistics under cluster sampling. *Remote Sens. Environ.* 60, pp. 258–269, 1997.

Stehman, S.V. Selecting and Interpreting Measures of Thematic Classification Accuracy. *Remote Sens. Environ.* 62, pp. 77–89 , 1997.

Stehman, S.V. and R.L. Czaplewski. Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sens. Environ.*, 64, in press, 1998.

Tenenbein, A.A. Double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics* 14, pp. 755–758, 1972.

Van Deusen, P.C. Correcting bias in change estimates from thematic maps. *Remote Sens. Environ.* 50, pp. 67–73, 1994.

Walsh, T.A. and T.E. Burk. Calibration of satellite classifications of land area. *Remote Sens. Environ.* 46, pp. 281–290, 1993.

Williamson, G.D. and M. Haber. Models for three-dimensional contingency tables with completely and partially cross-classified data. *Biometrics* 49, pp. 194–203, 1994.

not cited in text — should it be deleted from list?