

An investigation into the use of maximum posterior probability estimates for assessing the accuracy of large maps

Brian Steele¹ and Dave Patterson²
Dept. of Mathematical Sciences
University of Montana

Introduction

- The focus of this talk is on assessing classifier accuracy when the training sample is unreliable for this purpose
- This situation is particularly troublesome when classifiers are used for large-area land cover mapping
- 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 of label assignment measures one or more predictor variables on all map units by a remote sensing device such as a satellite, and land cover on a *training* sample of map units.
- A classification rule is then constructed from the training sample and used to predict land cover for the unsampled units using the remotely sensed data.

¹steele@mso.umt.edu

²3dapatterson@mso.umt.edu

- Accuracy assessment very important for effective map use
- Example: The USDA Forest Service and the Wildlife Spatial Analysis Lab at the University of Montana recently mapped 21.5 million hectares of forested mountains and rangeland in Montana and Idaho using 9 Landsat Thematic Mapper images
- Most of the training observations were collected by the Forest Service for purposes other than mapping land cover.
- The overall spatial distribution of training observations was highly irregular, largely because most were sampled from forested public lands with easy access.
- None of the training data sets constitute a probability sample, and post-classification sampling was not carried out.
- The reliability of accuracy estimates derived from the training samples is dubious
- Given that training samples are the only data source, how can we reduce the anticipated bias?
- More generally, are there alternatives to traditional accuracy estimators (e.g., cross-validation) that are resistant to population drift?

Classifiers and Posterior Probabilities

- \mathcal{P} denotes a population comprised of c classes, or groups, identified by labels $1, \dots, c$.
- An element $\mathbf{x} \in \mathcal{P}$ is a pair (\mathbf{t}, y) where y is the group label and \mathbf{t} is a p -vector of observations on a covariate vector.
- A training sample of size n collected from \mathcal{P} is denoted by

$$X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$$

- The probability that $\mathbf{x} \in \mathcal{P}$ belongs to group g , conditional on \mathbf{t} , is denoted by $P(y = g \mid \mathbf{t})$.
- Let η denote a classification rule obtained from X_n , and $\eta(\mathbf{t})$ denote a prediction of y obtained by evaluating the classification rule for $\mathbf{x} \in \mathcal{P}$.

- Generally, classification rules can be viewed as posterior probability estimators that assign \mathbf{x} to the group with the maximum posterior probability (MPP) estimate. That is,

$$\eta(\mathbf{t}) = \arg \max_g \hat{P}(y = g | \mathbf{t}),$$

where $\hat{P}(y = g | \mathbf{t})$ denotes an estimator of $P(y = g | \mathbf{t})$.

- Classifiers differ with respect to method of estimating posterior probabilities
- Example: the linear discriminant classifier estimator of $\hat{P}(y = g | \mathbf{t})$ estimates $P(y = g | \mathbf{t})$ by

$$\hat{P}(y = g | \mathbf{t}) = \frac{p_g f(\mathbf{t} | \bar{\mathbf{t}}_g, \hat{\Sigma})}{\sum_{j=1}^c p_j f(\mathbf{t} | \bar{\mathbf{t}}_j, \hat{\Sigma})}$$

where $f(\mathbf{t} | \bar{\mathbf{t}}_j, \hat{\Sigma})$ is the multivariate normal pdf evaluated at \mathbf{t} , given training sample estimates of $\boldsymbol{\mu}_j$, the common covariance matrix Σ , and p_j , the estimated prior probability of membership in group j

Classifier Accuracy

- Herein, the accuracy of a classifier is the probability that a randomly sampled $\mathbf{x} \in \mathcal{P}$ will be correctly classified. In the case of a finite population of N observations

$$\alpha = N^{-1} \sum_{\mathbf{x} \in \mathcal{P}} P[\eta(\mathbf{t}) = y]$$

- Let $\Psi(E)$ denote the indicator function of the event E
- The probability that $\eta(\mathbf{t})$ is correct is

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

- All of the indicator variables in this sum are 0 except the indicator of the group for which the probability of membership, $P(y = g | \mathbf{t})$, is maximal.
- For that group, say group j , $P(y = j | \mathbf{t}) = \max_g P(y = g | \mathbf{t})$. Hence,

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

- While this relationship is well-known, it is not used in practice for two reasons
 - 1) The exact posterior probabilities are rarely known, and unbiased estimators of the individual specific accuracy's $P[\eta(\mathbf{t}) = y]$ are rarely known
 - 2) Given a training sample collected by a probability sampling, cross-validation and bootstrapping is effective
- However, without a probability sample, or under population drift, resampling estimators are not reliable
- Post-classification is sometimes possible

An Alternative Approach

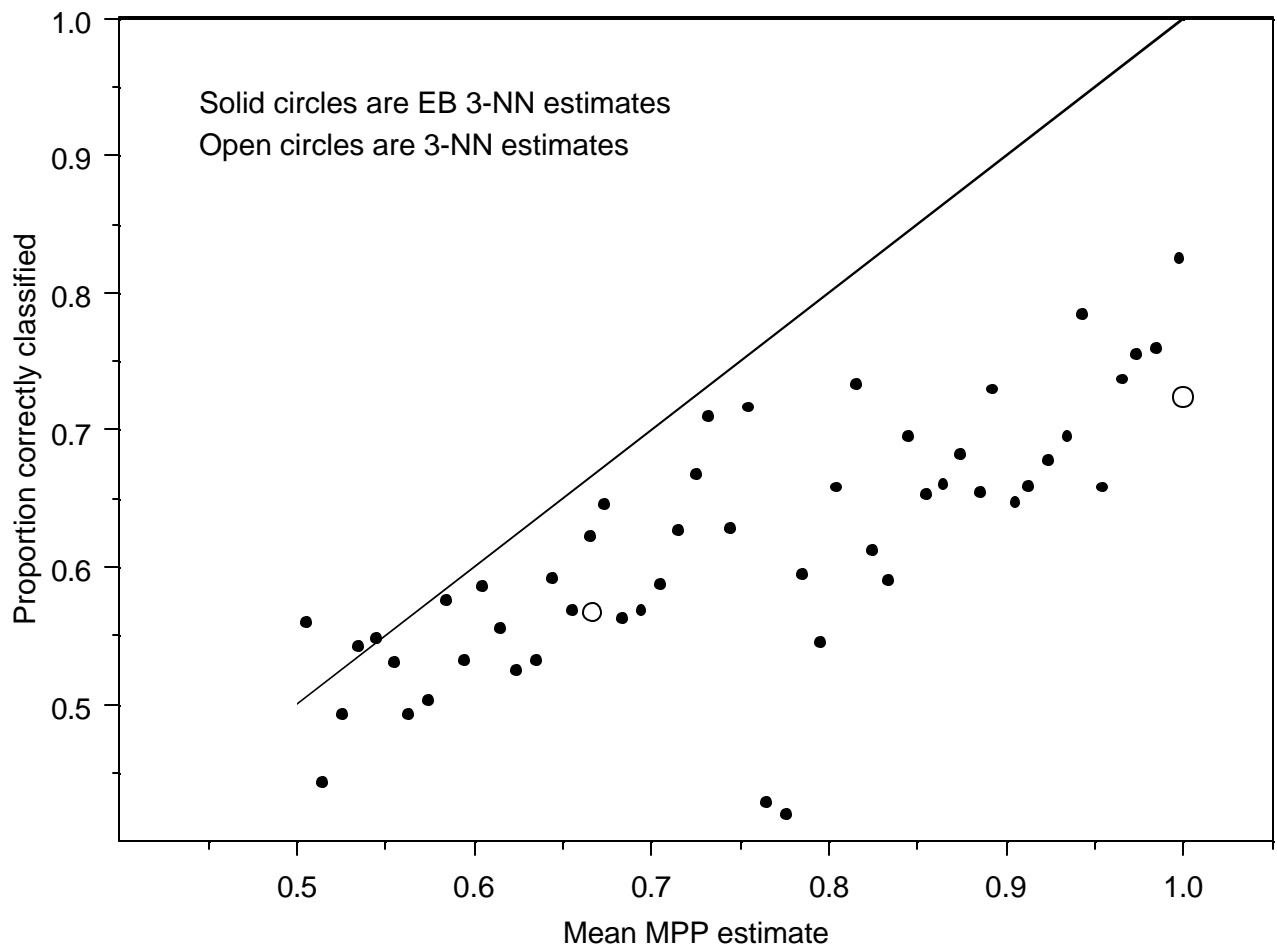
1. Compute $\hat{P}[\eta(\mathbf{t}) = y] = \max_g \hat{P}(y = g | \mathbf{t})$ for every map unit $\mathbf{x} \in \mathcal{P}$
2. Calibrate (correct for bias in) the estimates $\hat{P}[\eta(\mathbf{t}) = y]$ for every $\mathbf{x} \in \mathcal{P}$
3. Compute

$$\hat{\alpha} = N^{-1} \sum_{\mathbf{x} \in \mathcal{P}} \hat{P}_c[\eta(\mathbf{t}) = y]$$

where $\hat{P}_c[\eta(\mathbf{t}) = y]$ is the calibrated estimator of $P[\eta(\mathbf{t}) = y]$

Calibration

- Why calibrate? Classifiers are developed to maximize accuracy, not for unbiased MPP estimation
- For illustration, we simulated 10,000 outcomes $\Psi[\eta(\mathbf{t}) = y]$ and elementary estimates $\hat{P}[\eta(\mathbf{t}) = y]$ using 3-NN classifier and a smoothed 3-NN classifier
- Estimates are binned by the MPP estimates. For each bin of 200 pairs, we computed and plotted the proportion of correctly classified observations and the mean MPP estimate



- Calibration functions are developed by modeling the relationship between

$$P[\eta(\mathbf{t}) = y \mid \mathbf{t}] = E\{\Psi[\eta(\mathbf{t}) = y] \mid \mathbf{t}\}$$

and the MPP estimator $\hat{P}[\eta(\mathbf{t}) = y \mid \mathbf{t}]$.

- Herein the calibration model is

$$E\{\Psi[\eta(\mathbf{t}) = y] \mid \mathbf{t}\} = \beta \hat{P}[\eta(\mathbf{t}) = y \mid \mathbf{t}]$$

- The calibration function is constrained to pass through $(1/c, 1/c)$

- The constraint is motivated by

$$1/c \leq \max_g P(y = g | \mathbf{t}) = P[\eta(\mathbf{t}) = y | \mathbf{t}].$$

- We use the training sample to estimate the calibration coefficient
- The least squares estimator of β is

$$\hat{\beta} = \frac{\sum (p_i - c^{-1})(\Psi_i - c^{-1})}{\sum (p_i - c^{-1})^2},$$

where

$$p_i = \hat{P}[\eta(\mathbf{t}_i) = y_i | \mathbf{t}_i]$$

and

$$\Psi_i = \Psi[\eta(\mathbf{t}_i) = y_i]$$

- The pairs (p_i, Ψ_i) , $i = 1, \dots, n$ are computed after removing (\mathbf{t}_i, y_i) from the training sample used to compute the classifier
- The calibrated MPP estimator is

$$\begin{aligned} & \hat{P}_c[\eta(\mathbf{t}) = y | \mathbf{t}] \\ &= \begin{cases} \hat{\beta} \hat{P}[\eta(\mathbf{t}) = y | \mathbf{t}] + \frac{(1-\hat{\beta})}{c} & \text{if } \hat{\beta} \hat{P}[\eta(\mathbf{t}) = y | \mathbf{t}] + \frac{(1-\hat{\beta})}{c} \leq 1 \\ 1 & \text{if } \hat{\beta} \hat{P}[\eta(\mathbf{t}) = y | \mathbf{t}] + \frac{(1-\hat{\beta})}{c} > 1. \end{cases} \end{aligned}$$

- Post-classification samples can be used to compute calibration functions

ASimulation Study

- We compare
 - 1) 10 fold cross-validation estimators (from the training sample)
 - 2) post-classification estimators
 - 3) MPP estimators calibrated from training samples
 - 4) MPP estimators calibrated from post-classification samples
- \mathcal{P} ($N = 200,000$) was simulated by using the training observations to locate centers of probability mass in the covariate space
- Each simulated observation was generated by adding a multivariate normal random vector to a training observation, The group label was the same as the training observation
- Training samples were drawn according to three designs
 - 1) random sampling
 - 2) producing positively biased 10-fold CV estimates of α
 - 3) producing negatively biased 10-fold CV estimates of α
- A positively biased 10-fold CV estimate was obtained by preferentially sampling observations with large MPP estimates
- The linear discriminant classifier was used
- Exact classifier accuracy was computed by classifying every $\mathbf{x} \in \mathcal{P}$

Table 1. Effect of three sampling designs on the 10-fold cross-validation estimator $\hat{\alpha}_{CV}$. Tabled values are the true accuracy rates $\bar{\alpha}$ and the means of the 10-fold cross-validation estimator $\hat{\alpha}_{CV}$. 100 simulations.

Sampling design	Landsat Scene			
	P39/R27		P37/R29	
	$\bar{\alpha}$	$\overline{\hat{\alpha}_{CV}}(\mathbf{X}_n)$	$\bar{\alpha}$	$\overline{\hat{\alpha}_{CV}}(\mathbf{X}_n)$
Unbiased	67.9	68.0	59.0	59.0
Positive bias	66.9	75.8	58.1	66.3
Negative bias	66.3	53.1	59.2	51.1

Figure 1. Bias and root mean square error estimates (%) under random sampling. Each data point was obtained from 100 simulated populations of $N = 200,000$ observations.

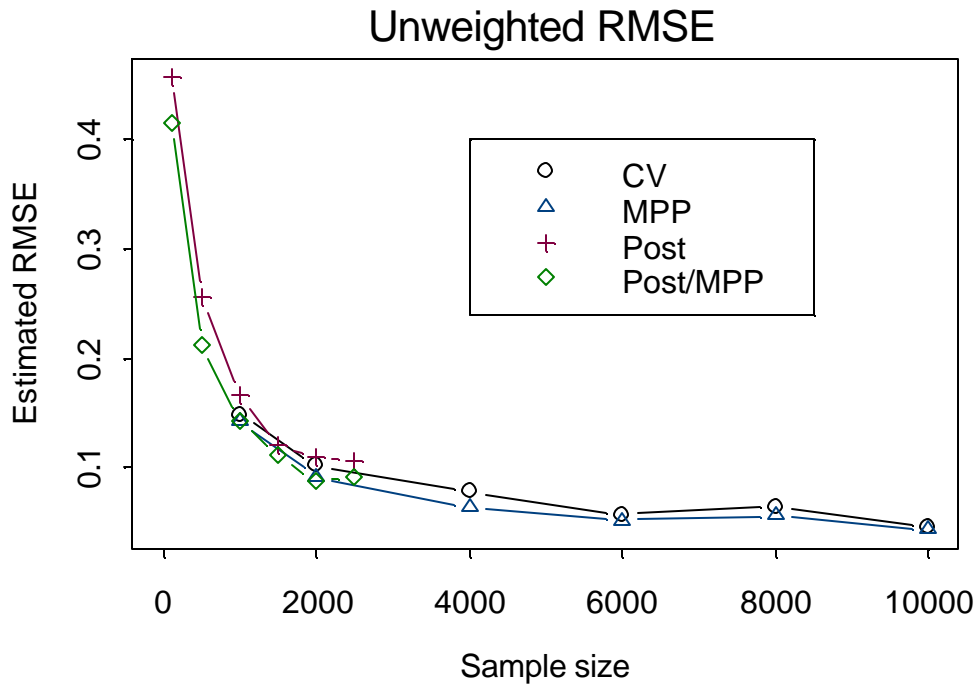
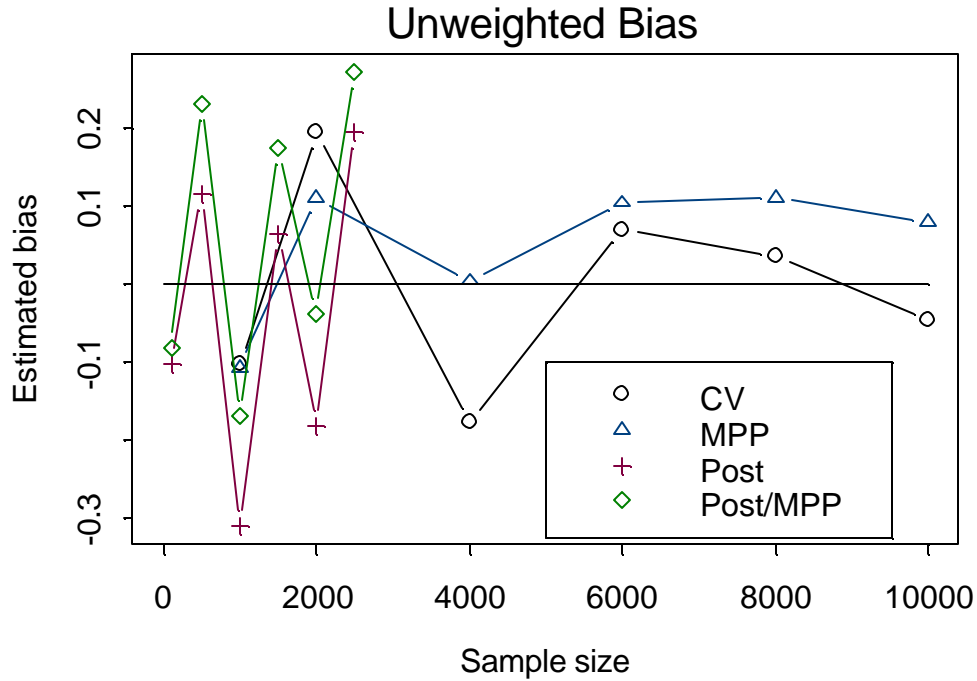


Figure 2. Bias and root mean square error estimates (%) under a positively biased sampling design. Each data point was obtained from 100 simulated populations of $N = 200,000$ observations.

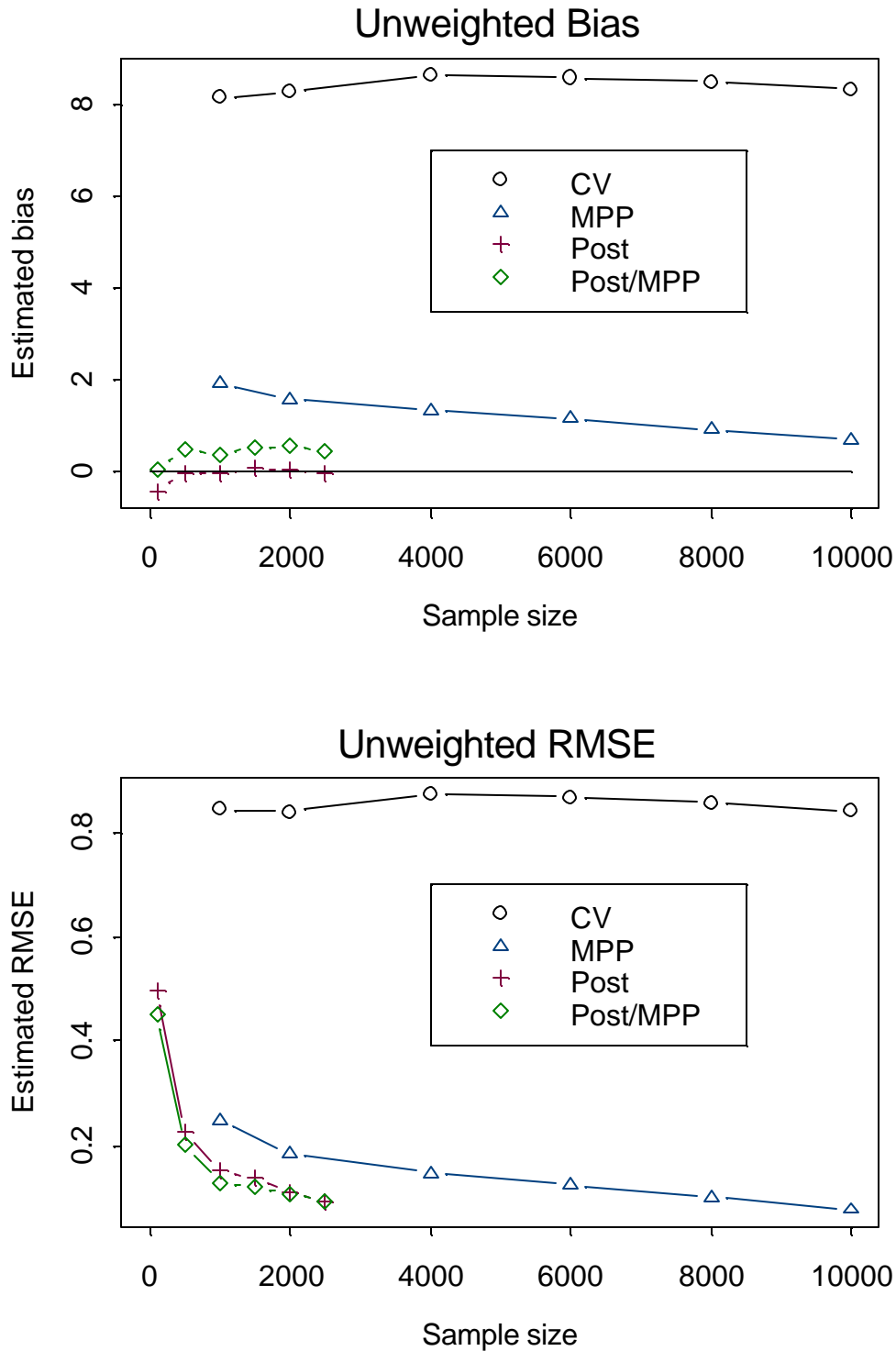
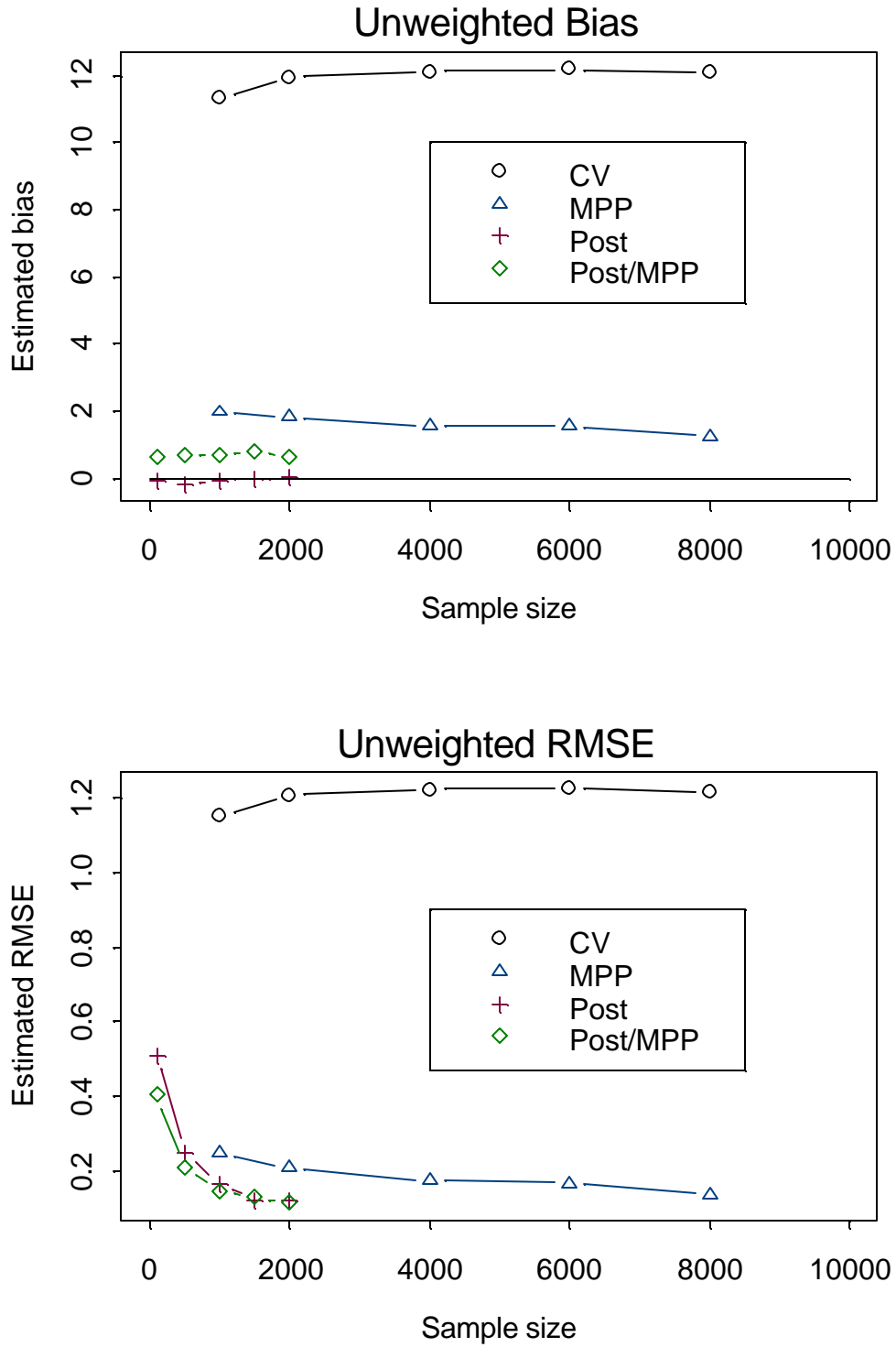


Figure 3. Bias and root mean square error estimates (%) under a negatively biased sampling design. Each data point was obtained from 100 simulated populations of $N = 200,000$ observations.



Conclusions

- Post-classification sampling is preferred over conventional resampling and MPP-based methods if cost is no object
- There is little lost in using calibrated MPP estimators compared to cross-validation estimators
- Bias and MSE may be substantially reduced if calibrated MPP estimators are used
- This strategy is very useful for large-scale land cover mapping when post-classification sampling is not practical

References

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