Remote Sensing of Environment 115 (2011) 715–724

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Satellite image-based maps: Scientific inference or pretty pictures?

Ronald E. McRoberts

Northern Research Station, U.S. Forest Service, Saint Paul, Minnesota 55108 USA

ARTICLE INFO

Article history:
Received 5 May 2010
Received in revised form 27 October 2010
Accepted 30 October 2010

Keywords:
Probability-based inference
Model-based inference
Forest inventory
Accuracy
Bias
Precision
Multinomial logistic regression

ABSTRACT

The scientific method has been characterized as having two distinct components, *Discovery* and *Justification*. Discovery emphasizes ideas and creativity, focuses on conceiving hypotheses and constructing models, and is generally regarded as lacking a formal logic. Justification begins with the hypotheses and models and ends with a valid scientific inference. Unlike Discovery, Justification has a formal logic whose rules must be rigorously followed to produce valid scientific inferences. In particular, when inferences are based on sample data, the rules of the logic of Justification require assessments of bias and precision. Thus, satellite image-based maps that lack such assessments for parameters of populations depicted by the maps may be of little utility for scientific inference; essentially, they may be just pretty pictures. Probability- and model-based approaches are explained, illustrated, and compared for producing inferences for population parameters using a map depicting three land cover classes: non-forest, coniferous forest, and deciduous forest. The maps were constructed using forest inventory data and Landsat imagery. Although a multinomial logistic regression model was used to classify the imagery, the methods for assessing bias and precision can be used with any classification method. For probability-based approaches, the difference estimator was used, and for model-based inference, a bootstrap approach was used.

Published by Elsevier Inc.

1. Introduction

Bunge (1967) asserts that science has a unique goal and a unique method that serve to distinguish it from non-science. With respect to method, he further asserts that "the scientific method is a mark of science,... no scientific method, no science." In characterizing the scientific method, Reichenbach (1938) emphasized the distinction between the context of *Discovery* and the context of *Justification*. The starting point for Discovery is an acknowledged problem or gap in the current state of knowledge, and the ending point is one or more hypotheses, models, or solutions proposed to fill the gap. The starting point for Justification is the set of hypotheses, models, or proposed solutions, and the ending point is a valid scientific inference.

An important distinction between Discovery and Justification is that whereas Justification has a well-accepted logic, Discovery is a creative enterprise for which no formal logic has been constructed. Strategies that have been used in Discovery include trial and error, systematic search, derivation from theory, illumination of the well-prepared mind, inspiration, serendipity, and more recently, artificial intelligence and data mining (McRoberts, 1989). Within Discovery, there are no right or wrong methods; in fact, Feyerabend (1975) asserts that "Anything goes!" Thus, the crux of Discovery is ideas, regardless of how they are generated. The U.S. Forest Service has characterized scientific research as a competition of ideas (USDA Forest Service, 2000), and Efron (2007) asserts that "ideas are the coin of the realm in the intellectual world."

Although "Anything goes!" in Discovery, such is decidedly not the case for Justification. The ending point of Justification is a scientific inference, and the logic of inference is formal and well-documented. The relevant Oxford English Dictionary definition of *infer* is "to accept from evidence or premises" (Simpson & Weiner, 1989). For most scientific problems, evidence in the form of complete enumerations of populations of interest would be prohibitively expensive, if not physically impossible. Thus, statistical procedures have been developed to infer values for population parameters from estimates based on observations from a sample of population units. In a sampling framework, inference requires expression of the relationship between the population parameter, μ, and its estimate, $\hat{\mu}$, in probabilistic terms (Dawid, 1983). For situations in which the intent is estimation, as opposed to hypothesis testing, these probabilistic expressions often take the form of the familiar $1-\alpha$ confidence intervals where $1-\alpha$ denotes the probability that confidence intervals constructed using data for all possible samples will include μ. Of crucial importance, construction of the confidence intervals requires use of the sample data to calculate the estimate, $\hat{\mu}$ and its variance estimate, $\hat{Var}(\hat{\mu})$. Failure to express $\hat{\mu}$ in probabilistic terms through ignorance, neglect, or the inability to estimate $\hat{Var}(\hat{\mu})$ prohibits characterization of the result as a valid scientific inference.

For natural resources problems, satellite imagery has become a well-accepted source of landscape information that contributes to the construction of land cover maps. The utility of these maps for scientific

E-mail address: rmcroberts@fs.fed.us.

0034-4257/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.rse.2010.10.013

inference varies depending on issues of bias and precision. Card (1982) cautions that "users will not and should not take a map at face value without some associated estimate of error." Switzer (1969) asserts that a map without an uncertainty assessment puts the user "in a position similar to one who is given a point estimate without any notion of its standard error." Thus, if bias and precision have not been assessed and reported, regardless of the reason, then maps may have little or no utility for scientific inference. The conclusion, which will certainly be disconcerting and unpleasant to some, is that if estimates of population parameters derived from maps cannot be characterized in probabilistic terms the map cannot serve as a basis for scientific inference; from an inferential perspective, the map may be little more than just a pretty picture.

The remote sensing literature is replete with examples of map accuracy assessments based on error matrices and measures such as overall accuracy, users' and producers' accuracies, and the Kappa index. However, with only few exceptions (Czaplewski & Catts, 1992; Gallego, 2004; McRoberts, 2006, 2010; Stehman, 2005, 2009), these assessments do not provide information on the bias or precision of estimates of parameters such as land cover class proportions or total land cover area for populations depicted by the maps. In a comprehensive review of accuracy assessment methods, Strahler et al. (2006, p. 39) note the need for "methods for estimating the accuracy of spatially aggregated products." The reasons for the lack of literature documentation of bias and precision assessments for estimates of areal population parameters and methods for conducting them are not apparent; perhaps they relate to lack of interest, to lack of expertise, or to lack of methods included in software packages. Regardless of the reasons, descriptions and illustrations of general methods for assessing the bias and uncertainty of estimates of parameters of populations depicted by maps that can be used with any approach to classification are warranted.

The objectives of the study are twofold: (1) to distinguish probability-based and model-based inference as they pertain to inferences derived from maps for areal population parameters, and (2) to describe and illustrate methods for assessing bias and precision for both probability- and model-based inference that can be used without regard to the classification method. The context of the analyses is a land cover map that depicts non-forest, coniferous forest, and deciduous forest classes and that was constructed using forest inventory data, Landsat satellite imagery, and a multinomial logistic regression model.

2. Data

The study area was defined by the portion of the row 27, path 27, Landsat scene in northern Minnesota, in the United States of America (USA) (Fig. 1). Land use for the study area consists of forest land dominated by aspen-birch and spruce-fir associations, agriculture, wetlands, and water. Landsat Thematic Mapper (TM) imagery was acquired for three dates corresponding to early, peak, and late seasonal vegetative stages: April 2000, July 2001, and November 1999. Preliminary analyses indicated that the normalized difference vegetation index (NDVI) (Rouse et al., 1973) and the tasseled cap (TC) transformations (brightness, greenness, and wetness) (Crist & Cicone, 1984; Kauth & Thomas, 1976) were superior to both the spectral band data and principal component transformations with respect to predicting the probability of the three land cover classes. Thus, 12 TM-based variables were used, NDVI and the three TC transformations for each of the three image dates. Within the study area, six 15-km × 15-km areas of interest (AOI) were selected using a systematic grid of locations.

The Forest Inventory and Analysis (FIA) program of the U.S. Forest Service conducts the national forest inventory of the USA. The program has established field plot centers in permanent locations using a sampling design that produces an equal probability sample (Bechtold & Patterson, 2005; McRoberts et al., 2005). The sampling design is based on a tessellation of the country into approximate 2400-ha (6000-ac) hexagons and features a permanent plot at a randomly selected location within each hexagon. Some states, including Minnesota in which the study area is located, provide additional funding to double the sampling intensity to approximately one plot per 1200 ha.

Each FIA plot consists of four 7.32-m (24-ft) radius circular subplots for a total area of 672.45 m^2. The subplots are configured as a central subplot and three peripheral subplots with centers located at 36.58 m (120 ft) and azimuths of 0^o, 120^o, and 240^o from the center of the central subplot. In general, locations of forested or previously forested plots are determined using global positioning system receivers, whereas locations of non-forested plots are verified using aerial imagery and digitization methods. The spatial configuration of the FIA subplots with centers separated by 36.58 m and the 30-m × 30-m spatial resolution of the TM /ETM+ imagery permits individual subplots to be associated with individual image pixels. For this study, only data for the central subplots of the FIA plots were used to avoid issues related to lack of independence among observations for subplots of the same plot.

Field crews observe species and measure diameter at-breast-height (dbh) (1.37 m, 4.5 ft) and height for all trees with dbh of 12.7 cm (5 in) or greater. Tree data are aggregated at the subplot level to obtain totals for basal area (BA) and other variables. Area and BA by land cover class are obtained for subplots by collapsing ground land use conditions into non-forest and multiple forest classes subject to the FIA definition of forest land: area of at least 0.4 ha (1 ac), continuous crown-to-crown width of at least 36.58 m (120 ft), and stocking of at least 10% where stocking refers to the extent to which the growth potential of a site is used by trees (Hansen & Hahn, 1992). For this study, data for subplots that were not completely forested or completely non-forested were not used. In addition, data for subplots that field crews determined to be forested but that had no trees with dbh of 12.7 cm or greater were not used. The 1907 central subplots in the study area, including the 94 subplots in the six AOIs (14–17 subplots per AOI), that satisfied these conditions were measured between 1999 and 2003.

For this study, subplots were assigned to three land cover classes: (1) *non-forest* (*NF*) for subplots whose land cover did not qualify as forest land as determined via interpretation of aerial photography or by field crews, (2) *coniferous forest* (*CF*) for forested subplots with proportions of total BA in coniferous trees of at least 0.50, and (3) *deciduous forest* (*DF*) for forested subplots with proportions of total BA in deciduous trees greater than 0.50. Of the 1907 central subplots in the study area, 607 were classified as NF, 604 were classified as CF, and 696 were classified as DF; of the 94 plots in the six AOIs, 24 were classified as NF, 34 were classified as CF, and 36 were classified as DF. A variogram analysis indicated that spatial correlation for cover class observations for central subplots of FIA plots was negligible.

3. Methods

All analyses were based on four underlying assumptions: (1) a finite population consisting of N units in the form of 30-m × 30-m TM pixels, (2) an equal probability sample of n population units in the form of observations of the central subplots of FIA plots, (3) ancillary data in the form of the 12 TM-based spectral transformations for each pixel and (4) adequate characterization of the 900-m^2 pixels by the 167.87-m^2 subplots. In the following sections, the terms *population unit* and *pixel* are used interchangeably.

Data for accuracy assessments and for estimating the bias and precision of estimators may come from the sample data, a subset of the sample data, or from an independent data set. Stehman and Czaplewski (1998), Stehman (2000, 2006, 2008), Strahler et al. (2006), and McRoberts (2010a,b) all provide design guidelines and

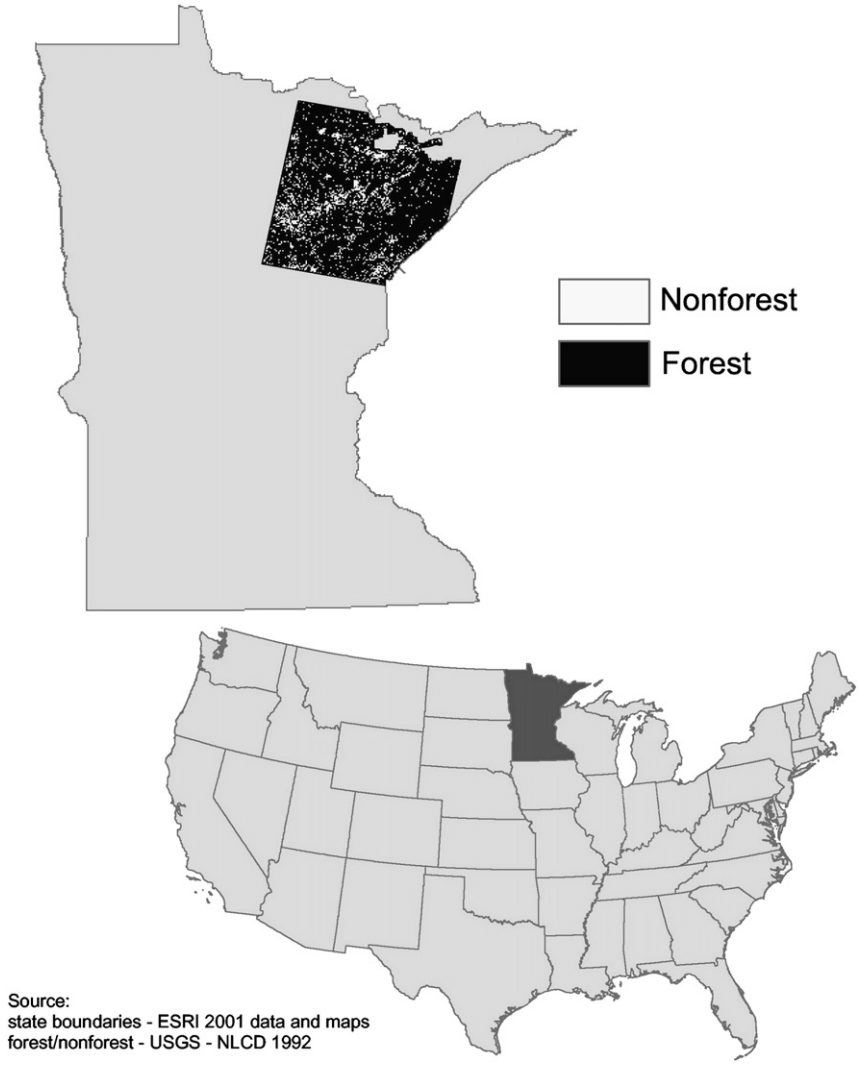

Fig. 1. Study area.

principles for acquiring independent data for accuracy assessment purposes. Although independent data are preferable, their acquisition may not be possible. In addition, splitting the sample data into model calibration and model validation subsets exacerbates the problem of insufficient sample sizes for small areas. Thus, a common alternative is simply to use all the sample data for both model calibration and validation; such is the case for this study. One consequence, however, is that accuracy assessments tend to be optimistic.

3.1. Logistic regression model

Two general approaches to classification of satellite images are common, *unsupervised classification* and *supervised classification*. With unsupervised classification, pixels are first aggregated into homogenous classes based on their image spectral values. Labels such as NF, CF, and DF are then assigned to the homogeneous classes based on expert opinion, ground observations for pixels in the classes, or other methods. Supervised classification is more sophisticated and entails estimating relationships between the land cover classes and the spectral values of the imagery. For this study, a supervised classification approach based on a multinomial logistic regression model was used. However, the bias and precision assessment methods that are proposed and illustrated are not specific to any particular classification approach.

The relationship between a dichotomous dependent variable, Y, and continuous independent variables, **X**, is often expressed in the form,

$$p_i = f(\mathbf{X}_i;\boldsymbol{\beta}),$$

where i indexes population units, p_i is the probability that $y_i = 1$, and $\boldsymbol{\beta}$ is a vector of parameters to be estimated (Agresti, 2007). The function, $f(\mathbf{X}_i;\boldsymbol{\beta})$, expresses the statistical expectation of Y in terms of **X** and $\boldsymbol{\beta}$ and is often formulated using the logistic function as,

$$p_i = f(\mathbf{X}_i;\boldsymbol{\beta}) = \frac{exp\left(\sum_{j=1}^{J}\beta_j x_{ij}\right)}{1 + exp\left(\sum_{j=1}^{J}\beta_j x_{ij}\right)}, \quad (1)$$

where $j = 1, 2, \ldots, J$ indexes the independent variables, and exp(.) is the exponential function.

The logistic regression model approach can be extended from dichotomous to polychotomous response variables (Agresti, 2007; McRoberts, 2009, 2010). For a response variable with $M \geq 3$ classes, one class is selected arbitrarily and designated the Mth class. The

estimate of the probability that the ith population unit belongs to the Mth class is,

$$\hat{p}_{M,i} = \frac{1}{1 + \sum_{m=1}^{M-1} \sum_{j=1}^{J} exp\left(\hat{\beta}_{mj} x_{ij}\right)}, \quad (2a)$$

where m indexes the classes of the response variable. The estimate of the probability that the i^{th} population unit belongs to the mth class (m<M) is,

$$\hat{p}_{m,i} = \hat{p}_{M,i}\, exp\left(\sum_{j=1}^{J} \hat{\beta}_{mj} x_{ij}\right). \quad (2b)$$

All parameters of the multinomial logistic regression model can be estimated simultaneously using a variety of software packages such as SAS with the CATMOD procedure or R with the VGAM library.

3.2. Probability-based inference

Probability-based inference is based on three assumptions: (1) population units are selected for the sample using a randomization scheme; (2) the probability of selection for each population unit is positive and known; and (3) the observation of the response variable for each population unit is a constant as opposed to a random variable. Properties of probability-based estimators are based on random variation resulting from the probabilities of selection of population units into the sample, thus the characterization of these estimators as probability-based (Hansen et al., 1983). Although probability-based inference is also characterized as design-based inference, the term *design* is not well-defined in the sense that it may refer simply to the selection of population units into the sample or additionally to the entire inferential process (Kendall & Buckland, 1982).

3.2.1. Pixel-level inference

With probability-based methods, land cover class observations and predictions are both discrete values of a categorical variable. An error matrix (Table 1) is used to compare numbers of observations and predictions by land cover class and provides the necessary information for four commonly used measures of accuracy: *overall accuracy* (OA), which is the proportion of observations correctly classified; *user's accuracy* (UA), which is the ratio of the number of correct predictions and the total number of predictions for a class; and *producer's accuracy* (PA), which is the ratio of the number of correct predictions and the total number of observations for a class. In addition, the *Kappa coefficient* (κ) (Congalton, 1991) is a measure of association describing the agreement between two classifications (Kraemer, 1983). When one classification is based on observations and the other is based on predictions, κ is considered a measure of accuracy and is estimated as,

$$\hat{\kappa} = \frac{n \sum_{m=1}^{M} n_{mm} - \sum_{m=1}^{M} (n_{m+} \cdot n_{+m})}{n^2 - \sum_{m=1}^{M} (n_{m+} \cdot n_{+m})}, \quad (3)$$

where n is the number of observations; M is the number of land cover classes; n_{mm} is the number of population units both observed and predicted to be in the mth land cover class; n_{m+} and n_{+m} are the numbers of population units observed in the mth land cover class and predicted to be in the mth land cover class, respectively; and the dot symbol (·) denotes multiplication. Perfect agreement between the observations and predictions is indicated when $\hat{\kappa} = 1$, whereas $\hat{\kappa} = 0$ indicates no agreement beyond that expected by chance. An assumption underlying κ is that the land cover class observations and predictions are independent which is not the case for this study. Therefore, estimates of κ for this study are acknowledged to be optimistic. Of crucial importance, a map accuracy assessment does not constitute an inference for a population parameter because it does not directly provide the necessary estimates of bias and precision.

With probability-based inference, there is no bias or uncertainty for the observed classes of individual population units selected for the sample because the observations are constants, apart from measurement error. No bias assessment for the land cover class predictions for individual population units not selected for the sample is possible apart from bias assessments made for population units in aggregate (Section 3.2.2).

3.2.2. Areal inference

Probability-based inference at the areal level relies on the sample data. Further, probability-based inference produces estimates, $\hat{\mu}$, of areal population parameters using estimators that are generally unbiased, although an estimate based on any one sample may deviate considerably from the true value. However, probability-based estimators do not generally produce the same estimates as are obtained by aggregating population unit predictions, and they often do not produce sufficiently precise estimates for small AOIs with small sample sizes.

For areal assessment of maps, the objective is typically to estimate population parameters such as the total area or proportion of a land cover class. Because the estimate of the total area of a class is simply the product of total area which is usually known and the estimate of the class proportion, the focus of this study is estimation of the proportion. For the mth land cover class, define,

$$y_{m,i} = \begin{cases} 1 & \text{if class m is observed for the } i^{th} \text{ population unit} \\ 0 & \text{if class m is not observed for the } i^{th} \text{ population unit} \end{cases};$$

and,

$$\hat{y}_{m,i} = \begin{cases} 1 & \text{if class m is predicted for the } i^{th} \text{ population unit} \\ 0 & \text{if class m is not predicted for the } i^{th} \text{ population unit} \end{cases}.$$

For sampling designs with equal probabilities of selection for each population unit, the simplest approach to probability-based areal inference for the proportion, μ_m, of an AOI with map class m is to use the familiar simple random sampling (SRS) estimators:

$$\hat{\mu}_{m,SRS} = \frac{1}{n} \sum_{i=1}^{n} y_{m,i}, \quad (4)$$

Table 1
Generic error matrix.

		Predicted class			Total	Producer's accuracy (PA)
		1	2	3		
Observed class	1	n_{11}	n_{12}	n_{13}	n_{1+}	$\frac{n_{11}}{n_{1+}}$
	2	n_{21}	n_{22}	n_{23}	n_{2+}	$\frac{n_{22}$
	3	n_{31}	n_{32}	n_{33}	n_{3+}	$\frac{n_{33}}{n_{3+}}$
Total		n_{+1}	n_{+2}	n_{+3}	n	
User's accuracy (UA)		$\frac{n_{11}}{n_{+1}}$	$\frac{n_{22}}{n_{+2}}$	$\frac{n_{33}}{n_{3+}}$	OA = $\frac{n_{11} + n_{22} + n_{33}}{n}$	

and,

$$\hat{Var}\left(\hat{\mu}_{m,SRS}\right) = \frac{\hat{\mu}_{m,SRS}\left(1-\hat{\mu}_{m,SRS}\right)}{n}, \quad (5)$$

where i indexes the n sample observations and $y_{m,i}$ is the observation for the ith population unit. Probability-based areal inference consists of inferring μ_m from the sample observations as $\hat{\mu}_{m,SRS} \pm t\sqrt{\hat{Var}\left(\hat{\mu}_{m,SRS}\right)}$ where $t \approx 2.0$ produces an approximate 95% confidence interval. For this study, finite population correction factors are ignored for these estimators because the sampling intensity is very small, on the order of 1:144,000. In addition, although the SRS variance estimators are known to be biased when used with systematic sampling designs, the bias is conservative in the sense that the estimators over-estimate the true variances (Särndal et al., 1992). The advantages of SRS estimators are that they are intuitive, familiar, and simple; the primary disadvantage is that they often produce large variance estimates, particularly for small areas with small sample sizes.

Model-assisted estimators rely on observations for population units selected for the sample and model predictions for population units not selected for the sample. However, because the validity of an inference is still based on the probability sample, the estimator is characterized as probability-based. Stehman (2009) provides an excellent discussion of the conceptual framework underlying the relationship between model-assisted estimators and map accuracy assessments. For equal probability samples, a naïve model-assisted estimator of the population proportion, μ_m, for the mth land cover class, is,

$$\hat{\mu}_{m,naive} = \frac{1}{N}\left(\sum_{i=1}^{n} y_{m,i} + \sum_{i=n+1}^{N} \hat{y}_{m,i}\right),$$

where the population has been indexed so that $i = 1,2,\ldots,n$ denotes the sampled population units with observations, $y_{m,i}$, and $i = n+1,n+2,\ldots,N$ denotes the remaining population units with predictions, $\hat{y}_{m,i}$. The bias of this estimator can be estimated as,

$$\hat{Bias}\left(\hat{\mu}_{m,naive}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_{m,i}-y_{m,i}\right) = \frac{n_{+m}-n_{m+}}{n}, \quad (6)$$

where n_{+m} and n_{m+} are obtained from the error matrix (Table 1) (McRoberts, 2010). The model-assisted difference estimator (Baffetta et al., 2009; Särndal et al., 1992) is defined as the difference between the naïve estimator and its bias estimate,

$$\hat{\mu}_{m,Dif} = \frac{1}{N}\left(\sum_{i=1}^{n} y_{m,i} + \sum_{i=n+1}^{N} \hat{y}_{m,i}\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_{m,i}-y_{m,i}\right), \quad (7a)$$

which can be approximated as,

$$\hat{\mu}_{m,Dif} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_{m,i} - \frac{n_{+m}-n_{m+}}{n}, \quad (7b)$$

(McRoberts, 2010). The variance of $\hat{\mu}_{m,Dif}$ can be approximated as,

$$\hat{Var}\left(\hat{\mu}_{m,Dif}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{y}_{m,i}-y_{m,i}\right)^2 = \frac{n_{+m}+n_{m+}-2n_{mm}}{n(n-1)}, \quad (8)$$

(McRoberts, 2010) where n_{+m}, n_{m+}, and n_{mm} are again obtained from the error matrix (Table 1). Thus, although error matrices do not constitute inferences and do not directly assess bias and precision of population parameters, they do provide much of the information necessary for doing so when using the model-assisted difference estimator.

The primary advantage of model-assisted estimators is that they have the potential to capitalize on the relationship between the land cover class observations and their predictions to reduce the variance of the population parameter estimate. However, as with all probability-based estimators, model-assisted estimators often suffer the effects of small sample sizes when used for small area estimation.

3.3. Model-based inference

With model-based inference, a model is used to produce predictions for all population units. Although model-based estimators are often more computationally intense, they produce maps as by-products; they produce estimates that are compatible with maps in that the estimates are calculated by aggregating population unit predictions; they may produce viable estimates for small areas; but they cannot be assumed to be unbiased. Model-based approaches to inference are based on the assumption of an entire distribution of possible observations for Y for each population unit, not just a single constant value as is the case for probability-based inference. This conceptual framework of a distribution of possible values for each population unit is often characterized as a superpopulation (Graubard & Korn, 2002; Rennolls, 1982). Whereas randomization for probability-based inference enters through the random selection of population units into the sample, randomization for model-based inference enters through the random realization of observations from the distributions for individual population units selected for the sample. Thus, model-based inference is often characterized as conditional on the sample.

Current approaches to model-based inference originated in the context of survey sampling and can be attributed to Mátern (1960), Brewer (1963), and Royall (1970). Given the origins of model-based inference in survey sampling, it is not surprising that forestry applications have often been in the context of forest inventory (Gregoire, 1998; Kangas, 2006; Mandallaz, 2008; Rennolls, 1982). An important aspect of the model-based approach is that when the model is correctly specified, the estimator is unbiased (Lohr, 1999); however, when the model is misspecified, the adverse effects on inference may be substantial (Hansen et al., 1983; Royall & Herson, 1973). Thus, much of the reported research on model-based inference has focused on selection of sampling designs and estimators that are robust to model misspecification (Chambers, 2003; Valliant et al., 2000).

The mean and standard deviation of the distribution of Y_m for the ith population unit may be denoted $\mu_{m,i}$ and $\sigma_{m,i}$, respectively. For model-based inference, an observation, $y_{m,i}$, for the ith population unit is expressed as,

$$y_{m,i} = \mu_{m,i} + \varepsilon_{m,i},$$

where $\varepsilon_{m,i}$ is the random deviation of the observation, $y_{m,i}$, from its mean, $\mu_{m,i}$. The mean, $\mu_{m,i}$, is expressed mathematically as,

$$\mu_{m,i} = f(\mathbf{X}_i;\boldsymbol{\beta}),$$

and the model,

$$y_{m,i} = f(\mathbf{X}_i;\boldsymbol{\beta}) + \varepsilon_{m,i},$$

is fit to the sample data where $\mathbf{X}_i$ is the vector of ancillary variables associated with the ith population unit, and $\boldsymbol{\beta}$ is a vector of parameters to be estimated. With model-based approaches, estimation often focuses on the mean, $\mu_{m,i}$, rather than the particular observed land cover class, $y_{m,i}$, which is a random realization from the distribution of which $\mu_{m,i}$ is the mean. With the multinomial logistic regression

model, $\mu_{m,i}$ is estimated as $\hat{\mu}_{m,i} = \hat{p}_{m,i}$ from Eqs. (2a) and (2b). A crucial component of bias assessment for model-based approaches is an evaluation of the quality of fit of the model to the data as a means of assessing the adequacy of the model specification (Section 4.2).

3.3.1. Pixel-level inference

With model-based approaches, there are no correct or incorrect land cover class predictions because the class observations are random realizations from distributions of possible observations. Therefore, error matrices and accuracy measures such as OA, UA, PA, and κ are not relevant for model-based approaches. However, whereas there is no uncertainty in land cover class observations for sample population units with the probability-based approach because observations are constants, such uncertainty does exist with model-based approaches because observations are realizations of random variables. Depending on the modelling approach, parametric methods for estimating the uncertainty of $\hat{\mu}_{m,i}$ for individual population units may be possible. For a binomial logistic regression model, McRoberts (2006, 2010) describe and illustrate a parametric method using Taylors series approximations for estimating variances. However, an appropriate approach for $M \geq 3$ is more complex and is highly dependent on the mathematical form of the model. Therefore, a more generic approach such as bootstrap resampling could be considered (Section 3.3.3).

3.3.2. Areal inference

With model-based inference, the population mean, μ_m, is estimated as,

$$\hat{\mu}_{m,Mod} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_{m,i} = \frac{1}{N}\sum_{i=1}^{N} f\left(\mathbf{X}_i;\hat{\boldsymbol{\beta}}\right), \tag{9}$$

where the subscript *Mod* denotes an estimator for model-based inference. With the model-based approach, $\hat{\mu}_{m,Mod}$ is compatible with the map, whereas estimates based on probability-based estimators may differ considerably from $\hat{\mu}_{m,Mod}$. The general form of an estimator of the variance of $\hat{\mu}_{m,Mod}$ is,

$$\hat{Var}\left(\hat{\mu}_{m,Mod}\right) = \frac{1}{N^2}\left[\sum_{i=1}^{N}\hat{Var}\left(\hat{\mu}_{m,i}\right) + 2\sum_{i<}^{N}\sum_{j}^{N}\hat{Cov}\left(\hat{\mu}_{m,i},\hat{\mu}_{m,j}\right)\right]$$

(McRoberts, 2006, 2010). Calculation of $\hat{Var}\left(\hat{\mu}_{m,i}\right)$ and $\hat{Cov}\left(\hat{\mu}_{m,i},\hat{\mu}_{m,j}\right)$ is complex, highly dependent on the mathematical form of the model, and possibly computationally intensive. Thus, resampling approaches to estimating $\mathrm{Var}\left(\hat{\mu}_{m,Mod}\right)$ also merit consideration (Section 3.3.3).

3.3.3. The bootstrap procedure

Resampling procedures such as the bootstrap are well-suited for complex and non-parametric model applications and for applications requiring assumptions whose validity is difficult to assess. The bootstrap resampling procedure was invented by Efron (1979, 1981, 1982) and further developed by Efron and Tibshirani (1994). All bootstrap methods depend on the notion of a *bootstrap sample*. For regression problems, Efron and Tibshirani (1994) describe three approaches to constructing a bootstrap sample. First, for a sample of n pairs $(y_i, \mathbf{X}_i)$ and the corresponding empirical distribution of the pairs, each with probability 1/n, a bootstrap sample is defined to be a sample of size n drawn with replacement from the empirical distribution. This approach is characterized as *bootstrapping pairs*. The pairs can also be expressed as $(\hat{y}_i + \varepsilon_i, \mathbf{X}_i)$ where $\varepsilon_i = y_i - \hat{y}_i$. The second and third approaches to bootstrapping focus on resampling the residuals, ε, and are characterized as *bootstrapping residuals*. The second approach draws a random sample with replacement from the empirical distribution of residuals and adds them back to the model predictions to form a bootstrap sample. The third approach draws the sample of residuals from a parametric model of the residual distribution. Assumptions underlying all bootstrapping approaches are that the resampling mimics the original sampling features such as clustering and the original data features such as heteroscedasticity and that both observations and residuals are independently and identically distributed (iid). In the case of heteroscedasticity, either the residuals must be studentized by dividing them by their respective standard deviations before resampling or the bootstrapping procedure must incorporate heteroscedasticity.

For model-based inference, randomization occurs in the realization of observations from the distribution for each population unit rather than randomization in the selection of population units into the sample. Thus, for model-based inference, bootstrapping residuals is the more intuitive approach, although the iid assumption may be difficult to accommodate.

Efron and Tibshirani (1994, page 113) comment that bootstrapping residuals is more sensitive to assumptions than bootstrapping pairs for which the only assumption is that the original pairs represent a random sample from the population of interest. However, they further note that when bootstrapping pairs, the results approach those obtained when bootstrapping residuals as n increases.

Regardless of the bootstrapping approach selected, the model is refit to the kth bootstrap sample, estimates are calculated for each population unit, and the population estimate, $\hat{\mu}^k_{boot}$, is calculated using Eq. (9). The overall bootstrap population estimate is then,

$$\hat{\mu}_{boot} = \frac{1}{n_{boot}}\sum_{k=1}^{n_{boot}}\hat{\mu}^k_{boot}, \tag{10}$$

where n_{boot} is the number of bootstrap samples. The bootstrap estimate of bias is defined as,

$$\hat{\mathrm{Bias}}_{boot}\left(\hat{\mu}\right) = \hat{\mu}_{boot} - \hat{\mu} \tag{11}$$

where $\hat{\mu}$ is the estimate obtained from the original sample. Of particular importance, this bias estimate pertains only to sampling issues; it does not address model misspecification. The bootstrap estimate of variance is calculated as,

$$\hat{\mathrm{Var}}_{boot}\left(\hat{\mu}\right) = \frac{1}{n_{boot}-1}\sum_{k=1}^{n_{boot}}\left(\hat{\mu}^k_{boot} - \hat{\mu}_{boot}\right)^2. \tag{12}$$

4. Analyses

For both the model-assisted and model-based approaches to inference, observations and TM data for the 1907 central subplots were used to estimate the parameters of the multinomial logistic regression model. A variogram analysis of studentized residuals indicated negligible spatial correlation. The model was then applied to the six AOIs using model parameter estimates based on the sample from the entire study area. Thus, both the model-assisted difference estimator and the model-based estimator are characterized as *synthetic* because they use data external to the AOIs to which they are applied (Gonzalez, 1973).

4.1. Probability-based inference

For each pixel in each AOI, the class, m, for which $\hat{p}_{m,i}$ from Eqs. (2a) and (2b) was the largest was assigned. For all subplots in the study area and for all subplots in the aggregation of the six AOIs, error matrices corresponding to sample observations and corresponding predictions were constructed (Table 1), and OA, PA, UA, and $\hat{\kappa}$ were calculated. For each class, m, and for each AOI and the aggregation of the six AOIs, the SRS estimates, $\hat{\mu}_{m,SRS}$ and $\hat{Var}\left(\hat{\mu}_{m,SRS}\right)$, were calculated using only observations for subplots with centers in the

AOIs. The SRS estimates served as the standard for comparison for the other estimators. In addition, for each class, m, and for each AOI and the aggregation of the six AOIs, $\hat{\mu}_{m,Dif}$, $\hat{Var}\left(\hat{\mu}_{m,Dif}\right)$, and $\hat{Bias}\left(\hat{\mu}_{m,naive}\right)$ were calculated with the constraint that the last two estimates were based only data for subplots with centers in the AOIs.

4.2. Model-based inference

For each pixel and for m = 1 (NF), m = 2 (CF), and m = 3 (DF), $\hat{\mu}_{m,i} = \hat{p}_{m,i}$ was calculated, and for each AOI and the aggregation of the six AOIs, $\hat{\mu}_{NF,Mod}$, $\hat{\mu}_{CF,Mod}$, and $\hat{\mu}_{DF,Mod}$ were calculated.

Because the issue of bias for model-based estimators is closely linked to correct model specification, the approach to assessing pixel-level bias focused on assessing the quality of fit of the model to the data at the subplot/pixel level. If the model is correctly specified, a graph of observations versus model predictions for a continuous response variable should feature points that lie along a line with intercept 0 and slope 1. However, because the data used to estimate the model parameters for this study are polychotomous, a slightly modified three-step approach was used: (1) all subplot observation/pixel model prediction pairs, $(y_{mi}, \hat{\mu}_{mi})$, were ordered with respect to $\hat{\mu}_{mi}$; (2) the ordered pairs were grouped into categories of equal numbers of pairs, and the group means of the subplot observations and of the pixel model predictions were calculated; and (3) a graph of the observation means versus the model prediction means was constructed. In the absence of model lack of fit a graph of the points defined by the corresponding group means of observations and group means of model predictions should lie along the 1:1 line and a simple linear model fit to the points should have intercept of approximately 0 and slope of approximately 1.

For both binomial and multinomial logistic regression models, bootstrapping residuals using the empirical distribution of residuals is not feasible (Appendix A). However, both the bootstrapping pairs approach and the parametric approach to bootstrapping residuals were used. For both approaches, bootstrap resamples of size n = 1907 were selected with replacement. For the parametric approach to bootstrapping residuals, a modification as per Lin et al. (2009) and Mittlböck and Schemper (2002) was used to accommodate the polychotomous distribution of the data and the heterogeneous residual variances. For each bootstrap sample, a random number, r, was drawn from a uniform (0,1) distribution for the ith population unit in the sample, and a bootstrap observation, $\tilde{y}_i$, was generated by assigning a land cover class to the unit as,

$$\tilde{y}_i = \begin{cases} 1(\text{NF}) \text{ if } 0 \le r \le \hat{\mu}_{NF,i} \\ 2(\text{CF}) \text{ if } \hat{\mu}_{NF,i} < r \le \hat{\mu}_{NF,i} + \hat{\mu}_{CF,i} \\ 3(\text{DF}) \text{ if } \hat{\mu}_{NF,i} + \hat{\mu}_{CF,i} < r \le 1 \end{cases},$$

where $\hat{\mu}_{NF,i}$, $\hat{\mu}_{CF,i}$, and $\hat{\mu}_{DF,i}$ were calculated using parameters estimated from the original data. Efron and Tibshirani (1994) recommend at least $n_{boot} = 200$ bootstrap samples. For this study, bootstrap resampling and calculation of $\hat{Bias}_{boot}(\hat{\mu})$ from Eq. (11) and $SE_{boot}(\hat{\mu}) = \sqrt{\hat{Var}_{boot}(\hat{\mu})}$ from Eq. (12) continued until $SE_{boot}(\hat{\mu})$ stabilized. Bootstrap estimates, $\hat{Bias}_{boot}(\hat{\mu})$ and $SE_{boot}(\hat{\mu})$, were calculated for each of the six AOIs and the aggregation of the six AOIs.

5. Results and discussion

5.1. Probability-based inference

Results of accuracy assessments for all subplots in the entire study area and for only the subplots in the six AOIs were similar (Tables 2a and 2b). For both datasets, PA, UA, and OA were small relative to the general preference for accuracies of 0.85 or greater (Anderson et al., 1976). However, the results are similar to those obtained by McRoberts (2009) who used similar data to compare multinomial logistic regression, disciminant analysis, and the k-Nearest Neighbors techniques. Multiple reasons can be advanced to explain these results. First, the CF and DF classes are not intrinsically discrete classes but rather are distinguished by a threshold for a continuous variable, BA. Thus, confusion between the CF and DF classes for BA ≈ 0.5 should be expected. Second, for GPS receivers of the type used by the FIA program, as many as half the subplots could be associated with an incorrect pixel (McRoberts, 2010b). Third, tree density on forest subplots varies considerably, even though the entire subplot is characterized as forest. Thus, completely forested subplots with sparse tree cover may be predicted to have smaller probabilities for the CF and DF classes than forested subplots with dense tree cover. Fourth, spectral differences for subplots with similar tree densities but different age structures or health conditions may produce different estimates for land cover classes. Fifth, subplots with tree cover may be characterized as NF if the minimum FIA requirements of a 0.4-ha patch with 36.58-m width and 10% stocking are not satisfied. Sixth, the observation for the smaller subplot may not adequately characterize the larger pixel. However, when the CF and DF classes were aggregated into a single forest class, forest/non-forest accuracy assessments produced OA = 0.88 for all subplots in the study area and OA = 0.89 for only the subplots in the six AOIs. These latter OAs were similar to those reported elsewhere for forest/non-forest classifications based on similar data but using different classification methods (Finley et al., 2008; Haapanen et al., 2004). These results suggest the primary cause for the relatively low accuracies may be due to the inability to distinguish the CF and DF classes, particularly for BA ≈ 0.5.

The SRS and model-assisted difference estimates of means were similar for the aggregation of the six AOIs but deviated substantially for some individual AOIs (Table 3). For the aggregation of the six AOIs, the model-assisted difference estimates were within two SRS standard errors of the SRS means. The standard errors for model-assisted difference estimates were generally, although not always, smaller than the SRS estimates, a result that reflects the quality of the class predictions obtained from the model. However, the naïve bias estimates for the model-assisted approach were as large in absolute value as 0.18 for individual AOIs. This result may be attributed to the relatively small numbers of subplots in the AOIs. For example, with only 14–17 subplots in each AOI, each misclassified plot has the potential to increase the naïve bias estimate by approximately 0.07. As expected, the naive bias estimates were smaller for the aggregation of the six AOIs, ranging in absolute value from 0.01 to 0.05.

5.2. Model-based inference

Graphs of the means of ordered groups of observations versus means of corresponding ordered groups of predictions suggested adequate fit of the model to the data and, therefore, little model misspecification (Fig. 2a,b,c). Simple linear models fit to the grouped data yielded $\hat{\alpha}_{NF} = 0.011$, $\hat{\alpha}_{CF} = 0.008$, and $\hat{\alpha}_{DF} = 0.017$ as estimates of the intercepts and $\hat{\beta}_{NF} = 0.966$, $\hat{\beta}_{CF} = 0.975$, and $\hat{\beta}_{DF} = 0.952$ as

Table 2a
Error matrix for all central subplots in study area.

		Predicted class			Total	Producer's accuracy (PA)
		NF	C	D		
Observed class	NF	444	46	117	607	0.731
	C	30	414	160	604	0.685
	D	38	146	512	696	0.736
Total		512	606	789	1907	
User's accuracy (UA)		0.867	0.683	0.645	OA = 0.718 $\hat{\kappa}$ = 0.650	

Table 2b
Error matrix for all central subplots in six AOIs.

		Predicted class			Total	Producer's accuracy (PA)
		NF	CF	DF		
Observed class	NF	17	1	6	24	0.708
	CF	2	28	4	34	0.824
	DF	1	8	27	36	0.750
Total		20	37	37	94	
User's accuracy (UA)		0.850	0.757	0.730	OA = 0.755 $\hat{\kappa}$ = 0.698	

estimates of the slopes. All these pairs of estimates are very similar to $\alpha = 0$ and $\beta = 1$ as would be expected for correct model specification.

For the aggregation of the six AOIs, $SE_{boot}(\hat{\mu})$ stabilized for $n_{boot} = 300$ for the parametric approach to bootstrapping residuals, but $n_{boot} = 500$ was necessary for $SE_{boot}(\hat{\mu})$ to stabilize for the bootstrapping pairs approach (Fig. 3a,b). For individual AOIs, similar results were obtained. Although stabilization for the same values of n_{boot} for individual AOIs as for the aggregation of the six AOIs may seem counter-intuitive, estimates for the multinomial logistic regression model used for estimating $\hat{\mu}_{Mod,m}$ for the individual AOIs were based on the same number of subplots and the same bootstrap observations as were used for estimating $\hat{\mu}_{Mod,m}$ for the aggregation of the six AOIs.

The model-based estimates of the mean, $\hat{\mu}_{Mod}$, were all within two SRS standard errors of the SRS estimates, $\hat{\mu}_{SRS}$, except for the single case for which $SE(\hat{\mu}_{SRS}) = 0$. The two bootstrapping approaches produced very similar estimates of $SE_{boot}(\hat{\mu})$ for both individual AOIs and the aggregation of the six AOIs. This result confirms the assertion of Efron and Tibshirani (1994, page 113) that results for bootstrapping pairs should approach results for bootstrapping residuals for large n. Although a comparison of the two sets of model-based standard errors to the two sets of probability-based standard errors is enticing, the two inferential approaches are based on such different underlying assumptions that such a comparison would be conceptually meaningless. Bootstrap estimates of bias, $\widehat{Bias}_{boot}(\hat{\mu})$, for the two approaches were nearly the same and essentially negligible. However, as previously noted, $\widehat{Bias}_{boot}(\hat{\mu})$ does not assess model misspecification which is why separate analyses to assess model lack of fit are necessary (Fig. 2a,b,c).

Model-based estimates of variances and standard errors for logistic regression models depend heavily on estimates of parameter variances and covariances which, in turn, depend heavily on sample sizes. Thus, if the sample size used for estimating model parameters is doubled, standard errors would be expected to decrease by a multiplicative factor of $\frac{1}{\sqrt{2}}$, whereas if the sample size were only half, standard errors would be expected to increase by a multiplicative factor of $\sqrt{2}$. This effect is reflected in the bootstrapping procedure in that parameters estimated from smaller or larger bootstrap samples would exhibit greater or lesser variability, respectively.

6. Conclusions

Multiple conclusions may be drawn from the study. First, for the aggregation of the six AOIs, the two probability-based approaches and the model-based approach produced similar estimates of proportions for the three land cover classes. However, the estimates differed considerably for the individual AOIs. In particular, as per Särndal et al. (1992, page 223), the model-assisted difference estimator should not be used for small sample sizes or unless classification accuracies are very large. Nevertheless, for areas with larger sample sizes such as the aggregation of the six AOIs, the model-assisted difference estimator can be used with any classification approach to produce valid statistical inferences. Second, the two bootstrapping approaches to estimating standard errors for the model-based approach produced nearly identical results and confirm the assertion of Efron and Tibshirani (1994). This finding indicates that bootstrapping pairs, for which underlying assumptions are minimal, can be used with classification approaches for which bootstrapping residuals is difficult or impossible. In particular, bootstrapping pairs, in conjunction with an assessment of model lack of fit, provides a mechanism leading to a valid model-based inference for any approach used to obtain areal population estimates. In summary, methods producing both valid probability-based and model-based inferences, regardless of the classification or estimation approach, were described and illustrated. These methods are sufficient to move a map from simply a pretty

Table 3
Estimates for six 15-km × 15-km AOIs individually and in aggregate.

AOI	n	Class	SRS		Model-assisted			Model-based				
			$\hat{\mu}_{srs}$	$SE(\hat{\mu}_{srs})$	$\hat{\mu}_{Dif}$	$SE(\hat{\mu}_{Dif})$	$\widehat{Bias}(\hat{\mu}_{naive})$ [a]	$\hat{\mu}_{Mod}$ Mod	Pairs bootstrapping		Residuals bootstrapping	
									$SE_{boot}(\hat{\mu}_{Mod})$	$\widehat{Bias}_{boot}(\hat{\mu}_{Mod})$ [b]	$SE_{boot}(\hat{\mu}_{Mod})$	$\widehat{Bias}_{boot}(\hat{\mu}_{Mod})$ [b]
1	15	NF	0.267	0.114	0.357	0.098	−0.133	0.291	0.012	−0.0008	0.012	−0.0008
		CF	0.333	0.122	0.214	0.098	0.000	0.270	0.011	0.0005	0.012	0.0000
		DF	0.400	0.127	0.429	0.138	0.133	0.439	0.015	0.0003	0.015	0.0008
2	14	NF	0.643	0.357	0.348	0.074	0.071	0.457	0.007	−0.0016	0.007	0.0000
		CF	0.357	0.128	0.461	0.074	−0.071	0.333	0.010	0.0012	0.010	0.0001
		DF	0.000	–	0.191	0.000	0.000	0.210	0.009	0.0004	0.009	−0.0001
3	17	NF	0.059	0.057	0.158	0.061	−0.059	0.143	0.009	−0.0005	0.010	0.0008
		CF	0.706	0.111	0.573	0.105	0.177	0.611	0.019	−0.0003	0.018	−0.0010
		DF	0.235	0.103	0.270	0.086	−0.118	0.254	0.017	0.0009	0.016	0.0002
4	17	NF	0.118	0.078	0.162	0.105	−0.059	0.166	0.011	0.0004	0.011	0.0000
		CF	0.353	0.116	0.260	0.136	0.177	0.404	0.018	0.0001	0.016	0.0004
		DF	0.529	0.121	0.578	0.146	−0.118	0.430	0.018	−0.0005	0.017	−0.0004
5	14	NF	0.214	0.110	0.323	0.000	0.000	0.387	0.012	−0.0009	0.013	−0.0007
		CF	0.214	0.110	0.263	0.148	−0.143	0.174	0.009	0.0003	0.010	0.0001
		DF	0.571	0.132	0.414	0.148	0.143	0.439	0.014	0.0006	0.014	0.0008
6	17	NF	0.294	0.111	0.298	0.105	−0.059	0.308	0.013	−0.0001	0.013	−0.0006
		CF	0.177	0.923	0.155	0.000	0.000	0.191	0.009	0.0001	0.008	−0.0004
		DF	0.529	0.121	0.547	0.105	0.059	0.501	0.014	0.0000	0.014	−0.0002
All	94	NF	0.255	0.045	0.277	0.034	−0.043	0.292	0.008	−0.0006	0.008	0.0000
		CF	0.362	0.050	0.312	0.041	0.032	0.331	0.009	0.0003	0.008	−0.0002
		DF	0.383	0.050	0.411	0.047	0.011	0.378	0.010	0.0003	0.009	0.0002

[a] Bias of naïve estimator incorporated into difference estimate of mean.
[b] Bias not incorporated into estimate of mean.

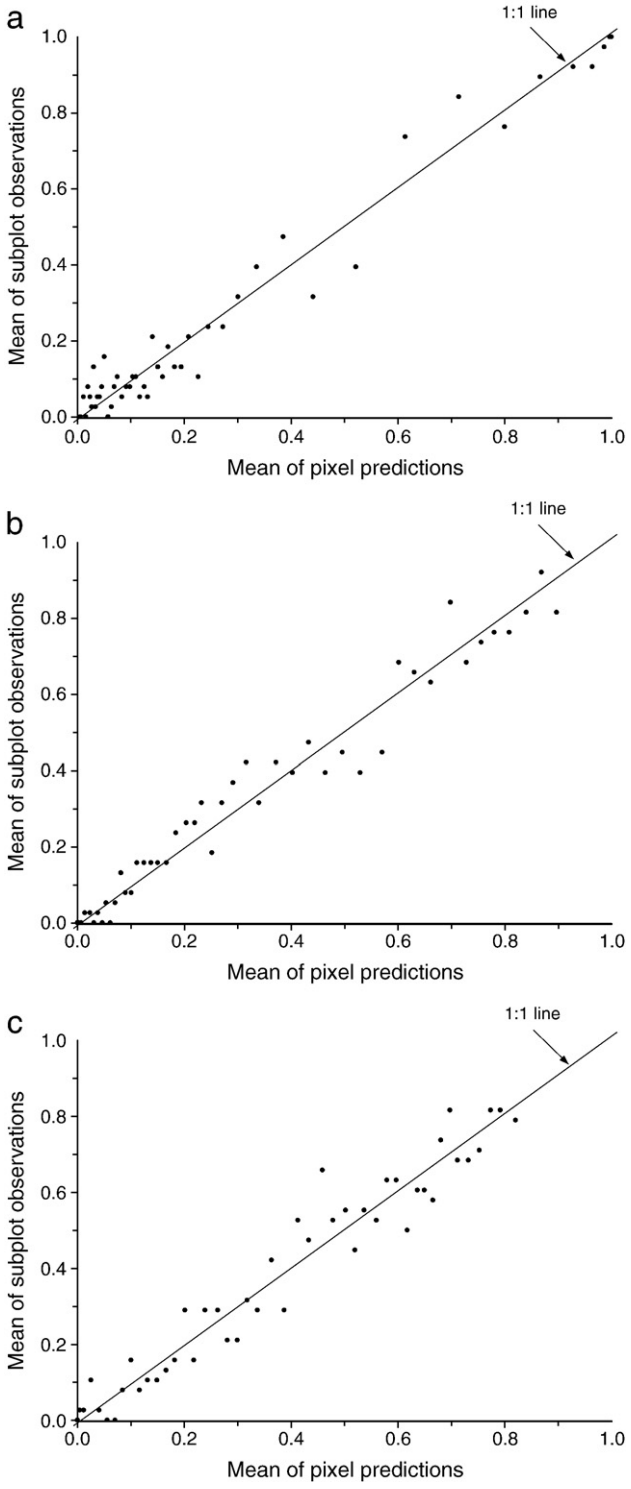

Fig. 2. a. Estimated probabilities of non-forest (NF) from Eqs. (2a) and (2b). b. Estimated probability of coniferous forest (CF) from Eqs. (2a) and (2b). c. Estimated probability of deciduous forest (DF) from Eqs. (2a) and (2b).

picture to a basis for scientific inference and to move a map-based estimation effort from Discovery to Justification.

Appendix 1

An assumption underlying bootstrapping of residuals is that the residuals are independently and identically (iid) distributed. Thus, in the presence of heteroscedasticity, the residuals must be studentized

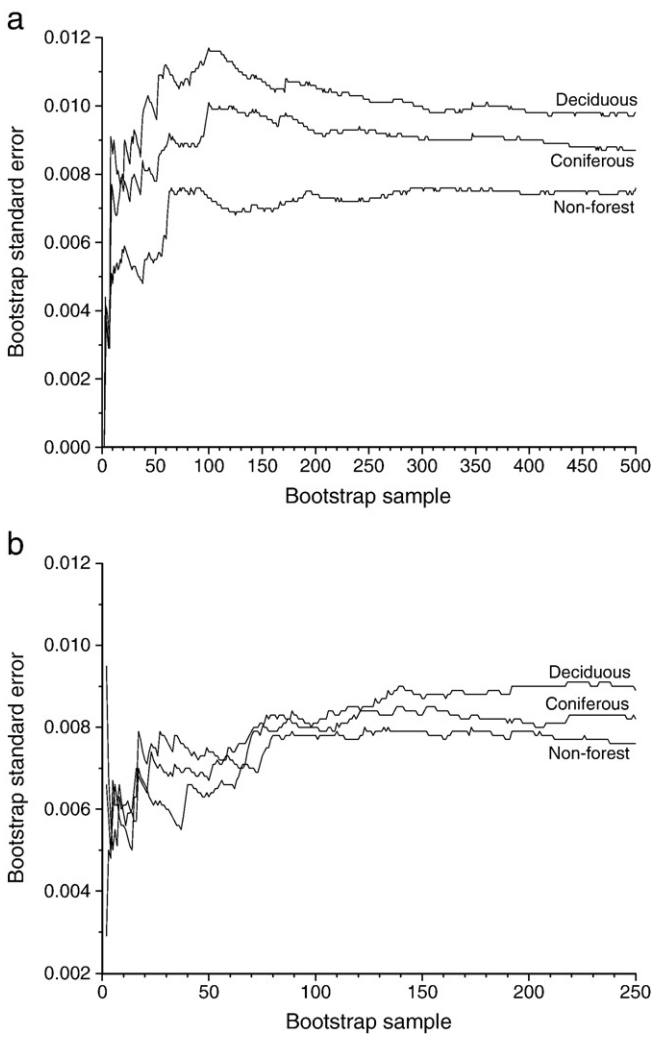

Fig. 3. a. Bootstrap standard error versus number of bootstrap samples for aggregation of six AOIs using bootstrapping pairs. b. Bootstrap standard error versus number of bootstrap samples for aggregation of six AOIs using parametric bootstrapping residuals.

by dividing them by estimates of their standard deviations. For a binomial variable, studentized residuals are of the form,

$$\frac{\varepsilon_i}{s_i} = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \begin{cases} \dfrac{1-\hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \sqrt{\dfrac{1-\hat{p}_i}{\hat{p}_i}} & \text{if } y_i = 1 \\ \dfrac{0-\hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = -\sqrt{\dfrac{\hat{p}_i}{1-\hat{p}_i}} & \text{if } y_i = 0 \end{cases}.$$

Thus, the studentized residual corresponding to any particular observation can take on only one of two values, depending on y_i. A simple example shows that the process of re-assigning residuals to model predictions leads to erroneous results. For $y_i = 1$ and $\hat{p}_i = 0.75$, the studentized residual is $\frac{\varepsilon_i}{s_i} = \sqrt{\frac{1-\hat{p}_i}{\hat{p}_i}} = \sqrt{\frac{0.25}{0.75}} \approx 1.73$. If this studentized residual is re-assigned to a prediction, $\hat{p}_j = 0.33$ with $s_j = \sqrt{\hat{p}_j\left(1-\hat{p}_j\right)} \approx 0.47$, the bootstrap observation becomes,

$$\tilde{y}_j = \hat{p}_j + s_j \frac{\varepsilon_i}{s_i} \approx 0.33 + 0.47(1.73) \approx 1.14,$$

which is impossible because y_j must be either 0 or 1. Similar results occur for a multinomial approach.

References

Agresti, A. (2007). *An introduction to categorical data analysis.* Hobkoen, NJ: Wiley-Interscience.

Anderson, J., Hardy, E., Roach, J., & Witmer, R. (1976). *A land use and land cover classification system for use with remote sensing data.* U.S.G.S. Professional Paper 964. Washington, DC: Government Printing Office.

Baffetta, F., Fattorini, L., Franceschi, S., & Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment*, *11*(3), 463–475.

Bechtold, W. A., & Patterson, P. L. (Eds.). (2005). *The enhanced Forest Inventory and Analysis program—National sampling design and estimation procedures. General Technical Report SRS-80.* Asheville, North Carolina: Southern Research Station, USDA Forest Service 85 pp. Available at:. http://www.srs.fs.usda.gov/pubs/ Last accessed: April 2009.

Brewer, K. R. W. (1963). Ratio estimation in finite populations: Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, *5*, 93–105.

Bunge, M. (1967). *Scientific research I.* New York: Springer-Verlag.

Card, D. H. (1982). Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, *48*(3), 431–439.

Chambers, R. (2003). An introduction to model-based survey sampling. Available at: http://www.eustat.es/prodserv/seminario_i.html Last accessed: November 2009.

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

Crist, E. P., & Cicone, R. C. (1984). Application of the tasseled cap concept to simulated Thematic Mapper data. *Photogrammetric Engineering and Remote Sensing*, *50*, 343–352.

Czaplewski, R. L., & Catts, G. P. (1992). Calibration of remotely sensed proportion or area estimates for misclassification error. *Remote Sensing of Environment*, *39*, 29–43.

Dawid, A. P. (1983). Statistical inference I. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, *Volume 4.* (pp. 89–105). New York: Wiley.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, *68*, 589–599.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* 38. Society of Industrial and Applied Mathematics CBMS-NSF Monographs.

Efron, B. (2007). The future of statistics. *Amstat News*, *363*, 47–50.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* New York: Chapman & Hall.

Feyerabend, P. (1975). *Against method* (third edition). New York: Verso.

Finley, A. O., Banerjee, S., & McRoberts, R. E. (2008). A Bayesian approach to multi-source forest area estimation. *Environmental and Ecological Statistics*, *15*(2), 1352–8505.

Gallego, F. J. (2004). Remote sensing and land cover estimation. *International Journal of Remote Sensing*, *25*(15), 3019–3047.

Gonzalez, M. E. (1973). Use and evaluation of synthetic estimates. *1973 Proceedings of the Social Statistics Section* (pp. 33–36). American Statistical Association.

Graubard, B. I., & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, *17*(1), 73–96.

Gregoire, T. (1998). Design-based and model-based inference: Appreciating the difference. *Canadian Journal of Forest Research*, *28*, 1429–1447.

Haapanen, R., Ek, A. R., Bauer, M. E., & Finley, A. O. (2004). Delineation of forest/nonforest land use classes using nearest neighbor methods. *Remote Sensing of Environment*, *89*(3), 265–271.

Hansen, M. H., & Hahn, J. T. (1992). Determining stocking, forest type, and stand-size class from forest inventory data. *Northern Journal of Applied Forestry*, *9*, 82–89.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-depending and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *78*, 776–793.

Kangas, A. (2006). Model-based inference. In A. Kangas & M. Maltamo (Eds.), *Forest inventory, methodology and applications* (pp. 39–52). Dordrecht, The Netherlands: Springer.

Kauth, R. J., & Thomas, G. S. (1976). The Tasseled Cap—A graphic description of the spectral–temporal development of agricultural crops as seen by Landsat. *Proceedings of the symposium on machine processing of remotely sensed data* (pp. 41–51). West Lafayette, IN: Purdue University.

Kendall, M. G., & Buckland, W. R. (1982). *A dictionary of statistical terms* (fourth edition). New York: Longman Group, Ltd.

Kraemer, H. C. (1983). Kappa coefficient. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*, *Volume 4.* (pp. 352–354). New York: Wiley.

Lin, Y., Newcombe, R. G., Lipsitz, S., & Carter, R. E. (2009). Fully specified bootstrap confidence intervals for the difference of two independent binomial proportions based on the median unbiased estimator. *Statistics in Medicine*, *28*, 2876–2890.

Lohr, S. (1999). *Sampling: design and analysis.*Pacific Grove, CA: Duxbury 494 pp.

Mandallaz, D. (2008). *Sampling techniques for forest inventories.* New York: Chapman & Hall 256 pp.

Mátern, B. (1960). *Spatial variation.* Medd. Statens Skogsforskningsinst. Band 49, No. 5. (Reprinted as volume 36 of the series Lecture notes in Statistics. 1986. New York: Springer-Verlag 151 pp.

McRoberts, R. E. (1989). A survey of research strategies. In R. A. Leary (Ed.), *Discovering new knowledge about trees and forests. General Technical Report NC-135.* (pp. 25–31). St. Paul, mN: U.S.Department of Agriculture, Forest Service, North Central Research Station.

McRoberts, R. E. (2006). A model-based approach to estimating forest area. *Remote Sensing of Environment*, *103*, 56–66.

McRoberts, R. E. (2009). A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. *Remote Sensing of Environment*, *113*, 532–545.

McRoberts, R. E. (2010a). Probability- and model-based approaches to estimating proportion forest using satellite imagery. *Remote Sensing of Environment*, *114*, 1017–1025.

McRoberts, R. E. (2010b). The effects of rectification and Global Positioning System errors on satellite image-based estimates of forest area. *Remote Sensing of Environment*, *114*, 1710–1717.

McRoberts, R. E., Bechtold, W. A. B., Patterson, P. L., Scott, C. T., & Reams, G. A. (2005). The enhanced Forest Inventory and Analysis program of the USDA Forest Service: historical perspective and announcement of statistical documentation. *Journal of Forestry*, *103*, 304–308.

McRoberts, R. E., Cohen, W. B., Næsset, E., Stehman, S. V., & Tomppo, E. O. (2010). Using remotely sensed data to construct and assess forest attribute maps and related spatial products. *Scandinavian Journal of Forest Research*, *25*, 340–367.

Mittlböck, M., & Schemper, M. (2002). Explained variation for logistic regression—Small sample adjustments, confidence intervals and predictive precision. *Biometrical Journal*, *44*, 263–272.

Reichenbach, H. (1938). *Experience and prediction.* Chicago: University of Chicago Press.

Rennolls, K. (1982). The use of superpopulation-prediction methods in survey analysis, with application to the British national census of woodlands and trees. In T. B. Brann, L. O. House IV, & H. G. Lund (Eds.), *In place resource inventories: Principles and practices* (pp. 395–401). Bethesda, MD: Society of American Foresters 9-14 August 1981, Orono, ME.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring vegetation systems in the great plains with ERTS. *Proceedings of the Third ERTS Symposium, NASA SP-351*, *Volume 1.* (pp. 309–317). Washington, DC: NASA.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*(2), 377–387.

Royall, R. M., & Herson, J. (1973). Robust estimation in finite populations II. *Journal of the American Statistical Association*, *68*(344), 890–893.

Särndal, C. -E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer-Verlag, Inc. 694 pp.

Simpson, J. A., & Weiner, E. S. C. (Preparers) (1989). *The Oxford English Dictionary, second edition.* Clarendon Press. Oxford. Volume 7, pp 923-924.

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

Stehman, S. V. (2005). Comparing estimators of gross change derived from complete coverage mapping versus statistical sampling of remotely sensed data. *Remote Sensing of Environment*, *96*, 466–474.

Stehman, S. V. (2006). Design, analysis, and inference for studies comparing thematic accuracy of classified remotely sensed data: A special case of map comparison. *Journal of Geographic Systems*, *8*, 209–226.

Stehman, S. V. (2008). Sample designs for assessing map accuracy. In D. Li, Y. Ge, & G. M. Foody (Eds.), *Accuracy in Geomatics, Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, *Volume II.* (pp. 8–15).

Stehman, S. V. (2009). Model-assisted estimation as a unifying framework for area estimation of land cover and land-cover change. *Remote Sensing of Environment*, *113*(11), 2455–2462.

Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment*, *64*, 331–344.

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., et al. (2006). *Global land cover validation: recommendations for evaluation and accuracy assessment of global land cover maps.* EUR 22156 EN. Luxembourg: Office for Official Publications of the European Communities.

Switzer, P. (1969). Mapping a geographically correlated environment. In G. P. Patil, E. C. Pielou, & E. E. Waters (Eds.), *Statistical ecology. Volume 1: Spatial Patterns and Statistical Distributions.* (pp. 235–267).

USDA Forest Service (2000). USDA Research and Development, Code of Scientific Ethics. Available at: http://www.fs.fed.us/research/pdf/fs_code_of%20_scientific_ethics.pdf Last accessed: February 2010.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference.* New York: Wiley 504 pp.