Maximum Posterior Probability Estimators of Map Accuracy

Brian M. Steele

Correspondence should be directed to B.M. Steele, Dept. of Mathematical Sciences, University of Montana, Missoula MT 59812, phone: 406-243-5396, fax: 406-243-2674, email: steele@mso.umt.edu

Abstract

Assessing the accuracy of land cover maps is often prohibitively expensive because of the difficulty of collecting a statistically valid probability sample from the classified map. Even when post-classification sampling is undertaken, cost and accessibility constraints may result in imprecise estimates of map accuracy. If the map is constructed via supervised classification, then the training sample provides a potential alternative source of data for accuracy assessment. Yet unless the training sample is collected by probability sampling, the estimates are, at best, of uncertain quality, and may be substantially biased. In this article, a new approach to map accuracy assessment based on maximum posterior probability estimators is discussed. These estimators may be used to reduce bias and increase precision when the training sample is collected without benefit of probability sampling, and also

to increase precision of estimates obtained from post-classification sampling. A calibrated maximum posterior probability estimate of map accuracy is computed by first estimating the probability that each map unit has been correctly classified. Then, the map unit estimates are calibrated using a function derived from either the training sample or a post-classification sample. Finally, estimates of map accuracy are computed as means of the calibrated map unit estimates. In addition to discussing maximum posterior probability estimators, this article reports on a simulation study comparing three approaches to estimating map accuracy: 1) post-classification sampling, 2) resampling the training sample via cross-validation, and 3) maximum posterior probability estimation. The simulation study showed substantial reductions in bias and improvements in precision when comparing calibrated maximum posterior probability and cross-validation estimators when the training sample was not representative of the map. In addition, combining an ordinary post-classification estimator and the maximum posterior probability estimator produced an estimator that was at least, and usually more precise than the ordinary post-classification estimator.

1 Introduction

This article is concerned with the problem of obtaining precise and unbiased accuracy estimates for land cover maps. There are two general approaches to assessing the accuracy of a map. The first, post-classification sampling, collects a probability sample of the classified map (Stehman and Czaplewski, 1998; Stehman, 2000). Accuracy is estimated by comparing observed and predicted land cover classes at the sample locations. The second approach is resampling of the training set, though this approach requires a training set that was collected by probability sampling to insure that the estimates are reliable. If so, then these data can be utilized for accuracy assessment without the cost of additional sampling. Analysis is carried out using a resampling algorithm such as cross-validation or bootstrap (Efron and Tibshirani, 1997; Hand, 1997; Schavio and Hand, 2000) to eliminate over-fitting as a substantive source of bias. In using either post-classification or resampling approaches, it is critical that the data constitute a sufficiently large probability sample of the classified map units to insure that inferences about map accuracy are precise and free, or nearly so, of bias (Hammond and Verbyla, 1996; Stehman and Czaplewski, 2003). However, probability sampling is often impractical because of time, cost and accessibility constraints (Foody, 2002). When probability sampling is carried out, the cost of obtaining sufficiently large samples for precise accuracy estimation may be unacceptable. Because this is a significant problem, strategies for improving accuracy estimators besides increasing sample size are needed (Stehman and Czaplewski, 2003).

With respect to the training sample, probability sampling is not a prerequisite for

constructing accurate maps because the classification objective is to construct a classifier for assigning class labels to unsampled population units. Consequently, statistically valid sampling designs are often not used to collect data for training the classification rule, and accuracy estimation is not carried out by resampling the training set.

For the purposes of this discussion, a land cover map is a finite set of map units each of which has been assigned a land cover class label via supervised classification. Thus, it is assumed that a subset of map units, the training sample, has been collected by ground visitation, and that for each sampling unit, observations on land cover class and one or more predictive variables are obtained. Observations on the predictive variables are available for all map units, and a classification rule uses these predictive variables to assign land cover to the unsampled map units. The classification rule is constructed from, or trained on, the training sample. In the examples that follow, the predictor variables are reflectance intensity recorded on 6 of 7 spectral bands measured by the Landsat Thematic Mapper (TM) satellite and measurements on elevation, slope and aspect. These data were collected for a mapping project covering 21.5 million hectares of forested mountains and rangeland within and adjacent to the Rocky Mountains in northwestern USA. The map was constructed from nine adjacent Landsat TM scenes, and each map consists of between $480,916$ and $727,864$ map units. The USDA Forest Service initiated the project and drew most of the training observations from existing data sets that were collected for purposes other than mapping land cover. Some, but not all of these training data

were collected by probability sampling of Landsat TM scene subregions. The overall spatial distribution of training observations was highly irregular, largely because most were sampled from public lands with easy access. Thus, privately owned lands and wilderness areas were sporadically sampled, and those observations that were collected usually were done so opportunistically. For these reasons, none of the training data sets constitute a probability sample. Moreover, post-classification sampling was not pursued because of financial constraints. Yet, the usefulness of the map is substantially diminished without trustworthy accuracy estimates and there is little guidance on how to proceed in this situation. An approach for assessing map quality recently was proposed by Baraldi et al. (2005) based on measuring fidelity among different mappings of selected regions of the map. Their method is qualitative, particularly since the extent to which fidelity coincides with accuracy is unknown.

In contrast, the method discussed herein estimates conventional measures of map accuracy via the *maximum posterior probability* approach (Steele et al., 2003). The central innovation of Steele et al.'s (2003) method is that the probability of correct classification is estimated for every map unit. Map accuracy, and user's accuracy are estimated by computing appropriately defined population means of the map unit estimates. While Steele et al., (2003) developed the statistical rationale supporting the method, they did not investigate bias nor precision of the accuracy estimators. In this article, the methodology proposed by Steele et al., (2003) is refined and extended to improve the precision

of estimators based on post-classification sampling. In addition, a simulation study investigating the performance of these estimators is reported on. In brief, compared to post-classification sampling and resampling estimators, maximum posterior probability estimators were found to be as, or more precise than post-classification sampling and resampling estimators, and much less biased when training samples were not representative of the classified map.

The key ideas addressed in this article are as follows. If a training sample collected without benefit of a statistical sampling design is used for estimating map accuracy, then the estimates may be substantially in error and hence, are unreliable. Additionally, if post-classification sampling is used to collect data for this purpose, then post-classification estimators will be imprecise if cost and accessibility problems limit the number of sample observations. There are two situations for which maximum posterior probability accuracy estimators are designed. The first of these situations occurs when a post-classification sample is unavailable and the training sample was not collected using a statistical sampling design. In this situation, the usual map accuracy estimators derived from the training sample (such as those based on cross-validation) may be biased, perhaps badly so depending on what criteria were used to select the sample. For example, accuracy estimators likely to be optimistically biased if observations that are ideal representatives a particular class are preferentially selected over observations that appear to have characteristics of two or more classes. The second situation occurs when

a post-classification sample has been collected using a statistically valid sampling design. Then, map accuracy can be estimated without bias, though accuracy estimators may lack sufficient precision if the sample size is small. The maximum posterior probability estimators discussed herein are aimed at eliminating or at least reducing the bias in the first situation. It cannot be said that maximum posterior probability estimators are unbiased in the first situation; however, the research discussed herein argues that these estimators are better than existing methods in this regard. It is also argued that when a post-classification sample is available (the second situation), then maximum posterior probability estimators usually will improve on the precision of post-classification accuracy estimators. In brief, the maximum posterior probability method estimates the probability of correct classification for *every* map unit, and computes simple averages of these values to obtain estimates of overall map accuracy and user's accuracy for individual classes. Bias and precision are improved on because accuracy is estimated for every map unit. To insure that this innovation works, individual map unit estimates are calibrated using the training sample. The cost, in terms of bias, of using a training sample collected without probability sampling for calibration is much less than using the sample directly for accuracy estimation via resampling methods such as cross-validation or bootstrapping. Hence, maximum posterior probability estimators represent a significant improvement over resampling methods when a statistical sampling design is not used for collecting the training sample. Moreover, when a post-classification sample is

available, the precision of map accuracy estimators may be improved on by the maximum posterior probability method.

The next section develops the statistical ideas and notation needed to discuss maximum posterior estimators in detail.

2 Classifiers and posterior probabilities

Consider a population $\mathcal{P}$ comprised of c classes, or groups, identified by labels $1, \ldots, c$. An element of $\mathcal{P}$ is a pair $\boldsymbol{x} = (\boldsymbol{t}, y)$ consisting of a covariate vector $\boldsymbol{t}$ and a group label y. In the examples below, covariate vectors consist of measurements on reflectance for Landsat TM bands and physical features, and the group labels identify land cover class. A training sample of size n collected from $\mathcal{P}$ is denoted by $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. The posterior probability that $\boldsymbol{x} \in \mathcal{P}$ belongs to group g, given the covariate vector $\boldsymbol{t}$, is denoted by $P(y = g \mid \boldsymbol{t})$. Let η denote a classification rule trained on $\boldsymbol{X}$, and $\eta(\boldsymbol{t})$ denote a prediction of y obtained by applying the classification rule to $\boldsymbol{x} \in \mathcal{P}$. For example, the elementary k-nearest neighbor (k-NN) classifier assigns $\boldsymbol{x}$ to the group with the largest plurality among the k nearest neighbors of $\boldsymbol{x}$ among the training observations where the distance between $\boldsymbol{x}$ and $\boldsymbol{x}_i \in \boldsymbol{X}$ is measured on the covariates using, say, Euclidean distance. The k-NN classifier can also be formulated as an estimator of the c posterior probabilities. Specifically, the k-NN estimator of $P(y = g \mid \boldsymbol{t})$ is the sample proportion of the k-nearest

neighbors in the training sample that belong to group g, and the classification rule assigns $\boldsymbol{x}$ to the group with the maximum posterior probability estimate. Any classifier that produces a quantitative score expressing the relative likelihood of membership in group g can be used to compute posterior probability estimates of group membership by adding a suitable constant to each score to insure that scores are non-negative and then dividing each score by the sum of the c shifted scores. The rule can then be defined in terms of the estimated posterior probabilities by assigning $\boldsymbol{x}$ to the group with the maximum posterior probability estimate.

A second example is the linear discriminant classifier (McLachlan 1992, Chap. 1). The estimator of $P\left(y = g \mid \boldsymbol{t}\right)$ is

$$\widehat{P}\left(y = g \mid \boldsymbol{t}\right) = \frac{p_g \exp\left[-\frac{1}{2} d(\boldsymbol{t}, \overline{\boldsymbol{t}}_g)\right]}{\sum_{j=1}^{c} p_j \exp\left[-\frac{1}{2} d(\boldsymbol{t}, \overline{\boldsymbol{t}}_j)\right]}, \tag{1}$$

where $\overline{\boldsymbol{t}}_g$ is the covariate sample mean computed from training observations belonging to group g, $d(\boldsymbol{t}, \overline{\boldsymbol{t}}_g) = (\boldsymbol{t} - \overline{\boldsymbol{t}}_g)^T \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{t} - \overline{\boldsymbol{t}}_g)$ is the squared Mahalanobis distance between $\boldsymbol{t}$ and $\overline{\boldsymbol{t}}_g$, p_g is the estimated prior probability of membership in group g, and $\widehat{\boldsymbol{\Sigma}}$ is the pooled covariance matrix computed from the training observations. This classifier is optimal if the covariate vectors are multivariate normal in distribution with a common covariance matrix and the estimates $\overline{\boldsymbol{t}}_g$ and $\widehat{\boldsymbol{\Sigma}}$ are replaced with the population parameters. However, the parameters are unknown in practical land cover mapping exercises and the distributional model is often unrealistic given that land cover classes vary with respect to homogeneity of dominant species and surface types. The Gaussian maximum

likelihood classifier has been preferred to the linear discriminant function for land cover mapping because separate covariance matrix estimates are computed for each group, thereby avoiding the assumption of common covariance.

The accuracy of a *classifier* refers to the probability of correctly classifying a population unit when using a classifier constructed from a probability sample. This definition addresses the performance of the classifier before the sample is collected and is relevant for evaluating the properties of the classifier. The covariate distribution, the classification rule, and the sampling design determine the accuracy of the classifier. Once the training sample is in hand, then interest shifts to the proportion of population units that are correctly classified by the rule constructed from the training sample. This rate is conditional on the training sample and implicitly, on the specific classifier derived from the sample, and so the focus is on the performance of this specific classifier on the population of map units (see McLachlan, 1992, sec., 10.1 and Efron and Tibshirani, 1997 for further discussion). As the population of interest is a collection of map units, this rate is called map accuracy.

3 Map accuracy

Commonly, the accuracy of a map is analyzed from a design-based perspective in which the map as a viewed as fixed population of map units, and the only random process related

to map accuracy is the selection of sample units used to estimate accuracy (Stehman 2000). Map accuracy then defined in a natural way as the proportion of correctly classified map units. Let α denote map accuracy, and note that the probability that a map unit drawn at random from the population will be correctly classified by the classifier η is equal to α. This perspective also leads to defining the user's accuracy rate for class g as the proportion of map units assigned to class g that have been correctly classified. This parameter, denoted by α^g is also equal to the probability of randomly sampling a correctly classified map unit from those classified as group g.

Some additional mathematical development is useful before discussing maximum posterior probability estimators. Map accuracy can be defined according to the formula

$$\alpha = N^{-1} \sum_{i=1}^{N} \Psi[\eta(\boldsymbol{t}_i) = y_i], \tag{2}$$

where N is the number of map units and $\Psi[\eta(\boldsymbol{t}_i) = y_i] = 1$ if y_i was correctly classified and is 0 if y_i was incorrectly classified. Generally, design-based accuracy estimators are functions of a subset of the outcomes $\Psi[\eta(\boldsymbol{t}_i) = y_i]$, $i = 1, \ldots, N$. For example, if $\boldsymbol{X}^p = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is a post-classification sample collected by randomly sampling $\mathcal{P}$, then the usual estimator of α is the proportion of sample units in $\boldsymbol{X}^p$ correctly classified by the rule η, say

$$\alpha(\boldsymbol{X}^p) = n^{-1} \sum_{i=1}^{N} \Psi[\boldsymbol{x}_i \in \boldsymbol{X}^p] \Psi[\eta(\boldsymbol{t}_i) = y_i], \tag{3}$$

where $\Psi[\boldsymbol{x}_i \in \boldsymbol{X}^p]$ is 1 if $\boldsymbol{x}_i$ is a sample unit and 0 if not. An estimator of the user's accuracy for class g is the proportion of sample units assigned to class g that are correctly

classified. i.e.,

$$\alpha^g(\boldsymbol{X}^p) = \frac{\sum_{i=1}^{N} \Psi[y_i = g]\Psi[\eta(\boldsymbol{t}_i) = g]\Psi[\boldsymbol{x}_i \in \boldsymbol{X}^p]}{\sum_{i=1}^{N} \Psi[\eta(\boldsymbol{t}_i) = g]\Psi[\boldsymbol{x}_i \in \boldsymbol{X}^p]}.$$

If probability sampling has been used to collect the post-classification sample, then $\Psi[\boldsymbol{x}_i \in \boldsymbol{X}^p]$ is a random variable and the probability of sampling $\boldsymbol{x}_i$ can be determined. From the design-based perspective, this is the only source of randomness; all other terms are fixed.

The development of maximum posterior probability estimators requires a shift from a design-based perspective to a model-based perspective of the map. Generally, a model-based analysis views the classified map as a single realization of a random process, and attempts to identify a model that describes the process, or some portion thereof. If the model is realistic, then it may be possible extract additional information about accuracy through the classifier behavior. To begin, consider an event A that occurs with probability $P(A) = p$. The indicator function Ψ is useful for analyzing events such as A; herein $\Psi(A)$ is defined according to

$$\Psi(A) = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur.} \end{cases}$$

Because A is an event, $\Psi(A)$ is a random variable, specifically, a Bernoulli random variable with parameter p, and the expectation of $\Psi(A)$ is $\mathrm{E}[\Psi(A)] = p$ (Ross, 1998, Chap. 4.7). That is, the expected value of the indicator random variable is equal to the probability of the event. This model can be applied to the event of classifying a map unit $\boldsymbol{x} = (\boldsymbol{t}, y)$.

It is useful to imagine that there are many map units with the specific covariate vector value $\boldsymbol{t}_0$, and that these units belong to more than one group. Further, suppose that we know only the value of $\boldsymbol{t}_0$ and are interested in whether the classifier will correctly classify the map unit. If the classifier correctly classifies $\boldsymbol{x}$, then $\Psi[\eta(\boldsymbol{t}_0) = y] = 1$, and if the prediction is incorrect, then $\Psi[\eta(\boldsymbol{t}_0) = y] = 0$. The previous discussion of the Bernoulli random variable implies that the probability of correctly classifying $\boldsymbol{x}$ given that $\boldsymbol{t} = \boldsymbol{t}_0$ is $P[\eta(\boldsymbol{t}_0) = y \mid \boldsymbol{t}_0] = \mathrm{E}\{\Psi[\eta(\boldsymbol{t}_0) = y \mid \boldsymbol{t}_0]\}$.

To develop maximum posterior probability estimators, a model-based perspective is adopted that views each covariate vector as potentially occurring with any of the c groups. For every possible value of $\boldsymbol{t}$, there is a probability distribution expressing the relative likelihood of membership within each group. This view is compatible with the notion of a covariate space in which the probability of group membership varies across the space. Classification errors occur because groups are not completely separated in the covariate space, and it will be shown that the group membership distributions determine the probability that $\boldsymbol{x}$ is correctly classified given $\boldsymbol{t}$. The probability of correctly classifying $\boldsymbol{x}$ using the rule η is denoted by $P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$. Generally, $P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$ is not equal to α, the accuracy of the map and the proportion of correctly classifier units, because $\boldsymbol{t}$ provides information that makes the probability of correct classification more precise. Before taking up these ideas at greater length, consider a simple example. Suppose that 60% of all possible map units with the covariate vector $\boldsymbol{t}_0$ belong to group 1, 30% belong

to group 2 and 10% to group 3. Then, if we know only $\boldsymbol{t}_0$ and these relative proportions, $\boldsymbol{t}_0$ always should be assigned to group 1 because this group is most likely to be correct. Moreover, the proportion of times that $\boldsymbol{t}_0$ will be correctly classified using this rule is 0.6 because 60% of the units with the covariate vector $\boldsymbol{t}_0$ belong to group 1, and hence, the probability of correct classification is $P[\eta(\boldsymbol{t}_0) = y \mid \boldsymbol{t}_0] = 0.6$. This argument is based on a model of the mapping process, and in particular, of the covariate space; in contrast, the design-based approach ignores the origins of the map, and is only concerned with what exists. A particular advantage of the design-based approach is that estimation of the variance of map accuracy estimators is usually tractable whereas under the model-based approach variance estimation generally is intractable. In the next section, the model-based perspective motivates a new approach to accuracy assessment via the map unit-specific probabilities of correct classification $P[\eta(\boldsymbol{t}_i) = y \mid \boldsymbol{t}_i], i = 1, \ldots, N$.

4 Maximum posterior probability

This section connects design- and model-based perspectives of map accuracy, and discusses the relationship between the probability of correct classification for an individual map unit and the group membership probabilities. To summarize, the formal definitions of map accuracy differ between the two perspectives, though the numerical values are indistinguishable in practical situations. More importantly, the model-based perspec-

tive leads to a probability model describing the outcome of classifying an individual map unit, and a definition of the probability of correct classification for the map unit. Specifically, the probability of correct classification is shown to be the maximum of the c group membership probabilities. This relationship motivates an estimator of the probability of correct classification for an individual map unit based on the maximum *estimated* probability of group membership.

Herein, the accuracy of a single prediction $\eta(\boldsymbol{t})$ is defined to be the conditional probability of correctly classifying the observation $\boldsymbol{x}$, given the sample $\boldsymbol{X}$, and the covariate vector $\boldsymbol{t}$. As discussed above, the notation for this quantity is $P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$. The random variable $\Psi[\eta(\boldsymbol{t}) = y]$ identifying the correctness of the classification of $\boldsymbol{x}$ is Bernoulli in distribution with probability $P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$. Because $\Psi[\eta(\boldsymbol{t}) = y]$ is a Bernoulli random variable, the probability of correct classification is also the expectation of the random variable, i.e., $P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = \mathrm{E}\{\Psi[\eta(\boldsymbol{t}) = y] \mid \boldsymbol{t}\}$. An alternative definition of map accuracy can be formulated by replacing the outcomes of classification in the definition of α [formula (2)] by the probabilities of correct classification. The resulting parameter is the population mean probability of correct classification $\alpha^* = N^{-1} \sum P[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}_i] = N^{-1} \sum \mathrm{E}\{\Psi[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}_i]\}$. The two parameters α and α^* can be compared from the model-based perspective. The difference $\alpha - \alpha^* = N^{-1} \sum (\Psi[\eta(\boldsymbol{t}_i) = y_i] - \mathrm{E}\{\Psi[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}]\})$ is not 0, though Chebychev's inequality can be used to argue that the difference between $\alpha - \alpha^*$ will be small for large

N. The reason that α and α^* are not equal is that the left-hand terms are the realized outcomes of classification and the right-hand terms are the probabilities of correct classification, though expressed as expectations rather than probabilities. The realized outcomes are either 0 or 1 whereas the probabilities are never 0 and rarely 1 (at least in real applications). In practice, N is usually large, and the difference between the two terms tends to be very small.

Estimators of the probability of correct classification $P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$ are motivated by their connection to the posterior probabilities of class membership $P(y = g \mid \boldsymbol{t})$, $g = 1, \ldots, c$. Specifically, the probability that $\boldsymbol{x} = (\boldsymbol{t}, y)$ is correctly classified, given $\boldsymbol{t}$, is the maximum posterior probability:

$$P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = \max_g P(y = g \mid \boldsymbol{t}) \tag{4}$$

(Huberty, 1994, Chap. 6; Glick, 1978; Ripley 1996, chap. 2). The Appendix provides a proof of this statement. This result is not immediately useful for accuracy estimation because the posterior probabilities of class memberships are almost always unknown. However, the classifier η allows construction of estimators $\widehat{P}(y = g \mid \boldsymbol{t})$, for each class $g = 1, \ldots, c$, and these can be used to define a plug-in estimator of $P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$, say

$$\widehat{P}[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = \max_g \widehat{P}(y = g \mid \boldsymbol{t}). \tag{5}$$

Then, α^* and α can be estimated by the population average $N^{-1} \sum \widehat{P}[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}_i]$. However, the maximum posterior probability estimator [equation (5)] is not unbiased.

In fact, the estimator may be badly biased, and consequently, it is necessary to calibrate the estimates to reduce the bias of the accuracy estimator (Ripley 1996, chap 2; Steele et al., 2003).

4.1 Calibration of maximum posterior probability estimates

The objective of calibration is to construct a transformation that reduces the bias of the maximum posterior probability estimators. This is accomplished by modeling the probability of correct classification for an individual map unit as a function of the maximum posterior probability estimate. The fitted calibration model takes a maximum posterior probability estimate as a input and outputs a fitted value that better estimates the probability of correct classification. Data for fitting the calibration model can come from either a post-classification sample or from the training sample. In either case, a cross-validation algorithm is used to classify and compute maximum posterior probability estimates for each observation.

Calibration begins with a simple, prototype model that states that the true probability of correct classification $P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$ is proportional to the maximum posterior probability estimate $\widehat{P}[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$. Following Steele et al., (2003), the calibration model is

$$P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = \beta \widehat{P}[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}], \tag{6}$$

where β is the proportionality constant, or calibation parameter. Because $P[\eta(\boldsymbol{t}) = y \mid$

$\boldsymbol{t}] = \mathrm{E}\{\Psi[\eta(\boldsymbol{t}) = y] \mid \boldsymbol{t}\}$, equation (6) is seen to be equivalent to the no-intercept linear regression model $\mathrm{E}(Y) = \beta x$ of the mean of a random variable Y given a covariate x if we equate $\mathrm{E}(Y)$ with $\mathrm{E}\{\Psi[\eta(\boldsymbol{t}) = y] \mid \boldsymbol{t}\}$ and x with $\widehat{P}[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$. Note that this model constrains the regression line to pass through the origin. Consequently, the least squares estimator of β is $\sum x_i y_i / \sum x_i^2$, where x_i and y_i are the elements of an observation pair (x_i, y_i). The least squares estimator for the calibration model [equation (6)] yields the calibration coefficient $b_0 = \sum p_i \psi_i / \sum p_i^2$ by substituting $p_i = \widehat{P}[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}_i]$ for x_i and $\psi_i = \Psi[\eta(\boldsymbol{t}_i) = y_i]$ for y_i (Steele et al., 2003). Methods of obtaining the calibration sample $C = \{(p_1,\psi_1), \ldots, (p_n, \psi_n)\}$ are discussed momentarily.

The calibration model can be improved by constraining it to pass through the pair (c^{-1}, c^{-1}) instead of the origin. This constraint is motivated by two properties of posterior probabilities. The first property states that all probabilities are non-negative, hence, $0 \leq P(y = g \mid \boldsymbol{t})$ for each $g = 1, \ldots, c$. The second property states that the sum of group membership probabilities is 1; that is, $\sum P(y = g \mid \boldsymbol{t}) = 1$. Together, these conditions imply that the maximum of the group membership probabilities is at least $1/c$; otherwise, the sum of the group membership probabilities will be less than 1. This lower bound carries over to the probability of correct classification, and thus, $c^{-1} \leq \max P(y = g \mid \boldsymbol{t}) = P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$. The same conditions and constraint holds for the maximum posterior probability estimator $\widehat{P}[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$ because probability estimators should always be constrained to obey the non-negativity and summation properties described

above. Constraining the calibration model to pass through the pair (c^{-1}, c^{-1}) insures that the calibrated estimate $\widehat{P}[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}]$ is not less than c^{-1}. The Appendix shows that this constraint implies that the least squares estimator of β is

$$b = \frac{\sum(p_i - c^{-1})(\psi_i - c^{-1})}{\sum(p_i - c^{-1})^2}. \tag{7}$$

The calibrated maximum posterior probability estimator is

$$\widehat{P}_c[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = \begin{cases} b\widehat{P}[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] + c^{-1}(1-b) & \text{if } b\widehat{P}[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] + c^{-1}(1-b) \leq 1 \\ 1 & \text{if } b\widehat{P}[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] + c^{-1}(1-b) > 1. \end{cases} \tag{8}$$

Simulation studies (not reported herein) indicated that this new calibration method is preferable to the no-intercept linear regression equation with respect to bias and mean square error when used for estimating α.

A calibration sample $C = \{(p_1, \psi_1), \ldots, (p_n, \psi_n)\}$ consists of n pairs of binary outcomes $\psi_i = \Psi[\eta(\boldsymbol{t}_i) = y_i]$ and maximum posterior probability estimates $p_i = \widehat{P}[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}]$. One approach to collecting C is post-classification sampling. For simplicity, assume that the post-classification sample is collected by simple random sampling, though other sampling designs may be preferable (Stehman and Czaplewski, 1998). The pair (ψ_i, p_i) is obtained from the ith post-classification observation $(\boldsymbol{t}_i, y_i)$ by using the classifier η to compute p_i and a class membership prediction $\eta(\boldsymbol{t}_i)$. The outcome ψ_i is determined by comparing $\eta(\boldsymbol{t}_i)$ to y_i. The calibration sample is then used to estimate the calibration coefficient b according to formula (7).

The second approach uses the training sample to generate the calibration set. The important difference between this and the post-classification approach is that now the training observations are used both for constructing the rule η and to evaluate the rule. Because this strategy of reusing the data often leads to over-fitting and bias in accuracy estimators, a cross-validation algorithm is used. A k-fold cross-validation procedure randomly partitions $\boldsymbol{X}$ as k subsets $\boldsymbol{X}^j$, $j = 1, \ldots, k$, consisting of approximately equal numbers of observations. All observations not in $\boldsymbol{X}^j$ are used to construct a classifier, and the classifier is applied to all observations in $\boldsymbol{X}^j$. This procedure is repeated for each holdout set $\boldsymbol{X}^j$, $j = 1, \ldots, k$. At completion, each observation in $\boldsymbol{X}$ has been held-out and classified exactly once, and the pairs $(p_1, \psi_1), \ldots, (p_n, \psi_n)$ comprise calibration sample C. In the simulation study, 10-fold cross-validation was used because preliminary analyses indicated little difference between using $k = 10$ and using values of k greater than 10, and because computational cost increases with k.

5 Estimators of map accuracy

Two accuracy estimators, and their application to estimating map accuracy and user's accuracies, are discussed in this section. The first of these, the cross-validation estimator, is a commonly used resampling estimator. Accuracy is estimated by predicting class membership for each training observation and comparing the predicted and actual

classes. To reduce bias associated with using an observation for both classifier construction and classifier evaluation, each observation is withdrawn from the data set before it is classified. Maximum posterior probability estimators are defined as means of the calibrated maximum posterior probability estimates across the population or subsets of the population. The key innovation is that these are *population* means, rather than *sample* means obtained from the training or post-classification sample.

As discussed in the previous section, k-fold cross-validation partitions the training sample $\boldsymbol{X}$ as k disjoint sets. In turn, each test set $\boldsymbol{X}^j$ is removed, and the remaining observations are used to construct a classifier η_{-j}. The classifier η_{-j} is used to classify all observations in $\boldsymbol{X}^j$. After all observations have been classified, accuracy is estimated by the proportion of correctly classified observations. The k-fold cross-validation estimator can be expressed as

$$\alpha_{CV}(\boldsymbol{X}) = n^{-1} \sum_{j=1}^{k} \sum_{i=1}^{n} \Psi[\boldsymbol{x}_i \in \boldsymbol{X}^j] \Psi[\eta_{-j}(\boldsymbol{t}_i) = y_i].$$

The first term in the sum, $\Psi[\boldsymbol{x}_i \in \boldsymbol{X}^j]$, is 1 if $\boldsymbol{x}_i$ is in the holdout sample $\boldsymbol{X}^j$, and 0 otherwise. The second term in the sum, $\Psi[\eta_{-j}(\boldsymbol{t}_i) = y_i]$, is 1 if $\boldsymbol{x}_i$ is correctly classified by the classifier η_{-j} constructed without the held-out observations, and is 0 otherwise. The estimator α_{CV} is nearly unbiased for large n if the training sample is a random sample of the population of map units (Ripley 1996, chap. 2). Cross-validation also provides an estimator of α^g, the proportion of map units assigned to class g that are correctly

classified. This estimator is

$$\alpha_{CV}^{g}(\boldsymbol{X}) = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\Psi[\boldsymbol{x}_i \in \boldsymbol{X}^j]\Psi[y_i = g]\Psi[\eta_{-j}(\boldsymbol{t}_i) = g]}{\sum_{j=1}^{k}\sum_{i=1}^{n}\Psi[\boldsymbol{x}_i \in \boldsymbol{X}^j]\Psi[\eta_{-j}(\boldsymbol{t}_i) = g]}.$$

Maximum posterior probability accuracy estimators use a calibration sample and the population covariates $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N$ to estimate map accuracy. A maximum posterior probability estimator is evaluated as follows. First, a calibration coefficient b is computed from a calibration sample. Then, the calibrated maximum posterior probability estimators of correct classification are evaluated for each population unit by evaluating the maximum posterior probability estimators $\widehat{P}[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}_i]$ using η, and then calibrating the estimates using formula (8). Finally, the maximum posterior probability estimator is evaluated by computing the mean of the N calibrated maximum posterior probability estimates $\widehat{P}_c[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}_i]$. If the calibration function is derived from the training sample $\boldsymbol{X}$, then the maximum posterior probability estimator of α is denoted by

$$\alpha_M(\boldsymbol{X}) = N^{-1}\sum_{i=1}^{N}\widehat{P}_c[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}_i].$$

A maximum posterior probability estimator of α^g is the mean of the maximum posterior probability estimators over those map units assigned to class g. The estimator is formally defined as

$$\alpha_M^{g}(\boldsymbol{X}) = \frac{\sum_{i=1}^{N}\widehat{P}_c[\eta(\boldsymbol{t}_i) = y_i]\Psi[\eta(\boldsymbol{t}_i) = g]}{\sum_{i=1}^{N}\Psi[\eta(\boldsymbol{t}_i) = g]}.$$

Two remarks regarding calibrated maximum posterior probability estimators should be made. First, the maximum posterior probability estimates for individual map units

can be used to produce a map of estimated map accuracy. For example, Steele et al. (2003) shows a gray-scale map of estimated map accuracy constructed from the estimates $\widehat{P}_c[\eta(\boldsymbol{t}_i) = y_i \mid \boldsymbol{t}_i]$, $i = 1, \ldots, N$. Secondly, if the results of automated classification are substantially manipulated by a scene analyst, then map accuracy estimates derived from the automated classification are of diminished value for describing accuracy of the final map. Because a manually modified map is different, and presumably more accurate than the original automatically classified map, estimates derived from the training set may not be accurate, and little can be said about the magnitude of the differences.

5.1 Variance estimation

It is desirable to accompany an estimate of map accuracy with a measure of uncertainty such as the estimated mean square error, variance, or standard deviation of the accuracy estimator. Recall that the training sample and classifier are considered fixed when the interest is in map accuracy, as it is in this article. Methods of estimating the variance of post-classification map accuracy estimators are well-established (for instance, see Nusser and Klaas, 2003, and Stehman and Czaplewski, 1998). When only the training sample is available for accuracy assessment, the problem of variance estimation is intractable. It is not known how to resample the training sample in such a way that the uncertainty associated with map accuracy estimators can be correctly simulated. One alternative is to assess the variance of a bootstrap estimator of the classifier accuracy in lieu of

map accuracy as there are methods of estimating the variance in a bootstrap estimator of classifier accuracy (Efron and Tibshirani, 1997 and McLachlan 1992, sec. 10.3.3). However, good variance estimators for cross-validation accuracy estimators are unknown (Bengio and Grandvalet 2004; McLachlan 1992, chap. 10.3). Moreover, it is not obvious how well existing methods for classifier accuracy assessment transfer to map accuracy assessment.

5.2 Combining estimators of map accuracy

It is sometimes possible to combine two accuracy estimators to produce a single estimator that is more accurate than the constituent estimators. In particular, if a post-classification sample is available, then it is feasible to combine a post-classification estimator with a maximum posterior probability estimator derived from the training sample. The accuracy of a combination estimator, as measured by the mean square error, is likely to be less than the constituent estimators provided that the mean square error of one classifier is not several times greater than the other.

Suppose that $\widehat{\alpha}(\boldsymbol{X})$ is an estimator of α. The mean square error of $\widehat{\alpha}(\boldsymbol{X})$ measures the expected squared difference between the estimator and the estimand α. That is, the mean square error of $\widehat{\alpha}(\boldsymbol{X})$ is $\mathrm{MSE}[\widehat{\alpha}(\boldsymbol{X})] = \mathrm{E}\{[\widehat{\alpha}(\boldsymbol{X}) - \alpha]^2\}$. Both the variance and bias of $\widehat{\alpha}(\boldsymbol{X})$ contribute to the mean square error of $\widehat{\alpha}(\boldsymbol{X})$ because $\mathrm{MSE}[\widehat{\alpha}(\boldsymbol{X})] = \mathrm{Var}[\widehat{\alpha}(\boldsymbol{X})] + \mathrm{B}[\widehat{\alpha}(\boldsymbol{X})]^2$, where $\mathrm{Var}(\widehat{\alpha})$ is the variance and $\mathrm{B}[\widehat{\alpha}(\boldsymbol{X})] = \mathrm{E}[\widehat{\alpha}(\boldsymbol{X})] - \alpha$ is

the bias of the estimator. The mean square error is a particularly useful measure of estimator quality because both the variance and bias are accounted for. Suppose that $\widehat{\alpha}_1 = \widehat{\alpha}_1(\boldsymbol{X}_1)$ and $\widehat{\alpha}_2 = \widehat{\alpha}_2(\boldsymbol{X}_2)$ are independent random variables and that one estimator is unbiased, say $\mathrm{B}[\widehat{\alpha}_1(\boldsymbol{X}_1)] = 0$. For example, if the samples are a post-classification sample collected by simple random sampling and a training sample, then the samples will be independent and the bias of the post-classification accuracy estimator $\alpha(\boldsymbol{X}^p)$ will be 0. Suppose that $0 \leq v \leq 1$, is a known constant. A third estimator is the linear combination of estimators $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ defined by $\widehat{\alpha}_3 = \widehat{\alpha}_3(\boldsymbol{X}_1, \boldsymbol{X}_2) = v\widehat{\alpha}_1 + (1-v)\ \widehat{\alpha}_2$. The estimator $\widehat{\alpha}_3$ has bias $\mathrm{B}(\widehat{\alpha}_3) = v\mathrm{B}(\widehat{\alpha}_1) + (1-v)\mathrm{B}(\widehat{\alpha}_2)$; moreover, Appendix 3 shows that the mean square error of $\widehat{\alpha}_3$ is $\mathrm{MSE}(\widehat{\alpha}_3) = v^2\mathrm{MSE}(\widehat{\alpha}_1) + (1-v)^2\mathrm{MSE}(\widehat{\alpha}_2)$, and that the value of v that minimizes $\mathrm{MSE}(\widehat{\alpha}_3)$ is $\mathrm{MSE}(\widehat{\alpha}_2)/[\mathrm{MSE}(\widehat{\alpha}_1)+\mathrm{MSE}(\widehat{\alpha}_2)]$. However, $\mathrm{MSE}(\widehat{\alpha}_1)$ and $\mathrm{MSE}(\widehat{\alpha}_2)$ are rarely known, and so a reasonable choice is to construct $\widehat{\alpha}_3$ by averaging $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ by setting $v = 1/2$. Then, $\mathrm{MSE}(\widehat{\alpha}_3) = [\mathrm{MSE}(\widehat{\alpha}_1)+\mathrm{MSE}(\widehat{\alpha}_2)]/4$. Some care may be taken because it is possible that $\mathrm{MSE}(\widehat{\alpha}_3)$ will greater than the smaller of $\mathrm{MSE}(\widehat{\alpha}_1)$ and $\mathrm{MSE}(\widehat{\alpha}_2)$ if one of the estimators has a mean square error larger than 3 times that of the other. Estimates of the mean square errors sometimes can be computed via bootstrapping (Efron and Tibshirani, 1993).

Combining estimators for accuracy assessment is feasible if a post-classification sample has been collected. Let $\alpha(\boldsymbol{X}^p)$ denote an estimator based on the post-classification sample $\boldsymbol{X}^p$, and suppose that the training sample is used to construct a calibrated

maximum posterior probability estimator $\alpha_M(\boldsymbol{X})$. Then, the estimator $\overline{\alpha}(\boldsymbol{X}, \boldsymbol{X}^p) = \frac{1}{2}[\alpha(\boldsymbol{X}^p)+\alpha_M(\boldsymbol{X})]$ may improve on both $\alpha(\boldsymbol{X}^p)$ and $\alpha_M(\boldsymbol{X})$ with respect to mean square error. The estimator $\overline{\alpha}(\boldsymbol{X}, \boldsymbol{X}^p)$ will be referred to as the post-classification maximum posterior probability estimator. Heuristically, averaging the two estimators reduces bias associated with $\alpha_M(\boldsymbol{X})$ and variance associated with $\alpha(\boldsymbol{X}^p)$.

6 Examples

A mapping project covering 21.5 million hectares of forested mountains and rangeland within Montana illustrates differences in map accuracy estimates in application. The map region is located east of the Continental Divide and includes portions of the Rocky Mountains and the adjacent Great Plains. The map is comprised of 9 Landsat Thematic Mapper scenes, each of which was classified separately before edge-matching. The intended uses of the map encompass commercial timber and range management, hydrology, and wildlife management. The USDA Forest Service, Northern Region, initiated the project and amalgamated training sets collected from a variety of inventory programs undertaken for purposes besides land cover mapping (primarily, USDA Forest Service Timber Stand Exam and Forest Inventory Analysis programs). Table 1 provides a brief summary of the numbers of polygons, training observations, and land cover types for each map. Figure 1 contrasts percent areal cover across the classified maps with the distri-

bution of training observations for the vegetative land cover types common to all scenes. These boxplots show that the percentages of training observations in the Douglas-fir, lodgepole, and Douglas-fir/lodgepole land cover types are greater than the corresponding areal percent covers, and correspondingly, the percentages of training observations in the grassland land cover types are smaller than the percent covers of these types across the maps. Post-classification sampling was not undertaken because of cost constraints. Additional details regarding mapping and data can be found in Steele et al., (2003) and Redmond et al., (2001).

Estimates of map accuracy (α) were computed using 10-fold cross validation, maximum posterior probability estimators without calibration, and maximum posterior probability estimators with calibration. Three classifiers with substantially different characteristics were used to construct maps which were then assessed for accuracy. The linear discriminant classifier was used because this classifier computes posterior probabilities of group membership by evaluating a multivariate normal model of the covariate vector distribution. The multivariate normal model seems at best to be a rough approximation of the actual covariate distribution. The 5-NN classifier was used as a contrast to the linear discriminant classifier because it is distribution-free, and hence no model is used in estimating the posterior probabilities. On the other hand, the maximum posterior probability estimator derived from the 5-NN classifier is imprecise because (ignoring ties) only four values for a maximum posterior probability estimate are possible: 2/5, 3/5, 4/5

and 5/5. Because of these limitations, both classifiers may perform poorly for maximum posterior probability estimation, though for much different reasons. The third classifier, the exact bagging aggregation (EB) 10-NN+MID classifier (Steele et al., 2003) had been identified as a consistently accurate classifier across the 9 scenes (Redmond et al., 2001). There are two components to this combination classifier. The first component, the exact bagging aggregation 10-NN classifier, is a smoothed version of the conventional 10-NN classifier that weights near neighbors according to their distance in the covariate space to the observation to be classified. The second component, called the mean inverse distance (MID) classifier, uses spatially near training observations to estimate the posterior probability of group membership (Steele, 2000; Steele et al., 2003). It is expected that the maximum posterior probability estimators derived from the combination classifier will be relatively accurate because the classifier is distribution-free, and more precise than the 5-NN classifier because the possible posterior probability values are not restricted to a small set. On the other hand, cross-validation accuracy estimators for this classifier are likely to be biased because the training observations are spatially clustered. Because of clustering, more spatial information is available for classification of the training observations than for other population units.

Estimates of map accuracy are summarized in Figure 2 by accuracy estimator, classifier, and Landsat TM scene. In this Figure, the Landsat TM scenes have been arranged on the horizontal axis in descending order according to the calibrated maximum poste-

rior probability accuracy estimates obtained for the linear discriminant classifier. This Figure shows that the estimates of map accuracy differed substantially among accuracy estimators. For example, among the 9 scene maps constructed using the linear discriminant classifier, differences in accuracy estimates varied from 0.3% to 10.5%, depending on which two accuracy estimators are compared (e.g., cross-validation versus calibrated maximum posterior probability estimators). When the linear discriminant and the 5-NN classifiers were used, the cross-validation estimates were consistently smaller than the maximum posterior probability-based estimators. This difference between accuracy estimates is attributed to non-representative training samples. Specifically, 69.4% of the training observations were in forest types whereas 42.7% of the classified map units were in forest types; moreover, the average of the calibrated maximum posterior probability rates for forest and nonforest classes are 58.3% and 74.7%, respectively. In contrast, the top panel shows that for the EB 10-NN+MID classifier, the calibrated maximum posterior probability estimates are substantially smaller (3.9 to 16.8%) than the estimates from cross-validation and maximum posterior probability without calibration. This difference is attributed to clumpiness of training observation sample locations because spatial aggregation leads to optimistically biased cross-validation accuracy estimates. Because the maximum posterior probability estimators are computed from all map units instead of only the training observations, this source of bias is avoided. When comparing across panels, the cross-validation accuracy estimates suggest rather large differences among

the three classifiers with respect to map accuracy. In contrast, the calibrated maximum posterior probability estimates indicate relatively small differences in accuracy among classifiers. This result demonstrates that the use of a biased accuracy estimator can produce misleading differences when comparing different classifiers or map classifications.

Calibration had a large effect on the map accuracy estimates. Specifically, the average difference between the calibrated and uncalibrated maximum posterior probability estimates was 11.8% and the maximum difference was 16.8%. In summary, these examples show substantial differences among all three accuracy estimators. However, there is little information in these data indicating which method is superior. The following simulation study addresses this question.

7 Simulation study

The simulation study compared the cross-validation, post-classification sample, calibrated maximum posterior probability and post-classification maximum posterior probability estimators. The classifiers were linear discriminant and 5-NN, and the training sample was drawn according to three designs aimed at assessing robustness against non-representative training samples. A sketch of the simulation study is given before discussing the details. Each simulation experiment consisted of generating a simulated population followed by $s = 100$ trials of drawing a training sample (3000 observations)

and post-classification sample (300 observations) from the population. The jth trial, $j = 1, \ldots, s$ produced a classifier η_j. For each of these classifiers, the exact map accuracy was computed by classifying every population unit and noting the proportion of correctly classified units. Then, the training and post-classification samples were used to compute estimates of map accuracy. Finally, bias and accuracy of the estimators were assessed by comparing the exact and estimated map accuracies across the trials. The results of the experiment were summarized for each population by averaging bias and accuracy over the s simulations.

The training sample size of 3000 was selected to be roughly the mean sample size of the nine training samples. The post-classification sample size of 300 was chosen to be an order of magnitude smaller than the training sample to roughly reflect the relative allocation of effort towards each sampling program in a practical mapping exercise. It should be noted that this difference implies that the post-classification estimator standard deviations ought to be approximately $\sqrt{10} \approx 3.16$ times larger than the cross-validation estimator standard deviations. In the simulation experiment, this relative proportion was approximately 2.6. This difference is attributed to additional variance induced by cross-validation, as this resampling algorithm is recognized to yield estimators with relatively large variances (Ripley 1996, chap. 2).

7.1 Evaluating accuracy estimators

An accurate estimator is precise and has small bias. Bias is the difference between the expectation of the estimator and the target parameter, and precision refers to the variability of the estimator about its expectation. Estimator accuracy is quantified by mean square error, and as discussed in section 5.1, both bias and precision contribute to mean square error. While none of these three quantities can be computed exactly unless the distribution of the estimator is known, simulation can be used to compute estimates. This is accomplished by generating a large number of realizations of the estimator, and estimating bias by computing the average difference between the estimates and the target parameter. Mean square error is estimated by the average squared difference between the estimates and the target parameter. Formally, let $\widehat{\alpha}$ denote an estimator of map accuracy α, and suppose that s sampling experiments are conducted under identical conditions. Each experiment, $j = 1, \ldots, s$, generates a target value α_j, the exact accuracy of the map constructed from the jth classifier, and an estimate $\widehat{\alpha}_j$ of accuracy. The bias and mean square error of $\widehat{\alpha}_j$ are $\widehat{\alpha}_j - \alpha_j$ and $(\widehat{\alpha}_j - \alpha_j)^2$, respectively. The bias of $\widehat{\alpha}$ is estimated by the average

$$s^{-1} \sum_{j=1}^{s} (\widehat{\alpha}_j - \alpha_j), \tag{9}$$

and the precision of $\widehat{\alpha}$ is estimated by the root mean square error

$$\mathrm{RMSE}(\widehat{\alpha}) = \sqrt{s^{-1} \sum\nolimits_{j=1}^{s} (\widehat{\alpha}_j - \alpha_j)^2}. \tag{10}$$

In addition to evaluating performance for estimators of α, performance was evaluated for estimators of individual group accuracy (α^g). Performance is reported for four land cover types that characterize important situations with respect to group-specific training sample sizes and group-specific accuracies. The representative land cover types and situations are: low cover grassland (typical accuracies and typical training sample sizes), lodgepole pine (low accuracies and relatively large sample sizes), mixed broadleaf (high accuracies and small sample sizes), and barren (variable accuracy and small sample sizes). This last type was selected in lieu of a type with poor accuracies and small sample sizes, as such types were not present in the original training samples because difficult-to-predict types with small training sample sizes tended to be combined with ecologically similar types before map construction. Though all three sampling designs (random, positively biased, and negatively biased) were used to generate data, results (below) are presented only for the random and positively sampling design when using the linear discriminant function for classification. The other combinations of design and classifier were omitted because the 5-NN classifier results were consistent with results obtained for the linear discriminant classifier, and the positively and negatively biased designs results were similarly consistent.

7.2 Simulation study details

The intention was to generate populations that were realistic with respect to the training sets. In other words, it should be plausible that the training samples might have been produced by sampling the simulated population. One approach is to sample from the classified scene maps. This approach was not pursued because relatively large fractions of the classified map units appear to be misclassified, as indicated by Figure 1, bottom panel. Anticipating the effect of these misclassifications on the simulation results is difficult, and separating these effects from the properties of the estimators is problematic. In addition, an initial investigation showed that training samples drawn from this population yielded accuracy rates that were much larger than the estimated accuracies for the classified scene maps. Thus, instead of sampling the classified scene maps, the training samples were used to generate a population of map units. This was accomplished locating centers of probability mass in the covariate space at each training observation, and generating additional observations about these centers. Specifically, given a training covariate vector $\boldsymbol{t}_i$ with membership in class g, a simulated observation $\boldsymbol{u}$ was created by adding a multivariate normal random vector $\boldsymbol{v}$ to $\boldsymbol{t}_i$, i.e., $\boldsymbol{u} = \boldsymbol{t}_i + \boldsymbol{v}$. The expected value of $\boldsymbol{v}$ was the zero vector, and the variance-covariance matrix of $\boldsymbol{v}$ was $\boldsymbol{S}_g$, where $\boldsymbol{S}_g$ was a diagonal matrix with the sample variances of the class g covariates on the diagonal. This scheme is a rapid method of generating a population much like the training sample, but denser in the covariate space. Two caveats are that $\boldsymbol{S}_g$ only roughly approximates of the

dispersion of class g observations about $\boldsymbol{t}_i$ because the elements of $\boldsymbol{S}_g$ measure dispersion about the group sample mean rather than $\boldsymbol{t}_i$, and because the off-diagonal elements of $\boldsymbol{S}_g$ were set to 0, correlation among covariates was ignored.

Each simulation experiment generated a set $\mathcal{D}$ of $N^* = 203,000$ observations. Approximately $N_g = \pi_g N^*$ observations belonged to class g, where π_g was the proportion of map units in the training sample belonging to class g. In the jth sampling trial, $\mathcal{D}$ was partitioned as a population $\mathcal{P}_j$ of $N = 200,000$ and a training sample $\boldsymbol{X}_j$ of $n = 3,000$ observations. A classifier η_j was constructed from $\boldsymbol{X}_j$, and α_j, the exact accuracy of η_j was computed as the proportion of observations in $\mathcal{P}_j$ correctly classified by η_j. The exact conditional group-specific accuracy rates were computed by determining the proportion of observations in $\mathcal{P}_j$ that were assigned to class g and were correctly classified by η_j. Then, resampling-based accuracy estimates of α were computed using $\boldsymbol{X}_j$. Finally, a post-classification sample $\boldsymbol{X}_j^p$ of 300 observations was drawn by simple random sampling of $\mathcal{P}_j$, and post-classification accuracy estimates were computed by determining the fraction of $\boldsymbol{X}_j^p$ that was correctly classified. Results of the $s = 100$ simulation experiments were summarized by computing the average bias and the root mean square error estimate [formulas (9) and (10)].

Three sampling designs were used to draw the training samples from $\mathcal{D}$. The first design was simple random sampling, a design which yields (nearly) unbiased cross-validation estimators of accuracy. The second and third designs were formulated to assess the bias

of the maximum posterior probability estimators when the training sample is not representative of the population. Specifically, these designs induced bias into the training sample cross-validation estimates. In this context, a positively biased sample design yields $\mathrm{B}[\alpha_{CV}(\boldsymbol{X})] > 0$, and a negatively biased sample design yields $\mathrm{B}[\alpha_{CV}(\boldsymbol{X})] < 0$, respectively. To induce bias, the probability that the observation $\boldsymbol{x}_i \in \mathcal{D}$ was selected as a training observation was determined by its estimated probability of correct classification. Positive bias was induced by preferentially selecting easy-to-classify observations, and negative bias was induced by selecting difficult-to-classify observations. The next paragraph provides details.

Let $\widehat{P}_{i,j} = \widehat{P}_c[\eta_{j-1}(\boldsymbol{t}_i) = y_i]$ denote the calibrated maximum posterior probability estimate of the probability that $\boldsymbol{x}_i \in \mathcal{D}$ will be correctly classified by the classifier η_{j-1}, and $p_{i,j}$ denote the probability that $\boldsymbol{x}_i$ is selected when drawing a training sample from the population for the jth trial of the sampling experiment. For $j = 1$, $p_{i,j}$ was set to n/N^*, for all $i = 1, \ldots, N^*$. For the simple random sampling design, $p_{i,j} = n/N^*$ for every i and j. To induce positive bias, $p_{i,j}$ was set to $p^+_{i,j} = n\widehat{P}^2_{i,j-1} / \sum_{k=1}^{N^*} \widehat{P}^2_{k,j-1}$, for $j > 1$. By using $p^+_{i,j}$ as the probability of including $\boldsymbol{x}_i$ in the sample, observations that are relatively more likely to be correctly classified tended to have inclusion probabilities larger than n/N^* whereas those that are less likely to be correctly classified tended to have inclusion probabilities smaller than n/N^*. Similarly, negative bias was induced by setting the inclusion probabilities to $p^-_{i,j} = n(1 - \widehat{P}_{i,j-1})^2 / \sum_{k=1}^{N^*} (1 - \widehat{P}_{k,j-1})^2$ for $j > 1$.

The true accuracy of the classifiers was also affected to some extent by this scheme; for example, using the $p_{i,j}^{+}$'s tended to increase classifier performance by approximately 5% compared to simple random sampling. When using the biased sampling designs, the p_{ij}'s initially tended to migrate away from n/N^* with j, but showed little trend after $j = 20$ trials.

7.3 Simulation study results

Linear discriminant and 5-NN classifiers produced exact accuracy rates that varied between 49.9 and 63.4% when simple random sampling was used to draw the training sample. Figures 3-7 summarize the results in detail. The linear discriminant classifier was used to produce the results shown in Figures 3 and 4 and the 5-NN classifier was used for Figures 5 and 6. In each Figure, the top row of panels shows estimator behavior when the training sample was drawn by simple random sampling, and the middle and bottom rows shows behavior when the positively bias and negatively biased sampling protocols were used to draw the training sample. In addition, the left column of panels show the results for the cross-validation $\alpha_{CV}(\boldsymbol{X})$ and calibrated maximum posterior probability $\alpha_M(\boldsymbol{X})$ estimators. The right column shows the results for the post-classification estimator $\alpha(\boldsymbol{X}^p)$ and the post-classification calibrated maximum posterior estimator $\overline{\alpha}(\boldsymbol{X}, \boldsymbol{X}^p)$.

Figure 3 shows that under simple random sampling (top panels), there is no evidence of bias in the cross-validation and calibrated maximum posterior probability estimators

of map accuracy, and that the average bias of post-classification estimator and post-classification maximum posterior estimator was generally less than 1%. The left middle panel shows that the positively biased sampling design produced substantial bias in the estimates obtained from the cross-validation estimator (the average bias was 12.2%) while the average bias of the calibrated maximum posterior probability estimator was much less, averaging 1.3% across the simulated populations. The right middle panel shows very little bias in the estimates obtained from either the post-classification estimator or the post-classification maximum posterior estimator. The bottom row of panels shows that under the negatively biased sampling design, the average bias of the calibrated maximum posterior probability estimator is much reduced compared to the cross-validation estimator (1.0 versus -7.3%), and that the average bias of the post-classification estimator was negligible and the average bias of the post-classification maximum posterior estimator was less than 2%.

Figure 4 shows root mean square error estimates of the four estimators of map accuracy. The estimated root mean square errors of calibrated maximum posterior probability estimator and post-classification maximum posterior estimator were consistently less than or equal to the estimates from cross-validation and post-classification sampling, respectively. In addition, when biased sampling designs were used to draw the training sample, the estimates for the calibrated maximum posterior probability estimator were substantially less than those for the cross-validation estimator.

Generally, the performance of the calibrated maximum posterior probability and post-classification maximum posterior estimators with respect to bias reduction was worse when the 5-NN classifier was used compared to the linear discriminant classifier. For example, the average biases associated with calibrated maximum posterior probability estimator were 5.1% (positively biased design) and -2.4% (negatively bias design) compared to 9.3 and -5.6%, respectively for the cross-validation estimator (Figure 5). Figure 6 shows that the root mean square error of the calibrated maximum posterior probability estimator is about half that of the cross-validation estimator under biased sampling, and that differences between the post-classification maximum posterior and the post-classification estimator are minor.

The performance of estimators of user's accuracy (α^g) was evaluated for four land cover types by comparing the root mean square errors of the four estimators. For brevity, results are presented only for the linear discriminant classifier using random and positively biased sampling designs, as results for the other combinations of classifier and design are consistent with these results. Figure 7 summarizes results when a random sampling design was used with the linear discriminant classifier, whereas Figure 8 shows results obtained from the positively biased sampling design and linear discriminant classifier. For each land cover type, estimated root mean square error of the estimators of α^g (%) are plotted against the simulation average of the true user's accuracy. On the horizontal axis, there are 9 distinct values corresponding to the 9 simulated populations.

In total, there are 36 pairs plotted in each panel as there are four accuracy estimators (cross-validation, calibrated maximum posterior probability, post-classification, and post-classification maximum posterior probability). The relationships among the four were fairly consistent among the land cover types. Under random sampling, the estimated root mean square errors of the cross-validation estimator were least, followed by the calibrated maximum posterior, post-classification calibrated maximum posterior probability, and lastly, the post-classification estimator. The root mean square error is a rough approximation of the average difference between the true user's accuracy rate and the estimated user's accuracy rate, hence, these simulation results show that the average error tends to be less than 1.5% for all methods. However, the barren land cover class does provide evidence that the average error occasionally may be substantially larger. Figure 8 shows results when the linear discriminant classifier was used with a positively biased sampling design. In contrast to the random sampling design, root mean square error estimates were generally smallest when the calibrated maximum posterior probability was used. The next best performer with respect to root mean square error was the post-classification maximum posterior probability estimator followed by cross-validation and post-classification. It should be noted though, that there is some variation in the relative ordering. Moreover, the relatively poor performance of the post-classification estimator is entirely a function of sample size as this estimator is unbiased for α^g.

8 Summary

This article has presented evidence that maximum posterior probability estimators of map accuracy are useful for two situations. In the first situation, the training sample are the only data collected in the course of the study, and moreover a statistical sampling design was not used to collect the training sample. In this case, maximum posterior probability estimators were demonstrated to yield map accuracy estimators with substantially less bias and mean square error than cross-validation, a standard method of resampling the training set. Bias reduction is attributed to reducing the role and importance of the training sample as the source of information on accuracy. Information is instead are obtained by estimating the probability of correct classification for every map unit, and computing map accuracy estimates as the mean of the map unit estimates. However, the training sample does have a role as it is used for calibrating the map unit estimates. This use of the training sample has a much reduced impact on bias compared to using the training sample as the sole source of accuracy information. The second situation occurs when a post-classification sample has been collected using a statistically valid sampling design. More precise map accuracy estimators were obtained by averaging the post-classification estimator with the maximum posterior probability estimator derived from the training sample. In general, averaging the post-classification estimator with the calibrated maximum posterior probability estimator is useful provided that the number of training observations is at least several times that of the number of post-classification

observations.

The simulation study showed that the root mean square error estimates for the maximum posterior probability estimator were consistently smaller than or equal to the root mean square error estimates for the cross-validation and post-classification estimator. Specifically, the root mean square error estimates for the calibrated maximum posterior probability estimators of overall map accuracy ranged between 10 and 70% of the root mean square error estimates of the cross-validation and post-classification estimators when the training sample was collected by a biased sampling design. When the design was unbiased (i.e., simple random sampling), then the root mean square error estimates for the calibrated maximum posterior probability and cross-validation estimators were approximately equal, and the root mean square error estimates for the post-classification maximum posterior probability estimator was slightly but consistently smaller than the root mean square error estimates for the post-classification estimator. With respect to estimating user's accuracies for individual land cover types, the performance of the maximum posterior probability estimators was clearly better when the linear discriminant classifier was used with a biased sampling design, but not so when the design was unbiased. In the simulation experiment, positive (negative) bias was introduced by selecting observations that were unusually easy (difficult) to classify. In the case of the 5-NN classifier, training observation posterior probabilities were significantly affected by biased sampling designs because these probabilities were entirely determined from the 5 nearest

neighboring observations and without benefit of a model (unlike the linear discriminant classifier). Calibration was not sufficient to reduce bias to negligible levels.

With any simulation study, the results are anecdotal in nature and dependent on the data used to generate the simulated populations and the simulation design. The simulation study discussed above was designed to generate data unfavorable to efficient posterior probability estimation and thereby test maximum posterior probability estimators under unfavorable conditions. While this study does provide evidence that maximum posterior probability estimators generally improve on cross-validation and post-classification accuracy estimators, it is also apparent that further innovations aimed at reducing maximum posterior probability estimator bias are desirable. Some possibilities currently under investigation are using group-specific calibration functions and nonlinear calibration equations.

Acknowledgments

I thank two anonymous reviewers for extensive and thoughtful comments and suggestions that led to substantial improvements in this article, and the U.S.D.A. Forest Service and the Wildlife Spatial Analysis Laboratory, Montana Wildlife Research Unit at the University of Montana for land cover mapping data. Special thanks go to Dave Patterson for helpful discussions regarding classification and accuracy estimation.

References

Baraldi, A., Bruzzone, L., & Blonda, P. (2005) Quality assessment of classification and cluster maps without ground truth knowledge. *IEEE Transactions on Geoscience and Remote Sensing*, **43**, 857-873.

Bengio, Y. & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation.*Journal of Machine Learning*, **5**, 1089-1105.

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

Efron, B. & Tibshirani, R.J. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, **92**, 548-560.

Foody, G.M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**,185-201.

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

Hammond, T.O. & Verbyla, D.L. (1996). Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing*, **17**, 1261-1266.

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

Huberty, C.J. (1994). *Applied Discriminant Analysis.* Wiley: New York.

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

Nusser, S.M. & Klaas, E.E. (2003). Survey methods for assessing land cover map accuracy. *Ecological and Environmental Statistics*, **10**, 309-332.

Redmond, R.R., Winne, J.C., & Fisher, C. (2001). Existing vegetation and land cover of west central Montana. Wildlife Spatial Analysis Laboratory, MTCWRU, University of Montana, Missoula MT. http://ku.wru.umt.edu/project/silcpage/esides3rpt.pdf.

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

Ross, S. (1998) *A First Course in Probability*, 5th ed. Upper Saddle River, New Jersey: Prentice Hall.

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

Steele, B.M., Patterson, D.A., & Redmond, R.L. (2003). Toward estimation of map accuracy without a probability test sample. *Ecological and Environmental Statistics*, **10**, 333-356.

Steele, B.M. & 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. (2000). Combining multiple classifiers: an application using spatial and remotely sensed information for land cover type mapping. *Remote Sensing of Environment*, **74**, 545-556.

Steele, B.M., Winne, J.C., & 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. (2000). Practical implications of design-based inference for thematic map accuracy assessment, *Remote Sensing of Environment*, **72**, 35-45.

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

Stehman, S.V. & Czaplewski, R.L. (2003). Introduction to the special issue on map accuracy. *Ecological and Environmental Statistics*, **10**, 301-308.

Stehman, S.V., Sohl, T.L., & Loveland, T.R. (2003) Statistical sampling to characterize recent United States land-cover change. *Remote Sensing of Environment*, **86**, 517-529.

Appendix A. Mathematical results

A.1. Claim: the probability that an observation $\boldsymbol{x} = (\boldsymbol{t}, y)$ *is correctly classified by the rule* η *is the maximum posterior probability of group membership as determined by* η*.*

Proof: the posterior probability of membership in group g is the probability that $\boldsymbol{x}$ belongs to group g, given the covariate vector $\boldsymbol{t}$ and η. These probabilities are denoted by $P(y = g \mid \boldsymbol{t})$, $g = 1, \ldots, c$. Thus, the claim can be stated as

$$P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = \max_g P(y = g \mid \boldsymbol{t}).$$

Let E_g denote the event that $\boldsymbol{x}$ belongs to group g and the rule assigns $\boldsymbol{x}$ to group g, i.e., $E_g = \{y = g, \eta(\boldsymbol{t}) = g\}$. Note that E_g and E_h are mutually exclusive for $g \neq h$. Then, the probability that $\eta(\boldsymbol{t})$ is correct is

$$P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = P\left(\cup E_g\right) = \sum_{g=1}^{c} P(E_g).$$

These probabilities are conditional on $\boldsymbol{t}$ and the classifier η. Therefore, the event $\{\eta(\boldsymbol{t}) = g\}$ is not a random event and $P(E_g) = P\left(y = g \mid \boldsymbol{t}\right) \Psi[\eta(\boldsymbol{t}) = g]$. Consequently,

$$P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = \sum_{g=1}^{c} P(y = g \mid \boldsymbol{t}) \Psi[\eta(\boldsymbol{t}) = g].$$

All of the terms $\Psi[\eta(\boldsymbol{t}) = g], g = 1, \ldots, c$, in this sum are 0 except for the indicator of the group, say j, which is predicted by $\eta(\boldsymbol{t})$. Thus

$$P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = P(y = j \mid \boldsymbol{t}).$$

The classifier will predict group j if and only if $P(y = j \mid \boldsymbol{t})$ is the largest posterior probability. Hence, $P(y = j \mid \boldsymbol{t}) = \max P(y = g \mid \boldsymbol{t})$, and

$$P[\eta(\boldsymbol{t}) = y \mid \boldsymbol{t}] = \max_{g} P(y = g \mid \boldsymbol{t}),$$

as claimed.

A.2. Derivation of the calibration coefficient

Suppose that $C = \{(p_1,\psi_1), \ldots, (p_n, \psi_n)\}$ is a calibration sample as discussed in the text, and recall that the least squares estimator of β is obtained by minimizing the error sum of squares with respect to β. In the case of the no-intercept model, the error sum of squares is $S(\beta) = \sum(\psi_i - \beta p_i)^2$. Minimization of $S(\beta)$ yields $\sum p_i\psi_i / \sum p_i^2$ as the estimator of β.

To constrain the fitted model to pass through the pair (c^{-1}, c^{-1}), p_i and ψ_i are transformed so that the no-intercept model (in terms of the transformed variables) conforms to the constraint. The transformed variables are $p_i^* = p_i - c^{-1}$ and $\psi_i^* = \psi_i - c^{-1}$ and the model is $\mathrm{E}(\psi^*) = \beta p^*$. Then, the least squares estimator of the calibration coefficient is

$$\begin{aligned} b &= \frac{\sum p_i^* \psi_i^*}{\sum p_i^{*2}} \\ &= \frac{\sum (p_i - c^{-1})(\psi_i - c^{-1})}{\sum (p_i - c^{-1})^2}. \end{aligned}$$

In addition, the predictive equation $\widehat{\psi}^* = b p^*$ can be expressed as

$$\widehat{\psi} - c^{-1} = b(p - c^{-1})$$

which implies $\widehat{\psi} = bp + c^{-1}(1 - b)$. Because the calibrated maximum posterior probability estimates should be bounded above by 1, the predictive equation [formula (8)] is constrained to prevent being estimates larger than one.

Remark The calibration coefficient β is estimated by the constrained least squares estimator. This estimator is used as it minimizes the sum of the squared residuals, and is

optimal in the sense that no other estimator will yield a smaller sum of squared residuals. Often in regression, it is assumed that the residuals are independent, normal and homoscedastic as these assumptions lead to distributional properties useful for inference. These properties are not needed for the purpose of calibration, and hence, no effort has been made to investigate whether the assumptions might hold.

A.3. Mean square error of a linear combination of estimators

Suppose that the independent random variables $\widehat{\alpha}_1 = \widehat{\alpha}_1(\boldsymbol{X}_1)$ and $\widehat{\alpha}_2 = \widehat{\alpha}_2(\boldsymbol{X}_2)$ are estimators of the parameter α, and that $0 \leq v \leq 1$, is a known constant. Further, suppose that $\widehat{\alpha}_1$ is unbiased for α. The mean square error of $\widehat{\alpha}_3 = \widehat{\alpha}_3(\boldsymbol{X}_1, \boldsymbol{X}_2) = v\widehat{\alpha}_1 + (1-v)\widehat{\alpha}_2$ is

$$\begin{aligned}
\mathrm{MSE}(\widehat{\alpha}_3) &= \mathrm{E}\{[\widehat{\alpha}_3(\boldsymbol{X}) - \alpha]^2\} \\
&= \mathrm{E}\{[v\widehat{\alpha}_1 + (1-v)\widehat{\alpha}_2 - \alpha)]^2\} \\
&= \mathrm{E}\{[v(\widehat{\alpha}_1 - \alpha) + (1-v)(\widehat{\alpha}_2 - \alpha)]^2\} \\
&= v^2\mathrm{E}[(\widehat{\alpha}_1 - \alpha)^2] + 2v(1-v)\mathrm{E}[(\widehat{\alpha}_1 - \alpha)(\widehat{\alpha}_2 - \alpha)] + (1-v)^2\mathrm{E}[(\widehat{\alpha}_2 - \alpha)^2].
\end{aligned}$$

Because $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ are independent, $\mathrm{E}[(\widehat{\alpha}_1 - \alpha)(\widehat{\alpha}_2 - \alpha)] = \mathrm{E}(\widehat{\alpha}_1 - \alpha)\mathrm{E}(\widehat{\alpha}_2 - \alpha)$, and because $\widehat{\alpha}_1$ is assumed to be unbiased, $\mathrm{E}(\widehat{\alpha}_1 - \alpha) = 0$. Thus,

$$\begin{aligned}
\mathrm{MSE}(\widehat{\alpha}_3) &= v^2\mathrm{E}[(\widehat{\alpha}_1 - \alpha)^2] + (1-v)^2\mathrm{E}[(\widehat{\alpha}_2 - \alpha)^2] \\
&= v^2\mathrm{MSE}(\widehat{\alpha}_1) + (1-v)^2\mathrm{MSE}[(\widehat{\alpha}_2).
\end{aligned}$$

The value of v that minimizes $\mathrm{MSE}(\widehat{\alpha}_3)$ is determined by differentiating $\mathrm{MSE}(\widehat{\alpha}_3)$ with respect to v and setting the derivative equal to zero and solving for v. Differentiation yields

$$0 = 2v\mathrm{MSE}(\widehat{\alpha}_1) - 2(1-v)\mathrm{MSE}[(\widehat{\alpha}_2).$$

Solving this equation yields the minimizing value v^* =$\mathrm{MSE}[(\widehat{\alpha}_2)/\{\mathrm{MSE}(\widehat{\alpha}_1)+\mathrm{MSE}[(\widehat{\alpha}_2)\}$.

Tables

Table 1. Summary of the classified scene maps. The number of map units is denoted by N, the number of observations in the training sample is denoted by n, and the number of classes is c. Additional details can be found in Steele et al., (2003).

Scene	N	n	c
P37/R29	$567,457$	2052	17
P38/R27	$592,439$	2462	17
P38/R28	$622,080$	3749	17
P38/R29	$521,981$	1550	16
P39/R27	$480,916$	2446	17
P39/R28	$666,092$	4242	18
P39/R29	$569,595$	1728	14
P40/R27	$674,331$	2995	19
P40/R28	$727,864$	3013	16

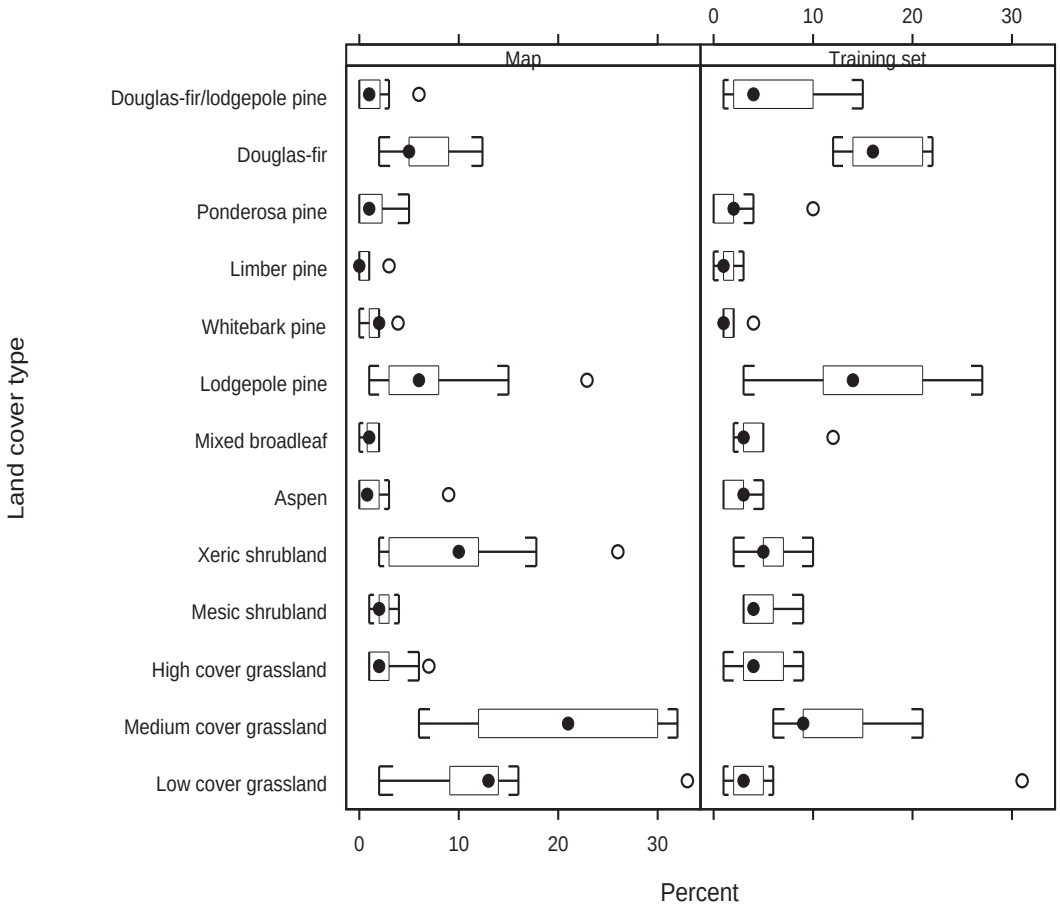

Figure 1: Box plots of percent map area coverage (left panel) and percentage of training observations (right panel) for each of the important vegetative land cover types across the study area. Each box plot was constructed from 9 values for percent cover within a Landsat TM scene map (left panel), or training sample size for a Landsat TM scene (right panel).

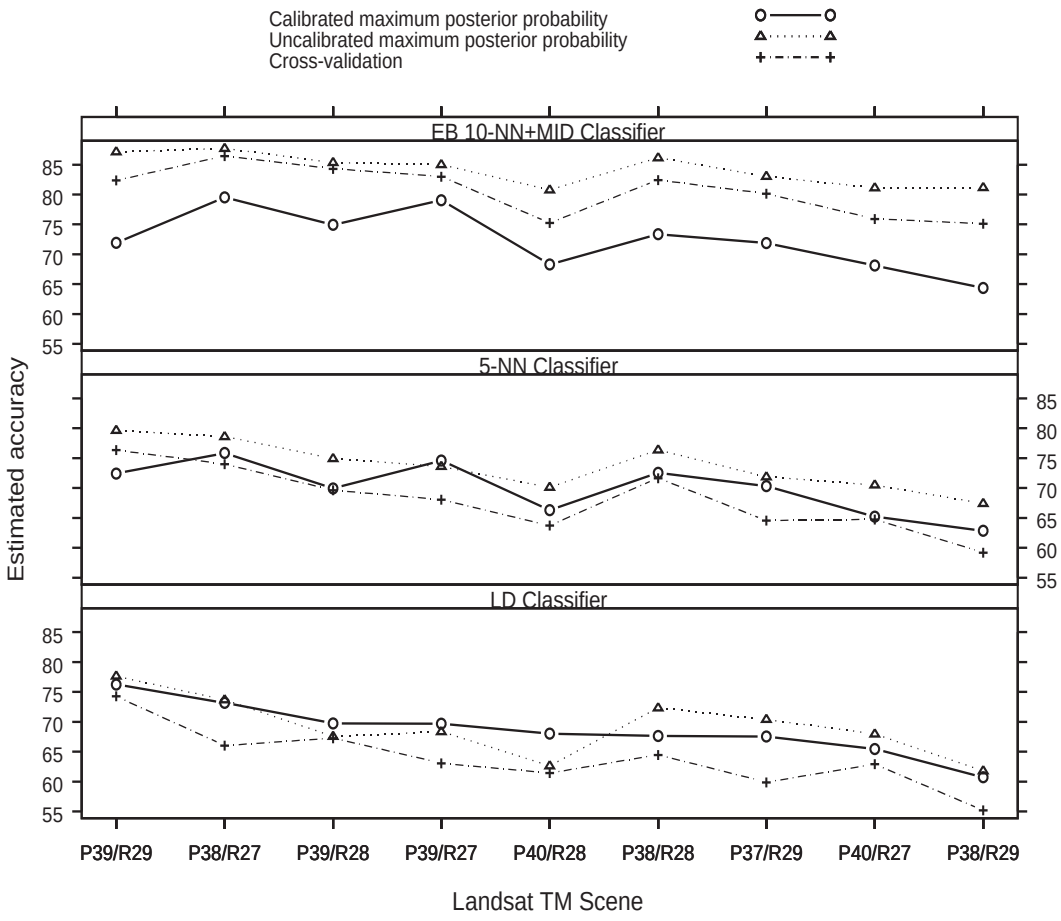

Figure 2: Estimates of map accuracy (%) plotted against Landsat TM scene. Landsat TM scenes are identified by path and row. Three accuracy estimators are shown: 10-fold cross-validation (denoted as $\alpha_{CV}(\boldsymbol{X})$ in the text), the uncalibrated maximum posterior probability estimator, and the calibrated maximum posterior probability estimator (denoted by $\alpha_M(\boldsymbol{X})$ in the text). Each panel shows the results for a different classifier. Note: the Landsat TM scenes have been arranged along the horizontal axis so that the accuracy estimates are in generally descending order when reading from left to right.

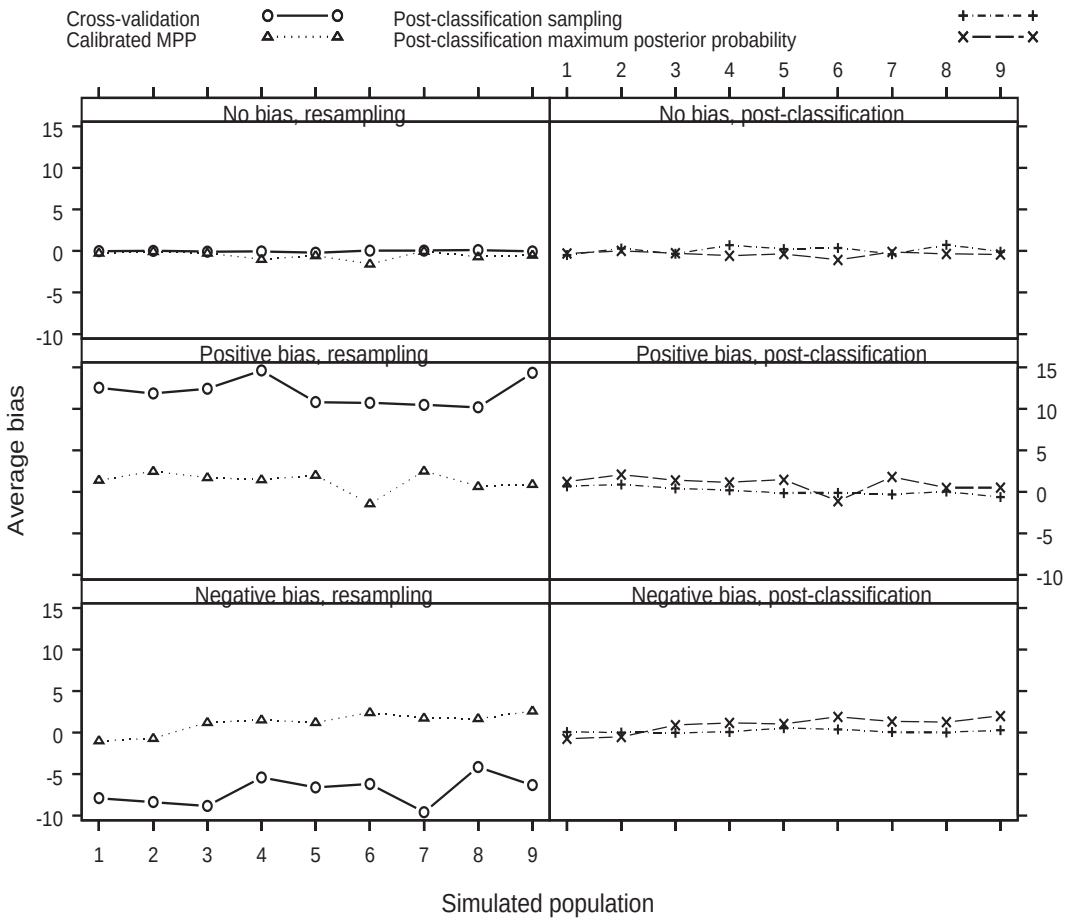

Figure 3: Average bias of four estimators of map accuracy (%) plotted by simulated population when using the linear discriminant classifier. The estimators are cross-validation $\alpha_{CV}(\boldsymbol{X})$, calibrated maximum posterior probability $\alpha_M(\boldsymbol{X})$, post classification sample $\alpha(\boldsymbol{X}^p)$, and post-classification maximum posterior probability $\overline{\alpha}(\boldsymbol{X}, \boldsymbol{X}^p)$. The top row of panels show average bias when the training sample was collected by simple random sampling; the middle row of panels show average bias when the training sample was collected using a positively biased sampling design, and the bottom row of panels show average bias when using a negatively biased design.

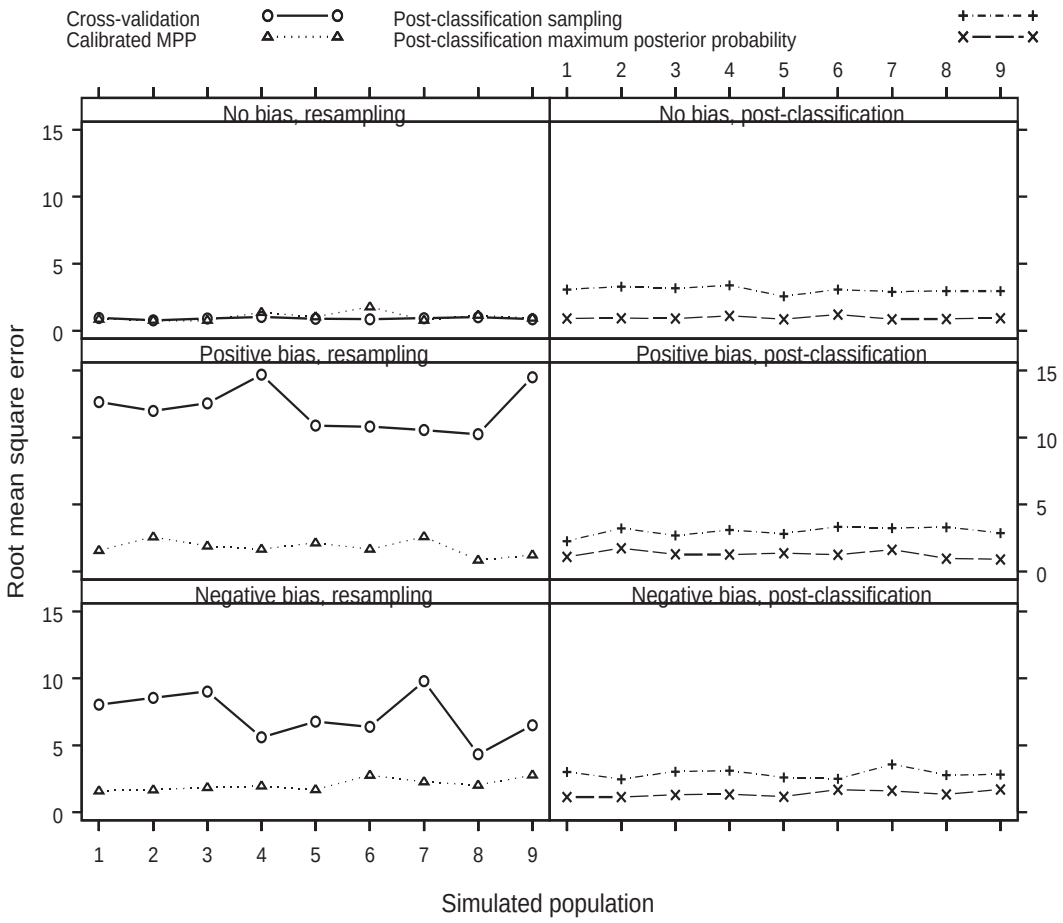

Figure 4: Estimates of root mean square error for estimators of map accuracy (%) plotted by simulated population when using the linear discriminant classifier. The legend for Figure 3 gives details.

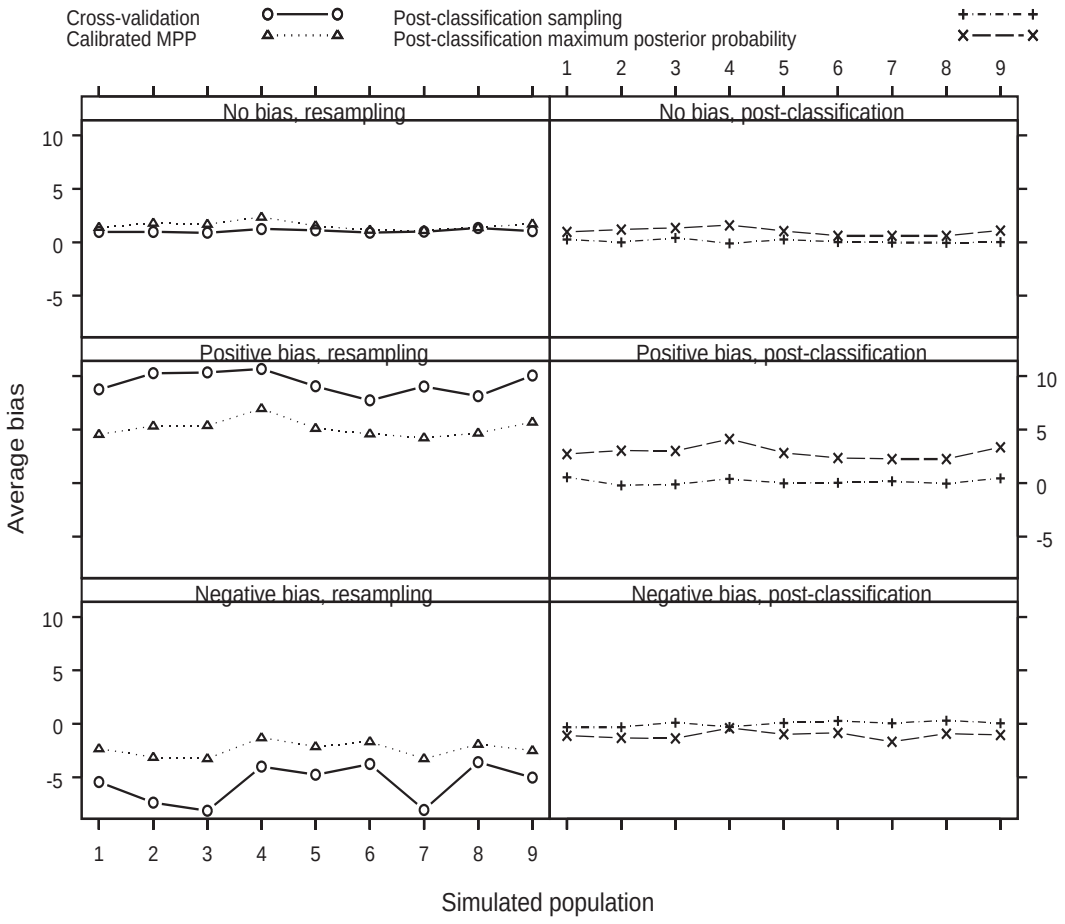

Figure 5: Average bias of the estimators of map accuracy (%) plotted by simulated population when using the 5-NN classifier. The legend for Figure 3 gives details.

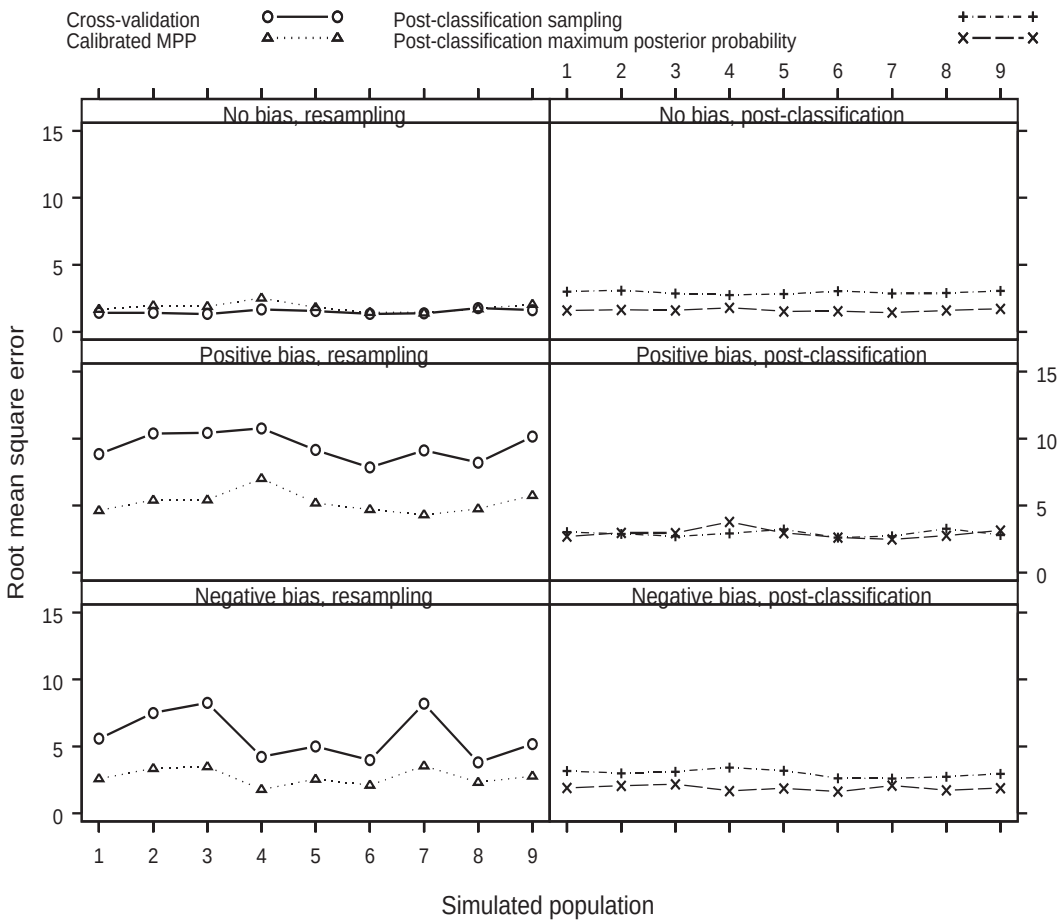

Figure 6: Estimates of root mean square error for estimators of map accuracy (%) plotted by simulated population when using the 5-NN classifier. The legend for Figure 3 gives details.

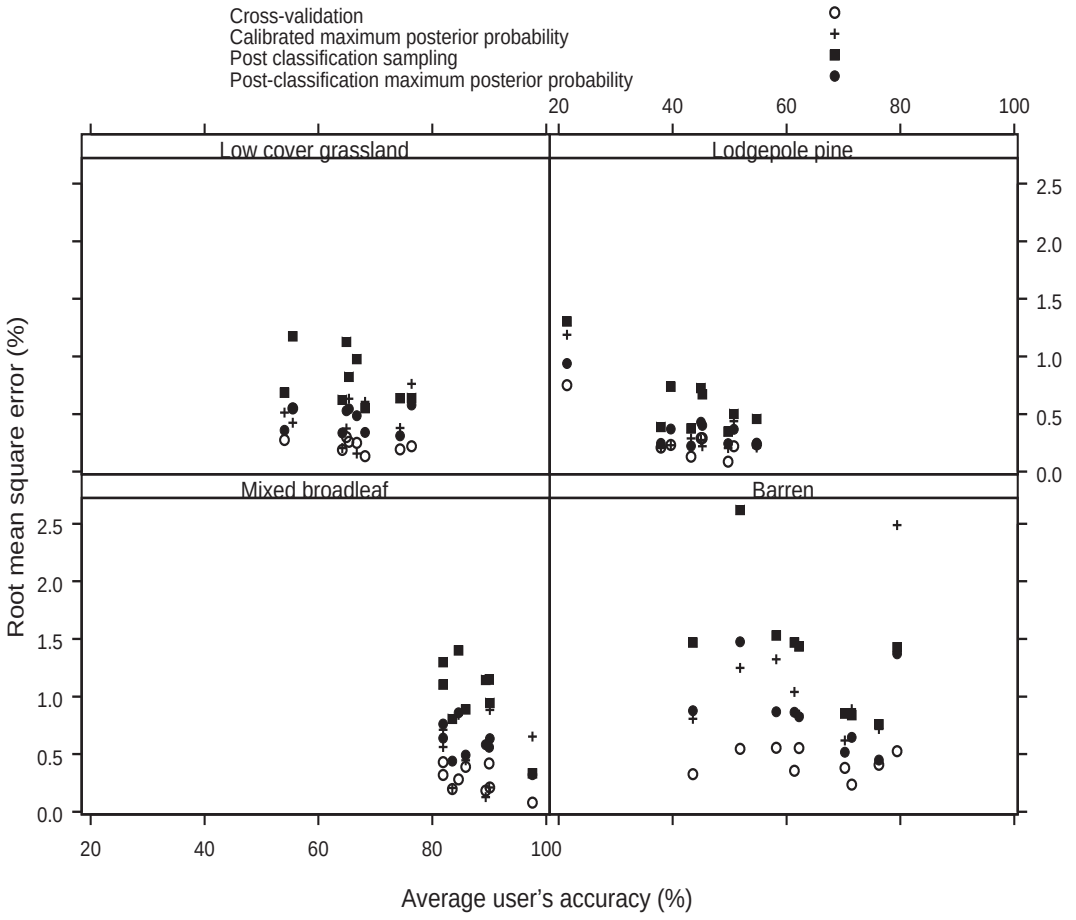

Figure 7: Estimates of root mean square error for user's accuracy estimators (%) plotted against average user's accuracy (%). Values were obtained for 9 simulated populations using the linear discriminant classifier and training samples drawn by random sampling. Results for four land cover types with different characteristics (see text for details) are shown.

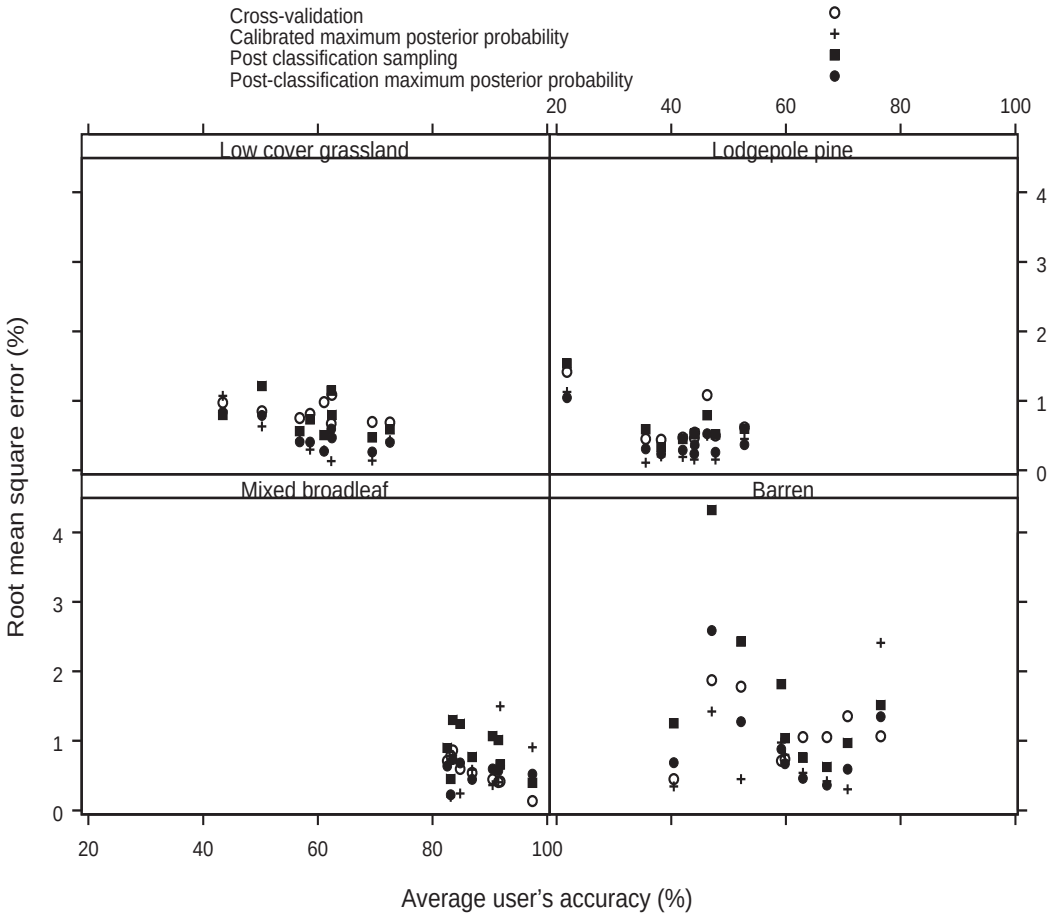

Figure 8: Estimates of root mean square error for user's accuracy estimators (%) plotted against average user's accuracy (%). Values were obtained for 9 simulated populations using the linear discriminant classifier and training samples drawn by a biased sampling plan. Results for four land cover types with different characteristics (see text for details) are shown.