Empirical coverage of model-based variance estimators for remote sensing assisted estimation of stand-level timber volume

Johannes Breidenbach [a,*], Ronald E. McRoberts [b], Rasmus Astrup [a]

[a] Norwegian Institute of Bioeconomy Research (NIBIO), Postboks 115, 1431 Ås, Norway
[b] USDA Forest Service, Forest Inventory and Analysis, St. Paul, MN, United States

ABSTRACT

Due to the availability of good and reasonably priced auxiliary data, the use of model-based regression-synthetic estimators for small area estimation is popular in operational settings. Examples are forest management inventories, where a linking model is used in combination with airborne laser scanning data to estimate stand-level forest parameters where no or too few observations are collected within the stand. This paper focuses on different approaches to estimating the variances of those estimates. We compared a variance estimator which is based on the estimation of superpopulation parameters with variance estimators which are based on predictions of finite population values. One of the latter variance estimators considered the spatial autocorrelation of the residuals whereas the other one did not. The estimators were applied using timber volume on stand level as the variable of interest and photogrammetric image matching data as auxiliary information. Norwegian National Forest Inventory (NFI) data were used for model calibration and independent data clustered within stands were used for validation. The empirical coverage proportion (ECP) of confidence intervals (CIs) of the variance estimators which are based on predictions of finite population values was considerably higher than the ECP of the CI of the variance estimator which is based on the estimation of superpopulation parameters. The ECP further increased when considering the spatial autocorrelation of the residuals. The study also explores the link between confidence intervals that are based on variance estimates as well as the well-known confidence and prediction intervals of regression models.

1. Introduction

The use of airborne laser scanning (ALS) data in operational forest management inventories (FMI) has a long tradition in the Nordic countries (Maltamo & Packalen, 2014; Næsset, 2014). Usually, the area-based approach (ABA) is adopted, where the study area is gridded into small cells for which height and density metrics are calculated from ALS data (Næsset, 1997, 2014). A model linking the variable of interest, such as timber volume, to the ALS metrics is estimated using field sample plots where the variable of interest and the ALS data are both available. The linking model is then applied to the grid cells to map the timber volume. A main product of an FMI is a map of mean stand-level timber volume where the mapped value for each stand is calculated as the mean of timber volume predictions for grid cells whose centers are in the stand.

Although the ABA was developed using ALS data, it is well suited for the use with other remote sensing data providing high-resolution height information. For example, photogrammetric image matching data are increasingly popular to estimate forest parameters using the ABA due to the increasing availability of high-quality digital terrain models as well as improved hard- and software (e.g., Bohlin, Wallerman, & Fransson, 2012; Breidenbach & Astrup, 2012; Vastaranta et al., 2013).

In the terms of survey sampling, the ABA is one form of small area estimation (SAE), since the stands are so small or so remote that few if any sample plots are located within them. From the perspective of SAE, ALS metrics are auxiliary data and aggregating the predictions for grid-cells at stand-level is a synthetic estimate for a small area (Rao, 2003, p. 46). This estimate is termed synthetic because only model predictions are used, with no correction for model prediction errors. While design-based estimators are generally preferred in forest inventories if enough field observations are available because they are asymptotically unbiased, synthetic estimators are generally model-based (Chambers & Clark, 2012, p. 169). The basic difference between model-based and design-based inference is the source of randomness (Kangas, 2006). Whereas randomness is introduced by sample selection in design-based inference and observations are assumed to be fixed values, the observations are assumed to be a random realization of a joint distribution known as the superpopulation in model-based inference. One consequence of the differences in these underlying assumptions is that probability samples are not necessary for model-based estimators.

* Corresponding author.
  E-mail address: Johannes.Breidenbach@nibio.no (J. Breidenbach).

For an introduction to model-based inference and a comparison to design-based inference see Gregoire (1998).

Although studies aiming at small area estimation using remotely sensed data are plentiful, the uncertainty of estimates is often ignored for management applications. While the number of SAE studies in forestry including inference is increasing (e.g., Breidenbach & Astrup, 2012; Goerndt, Monleon, & Temesgen, 2013; Lappi, 2001; Magnussen, Mandallaz, Breidenbach, Lanz, & Ginzler, 2014; Steinmann, Mandallaz, Ginzler, & Lanz, 2013), the number of studies that provide methods for stand-level inference using synthetic estimators is small (e.g., Kangas, 1996; Mandallaz, 2013; McRoberts, 2006).

In this study, we focus on synthetic estimation, which is relevant for small areas that frequently contain no or too few observations to apply other estimators. The context of the study is model-based inference which assumes that an entire distribution of observations is possible for each population unit. In this context, prediction of an individual observation (a finite population value) is distinguished from estimation of the expected value of the distribution of observations (a superpopulation parameter) (Kangas, 2006, p. 40). Although the prediction of an observation and the estimate of its expected value are the same for models relevant in our context, the variance estimates may be quite different. This paper focuses on different approaches to estimating the variances.

A variance estimator based on the estimation of superpopulation parameters just considers the variance resulting from the estimation of the model parameters and is therefore independent of stand size (e.g., Kangas, 1996; Mandallaz, 1991; McRoberts, Andersen, & Næsset, 2014).

Kangas (2006) and McRoberts (2006) described a variance estimator which is based on predictions of finite population values rather than estimates of superpopulation parameters. The estimated variance is therefore dependent on stand size. Kangas (2006) described the basic form of the variance estimator in a general setting, not specifically for SAE. McRoberts (2006) extended the variance estimator for spatial autocorrelation and applied it to the binary variable forest/non-forest. We modify the variance estimator described by McRoberts (2006) for application to a continuous response variable for which we accommodate heteroskedasticity and spatial autocorrelation.

The aim of this study is to compare a variance estimator based on the prediction of superpopulation parameters with variance estimators based on predictions of finite population values in the context of synthetic estimation. Furthermore, we link the variance estimators to the concepts of prediction intervals and confidence intervals well-known from regression analysis.

In a case study, we use Norwegian National Forest Inventory (NFI) data to estimate stand-level mean timber volume. Photogrammetric image matching data processed using the ABA serve as auxiliary information. To compare estimators, the empirical coverage proportions (ECP) of the confidence intervals based on the different variance estimators are obtained using independent validation data.

2. Methods

2.1. Estimators

A linking model describes the statistical relation between the response (variable of interest) denoted y and the auxiliary variables x which, in this case, are obtained from remotely sensed data

$$y_i = f(\boldsymbol{X}_i, \boldsymbol{\beta}) + \varepsilon_i, \quad i = \{1, ..., n\}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2 \boldsymbol{W}) \tag{1}$$

where i indexes observations, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \cdots, \boldsymbol{X}_n^T)^T = (1 \ \boldsymbol{x}_1 \cdots \boldsymbol{x}_p)$ is a $n \times (p + 1)$ design matrix, p is the number of auxiliary (explanatory) variables, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ is a vector of model parameters to be estimated, and ε_i is a residual.

The residual variance is expressed as the product of σ_ε^2 and a $n \times n$ matrix $\boldsymbol{W}$ where σ_ε^2 is the mean square residual. In the case of homogeneous variances, $\boldsymbol{W}$ is an identity matrix ($w_{ii} = 1$). In the case of heteroskedsticity, the diagonal elements contain appropriate weights w_{ii} that result from a variance model. In the case of autocorrelation, also the off-diagonal elements contain appropriate weights w_{ij} that result from a model describing the correlation pattern among the residuals.

To simplify the following estimators, we assume a linear model f. While we assume that the auxiliary variable is available wall-to-wall in the areas of interest, the response is only observed at a sample of the population. In forest inventories, the response is typically observed at n sample plots systematically distributed over the landscape with distances between plots in the range 100–1000 m.

In general, synthetic estimators describe a group of estimators for small areas that are based on a population level model, assuming that the characteristics of the large area hold for the small areas (Gonzalez, 1973; NCHS, 1968). This means that differences in estimates for different areas are explained by differences in the auxiliary variables rather than differences in relationships between the response and auxiliary variables (Särndal, Swensson, & Wretman, 1992, p. 411). Synthetic estimators are potentially biased, but the bias can be small if the linking model holds in the small area.

Suppose, enough observations were available within a small area to support fitting a local linking model just for the small area. If the estimated model parameters of the local linking model are very similar to those for the linking model fitted to the large area, the bias of the synthetic estimator would be small. However, usually large numbers of observations within stands are not available in operational forest inventories. The bias of the synthetic estimator will therefore usually remain unknown.

The regression-synthetic estimator (Rao, 2003, p. 46), as one specific synthetic estimator, is the mean of predictions of a linking model for units within a small area. If the linking model is a linear regression model as assumed in this study, the mean of the model predictions equals the product of the means of the auxiliary variables ($\overline{\boldsymbol{X}}$) and the regression coefficient estimates

$$\widehat{\overline{Y}}_m = \overline{\boldsymbol{X}}_m^T \hat{\boldsymbol{\beta}} = \frac{1}{N_m} \sum_{i=1}^{N_m} f(\boldsymbol{X}_i, \hat{\boldsymbol{\beta}}) = \frac{1}{N_m} \sum_{i=1}^{N_m} \hat{\mu}_i \tag{2}$$

where $i = \{1, ..., N_m\}$, N_m is the number of population elements within small area m, and $m = \{1, ..., M\}$ where M is the total number of small areas. The upper case letter $\widehat{\overline{Y}}_m$ is used for the small area-level estimate which is the estimated mean of predictions for population elements $\hat{\mu}_i$. Very small areas can also consist of only one population element in which case $\widehat{\overline{Y}}_m = \hat{\mu}_i$, $i = 1$. The notation $\hat{\mu}$ indicates that the model prediction is an estimate of the superpopulation parameter (the expected value of the linking model given the explanatory variables), not a prediction of the observation, $\hat{y}$. The first representation of estimator (2) is applicable for linear models (Kangas, 1996; Mandallaz, 1991), the second and third representation is more generally valid (e.g., also for nonlinear models) (McRoberts, 2006; McRoberts, Næsset, & Gobakken, 2013).

In the ABA, the population elements are often designated grid-cells and the small areas are typically stands. Typically, some of the grid cells will overlap with the sample plots used to fit the linking model (1). While Næsset (1997) was among the first to apply the regression-synthetic estimator in the ABA, Kangas (1996), McRoberts (2006), Mandallaz (2013), and McRoberts et al. (2013) have, among others, described the variance of the estimator in a forest inventory setting. Because estimator (2) is the mean of the predictions, the variance is the two-dimensional mean of the covariances of the predictions

$$\widehat{Var}_p(\widehat{\overline{Y}}_m) = \overline{\boldsymbol{X}}_m^T \hat{\Sigma} \overline{\boldsymbol{X}}_m = \frac{1}{N_m^2} \sum_{i=1}^{N_m} \sum_{j=1}^{N_m} \widehat{Cov}(\hat{\mu}_i, \hat{\mu}_j) \tag{3}$$

where the subscript p identifies the estimator as considering parameter uncertainty, $i = j = \{1, ..., N_m\}$ index grid cells and $\hat{\Sigma}$ is the estimated covariance matrix of the parameter estimates of the linking model (e.g., Fahrmeir, Kneib, & Lang, 2007). This variance estimator incorporates the uncertainty in the estimate of the mean due to the uncertainty in the estimates of the regression parameters. The general representation of the estimator is given by McRoberts (2006), and a specific representation for linear models is given by Kangas (1996) and Mandallaz (2013).

In model-based inference, one can either estimate the superpopulation parameter (expected value, μ) or predict the finite population value (observation, y) for each grid-cell. For the estimate of a mean within a small area, both approaches are equivalent and estimator (2) can be written

$$\hat{\overline{Y}}_m = \frac{1}{N_m} \sum_{i=1}^{N_m} \hat{\mu}_i = \frac{1}{N_m} \sum_{i=1}^{N_m} \hat{y}_i. \qquad (4)$$

However, for the variance estimate, there is a difference as we will see below. Because the variance estimator (3) depends only on the explanatory variables and the parameter covariance matrix, this estimator is independent of stand size (McRoberts et al., 2014). Although this is correct when estimating the superpopulation parameter, this property is somewhat counter-intuitive from a design-based perspective for which one would assume that the uncertainty of an estimate decreases as the number of grid cells within a stand increases. Furthermore, forest stands often consist of only few grid cells and ignoring the residual error variance may not be adequate. An estimate of the residual variance can be added to the variance of the synthetic estimator to compensate for that error term (Prasad & Rao, 1990)

$$\widehat{Var}_{pr}\left(\hat{\overline{Y}}_m\right) = \widehat{Var}_p\left(\hat{\overline{Y}}_m\right) + \frac{\hat{\sigma}_\varepsilon^2}{N_m} \qquad (5)$$

where the subscript pr identifies the estimator as considering parameter and residual uncertainty.

Heteroskedasticity can be incorporated by

$$\widehat{Var}_{prh}\left(\hat{\overline{Y}}_m\right) = \widehat{Var}_p\left(\hat{\overline{Y}}_m\right) + \frac{1}{N_m^2} \sum_{i=1}^{N_m} \hat{\sigma}_i^2 \qquad (6)$$

where $\hat{\sigma}_i^2 = \hat{\sigma}_\varepsilon^2 w_i$ is estimated from the model residual variance using appropriate weights. The subscript prh identifies the estimator as considering parameter uncertainty and heteroskedastic residual uncertainty.

In general, heteroskedsticity and autocorrelation within a small area can be accommodated by

$$\begin{aligned}\widehat{Var}_{prhs}\left(\hat{\overline{Y}}_m\right) &= \frac{1}{N_m^2} \sum_{i=1}^{N_m} \sum_{j=1}^{N_m} \left(\widehat{Cov}\left(\hat{\mu}_i, \hat{\mu}_j\right) + \widehat{Cov}\left(\varepsilon_i, \varepsilon_j\right)\right) \\ &= \widehat{Var}_p\left(\hat{\overline{Y}}_m\right) + \frac{1}{N_m^2} \sum_{i=1}^{N_m} \sum_{j=1}^{N_m} \hat{\sigma}_i \hat{\sigma}_j \hat{\varrho}_{ij}\end{aligned} \qquad (7)$$

where $\hat{\varrho}_{ij}$ is the estimated correlation between two grid cell residuals i and j due to (spatial) autocorrelation. The subscript $prhs$ identifies the estimator as considering parameter uncertainty and spatially correlated heteroskedastic residual uncertainty. McRoberts (2006) described the variance estimator in the context of estimating forest area and thus for the binary variable forest/non-forest. Estimator (7) is the most general form of the synthetic variance estimators described here. Estimator (7) reduces to estimator (6) by ignoring spatial autocorrelation ($\hat{\varrho}_{ij} = 0$ $\forall i \neq j$) and to estimator (5) by further assuming a homoscedastic variance structure ($\hat{\sigma}_i = \hat{\sigma}_j = \hat{\sigma}_\varepsilon$). We obtain estimator (3) by ignoring

the effect of residual error ($\hat{\sigma}_i = \hat{\sigma}_j = \hat{\sigma}_\varepsilon = 0$). Estimators for totals can be obtained by omitting the terms $\frac{1}{N_m}$ or $\frac{1}{N_m^2}$ in the estimators (2)–(7).

2.2. Confidence intervals based on variance estimates vs. confidence intervals and prediction intervals known from regression analysis

A confidence interval (CI) is an interval around an estimate of a population parameter which will or will not include the true population parameter. If many CIs are estimated based on independent samples from the same population, a proportion of them will include the true population parameter. The proportion of CIs that include the true population parameter if all assumptions are met is determined by the confidence level given by $100(1 - \alpha)$ % which, in forest inventories, is typically set to 95% and thus $\alpha = 0.05$. The coverage probability is the proportion of CIs that actually contain the population parameter of interest.

A CI of an estimate can be constructed as

$$CI\left(\hat{\overline{Y}}_m\right) = \hat{\overline{Y}}_m \pm t_{n-1} \sqrt{MSE\left(\hat{\overline{Y}}_m\right)} \qquad (8)$$

where MSE is the mean squared error and t is the upper $\alpha/2$ point of the Student's t distribution with $n - 1$ degrees of freedom (Thompson, 2002, p. 30). The MSE is the sum of variance and squared bias. Since the bias cannot be determined in synthetic estimation, we will use the variance estimates instead of MSEs in Eq. (8). The term $\sqrt{MSE(\cdot)}$ is also known as the standard error (SE) which is a common measure of uncertainty. Half CIs ($CI(\cdot)/2$, $\alpha = 0.05$) are consequently $t_{n-1}SE \approx 2SE$. The notation of the CIs follows the notation of the variance estimators. For example, $CI_{prhs}(\hat{\overline{Y}}_m)$ is a CI based on variance estimator (7).

Two additional kinds of intervals are well-known from regression analysis (e.g., Fahrmeir et al., 2007): (i) The CI, similar to the description above, is the interval around the regression line in which the expected value is likely to be if the regression were repeated with other samples taken from the same population; (ii) in contrast to a CI which focuses on inference for the superpopulation parameter, μ, a prediction interval (PI) focuses on an inference for a finite population value, y. Thus, the PI is the interval around the regression line in which the observations are likely to be if the regression were repeated with other samples taken from the same population.

The CI of a linear regression such as the linking model (1) is given by

$$CI(\hat{\mu}(\boldsymbol{x}_0)) = \hat{\mu}(\boldsymbol{x}_0) \pm t_{n-(p+1)} \sqrt{\boldsymbol{x}_0^T \left(\hat{\sigma}_\varepsilon^2 \left(\boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X}\right)^{-1}\right) \boldsymbol{x}_0} \qquad (9)$$

where $\hat{\mu}(\boldsymbol{x}_0)$ is the prediction of the response given the vector of explanatory variables $\boldsymbol{x}_0$ at an arbitrary position for which the CI should be estimated, p is the number of parameters in the linking model, and $\boldsymbol{X}$ is the design matrix of the linking model. Recalling that the covariance matrix of the parameter estimates of the linking model can be estimated by $\hat{\Sigma} = \hat{\sigma}_\varepsilon^2 (\boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X})^{-1}$, we see that $CI_p(\hat{\overline{Y}}_m)$ is numerically similar to $CI(\hat{\mu}(\boldsymbol{x}_0))$ if the mean of the auxiliary variables in stand m is used to estimate the CIs ($\overline{\boldsymbol{X}}_m = \boldsymbol{x}_0$). The difference between the two CIs is the degrees of freedom for the Student's t distribution which is $n - 1$ for $CI_p(\hat{\overline{Y}}_m)$ and $n - p$ for $CI(\hat{\mu}(\boldsymbol{x}_0))$. However, under practical conditions with a sufficient number of observations and a reasonable number of auxiliary variables, the two CIs can be considered equivalent.

The PI is given by

$$PI(\hat{y}(\boldsymbol{x}_0)) = \hat{y}(\boldsymbol{x}_0) \pm t_{n-(p+1)} \sqrt{\boldsymbol{x}_0^T \left(\hat{\sigma}_\varepsilon^2 \left(\boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X}\right)^{-1}\right) \boldsymbol{x}_0 + \hat{\sigma}_\varepsilon^2}. \qquad (10)$$

Comparing Eqs. (5) and (10), we see that $CI_{pr}(\hat{\overline{Y}}_m)$ is numerically similar to $PI(\hat{y}(\boldsymbol{x}_0))$ if the stand consists of only one population unit

$(N_m = 1)$ and $\overline{X}_m = x_0$. As for the CIs the difference between $CI_{pr}(\widehat{\overline{Y}}_m)$ and $PI(\hat{y}(x_0))$ is the degrees of freedom for the Student's t distribution. For stands with $N_m > 1$, $CI_{pr}(\widehat{\overline{Y}}_m)$ will be between $CI(\hat{\mu}(x_0))$ and $PI(\hat{y}(x_0))$.

3. Case study

3.1. Overview

The aim of the case study was to compare the variance estimators previously described with respect to empirical coverage proportions (ECPs) for estimates of mean stand-level timber volume. NFI plot data were used to fit the linking model used with the estimators. The study area was located in Vestfold county, in southern Norway. The validation data were independent of the calibration data as they were obtained from 64 stands, each of which included 5–7 sample plots. The validation stands were created by assuming that one stand consists only of the sample plots within the stand. The validation data are mainly used to calculate the coverage proportion of the confidence intervals obtained from the different variance estimators. Validation data may, in operational applications, not be available.

3.2. Calibration data

Data for n = 131 sample plots in Vestfold county measured between 2009 and 2013 were used to fit the linking model (Breidenbach & Astrup, 2012). General information on the NFI can be found in Tomter, Hylen, and Nilsen (2010) and Kolshus (2014, Chapter 7.2). The NFI uses an interpenetrating panel design where one-fifth of the sample plots are surveyed every year. The NFI uses permanent, circular, 250-m² sample plots on which all trees with diameter at breast height (dbh, 1.3 m) ≥ 5 cm are measured. Trees for height measurements are selected using a relascope with an expected number of 10 trees per plot. All heights are measured if the plot contains 10 or fewer trees. Heights of the remaining trees are estimated using height curves based on the measured trees. Timber volumes of single trees are estimated using species-specific volume models with dbh and height as explanatory variables (Braastad, 1966; Brantseg, 1967; Vestjordet, 1967). The variable of interest in this study, timber volume per hectare, is the sum of the single tree timber estimates per sample plot scaled to a per unit area. Uncertainties of the single tree timber volume model predictions are ignored.

Since the validation data were obtained in 2011, the timber volumes observed at the NFI sample plots were linearly interpolated and extrapolated to 2011. Without interpolation and extrapolation, the model fit to NFI data would have exhibited a systematic lack of fit when compared to the validation data. Details of the interpolation and extrapolation are in the Appendix A. Table 1 gives an overview of plot level characteristics of the response variable.

3.3. Validation data

Stand polygons were available from a previous forest management inventory (FMI). A first step in Norwegian FMIs is stand delineation based on manual photo interpretation. Data for 382 sample plots clustered in 64 stands were used as validation data; 5–7 plots were measured in each stand. The stands were selected under some constraints: (i) the stands had to be between 1–3 ha in size and compact with $\sqrt{A}/P > 0.2$ where A is the area and P is the perimeter of the polygon to minimize problems with defining stand borders in the field; (ii) in order to assure a wide range of volumes in the validation data, the same number of stands were randomly selected from two strata. The strata were formed based on the estimated volume in the FMI which was greater than 150 m³/ha in one stratum and less than or equal to 150 m³/ha in the other. A total of 34, 15 and 15 stands were selected

Table 1
Characteristics of the variable of interest (timber volume, m³/ha).

	Mean	SD	Max
Calibration plots	164.75	124.71	756.32
Validation plots	193.02	141.23	947.80
Validation stands	193.53	113.82	547.54

in the municipalities Lardal, Holmestrand, and Stokke in Vestfold county.

A 20-m × 20-m grid was superimposed over the selected stands, and 5–7 randomly selected grid intersections per stand were used as plot locations. Field crews navigated to the sample plots using hand-held GPS receivers. The established plot position was recorded with a differential GPS resulting in accuracies that are assumed to be on the order of one meter. Field work was carried out in 2011 following the NFI field protocol described in the section above. Parts of the data were used in previous studies by Solberg, Astrup, Breidenbach, Nilsen, and Weydahl (2013) and Breidenbach and Astrup (2014). Table 1 gives an overview of plot and stand level characteristics of the response variable. The validation plots do not coincide with the calibration plots.

3.4. Aerial photogrammetry data

Overlapping digital aerial images with a ground sampling distance (resolution) of 20 cm were acquired in 2010 using an UltraCamX sensor. The images were photogrammetrically processed using the default NGATE setting of SocetSet 5.5.0 which resulted in a digital surface model (DSM) of 1 m resolution. The DSM was normalized to heights above ground by subtracting a digital terrain model (DTM). The best available DTM was used for normalization. For the largest part of the study area, this was the Norwegian standard DTM with a resolution of 10 m. In Lardal municipality, an ALS DTM with 1 m resolution was available. Heights above ground estimated from aerial photogrammetry data are denoted AP heights. The data were interpolated to 20 cm × 20 cm grids and delivered with true ortho photographs by the data provider. More details of the data can be found in Breidenbach and Astrup (2012).

A variety of height and density metrics for the AP heights above ground (measured in dm) within the sample plots was calculated using FUSION (McGaughey, 2014). All statistical calculations were carried out using the R software for statistical computing (R Development Core Team, 2014).

3.5. Linking model

After comparison of several model alternatives, a simple linear linking model with intercept term and the mean of the AP heights as the only explanatory variable was fit to the calibration data:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = \{1, \ldots, n = 131\} \tag{11}$$

where i indexes sample plots, y_i is timber volume (m³/ha), β is a regression coefficient, x_i is the mean AP height. To accommodate heteroskedasticity, we fitted a weighted model (see Eq. (1)). Because residual variances were proportional to the explanatory variable, each observation in the regression model was weighted by the inverse of $\sigma_i^2 = \sigma_\varepsilon^2 x_i$ where σ_ε^2 is the residual variance. Spatial autocorrelation was not considered in the model since the sample plots have a minimum separation distance of 3 km which is greater than the range of

spatial correlation. Besides the coefficient of determination, the root mean squared error (RMSE) was used as a measure of model fit:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (12)$$

where y_i is the observed and $\hat{y}_i$ the predicted value. RMSE% is defined as the RMSE as a percentage of the mean observed volume at the NFI plots.

The linking model had a $R^2 = 0.69$, RMSE $= 75.0$ m^3/ha and RMSE% $= 46\%$. The estimated parameters were $\hat{\beta}_0 = 25.43$ (p-value < 0.001), $\hat{\beta}_1 = 1.73$ (p-value < 0.001) and $\hat{\sigma}_\varepsilon = 8.7$ m^3/ha. The estimated parameter covariance matrix used in the variance estimators was

$$\hat{\Sigma} = \begin{array}{c c c} & \hat{\beta}_0 & \hat{\beta}_1 \\ \hat{\beta}_0 & 19.594 & -0.244 \\ \hat{\beta}_1 & -0.244 & 0.010 \end{array} . \qquad (13)$$

3.6. Modeling spatial autocorrelation using validation data

The calibration data (NFI plots) which were interpolated and extrapolated to 2011 coincided well with the validation data (Fig. 1). The generalized least squares method was used to fit a model to the sample plot data for the validation stands to estimate the spatial autocorrelation structure of the residuals within stands:

$$y_{mi} = \beta_0 + \beta_1 x_{mi} + \varepsilon_{mi}, \quad \varepsilon_m \sim N(0, \sigma_\varepsilon^2 \boldsymbol{W}_m), \\ m = \{1, ..., 64\} \qquad (14)$$

where m indexes stands and i indexes plots within stands. The stand-specific weighting matrix $\boldsymbol{W}_m$ has the diagonal elements $\hat{\sigma}_i^2 = \hat{\sigma}_\varepsilon^2 x_i$. The off-diagonal elements of $\boldsymbol{W}_m$ are given by $\hat{\sigma}_i \hat{\sigma}_j \hat{\varrho}_{ij}$. We assumed a stationary and isotropic Gaussian process without nugget to estimate the correlation among pairs of residuals within a stand

$$\hat{\varrho}_{ij} = \exp\left[-\left(s_{ij}/\hat{\rho}\right)^2\right] \qquad (15)$$
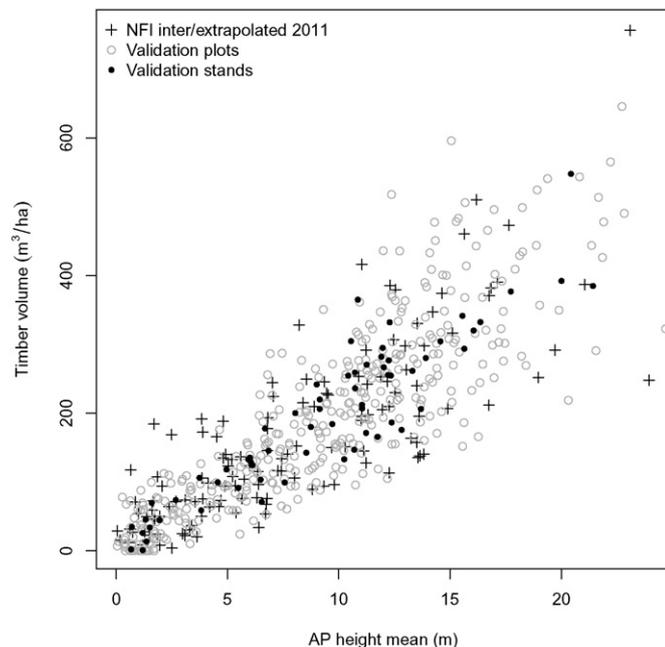

Fig. 1. Timber volume versus mean height derived from aerial photogrammetry (AP) for NFI and validation data.

where $\hat{\rho}$ is the estimated range and s_{ij} is the Euclidean distance between two sample plots (Pinheiro & Bates, 2002, p. 232). The gls function in the R package nlme (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2013) was used to fit the model. The spatial autocorrelation function can also be localized as described for example by Räty, Heikkinen, and Kangas (2011). However, to focus on the estimators, we chose a relatively simple stationary model.

A range ($\hat{\rho}$) of 23.0 m of the spatial autocorrelation was estimated from the validation data.

3.7. Application of the estimators

The validation plots were treated as grid cells in the estimators, such that a stand consists of only the grid cells. The advantage of this procedure is that observations (rather than estimates) at stand level are available. The disadvantage is that the stands are rather small because they consist of only 5–7 sample plots. This means in the estimators $N_m = n_m$ and $\sum N_m = 382$.

The ECP is defined as the proportion of stands for which the observed mean timber volume is within the CI. The CIs were estimated such that the coverage proportion of all estimators should be 95% by letting $\alpha = 0.05$. The spatial autocorrelation structure described in (14) was used in the variance estimator $\widehat{Var}_{prhs}(\widehat{\overline{Y}}_m)$ (7). The ECP of $\widehat{Var}_{prhs}(\widehat{\overline{Y}}_m)$ was estimated with the spatial range $\hat{\rho} = 23.0$ m estimated using the validation data. In addition, the ECP of $\widehat{Var}_{prhs}(\widehat{\overline{Y}}_m)$ was estimated with different alternative values of ρ between 0 and 120 m to evaluate the sensitivity of the estimates to spatial autocorrelation.

To compare the CIs of different estimators given different stand sizes, stands were simulated. This procedure is not a part of the validation; it merely shows the trend in the CIs given stand size. The simulated stands were square and all grid cells were assigned the average AP mean height observed at the validation plots (10 m). The only difference between the simulated stands was their size which ranged between one and $40 \times 40 = 1600$ grid cells (stand sizes of 256 m^2 to 40.64 ha).

4. Results

The average of all regression-synthetic mean stand-level timber volume estimates was 186.0 m^3/ha. While the linking model resulted in reasonable estimates for most stands, mean stand-level timber volume was clearly underestimated for the stands with the largest and fifth-largest observed timber volume (Fig. 2).

None of the CIs had an ECP of 95% without adjustments (Table 2). The ECP of $CI_p(\widehat{\overline{Y}}_m)$, which included only parameter uncertainty, was especially low (37%), whereas the ECP of $CI_{prhs}(\widehat{\overline{Y}}_m)$, which included parameter uncertainty and spatially correlated heteroskedastic residual uncertainty with a range of 23.0 m was 91%. ECPs of $CI_{prhs}(\widehat{\overline{Y}}_m)$ were also estimated for other values of the range of spatial autocorrelation. The ECPs ranged from 87% to 98% with a range of spatial correlation of 0 to 120 m (Fig. 3).

The PI had an ECP of 100% but was more than 50% bigger than $CI_{prhs}(\widehat{\overline{Y}}_m)$. The sizes of $CI_{pr}(\widehat{\overline{Y}}_m)$, $CI_{prh}(\widehat{\overline{Y}}_m)$ and $CI_{prhs}(\widehat{\overline{Y}}_m)$ decrease with stand size if all other parameters are held constant. The larger the spatial autocorrelation, the slower is the decrease of the CI length. For the stands simulated to visualize the development of the CIs given different stand sizes, $CI_{prh}(\widehat{\overline{Y}}_m)$ was very close to $CI_p(\widehat{\overline{Y}}_m)$ for stands of 10 ha and larger. The dependence of the CIs obtained from different estimators on stand size is shown in Fig. 4.

Ignoring the residual variability ($CI_p(\widehat{\overline{Y}}_m)$) or the spatial correlation ($CI_{prh}(\widehat{\overline{Y}}_m)$) resulted in much smaller CIs than $CI_{prhs}(\widehat{\overline{Y}}_m)$ which, however, would come with the price of a lower ECP.
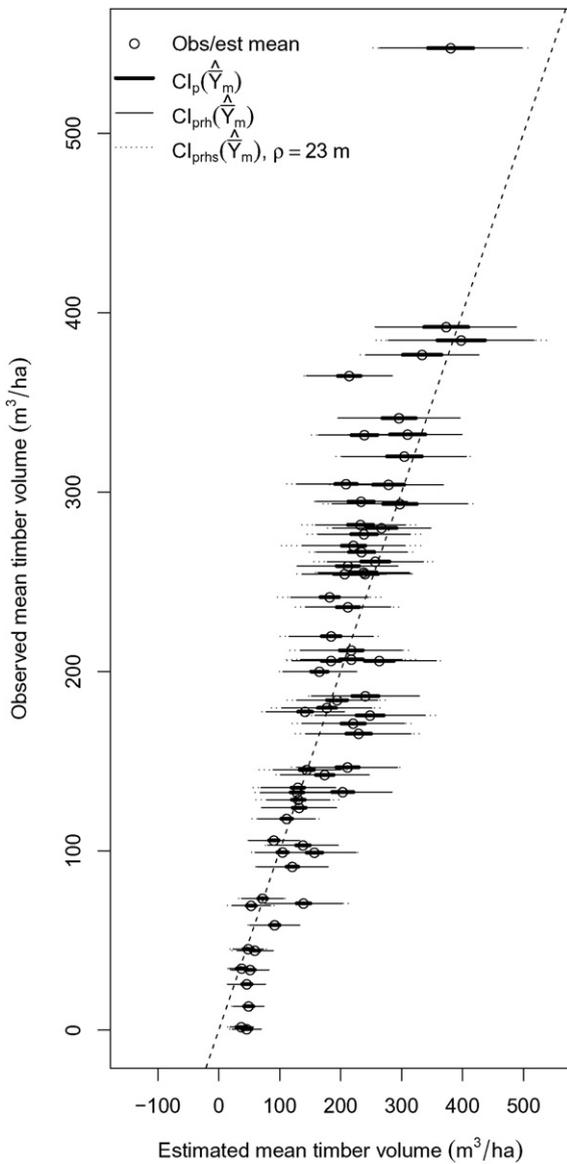
Fig. 2. Observed versus estimated mean timber volume on stand level with 95% confidence intervals based on different estimators and 1:1 line (dashed diagonal line).

Fig. 3. Development of the ECP of $CI_{prhs}(\widehat{\overline{Y}}_m)$ given the spatial range of autocorrelation (ρ).

5. Discussion

In this study, we compared model-based variance estimators for regression-synthetic estimates of small areas. One estimator was based on the estimation of superpopulation parameters (Kangas, 1996; Mandallaz, 2013; McRoberts et al., 2014), the other set of estimators was based on the prediction of finite population values (McRoberts, 2006). All compared variance estimators can be obtained from the most general variance estimator ($\widehat{Var}_{p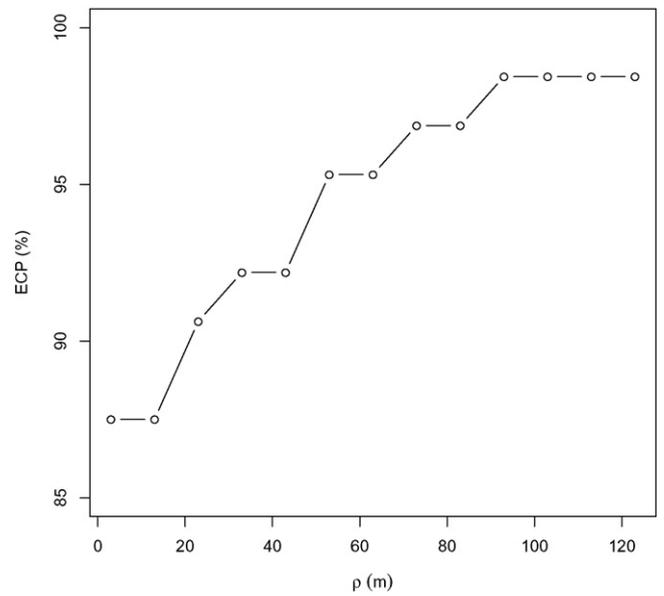rhs}(\widehat{\overline{Y}}_m)$, estimator (7)) by ignoring spatial autocorrelation ($\widehat{Var}_{prhs}(\widehat{\overline{Y}}_m)$, estimator (6)) and heteroskedasticity ($\widehat{Var}_{pr}(\widehat{\overline{Y}}_m)$, estimator (5)). Ignoring the influence of residual variation in general yields the estimator based on the estimation of superpopulation parameters ($\widehat{Var}_p(\widehat{\overline{Y}}_m)$, estimator (3)).

CIs obtained from the variance estimators based on the prediction of finite population values decrease with the size of the small areas and are between the PI and CI of the linking model. With increasing size of the small areas, all compared variance estimators converge. How fast $\widehat{Var}_{prhs}(\widehat{\overline{Y}}_m)$ approaches $\widehat{Var}_p(\widehat{\overline{Y}}_m)$ depends on the range of spatial autocorrelation and the residual variance of the linking model.

The estimator $\widehat{Var}_p(\widehat{\overline{Y}}_m)$ (Kangas, 1996; Mandallaz, 2013; McRoberts et al., 2014), which is numerically almost equivalent to the CI of the linking model, only considers the uncertainty in model parameter estimates and underestimated the variance of these very small stands (<1 ha). The estimator is, however, still appropriate for large stands. Considering the residual variance ($\widehat{Var}_{prh}(\widehat{\overline{Y}}_m)$) and spatial autocorrelation among residuals ($\widehat{Var}_{prhs}(\widehat{\overline{Y}}_m)$) improved the ECP compared to $\widehat{Var}_p(\widehat{\overline{Y}}_m)$. In our case study stands would have to be larger than 10 ha before the $\widehat{Var}_{prhs}(\widehat{\overline{Y}}_m)$ would be similar to $\widehat{Var}_p(\widehat{\overline{Y}}_m)$. Using the PI as an estimate of uncertainty would be very conservative as it represents the uncertainty to be expected on the grid-cell level.

McRoberts (2006) described both, $\widehat{Var}_p(\widehat{\overline{Y}}_m)$ and $\widehat{Var}_{prhs}(\widehat{\overline{Y}}_m)$, in the context of estimating forest area within small areas using Landsat images. In this study we extended the estimator based on the prediction of finite population parameters for heteroskedasticity and applied it to

Table 2
Empirical coverage proportions of the estimators.

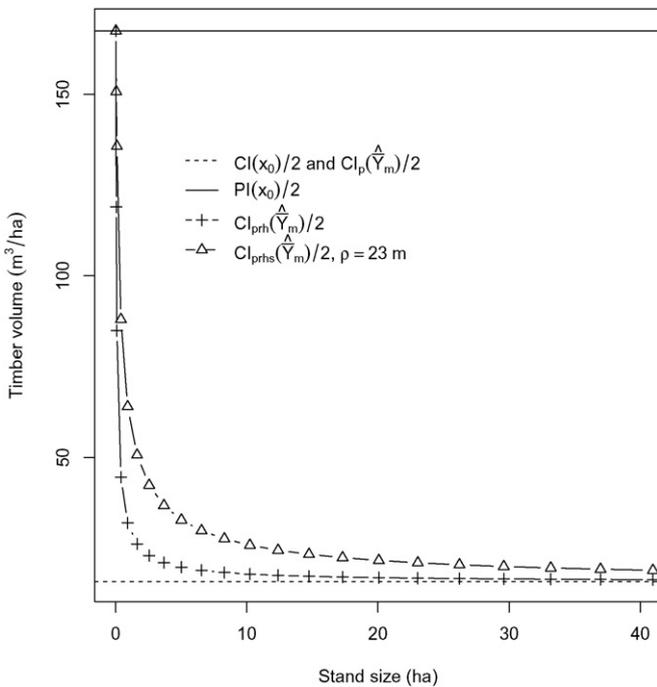| Estimator	Average half interval (m³/ha)	Average half interval (%)	# plots within interval	ECP (%)
$CI(\hat{\mu}(\boldsymbol{x}_0))$	17.1	9.8	24	37.5
$CI_p(\widehat{\overline{Y}}_m)$	17.1	9.8	24	37.5
$CI_{prh}(\widehat{\overline{Y}}_m)$	68.2	39.7	56	87.5
$CI_{prhs}(\widehat{\overline{Y}}_m), \rho = 23$ m	78.2	45.7	58	90.6
$PI(\hat{y}(\boldsymbol{x}_0))$	159.8	93.8	64	100.0

Fig. 4. Development of the CIs obtained from different estimators for simulated stands with an average AP mean height but different sizes.

a metric response variable. Since McRoberts (2006) applied both estimators to very large "small" areas, the variance estimates were similar up to the third digit and the estimator based on the prediction of finite population values was not further considered. Here we see that the estimator $\widehat{Var}_{prhs}(\overline{\widehat{Y}}_m)$ has the property of being dependent on the size of the small area.

Bias resulting from model lack of fit is not considered in synthetic estimation and stand-level observations are necessary to accommodate the fact that population-level models for various reasons do not fit to some stands. Also the variance estimators analyzed in this study do not change this fundamental problem. Bias is therefore also the reason why the ECP of $\widehat{Var}_{prhs}(\overline{\widehat{Y}}_m), \hat{\rho} = 23$ m is < 95%. Stands for which the linking model does not fit will necessarily result in a reduced ECP. Often, stand-level observations will be scarce because stands are too small and measurements of sample plots too expensive. If stand-level observations are available nonetheless, a variety of methods exist to improve the estimates (e.g., Goerndt, Monleon, & Temesgen, 2011; Magnussen et al., 2014; Mandallaz, 1991). As the linking model plays a prominent role in the model-based estimators discussed here, its properties need to be checked carefully. This includes tests for multicollinearity if more than one explanatory variable is selected.

The range and structure of the spatial autocorrelation process needs to be estimated to use $\widehat{Var}_{prhs}(\overline{\widehat{Y}}_m)$. This needs to be considered in the survey design. NFI data will usually not be suitable for this purpose because the distance between single plots is frequently selected in a way that prevents estimation of spatial autocorrelation.

All implemented estimators are available in an R package (Breidenbach, 2015).

6. Conclusions

Variance estimators based on the prediction of finite population values are recommended for small areas containing few grid cell predictions (e.g., small stands). They are applicable to small areas containing many or few grid cell predictions and decrease with the number of grid cells. Confidence intervals (CIs) obtained from variance estimators

based on the prediction of finite population values are between the CI and prediction interval (PI) of the linking model. For small areas containing many grid cell predictions (e.g., large stands), the CI of the linking model can be used as a simple-to-calculate indicator of uncertainty because it is practically equivalent to all variance estimators considered in this study.

Appendix A

The NFI plots (calibration data) were measured between 2008 and 2012. The validation plots were measured in 2011. In order to make the data sets match, the response variable (timber volume per hectare) of the NFI plots was interpolated or extrapolated to the year 2011 based on the two last observations on the respective NFI plot.

No interpolation or extrapolation was needed for 24 NFI plots which were measured in 2011. On all other plots except for those where selective harvests or mortality occurred, linear interpolation or extrapolation based on the two last observations at a NFI sample plot was used to estimate the timber volume in 2011. Extrapolation of one and two years was needed for 30 and 27 sample plots.

Selective harvests or mortality occurred on 13 NFI plots. For those plots, the last observed volume was used as the interpolated (6 plots) and extrapolated (7 plots) timber volume of 2011.

References

Bohlin, J., Wallerman, J., & Fransson, J. E. (2012). Forest variable estimation using photogrammetric matching of digital aerial images in combination with a high-resolution DEM. *Scandinavian Journal of Forest Research*, *27*(7), 692–699.

Braastad, H. (1966). *Volume tables for birch. (In Norwegian with English summary)*. Oslo: Meddr. Norske SkogforsVes, 265–365.

Brantseg, A. (1967). *Volume functions and tables for Scots pine. South Norway. (In Norwegian with English summary)*. Meddr. Norske SkogforsVes, 695–739.

Breidenbach, J. (2015). JoSAE: Functions for some unit-level small area estimators and their variances. R package version 0.2.3. URL http://CRAN.R-project.org/package=JoSAE

Breidenbach, J., & Astrup, R. (2012). Small area estimation of forest attributes in the Norwegian National Forest Inventory. *European Journal of Forest Research*, *131*(4), 1255–1267.

Breidenbach, J., & Astrup, R. (2014). The semi-individual tree crown approach. In M. Maltamo, E. Næsset, & J. Vauhkonen (Eds.), *Forestry Applications of Airborne Laser Scanning*. Springer.

Chambers, R., & Clark, R. (2012). *An introduction to model-based survey sampling with applications*. Oxford University Press.

Fahrmeir, L., Kneib, T., & Lang, S. (2007). *Regression*. Springer.

Goerndt, M. E., Monleon, V. J., & Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Canadian Journal of Forest Research*, *41*(6), 1189–1201.

Goerndt, M. E., Monleon, V. J., & Temesgen, H. (2013). Small-area estimation of county-level forest attributes using ground data and remote sensed auxiliary information. *Forest Science*, *59*(5), 536–548.

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

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: Appreciating the difference. *Canadian Journal of Forest Research*, *28*(10), 1429–1447.

Kangas, A. (1996). Small-area estimates using model-based methods. *Canadian Journal of Forest Research*, *26*(5), 758–766.

Kangas, A. (2006). Model-based inference. In M. Maltamo, & A. Kangas (Eds.), *Forest Inventory* (pp. 39–52). Springer.

Kolshus, H. H. (Ed.). (2014). *Greenhouse Gas Emissions 1990–2012, National Inventory Report*. The Norwegian Environment Agency.

Lappi, J. (2001). Forest inventory of small areas combining the calibration estimator and a spatial model. *Canadian Journal of Forest Research*, *31*(9), 1551–1560.

Magnussen, S., Mandallaz, D., Breidenbach, J., Lanz, A., & Ginzler, C. (2014). National forest inventories in the service of small area estimation of stem volume. *Canadian Journal of Forest Research*, *44*(9), 1079–1090.

Maltamo, M., & Packalen, P. (2014). Species-specific management inventory in Finland. In M. Maltamo, E. Næsset, & J. Vauhkonen (Eds.), *Forestry Applications of Airborne Laser Scanning.* Springer.

Mandallaz, D. (1991). *A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models.* Diss. eth no. 9378, Ph. D. thesis ETH Zürich, Switzerland: Academic Press (Chair of forest inventory and planning).

Mandallaz, D. (2013). Design-based properties of some small-area estimators in forest inventory with two-phase sampling. *Canadian Journal of Forest Research*, *43*(5), 441–449.

McGaughey, R. (2014). *Fusion.* USDA forest service.

McRoberts, R. E. (2006). A model-based approach to estimating forest area. *Remote Sensing of Environment*, *103*(1), 56–66.

McRoberts, R., Andersen, H. -E., & Næsset, E. (2014). Using airborne laser scanning data to support forest sample surveys. In M. Maltamo, E. Næsset, & J. Vauhkonen (Eds.), *Forestry Applications of Airborne Laser Scanning.* Springer.

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

Næsset, E. (1997). Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sensing of Environment*, *61*(2), 246–253.

Næsset, E. (2014). Area-based inventory in Norway — From innovation to an operational reality. In M. Maltamo, E. Næsset, & J. Vauhkonen (Eds.), *Forestry Applications of Airborne Laser Scanning.* Springer.

NCHS (1968). *Synthetic state estimates of disability. PHS Pub. No. 1759,* National Center for Health statistics.

Pinheiro, J., & Bates, D. (2002). *Mixed-effects models in S and S-plus.* Springer (528 pp.).

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2013). *nlme: Linear and non-linear mixed effects models.* R package version 3.1-113.

Prasad, N., & Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, *85*(409), 163–171.

R Development Core Team (2014). *R: A language and environment for statistical computing.*

Rao, J. N. K. (2003). *Small area estimation.* Wiley-Interscience.

Räty, M., Heikkinen, J., & Kangas, A. (2011). Kriging with external drift in model localization. *Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS)*, *3*(1), 1–14.

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

Solberg, S., Astrup, R., Breidenbach, J., Nilsen, B., & Weydahl, D. (2013). Monitoring spruce volume and biomass with InSAR data from TanDEM-X. *Remote Sensing of Environment*, *139*, 60–67.

Steinmann, K., Mandallaz, D., Ginzler, C., & Lanz, A. (2013). Small area estimations of proportion of forest and timber volume combining LiDAR data and stereo aerial images with terrestrial data. *Scandinavian Journal of Forest Research*, *28*(4), 373–385.

Thompson, S. (2002). *Sampling* (2nd ed.). Wiley-Interscience (367 pp.).

Tomter, S., Hylen, G., & Nilsen, J. -E. (2010). Norway's national forest inventory. In E. Tomppo, T. Gschwantner, & M. Lawrence (Eds.), *National forest inventories: Pathways for common reporting* (pp. 411–424). Verlag: Springer.

Vastaranta, M., Wulder, M. A., White, J. C., Pekkarinen, A., Tuominen, S., Ginzler, C., et al. (2013). Airborne laser scanning and digital stereo imagery measures of forest structure: Comparative results and implications to forest mapping and inventory update. *Canadian Journal of Remote Sensing*, *39*(5), 382–395.

Vestjordet, E. (1967). *Functions and tables for volume of standing trees. Norway spruce. (In Norwegian with English summary).* Meddr. Norske SkogforsVes, 543–574.