Toward Estimation of Map Accuracy Without a Probability Test Sample

Brian M. Steele, David A. Patterson and Roland L. Redmond

Abstract

The time and effort required of probability sampling for accuracy assessment of large-scale land cover maps often means that probability test samples are not collected. Yet, map usefulness is substantially reduced without reliable accuracy estimates. In this article, we introduce a method of estimating the accuracy of a classified map that does not utilize a test sample in the usual sense, but instead estimates the probability of correct classification for each map unit using only the classification rule and the map unit covariates. We argue that the method is an improvement over conventional estimators, though it does not eliminate the need for probability sampling. The method also provides a new and simple method of constructing accuracy maps. We illustrate some of problems associated with accuracy assessment of broad-scale land cover maps, and our method, with a set of nine Landsat Thematic Mapper satellite image-based land cover maps from Montana and Wyoming, USA.

Keywords: classification, discriminant analysis, posterior probabilities, spatial data

1. Introduction

A land cover map is constructed by partitioning a geographic area of interest into a finite set of map units and assigning a land cover class label to each unit. For maps with many units, label assignment or classification must be done automatically. A popular method is to measure one or more predictor variables on all map units by a remote sensing device such as a satellite, and land cover on a sample of map units. A classification rule is then constructed from the latter "training" sample and used to predict land cover for the unsampled units using the remotely sensed data. Inevitably, some map units will be incorrectly classified, and thus accuracy assessment is important, if not essential, for interpretation of the results.

There are two general approaches to thematic accuracy assessment. Given sufficient resources and time, the most desirable approach is to collect an additional set of test data by post-classification sampling (Stehman and Czaplewski, 1998; Stehman, 2000) and to compare observed and predicted land cover classes at the sample locations. Alternatively, both cross-validation and bootstrapping may be used to assess map accuracy from the same training set used to classify the imagery (Efron and Tibshirani, 1997; Hand, 1997; Schavio and Hand, 2000). Herein, this approach is referred to as resampling, as the training set is sampled to obtain the estimates.

Both approaches can lead to biased accuracy estimators depending on how the post-classification or training sample is chosen. For example, if regions, or land cover classes that are easy-to-classify are disproportionately sampled relative to their areal extent, then map accuracy estimates are likely to be optimistically biased (Hammond

and Verbyla, 1996). Conversely, if sample locations tend to fall disproportionately in difficult-to-classify regions or land cover classes, then accuracy estimates may be pessimistically biased. To avoid bias, the sample should be a simple random sample or chosen by some other probability sampling plan (such as a stratified random sample) whose structure can be adjusted for in computing the accuracy estimates. In practice, it may not be feasible to collect either a training set or a post-classification set by probability sampling, particularly when the map area is large (Vogelmann *et al*, 1998; Zhu *et al*, 1999). Consequently, training data are sometimes assembled from a variety of existing sources, some of which may be probability samples from subregions of the larger area, whereas others may be collected without intentional bias, but also without benefit of a probabilistic sampling plan.

In this article, we discuss a mapping project covering 21.5 million hectares of forested mountains and rangeland within and adjacent to the Rocky Mountains in northwestern USA. The intended uses of the map encompass commercial timber and range management, hydrology, fisheries, and wildlife management. As shown in Figure 1, the geographic extent of the area was defined by the perimeter of nine adjacent Landsat TM scenes. A single date image for each scene was classified independently, then all nine classifications were edge-matched in a geographic information system (GIS) to create a seamless land cover database for the entire area. Further details about the classification methods will be presented in later sections. Suffice it to say for now that the USDA Forest Service initiated the project and provided most of the training observations from existing data that were collected for purposes other than

mapping land cover. Some, but not all of these training data were collected by probability sampling of TM scene subregions. The overall spatial distribution of training observations was highly irregular, largely because most were sampled from public lands that were reasonably accessible; in other words, privately owned lands and wilderness areas were not as well-sampled due to their relative inaccessibility and/or high cost of access. For these reasons, none of the nine training data sets constitute a probability sample, and post-classification sampling apparently was not entertained due to its anticipated high cost. Yet without some measure of accuracy, the value of the resulting land cover map and its underlying digital database is greatly diminished.

Insert Figure 1 about here

A statistician presented with a request to assist in assessing map accuracy in this situation ultimately has two choices: decline with the admonishment that next time, some of the budget should be reserved for post-classification sampling, or accept the request and attempt to deal with the limitations of the training data. We have chosen the latter course, and propose a method of estimating accuracy for situations in which a probability test sample does not exist, either because the training data do not constitute a probability sample, or because a post-classification probability sample was not collected. We argue that the proposed method is substantially better than resampling methods in the absence of a probability sample. More importantly, by using our method with a small post-classification sample, it may be possible obtain accurate estimates with reduced effort and affordable cost.

Terminology and notation are introduced in Section 2, and Section 3 describes the classifiers. The proposed method of accuracy estimation is developed in Section 4, and in Section 5, the method is applied to the land cover mapping project. Before proceeding further, we present an example from one of the nine Landsat TM scenes to illustrate the pitfalls of using resampling methods with non-uniformly spatially distributed training observations to estimate land cover map accuracy.

1.1 An Example

Figure 2 shows the spatial distribution of 1422 training observations within a portion ($\sim 25\%$) of one TM scene (Path 39/Row 28; see Figure 1), along with the results of two supervised classifications of land cover—one based on a covariate-only classifier, (a weighted k-nearest neighbor), and the other a spatial combination classifier which extracts information from training observation location (explained further in Section 3). The training observations exhibit a high degree of spatial clustering, with most falling in coniferous forest types on mountainous slopes. However, there are some observations clustered in relatively rare but ecologically important riparian forests. Despite the fact that both land cover maps appear to be very similar, cross-validation estimates of thematic accuracy differ by more than 10% (75.5% for the covariate-only classifier and 86.5% for the spatial combination classifier). How are we to interpret and explain this difference? First of all, even in this transitional landscape where the Rocky Mountains meet the Great Plains, it is safe to assume that certain land cover types are likely to occur in close proximity to each other—coniferous forest types occur mostly

on the mountainous hillsides whereas shrub and grass types are usually found on gentler slopes and at lower elevations. Thus, we might also assume that within local clusters of training observations, spatial information has been successfully extracted, and that the accuracy of the spatial combination classifier should be greater than that of conventional classifiers, at least within these clusters. But because the training observations tend to be spatially clustered, they are not representative of the overall polygon set with respect to spatial information. Consequently, accuracy estimates based on resampling (i.e., cross-validation or bootstrapping) may be optimistically biased. We therefore contend that resampling methods are potentially misleading when used with spatial combination classifiers.

Insert Figure 2 about here

2. Terminology and Notation

Suppose that a training sample $X_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of size n has been collected by sampling a population $\mathcal{P} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ of N elements. Each element belongs to one of c classes, or groups, identified by labels $1, \ldots, c$, and n_g is the number of training observations in class g. For our purposes, $\mathcal{P}$ is a set of map polygons. The observation $\boldsymbol{x}_i \in \mathcal{P}$ is a triple $(\boldsymbol{t}_i, y_i, z_i)$ where z_i is a pair of coordinates identifying the location of the polygon centroid, y_i is the land cover class of the polygon, and $\boldsymbol{t}_i$ is a multidimensional covariate vector consisting of observations on remotely sensed and terrain variables (e.g., elevation). For all $\boldsymbol{x}_i$ in $\mathcal{P}$, $\boldsymbol{t}_i$ and z_i are known, whereas y_i is unknown except for those observations in X_n. The probability that $\boldsymbol{x}_i$ belongs to

class (or group) g, given t_i, is denoted by $P_g(\boldsymbol{t}_i) = P(y_i = g \mid \boldsymbol{t}_i)$. A classifier can be viewed as an estimator of $P_1(\boldsymbol{t}_i), \ldots, P_c(\boldsymbol{t}_i)$ which assigns $\boldsymbol{x}_i$ to the class with the largest posterior probability estimate.

The estimator of $P_g(\boldsymbol{t}_i)$ produced by a k-NN Euclidean distance classifier (Hand, 1997, Chap. 5) is the sample proportion of the k-nearest neighbors belonging to group g, where the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j \in \mathcal{P}$ is the Euclidean distance between $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$. To develop these ideas further, let $\boldsymbol{t}_{i,j}$ denote the jth closest observation to $\boldsymbol{t}_i$ among the training observation covariates $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$, and let $y_{i,j}$ denote the class label of $\boldsymbol{t}_{i,j}$. Then, the k-NN estimate of $P_g(\boldsymbol{t}_i)$ is

$$P_g^{k\text{NN}}(\boldsymbol{t}_i) = \frac{1}{k}\sum_{j=1}^{k} \Psi(y_{i,j} = g), \tag{1}$$

where $\Psi(E)$ is the indicator function of the event E. Ties among the maximum posterior probability estimates may be broken increasing the neighborhood size and recomputing formula (1).

More generally, let η denote an arbitrary classification rule obtained from X_n, and $\eta(\boldsymbol{x}_i)$ denote a prediction of y_i obtained by evaluating the classification rule for a randomly sampled polygon $\boldsymbol{x}_i \in \mathcal{P}$. For example, the k-NN classifier assignment for $\boldsymbol{x}_i$ is $\eta(\boldsymbol{x}_i) = \arg\max_g P_g^{k\text{NN}}(\boldsymbol{t}_i)$. The *conditional error rate* of η (sometimes called the true error rate) is $\varepsilon = E\Psi[\eta(\boldsymbol{x}_i) \neq y_i]$. Here, expectation is over the sampling distribution of a random observation $\boldsymbol{x}_i \in \mathcal{P}$, whereas X_n, and hence η, are fixed. The unconditional error rate, obtained as the expectation of ε over the distribution of X_n, is less often of interest in the classification arena, though it is important when the

objective is to find classifiers that perform well over a family of training samples. Ripley (1996, Chap. 2), and McLachlan (1992, Chap. 10) provide further discussion.

3. Classifiers

This section discusses the classifiers used in this study for land cover mapping. These, and other polygon-based land cover mapping problems, are somewhat unusual in two respects. Unlike most classification problems, the set to be classified, $\mathcal{P}$, is known in all respects besides class membership. This difference changes the problem of assessing classifier accuracy. Secondly, k-NN classifiers consistently have been found to be the most accurate classifiers with regard to cross-validation accuracy estimates when compared to binary tree, linear and quadratic discriminant, and logistic discriminant methods, and variants thereof (Steele and Patterson, in press). The k-NN classifiers are relatively old (two important and early papers are Fix and Hodges [1951] and Cover and Hart [1967]), and are no longer prominent in classifier research. Yet, k-NN methods performed well in this study because most of the covariates are satellite measurements of energy reflectance in specific bands of the electromagnetic spectrum (Ma *et al*, 2001), and reflectance is measured on a common scale. Moreover, the bands are of about equal value for classification. Consequently, the k-NN metric, Euclidean distance, is a very effective measure of similarity between observations compared to metrics used by other classifiers. Additionally, because the land cover class system is a human-imposed partitioning of

inter-grading vegetation communities, it is unlikely that there are many linear or distinct boundaries between classes in the covariate space. As a result, data-based methods perform better than model-based classifiers, such as linear discriminant functions, and the k-NN methods are apparently best among the distribution-free methods. In the next two sections, we present a new result regarding a weighted k-NN classifier, and a new method of extracting information from training observation location. See Steele and Patterson (in press) for additional information.

3.1 The Exact Bootstrap Aggregation k-NN Classifier

Recently, there has been a substantial amount of research in the classification arena on methods of combining ensembles of classifiers as a single rule (see Hastie *et al*, 2001 or Optiz and Maclin, 1999 for an overview). In this section, we use one of these methods, bootstrap aggregation or "bagging" (Breiman, 1996), on the k-NN classifier. The purpose of bagging a classifier is to reduce the variance of the class membership probability estimator with the hope of improving classifier performance. Bagging is carried out by drawing a bootstrap sample (a sample of size n drawn randomly with replacement) from the training sample X_n. The desired classifier is constructed from the bootstrap training sample and applied to an unclassified observation $\boldsymbol{x}_i$ with covariate vector $\boldsymbol{t}_i$ to give a vector $[P_1^{*1}(\boldsymbol{t}_i) \cdots P_c^{*1}(\boldsymbol{t}_i)]$ of class membership probability estimates. This process is repeated B times to generate B vectors $[P_1^{*1}(\boldsymbol{t}_i) \cdots P_c^{*1}(\boldsymbol{t}_i)], \ldots, [P_1^{*B}(\boldsymbol{t}_i) \cdots P_c^{*B}(\boldsymbol{t}_i)]$ of class membership probability estimates. These vectors will differ because the bootstrap training samples will vary

from iteration to iteration. The bagging estimate of the probability that $\boldsymbol{x}_i$ comes from class g is the mean

$$P_g^{\text{BA}}(\boldsymbol{t}_i) = B^{-1}\sum\nolimits_{b=1}^{B} P_g^{*b}(\boldsymbol{t}_i), \tag{2}$$

and the bagging classifier is given by $\eta^{\text{BA}}(\boldsymbol{t}_i) = \arg\max_g P_g^{\text{BA}}(\boldsymbol{t}_i)$.

Formula (2) actually is a Monte Carlo approximation to the exact bagging estimate defined by $E_{\widehat{F}} P_g^*(\boldsymbol{t}_i)$ where $\widehat{F}$ is the empirical distribution function placing probability n^{-1} at each $\boldsymbol{x}_i \in X_n$ and $P_g^*(\boldsymbol{t}_i)$ is an estimator of $P_g(\boldsymbol{t}_i)$ obtained from a random sample from $\widehat{F}$. The Monte Carlo approximation approaches the exact bagging estimate as $B \to \infty$ and is used because the true bagging estimate is almost always computationally intractable. However, with the k-NN classifier, the exact bagging estimate can be computed analytically and the Monte Carlo approximation is not necessary.

When the k-NN classifier is used with $\boldsymbol{t}_i$, then, the exact bagging estimate of the probability of membership in class g is

$$E_{\widehat{F}} P_g^{k\text{NN}}(t_i) = k^{-1}\sum_{j=1}^{k} E_{\widehat{F}}\Psi(y_{i,j} = g). \tag{3}$$

Note that $E_{\widehat{F}}\Psi(y_{i,j} = g)$ is the probability that the jth nearest neighbor of $\boldsymbol{x}_i$ in a bootstrap sample from X_n belongs to class g. This term is not difficult to calculate because there are only n possible jth closest neighbors. Specifically, let $\boldsymbol{t}_{i,j}^*$ denote the jth closest covariate to $\boldsymbol{t}_i$ among a bootstrap sample X_n^* drawn randomly and

with replacement from X_n. Let $P_{\widehat{F}}(\boldsymbol{t}^*_{i,j} = \boldsymbol{t}_{i,h})$ denote the probability that $\boldsymbol{t}_{i,h}$, the h-closest covariate vector to $\boldsymbol{t}_i$ among the training sample covariates, is the jth closest among X^*_n. Then,

$$E_{\widehat{F}}\Psi(y_{i,j} = g) = \sum_{h=1}^{n} P_{\widehat{F}}(\boldsymbol{t}^*_{i,j} = \boldsymbol{t}_{i,h})\Psi(y_{i,h} = g),$$

and

$$E_{\widehat{F}}P_g^{k\mathrm{NN}}(\boldsymbol{t}_i) = \frac{1}{k}\sum_{j=1}^{k}\sum_{h=1}^{n} P_{\widehat{F}}(\boldsymbol{t}^*_{i,j} = \boldsymbol{t}_{i,h})\Psi(y_{i,h} = g).$$

To calculate $P_{\widehat{F}}(\boldsymbol{t}^*_{i,j} = \boldsymbol{t}_{i,h})$, note that $\boldsymbol{t}^*_{i,j} = \boldsymbol{t}_{i,h}$ will occur if $T^*_n = \{\boldsymbol{t}^*_{i,1}, \ldots, \boldsymbol{t}^*_{i,n}\}$ contains at most $j-1$ copies of $\boldsymbol{t}_{i,1}, \ldots, \boldsymbol{t}_{i,h-1}$ and at least j copies of $\boldsymbol{t}_{i,1}, \ldots, \boldsymbol{t}_{i,h}$. If a ranges over the number of copies of $\boldsymbol{t}_{i,1}, \ldots, \boldsymbol{t}_{i,h-1}$ and is constrained to $\{0, \ldots, j-1\}$, and b ranges over the number of copies of $\boldsymbol{t}_{i,h}$ while constrained $\{j - a, \ldots, n-a\}$, then there must be $n-a-b$ copies of $\boldsymbol{t}_{i,h+1}, \ldots, \boldsymbol{t}_{i,n}$ contained in T^*_n. Hence,

$$\begin{aligned} P_{\widehat{F}}(\boldsymbol{t}^*_{i,j} = \boldsymbol{t}_{i,h}) &= \sum_{a=0}^{j-1}\sum_{b=j-a}^{n-a} \binom{n}{a}\left(\frac{h-1}{n}\right)^a \binom{n-a}{b}\left(\frac{1}{n}\right)^b \left(\frac{n-h}{n}\right)^{n-a-b} \\ &= \frac{1}{n^n}\sum_{a=0}^{j-1}\sum_{b=j-a}^{n-a} \binom{n}{a}(h-1)^a \binom{n-a}{b}(n-h)^{n-a-b}. \end{aligned}$$

The exact bootstrap aggregation k-NN classifier was previously developed by Steele and Patterson (2000) as a smoothed version of the ordinary k-NN classifier. In that article, the discrete outcomes $\Psi(y_{i,j} = g)$ in formula (1) were replaced with

smoothed versions given by the resampling expectations $E_{\widehat{F}}\Psi(y_{i,j}=g)$. Henceforth, the exact bootstrap aggregation k-NN classifier is denoted as EB k-NN.

3.2. Spatial Classifiers

Spatial information is present and potentially useful for classification when land cover class abundance varies with such factors as climate, topography, and soil class (Carpenter *et al*, 1999; Steele, 2000; Steele and Redmond, 2001). Contextual allocation methods (Griffith and Fellows, 1999; Kartikeyan *et al*, 1994; Lee, 2000; Stuckens *et al*, 2000) are useful for extracting spatial information when there are consistent patterns of positive spatial association among adjacent map units, as is often the case when the map is a lattice of pixels. Such methods are not appropriate when the map consists of polygons formed by segmentation of a pixel image (Ma *et al*, 2001). The reason for this is that image segmentation tends to produce polygon boundaries that coincide with changes in spectral reflectance in land cover class. Consequently, adjacent polygons are not predictably similar, and spatial association among adjacent polygons is weak or absent.

However, there often are patterns in the distribution of land cover class over areas encompassing more than a few polygons. For example, orographic effects may induce differences in the distribution of land cover class between lee and windward sides of a mountain range. If the training sample is representative of the spatial distribution of land cover classes, then these patterns will also exist within the training sample. Steele (2000), Steele and Patterson (in press), and Steele and Redmond

(2001) have developed an approach to extracting spatial information from the relative abundance and proximity of training observations in the vicinity of an unclassified polygon. Their approach is motivated by the application of Bayes rule when the conditional probability density functions $p(\boldsymbol{t}_i \mid y_i = g)$ and the prior probabilities $\pi_g = P(y_i = g)$, $g = 1, \ldots, c$, are known. The Bayes rule minimizes the total probability of misclassification by assigning $\boldsymbol{x}_i$ to the class with the largest posterior probability

$$P(y_i = g \mid \boldsymbol{t}_i) = \frac{\pi_g p(\boldsymbol{t}_i \mid y_i = g)}{\sum_{j=1}^{c} \pi_j p(\boldsymbol{t}_i \mid y_i = j)}$$

(McLachlan 1992, Chap. 1). The k-NN classifier (formula [1]) is a plug-in version of Bayes rule which assumes non-informative priors, that is, $\pi_g = c^{-1}$ for $j = 1, \ldots, g$. Consequently, the k-NN posterior probability estimator is $P^{k\text{NN}}(y_i = g \mid \boldsymbol{t}_i) = P_g^{k\text{NN}}(\boldsymbol{t}_i)$. The assumption of non-informative prior probabilities is appropriate if there is no information regarding the class membership probabilities besides that carried by the covariate vectors.

In some instances, the relative frequency of occurrence of the c classes in a spatial neighborhood about the location z_i of $\boldsymbol{x}_i$, is informative for classification; if so, then this information may be expressed as local prior class probabilities $\pi_g(z_i)$, $g = 1, \ldots, c$. We propose a local Bayes rule which assigns $\boldsymbol{x}_i$ to the class with the largest posterior probability

$$P(y_i = g \mid \boldsymbol{t}_i, z_i) = \frac{\pi_g(z_i)\, p(\boldsymbol{t}_i \mid y_i = g)}{\sum_{j=1}^{c} \pi_j(z_i)\, p(\boldsymbol{t}_i \mid y_i = j)}. \qquad (4)$$

This Bayes rule minimizes the probability of misclassifying a random observation given the local prior probabilities (McLachlan 1992, Chap. 1). In practice, $\pi_g(z_i)$ and $p(\boldsymbol{t}_i \mid y_i = g)$ are rarely known. The same difficulty arises in the conventional classification problem, and a common approach, which we also use, is to plug in estimates of the prior and conditional probability densities into formula (4). In this study, the local class priors are estimated using Steele and Patterson's (in press) mean inverse distance (MID) estimator. We defined the mean inverse distance from an observation with location z_i to class g to be

$$\overline{d}_g(z_i) = \frac{1}{n_g} \sum_{k=1}^{n} \Psi(y_k = g)\, d_i^E(z_k)^{-2}, \tag{5}$$

where $d_i^E(z_k)$ is the Euclidean distance between locations z_i and z_k. The MID classifier estimator of $\pi_g(z_i)$ is the normalized mean inverse distance to class g given by

$$\widehat{\pi}_g(z_i) = \frac{\overline{d}_g(z_i)}{\sum_{j=1}^{c} \overline{d}_j(z_i)}.$$

The MID estimates may be combined with any conventional classifier. For example, the EB k-NN+MID estimator of $P(y_i = g \mid \boldsymbol{t}_i, z_i)$ is

$$P^{\text{EB } k\text{NN+MID}}(\boldsymbol{x}_i) = \frac{\widehat{\pi}_g(z_i) E_{\widehat{F}} P_g^{k\text{NN}}(\boldsymbol{t}_i)}{\sum_{j=1}^{c} \widehat{\pi}_j(z_i) E_{\widehat{F}} P_j^{k\text{NN}}(\boldsymbol{t}_i)},$$

and the k-NN+MID classifier is $\eta^{\text{EB } k\text{NN+MID}}(\boldsymbol{x}_i) = \arg\max_g P^{\text{EB } k\text{NN+MID}}(\boldsymbol{x}_i)$.

Several remarks are in order. The idea of combining information from different sources (e.g., terrain and remotely sensed variables) using prior and posterior probability estimates is not new; for an example, see McIver and Friedl (2002). The mean inverse distance measure of spatial density implies that if the training observations are regarded as lights of equal intensity, then $\bar{d}_g(z_i)$ is the average illumination generated by the lights of class g at z_i. The use of the mean inverse distance as a measure of spatial density also is not new; for example, see Watson and Philip (1985). Steele and Redmond (2001) proposed a different method of estimating the $\pi_g(z_i)$'s by using the ranks of the distances between z_i and the nearest training observation in class g. Indicator kriging (Cressie 1993, Chap. 5) is a possible alternative to mean inverse distance weighting for estimating local prior probabilities. Because spatial classifiers extract spatial information from the training observations, we expect the accuracy of spatial classifiers to be positively associated with training observation density, given that all other factors are constant. This implies that resampling methods are at risk of producing optimistically biased accuracy estimates if training observations are spatially clustered.

4. Accuracy Estimation

Map accuracy may be quantified as proportion of the map that has been correctly classified. If the map is represented by a set $\mathcal{P}$ of N units, then map accuracy is

$$\alpha = \sum_{i=1}^{N} f_i \Psi[\eta(\boldsymbol{x}_i) = y_i], \tag{6}$$

where f_i is the fraction of the map area occupied by $\boldsymbol{x}_i \in \mathcal{P}$. When the objective is to estimate α, inference usually is conducted from a finite population standpoint using a post-classification sample. Inferential methods for this problem are discussed by Congalton (1991), Stehman (1997a,b) and Stehman and Czaplewski (1998). When the objective is to assess the performance of η in a more general context, a super-population view is adopted (see Stehman, 2000 for discussion of the distinctions). For example, it may be useful to view $\mathcal{P}$ as a realization of a random process governed by factors such as temporal changes in satellite imagery and plant phenology. In this context, accuracy assessment usually is conducted by applying resampling methods to the training sample X_n (see Hand, 1997, Chap. 7; McLachlan, 1992, Chap. 9; Schavio and Hand, 2000). A limitation of this approach is that X_n must be collected by probability sampling; even so, accuracy assessment methods should guard against bias induced by using X_n for both rule construction and accuracy assessment.

A widely-used resampling method is n-fold cross-validation (Efron and Tibshirani, 1993, Chap. 17; McLachlan, 1992, Chap. 9). This method sequentially removes $\boldsymbol{x}_j$, $j = 1, \ldots, n$, from X_n, and constructs the jth holdout rule η_j from the remaining $n - 1$ observations. Then, $\boldsymbol{x}_j$ is classified using η_j and the outcome $\Psi[\eta_j(\boldsymbol{x}_j) = y_j]$ is recorded. The n-fold cross-validation accuracy estimator is the sample mean of the outcomes; we denote this estimator by

$$\widehat{\alpha}^{CV} = n^{-1} \sum_{j=1}^{n} \Psi[\eta_j(\boldsymbol{x}_j) = y_j]. \tag{7}$$

When X_n is obtained by random sampling, then $\widehat{\alpha}^{CV}$ is nearly unbiased (Krzanowski,

2001). In this study, the training sets are not probability samples; moreover, the training observations exhibit varying degrees of spatial clustering and, hence, may produce incorrect accuracy estimates if resampling methods are used.

Lacking post-classification test samples and probabilistically sampled training sets, we are forced to turn elsewhere for accuracy estimates. We propose to estimate the accuracy of the class predictions for each polygon $\boldsymbol{x}_i \in \mathcal{P}$, and combine these estimates to estimate map accuracy. While the methods of estimating the probability of correct classification are not new, the application of these methods to the problem of assessing map accuracy is novel.

4.1 Estimating the Probability of Correct Classification for Individual Map Units

Herein, we adopt a super-population view, and regard $\mathcal{P}$ as a single realization of a random process. The estimand of interest is no longer α (formula 6), but rather the area-weighted mean conditional probability of correct classification

$$\alpha^P = \sum_{i=1}^{N} f_i P[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i]. \tag{8}$$

From the super-population view, α^P is interpreted as the probability that a map unit selected at random from the super-population will be correctly classified by the rule η derived from X_n. The parameters α and α^P are closely related because $\mathcal{P}$ is regarded as a random observation from the super-population; hence, $E(\alpha) = \alpha^P$.

To develop the method, we first suppose that η is the Bayes rule. Then, $\eta(\boldsymbol{x}_i) = \arg\max_g P(y_i = g \mid \boldsymbol{x}_i)$, where $P(y_i = g \mid \boldsymbol{x}_i)$ is the posterior probability of membership in class g. Moreover, the probability that $\boldsymbol{x}_i$ is correctly classified is the maximum posterior probability; i.e.,

$$P[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i] = \max_g P(y_i = g \mid \boldsymbol{x}_i) \tag{9}$$

(Ripley 1996, Chap. 2). To estimate $P[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i]$, we substitute estimates $\widehat{P}(y_i = g \mid \boldsymbol{x}_i)$ of $P(y_i = g \mid \boldsymbol{x}_i)$ in place of the true values in formula (9). For example, if η is a k-NN rule, then $\widehat{P}(y_i = g \mid \boldsymbol{x}_i) = P_g^{k\text{NN}}(\boldsymbol{t}_i)$, and the estimated probability that $\boldsymbol{x}_i$ is correctly classified is $\widehat{P}[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i] = \max_g P_g^{k\text{NN}}(\boldsymbol{t}_i)$. The use of these plug-in posterior probabilities for accuracy estimation is discussed by Basford and McLachlan (1985) and Ripley (1996, Chap. 2). Related topics are discussed by Ganesalingam and McLachlan (1980), Glick (1978), Hand (1986), McLachlan (1992, Chap. 10), Toussaint (1974) and Weiss (1991). Finally, the super-population view of map accuracy estimation recognizes α^P as an optimal estimator of α because $P[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i]$, the conditional expectation of $\Psi_i = \Psi[\eta(\boldsymbol{x}_i) = y_i]$, minimizes the expected mean square error $E(\Psi_i - \widehat{\Psi}_i)^2$.

It is important to recognize that the plug-in estimator $\max_g \widehat{P}(y_j = g \mid \boldsymbol{x}_j)$ may be severely biased when $\boldsymbol{x}_j$ belongs to the training set used to construct η. However, according to Ripley (1996, Chap. 2), the bias is expected to be small when applied to an observation that was not used to construct η. To address the risk of optimistic bias

when estimating α^P, we discuss two methods of calibrating $\widehat{P}[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i]$ in the next section.

4.2 Calibrating the Estimated Probabilities of Correct Classification

Dawid (1982), Cox (1958), and Ripley (1996, Chap. 2) discuss calibration methods for reducing bias of estimators of the probability of an event. We propose to use the training set to assess the effect of calibration on the estimator of α^P. This assessment is carried out by regressing the binary leave-one-out outcomes $\Psi[\eta_j(\boldsymbol{x}_j) = y_j]$, $j = 1, \ldots, n$, on the leave-one-out probability estimates of correct classification $\widehat{P}[\eta_j(\boldsymbol{x}_j) = y_j \mid \boldsymbol{x}_j]$. The usefulness of such comparisons depends on the degree of correspondence between X_n and $\mathcal{P}$ with respect to calibration. If X_n were a random sample from $\mathcal{P}$, then the two methods should yield nearly unbiased estimates. However, in this study, this is not the case, and there is a risk that a calibration function derived from X_n may not be effective at calibrating the correct classification probabilities for $\mathcal{P}$. With this caveat in mind, though, it is illustrative to compare accuracy estimates derived from the calibrated and uncalibrated probabilities of correct classification, and ordinary cross-validation.

Linear and logistic regression were used to derive calibration functions from the training set. To set up the calibration functions, let $\widehat{p}_j = \widehat{P}[\eta_j(\boldsymbol{x}_j) = y_j \mid \boldsymbol{x}_j]$ denote the estimated probability that $\boldsymbol{x}_j \in X_n$ is correctly classified by the holdout rule η_j, and $O_j = \Psi[\eta_j(\boldsymbol{x}_j) = y_j]$, $j = 1, \ldots, n$, denote the outcome of classifying $\boldsymbol{x}_j$ by η_j. For $\boldsymbol{x}_i \in \mathcal{P}$, the linear calibration function specifies that the calibrated estimate of

$P[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i]$ is

$$\widehat{p}_i^{lin} = \begin{cases} \widehat{\beta}\widehat{p}_i, & \text{if } \widehat{\beta}\widehat{p}_i < 1 \\ 1, & \text{if } \widehat{\beta}\widehat{p}_i \geq 1. \end{cases}$$

The coefficient $\widehat{\beta}$ is determined by minimizing $\sum(O_j - \beta\widehat{p}_j)^2$ with respect to β; hence, $\widehat{\beta} = \sum O_j \widehat{p}_j / \sum \widehat{p}_j^2$. Cox (1958) proposed to calibrate probability estimates using logistic regression. Logistic regression is justified under the assumption that O_j, $j = 1, \ldots, n$, are independent Bernoulli random variables with expectations $P[\eta_j(\boldsymbol{x}_j) = y_j \mid \boldsymbol{x}_j]$. Then, the logistic calibration model is $P[\eta_j(\boldsymbol{x}_j) = y_j \mid \boldsymbol{x}_j] / P[\eta_j(\boldsymbol{x}_j) \neq y_j \mid \boldsymbol{x}_j] = [\widehat{p}_j / (1 - \widehat{p}_j)]^{\gamma}$, and the calibrated estimate of $P[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i]$ is $\widehat{p}_i^{log} = \left\{1 + [(1 - \widehat{p}_i)/\widehat{p}_i]^{\widehat{\gamma}}\right\}^{-1}$, where the calibration coefficient $\widehat{\gamma}$ is computed by logistic regression.

4.3 Assessing the Effect of Calibration

Lacking post-classification test samples, we use the training sets to assess the effect of calibration. As remarked earlier, n-fold cross-validation yields nearly unbiased accuracy estimators when X_n is a random sample from $\mathcal{P}$. Therefore, a comparison of accuracy estimates derived from cross-validation and estimated probabilities of correct classification provides information regarding the need for, and effectiveness of, calibration. In this study, though, formal inferences cannot be drawn regarding map accuracy and calibration because the training sets are not probability samples of the polygon sets. Therefore, the training sample n-fold cross-validation accuracy estimate, $\widehat{\alpha}^{CV}$, is viewed as a fixed quantity, and the training sample means of the uncalibrated

and calibrated estimated probabilities of correct classification are viewed as approximations to $\widehat{\alpha}^{CV}$. Comparisons of $\widehat{\alpha}^{CV}$ and the uncalibrated, linear, and logistic calibrated approximations provide three mean differences and mean squared differences from each training set.

The area-weighted mean difference between the uncalibrated correct classification probability estimates and the n-fold cross-validation estimates is $D_u = \sum_{g=1}^{c} \widehat{f}_g \sum_{j=1}^{n} \Psi[\eta_j(\boldsymbol{x}_j) = g](O_j - \widehat{p}_j)$, where $\widehat{f}_g = \sum_{i=1}^{N} f_i \Psi[\eta(\boldsymbol{x}_i) = g]$ is the fraction of the entire map area assigned to class g by η. The area-weighted mean square difference, S_u, is computed similarly except that the differences $O_j - \widehat{p}_j$, $j = 1, \ldots, n$, are replaced by squared differences. Mean differences and mean squared differences between the calibrated correct classification probability estimates and the n-fold cross-validation estimates are computed in the same manner as for the uncalibrated estimates except that $\widehat{p}_i$ is replaced by $\widehat{p}_i^{lin}$ and $\widehat{p}_i^{log}$. The mean differences and root squared mean differences are denoted by D_{lin}, D_{log}, S_{lin} and S_{log}.

4.4 Class and Overall Accuracy Estimates

In this section, we present estimators of α^P, the area-weighted mean conditional probability of correct classification [formula (8)], and class-specific accuracy. The estimators of α^P are plug-in versions of formula (8); for example, if $P[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i]$ is estimated by the uncalibrated estimate $\widehat{p}_i$, $i = 1, \ldots, N$, then the estimator is $\widehat{\alpha}^P = \sum_{i=1}^{N} f_i \widehat{p}_i$. For our purposes, class-specific accuracy is

defined as $P[y_i = g \mid \eta(\boldsymbol{x}_i) = g]$, the conditional probability that an observation predicted to belong to class g is actually a member of class g. A popular alternative measure is $P[\eta(\boldsymbol{x}_i) = g \mid y_i = g]$, the conditional probability that an observation belonging to class g is predicted to belong to class g. In the remote sensing literature, these probabilities are commonly described as users' and producers' accuracy, respectively. The remainder of this section discusses several estimators useful for assessing map accuracy.

We propose to estimate users' accuracy for class g by the area-weighted mean accuracy of those polygons predicted to belong to class g. An estimator computed from the uncalibrated estimates of $P[\eta(\boldsymbol{x}_i) = y_i \mid \boldsymbol{x}_i]$ is

$$\widehat{\alpha}_g^P = \sum_{i=1}^{N} f_i \widehat{p}_i \Psi[\eta(\boldsymbol{x}_i) = g] \Big/ \sum_{i=1}^{N} f_i \Psi[\eta(\boldsymbol{x}_i) = g].$$

Linear and logistic calibrated versions of $\widehat{\alpha}_g^P$ are obtained by replacing $\widehat{p}_i$ by $\widehat{p}_i^{lin}$ or $\widehat{p}_i^{log}$, respectively. A slightly different estimator of $P[y_i = g \mid \eta(\boldsymbol{x}_i) = g]$ was proposed by Basford and McLachlan (1985) for assessing the results of cluster analysis.

Often, map accuracy assessment utilizes a users' confusion matrix. These $c \times c$ matrices contain estimates of the conditional probabilities $P[y_i = g \mid \eta(\boldsymbol{x}_i) = h]$, for $1 \leq g, h \leq c$ (Congalton, 1992). Confusion matrices are commonly used for 1) evaluating training data and quickly identifying confusion among observations, and 2) understanding why certain types are not classifying well, and related to this, helping to decide whether to collapse classes or use fuzzy sets (e.g., Gopal and Woodcock,

1994; Foody, 1996). A users' confusion matrix can be computed without recourse to resampling methods by estimating the area-weighted mean probability of membership in class g across all polygons predicted to belong to class h. This calculation yields

$$\widehat{P}[y_i = g \mid \eta(\boldsymbol{x}_i) = h\,] = \frac{\sum_{i=1}^{N} f_i \widehat{P}[y_i = g \mid \boldsymbol{x}_i]\Psi[\eta(\boldsymbol{x}_i) = h\,]}{\sum_{i=1}^{N} f_i \Psi[\eta(\boldsymbol{x}_i) = h\,]}, \qquad (10)$$

where $\widehat{P}[y_i = g \mid \boldsymbol{x}_i]$ is the estimated probability of membership in class g obtained from η. Unfortunately, there is no obvious simple way to calibrate misclassification probability estimates.

In this study, it is helpful to estimate accuracy for two important categories of land cover classes, forested and nonforested (primarily xeric grassland and shrubland land cover classes). The remainder of the land cover classes are categorized as nonvegetated (primarily rock and barren soil land cover classes). The users' accuracy estimate for the nonforest category is

$$\widehat{\alpha}_{NF}^{P} = \sum_{i=1}^{N} f_i \widehat{p}_i \Psi[\eta(\boldsymbol{x}_i) \in I_{NF}] \Big/ \sum_{i=1}^{N} f_i \Psi[\eta(\boldsymbol{x}_i) \in I_{NF}]$$

where I_{NF} is the index set of nonforested land cover classes. Estimates for forested and nonvegetated categories are analogously defined.

Finally, from the training sample, the n-fold cross-validation estimate of users' accuracy for class g is

$$\widehat{\alpha}_g^{CV} = \sum_{j=1}^{n} \Psi[\eta_j(\boldsymbol{x}_j) = y_j]\Psi[y_j = g] \Big/ \sum_{j=1}^{n} \Psi[\eta_j(\boldsymbol{x}_j) = y_j].$$

The cross-validation estimate of map accuracy is a weighted average of the class-specific accuracy estimates $\widehat{\alpha}^{CV} = \sum \widehat{f}_g \alpha_g^{CV}$.

4.5 Accuracy Maps

Estimates of α^P and α_g^P, $g = 1, \ldots, c$, provide useful summaries of overall and class-specific accuracy rates. Yet, there may be substantial spatial variation in accuracy across a given map (Campbell 1981; Congalton 1988) and these statistics contain no spatial information. Recently, methods for estimating and mapping local accuracy rates have been developed by Steele *et al* (1998) and Kyriakidis and Dungan (2001). Steele *et al* (1998) adopt a super-population view and estimate the probability that the predicted class is correct at a training sample location by bootstrap resampling. These estimates are interpolated from the training observation locations to a superimposed lattice via kriging, and a map is constructed from the lattice estimates. A limitation of this method is that interpolated estimates may be poor in regions with too few training observations. Kyriakidis and Dungan (2001) use a post-classification test sample $U_m = \{u_1, \ldots, u_m\}$ and kriging to combine class-specific accuracy estimates with the outcomes $\Psi[\eta(u_i) = y_i]$, $i = 1, \ldots, m$. On the one hand, Kyriakidis and Dungan's (2001) method provides additional information in sample-deficient areas by using global summaries. However, these global estimates may be inaccurate in some localities.

The estimated probabilities of correct classification for individual polygons provide a simple alternative method of accuracy mapping. Specifically, estimated accuracy is represented by a gray-scale map constructed from the estimates $\widehat{p}_i$, $i = 1, \ldots, N$ (or calibrated versions, thereof). These estimates provide a finer scale of resolution than those provided by the methods of Steele *et al* (1998) or Kyriakidis and Dungan (2001), and with very little computational effort (see Figure 3 for an example).

Insert Figure 3 about here

5. Example

To further illustrate our new method of accuracy assessment, we present its application to a broad-scale land cover mapping project mentioned previously in Section 1 and shown in Figure 1. The study area was topographically diverse and included some very rugged, mountainous terrain along and adjacent to the North American Continental Divide. Coniferous forests predominated on the mountain slopes, at least below timberline, whereas the lower elevation valleys and plains were predominately shrub and grass rangelands, interspersed with agriculture. More details about the study area and its vegetation can be found in Arno (1979).

The underlying digital database was constructed from the classification of nine Landsat Thematic Mapper (TM) scenes, each of which was classified separately and in general accordance with methods described by Ma *et al* (2001) and Steele and Redmond (2001). The unsupervised classification and segmentation of these nine

images produced nearly 4.27 million unique map units (unduplicated after edge-matching); and these ranged in size from a single 30 m^2 pixel to patches up to 202 ha.

Training data sets for the supervised classification of each TM scene were assembled from a variety of existing sources; most of the data came from the USDA Forest Service Timber Stand and Management Record System (TSMRS), its Forest Inventory and Analysis program (FIA), and the Natural Resource Conservation Service Natural Resource Inventory program (NRI). Observations were screened for positional and covariate errors by plotting position on a false-color composite of Landsat TM channels 4, 5, and 3 (R,G,B). Observations with recorded covariate and land cover class labels that were grossly inconsistent with the false-color composite image were removed from the training set. Despite the fact that some of these existing data were collected according to a probability sampling plan (e.g., the FIA and NRI data), the spatial distribution of training observations over each scene was irregular and tended to correspond with patterns of land ownership (e.g., relatively few data were available for lands in private ownership).

5.1 TM Scene Classifications

In the analyses presented below, we used all available covariates for classification and three classifiers: the EB 10-NN, the MID spatial, and EB 10-NN+MID rule. Covariates were spectral reflectance intensity for TM channels 1-5 and 7, MNDVI (Nemani, et al. 1993), scaled elevation, slope, and a measure of solar insolation $s \times \cos(2\pi[(a+135)/360])$, where a is aspect (degrees) and s is percent slope.

Table 1 summarizes the map accuracy results obtained from the linear calibrated estimates of the probability of correct classification. The estimates for the nine Landsat scenes ranged from 64.5 to 80.3%, with five of the nine estimates falling between 72 and 76%. Estimates of mapaccuracy for regions classified as nonforest varied between 67.1 and 83.8%, whereas the accuracy estimates ranged from 58.1 to 75.4% for forested regions (not shown in Table 1).

Insert Table 1 about here

Information regarding the usefulness of spatial information for classification is expressed by differences in estimated map accuracy between the EB 10-NN+MID and the EB 10-NN classifier. Table 1 shows that these differences varied between 0.3 and 4.5%. Alone, the MID spatial classifier was a poor classifier, with accuracy estimates ranging from 19.7 to 34.9% (Table 1). These estimates reflect the sensitivity of MID spatial classifier to sample size; in fact, Pearson's correlation between map accuracy estimate and sample size is $r = .71$ $(n = 9)$. With respect to the EB 10-NN+MID classifier, n-fold cross-validation map accuracy estimates (not shown) tended to be substantially larger (78.0 to 90.4%) than those obtained from linear calibrated estimates of the probability of correct classification (Table 1). It appears that cross-validation tends to exaggerate the value of spatial information; specifically, differences between the EB 10-NN+MID and the EB 10-NN cross-validation accuracy estimates ranged from 6.1 to 11.1%, compared to a range of differences of 0.3 to 4.5% obtained from the linear calibrated estimates of the probability of correct classification.

The accuracy of our proposed method depends on the bias of the estimators of the probability of correct classification. Unfortunately, we cannot assess bias without a probability training or post-classification sample. That said, each training set is largely a collection of probability samples drawn from subregions within the larger TM scene. The training samples differ from the polygon set in one other obvious way: the training observations tend to be spatially clustered. Of course, there may be other, less obvious differences. If these differences are unimportant with respect to difficulty of classification, then map accuracy estimates for a covariate-only classifier (specifically, EB 10-NN) should be approximately equal when the estimates are computed by n-fold cross-validation as when computed from map unit estimates of the probability of correct classification. These estimates, for nonforest and forest lifeforms, by TM scene, are graphed in Figure 4. The n-fold cross-validation estimates are slightly but consistently larger than the estimates obtained from the linear calibrated probability of correct classification. In particular, the sample mean of the nine differences for all land cover classes (i.e., $\widehat{\alpha}^{CV} - \widehat{\alpha}^{P}$) was 2.67% (SE $=0.99\%$). This result suggests that $\widehat{\alpha}^{P}$ is not optimistically biased. Based on earlier work on posterior probabilities of correct classification (see Ripley, 1996, Chap. 2), it seems unlikely that $\widehat{\alpha}^{P}$ is pessimistically biased. We suspect instead that there is a small amount of optimistic bias in $\widehat{\alpha}^{CV}$ originating from differences between training and polygon observations.

Insert Figure 4 about here

The results for a single TM scene, P39/R28, are broken down further in Table 2 to examine and illustrate the effects of calibration on the different land cover classes

and lifeforms. Table 2 shows seven sets of accuracy rates for the EB 10-NN+MID classifier. Values in the leftmost column were obtained by computing correct classification probability estimates using the polygon set; these were then calibrated using linear and logistic calibration functions derived from the training set. Also shown are estimates obtained from the training set by computing uncalibrated and calibrated correct classification probability estimates. Although there was relatively little difference between the calibrated and uncalibrated estimates within the polygon estimates, and similarly within training sample estimates, the differences between the same estimate derived from the two different sets were much larger (Table 2). For example, the cross-validation estimate of map accuracy was $\widehat{\alpha}^{CV} = 86.5\%$ whereas the polygon set of linear calibrated correct classification probability estimates yielded the estimate $\widehat{\alpha}^{P} = 76.0\%$. This pattern of differences was, with a few exceptions, relatively consistent across land cover type and lifeform. For brevity, Table 3 shows only two columns of the users' confusion matrices computed from the polygon class membership estimates (equation 10), and by cross-validation. The estimates show a similar pattern of confusion, though the polygon estimates of the probabilities of misclassification tend to be larger than the cross-validation estimates. Finally, Figure 3 shows a portion of the accuracy map for TM Scene P39/R28 corresponding to Figure 2.

Insert Table 2 about here

Insert Table 3 about here

5.2 Assessment of Calibration

The effect of calibration on map accuracy estimates was investigated using all nine training sets. Figure 5 shows mean differences (D_u) between the map accuracy estimates computed by n-fold cross-validation ($\widehat{\alpha}^{CV}$) and estimates computed from the uncalibrated correct classification probability estimates, by TM scene. Also shown are the mean differences between the map accuracy estimates derived from cross-validation and the linear and logistic calibrated correct classification probability estimates (D_{lin} and D_{log}). The uncalibrated differences D_u are consistently larger than D_{lin}, and D_{log}; specifically, the averages of D_u, D_{lin} and D_{log} are 4.54, 0.88 and 0.82%, respectively. Figure 6 shows the root mean square differences S_u, S_{lin} and S_{log}, by TM scene. The averages of S_u, S_{lin} and S_{log} are 7.29, 5.79 and 5.77%, respectively. These results, showing relatively small differences between map accuracy estimates computed from n-fold cross-validation and the training sample-derived probability of correct classification estimates, are anecdotal evidence that the polygon-based probability of correct classification estimates are useful for accuracy assessment.

Insert Figures 5 and 6 about here

6 Discussion

In this article, we have proposed a new approach to assessing map accuracy that focuses on estimating the probability of correct classification for each map unit. These probabilities are then calibrated, if necessary, and averaged to obtain map accuracy

estimates. The effectiveness of this approach depends on the estimators of the correct classification probabilities being nearly unbiased. We described methods of calibrating the estimated probabilities of correct classification, and a method of assessing the effectiveness of calibration. Unfortunately, our examples must be viewed as anecdotal because calibration could not be formally evaluated without a probability sample.

In light of these comments, what is the value of the proposed method? For the example presented herein, lacking probability training and test samples, we argued that the map accuracy estimates derived from the probability of correct classification estimates are better than those derived by resampling the training set because spatial clustering of the training observations introduced a substantial amount of error into the estimates. This source of error was avoided by using the polygons rather than the training observations for estimating accuracy. In addition, by estimating the probability of correct classification for each polygon, a new, simple, and informative method of producing accuracy maps is immediately available. Lastly, and most importantly, the proposed method may be more efficient than conventional post-classification assessment programs if a probabilistically sampled training set, or even a small post-classification test set is used for calibration. Further investigation of this method, and others aimed at improving map accuracy estimates, are urgently needed given the rapidity at which broad-scale digital land cover maps are being adopted for regional ecosystem management.

Acknowledgments

This research was funded in large part by the U.S.D.A. Forest Service (contracts PNW97-0511-2 & 53-0343-0-0010). We are particularly grateful to Ken Brewer, Ron Brohman, Dwight Chambers, Jeff DiBenedetto, Mark Jensen, and Judy Tripp for coordinating various aspects of the project for the Forest Service. The image classifications were carried out by Chip Fisher, Chris Winne, and Gary Gooch at the Wildlife Spatial Analysis Lab. We also thank Jim Schumacher for preparing the first three figures and Chip Fisher for his patience with many data transfers. We thank two anonymous reviewers for helpful comments.

References

Arno, S.F. (1979) Forest regions of Montana. USDA Forest Service Research Paper INT-218, Ogden, UT.

Basford, K.E. and McLachlan, G.J. (1985) Estimation of allocation rates in a cluster analysis context. *Journal of the American Statistical Association*, **80**, 286-93.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **26**, 123-40.

Campbell, J. (1981) Spatial autocorrelation effects upon the accuracy of supervised classification of land cover. *Photogrammetric Engineering and Remote Sensing*, **47**, 355-63.

Carpenter, G.A., Gopal, S., Macomber, S., Martens, S., Woodcock, C.E., and Franklin, J. (1999) A neural network method for efficient vegetation mapping. *Remote Sensing of Environment*, **70**, 326-38.

Congalton, R.G. (1988) Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, **54**, 587-92.

Congalton, R.G. (1991) A review of assessing the accuracy of classification of remotely sensed data. *Remote Sensing of Environment*, **37**, 35-46.

Cover, T.M. and Hart, P.E. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**, 21-7.

Cox, D.R. (1958) Two further applications of a model for binary regression. *Biometrika*, **45**, 562-65.

Cressie, N.A.C. (1993) Statistics for Spatial Data. Wiley, New York.

Dawid, A.P. (1982) The well-calibrated Bayesian (with discussion). *Journal of the American Statistical Association*, **77**, 605-13.

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Efron, B. and Tibshirani, R. (1997) Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, **92**, 548-60.

Fisher, P.F. (1994) Visualization of the reliability in classified remotely sensed images. *Photogrammetric Engineering and Remote Sensing*, **60**, 905-10.

Fix, E. and Hodges, J.L. (1951) Discriminatory analysis—nonparametric discrimination: consistency properties. Report no. 4, US Air Force School of Aviation Medicine, Random Field, TX.

Foody, G.M. (1996) Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. *International Journal of Remote Sensing*, **17**, 1317-40.

Ganesalingam, S. and McLachlan. G.J. (1980) Error rate estimation on the basis of posterior probabilities. *Pattern Recognition*, **12**, 405-13.

Glick, N. (1978) Additive estimators for probabilities of correct classification. *Pattern Recognition*, **10**, 211-22.

Gopal, S., and Woodcock, C. (1994) Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, **60**, 181-8.

Griffith, D.A. and Fellows, P.L. (1999) Pixels and eigenvectors: classification of Landsat TM imagery using spectral and locational information. In *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, K. Lowell and A. Jaton (eds), Ann Arbor Press, Chelsea MI.

Hammond, T.O. and Verbyla, D.L. (1996) Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing*, **17**, 1261-66.

Hand, D.J. (1986) Recent advances in error rate estimation. *Pattern Recognition Letters*, **4**, 335-46.

Hand, D.J. (1997) *Construction and Assessment of Classification Rules.* Wiley, New York.

Hastie, T., Tibshirani, T. and Freidman, J. (2001) *The Elements of Statistical Learning*. Springer, New York.

Kartikeyan, B., Gopalakrishna, B., Kalaburme, M.H., and Majumdar, K.L. (1994) Contextual techniques for classification of high and low resolution remote sensing data. *International Journal of Remote Sensing*, **15**, 1037-51.

Krzanowski, W.J. (2001) Data-based interval estimation of classification error rates. *Journal of Applied Statistics*, **5**, 585-95.

Kyriakidis, P.C. and Dungan, J.L. (2001) A geostatistical approach for mapping thematic classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions. *Environmental and Ecological Statistics*, **8**, 311-330.

Lee, T.C.M. (2000) A minimum description length-based image segmentation procedure, and its comparison with a cross-validation-based segmentation procedure. *Journal of the American Statistical Association*, **95**, 259-70.

Ma, Z., Hart, M.M., and Redmond, R.L. (2001) Mapping vegetation across large geographic areas: Integration of remote sensing and GIS to classify multisource data. *Photogrammetric Engineering and Remote Sensing*, **67,** 295-307.

McIver, D.K. and Friedl, M.A. (2002) Using prior probabilities in decision-tree classification of remotely sensed data. *Remote Sensing of Environment*, **81**, 253-61.

McLachlan, G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

Nemani, R.R., Pierce, L., Running, S.W., and Band, L.E. (1993) Forest ecosystem processes at the watershed scale: sensitivity to remotely-sensed leaf-area index estimates. *International Journal of Remote Sensing*, **14**, 2519-34.

Opitz, D. and Maclin, R. (1999) Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, **11**, 169-98.

Ripley, B.D. (1996) *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

Schavio, R.A. and Hand, D.J. (2000) Ten more years of error rate research. *International Statistical Review*, **68**, 295-310.

Steele, B.M. (2000) Combining multiple classifiers: An application using spatial and remotely sensed information for land cover type mapping. *Remote Sensing of Environment*, **74**, 545-56.

Steele, B. M. and Patterson, D.A. (2000) Ideal bootstrap estimation of expected prediction error for k-nearest neighbor classifiers: Applications for classification and error assessment. *Statistics and Computing*, **10**, 349-55.

Steele, B. M. and Patterson, D.A. (in press) Land cover mapping using combination and ensemble classifiers. *Proceedings of the 33rd Symposium on the Interface*. Interface Foundation of North America,Fairvax Station VA.

Steele, B.M., and Redmond, R.L. (2001) A method of exploiting spatial information for improving classification rules: Application to the construction of polygon-based land cover maps. *International Journal of Remote Sensing*, **22**, 3143-66.

Steele, B.M., Winne, J.C., and Redmond, R.L. (1998) Estimation and mapping of misclassification probabilities for thematic land cover maps. *Remote Sensing of Environment*, **66**, 192-202.

Stehman, S.V. (1997a) Selecting and interpreting measures of thematic map accuracy. *Remote Sensing of Environment*, **62**, 77-89.

Stehman, S.V. (1997b) Thematic map accuracy assessment from the perspective of finite population sampling. *International Journal of Remote Sensing*, **16**, 589-93.

Stehman, S.V. (2000) Practical implications of design-based inference for thematic map accuracy assessment. *Remote Sensing of Environment*, **72**, 35-45.

Stehman, S.V. and Czaplewski, R.L. (1998) Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, **64**, 331-44.

Stuckens, J., Coppin, P.R., and Bauer, M.E. (2000) Integrating contextual information with per-pixel classification for improved land cover classification. *Remote Sensing of Environment*, **71**, 282-96.

Toussaint, G.T. (1974) Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory*, **IT-20**, 472-79.

Vogelmann, J.E. Sohl, T.L., Campbell, P.V., and Shaw, D.M. (1998) Regional land cover characterization using Landsat Thematic Mapper data and ancillary data sources. *Environmental Monitoring and Assessment*, **51**, 415-28.

Watson, D.F. and Philip, G.M. (1985) A refinement of inverse distance weighted interpolation. *Geo-Processing*, **2**, 315-27.

Weiss, S.M. (1991) Small sample error rate estimation for k-NN classifiers. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, **13**, 285-89.

Zhu, Z., Yang, L., Stehman, S.V., and Czaplewski, R.L. (1999) Designing an accuracy assessment for a USGS regional land cover mapping program. In *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, K. Lowell and A. Jaton (eds), Ann Arbor Press, Chelsea MI.

Tables

Table 1. Map accuracy summary statistics. Estimates were computed using the linear calibrated estimated probabilities of correct classification. The tabled values are numbers of map units (N) subjected to supervised classification, training observations (n), and classes (c). Also shown are map accuracy estimates for three classifiers, the exact bagging 10 nearest neighbor (EB 10-NN), the mean inverse distance (MID) spatial classifier, and the spatial combination classifier formed from the EB 10-NN and MID spatial classifiers (EB 10-NN+MID).

Scene	N	n	c	EB 10-NN	MID	EB 10-NN+MID
P37/R29	567457	2052	17	71.1	19.7	72.8
P38/R27	592439	2462	17	76.7	29.4	80.3
P38/R28	622080	3749	17	71.7	30.2	73.8
P38/R29	521981	1550	16	61.4	21.8	64.5
P39/R27	480916	2446	17	77.3	29.4	80.1
P39/R28	666092	4242	18	71.5	34.9	76.0
P39/R29	569595	1728	14	71.0	29.1	72.2
P40/R27	674331	2995	19	71.7	26.1	72.0
P40/R28	727864	3013	16	65.4	31.7	68.9

Table 2. Accuracy estimates for Landsat scene P39/R28. The calibration method is identified in the column heading. Polygon estimates are computed from the average estimated probabilities of correct classification using the $N = 666092$ polygon values, and training sample estimates are computed analogously for the $n = 4242$ training sample observations. The rightmost column (C.V.) shows the n-fold cross-validation estimates. Percent area estimates were determined by using the EB 10-NN+MID rule to classify polygons and summing the polygon areas according to predicted land cover class.

		Polygon Estimates Calibration Method			Training Sample Estimates Calibration Method			
Group	% Area	None	Logistic	Linear	None	Logistic	Linear	C.V.
LC[1] grassland	15.0	88.3	86.9	87.1	96.2	95.2	94.9	90.8
MC[2] grassland	24.5	81.9	80.1	80.9	92.6	91.2	91.4	91.4
HC[3] grassland	6.5	77.6	76.0	76.6	90.5	89.2	89.4	88.0
Mesic shrubland	1.6	61.9	60.9	61.1	82.4	80.4	81.3	82.0
Sagebrush[4]	10.9	68.3	66.6	67.5	85.0	82.8	83.9	82.0
Aspen	0.4	60.8	60.2	60.1	90.1	89.7	89.9	94.7
Mixed Broadleaf	1.8	77.8	76.2	76.8	96.3	95.5	95.0	95.0
Lodgepole pine (LP)	7.3	71.2	69.5	70.3	81.0	79.0	80.0	77.0
Whitebark pine	1.6	76.1	74.3	75.1	90.0	89.0	88.8	84.4
Limber pine	0.3	60.6	60.0	59.8	91.3	89.7	90.2	81.0
Ponderosa pine (PP)	0.4	65.1	63.8	64.2	75.7	73.6	74.7	65.2
Douglas-fir (DF)	10.9	74.2	72.5	73.3	87.8	86.0	86.7	93.5
Rocky Mtn juniper	1.4	62.4	61.3	61.6	90.0	88.9	88.8	75.0
DF-LP	6.9	70.4	68.6	69.4	77.5	75.4	76.5	77.5
DF-PP	1.5	65.8	64.5	64.9	82.1	80.1	81.1	74.8
Subalpine fir/spruce	0.5	69.0	67.3	68.1	76.5	75.0	75.5	63.2
Mixed xeric conifer	6.5	74.0	72.1	73.1	79.7	77.4	78.6	76.5
Rock	2.1	85.2	84.0	84.1	96.8	96.4	95.5	98.3
Lifeform								
Nonforested	58.5	80.0	78.3	79.0	91.6	90.2	90.4	88.8
Forested	39.4	72.1	70.4	71.1	83.5	81.6	82.4	82.5
Overall	100.0	77.0	75.3	76.0	88.5	86.9	87.4	86.5

[1] Very low cover; [2] low to moderate cover; [3] moderate to high cover;
[4] Sagebrush/xeric shrubland.

Table 3. Entries from the users' confusion matrices obtained from the polygon estimates of the probability of class membership, and by cross-validation. Tabled values are the estimated probability ($\times 100$) that an polygon predicted to belong to a row class is actually a member of column class.

	Actual Class			
	Polygon Estimates		Cross-validation	
Predicted Class	LC[1] grassland	MC[2] grassland	LC[1] grassland	MC[2] grassland
LC[1] grassland	88	3	91	9
MC[2] grassland	2	82	2	91
HC[3] grassland	0	2	0	3
Mesic shrubland	0	0	0	0
Sagebrush[4]	6	17	0	15
Aspen	0	1	0	0
Mixed Broadleaf	0	4	0	0
Lodgepole pine (LP)	0	0	0	0
Whitebark pine	0	3	0	0
Limber pine	0	5	0	0
Ponderosa pine (PP)	0	2	0	0
Douglas-fir (DF)	0	1	0	0
Rocky Mtn juniper	1	11	0	4
DF-LP	0	0	0	0
DF-PP	0	2	0	0
Subalpine fir/spruce	0	0	0	0
Mixed xeric conifer	0	0	0	0
Rock	2	4	0	1

[1] Very low cover; [2] low to moderate cover; [3] moderate to high cover; [4] Sagebrush/xeric shrubland.

Figures

Figure 1. The 21.5 million hectare study area is shown as the outer perimeter of the nine Landsat TM scenes and in relation to the states of Idaho, Montana and Wyoming, USA . The white inset box shows the area mapped in Figures 2 and 3 below.

Figure 2. Land cover classes predictions of the EB 10-NN (right panel) and the EB 10-NN+MID (left panel) classifiers for a portion TM scene Path 39/Row 28 (see Figure 1), and the location of the 1422 training observations contained within the area.

Figure 3. Estimated accuracy of the land cover map shown in the left panel of Figure 2. Note that manually labeled polygons are shown in white because accuracy estimates are unavailable using the method described herein.

Figure 4. Estimates of map accuracy for the nonforest and forest lifeforms obtained from the EB 10-NN classifier plotted against TM scene. The n-fold cross-validation (CV) estimates are denoted by $\widehat{\alpha}_{NF}^{CV}$ and $\widehat{\alpha}_{F}^{CV}$ in the text and the accuracy estimates computed from the linear calibrated probability estimates (PE) of correct classification are denoted by $\widehat{\alpha}_{NF}^{P}$ and $\widehat{\alpha}_{F}^{P}$.

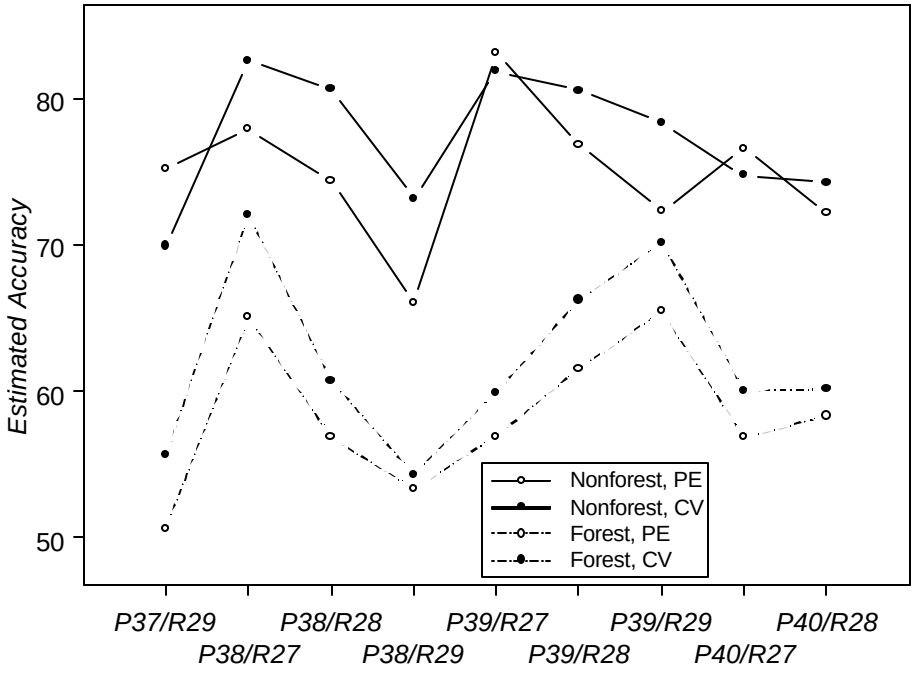


Figure 5. Mean differences, D_u, D_{lin} and D_{log}, between cross-validation estimates of map accuracy ($\widehat{\alpha}^{CV}$), and the uncalibrated, linear calibrated, and logistic calibrated versions of $\widehat{\alpha}^{P}$, computed from the training sets.

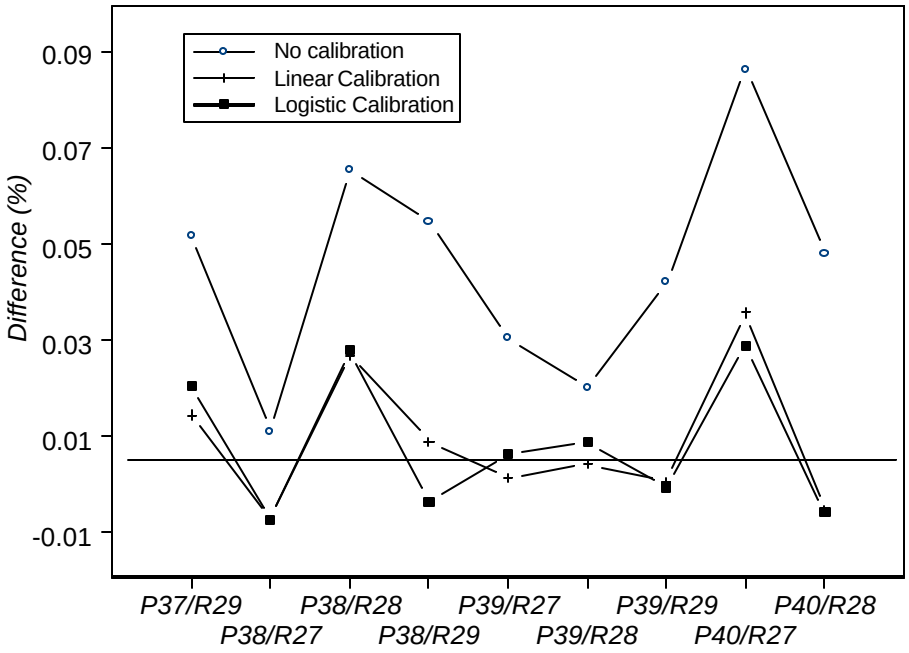
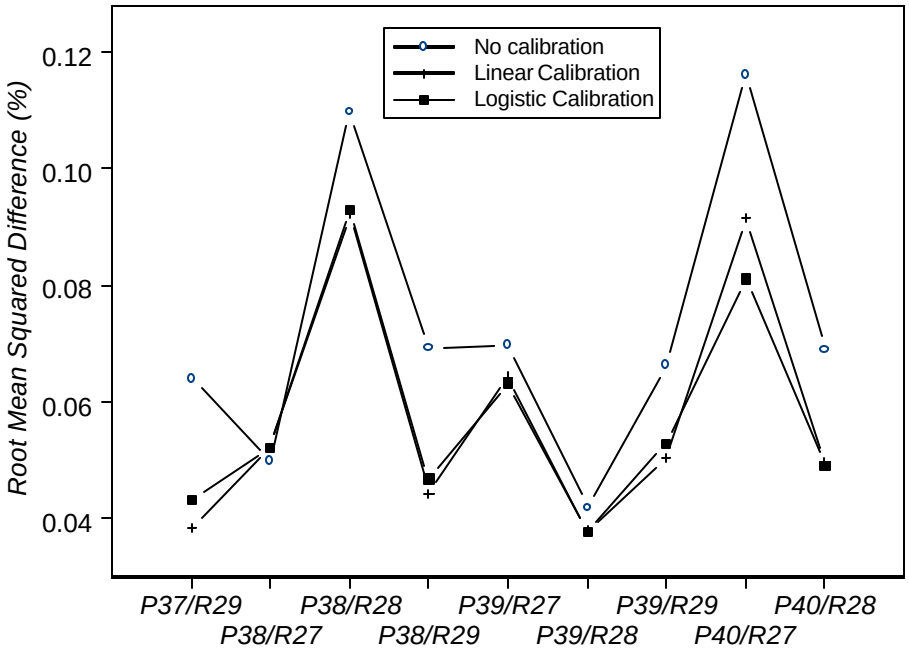


Figure 6. Root mean squared differences, S_u, S_{lin} and S_{log}, between cross-validation estimates of map accuracy ($\widehat{\alpha}^{CV}$), and the uncalibrated, linear calibrated, and logistic calibrated versions of $\widehat{\alpha}^{P}$, computed from the training sets.



Biographies

David Patterson is Professor in the Department of Mathematical Sciences at the University of Montana where he has been on the faculty since 1985. He received his Ph.D. in Statistics from the University of Iowa in 1984. Besides his work in supervised classification, he has collaborated on a variety of applied projects, principally relating to the contingent valuation method and the economic valuation of non-market natural resources and to sampling and population size in biology. He is currently on sabbatical in the Science Center at Glacier National Park under the Sabbatical in the Parks program.

Roland Redmond is a Research Associate Professor in the Division of Biological Sciences at UM, and currently directs the Wildlife Spatial Analysis Laboratory at the Montana Cooperative Wildlife Research Unit.

Brian Steele is an Assistant Professor in the Department of Mathematical Sciences at the University of Montana where he has been on the faculty since 1998. He received his Ph.D. in Mathematics from the University of Montana in 1995. He has worked on land cover classification topics since 1997, and has collaborated on a number of projects in ecology and wildlife biology in the past decade.