Forest Ecosystems

RESEARCH Open Access

Benefits of past inventory data as prior information for the current inventory

Annika Kangas[1*], Terje Gobakken[2] and Erik Næsset[2]

Abstract

Background: When auxiliary information in the form of airborne laser scanning (ALS) is used to assist in estimating the population parameters of interest, the benefits of prior information from previous inventories are not self-evident. In a simulation study, we compared three different approaches: 1) using only current data, 2) using non-updated old data and current data in a composite estimator and 3) using updated old data and current data with a Kalman filter. We also tested three different estimators, namely i) Horwitz-Thompson for a case of no auxiliary information, ii) model-assisted estimation and iii) model-based estimation. We compared these methods in terms of bias, precision and accuracy, as estimators utilizing prior information are not guaranteed to be unbiased.

Results: The largest standard errors were obtained when neither prior information nor auxiliary information were used. If a growth model was not applied to update the old data, the resulting composite estimators were biased. Largest RMSEs were obtained using non-updated prior information in a composite estimator. Using the ALS data as auxiliary information produced smaller RMSE than using prior information from the old inventory. The smallest RMSEs were obtained when both the auxiliary data and updated old data were used. With growth updating the bias can be substantially reduced, although design-unbiasedness of the estimator cannot be guaranteed.

Conclusions: Prior information from old inventory data can be useful also when combined with highly accurate auxiliary information, when both data sources are efficiently used. The benefits obtained from using the old data will increase if the past harvests can be detected without errors from changes in the auxiliary data instead of being predicted with models.

Keywords: Data fusion, Kalman filtering

Introduction

In forest inventory, many types of information can be used in addition to an actual sample of observations. There are at least two good reasons for using such information in forest inventory: 1) we can either improve the accuracy (mean square error, MSE) of the estimates while keeping the costs the same level as before, or 2) we can reduce the costs without reducing the accuracy. Obviously, this necessitates the auxiliary data to be cheap or free (i.e. the costs are assumed to be sunk costs from some previous use of the data). It is possible to use remotely sensed data, e.g. from satellite images or airborne laser scanning (ALS), as auxiliary information using stratification, model-assisted or model-based frameworks (e.g. Gregoire et al. 2011; Ståhl et al. 2011). In addition, it is possible to combine current data from most recent forest inventory with old data from previous inventories or existing models constructed from old data as prior information (e.g. Tomppo and Halme 2004; McRoberts et al. 2014).

One method for using old data in forest inventory is sampling with partial replacement (SPR, Ware and Cunia 1962). For estimating the current population mean, two independent estimates are combined to form a single linear unbiased estimator. The weight placed on the two estimates depends on the correlation between

* Correspondence: annika.kangas@luke.fi
[1]Bioeconomy and Environment, Natural Resources Institute Luke (Finland), Yliopistokatu 6, 80100 Joensuu, Finland
Full list of author information is available at the end of the article

the re-measured plots on first and second occasions and on the estimated variances of the population parameter estimators on these two occasions.

In the case of three or more successive inventories, SPR results in quite complicated estimators. However, Bickford et al. (1963) published estimators based on a different approach, namely on composites of two estimators, weighted inversely to their variances (Meier 1953). The first estimator is based on the data from old field plots updated to the current occasion using the change observed in re-measured plots through a regression estimator, and the other on the current field plots. Scott (1984) extended this approach to include also change estimation. Scott and Köhl (1994) used a similar approach to provide composite estimators also for a stratified case, by which it would be possible to apply auxiliary information from both remotely sensed data and from previous inventories.

When information of growth is available in the form of a growth model, it can be utilized in a Kalman filter (Dixon and Howitt 1979). In a Kalman filter approach, the old data from previous inventories are updated with growth and harvest information and the updated data are used as prior information. The growth model used in Dixon and Howitt (1979) was crude; it simply gave the proportional change of the state vector over time, and the harvests were assumed known control actions. Kangas (1991) used data updated by tree-wise growth models and stand-level harvest models as prior information. However, for estimating the precision of the resulting estimates, all changes were described using a single proportional change of state, which was used in a Kalman filter-type of analysis.

More advanced types of Kalman filters can be applied by allowing for non-linear growth models (e.g. Ehlers et al. 2013). However, when accurate auxiliary information such as that provided by ALS or digital aerial photogrammetry data is available, utilizing prior information from old data may produce only marginally smaller MSE than using only the most recent data (Nyström et al. 2015).

Even if we have accurate current data, there is still merit to see if the overall performance can be improved by using old data. The aim of this study was to assess if prior information from old inventory enhances the accuracy of the results in a case where auxiliary information from ALS is available. We compared three different approaches: 1) using only current data, 2) using non-updated old data and current data in a composite estimator and 3) using updated old data and current data with a Kalman filter. We tested three different estimators, namely i) Horwitz-Thompson for a case of no auxiliary information, ii) model-assisted estimation and iii) model-based estimation. We compared these methods in terms of bias, precision and accuracy, as estimators utilizing prior information are not guaranteed to be unbiased.

Materials

The field data

The empirical part of this study was based on data from Våler Municipality in south-eastern Norway. The study area (altogether 853 ha) is located in a boreal forest region. The forest is actively managed, with Norway spruce (*Picea abies* (L.) Karst.) and Scots pine (*Pinus sylvestris* L.) as the dominant species.

Prior to the field inventory, photo interpretation was adopted to delineate the study area into forest stands, each belonging to one of four classes related to stand age and species dominance: 1) recently regenerated forests, 2) young forests, 3) mature, spruce-dominated forests, and 4) mature, pine-dominated forests. Only the strata 2–4 were included in this study due to deficient data collected for stratum one in 1999. As part of the plots were harvested during 1999–2010, recently regenerated stands were, however, also included in the analysis. A sample survey was conducted according to a systematic design with random start with sampling intensities approximately equal for the first three strata, but for the fourth stratum the intensity was only one third of that of the other three strata (Næsset 2002; Næsset et al. 2013, 2015).

Measurements were obtained for 178 systematically distributed, circular, 200-m^2 (radius 7.98 m) forest inventory plots measured in 1999 and 2010. Four plots were discarded from the analysis due to missing values. Tree-level aboveground biomass (*AGB*) was predicted using allometric models based on field observations of species and measurements of diameter at-breast-height (1.3 m) and height (Marklund 1988). Plot-level *AGB* was then predicted as the sums of individual tree *AGB* predictions, scaled to per-hectare values, and used as ground reference.

Wall-to-wall ALS data were acquired for the study area in 1999 and 2010. Pulse density was approximately 1.2 pulses per m^2 in 1999 and 7.3 pulses per m^2 in 2010. Empirical distributions of first and single echo heights were constructed for the 200-m^2 circular plots. The entire study area was tessellated into 200 m^2 regular squares (cells) and similar ALS echo distributions were constructed for each cell. A threshold of 1.3 m above the ground surface was used to remove the effects of echoes from ground vegetation whose biomass is not included in tree-level biomass. For each plot and cell, heights corresponding to the 0th, 10th, 20th, …, 90th percentiles ($p0$, $p10$, $p20$, …, $p90$) of the ALS height distributions were calculated. Furthermore, several measures of canopy density were derived. The range between 1.3 m above ground and the 95 percentile was divided into 10 vertical fractions of equal

height. Canopy densities were then calculated as the proportions of echoes with heights above fraction 0 (> 1.3 m), 1, …, 9 to total number of echoes (*d*0, *d*1, …, *d*9). Maximum value (*hmax*), mean value (*hmean*), and coefficient of variation (*hcv*) were also computed. Thus, 23 ALS metrics were available as explanatory variables. Næsset et al. (2013) provide more details for the study area and the dataset.

The copula population

We used the data from Våler to construct a simulated copula population. In the C vine copula, a multivariate distribution of the variables is formed. This is based on pair copulas that describe dependencies between each pair of the variables when the marginal distributions of these variables are transformed to uniform distributions (see Aas et al. 2009). The pairs are formed using a specific tree structure of the variables depicting the strength of the dependencies. For construction of the copula, we used the same approach as Myllymäki et al. (2017) and Kangas et al. (2016). That is, we calculated the empirical marginal distributions for the variables *AGB*, *p*0, *p*20, *p*40, *p*60, *p*80, *d*2, *d*4, *d*6 and *d*8 from the data using the logspline package in R (Kooperberg 2015, R Core team 2014) and estimated the C vine copula using the VineCopula package in R (Schepsmeier et al. 2015). In the current study, we included the *AGB* and the ALS metrics from both occasions to be able to analyse the case of using prior information. To our knowledge, this is the first simulation study involving also change.

The copula model was used to simulate 10,000 uniformly distributed observations with the modelled (pairwise) dependencies. These 10,000 observations can be interpreted as 200 m^2 grid cells mimicking similar cells in actual ALS data acquisition in an area of 200 ha. The copula population was then obtained by calculating the quantiles of the empirical distributions at those simulated uniformly distributed values. Figure 1 shows the dependency between the simulated *AGB*s on the two occasions in time. It reflects both the growth of the plots between the two points in time (dots above the red line) and the cuttings (dots below the red line).

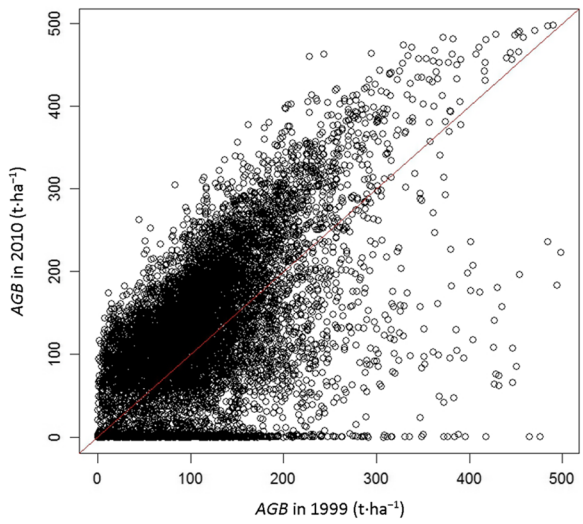

Fig. 1 The simulated values from the Copula population for the aboveground biomass (*AGB*) for 1999 and 2010

estimator of the meanMethods

Estimators to be compared

First, it is possible to use the field sample from both time points with a Horwitz-Thompson (HT) estimator, and make a composite of these two estimates. The HT estimator of total *AGB* is (e.g. Särndal et al. 1992, p 42)

$$\hat{t}_{\mathrm{HT}} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}, \tag{1}$$

where y_i is the *AGB* of cell i and π_i is the inclusion probability of cell i. Assuming simple random sampling without replacement this inclusion probability is n/N. The estimator of the mean is

$$\hat{\bar{y}}_{\mathrm{HT}} = \frac{1}{A}\hat{t}_{\mathrm{HT}}, \tag{2}$$

where A is the total area. Its variance estimator is (e.g. Särndal et al. 1992, p. 43)

$$\hat{Var}(\hat{\bar{y}}_{\mathrm{HT}}) = \frac{1}{A^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}, \tag{3}$$

where π_{ij} is the joint inclusion probability of cells i and j. In the case of simple random sampling, when $i = j$, this joint probability is π_i, otherwise it is $n(n-1)/N(N-1)$. These formulas can be extended also to stratified sampling.

Another option is to utilize auxiliary information and adopt a model-assisted estimator. Then, the difference estimator for the mean *AGB* is (e.g. Särndal et al. 1992, p 222)

$$\hat{\bar{y}}_d = \frac{1}{A}\left(\sum_{i=1}^{N}\hat{y}_i + \sum_{i=1}^{n}\frac{e_i}{\pi_i}\right), \tag{4}$$

where $\hat{y}_i$ is the model prediction for *AGB* in cell i and $e_i = y_i - \hat{y}_i$. Its variance estimator (the simplified estimator assuming g-weights to be 1 for all i, Särndal et al. 1992, p 362) is

$$\hat{Var}(\bar{\hat{y}}_d) = \frac{1}{A^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{e_i}{\pi_i}\frac{e_j}{\pi_j}. \quad (5)$$

Yet another option is to use a model-based estimator for the mean, i.e.,

$$\hat{\mu}_{MB} = \frac{1}{A}\left(\sum_{i=1}^{N}\hat{y}_i\right) = \frac{1}{A}\left(\sum_{i=1}^{N} f\left(X_i,\hat{\beta}\right)\right), \quad (6)$$

This is equal to the first part of the estimator in Eq. 4, meaning that the model predictions are used without calibrating with the sample data. Its variance can be estimated with (Ståhl et al. 2011, Eq. 7)

$$\hat{Var}(\hat{\mu}_{MB}) = \sum_{k=1}^{p}\sum_{l=1}^{p} Cov\left(\hat{\beta}_k,\hat{\beta}_l\right)\hat{f}'_k\hat{f}'_l, \quad (7)$$

where $f'_k = \frac{\partial f(x,\beta)}{\partial\beta_k}$ is the partial derivative of the model f with respect to parameter β_k and $Cov(\hat{\beta}_k,\hat{\beta}_l)$ is covariance of the model parameters k and l. When the model is linear, the means of partial derivatives are equal to means of the independent variables leading to

$$\hat{Var}(\hat{\mu}_{MB}) = \sum_{k=1}^{p}\sum_{l=1}^{p} Cov\left(\hat{\beta}_k,\hat{\beta}_l\right)\bar{x}_k\bar{x}_l = \bar{\boldsymbol{X}}^T\boldsymbol{Cov}(\boldsymbol{\beta})\bar{\boldsymbol{X}}$$

(Kangas 2006). The residual errors of the model are assumed to have a negligible effect on the variance, meaning that the population model mean instead of a finite population mean is estimated (Ståhl et al. 2011, p. 99). If the residual errors are spatially correlated, that would introduce an additional term (McRoberts et al. 2018). In the current case, in the context of a relatively small area, the spatial correlation is likely to have a significant effect. It was, however, assumed negligible, as no mechanism to produce a specific spatial structure to the simulated data was available.

When two HT estimators (or model-assisted or model-based estimators) from two time points are available, a composite estimator can be formulated as

$$\bar{\hat{y}}_c = \alpha\bar{\hat{y}}_1 + (1-\alpha)\bar{\hat{y}}_2, \quad (8)$$

where subscripts c, 1 and 2 denote the composite estimator, the estimator for the first time point and the estimator for the second time point, respectively. α is calculated from the variances of the two estimates as

$$\alpha = \frac{w_t}{w},$$

where

$$w_t = var(\bar{\hat{y}}_t)^{-1} \text{ and } w = var(\bar{\hat{y}}_1)^{-1} + var(\bar{\hat{y}}_2)^{-1},$$

to obtain a composite where the individual estimators have the larger weight, the smaller the variance (e.g. Meier 1953; Scott and Köhl 1994). The variance for this composite estimator is (Shahar 2017)

$$Var(\bar{\hat{y}}_c) = \sum_{t=1}^{2} w_t^2 Var(\bar{\hat{y}}_t) = \frac{1}{\sum_{t=1}^{2}\frac{1}{Var(\bar{\hat{y}}_t)}}, \quad (9)$$

where t denotes the time points 1 and 2. When the variances are estimated, an unbiased estimator is (Meier 1953; Scott and Köhl 1994)

$$\hat{Var}(\bar{\hat{y}}_c) = \frac{\left(1+\frac{4}{w^2}\sum_{t=1}^{2}\frac{w_t(w-w_t)}{m_t}\right)}{w}, \quad (10)$$

where m_t denotes the degrees of freedom for $\bar{\hat{y}}_t$. In the studies case, the two samples were independent of each other, and also the estimates were therefore independent. If the estimators were correlated, the weights would be more complicated (Grafström et al. 2019).

It is also possible to use a Kalman filter to update the previous sample data using a growth model and combine it with the new sample data information. The growth model can be written using notation from (Ehlers et al. 2013; Nyström et al. 2015) as

$$x_{t+1} = ax_t + bu_t + e_t, \quad (11)$$

where e_t is an error term normally distributed with zero mean and variance q_t^2, x_t is the vector of the (random) state variables at time t, a is a coefficient describing the growth, u_t describes control actions and the coefficient b their impact. Instead of using fixed coefficients a and b to describe the growth and harvests, the changes can also be described with a (possibly non-linear) model g

$$x_{t+1} = x_t + g(x_t,\beta) + e_t \quad (12)$$

The model of the sampling system can be written as

$$y_t = x_t + v_t$$

The error term, v, is also normally distributed with zero mean and variance r_t^2. The residual error term can be interpreted as describing the sampling error for the new data in the model-based framework (see e.g. Cassel et al. 1977).

The Kalman estimator of the state vector can be calculated by the following procedure. The Kalman filter has

a prediction step and an update step that follow each other in sequence. The predicted conditional mean given all the data through time t is

$$x_{t+1|t} = ax_{t|t} + bu_t \quad (13)$$

and the conditional variance $p^2_{t+1|t}$ is

$$p^2_{t+1|t} = a^2 p^2_{t|t} + q^2_t \quad (14)$$

where $p^2_1 = r^2_1$. A sample is then taken to obtain y_{t+1}. The predicted value will almost never be the same as the observed value, so a residual vector η_{t+1} can be defined as

$$\eta_{t+1} = y_{t+1} - x_{t+1|t} \,. \quad (15)$$

The prior information, $x_{t+1|t}$, and the sample information, η_{t+1}, are then combined in the update cycle to yield

$$x_{t+1|t+1} = x_{t+1|t} + K_{t+1}\eta_{t+1} = (1-K_{t+1})x_{t+1|t} + K_{t+1}y_{t+1} \quad (16)$$

where

$$K_{t+1} = \frac{p^2_{t+1|t}}{p^2_{t+1|t} + r^2_{t+1}}$$

is the Kalman gain, and the variance of the assimilated value is

$$p^2_{t+1|t+1} = (1-K_{t+1})p^2_{t+1|t}.$$

Models used

In this study, we estimated the external models (i.e. models estimated from a dataset independent of the sample at hand) to be used in model-assisted estimation from the Våler plot data. As the copula population is simulated based on the same data, the models are not truly independent from the simulated data. However, the external model is fixed across the simulated samples. All the models were estimated using weighted regression to account for heteroscedasticity. This was carried out iteratively: the weights were estimated from the OLS model residuals, and the inverses of squared residuals were then used as weight in WLS.

For 1999, the estimated external model for AGB (t·ha^{-1}) was $AGB_{1999} = \beta_0 + \beta_1 p_{20\_1999} + \beta_2 p_{80\_1999} + \beta_3 d_{8\_1999} + \varepsilon_{1999}$ (see Table 1). The residual standard error $RSE = 32.77$, $R^2 = 0.7827$ and adjusted $R^2 = 0.7784$. The residuals of this model are presented in Fig. 2.

Table 1 The coefficients for a model of AGB in 1999

	Estimate	Std. Error	t value
Intercept	−55.18647	7.561298	−7.298545
p_{20_1999}	9.555293	1.39298	6.859605
p_{80_1999}	6.014979	0.7931157	7.583987
d_{8_1999}	82.058	21.02881	3.902171

For 2010, the estimated external model for AGB (t·ha^{-1}) was $AGB_{2010} = \beta_0 + \beta_1 p_{20\_2010} + \beta_2 p_{60\_2010} + \beta_3 d_{2\_2010} + \varepsilon_{2010}$ (see Table 2). The residual standard error $RSE = 41.81$, $R^2 = 0.2044$ and adjusted $R^2 = 0.8010$. The residuals of this model are presented in Fig. 3. The effect of cuttings after 1999 can be detected from the zero biomass measured in 2010, and also from the greater residual error than that observed in 1999.

The changes between 1999 and 2010 include both growth of the plots and the effect of harvests. Especially the effect of harvests is difficult to predict with a model, but unless the harvests cannot be assumed as known control actions, a model capable for predicting both is necessary for the Kalman filter approach.

The change model can be constructed in several different ways. The first option is to rely on the variables describing the growing stock, in this case the AGB from 1999, which is the typical way to make a growth model. Such a model would allow for predicting changes happening after either 1999 or 2010 inventory, using the AGB from the respective inventory. Another option is to utilize both the growing stock estimate and the ALS metrics. If only the metrics from 1999 are used in the model, the model allows for predicting the changes both after the 1999 or 2010 inventory using the respective metrics. If both 1999 and 2010 metrics are used in a model, the model can only be used to estimate the past changes between 1999 and 2010. However, such model is likely to be more accurate, as the differences in the 1999 and 2010 metrics enable close to direct detection of the changes.

In this case, the first option produced large standard errors, especially with respect to the harvests. Therefore, change (C) was predicted based on the observed AGB in 1999 and the ALS metrics. The first model using only 1999 ALS metrics for change in AGB (t·ha^{-1}) is $C1 = \beta_0 + \beta_1 AGB_{1999} + \beta_2 p_{60\_1999} + \beta_3 p_{80\_1999} + \beta_4 d_{2\_1999} + \beta_5 d_{6\_1999} + \varepsilon_{C1}$ (see Table 3).

The residual standard error $RSE = 34.50$, multiple $R^2 = 0.7792$, and adjusted $R^2 = 0.7726$. The residuals of this model are presented in Fig. 4. It is notable that the standard error of the change model is actually a little bit greater than that of the model for the AGB in 1999. The model is to some extent also able to capture the cuttings in addition to the growth. The predicted AGB in 2010 using the true values of AGB in 1999 and the predicted change are presented in Fig. 5. In some plots, the

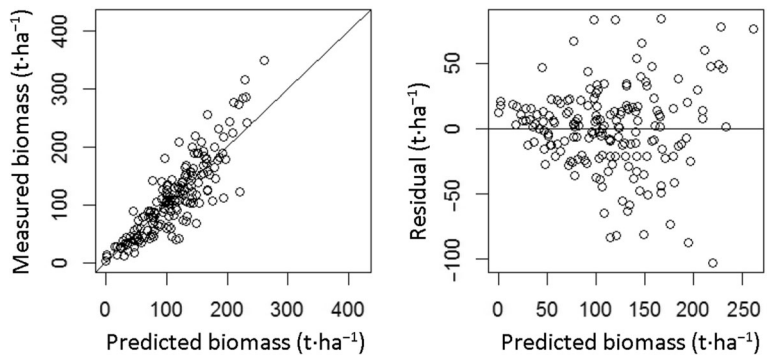

Fig. 2 The predicted *AGB* versus ground reference *AGB* in 1999 (left) and the residual plot (right)

predicted *AGB* is negative, but overall the model behaves logically. Negative predictions can, when the model is applied, be adjusted to zero, but in this case such correction was not made.

To analyse the effect of the change model, we estimated an alternative model for change *C* using also the 2010 ALS metrics as predictors. The model for change in *AGB* (t·ha^{-1}) is $C2 = \beta_0 + \beta_1 AGB_{1999} + \beta_2 p_{20\_2010} + \beta_3 d_{2\_2010} + \beta_4 d_{2\_1999} + \varepsilon_{C2}$ (see Table 4).

The residual standard error $RSE = 30.27$, $R^2 = 0.8218$ and adjusted $R^2 = 0.8176$. The residuals are presented in Fig. 6. In this model, the change in the density metric d_2 between 1999 and 2010 can be interpreted to describe the effect of harvests.

Simulations

We present different approaches to utilize data from the old inventory as prior information and assess their accuracy in a simulation study. In the copula population a simple random sample of size $n = 100$ was simulated $s = 5000$ times. Independent samples of size n were selected from the 1999 data and the 2010 data in order to calculate the results utilizing prior information, i.e. no re-measurements were assumed. The simulated (true) variance was calculated as the variance among the 5000 realizations of the sample. The bias was calculated as the difference of the true mean and the mean of the estimates of mean from these 5000 realizations and RMSE was calculated from them with

$$RMSE = \sqrt{var(\hat{\bar{y}}) + bias(\hat{\bar{y}})^2} \quad (17)$$

Table 2 The coefficients for a model of *AGB* in 2010

	Estimate	Std. Error	*t* value
Intercept	− 65.96272	10.65852	−6.188734
p_{20_2010}	14.41628	1.528627	9.430866
p_{60_2010}	3.469934	1.070011	3.242897
d_{2_2010}	35.11362	14.96293	2.346708

The estimated variance is a mean of sample variance estimates over these realizations.

In a case of model assisted inference, an external model (i.e. a model estimated from independent data previous to the sampling) is recommended, as using a model estimated from the sample at hand (internal model) has shown to lead to underestimation of variance (e.g. Kangas et al. 2016). While it is possible to reduce the underestimation by using a fixed mathematical form of a model, (i.e. the mathematical form of the model is assumed external while the coefficients are internal), we used an external model for the model-assisted estimation.

In the case of model-based inference, however, the inference is solely based on the model estimated from the sample. Thus, in model-based estimation, using an external model would mean that the sample at hand does not have any effect on the variance estimates, as all the terms in Eq. 7 would be fixed. Therefore, for all occasions, we used a model estimated from the sample for the model-based approach.

In a case of a change model, either an internal or an external model is applicable. Here both the change models (with and without 2010 ALS metrics) were assumed to be external, and the same model was used in all cases where a change was predicted (i.e. both for the model-based and the model-assisted approach). This is justified, as the growth models used for prediction are typically based on separate experimental data sets rather than inventory data. Moreover, change models estimated from the simulated samples proved to be fairly unstable. Both of the external change models had a mean error quite close to zero in the Copula population, with the mean change of 15.79 t·ha^{-1} for the population, and 15.03 t·ha^{-1} with the first change model and 15.31 t·ha^{-1} with the second change model.

Results

The HT estimator using solely the 2010 field data (i.e. without the ALS data or the old inventory data), produced the largest estimated and simulated variances

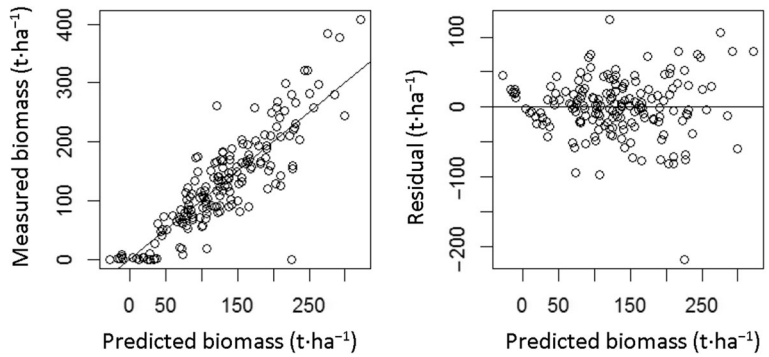

Fig. 3 The predicted versus ground reference *AGB* in 2010 (left) and the residual plot (right)

(Table 5). Both the model-assisted and model-based estimators produced markedly (39%–40%) smaller variances (simulated and estimated) than the HT estimator. These results were obtained even though a linear model with a good fit for *AGB* in 2010 was difficult to obtain as there were many plots with near-zero *AGB* (clear-felled after 1999) in the data. The reduction in variance was slightly larger for the model-based with respect to the simulated variances, but the difference was minor. In the case of linear models used here, the model-assisted estimator with internal model would actually produce an identical result to the model-based estimator. Since the mean of errors within the sample is zero using the internal model, the estimate (Eq. 4) adjusting the estimate for errors in the model predictions is the same as the model-based estimate (Eq. 6) not including an adjustment part. This does not, however, hold for non-linear models.

A composite of 1999 and 2010 HT estimates had clearly smaller variance than the HT estimate using solely the 2010 data. Introducing prior information in the form of old data reduced the variance almost as much as utilizing the auxiliary information from ALS: the simulated (true) variance was 5.66 compared to 5.24 in model-assisted. However, as the old sample plots were not updated, the resulting composite estimator for the 2010 *AGB* was clearly biased, and if the bias is taken into account, using purely 2010 data would be a better choice.

Table 3 The coefficients for a change model using ALS metrics from 1999

	Estimate	Std. Error	*t* value
Intercept	− 105.0843	12.77072	−8.228538
AGB_{1999}	−0.4174454	0.07490027	−5.57335
p_{60_1999}	−4.218624	3.475821	−1.213706
p_{80_1999}	5.513255	3.006514	1.83377
d_{2_1999}	158.1842	27.59308	5.73275
d_{6_1999}	76.81712	33.87098	2.267933

When both the prior information and the auxiliary ALS data were utilized, the variances were further reduced. The simulated variance of a composite of two model-assisted estimates was 3.80, i.e. markedly smaller than the pure model-assisted (5.24), but the biases were large. The results clearly show that composite estimators for which the growth and cuttings are not accounted for are highly biased.

A Kalman estimate based on HT estimator of *AGB* in 2010 and updated HT estimator from 1999 reduced both the bias and the variance, so that the root mean square error RMSE of the estimate was 6.16 compared to 5.24. However, even in this case utilizing model-assisted or model-based estimation instead of utilizing the prior information from old data is advisable.

The smallest RMSE estimates were obtained when the model-assisted approach was combined with a Kalman filter. The estimated variance was 4.23, and the simulated variance 3.73, indicating that the Kalman filter variance estimate was in this setting conservative. The Kalman filter with a model-based approach had an estimated variance of 4.12 and the simulated variance was 3.72, i.e. almost identical to the model-assisted results.

When an alternative growth model *C*2 with also 2010 ALS metrics included was used in the Kalman filter, the RMSEs were somewhat improved compared to those obtained with model *C*1 due to the improved accuracy of the model (Table 6). This means that the usefulness of the Kalman approach is highly dependent on the accuracy of the change models.

Discussion

The results of this study confirm that it is not self-evident to reduce the RMSE of the population parameters by using prior data from previous inventories, if the estimation is already enhanced with accurate current remotely sensed auxiliary information. It is possible to improve the results by using a Kalman filter type of approach, but that requires that the auxiliary

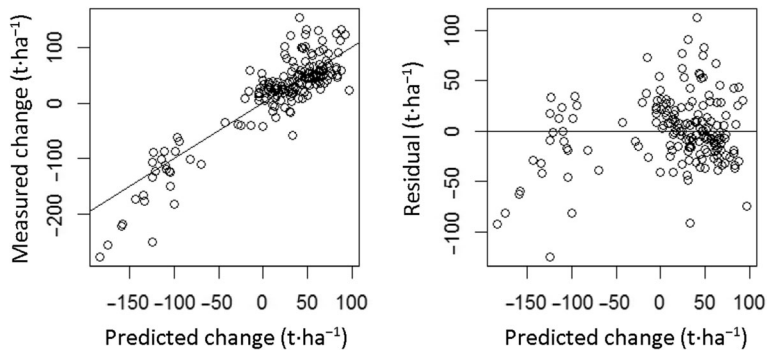

Fig. 4 The predicted versus observed change from 1999 to 2010 (left) and the residual plot (right)

information obtainable from remote sensing is also utilized efficiently in the analysis.

It is clear from the results that a composite estimation using old and current data is not a feasible approach when the time interval between the acquisitions of these two data sets is as large as in the current study (11 years). It is possible that a composite estimator without the updating would be useful, if the interval was markedly shorter, and if the plots influenced by harvest could be correctly detected. If permanent plots were available, a regression estimator proposed by Bickford et al. (1963) could also be used for updating, but in this study the two samples were assumed independent. Otherwise, it is clear that updating the data using a growth model would be highly advisable. However, even in this case the results are likely to be the more accurate the shorter the time interval, meaning smaller variance of the predicted change. In addition, it is unlikely that an external change model would be correctly specified for the target population in a real case. That would involve considerations as to how large a bias in the estimators would be acceptable. Depending on the use of the data, it may not be enough if the RMSEs of the estimates can be reduced when using old data as prior information: it is possible that in some applications even a small bias is unacceptable.

Typically, in Kalman filtering it has been assumed that the sample estimate is a random sampling estimate. However, there is nothing in the method that prevents using model-assisted or model-based estimator as the starting point, which is updated as in the Kalman filter. Then the resulting estimate can be combined with another model-assisted or model-based sampling estimate to obtain Kalman gains. If the applied growth model is linear, this is straightforward. If the growth model is non-linear, it has to be linearized with a Taylor series approximation (e.g. Ehlers et al. 2013) or by computing the average change as in Kangas (1991). It would also be possible to utilize stratification or post-stratification instead of model-assisted or model-based estimation, which might involve simpler estimators in case of non-linear change models.

It is often argued that comparing model-assisted and model-based approaches is not useful, as the underlying assumptions are very different, thereby causing different interpretations of uncertainty. However, here we compared the simulated (true) estimates of variance and RMSE, describing how well these approaches can estimate

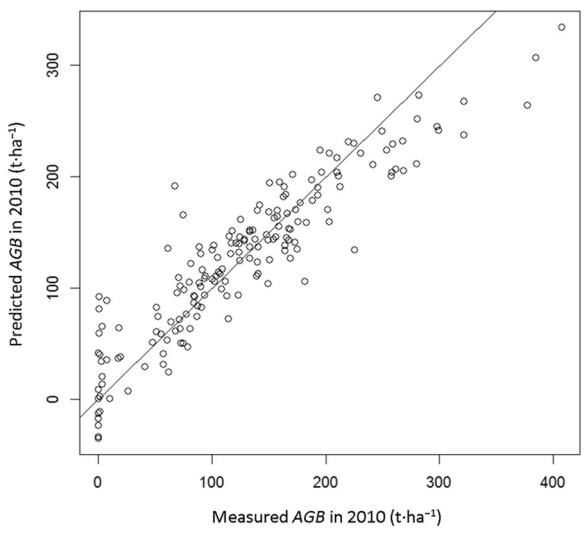

Fig. 5 The predicted *AGB* in 2010, based on *AGB* in 1999 and the change estimate

Table 4 The coefficients for a change model using ALS metrics from 1999 and 2010

	Estimate	Std. Error	*t* value
Intercept	−84.02443	8.182435	−10.26888
AGB_{1999}	−0.4180629	0.0610819	−6.844301
p_{20_2010}	7.467099	1.109191	6.732025
d_{2_2010}	−34.96214	16.82324	−2.078205
d_{2_1999}	159.6528	12.93541	12.3423

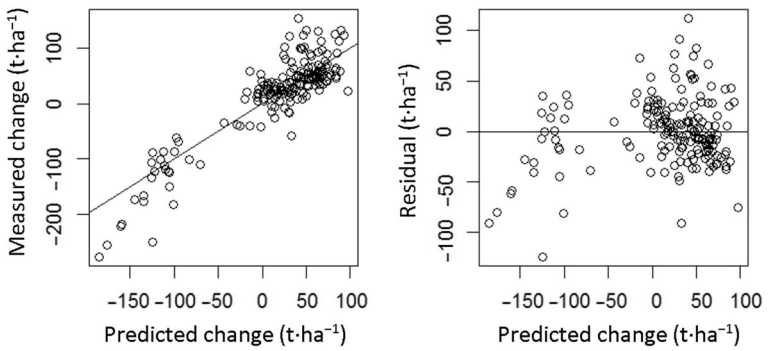

Fig. 6 The predicted versus observed change of aboveground biomass (*AGB*) from 1999 to 2010 (left) and the residual plot (right) using 1999 and 2010 ALS metrics

the variable of interest, the mean of *AGB* in 2010. Irrespective of the interpretation of the uncertainty, this is the important issue for the users of the data. Moreover, when the old inventory data are updated with a growth model, the end-result is a hybrid estimate involving the sampling error of the original estimate and the model-based prediction error (e.g. Melo et al. 2018). If the errors in the within-plot estimations (such as the allometric biomass models) were included, all the cases considered would be hybrid estimators (e.g. Ståhl et al. 2014, Corona et al. 2014, Fortin et al. 2016, Ståhl et al. 2016, Holm et al. 2017). In the case of hybrid estimators, the differences in the theoretical foundations of the design-based and model-based approach are ignored. In the current case, these two approaches produced also empirically very similar results on average.

In this study, we did not include the uncertainties of the allometric models. In this study, we assumed that these errors are negligible when compared to the errors in the change predictions. However, when the interval between the past and current inventories gets shorter, the relative importance of the uncertainties due to the allometric models will increase (see e.g. Chen et al. 2015, 2016). In addition, in this study we ignored the effect of spatial correlation. It can be assumed that the smaller the area over which the results are calculated, the larger is the effect of this spatial correlation (e.g. McRoberts et al. 2018). Both of these aspects need to be studied in the future. Furthermore, the possibility that the two estimators are not independent (like if part or all of the plots are permanent), also needs to be addressed in the future (Grafström et al. 2019).

The most problematic issue in the updating of the data for the Kalman filtering is the harvests. In this study, a fairly simple linear model was utilized. However, even with such a simple model it proved to be possible to also predict the harvests happening between 1999 and 2010 (i.e. the model also predicted negative changes). In this study, first a model (*C*1) based on the 1999 ALS metrics, and secondly a model (*C*2) based on both 1999 and 2010 ALS metrics were tested. If the purpose is to estimate past changes in order to update the old (1999) data to the current year (2010), it would be possible to utilize the second type of model. If the purpose would be to predict also the harvests happening after 2010 (e.g. between 2010 and time point *t*3, if ALS data from that

Table 5 The results from the simulation using the change model 1 for the Kalman filter. The mean of sample means of *AGB* (t·ha^{-1}) and sample variance estimates, the simulated variance, the bias and the RMSE

		Mean	Estimated std	Simulated std	Bias	RMSE
	True	128.93				
Current occasion	HT	128.87	8.72	8.60	0	8.60
	Model-assisted	128.90	5.31	5.24	0	5.24
	Model-based	128.39	5.12	5.13	−0.54	**5.16**
Composite of two occasions	HT	118.97	5.39	5.66	−9.96	11.45
	Model-assisted	119.45	3.28	3.80	−9.48	**10.22**
	Model-based	119.54	3.24	3.86	−9.59	10.34
Kalman filter	HT	128.42	6.23	5.86	−0.51	5.88
	Model-assisted	128.80	4.50	4.00	−0.13	**4.00**
	Model-based	128.32	4.37	3.98	−0.61	4.03

Table 6 The Kalman filter results from the simulation using the change model 2. The mean of sample means of AGB ($t{\cdot}ha^{-1}$) and sample variance estimates, the simulated variance, the bias and the RMSE

	Mean	Estimated std	Simulated std	Bias	RMSE
HT	128.53	5.66	6.16	−0.40	6.17
Model-assisted	128.85	4.23	3.73	−0.08	**3.73**
Model-based	128.37	4.12	3.72	−0.51	3.76

time point are unavailable), only the first type of model is applicable. In this case, the latter model produced more accurate change estimates, as the differences in the density metrics ($d2$) between the two time points were able to describe the harvests. This model consequently produced more accurate estimates for the current (2010) *AGB*. It means that all the information available for the updating should be included in the analysis.

The prediction of the harvests is also likely to increase the error variance of the change estimates, so that improved accuracy would be obtained if the harvests were directly observed from the differences between remote sensing materials and used as known control actions rather than predicted using a model as was done here. This is possible for clearcuts, which can be accurately delineated from differences between two satellite images (Pitkänen et al. 2020). Moreover, if the change model would reflect purely growth, it would be possible to utilize relative errors (as in Ehlers et al. 2013) rather than absolute errors. This is important, as the errors in predicted growth are often heteroskedastic. Then, relative error may reflect the true situation better. Such an approach was not feasible to adopt in the current study, as the model also predicted harvests, and part of the change estimates were negative.

In the Kalman filtering approach, the errors in the growth model are an important source of error. The simulated variances for both model-assisted and model-based with a composite model were smaller than those of the Kalman filter counterparts, as the Kalman filter variance estimates also include the error of the growth model. Optimal weight for the composite estimator would be obtained, if the bias (and therefore also RMSE) was known, but this is normally not the case. On the other hand, as the bias was mostly removed in the Kalman filter approach, the RMSEs of Kalman filter estimates were clearly smaller than those of the composite estimates. While using the variances provides optimal weights, they also can complicate estimation if the weights vary across the domains (Scott and Köhl 1994). Therefore, non-optimal weights based on simply the number of plots might be useful (Scott personal communication 2019).

In the current study, the Kalman filtering approach was used to estimate the mean *AGB* in the whole population. If this approach was to be used in a real NFI setting (Tomppo et al. 2010), it would require that all variables of interest can be updated to the date of the current inventory. This generally requires a growth and yield simulator with a tree-level growth models (e.g. Kangas 1991). If this requirement can be fulfilled, then the approach is applicable in NFI with the same premises than model-based estimation in general is applicable in NFI.

However, this kind of approach might be more useful for results calculated for smaller domains/categories, for instance for a small area or for a rare tree species. In that kind of situation, it might be useful to use plots measured from a longer period of years than normally in a case of continuous panel inventory. In NFIs, often a continuous panel inventory with 5-year interval is used, but using plots from a 10-year period would be possible. Plots measured during the 5-year interval might not be updated, but if a longer period is used, updating the data would be advisable. In some countries, the inventories are separate campaigns like in this study, and in such a case it might be useful to use data from two or more campaigns for the smaller domains/categories. The usefulness of old inventory data for small area estimation remains to be studied in the future.

In previous studies concerning the Kalman filter, utilization of prior information has mostly been tested in a setting where the interest has been in the individual plot or pixel level results (e.g. Ehlers et al. 2013). It is likely that in such a setting improving the accuracy using the old data is markedly more difficult than in the studied case. This is because in a sampling setting, the prior information can be interpreted as increasing the number of plots in the analysis, which improves the estimates for the population mean and total. In a pixel level analysis such interpretation cannot be made.

Conclusion

Prior information from old inventory data can be useful also when combined with highly accurate auxiliary information, when both data sources are efficiently used.

Acknowledgements

We wish to thank Dr. Charles T. Scott for helpful comments on the earlier phase of the manuscript.

Authors' contributions

Professor Kangas estimated the models used, produced the copula population and made the simulations. She was also mainly responsible for the writing of the paper. Professors Næsset and Gobakken provided the original field data, and took part in interpreting the results and writing the paper. The author(s) read and approved the final manuscript.

Funding
The analysis and interpretation of the results was partly funded by the Finnish Ministry of Agriculture and Forestry key project "Wood on the move and new products from forests" and partly by the Norwegian Forest Trust Fund (Skogtiltaksfondet) and the Forest Development Fund (Utviklingsfondet for skogbruket).

Availability of data and materials
The simulated population is available as. Rdata file from the corresponding author.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Bioeconomy and Environment, Natural Resources Institute Luke (Finland), Yliopistokatu 6, 80100 Joensuu, Finland. [2]Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway.

Received: 16 August 2019 Accepted: 23 March 2020
Published online: 07 April 2020

References

Aas K, Czado C, Frigessi A, Bakken H (2009) Pair-copula constructions of multiple dependence. Insur Math Econ 44:182–198
Bickford CA, Mayer CE, Ware KD (1963) An efficient sampling design for forest inventory: the northeastern forest resurvey. J For:826–833
Cassel CM, Särndal CE, Wretman JH (1977) Foundations on inference in survey sampling. Wiley, New York, p 192
Chen Q, Laurin GV, Valentini R (2015) Uncertainty of remotely sensed aboveground biomass over an Afrian tropical forest: propagating errors from trees to plots to pixels. Remote Sens Environ 160:134–143
Chen Q, McRoberts RE, Wang C, Radtke PJ (2016) Forest aboveground biomass mapping and estimation across multiple spatial scales using model-based inference. Remote Sens Environ 184:350–360
R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
Corona P, Fattorini L, Franceschi S, Scrinzi G, Torresan C (2014) Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: model-based, design-based, and hybrid perspectives. Can J For Res 44:1303–1311
Dixon B, Howitt R (1979) Continuous forest inventory using a linear filter. For Sci 25:675–698
Ehlers S, Grafström A, Nyström K, Olsson H, Ståhl G (2013) Data assimilation in stand-level forest inventories. Can J For Res 43:1104–1113
Fortin M, Manso R, Calama R (2016) Hybrid estimation based on mixed-effects models in forest inventories. Can J For Res 46:1310–1319
Grafström A, Ekström M, Jonsson BG, Esseen P-A, Ståhl G (2019) On combining independent probability samples. Surv Methodol 45:349–364
Gregoire TG, 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. Can J For Res 41:83–95
Holm S, Nelson R, Ståhl G (2017) Hybrid three-phased estimators for large-area forest inventory using ground plots, airborne lidar and space lidar. Remote Sens Environ 197:85–97
Kangas A (1991) Updated measurement data as prior information in forest inventory. Silva Fenn 25:181–191
Kangas A (2006) Model-based inference. In: Kangas a, Maltamo M (eds) Forest inventory, methods and applications. Managing Forest ecosystems Vol. 10. Springer, Dordrect, p 362
Kangas A, Myllymäki M, Gobakken T, Næsset E (2016) Model-assisted forest inventory with parametric, semi-parametric and non-parametric models. Can J For Res 46:855–868
Kooperberg C (2015) logspline: Logspline Density Estimation. Routines. R package version 2.1.8. http://CRAN.R-project.org/package=logspline.
Marklund LG (1988) Biomassafunktioner för tall, gran och björk i Sverige, Sveriges lantbruksuniversistet, Institutionen för skogstaxering, rapport 45, p 73 ISSN 0348-0496
McRoberts R, Liknes GC, Domke GM (2014) Using a remote sensing-based, percent tree cover map to enhance forest inventory estimation. Forest Ecol Manag 331:12–18
McRoberts RE, Næsset E, Gobakken T, Chirici G, Condés S, Hou Z, Saarela S, Qi C, Ståhl S, Walters BF (2018) Assessing components of model-based mean square error estimator for remote sensing assisted forest applications. Can J For Res 48:642–649
Meier P (1953) Variance of a weighted mean. Biometrics 9:59–73
Melo L, Schneider R, Fortin M (2018) Estimating model-and sampling-related uncertainty in large-area growth predictions. Ecol Model 390:62–69
Myllymäki M, Gobakken T, Næsset E, Kangas A. (2017) The efficiency of post-stratification compared to model-assisted estimation. Can J For Res 47: 515–526.
Næsset E (2002) Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sens Environ 80:88–99
Næsset E, Bollandsås OM, Gobakken T, Gregoire TG, 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 Sens Environ 128:299–314
Næsset E, Bollandsås OM, Gobakken T, Solberg S, McRoberts RE (2015) The effects of field plot size on model-assisted estimation of aboveground biomass change using multitemporal interferometric SAR and airborne laser scanning data. Remote Sens Environ 168:252–264
Nyström M, Lindgren N, Wallerman J, Grafström A, Muszta A, Nyström K, Bohlin J, Willén E, Fransson JES, Ehlers S, Olsson H, Ståhl G (2015) Data assimilation in forest inventory: first empirical results. Forests 6:4540–4557
Pitkänen T, Sirro L, Häme L, Häme T, Törmä M, Kangas A (2020) Errors related to the automatized satellite-based change detection of boreal forests in Finland. Int J Appl Earth Obs Geoinform 86:102011
Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer-Verlag, New York, p 694
Schepsmeier U, Stoeber J, Brechmann EC, Graeler B (2015) VineCopula: statistical inference of vine copulas. R package version 1.6. http://CRAN.R-project.org/package=VineCopula
Scott C (1984) A new look at sampling with partial replacement. For Sci 30(1): 157–166
Scott C, Köhl M (1994) Sampling with partial replacement and stratification. For Sci 40(1):30–46
Shahar D (2017) Minimizing the variance of a weighted average. Open J Stat 7: 216–224
Ståhl G, Heikkinen J, Petersson H, Repola J, Holm S (2014) Sample-based estimation of greenhouse gas emissions from forests – a new approach to account for both sampling and model errors. For Sci 60(1):3–13
Ståhl G, Holm S, Gregoire TG, Gobakken T, Næsset E, Nelson R (2011) Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway. Can J For Res 41:96–107
Ståhl G, Saarela S, Schnell S, Holm S, Breidenbach J, Healey SP, Patterson PL, Magnussen S, Næsset E, McRoberts RE, Gregoire TG (2016) Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. Forest Ecosyst 3:5. https://doi.org/10.1186/s40663-016-0064-9
Tomppo E, Gschwantner T, Lawrence M, McRoberts RE (2010) National Forest Inventories - pathways for common reporting. Springer, Dordrech
Tomppo E, Halme M (2004) Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach. Remote Sens Environ 92:1–20
Ware K, Cunia T (1962) Continuous forest inventory with partial replacement of samples. Forest Sci Monograph 3:40