Accuracy and Precision for Remote Sensing Applications of Nonlinear Model-Based Inference

Ronald E. McRoberts, Erik Næsset, and Terje Gobakken

***Abstract*—In a forest inventory context, estimation for small areas and for remote and inaccessible regions may be problematic using traditional probability- or design-based inference because acquisition of sufficiently large samples to satisfy precision requirements is financially and/or logistically difficult. These problems can often be partially alleviated for inventory applications by enhancing inferences using models and remotely sensed independent variables. However, estimates obtained using probability-based, model-assisted estimators may still suffer detrimental effects as the result of small sample sizes. Model-based inference has the potential to alleviate these problems because precision is affected by other factors such as model specification. Nevertheless, model specification in the form of selection of independent variables often focuses exclusively on quality of fit with little consideration given to the precision of estimates of areal population parameters. Model-based inference is illustrated for two forest inventory applications, estimation of mean proportion forest area using Landsat-based independent variables for a study area in the USA and estimation of mean growing stock volume per unit area using lidar-based independent variables for a study area in Norway. Variations of a nonlinear logistic regression model are used for both applications. The results indicate selection of subsets of remotely sensed independent variables to maximize precision had negligible effects on the quality of fit of the models to the data and on estimates of means but substantial proportional beneficial effects on precision.**

***Index Terms*—Landsat, lidar, variable selection.**

I. Introduction

FOREST area and growing stock volume are the two primary national forest inventory (NFI) variables for which parameters are estimated and reported. In addition, these variables, and variables that may be based on them such as biomass and aboveground carbon, are of primary interest in the context of climate change. Traditional NFI approaches to estimating means and totals for these variables rely on probability- or design-based inference which requires probability-based samples. For many inventory and monitoring applications, including in countries with long-established NFIs and tropical developing countries developing carbon-accounting programs, these inferences can be enhanced using models in combination with remotely sensed data [1]–[4]. However, because these approaches still rely substantially on increasing sample sizes to reduce variances of estimates, they have less utility for small area estimation or for remote and inaccessible regions for which even modest sample sizes may be logistically impossible, prohibitively expensive, or both.

Model-based approaches to inference rely on correct model specification for validity rather than probability samples and have the potential to alleviate some problems related to sample size. Current approaches to model-based inference originated in the context of survey sampling [5]–[7]. Given this origin, the frequency of forestry inventory applications of model-based inference is not surprising [3], [8]–[13]. If a small, non-probability sample is sufficient to facilitate selection of a correctly specified model and precise estimation of the model parameters, then inferences for small areas or remote, inaccessible regions may be possible when it is not feasible for probability-based inference [12], [14].

With both probability- and model-based inferential approaches that use models, most attention is focused on selecting a model that adequately characterizes the relationship between observations of response variables and independent variables.

Although selection among competing mathematical forms of nonlinear models may be difficult, selection among competing independent variables for a particular mathematical form is not necessarily difficult. Selection is typically guided by the degree of model adequacy, often characterized as *quality of fit* of the model to the data or inversely *lack of fit*. For model-based inference, quality of fit is particularly important because the unbiasedness of model-based estimators depends on correct model specification. Multiple measures of quality of fit have been proposed and used for both assessing overall quality of fit and comparing models that include different numbers of independent variables and model parameters. The most basic measure is often the sum of squared residual deviations $(\mathrm{SS_{res}})$ between observations and model predictions. Vanclay and Skovsgaard [15] proposed *model efficiency* which is based on the ratio of $\mathrm{SS_{res}}$ and the sum of squared deviations between observations and their mean. Mean square residual $(\mathrm{MS_{res}})$ is the ratio of $\mathrm{SS_{res}}$ and the difference between the number of observations and the number of model parameters. Thus, $\mathrm{MS_{res}}$ includes a penalty for including more model parameters as do the Akaike Information Criterion (AIC) and Mallow's $\mathrm{C_p}$. AIC incorporates both the maximized value of the likelihood function and the number of model parameters [16], whereas $\mathrm{C_p}$ uses the ratio of $\mathrm{SS_{res}}$ for all independent variables to $\mathrm{SS_{res}}$ for a subset of independent variables and the number of model parameters [17]. Although all these measures assess general quality of fit, they do not directly indicate portions of the range of either the dependent

Manuscript received March 29, 2012; revised July 06, 2012, September 27, 2012; accepted October 16, 2012. Date of publication January 08, 2013; date of current version March 11, 2013.

R. E. McRoberts is with the Northern Research Station, U.S. Forest Service, Saint Paul, MN 55108 USA (e-mail: rmcroberts@fs.fed.us).

Erik Næsset and Terje Gobakken are with the Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, 1432 Ås, Norway.

Digital Object Identifier 10.1109/JSTARS.2012.2227299

or independent variables exhibiting lack of fit. The latter task typically entails analyses of residual deviations and/or graphical techniques. A useful approach for this task is described in Section III.D.3).

Much less attention is focused on model selection to maximize precision than is focused on model selection to maximize quality of fit. For model-based inference, the precision of areal estimates depends on the precision of estimates for individual population units which, in turn, depends on the mathematical form of the model, the values of the independent variables, residual variation, and the covariances of the model parameter estimates. For inventory, natural resource, and environmental survey applications that rely on probability samples, little control can typically be exercised over the values of the independent variables. Further, the number of available remotely sensed independent variables is often relatively large in which case assessment focuses on selection of the most parsimonious model with respect to quality of fit of the model to the data. For a particular model form, the effects of different combinations of remotely sensed independent variables on the precision of areal estimates are seldom, if ever, explicitly addressed.

The study focused on estimation of areal population means, $\hat{\mu}$, and variances, $\hat{\mathrm{Var}}(\hat{\mu})$, for purposes of constructing model-based inferences in the form of approximate 95% confidence intervals calculated as $\hat{\mu} \pm 2 \cdot \sqrt{\hat{\mathrm{Var}}(\hat{\mu})}$. In particular, the objective was narrowly focused on the effects of selection of remotely sensed independent variables on the quality of fit of nonlinear models and variances of estimates of areal population means. Because the study was narrowly focused, only simple and intuitive quality of fit criteria and parameter estimation techniques were used to avoid confounding their effects with the effects on precision of independent variable selection. Two datasets were used: observations of forest/non-forest (FOR) from NFI plots and Landsat satellite imagery for a study area in northeastern Minnesota in the United States of America (USA) and observations of growing stock volume (VOL) from NFI plots and lidar data for a study area in Hedmark County, Norway.

II. Data

A. Minnesota

The Minnesota study area was defined by the portion of the row 27, path 27, Landsat scene in northeastern Minnesota, USA (Fig. 1). Imagery was acquired for three dates corresponding to early, peak, and late seasonal vegetative stages: April 2000, July 2001, and November 1999. Spectral data in the form of the normalized difference vegetation index (NDVI) transformation [18] and the three tasseled cap transformations (TC1:brightness, TC2:greenness, and TC3:wetness) [19], [20] for each of the three image dates were used. Land use for the study area consists of forest land dominated by aspen-birch (*Populus spp.-Betula spp.*) and spruce-fir (*Picea spp.-Abies spp.*) associations, agriculture, wetlands, and water. Within the study area, a 15-km radius area of interest (AOI) was selected and served for analytical purposes as a small area within the larger study area.

Field inventory data were obtained for plots established by the Forest Inventory and Analysis (FIA) program of the U.S. Forest Service which conducts the NFI of the USA. The FIA program has established field plot centers in permanent locations using an equal probability sampling design based on a tessellation of the USA into approximate 2400-ha (6000-ac) hexagons that features a permanent plot at a randomly selected location in each hexagon [21]. Each FIA plot consists of four 7.32-m (24-ft) radius circular subplots that are configured as a central subplot and three peripheral subplots with centers located at distances of 36.58 m (120 ft) and azimuths of 0°, 120°, and 240° from the center of the central subplot. In general, centers of forested, partially forested, or previously forested plots are determined using global positioning system (GPS) receivers, whereas centers of non-forested plots are verified using aerial imagery.

Field crews visually estimate the proportion of each subplot that satisfies the FIA definition of forest land: minimum area of 0.4 ha (1.0 ac), minimum crown cover of 10%, minimum crown cover width of 36.6 m (120 ft), and forest land use. Data were obtained between 1999 and 2003 for 2266 plots in the study area. To avoid issues of spatial correlation among observations for subplots of the same plots, data for only the central subplot of each plot were used for this study. Further, to reduce problems associated with observations for the approximate 168-m^2 subplots failing to adequately characterize the larger 900-m^2 Landsat pixels, observations of FOR were restricted to subplots that were either completely forested or completely non-forested. Finally, field crews classify plots with respect to land use rather than land cover. Thus, recently harvested plots with no tree cover are still classified as forest if regeneration is expected. To avoid difficulties distinguishing between non-forest plots and forest plots without tree cover, data for the latter plots were also deleted. All deleted subplots were considered to be a random selection from the original sample. Following deletions, observations for 1902 of the original 2266 subplots remained. Observations of FOR data for central subplots were combined with the values of the spectral transformations for pixels containing subplot centers. For purposes of consistency between the Minnesota and Hedmark (Section II.B) datasets, FIA central subplots are characterized as plots in the remaining sections.

B. Hedmark

The study area is mostly in the municipalities of Åmot and Stor-Elvdal in Hedmark County, Norway (Fig. 2). The study area includes 2385 km^2 and features altitudinal variations ranging from 204-1134 m above sea level (asl) with a mean of 570 m asl. The dominant tree species are Norway spruce (*Picea abies* (L.) Karst.) and Scots pine (*Pinus sylvestris* L.). The study area was tessellated into square 250-m^2 lidar cells that served as population units. Within the study area, a 15-km radius AOI was selected for analysis.

Airborne lidar data were acquired between 15 July 2006 and 12 September 2006 from a height of approximately 1700 m with average aircraft speed of 75 ms^{-1}. The pulse repetition frequency was 50 kHz, the scan frequency was 31 Hz, the maximum scan angle was 16°, which corresponded to an average swath width of approximately 975 m, the mean footprint diameter was approximately 50 cm, and the average point density was 0.7 m^{-2}. Data for only single echoes or the first of multiple

echoes were used. For each plot and population unit, height distributions were estimated for first echoes from tree canopies, i.e., heights greater than 2 m. Echoes with heights less than 2 m were considered to have been reflected from non-tree objects such as shrubs, grass, or the ground. For each plot and population unit, heights corresponding to the 10th, 20th, ..., 100th percentiles of the distributions were denoted $h_{10}, h_{20}, \ldots, h_{100}$, respectively. Canopy densities were calculated as the proportions of echoes with heights greater than 0%, 10%, ..., 90% of the range between 2 m above the ground and the 95th percentile height and denoted $d_0, d_{10}, \ldots, d_{90}$, respectively [22]. To avoid problems of excessive correlation among observations of different lidar variables, only h_{10}, h_{30}, h_{50}, h_{70}, h_{100}, d_0, d_{20}, d_{40}, d_{60}, and d_{90} were used for this study.

Field measurements were obtained for 145 circular 250-m^2 Norwegian NFI field plots located at the intersections of a 3-km × 3-km grid [23] and measured between 2005 and 2007. On each plot, all trees with diameters at-breast-height (dbh, 1.3 m) of at least 5 cm were callipered. Heights were measured on an average of 10 sample trees per plot selected with probability proportional to stem basal area, and heights were predicted using height-dbh models for trees whose heights were not measured [24], [25]. The volume of each sample tree was estimated using species-specific volume models with dbh and either measured height or predicted height as predictor variables [26]–[28]. The ratio of the mean volume obtained using predicted heights and the mean volume obtained using measured heights was used to adjust the former volume. Total plot volume was obtained by adding volume estimates for individual trees. Although the plot volumes are estimates, the uncertainties of the estimates were considered negligible relative to other sources of uncertainty and were ignored. A variogram analysis indicated no meaningful spatial correlation among VOL plot observations. Differential Global Navigation Satellite Systems (GPS and the Russian GLONASS) were used to determine the position of the center of each plot.

III. Methods

All analyses were based on three underlying assumptions: (1) a finite population consisting of N units in the form of the square, 900-m^2 Landsat pixels for the Minnesota study area or the square 250-m^2 lidar cells for the Hedmark study area, (2) a sample of n population units in the form of pixels or lidar cells that contain NFI plot centers, and (3) availability of auxiliary data in the form of Landsat spectral transformations for all pixels or lidar height and density metrics for all lidar cells. In the following sections, the term *population unit* is considered synonymous with the terms *pixel* and *lidar cell*.

A. Binomial Logistic Regression Model

The relationship between a dichotomous dependent variable such as FOR, here denoted Y (assumes values $y = 0$ and $y = 1$), and continuous independent variables, $\mathbf{X}$, is often expressed in the form

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

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 [29]. The function, $f(\mathbf{X}_i; \boldsymbol{\beta})$, expresses the statistical expectation of Y in terms of $\mathbf{X}$ 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 $\exp(\cdot)$ is the exponential function. For future reference, the model expressed by (1) is designated the *binomial logistic regression model*. The parameters are estimated by maximizing the natural logarithm of the likelihood, L,

$$l = \ln(L) = \sum_{i=1}^{n} p_i^{y_i}(1 - p_i)^{(1-y_i)} = \sum_{i=1}^{n} f(\mathbf{X}_i; \boldsymbol{\beta})^{y_i}[1 - f(\mathbf{X}_i; \boldsymbol{\beta})]^{1-y_i},$$

and the parameter covariance matrix is estimated as

$$\hat{V}_{\hat{\beta}} = -D^{-1}, \quad (2)$$

where the elements of D are [29]

$$d_{ij} = \sum_{i=1}^{n} (\partial^2 l(\mathbf{X}_i; \hat{\boldsymbol{\beta}}))/(\partial\beta_j \partial\beta_k) = \sum_{i=1}^{n} \hat{y}_i(1 - \hat{y}_i)x_{ij}x_{ik}.$$

B. Asymptotic Logistic Regression Model

A nonlinear logistic regression model was used to describe the relationship between VOL for NFI plots and associated lidar metrics. The model has the mathematical form

$$y_i = f(\mathbf{X}_i; \boldsymbol{\beta}) = \frac{\beta_{J+2}}{1 + \exp\left(\beta_{J+1} + \sum_{j=1}^{J} \beta_j x_{ij}\right)} + \varepsilon_i, \quad (3)$$

where i indexes population units, x_{ij} is the jth lidar metric, the βs are parameters to be estimated, and ε_i is a residual term. For future reference, the model expressed by (3) is designated the *asymptotic logistic regression model* to distinguish it from the binomial logistic regression model. An advantage of the nonlinear model expressed by (3) over a linear model is that all predictions are non-negative and are constrained by the horizontal asymptote, β_{J+2}, which is estimated from the sample data.

The model parameters were estimated using iteratively re-weighted least squares techniques, and the parameter covariance matrix was estimated as

$$\hat{V}_{\hat{\beta}} = (Z_i' W Z_i)^{-1}, \quad (4)$$

where $z_{ij} = (\partial f(\mathbf{X}_i; \hat{\boldsymbol{\beta}}))/(\partial\beta_j)$ and W is a matrix of weights. Because spatial correlation among observations was negligible for the Hedmark dataset, W can be expressed as a diagonal matrix with $w_{ii} = \hat{\sigma}_i^{-2}$. Although the variance estimator expressed by (4) is based on a first-order Taylor series approximation, the

intrinsic nonlinearity in the mathematical form of the model is negligible, meaning that the first-order variance approximations are expected to be quite good [30], [31].

C. Model-Based Inference

The assumptions underlying model-based inference differ considerably from the assumptions underlying the more familiar probability- or design-based inference. With model-based inference, the observation for a population unit is a random variable whose value is considered a realization from a distribution of possible values, rather than a fixed value as is the case for probability-based inference. Further, the basis for model-based inference is the validity of the model, not the probabilistic nature of the sample as is the case for probability-based inference. Randomization for model-based inference enters through the random realizations from the distributions for individual population units, whereas randomization for probability-based inference enters through the random selection of population units into the sample. One of the greatest impacts of the different assumptions underlying model-based inference and probability-based inference is the form of variance estimators, regardless of the nature of the sample.

For model-based inference, the mean and standard deviation of the distribution of Y for the ith population unit may be denoted μ_{i} and σ_{i}, respectively. The mean is estimated as $\hat{\mu}_{\mathrm{i}} = \hat{p}_{\mathrm{i}}$ obtained using (1) for the binomial logistic regression model or as $\hat{\mu}_{\mathrm{i}} = \hat{\mathrm{y}}_{\mathrm{i}}$ obtained using (3) for the asymptotic logistic regression model. The standard deviation is estimated as the residual standard deviation obtained from deviations between observations and model predictions. The model-based estimator of the population mean is based on the set of estimates, $\{\hat{\mu}_{\mathrm{i}}, \mathrm{i} = 1, 2, \ldots, \mathrm{N}\}$, of the means for individual population units:

$$\hat{\mu} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \hat{\mu}_{\mathrm{i}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{f}(\mathbf{X}_{\mathrm{i}}; \hat{\boldsymbol{\beta}}), \tag{5}$$

where $\mathrm{f}(\mathrm{X}_{\mathrm{i}}; \hat{\beta})$ is from either (1) or (3) and is evaluated at the parameter estimates. The corresponding variance can be estimated as

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{\mu}) = \frac{1}{\mathrm{N}^2} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}} \hat{\mathrm{C}}\mathrm{ov}(\hat{\mu}_{\mathrm{i}}, \hat{\mu}_{\mathrm{j}}), \tag{6}$$

where

$$\hat{\mathrm{C}}\mathrm{ov}(\hat{\mu}_{\mathrm{i}}, \hat{\mu}_{\mathrm{j}}) = \mathrm{Z}_{\mathrm{i}}' \hat{\mathrm{V}}_{\hat{\beta}} \mathrm{Z}_{\mathrm{i}},$$

and again $z_{\mathrm{ij}} = (\partial \mathrm{f}(\mathbf{X}_{\mathrm{i}}; \hat{\boldsymbol{\beta}}))/(\partial \beta_{\mathrm{j}})$.

D. Analyses

1) Synthetic Estimators: For this study, relationships between response variables and independent variables were assumed to be the same for the entirety of each study area. Thus, model parameters were estimated using data for all plots in the study area, even though the model was applied only to the smaller 15-km radius AOIs. The corresponding statistical estimators are characterized as *synthetic estimators* because they apply information for the entire study areas to subsets of the study areas [32].

2) Independent Variable Selection: Independent variables were selected on the basis of measures of the relationships between observations and model predictions. For the binomial logistic regression model, the selection criterion was *accuracy* defined as the proportion of correct categorical classifications of sample observations where

$$\hat{\mathrm{y}}_{\mathrm{i}} = \begin{cases} 0 & \text{if } \hat{\mathrm{p}}_{\mathrm{i}} < 0.5 \\ 1 & \text{if } \hat{\mathrm{p}}_{\mathrm{i}} \geq 0.5 \end{cases}.$$

For the asymptotic logistic regression model, the selection criterion was *model efficiency* [15] denoted and calculated as

$$\mathrm{Q}^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{mean}}},$$

where

$$\mathrm{SS}_{\mathrm{res}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})^2$$

and

$$\mathrm{SS}_{\mathrm{mean}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y}_{\mathrm{i}} - \bar{\mathrm{y}})^2.$$

Independent variables were selected using an iterative procedure whereby the single independent variable producing the greatest value of the selection criterion was first selected. Next, all combinations of the first selected variable with one of the remaining variables were evaluated, and the combination producing the greatest value of the selection criterion was selected. This procedure of selecting the independent variable that best augmented the previously selected independent variables continued until the combination producing the maximum value of the selection criterion was obtained for each number of independent variables. Although this iterative approach does not guarantee selection of the optimal combination of independent variables, assessment of all possible combinations is extremely laborious for nonlinear models, and more sophisticated approaches such as backward and forward selection are difficult to implement for nonlinear models. For the best combination of each number of independent variables obtained using the iterative procedure, $\hat{\mu}$ was calculated using (5), and $\hat{\mathrm{V}}\mathrm{ar}(\hat{\mu})$ was calculated using (6).

3) Quality of Fit of Models to Data: Model-based estimators are not necessarily unbiased, or even nearly unbiased, as is often the case for probability-based estimators. Because unbiasedness for model-based estimators is closely linked to correct model specification [33], the quality of fit of the model to the data was further assessed using an additional criterion which was estimated using a three-step approach: (1) all pairs of observations and predictions, $(\mathrm{y}_{\mathrm{i}}, \hat{\mathrm{y}}_{\mathrm{i}})$, were ordered with respect to $\hat{\mathrm{y}}_{\mathrm{i}}$; (2) the ordered pairs were grouped into categories of approximately equal numbers of pairs, and the group means of the observations and of the predictions were calculated; and (3) a graph of the observation means versus the prediction means was constructed [34]. If the model is correctly specified, the graph of means of observations versus means of model predictions

Fig. 1. Minnesota study area with plot locations.

should feature points that lie along the 1:1 line with intercept 0 and slope 1.

IV. Results and Discussion

Graphs of means of ordered observations versus means of ordered predictions indicated no substantial lack of fit of the models to the data (Figs. 3, 4). For the bionomial logistic regression model, the accuracies of the FOR predictions were approximately 0.90 which is comparable to results obtained for similar studies in the same study area, albeit using different techniques [35], [36]. For the asymptotic logistic regression model, Q^2 values were approximately 0.84 which is comparable to results obtained for similar studies [37], [38].

For the Minnesota AOI using the binomial logistic regression model with Landsat data to predict FOR, accuracies stabilized at approximately 0.90–0.91 for four or more remotely sensed independent variables, and corresponding estimates of the mean ranged only from 0.8340 to 0.8428 (Table I). For four or more remotely sensed independent variables, standard errors ranged from 0.0063 to 0.0076 and as proportions of estimates of means were less than 0.01. Although the standard errors were small and their range was small, the ratio of the largest to the smallest was 1.21 which represents a substantial proportional difference. The smallest standard error, 0.0063, was found for seven remotely sensed independent variables, whereas the largest standard error, 0.0076, was found when using all 12 variables. Of particular interest, the combination of independent variables that maximized precision was not the same combination that maximized accuracy. However, selection of the combination of independent variables that maximized precision had only minimal effects on accuracy or estimates of the AOI mean, but reduced the standard error of the areal mean proportionally by more than 0.20.

For the Hedmark AOI using the asymptotic logistic regression model with lidar data to predict VOL, Q^2 values stabilized at approximately 0.84 to 0.86 for four or more remotely sensed independent variables, and corresponding estimates of the means ranged only from 85.1114 to 87.4914 (Table II). For four or more remotely sensed independent variables, standard errors ranged from 3.8211 to 4.1652 and as proportions of estimates of means ranged from 0.045 to 0.050. Although the standard errors were small and their range was small, the ratio of the largest to the smallest was 1.09 which, again, could be considered to represent a substantial proportional difference. The smallest standard error, 3.8211, was found for four remotely sensed independent variables whereas the largest standard error, 4.1652, was found when using all 10 variables. As for the Minnesota study area, the combination of independent variables that

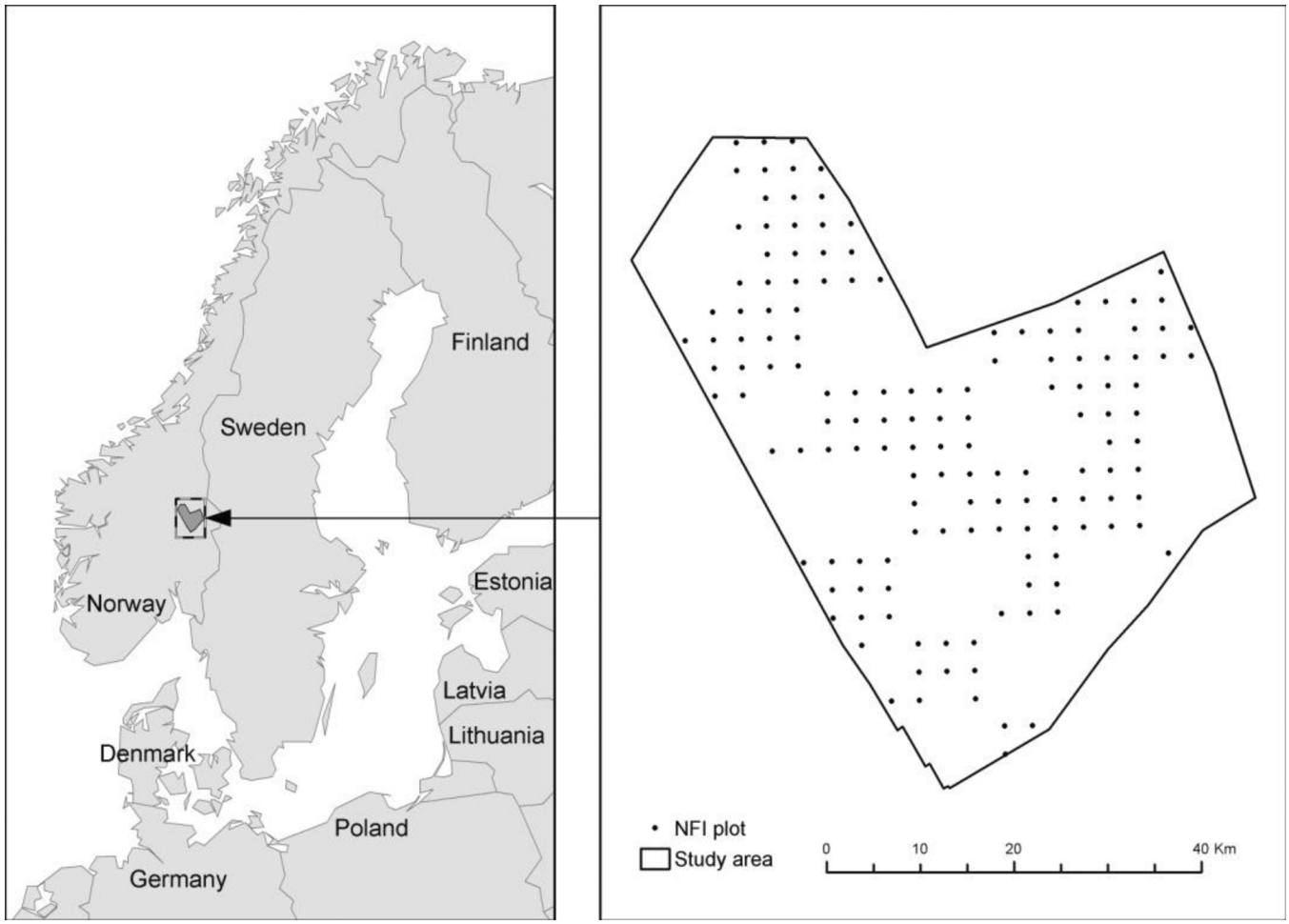

Fig. 2. Hedmark study area with plot locations.

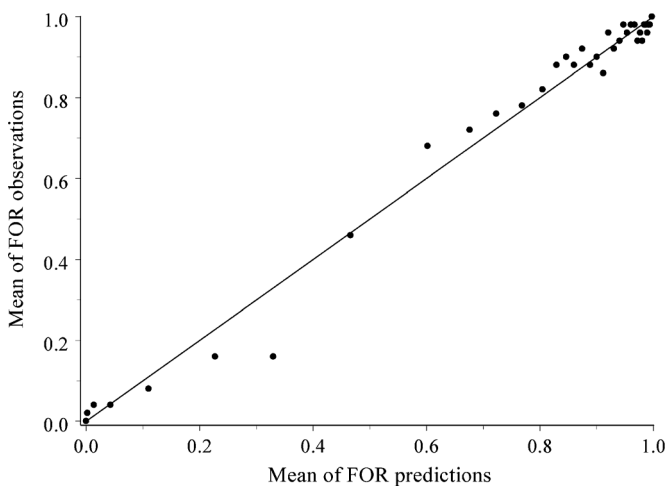

Fig. 3. Means of forest/non-forest (FOR) observations versus means of binomial logistic regression model predictions.

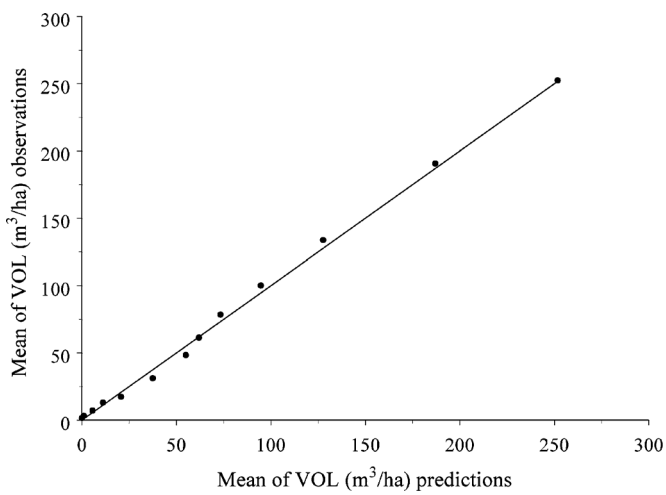

Fig. 4. Means of growing stock volume (VOL) observations versus means of asymptotic logistic regression model predictions.

maximized precision was not the same combination that maximized the measure of quality of fit of the model to the data. However, again, selection of the combination of independent variables that maximized precision had only minimal effects on quality of fit or estimates of the AOI mean, but reduced the standard error of the AOI mean proportionally by nearly 0.09.

The study was narrowly focused on assessing the effects of selection of independent variables on the precision of area estimates. In support of this narrow focus, only simple and intuitive selection criteria and parameter estimation techniques were used to avoid confounding their effects with the effects on precision. Criteria other than accuracy or modeling efficiency would

TABLE I
RESULTS FOR FOR USING BINOMIAL LOGISTIC REGRESSION MODEL AND LANDSAT DATA

No. of variables	Accuracy	Mean	SE	Variables*
1	0.8170	0.8050	0.0073	10
2	0.8512	0.8363	0.0066	10-2
3	0.8896	0.8219	0.0071	10-2-11
4	0.9017	0.8340	0.0070	10-2-11-7
5	0.9048	0.8383	0.0067	10-2-11-7-12
6	0.9054	0.8412	0.0063	10-2-11-7-12-8
7	0.9075	0.8406	0.0065	10-2-11-7-12-8-5
8	0.9075	0.8388	0.0066	10-2-11-7-12-8-5-1
9	0.9096	0.8399	0.0070	10-2-11-7-12-8-5-1-6
10	0.9096	0.8388	0.0072	10-2-11-7-12-8-5-1-6-3
11	0.9075	0.8428	0.0075	10-2-11-7-12-8-5-1-6-3-4
12	0.9064	0.8413	0.0076	10-2-11-7-12-8-5-1-6-3-4-9

*Variable definitions: 1=$TC1_{spr}$, 2=$TC2_{spr}$, 3=$TC3_{spr}$, 4=$TC1_{sum}$, 5=$TC2_{sum}$, 6=$TC3_{sum}$, 7=$TC1_{fal}$, 8=$TC2_{fal}$, 9=$TC3_{fal}$, 10=$NDVI_{spr}$, 11=$NDVI_{sum}$, 12=$NDVI_{fal}$ where *spr* designates spring, *sum* designates summer, *fal* designates fall (leaf off).

TABLE II
RESULTS FOR VOL USING NONLINEAR LOGISTIC REGRESSION MODEL AND LIDAR DATA

No. of variables	Q^2	Mean	SE	Variables
1	0.7247	82.5900	3.4390	d_{90}
2	0.7844	85.8967	3.6826	d_{90}-h_{100}
3	0.8388	84.6985	3.7648	d_{90}-h_{100}-d_0
4	0.8437	85.1114	3.8211	d_{90}-h_{100}-d_0-h_{10}
5	0.8441	85.0002	3.8339	d_{90}-h_{100}-d_0-h_{10}-d_{20}
6	0.8452	85.2394	3.8482	d_{90}-h_{100}-d_0-h_{10}-d_{20}-d_{40}
7	0.8480	84.8954	3.8379	d_{90}-h_{100}-d_0-h_{10}-d_{20}-d_{40}-d_{60}
8	0.8490	85.2912	3.8602	d_{90}-h_{100}-d_0-h_{10}-d_{20}-d_{40}-d_{60}-h_{70}
9	0.8501	86.1885	3.9609	d_{90}-h_{100}-d_0-h_{10}-d_{20}-d_{40}-d_{60} -h_{70}-h_{50}
10	0.8561	87.4914	4.1652	d_{90}-h_{100}-d_0-h_{10}-d_{20}-d_{40}-d_{60} -h_{70}-h_{50}-h_{30}

be expected to produce slightly different independent variable selections, slightly different parameter estimates, and slightly different residual variation. However, the fundamental nature of the problem would be unchanged in the sense that the precision of areal estimates and estimates for individual population units would still depend on the same factors, particularly selection of independent variables and covariances of the model parameter estimates. For linear models, Taylor series methods produce exact parameter covariance estimates, whereas for nonlinear models the covariance estimates are only approximations. Factors that affect the accuracy of approximations are primarily the mathematical form of the model and the quality of fit of the model to the data. As previously noted, the mathematical form of the logistic regression model is known to have desirable properties in this regard [30], [31]. In addition, the accuracies and Q^2 values for this study were sufficiently large to minimize adverse effects. However, when Taylor series covariance approximations are poor, Monte Carlo approaches are viable alternatives [39]. Finally, many forest inventories use clusters of plots leading to large correlations among observations for plots in the same cluster. The estimation complexity necessary to accommodate these correlations was avoided for this study by using data for only one plot in each cluster. Other studies on model-based inference using FIA data indicate negligible effects on areal precision when the correlations are accommodated in the estimation procedure [12].

V. CONCLUSION

Four conclusions may be drawn from the study. First, nonlinear models are useful for describing relationships between NFI attributes and remotely sensed independent variables. Second, model-based inference, which is not routinely used for NFI applications, has the potential to produce credible estimates and confidence intervals for parameters associated with NFI variables. Applications for small and inaccessible areas for which large samples are not feasible are of particular interest. However, care must be exercised to assess model quality of fit and to use correct variance estimators. Third, the standard errors of estimates of means were uniformly small for these areas of interest. Fourth, selection of subsets of remotely sensed independent variables that are optimal with respect to precision produced only negligible effects on quality of fit of the model to the data and estimates of population means, but had substantial beneficial proportional effects on standard errors. Of interest, for both study areas and datasets, the combination of independent variables that maximized precision of the areal mean was not the same combination that maximized quality of fit of the model to the data.

REFERENCES

[1] R. E. McRoberts, "Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data," *Remote Sens. Environ.*, vol. 114, pp. 1017–1025, 2010.

[2] T. G. Gregoire, G. Ståhl, E. Næsset, T. Gobakken, R. Nelson, and S. Holm, "Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway," *Can. J. Forest Res.*, vol. 41, pp. 83–95, 2011.

[3] G. Ståhl, S. Holm, T. G. Gregoire, T. Gobakken, E. Næsset, and R. Nelson, "Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway," *Can. J. Forest Res.*, vol. 41, pp. 96–107, 2011.

[4] R. E. McRoberts, T. Gobakken, and E. Næsset, "Post-stratified estimation of forest area and growing stock volume using lidar-based stratifications," *Remote Sens. Environ.*, vol. 125, pp. 157-166.

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

[6] K. R. W. Brewer, "Ratio estimation in finite populations: Some results deductible from the assumption of an underlying stochastic process," *Austral. J. Statist.*, vol. 5, pp. 93–105, 1963.

[7] R. M. Royall, "On finite population sampling theory under certain linear regression models," *Biometrika*, vol. 57, no. 2, pp. 377–387, 1970.

[8] K. Rennolls, "The use of superpopulation-prediction methods in survey analysis, with application to the British National Census of Woodlands and Trees," in H. G. Lund (Ed.), *In Place Resource Inventories: Principles and Practices*. 9–14 Aug. 1981, Orono, ME, Bethesda, MD: Society of American Foresters, pp. 395−401, 1982.

[9] A. Kangas, "Small area estimation using model based methods," *Can. J. Forest Res.*, vol. 26, pp. 758–766, 1996.

[10] T. G. Gregoire, "Design-based and model-based inference: Appreciating the difference," *Can. J. Forest Res.*, vol. 28, pp. 1429–1447, 1998.

[11] A. Kangas and M. Maltamo, *Forest Inventory, Methodology and Applications*. Dordrecht, The Netherlands: Springer, 2006.

[12] R. E. McRoberts, "A model-based approach to estimating forest area," *Remote Sens. Environ.*, vol. 103, pp. 56–66, 2006.

[13] D. Mandallaz, *Sampling Techniques for Forest Inventories*. New York: Chapman & Hall, 2008.

[14] R. E. McRoberts, E. Tomppo, and E. Næsset, "Advances and emerging issues in national forest inventories," *Scand. J. Forest Res.*, vol. 25, pp. 368–381, 2010.

[15] J. K. Vanclay and J. P. Skovsgaard, "Evaluating forest growth models," *Ecological Modelling*, vol. 98, pp. 1–12, 1997.

[16] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Contr.*, vol. 19, pp. 716–723, 1974.

[17] C. L. Mallows, "Some comments on C_P," *Technometrics*, vol. 15, pp. 661–675, 1973.

[18] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the great plains with ERTS," in *Proc. 3rd ERTS Symp.*, Washington, DC, 1973, vol. 1, pp. 309–317, NASA SP-351.
[19] R. J. Kauth and G. S. Thomas, "The tasseled cap—A graphic description of the spectral-temporal development of agricultural crops as seen by Landsat," in *Proc. Symp. Machine Processing of Remotely Sensed Data*, Purdue Univ., West Lafayette, IN, 1976, pp. 41–51.
[20] E. P. Crist, R. C. Cicone, and R. C. , "Application of the tasseled cap concept to simulated thematic mapper data," *Photogramm. Eng. Remote Sens.*, vol. 50, pp. 343–352, 1984.
[21] R. E. McRoberts, W. A. Bechtold, P. L. Patterson, C. T. Scott, and G. A. Reams, "The enhanced forest inventory and analysis program of the USDA forest service: Historical perspective and announcement of statistical documentation," *J. Forestry*, vol. 103, no. 6, pp. 304–308, 2005.
[22] T. Gobakken and E. Næsset, "Assessing effects of laser point density, ground sampling intensity, and field plot sample size on biophysical stand properties derived from airborne laser scanner data," *Can. J. Forest Res.*, vol. 38, pp. 1095–1109, 2008.
[23] S. M. Tomter, G. Hylen, and J.-E. Nilsen, "Development of Norway's national forest inventory," in *National Forest Inventories—Pathways for Common Reporting*, E. Tomppo, T. Gschwantner, M. Lawrence, and R. E. McRoberts, Eds. New York: Springer, 2010, pp. 411–424.
[24] A. Fitje and E. Vestjordet, "Stand height curves and new tariff tables for Norway spruce," *Reports of the Norwegian Forest Research Institute*, vol. 34, pp. 23–62, 1977, (in Norwegian with English summary).
[25] E. Vestjordet, "Merchantable volume of Norway spruce and Scots pine based on relative height and diameter at breast height or 2.5 m above stump level," *Meddelelser fra Det Norske Skogforsøksvesen*, vol. 25, pp. 411–459, 1968, (in Norwegian with English Summary.).
[26] H. Braastad, "Volume tables for birch," *Meddelelser fra Det Norske Skogforsøksvesen*, vol. 21, pp. 23–78, 1966, (in Norwegian with English summary).
[27] A. Brantseg, "Volume functions and tables for Scots pine. South Norway," *Meddelelser fra Det Norske Skogforsøksvesen*, vol. 22, pp. 689–739, 1967, (in Norwegian with English summary).
[28] E. Vestjordet, "Functions and tables for volume of standing trees. Norway spruce," *Meddelelser fra Det Norske Skogforsøksvesen*, vol. 22, pp. 539–574, 1967, (in Norwegian with English summary).
[29] A. Agresti, *An Introduction to Categorical Data Analysis*. Hoboken, NJ: Wiley-Interscience, 2007.
[30] D. A. Ratkowsky, *Nonlinear Regression Modelling: A Unified Practical Approach*. New York: Dekker, 1983.
[31] L. M. Haines, T. E. Brien, and G. P. Y. Clark, "Kurtosis and curvature measures for nonlinear regression models," *Statistica Sinica*, vol. 14, pp. 547–570, 2004.
[32] C. Särndal, B. Swensson, and J. Wretman, *Model-Assisted Survey Sampling*. New York: Springer-Verlag, 1992.
[33] S. Lohr, *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury, 1999.
[34] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York: Wiley, 1989.
[35] R. Haapanen, "Delineation of forest/nonforest land use classes using nearest neighbor methods," *Remote Sens. Environ.*, vol. 89, pp. 265–271, 2004.
[36] A. O. Finley, S. Banerjee, and R. E. McRoberts, "A Bayesian approach to multi-source forest area estimation," *Environmental and Ecological Statistics*, vol. 15, no. 2, pp. 241–258, 2008.
[37] E. Næsset, "Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data," *Remote Sens. Environ.*, vol. 80, pp. 88–99, 2002.
[38] E. Næsset and T. Gobakken, "Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser," *Remote Sens. Environ.*, vol. 112, pp. 3079–3090, 2008.
[39] R. E. McRoberts and V. C. Lessard, "Estimating the uncertainty in diameter growth model predictions and its effects on the uncertainty of annual inventory estimates," in *Proc. 2nd Annual Forest Inventory and Analysis Symp.*, G. A. Reams, R. E. McRoberts, and P. C. Van Deusen, Eds., General Technical Report SRS–47. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station, pp. 70-75, 2010.

Ronald E. McRoberts received the Ph.D. degree in biostatistics from the University of Minnesota.

He is a mathematical statistician with the Forest Inventory and Analysis program of the Northern Research Station, U.S. Forest Service, St. Paul, MN. His areas of research interest include nonlinear modeling, nearest neighbors techniques, and remote sensing-based inference. He has published approximately 100 papers in international peer-reviewed journals, has consulted on forest inventory around the world, and has served in leadership positions with IUFRO and ISPRS.

Dr. McRoberts is a Fellow of the American Statistical Association, an elected member of the International Statistical Institute, and an elected member of the Italian Academy of Forest Science.

Erik Næsset received the M.Sc. degree in forestry and the Ph.D. degree in forest inventory from the Agricultural University of Norway, Ås, Norway, in 1983 and 1992, respectively.

His major field of research is forest inventory and remote sensing, with particular focus on operational management inventories, sample surveys, photogrammetry, and airborne light detection and ranging (LiDAR). He has played a major role in developing and implementing airborne LiDAR in operational forest inventory. He has been the Leader and Coordinator of more than 50 research programs funded by the Research Council of Norway, the European Union, and private forest industry. He has published around 100 papers in international peer-reviewed journals. His teaching includes lectures and courses in forest inventory, remote sensing, geographic information systems, forest planning, and sampling techniques.

Terje Gobakken received the M.Sc. degree and the Ph.D. degree from the Agricultural University of Norway, Ås, in 1995 and 2001, respectively.

He has been working at the Norwegian National Forest Inventory and he is now a Professor at the Norwegian University of Life Sciences. His major field of research is forest inventory and airborne light detecting and ranging (LiDAR). He is teaching courses in GIS and long-term forest planning.