

# Estimating Map Accuracy without a Spatially Representative Training Sample

David A. Patterson, Mathematical Sciences, The University of Montana-Missoula

Brian M. Steele, Mathematical Sciences, The University of Montana-Missoula

Roland L. Redmond, Forestry, The University of Montana-Missoula

## Abstract

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. Land cover maps covering millions of acres consisting of millions of units are often constructed from satellite remotely-sensed data. A classification rule is constructed from a training sample of ground-truthed map units. Because of the expense of collecting a spatially representative training sample for such a large map, the training sample is often drawn from a variety of existing data collected for purposes other than mapping land cover. The spatial distribution of the training sample tends therefore to be highly irregular. It is crucial to estimate the accuracy of the resulting map both overall and on a smaller scale since accuracy may vary spatially and by land cover type. Traditional methods of assessing accuracy, such as cross-validation, may be biased because of the spatial irregularity of the training sample if the classification rule uses spatial information. To reduce bias, we suggest methods of estimating overall map accuracy and unit-by-unit accuracy by using calibrated estimates of the posterior probability of correct classification for each map unit.

## 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. 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 determine actual land cover by ground inspection 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 essential for interpretation of the results. A standard approach is to use cross-validation or bootstrapping to assess map accuracy from the same training set used to classify the imagery (Efron and Tibshirani, 1997).

The work in this paper was motivated by a mapping project covering 21.5 million hectares of forested mountains and rangeland within the northern Rocky Mountains. It was initiated by the USDA Forest Service, Northern Region. The intended uses were primarily natural resources management, include timber, range, water, fish, and wildlife. Figure 1 shows 9 Landsat Thematic Mapper (TM) scenes which were the focus of this study, each of which was classified separately. Each scene consisted of over 30 million 30 m.<sup>2</sup> pixels. Through preprocessing using an unsupervised classifier, the scenes were each segmented into from 480,000 to 730,000 homogeneous polygons, ranging in size from a single pixel to 202 hectares. The polygons were then classified into 14 to 19 land cover classes using a supervised classifier based on the following covariates: spectral reflectance intensity for TM channels 1-5 and 7, elevation, slope, a measure of solar insolation, and a vegetation index. We also used a classifier which incorporated spatial information (discussed in Section 2) and a comparison of the non-spatial and spatial

classifiers was an important component of this project. The training sets were assembled from a variety of existing sources unrelated to the project, principally USDA Forest Service Timber Stand Exam and Forest Inventory Analysis programs. As such, none of the training sets, which consisted of about 1500 to 4300 polygons per scene, could be viewed in aggregate as a probability sample from an entire scene. Moreover, the spatial distribution of the training observations over each scene was irregular, largely because there are privately-held regions that were not sampled because of lack of accessibility. This is illustrated in Figure 2, which shows a portion of one scene (classified with two different classifiers as discussed later), with the black dots representing the training sample. Probability sampling was not conducted because of the cost and logistics of sampling such a large area. We suspect that the training sample points are not unrepresentative, except spatially, of the entire scene and, therefore, one is tempted to act as if the training sample is a random sample and to assess the accuracy of a classifier by cross-validation or bootstrap in the usual way. However, there is a large potential bias in using this approach, particularly with a classifier which incorporates spatial information. For example, if regions or land cover classes that are homogeneous or 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 heterogeneous or difficult-to-classify regions or land cover classes, then accuracy estimates may be pessimistically biased. Our goal, therefore, was to develop techniques for better assessing the accuracy of classifiers, both spatial and non-spatial, from non-probability training samples.

The classifiers used in this project are discussed in Section 2, the proposed method of accuracy estimation is discussed in Section 3, the results are presented in Section 4 and discussed in Section 5.

## 2. Classifiers

We have studied the performance of a variety of non-spatial classifiers in our land cover classification problems. We have found  $k$ -nearest neighbor ( $k$ -NN) classifiers to consistently have the highest 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). Reasons for the superiority of  $k$ -NN methods may be that the satellite measurements on reflectance for the different bands of the electromagnetic spectrum are measured on a common scale and that the bands tend to have about equal value for classification. Consequently, the  $k$ -NN metric (Euclidean distance) is an effective measure of similarity between observations. In particular, we use a “bagged” version of the  $k$ -NN classifier. Before discussing the bagged  $k$ -NN classifier, we first introduce some notation.

Let  $\mathbf{x}_1, \dots, \mathbf{x}_n$  denote the training of size  $n$  collected by sampling a population of  $N$  elements. Each element belongs to one of  $c$  classes, or groups. The land cover class of  $\mathbf{x}_i$  is denoted by  $y_i$ . In our study,  $\mathbf{x}_i$  is the vector of remotely sensed spectral reflectances and terrain variables, including geographic location. The population is the entire set of map polygons in a scene. The value of the covariate vector  $\mathbf{x}$  is known for all polygons in the scene, while the land cover class  $y$  is known only for observations in the training sample. The conditional probability that a

random observation  $\mathbf{x}$  belongs to class (or group)  $g$ , given  $\mathbf{x}$ , is denoted by  $P_g(\mathbf{x}) = \Pr[y = g | \mathbf{x}]$ . A classifier can be viewed as an estimator of  $[P_1(\mathbf{x}), \dots, P_c(\mathbf{x})]$  which assigns  $\mathbf{x}$  to the class with the largest posterior probability estimate.

The estimator of  $P_g(\mathbf{x})$  produced by a  $k$ -NN Euclidean distance classifier is the sample proportion of the  $k$  nearest neighbors belonging to group  $g$ . More formally, let  $y_{[j]}$  denote the class label of the  $j^{\text{th}}$  closest neighbor of  $\mathbf{x}$ . Then, the  $k$ -NN estimate of  $P_g(\mathbf{x})$  is

$$\hat{P}_g^{k\text{NN}}(\mathbf{x}) = k^{-1} \sum_{j=1}^k \Psi(y_{[j]} = g) \quad (1)$$

where  $\Psi(E)$  is the indicator of the event  $E$ .

The purpose of bagging (Breiman 1996) a classifier is to reduce the variance from training sample to training sample of the class membership probability estimates 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. The desired classifier is constructed from the bootstrap training sample and applied to an unclassified observation  $\mathbf{x}$  to give a vector of class membership probability estimates. This process is repeated  $B$  times and the  $B$  vectors are averaged to give the bagged posterior probability estimates that  $\mathbf{x}$  comes from each of the classes. The bagging classifier classifies  $\mathbf{x}$  into the class with the highest bagged estimated posterior probability.

Bagging by repeatedly drawing bootstrap samples is a Monte Carlo approximation to the exact bagging estimate defined by  $E_{\hat{F}}[\hat{P}_g^*(\mathbf{x})]$  where  $\hat{F}$  is the empirical distribution function placing probability  $n^{-1}$  at each  $\mathbf{x}_i$  in the training sample, and  $\hat{P}_g^*(\mathbf{x})$  is an estimator of  $P_g(\mathbf{x})$  obtained from a random sample from  $\hat{F}$ . The Monte Carlo approximation is used because the exact bagging estimate is almost always too intractable to calculate. However, with the  $k$ -NN classifier, the exact bagging estimate can be computed analytically. The exact bagging estimate of the posterior probability of membership in class  $g$  for the  $k$ -NN classifier is

$$\hat{P}_g^{\text{EB } k\text{NN}}(\mathbf{x}) = k^{-1} \sum_{j=1}^n w_j \Psi(y_{[j]} = g) \quad (3)$$

where

$$w_j = n^{-n} \sum_{h=1}^k \sum_{a=0}^{h-1} \sum_{b=h-a}^{n-a} \binom{n}{a} (j-1)^a \binom{n-a}{b} (n-j)^{n-a-b}.$$

The details of this derivation are in Steele and Patterson (2000) who developed the exact bagging  $k$ -NN (hereafter denoted EB  $k$ -NN) as a smoothed version of the ordinary  $k$ -NN classifier. A comparison of equations (1) and (3) shows how bagging smoothes the estimated posterior probabilities.

We did not initially incorporate spatial information into the EB  $k$ -NN classifier. There is a large amount of positive spatial correlation when the data are considered on the pixel level since adjacent pixels are apt to be similar. However, when the data are considered at the polygon level the amount of spatial information is considerably reduced because image segmentation tends to produce polygon boundaries that coincide with changes 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 classes at a larger scale. For example, the distribution of land cover classes differs between lee and windward sides of a mountain range. Steele (2000) has developed an approach to extracting spatial information from the relative abundance and proximity of training observations in the vicinity of an unclassified polygon. His approach is motivated by the application of Bayes rule. The Bayes rule assigns  $\mathbf{x}$  to the class with the largest posterior probability

$$P(y = g | \mathbf{x}) = \frac{\pi_g f(\mathbf{x} | y = g)}{\sum_{j=1}^c \pi_j f(\mathbf{x} | y = j)}$$

where  $f(\mathbf{x} | y = g)$  is the probability density function and  $\pi_g$  is the prior probability for group  $g$ . The  $k$ -NN classifier (equation 1) is a plug-in version of the Bayes rule which assumes equal priors ( $\pi_g = c^{-1}$ ,  $g = 1, \dots, c$ ) and which estimates  $f(\mathbf{x} | y = g)$  by the proportion of the  $k$  nearest neighbors of  $\mathbf{x}$  which are from group  $g$ . Similarly, the EB  $k$ -NN classifier assumes equal priors and estimates  $f(\mathbf{x} | y = g)$  by a weighted proportion of the neighbors of  $\mathbf{x}$  which are from group  $g$ . The approach of Steele (2000) is to incorporate spatial information into the prior probabilities  $\pi_g$ ,  $g = 1, \dots, c$ . These prior probabilities are different for each polygon to be classified and depend on the relative frequency of the  $c$  classes in a spatial neighborhood about the geographic location of  $\mathbf{x}$ . Steele refers to these as local prior class probabilities, denoted  $\pi_g(\mathbf{x})$ ,  $g = 1, \dots, c$ . In this study, the local class priors are estimated using Steele and Patterson's (in press) mean inverse distance (MID) estimator. We define the mean inverse distance from an observation  $\mathbf{x}$  to group  $g$  to be

$$\bar{d}_g(\mathbf{x}) = \frac{1}{n_g} \sum_{i=1}^n \Psi(y_i = g) D^{-2}(\mathbf{x}, \mathbf{x}_i)$$

where  $D^2(\mathbf{x}, \mathbf{x}_i)$  is the squared Euclidean distance between the geographic locations of  $\mathbf{x}$  and  $\mathbf{x}_i$  and  $n_g$  is the number of training observations from group  $g$ . The MID estimated prior probability that  $\mathbf{x}$  belongs to group  $g$  is the normalized mean inverse distance to class  $g$  given by

$$\hat{\pi}_g(\mathbf{x}) = \frac{\bar{d}_g(\mathbf{x})}{\sum_{j=1}^c \bar{d}_j(\mathbf{x})}$$

The MID estimates may be combined with any conventional classifier. In particular, the  $k$ -NN+MID estimator of the posterior probability that  $\mathbf{x}$  belongs to group  $g$  is

**Table 1.** Area-weighted accuracy estimates from  $n$ -fold cross-validation (in percent)

Scene	EB 10-NN	EB 10-NN+MID
3729	68.0	81.4
3827	81.3	90.4
3828	76.4	82.5
3829	64.7	78.0
3927	77.3	86.2
3928	75.3	86.5
4027	71.3	78.2

$$\hat{p}^{\text{EB } k\text{NN}+\text{MID}}(\mathbf{x}) = \frac{\hat{\pi}_g(\mathbf{x})E_{\hat{F}}[\hat{P}_g^{k\text{NN}}(\mathbf{x})]}{\sum_{j=1}^c \hat{\pi}_j(\mathbf{x})E_{\hat{F}}[\hat{P}_j^{k\text{NN}}(\mathbf{x})]}. \quad (4)$$

The EB  $k$ -NN+MID classifier classifies  $\mathbf{x}$  into the class with the highest posterior probability.

The EB  $k$ -NN and the EB  $k$ -NN+MID with  $k = 10$  were compared on seven scenes. Figure 2 shows a comparison of the resulting land cover maps for a portion of one scene. Accuracy rates as estimated by  $n$ -fold cross-validation on the training sample are reported in Table 1. Because the polygons are different sizes, all accuracy rates are area-weighted, that is, they report the percentage of the total area of the polygons in the training sample correctly classified by  $n$ -fold CV. A comparison of these percentages reveals an apparently large improvement in accuracy when the spatial information is incorporated into the EB 10-NN rule. The improvements range from about 6 percentage points to over 13 percentage points.

While some improvement in accuracy rates is expected with the addition of spatial information, the size of these differences is startling, since the amount of spatial information at the polygon level seems limited. We suspect that cross-validation is inflating the accuracy rates for the EB  $k$ -NN+MID classifier because of the spatial clustering in the training samples. The amount of spatial information available for a training sample point tends to be greater than the amount of spatial information available for an arbitrary polygon in the population because training sample points tend to be near each other (see Figure 2). Figure 3 illustrates this for scene 3928; the distance to the nearest (other) training sample point tends to be much smaller for training sample points than for polygons in the entire scene.

The potential bias in using CV (or bootstrap resampling) to estimate accuracy rates with a spatially clustered training sample led us to use the estimated posterior probabilities (equations 3 and 4) to estimate the probability of correct classification for each polygon in the population. The maximum group posterior probability for an observation is an estimate of its probability of correct classification. These can be averaged over all polygons in the scene (weighted by area) to give an overall accuracy estimate. For the EB  $k$ -NN+MID classifier, the maximum posterior probabilities would reflect the lesser amount of spatial information available for the polygons not

in the training sample and thus might lead to better estimates of the overall accuracy rate. In fact, a comparison of the maximum group posterior probabilities for the training sample and for the population for one scene (Figure 4) indicates that they reflect this difference.

Using the estimated posterior probabilities for accuracy estimation is discussed by Ripley (1996, Chap. 2) and McLachlan (1992, Chap. 2). Although the maximum group posterior probability may be severely biased as an estimate of the probability of correct classification for an observation in the training sample, the bias is expected to be small for an observation not from the training sample, according to Ripley (1996). Ripley suggests calibrating the maximum estimated posterior probabilities to reduce bias. Calibration is addressed in the next section.

### 3. Calibration of the estimated probabilities of correct classification

Cox (1958) and McLachlan and Basford (1988, Chap. 5), among others, discuss calibration methods for reducing bias of estimators of the probability of an event. We propose constructing calibration functions from the training set to reduce bias in the estimated probabilities of correct classification in the following way. Let  $Z_i$  be a 0/1 indicator of whether the  $i^{\text{th}}$  training observation  $\mathbf{x}_i$  is correctly classified by  $n$ -fold cross-validation using whichever classifier is being evaluated. Let  $\hat{P}_{\max}(\mathbf{x}_i) = \max_{g=1,\dots,c} \hat{P}_g(\mathbf{x}_i)$  be the maximum group estimated posterior probability for  $\mathbf{x}_i$ , estimated by leaving  $\mathbf{x}_i$  out of the training sample. The calibration function is estimated by regressing  $Z_i$  on  $\hat{P}_{\max}(\mathbf{x}_i)$ ,  $i = 1, \dots, c$ . We tried several different methods, including linear and logistic regression, among others. Results reported here are for a simple linear calibration model with no intercept, fit by least squares (the particular calibration model did not make a great deal of difference in the results). The linearly calibrated probability of correct classification for any observation  $\mathbf{x}$  in the population is

$$\hat{P}^{\text{Lin}}(\mathbf{x}) = \begin{cases} \hat{\beta} \hat{P}_{\max}(\mathbf{x}), & \text{if } \hat{\beta} \hat{P}_{\max}(\mathbf{x}) < 1 \\ 1, & \text{if } \hat{\beta} \hat{P}_{\max}(\mathbf{x}) \geq 1 \end{cases}$$

where  $\hat{\beta}$  is the estimated slope coefficient from the least squares fit. These calibrated probabilities can be averaged over all polygons in the scene to estimate the overall accuracy rate.

### 4. Results

Table 2 summarizes the results for seven scenes. The  $n$ -fold cross-validation results from Table 1 are also included for comparison. It is apparent that using the uncalibrated posterior probability estimate of accuracy makes a bigger difference for the spatial classifier (EB 10-NN+MID) than for the non-spatial classifier (EB 10-NN). This is consistent with our presumption that the  $n$ -fold CV estimate is optimistically biased for the spatial classifier because of the spatial clustering in the training sample. In addition, calibration has a bigger effect for the spatial classifier than for the non-spatial one. The end result is that the difference in estimated accuracy between the two classifiers ranges from 0.3 to 4.5 percentage points using the linearly calibrated estimates as compared to 6.1 to 13.4 percentage points using the  $n$ -fold CV estimates. The former are more consistent with our a priori feeling that there is some value in the spatial

**Table 2.** Area-weighted overall accuracy estimates (percentage) for the two classifiers. “CV” is  $n$ -fold cross-validation on the training sample, “Uncal.” is the mean (over all polygons in the scene) uncalibrated estimated probability of correct classification, and “Lin. Cal.” is the mean linearly calibrated estimated probability of correct classification.

Scene	$N$	$n$	EB 10-NN			EB 10-NN + MID		
			CV	Uncal.	Lin.Cal.	CV	Uncal.	Lin.Cal.
3729	567457	2052	68.0	70.6	71.1	81.4	75.6	72.8
3827	592439	2462	81.3	76.2	76.7	90.4	81.5	80.3
3828	622080	3749	76.4	70.7	71.7	82.5	77.1	73.8
3829	521981	1550	64.7	60.8	61.4	78.0	68.2	63.4
3927	480916	2446	77.3	76.6	77.3	86.2	82.3	80.1
3928	666092	4242	75.3	70.3	71.5	86.5	77.0	76.0
4027	674331	2995	71.3	70.8	71.7	78.2	77.1	72.0

information, but not as much as indicated by the differences in the  $n$ -fold CV estimates of accuracy.

## 5. Discussion and conclusions

The calibrated estimates of accuracy based on estimated posterior probabilities appear to more accurately reflect the true accuracy of the spatial and non-spatial classifiers than do the  $n$ -fold CV estimates. We have no way of verifying this since we don’t have independent test samples, but the estimates are consistent with this hypothesis. The procedure implicitly assumes that the training sample is representative of the population of polygons as a whole, except spatially. More work, including simulations, will be needed to see under what conditions this procedure works.

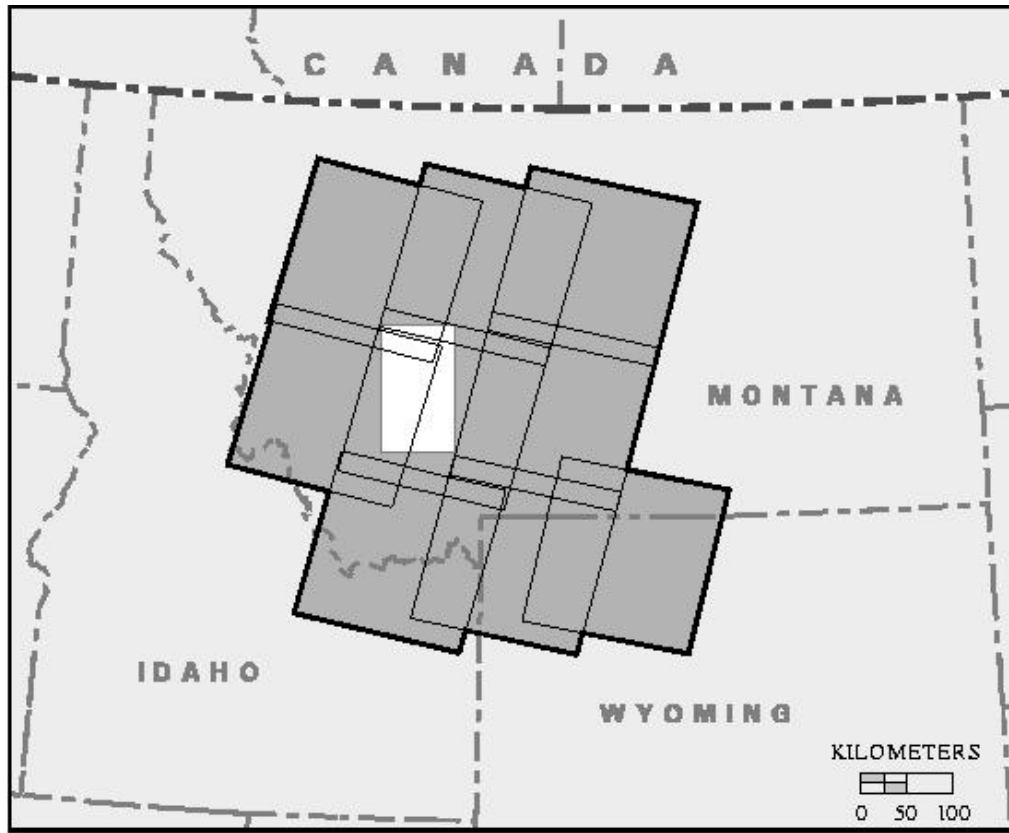
Using polygon based posterior probability estimates of accuracy has another advantage: they can be plotted polygon by polygon to give an accuracy map. This is not possible with cross-validation methods of accuracy assessment, except with considerable spatial smoothing. An example of such a map is displayed in Figure 5. Polygon by polygon estimates of accuracy are of considerable value to users working with a small section of the map, for whom the overall accuracy rate is less relevant.

Finally, calibration may be a valuable tool to use even when a post-classification random test sample is available for accuracy estimation. The post-classification sample could be used to calibrate the posterior probability estimates of correct classification to yield, again, accuracy estimates on an individual polygon level, and an improved estimate of overall accuracy. It would be a particularly valuable tool when only a small post-classification sample was possible.

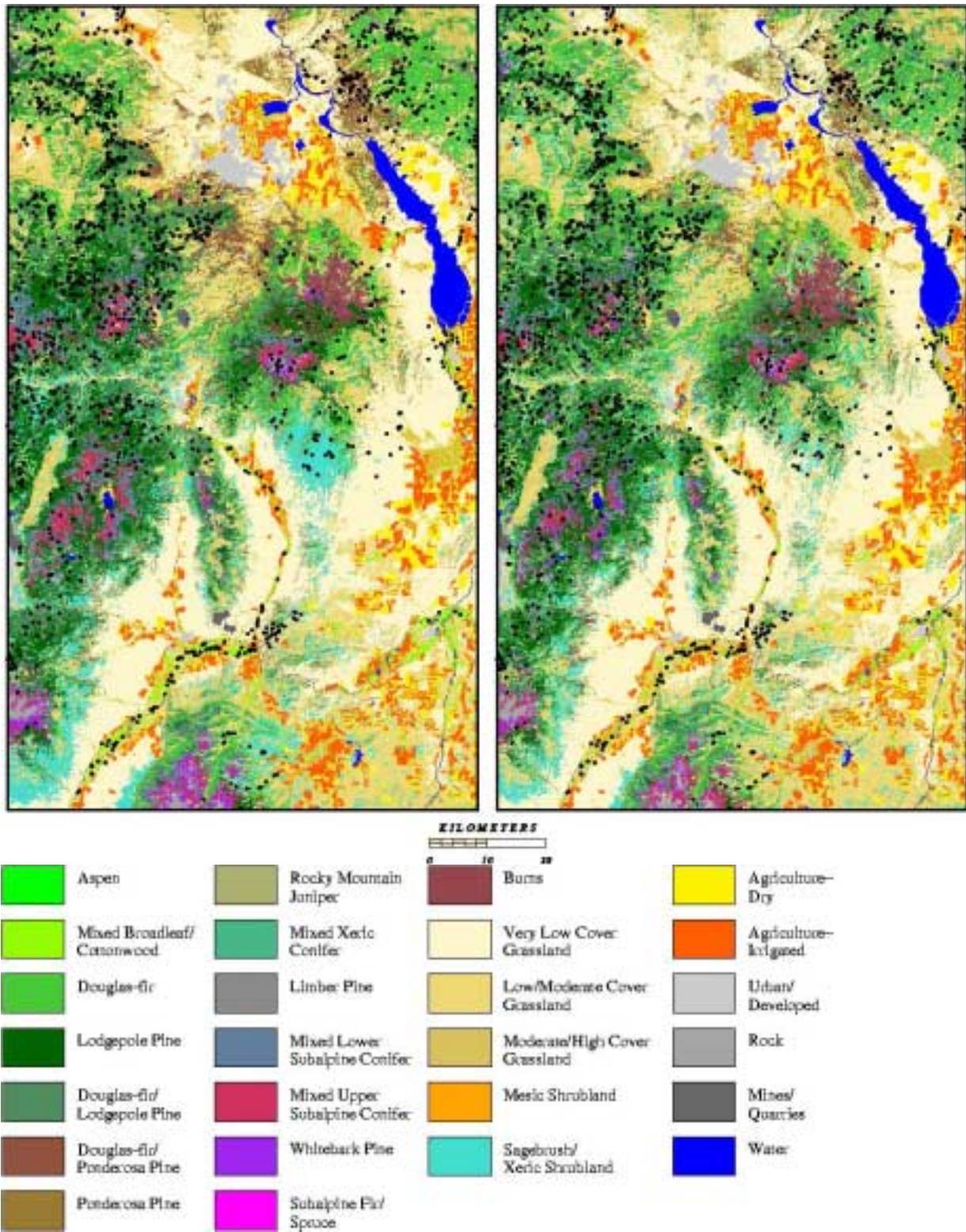
## References

- 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**, 123-140.

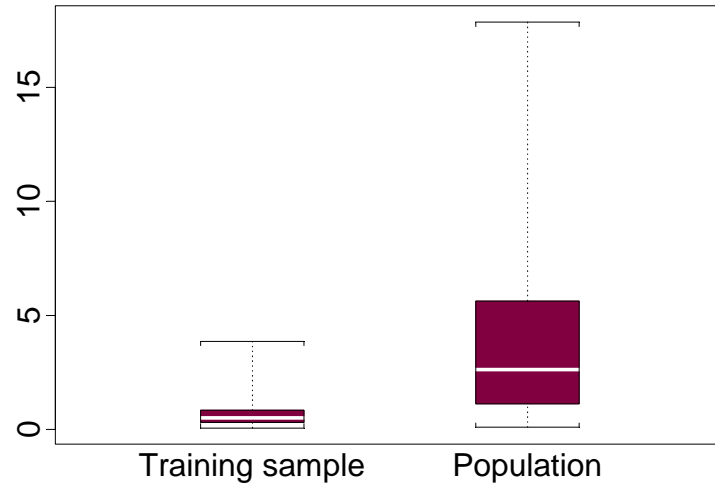
- Cox, D.R. (1958) Two further applications of a model for binary regression. *Biometrika*, **45**, 562-5.
- Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- Hammond, T.O. and Verbyla, D.L. (1996) Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing*, **17**, 1261-66.
- McLachlan, G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.
- Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.
- 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**, 349-355.
- Steele, B.M. and Patterson, D.A. (in press) Land cover mapping using combination and ensemble classifiers. *Proceedings of the 33<sup>rd</sup> Symposium on the Interface*. Interface Foundation of North America, Fairfax Station, VA.



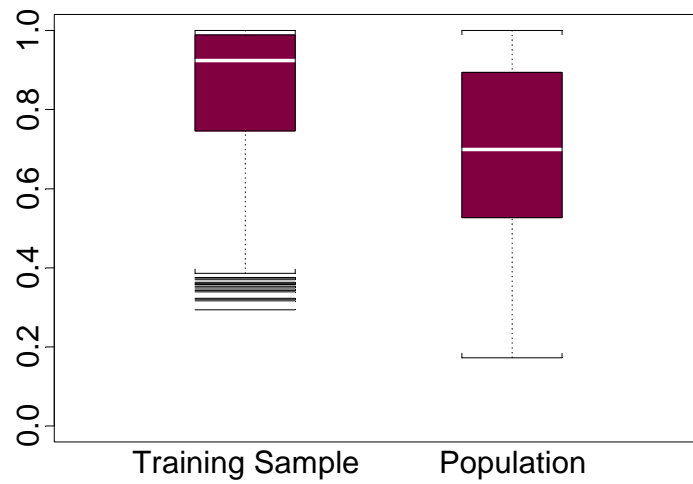
**Figure 1.** The nine Landsat TM scenes which together constitute the 21.5 million hectare study area. The white inset box shows the area mapped in Figure 2.



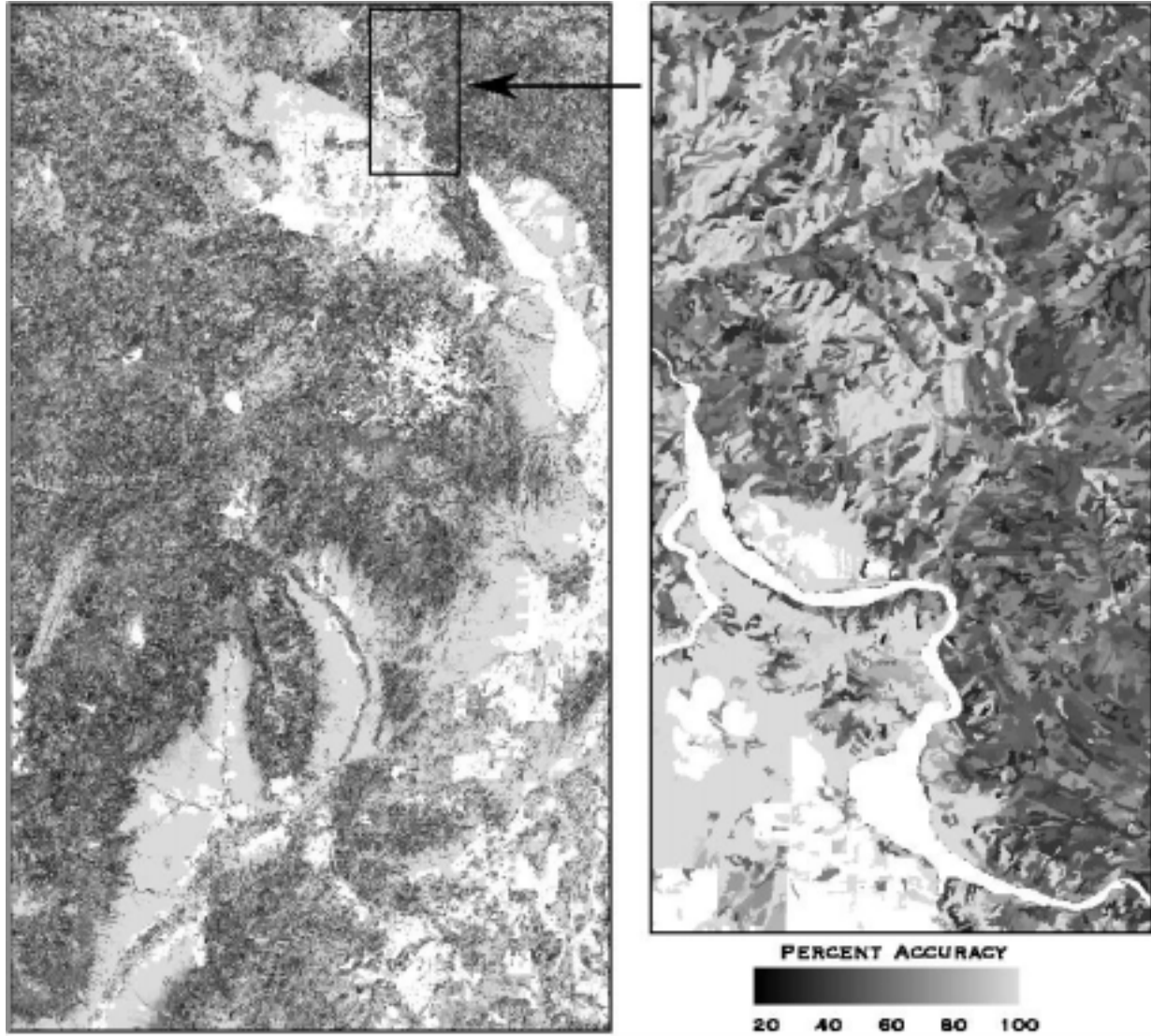
**Figure 2.** Land cover class predictions of the EB 10-NN (right panel) and the EB 10-NN+MID (left panel) classifiers for a portion of scene 3928 (see Figure 1 for location). Black dots indicate location of the 1422 training observations within the area.



**Figure 3.** Boxplots of distance (km) to nearest training sample point for scene 3928. This shows that training sample points tend to be closer to another training sample point than do polygons in the population.



**Figure 4.** EB 10-NN+MID estimated posterior probabilities of correct classification for all polygons in scene 3928. These reflect the lower amount of spatial information for the population versus the training sample.



**Figure 5.** Estimated accuracy of the land cover map shown in the left panel of Figure 2. Accuracies are linearly calibrated estimated posterior probabilities of correct classification for the EB 10-NN+MID classifier.