Remote Sensing of Environment 154 (2014) 226–233

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Estimation for inaccessible and non-sampled forest areas using model-based inference and remotely sensed auxiliary information

Ronald E. McRoberts [a], Erik Næsset [b], Terje Gobakken [b]

[a] Northern Research Station, U.S. Forest Service, 1992 Folwell Avenue, Saint Paul, MN USA
[b] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway

ARTICLE INFO

Article history:
Received 9 May 2014
Received in revised form 14 August 2014
Accepted 15 August 2014
Available online xxxx

Keywords:
Landsat
Lidar
Precision

ABSTRACT

For remote and inaccessible forest regions, lack of sufficient or possibly any sample data inhibits estimation and construction of confidence intervals for population parameters using familiar probability- or design-based inferential methods. Although maps based on remotely sensed data may provide information on the distribution of resources, map-based estimates are subject to classification and prediction error, and map accuracy measures do not directly inform the uncertainty of the estimates. Model-based inference does not require probability samples and when used with synthetic estimation can circumvent small or no-sample difficulties associated with probability-based inference. The study focused on estimating proportion forest area using Landsat data for a study area in Minnesota, USA, and aboveground biomass using airborne laser scanning data for a study area in Hedmark County, Norway. For both study areas, model-based inference was used to estimate the components necessary for constructing confidence intervals for population means for non-sampled areas. The estimates were compared to simple random sampling, model-assisted, and model-based estimates that would have been obtained if the areas had been sampled. All estimates were within two simple random sampling standard errors of each other, thereby illustrating the utility of model-based inference for non-sampled areas.

Published by Elsevier Inc.

1. Introduction

1.1. Background and motivation

Technical objectives for sample surveys, of which a forest inventory is an example, include construction of inferences in the form of confidence intervals for population parameters. The Oxford English Dictionary defines the term *infer* as "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 this context, 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 1–α confidence intervals,

$$\hat{\mu} \pm t_{1-\alpha} \cdot \sqrt{\hat{\mathrm{Var}}(\hat{\mu})}, \quad (1)$$

where 1–α denotes the probability that confidence intervals constructed using data for all possible samples will include μ. Thus, the inference problem focuses on $\hat{\mu}$ and $SE(\hat{\mu}) = \sqrt{\hat{\mathrm{Var}}(\hat{\mu})}$.

Two approaches to inference are relevant, the familiar probability- or design-based inference and model-based inference. Probability-based inference requires a probability sample and for sufficiently large samples produces estimates with acceptable precision. However, when only small samples can be acquired, particularly for highly variable populations, probability-based inference may fail to produce acceptably precise results. In addition, when no ground sampling is possible because the area of interest is remote or inaccessible and other information such as fine resolution remotely sensed data is lacking, probability-based inference is not possible. Examples include some tropical forests to be surveyed under the auspices of programs such as the United Nations initiative on Reducing Emissions due to Deforestation and Forest Degradation in developing countries and large, remote boreal regions such as interior Alaska in the United States of America (USA).

A general consensus is that inference for remote and inaccessible regions must rely on remotely sensed data, possibly in the form of maps. Of importance, however, maps only rarely accurately depict populations and provide no direct estimates of population parameters that are the primary survey objectives. Further, even if map unit predictions are aggregated to produce an estimate, map accuracy indices provide no direct information regarding the bias of the estimator resulting from classification and prediction errors or the precision of the estimate (McRoberts, 2011) and, therefore, cannot directly contribute to constructing inferences.

http://dx.doi.org/10.1016/j.rse.2014.08.028
0034-4257/Published by Elsevier Inc.

An alternative form of inference, characterized as model-based inference, has the potential to circumvent at least some of the difficulties associated with survey inference for remote and inaccessible regions. The validity of model-based inference is based on correct model specification rather than probability samples. Therefore, when combined with synthetic estimation which uses information external to the area of interest (Särndal et al., 1992), model-based inference can be used for remote and inaccessible regions for which probability samples are logistically difficult or financially prohibitive.

1.2. Objectives

The primary objective was to compare estimates obtained using model-based inference for a study area lacking sample data to estimates obtained using both model-based and model-assisted inference for the same study area when sample data were available. For a study area in Minnesota, USA, Landsat data were used to construct inferences for proportion forest area, and for a study area in Hedmark County, Norway, airborne laser scanning (ALS) data were used to estimate mean aboveground biomass per unit area (AGB).

2. Data

2.1. Minnesota, USA, study area

The study area was defined by the portion of the row 27, path 27, Landsat scene in northern Minnesota, USA, that was cloud-free for 16 July 2002 (Fig. 1). The 30-m × 30-m image pixels served as population units. Four smaller, 700-km^2 areas of interest (AOI) within the study area were also selected. Spectral data in the form of the normalized difference vegetation index (NDVI) transformation (Rouse, Haas, Schell, & Deering, 1973) and the three tasseled cap (TC) transformations (brightness, greenness, and wetness) (Crist & Cicone, 1984; Kauth & Thomas, 1976) were used as auxiliary information.

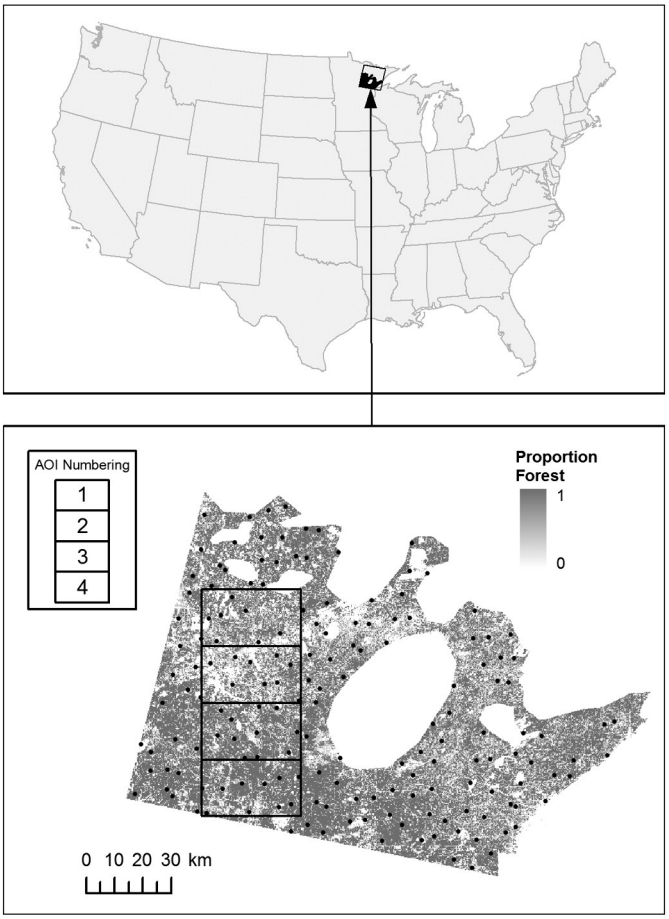

Fig. 1. Minnesota study area with four 700-km^2 areas of interest and inventory plots.

Ground data were obtained for plots established by the Forest Inventory and Analysis (FIA) program of the U.S. Forest Service which conducts the national forest inventory (NFI) of the USA. The FIA program has established field plot centers in permanent locations using a sampling design that is regarded as producing an equal probability sample. 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 (McRoberts, Bechtold, Patterson, Scott, & Reams, 2005; McRoberts, Hansen, & Smith, 2010). In general, centers of forested, partially forested, or previously forested plots are determined using global positioning system (GPS) receivers with accuracies of 10 m or greater, and centers of non-forested plots are verified using aerial imagery and digitization methods. 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. Subplot-level proportion forest was combined with the values of the spectral transformations for pixels containing subplot centers.

Because the smaller 168.3-m^2 subplots may not adequately characterize the larger 900-m^2 TM pixels, subplots whose observations were not completely forested or completely non-forested were deleted and assumed to be missing at random. In addition, to avoid issues related to spatial correlation among observations of subplots of the same plot, data for only the central subplot of each plot were used for this study. Subsequent to deletions, data for 168 plots measured in 2002 were available for the study. For future reference, the term *plot* refers to the central subplot of each FIA plot cluster.

2.2. Hedmark, Norway, study area

The study area was in the municipalities of Åmot and Stor-Elvdal in Hedmark County, Norway (Fig. 2). Four smaller, 100-km^2 AOIs within the study area were selected. The entire study area was tessellated into square 250-m^2 cells that served as population units.

ALS data were acquired between 15 July 2006 and 12 September 2006 with average density of 0.7 pulses/m^2. For each plot and population unit, height distributions were estimated for first echoes with heights greater than 2 m, and two sets of ALS metrics were calculated (Gobakken & Næsset, 2008). The first set of metrics consisted of heights corresponding to the 10th, 20th, …,100th percentiles of the distributions which were denoted h_{10}, h_{20}, …, h_{100}, respectively. The second set of metrics consisted of canopy densities calculated as the proportions of the same echoes with heights greater than 0%, 10%, …, 90% of the range between 2 m above the ground and the 95th percentile height and were denoted d_0, d_{10}, …,d_{90}, respectively.

Field measurements were obtained for 145 circular 250-m^2 Norwegian NFI field plots 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. AGB was estimated at the plot-level using models, and any model prediction errors were ignored. Differential Global Navigation Satellite Systems (GPS and the Russian GLONASS) were used to determine the positions of the centers of plots with accuracies on the order of a few cm. Gobakken et al. (2012) describe this dataset in greater detail.

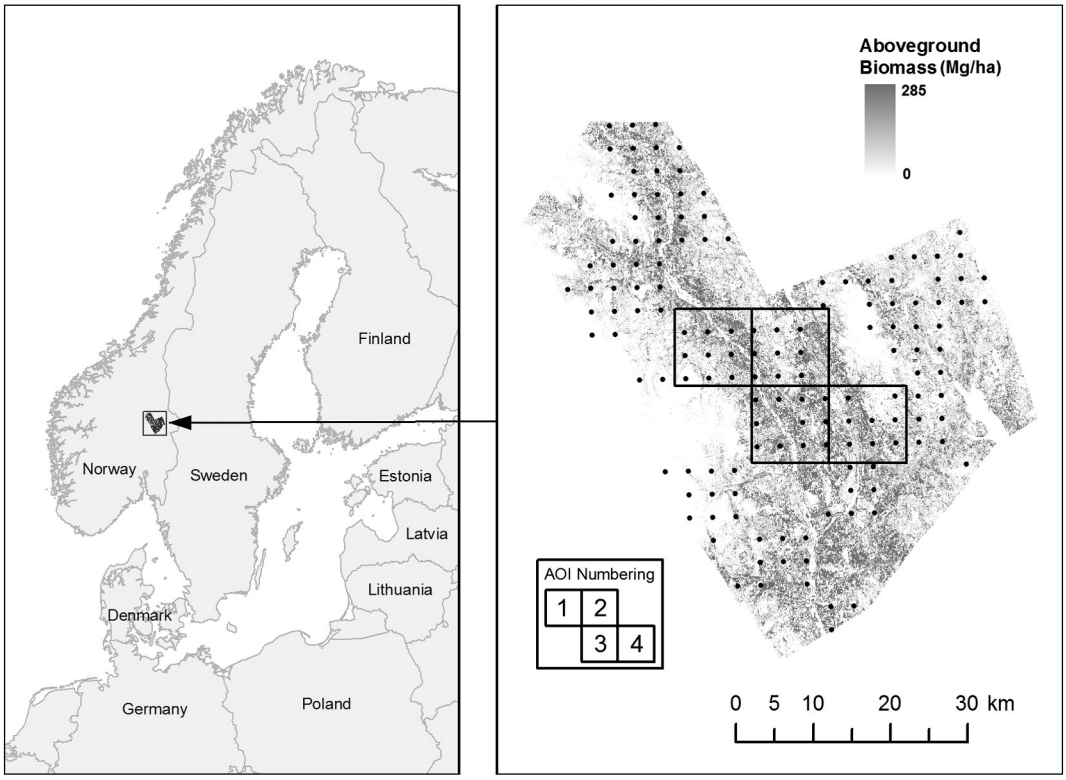

Fig. 2. Hedmark study area with four 100-km^2 areas of interest and inventory plots.

3. Methods

3.1. Framework

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

For both the Minnesota and Hedmark County study areas, the four smaller AOIs, individually and in aggregate, were used to simulate inaccessible areas for which sampling was not possible. For each of these AOIs and combinations of them, model-based inference based on a model constructed using only data external to the aggregation of the four AOIs was used to estimate means and variances which could then be used to construct confidence intervals. For comparison purposes, model-based estimates based on the model constructed using data for the entire study areas were calculated as were probability-based simple random sampling and model-assisted estimates that used the same data and model. Probability-based approaches are described first, because they are more familiar and provide a basis for comparison with the less familiar model-based approaches.

3.2. Probability-based inference

Hansen, Madow, and Tepping (1983) apparently coined the term *probability-based* as an alternative to the more familiar term *design-based*. Because the basis for inference is not just a *design* for sampling, but rather a design that features a randomization scheme and positive probabilities of selection for all population units, the term *probability-based* is considered by some to better characterize the basis for inference.

3.2.1. Assumptions

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 of each population unit into the sample is positive and known; and (3) the value of the response variable for each population unit is a constant value 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.

3.2.2. Estimators

The simplest and most familiar probability-based estimators are the *simple random sampling* (SRS) estimators. Estimators for the population mean rely only on the sample data and are expressed as,

$$\hat{\mu} = \frac{1}{n}\sum_{i\in S} y_i \tag{2}$$

and

$$\hat{Var}(\hat{\mu}) = \frac{1}{n(n-1)}\sum_{i\in S}(y_i - \hat{\mu})^2, \tag{3}$$

where the notation $i \in S$ indicates that the ith population unit is included in the sample, S, and y_i is the corresponding observation of the response variable. Because of the small sampling intensity, finite population correction factors were ignored.

With probability-based *model-assisted* approaches, a model of the relationship between the response variable, Y, and auxiliary variables, **X**, is formulated as,

$$y = f(\mathbf{X};\boldsymbol{\beta}) + \varepsilon, \tag{4}$$

where f($\mathbf{X}$; $\boldsymbol{\beta}$) is the model expression of the relationship between Y and $\mathbf{X}$, $\boldsymbol{\beta}$ is a set of parameters to be estimated, and ε is a random residual term with mean zero. An initial estimate of the population mean, μ, is calculated as,

$$\hat{\mu}_{initial} = \frac{1}{N}\sum_{i \in U} \hat{y}_i, \tag{5}$$

where $\hat{y}_i$ is a model prediction obtained using Eq. (4) with the parameter estimates. However, systematic classification or prediction errors may induce bias into this estimator which, for equal probability samples, can be estimated as,

$$\hat{\text{Bias}}(\hat{\mu}_{\text{initial}}) = \frac{1}{n}\sum_{i \in S} (\hat{y}_i - y_i). \tag{6}$$

The model-assisted regression estimator (Särndal et al., 1992, Section 6.5) is defined as the difference between the initial estimator and the estimator of its bias and is expressed as,

$$\hat{\mu} = \frac{1}{N}\sum_{i \in U} \hat{y}_i - \frac{1}{n}\sum_{i \in S} (\hat{y}_i - y_i). \tag{7}$$

Under the assumptions that N is both large and much larger than n, the variance estimator can be approximated as,

$$\hat{\text{Var}}(\hat{\mu}) = \frac{1}{n(n-1)}\sum_{i \in S} (\varepsilon_i - \bar{\varepsilon})^2, \tag{8}$$

where $\varepsilon_i = \hat{y}_i - y_i$ is the classification or prediction error, and $\bar{\varepsilon}$ is the mean of the errors. Finite population correction factors were again ignored based on the small sampling intensity. When systematic sampling rather than simple random sampling is used, variances may be overestimated (Särndal et al., 1992, p. 83). Model-assisted estimators have become popular for use with remotely sensed data, particularly ALS data (Andersen, Reutebuch, McGaughey, d'Oliveira, & Keller, 2014; d'Oliveira, Reutebuch, McGaughey, & Andersen, 2012; Gregoire et al., 2011; McRoberts, Næsset, & Gobakken, 2013a, 2013b; Næsset, Bollandsås, Gobakken, Gregoire, & Ståhl, 2013; Næsset, Gobakken, Bollandsås, Gregoire, Nelson, & Ståhl, 2013; Næsset et al., 2011; Strunk, Reutebuch, Andersen, Gould, & McGaughey, 2012).

3.3. Model-based inference

3.3.1. Assumptions

The assumptions underlying model-based inference differ considerably from the assumptions underlying probability-based inference. First, 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 constant as is the case for probability-based inference. The conceptual framework with a distribution of possible values for each unit in a finite population is characterized as a superpopulation, and model-based inference is occasionally characterized as superpopulation inference (Graubard & Korn, 2002). Second, the basis for model-based inference is correct specification of the model, not the probabilistic nature of the sample as is the case for probability-based inference. In fact, purposive, non-probability samples may produce entirely valid model-based inferences. For example, the sample may be selected to maximize the precision of the model parameter estimates or the precision of model predictions. However, a probability sample provides modest assurance that the ranges of values of independent variables in the sample data are similar to the ranges in the population to which the model is applied (Särndal, 1978). Randomization for model-based inference enters through the random realizations from the distributions for population units, whereas for probability-based inference randomization enters through the random selection of population units into the sample.

Current approaches to model-based inference originated in the context of survey sampling and can be attributed to Matérn (1986), 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 (Andersen et al., 2014; Gregoire, 1998; Kangas & Maltamo, 2006; Mandallaz, 2008; McRoberts, 2006, 2010; McRoberts et al., 2013a, 2013b; Rennolls, 1982; Ståhl et al., 2011).

3.3.2. Estimators

For model-based inference, the mean and standard deviation of the distribution of Y for the i^{th} population unit are denoted μ_i and σ_i, respectively. The mean is estimated as $\hat{\mu}_i = f\left(\mathbf{X}_i; \hat{\boldsymbol{\beta}}\right)$ where f(.) is from Eq. (4). Of importance, although the same model is used to calculate the estimate for the i^{th} population unit for both model-assisted and model-based approaches, for the latter approach the estimate is for the mean of the distribution for that unit, not the observed value; hence the notation $\hat{\mu}_i$ rather than $\hat{y}_i$ as is used for model-based inference. The standard deviation, σ_i, is estimated as the residual standard deviation obtained from deviations between observations and model predictions of the mean for all population units in the sample with the same values of the auxiliary variables.

The model-based estimator of the population mean is based on the set of estimates, $\{\hat{\mu}_i, i = 1, 2, \ldots, N\}$, of the means for individual population units and is expressed as,

$$\hat{\mu} = \frac{1}{N}\sum_{i \in U} \hat{\mu}_i, \tag{9}$$

with variance estimator,

$$\hat{\text{Var}}(\hat{\mu}) = \frac{1}{N^2}\sum_{i \in U}^{N}\sum_{j \in U}^{N} \hat{\text{Cov}}\left(\hat{\mu}_i, \hat{\mu}_j\right). \tag{10}$$

The covariance terms in Eq. (10) can be approximated using a first-order Taylor series as,

$$\hat{\text{Cov}}\left(\hat{\mu}_i, \hat{\mu}_j\right) = Z_i' \hat{V}_{\hat{\beta}} Z_j, \tag{11}$$

where $z_{ik} = \frac{\partial f(\mathbf{x}_i; \hat{\boldsymbol{\beta}})}{\partial \beta_k}$ and $\hat{V}_{\hat{\beta}}$ is the covariance matrix for the model parameter estimates.

The primary advantages of model-based inference are two-fold. First, inference is based on correct model specification, not the nature of the sample or sample size. Therefore, inference is possible for study areas with non-probability samples, small samples, and even no sample. For the latter two cases, the model is developed using data either partially or completely external to the study area, a practice characterized as *synthetic estimation* (Särndal et al., 1992, p. 399). Second, uncertainty can be estimated for any population unit, whereas with probability-based inference errors are known only for population units in the sample. However, as discussed in subsequent sections, bias must be assessed in a different manner, and variance estimation can be complex and computationally intensive.

3.3.3. Diagnostics

An important feature of model-based inference is that if the model is correctly specified, the population estimators are unbiased (Lohr, 1999), but if the model is misspecified, the adverse effects on inference may be substantial (Hansen et al., 1983; Royall & Herson, 1973; Särndal et al., 1992, p 411). With model-based inference, differences between observations and model predictions are not prediction errors as is the case for model-assisted inference, but rather differences are simply random deviations between particular realizations and the means of the distributions for population units. Thus, whereas bias is defined in terms of prediction errors for model-assisted inference, it is defined in terms of

model *lack of fit* or systematic mischaracterization of the means for individual population units for model-based inference. An important consequence is that lack of fit must be more carefully assessed for model-based inference, because there is no bias correction term as is the case for model-assisted inference.

An obvious visual diagnostic for assessing model lack of fit is a graph of observations versus predictions. Lack of fit is indicated if the points on the graph fail to lie along the 1:1 line with intercept of 0 and slope of 1. If large numbers of observations are available for each combination of values of the independent variables, lack of fit can be readily assessed in a statistically rigorous manner by comparing uncertainty due to lack of fit and uncertainty due to residual variation (Draper & Smith, 1981, Section 1.5). For large sample sizes, these two uncertainties can be distinguished, but for small samples with only a single observation for each set of values of independent variables, this distinction cannot be made.

For linear models, the covariance estimator of Eq. (11) is exact, but for nonlinear models it is only a first-order Taylor series approximation as is the model parameter covariance matrix, $\hat{V}_{\hat{\beta}}$. Depending on the model and data, one or both of these approximations may be poor (Bates & Watts, 1988, Fig. 6.18).

4. Analyses

4.1. Models

For the Minnesota study area, the relationship between the binary forest/non-forest observations ($y = 0$ denotes non-forest; $y = 1$ denotes forest) and the Landsat transformations was represented using a *binomial logistic regression model* of the form,

$$p_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}\right)} + \varepsilon_i, \tag{12}$$

where i indexes population units, p_i denotes the probability that $y_i = 1$, x_{ij} is the value of the j^{th} Landsat transformation, the βs are parameters to be estimated, and ε_i is a residual term. Maximum likelihood methods were used to estimate the model parameters (Agresti, 2007; McRoberts & Walters, 2012).

For the Hedmark study area, a nonlinear logistic regression model was used to describe the relationship between AGB and the associated ALS metrics. The model has the mathematical form,

$$y_i = f(\mathbf{X}_i;\boldsymbol{\beta}) = \frac{\alpha}{1 + \exp\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}\right)} + \varepsilon_i, , \tag{13}$$

where i indexes population units, x_{ij} is the j^{th} ALS metric, α and the βs are parameters to be estimated, and ε_i is a residual term. For future reference, the model expressed by Eq. (13) is designated the *asymptotic logistic regression model* to distinguish it from the binomial logistic regression model described by Eq. (12). Model parameters were estimated using nonlinear least squares techniques. Variables selected for the model are the same as reported by McRoberts et al. (2013a).

For both study areas, the models were fit twice, once using data for the entire study area and once using only data external to the aggregation of the four AOIs.

4.2. Model assessment

The models and their parameter estimates were assessed using four techniques. First, the quality of fit of the models to the data was assessed by graphing observations versus predictions. In addition, the prediction-observations pairs were ordered by the value of the prediction and aggregated into groups of size 10. The mean of the observations and the mean of the predictions were calculated for each group, and the observation means were graphed against the prediction means. If the models are correctly specified, both graphs should feature points that lie along a line with intercept 0 and slope 1. Second, mean deviations between plot observations and model predictions were calculated over all plots in the study areas and over all plots in the aggregations of the AOIs.

Third, the effects of curvature in the model prediction surface on the quality of the Taylor series variance approximations were assessed using the measures proposed by Bates and Watts (1988). A comprehensive discussion of nonlinear model curvature is beyond the scope of this study, but is available in both Bates and Watts (1988) and Ratkowsky (1983). The salient issues are that Taylor series variance approximations assume that the surface defined as the model predictions versus model parameter values is a multi-dimensional plane in the vicinity of the parameter estimates and that the predictions change linearly with changes in the parameter values. Deviations from these linearity assumptions cause variance estimates and confidence regions based on Taylor series approximations to deviate from the true estimates and regions. Bates and Watts (1988) proposed two measures of surface curvature to assess the statistical significance of deviations from linearity; if neither measure is statistically significant, then the Taylor series approximations may be considered adequate. Although both Ratkowsky (1983, Section 4.3) and Haines, Brien, and Clark (2004) suggest that the effects of curvature are not severe for logistic models, curvature measures were calculated nevertheless because the models used for this study have more parameters than the models evaluated previously. Specifically, curvature measures were calculated for the asymptotic logistic regression model which was fit using nonlinear least squares techniques, but not for the binomial logistic model which was fit with more general maximum likelihood techniques and for which curvature assessment techniques are not readily available.

Fourth, as a further diagnostic, standard errors obtained using the Taylor series variance approximations were compared to standard errors obtained using a Monte Carlo bootstrap analysis (Efron & Tibshirani, 1994; McRoberts & Westfall, 2014). With this approach, the data are resampled with replacement until the original sample size is reached; the model parameters are estimated by fitting the model to the resampled data; the model is applied to estimate the population mean; and the procedure is replicated until the mean of the estimated bootstrap means and the variance over replications stabilize. Of importance, the bootstrap resampling must mimic the original sampling scheme. For the aggregations of the four AOIs for each study area, standard error estimates obtained using Taylor series approximations expressed by Eqs. (10) and (11) were compared to the bootstrap variance estimates.

4.3. Variance estimation

A disadvantage of model-based estimators is that the parametric variance estimators of Eqs. (10) and (11) require calculation of derivatives and considerable computational intensity as a result of the double summation, particularly for large study areas. For example, in aggregate, the four Minnesota AOIs consist of more than 1.5×10^6 population units, which means that the number of covariance calculations necessary for Eq. (10) is on the order of 10^{12}. However, Eq. (10) is just a two-dimensional mean over all units in the population, and $\hat{\text{Var}}(\hat{\mu})$ can be approximated by sampling from the population. McRoberts (2010), McRoberts et al. (2013b) estimated $\text{Var}(\hat{\mu})$ using only the population units located at the intersections of an equally-spaced, two-dimensional, perpendicular grid superimposed on the study area, and reported that the detrimental effects were negligible for grid widths as great as 10 population units, although the maximum acceptable grid width will depend on the population size. For this study, instead of using a grid

superimposed on the AOIs, Var($\hat{\mu}$) was estimated using only every m^{th} population unit from the original ordering of the population units in the datasets where m = 1, 2, 10, 25, 50, and 100. This approach decreases computational intensity by a factor of m^2.

4.4. Comparisons

The models constructed using data for the entire study areas were used to calculate predictions for all population units with centers in the four smaller AOIs. The model-assisted means, biases, and variances were then estimated as per Eqs. (6), (7), and (8) using only data for plots whose centers were in the AOIs. The same model served as the basis for estimating the model-based means and variances using Eqs. (9) and (10). The resulting model-assisted and model-based estimates are the estimates that would have been obtained if sample data for the AOIs were available. In addition, the models constructed using data for only the portions of the study area external to the aggregation of the four AOIs were also used to calculate predictions for all population units with centers in the AOIs, and the corresponding model-based means and variances were estimated using Eqs. (9) and (10). The latter estimates are the estimates that would have been obtained if sample data for the entire study area were not available. Primary interest is in comparisons of the model-based estimates obtained using only data external to the AOIs and both the model-based and the model-assisted estimates that used data for the entire study areas.

5. Results and discussion

5.1. Model assessment

Graphs of response variable observations versus corresponding model predictions indicated no systematic lack of fit (Figs. 3, 4). For the entire Minnesota study area, the mean deviation between plot proportion forest observations and model predictions was −0.0003, representing less than 0.1% of the plot-based mean. For the entire Hedmark study area, the mean deviation was −2.7681 Mg/ha, representing less than 4% of the plot-based mean. The combination of the graphs and the relatively small mean deviations suggest little meaningful model lack of fit.

5.2. Estimates of the population mean

Comparisons of the model-based estimates, $\hat{\mu}_{MB-Ext}$, based on the model constructed using only data external to the AOIs to estimates based on data for entire study area were of primary interest.

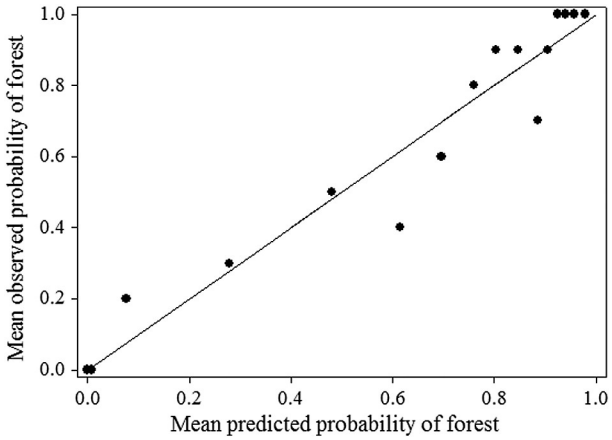

Fig. 3. Model accuracy for Minnesota study area.

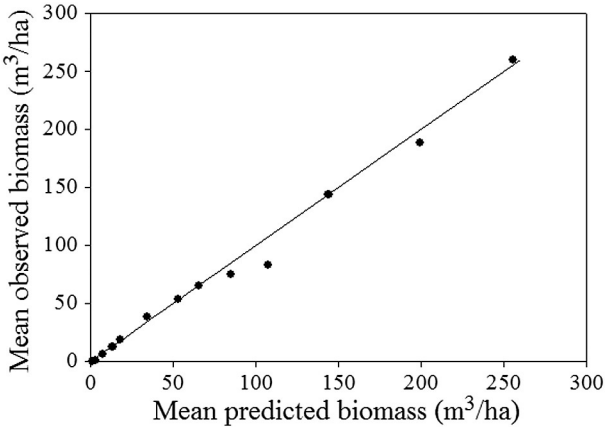

Fig. 4. Model accuracy for Hedmark study area.

For both study areas, the $\hat{\mu}_{MB-Ext}$ estimates tended to be smaller than the $\hat{\mu}_{MB-All}$ estimates obtained using the model constructed using data for the entire study area (Tables 1, 2). This result was attributed to slight differences in the model parameter estimates which, in turn, were attributed to slightly greater values of the response variable for the plots in the aggregation of the AOIs relative to the plots throughout the rest of the study area. In particular, proportion forest area for plots in the aggregation of the Minnesota AOIs was larger than the mean for all the plots by slightly more than 0.02, and mean AGB for plots in the aggregation of the Hedmark AOIs was larger than the mean for all plots by 32.70 Mg/ha. However, despite the systematic trends and the rather large differences in the AGB plot means, the differences between the $\hat{\mu}_{MB-Ext}$ estimates and the other estimates were seldom statistically significant. For the Minnesota study area, the $\hat{\mu}_{MB-Ext}$ estimates for all AOIs and their aggregations were within 0.7 model-based SEs of the $\hat{\mu}_{MB-All}$ estimates and within 1.2 SRS SEs of the $\hat{\mu}_{SRS}$ estimates. With only one exception the $\hat{\mu}_{MB-Ext}$ estimates were within 1.4 model-assisted SEs of the $\hat{\mu}_{MA}$ estimates (Table 1). For the Hedmark study area, the $\hat{\mu}_{MB-Ext}$ estimates for all AOIs and their aggregations were within 1.3 model-based SEs of the $\hat{\mu}_{MB-All}$ estimates and within 1.2 SRS SEs of the $\hat{\mu}_{SRS}$ estimates. With only one minor exception, the $\hat{\mu}_{MB-Ext}$ estimates were within 2.0 model-assisted SEs of the $\hat{\mu}_{MA}$ model-assisted estimates (Table 2). Finally, these comparisons do not consider the uncertainty in the $\hat{\mu}_{MB-Ext}$ estimates which would reduce the statistical significance of differences even more. The important result is that the model-based AOI estimates of the means based on the model constructed using only data external to the AOIs were comparable to estimates based on the model constructed using data for the entire study area.

5.3. Estimates of the standard errors

For the Hedmark study area, the assessment of curvature in the model prediction space indicated no statistical significance. This result was confirmed by the Monte Carlo bootstrap simulations for which the model-based SE for 20 resamples was 4.98 Mg/ha which compared very favorably to the estimate of 5.06 Mg/ha calculated using Eq. (10). For the Minnesota study area, the Monte Carlo bootstrap SE for 50 resamples was 0.026, the same as the SE calculated using Eq. (10). These results indicate that the Taylor series variance approximations can be used with confidence.

The Taylor series variance estimates calculated using only every m^{th} population unit differed very little from the estimates calculated using all population units (m = 1) (Table 3). For the Minnesota AOIs, proportional differences were less than 0.02 for $m \leq 100$, and for the Hedmark AOIs, proportional differences were less than 0.01 for $m \leq 50$. These results indicate that computational intensity may be substantially reduced with no loss in the quality of the SE estimates.

Table 1
Estimates of mean probability of forest.

Area of interest (AOI)	Area size (km^2)	Sample size	All plots						External plots	
			SRS		Model-assisted		Model-based		Model-based	
			$\hat{\mu}_{SRS}$	$SE(\hat{\mu}_{SRS})$	$\hat{\mu}_{MA}$	$SE(\hat{\mu}_{MA})$	$\hat{\mu}_{MB-All}$	$SE(\hat{\mu}_{MB-All})$	$\hat{\mu}_{MB-Ext}$	$SE(\hat{\mu}_{MB-Ext})$
1	700	7	0.571	0.202	0.534	0.155	0.583	0.029	0.575	0.035
2	700	9	0.556	0.176	0.528	0.124	0.504	0.028	0.499	0.034
3	700	11	0.818	0.122	0.815	0.056	0.696	0.027	0.679	0.034
4	700	10	0.700	0.153	0.707	0.061	0.712	0.027	0.698	0.033
1-2	1400	16	0.563	0.128	0.535	0.094	0.543	0.026	0.537	0.035
2-3	1400	20	0.700	0.105	0.676	0.063	0.600	0.026	0.589	0.032
3-4	1400	21	0.762	0.095	0.764	0.043	0.704	0.027	0.704	0.027
1-2-3	2100	27	0.667	0.093	0.638	0.061	0.594	0.027	0.584	0.033
2-3-4	2100	30	0.700	0.085	0.686	0.046	0.637	0.026	0.625	0.032
1-2-3-4	2800	37	0.676	0.078	0.655	0.047	0.624	0.026	0.613	0.032

Table 2
Estimates of mean aboveground biomass per unit area (Mg/ha).

Area of interest (AOI)	Area size (km^2)	Sample size	All plots						External plots	
			SRS		Model-assisted		Model-based		Model-based	
			$\hat{\mu}_{SRS}$	$SE(\hat{\mu}_{SRS})$	$\hat{\mu}_{MA}$	$SE(\hat{\mu}_{MA})$	$\hat{\mu}_{MB-All}$	$SE(\hat{\mu}_{MB-All})$	$\hat{\mu}_{MB-Ext}$	$SE(\hat{\mu}_{MB-Ext})$
1	100	9	109.05	48.49	125.35	14.21	99.11	4.38	95.82	5.34
2	100	9	66.20	26.33	98.96	3.94	104.94	4.89	101.07	5.89
3	100	11	114.38	35.07	123.93	9.94	128.08	6.21	121.06	7.44
4	100	8	140.26	40.77	126.10	13.90	102.78	4.91	98.34	5.96
1-2	200	18	87.62	27.26	112.10	8.15	101.97	4.62	98.44	5.61
2-3	200	20	92.70	22.75	111.54	5.62	116.51	5.53	111.06	6.64
3-4	200	19	125.28	26.01	122.93	8.58	115.52	5.53	109.69	6.67
1-2-3	300	29	97.77	21.28	115.42	6.33	110.62	5.13	105.98	6.21
2-3-4	300	28	106.29	20.01	115.04	6.02	111.93	5.31	106.82	6.40
1-2-3-4	400	37	109.96	18.82	117.46	5.85	108.72	5.06	104.07	6.12

For both study areas, the model-assisted SEs were considerably smaller than SEs for the SRS estimates, indicating the utility of the ALS data for increasing precision (Table 2); similar results have already been widely reported (e.g., d'Oliveira et al., 2012; McRoberts et al., 2013b; Næsset et al., 2011). As expected, aggregations of AOIs and the consequential increase in sample sizes, produced reductions in both the SRS and model-assisted SEs. However, such was generally not the case for the model-based SEs which were generally similar, regardless of the sample sizes associated with the AOIs and their aggregations. This phenomenon is explained by the fact that the model-based SEs are primarily influenced by the parameter covariance estimates which, in turn, are influenced by the mathematical form of the model, the residual variation, the size of the sample used to construct the model, the distribution of values of the independent variables in the sample, and the parameter estimates. These features are unchanging for any region within the larger study area because of the synthetic approach to estimation. The only feature that changes is the distribution of values of the independent variables in the dataset to which the models are applied. However, because of general similarity of the resource throughout the study area, the distributions of values of the independent variables for the entire study area may be expected to be generally similar to the distribution for the AOIs. The larger model-based SEs for the estimates based on the model constructed using only data for the external plots relative to SEs for estimates based on the model constructed using data for the entire study area is attributed to the smaller sample size available for the former model.

Table 3
Effects of sampling on model-based standard errors.

AOI	m*						
	1	2	5	10	25	50	100
Minnesota study area							
1	0.0294	0.0294	0.0293	0.0293	0.0292	0.0292	0.0289
2	0.0281	0.0281	0.0281	0.0281	0.0282	0.0281	0.0280
3	0.0273	0.0273	0.0273	0.0273	0.0274	0.0278	0.0274
4	0.0268	0.0268	0.0268	0.0267	0.0267	0.0269	0.0271
Hedmark study area							
1	4.375	4.372	4.384	4.404	4.378	4.399	4.435
2	4.884	4.886	4.882	4.867	4.928	4.890	4.840
3	6.214	6.208	6.221	6.214	6.244	6.217	6.080
4	4.910	4.911	4.909	4.923	4.916	4.935	5.155

* Estimates are calculated using samples consisting of every m^{th} population unit.

6. Conclusions

The primary conclusion drawn from the study was that model-based inference, together with synthetic estimation, is a relevant and useful approach for estimation and inference for non-sampled areas. Applications, as noted, include remote tropical and boreal forests for which access is difficult and sampling is either limited or even impossible. Several pre-cautions merit consideration: first, data for a region similar to the non-sampled area must be available; second, model lack of fit must be carefully evaluated; and third, if nonlinear models are used, the validity of Taylor series variance approximation should be evaluated. Subject to these constraints, model-based inference merits greater consideration for a variety of applications.

Acknowledgement

The authors thank Mr. Brian F. Wilson of the Northern Research Station, U.S. Forest Service, for assistance with graphics.

References

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

Andersen, H. -E., Reutebuch, S. E., McGaughey, R., d'Oliveira, M., & Keller, K. (2014). Monitoring selective logging in western Amazonia with repeat LIDAR flights. *Remote Sensing of Environment*, *151*, 157–165.

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications.* New York: Wiley.

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.

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.

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

d'Oliveira, M. V. N., Reutebuch, S. E., McGaughey, R. J., & Andersen, H. E. (2012). Estimation of forest biomass and identifying low-intensity logging areas using airborne scanning lidar in Antimary State Forest, Acre State, Western Brazilian Amazon. *Remote Sensing of Environment*, *124*, 479–491.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap.* Boca Raton, FL: Chapman and Hall/CRC.

Gobakken, T., & Næsset, E. (2008). Assessing effects of laser point density, ground sampling intensity, and field plot sample size on biophysical stand properties derived from airborne laser scanner data. *Canadian Journal of Forest Research*, *38*, 1095–1109.

Gobakken, T., Næsset, E., Nelson, R., Bollandsås, O. M., Gregoire, T. G., Ståhl, G., et al. (2012). Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning. *Remote Sensing of Environment*, *123*, 443–456.

Graubard, B. I., & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, *17*(1), 73–196.

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

Gregoire, T. G., Ståhl, G., Næsset, E., Gobakken, T., Nelson, R., & Holm, S. (2011). Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. *Canadian Journal of Forest Research*, *41*, 83–95.

Haines, L. M., Brien, T. E., & Clark, G. P. Y. (2004). Kurtosis and curvature measures for nonlinear regression models. *Statistica Sinica*, *14*, 547–570.

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

Kangas, A., & Maltamo, M. (2006). *Forest inventory, methodology and applications.* 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, Indiana: Purdue University.

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

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

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

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

McRoberts, R. E. (2010). Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. *Remote Sensing of Environment*, *114*, 017–1025.

McRoberts, R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sensing of Environment*, *115*, 715–724.

McRoberts, R. E., Bechtold, W. A., 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., Hansen, M. H., & Smith, W. B. (2010). Country report: United States of America. In E. Tomppo, T. Gschwantner, M. Lawrence, & R. E. McRoberts (Eds.), *National forest inventories: pathways for common reporting* (pp. 567–582). Heidelberg: Springer.

McRoberts, R. E., Næsset, E., & Gobakken, T. (2013a). Accuracy and precision for remote sensing applications of nonlinear model-based inference. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(1), 27–34.

McRoberts, R. E., Næsset, E., & Gobakken, T. (2013b). Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sensing of Environment*, *128*, 268–275.

McRoberts, R., & Walters, B. F. (2012). Statistical inference for remote sensing-based estimates of net deforestation. *Remote Sensing of Environment*, *124*, 394–401.

McRoberts, R. E., & Westfall, J. A. (2014). Effects of uncertainty in model predictions of individual tree volume on large area volume estimates. *Forest Science*, *60*(1), 34–42.

Næsset, E., Bollandsås, O. M., Gobakken, T., Gregoire, T. G., & Ståhl, G. (2013). Model-assisted estimation of change in forest biomass over an 11 year period in a sample survey supported by airborne LiDAR: A case study with post-stratification to provide "activity data". *Remote Sensing of Environment*, *128*, 299–314.

Næsset, E., Gobakken, T., Bollandsås, O. M., Gregoire, T. G., Nelson, R., & Ståhl, G. (2013). Comparison of precision of biomass estimates in regional field sample surveys and airborne LiDAR-assisted surveys in Hedmark County, Norway. *Remote Sensing of Environment*, *130*, 108–120.

Næsset, E., Gobakken, T., Solberg, S., Gregoire, T. G., Nelson, R., Ståhl, G., et al. (2011). Model-assisted regional forest biomass estimation using LiDAR and InSAR as auxiliary data: A case study from a boreal forest area. *Remote Sensing of Environment*, *115*, 3599–3614.

Ratkowsky, D. A. (1983). *Nonlinear regression modelling: a unified practical approach.* New York: Dekker.

Rennolls, K. (1982). 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* (pp. 395–401). Orono, ME, Bethesda, MD: Society of American Foresters (9–14 Aug. 1981).

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, Vol. 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. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, *5*, 27–52.

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

Simpson, J. A., & Weiner, E. S. C. (Preparers) (1989). *The Oxford English Dictionary*, *2nd edition.* Oxford: Clarendon Press. Volume 7, pp 923-924.

Ståhl, G., Holm, S., Gregoire, T. G., Gobakken, T., Næsset, E., & Nelson, R. (2011). Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway. *Canadian Journal of Forest Research*, *41*, 96–107.

Strunk, J., Reutebuch, S., Andersen, H. -E., Gould, P., & McGaughey, R. (2012). Model-assisted forest yield estimation with light detection and ranging. *Western Journal of Applied Forestry*, *27*, 53–59.