Forest Ecology and Management 401 (2017) 295–303

Contents lists available at ScienceDirect

Forest Ecology and Management

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

Multivariate inference for forest inventories using auxiliary airborne laser scanning data

Ronald E. McRoberts [a,*], Qi Chen [b], Brian F. Walters [a]

[a] Northern Research Station, U.S. Forest Service, Saint Paul, MN, USA
[b] Department of Geography, University of Hawai'i at Mānoa, Honolulu, HI, USA

ARTICLE INFO

Article history:
Received 12 April 2017
Received in revised form 8 July 2017
Accepted 10 July 2017
Available online 22 July 2017

Keywords:
k-Nearest Neighbors
Stratified estimator
Model-assisted regression estimator

ABSTRACT

National forest inventories have a long history of using remotely sensed auxiliary information to enhance estimation of forest parameters. For this purpose, aerial photography and satellite spectral data have been shown to be effective as sources of information in support of stratified estimators. These spectral-based stratifications are much more effective for reducing variances for forest area-related parameters than for parameters related to continuous attributes such as volume and biomass. For variables related to the latter attributes, stratified estimators using airborne laser scanning auxiliary data are much more effective, but are less effective than model-assisted estimators using the same auxiliary data. For inventory applications, however, stratified estimators using the same stratification for all response variables are naturally multivariate, whereas model-assisted estimators are not. A consequence is that multiple, univariate applications of model-assisted estimators cannot ensure compatibility among estimates of inventory parameters related to variables such as forest area, growing stock volume, and tree density.

The objectives of the study were twofold: (1) to optimize a multivariate, k-NN approach for simultaneously predicting multiple forest inventory variables; and (2) to compare multivariate model-assisted generalized regression estimators using optimized k-NN predictions to post-stratified estimators with respect to inferences in the form of confidence intervals for multiple forest inventory parameters. The analyses included use of airborne laser scanning data as auxiliary information and the multivariate k-NN technique for prediction in support of the model-assisted estimators. The study area was in north central Minnesota in the USA and is characterized by both lowland and upland forest types interspersed with wetlands and lakes.

The first primary result was that the optimized k-NN technique in combination with a model-assisted estimator produced compatible multivariate estimates of population means for six inventory parameters. Second, variances for the multivariate model-assisted estimators were smaller by 23%–35% than variances for a post-stratified estimator. These results warrant serious consideration of this approach for operational implementation by national forest inventories.

Published by Elsevier B.V.

1. Introduction

National forest inventories (NFI) in the Nordic countries and the United States of America (USA) have a long history of using remotely sensed auxiliary information to enhance inferences in the form of confidence intervals for forest inventory parameters. Aerial photography served as the earliest source of remotely sensed information for constructing stratifications for this purpose. Bickford (1953, 1960) in the USA and Poso (1972) in Finland constructed strata based on interpreted aerial photography to support stratified estimation. More recently, satellite imagery has served as the source of information for constructing stratifications. With this approach, the imagery is classified with respect to a forest attribute of interest, and the classes, or aggregations of the classes, serve as strata (Poso et al., 1984, 1987; Hansen and Wendt, 2000; McRoberts et al., 2002, 2006). McRoberts et al. (2012) provide more details on the history of using aerial photography and satellite imagery to support stratified estimators for forest inventory applications. Although stratifications based on aerial photography and satellite imagery have been shown to be effective for increasing the precision of estimators of forest area, their effectiveness for attributes such as growing stock volume and biomass is considerably less (McRoberts et al., 2006).

* Corresponding author.
E-mail address: rmcroberts@fs.fed.us (R.E. McRoberts).

http://dx.doi.org/10.1016/j.foreco.2017.07.017
0378-1127/Published by Elsevier B.V.

The advent of airborne laser scanning (ALS) data has introduced new possibilities for using remotely sensed auxiliary information to increase the precision of estimators of parameters related to forest volume and biomass. Næsset (2002) reported that 80–93% of the variability in field measured forest volume could be explained by models that use ALS metrics, and Næsset and Gobakken (2008) reported that 88% of the variability in aboveground biomass could be explained with models using similar metrics. These results have been confirmed in multiple additional studies (Li et al., 2008; Zhao et al., 2009; Frazer et al., 2011). McRoberts et al. (2012, 2013) demonstrated that ALS-based stratifications increase precision for estimators of growing stock volume comparable to the increases satellite image-based stratifications produce for forest area. However, for continuous forest attributes such as growing stock volume, model-assisted estimators using ALS data increase precision by even more than stratified estimators (McRoberts et al., 2013). With model-assisted estimators, an initial estimate based on model predictions for all population units is adjusted using differences between sample unit observations and predictions to compensate for systematic prediction error.

For operational purposes, NFIs require compatibility among estimates of parameters for different attributes. For example, for a particular estimation unit, a small estimate of forest area should not accompany a large estimate of growing stock volume. Such problems do not arise with stratified estimators using the same stratification because the stratifications only provide weights which are applied equally to observations for all response variables. Model-assisted estimators, on the other hand, require predictions for all response variables for all population units, and if a multivariate prediction approach is not used, then inevitably incompatible predictions such as large growing stock volume for a population unit that is predicted to have no forest cover will occur.

Multivariate regression methods typically require multivariate normally distributed response variables, a condition that is seldom satisfied for forest inventory variables. An alternative that has become very popular for use with remotely sensed data for forest inventory applications is the multivariate, non-parametric k-Nearest Neighbors (k-NN) technique (Chirici et al., 2016). Among the reported multivariate applications of k-NN, Temesgen et al. (2003) and LeMay and Temesgen (2005) predicted basal area and tree density using variables that included crown closure, height, age and ecological zone. McRoberts et al. (2007) and McRoberts (2009) predicted basal area, tree density and volume using Landsat metrics and used model-based inference to estimate small area means and their standard errors. Nothdurft et al. (2009) and Breidenbach et al. (2010) predicted total and three species-specific timber volumes for stands using ALS and optical data. Dash et al. (2015) predicted basal area, tree density, volume, and height using lidar metrics and estimated stand-level means and standard errors using the same approach to model-based inference. These studies established the utility of k-NN for multivariate prediction and for small area, model-based inference. However, none of these studies focused on larger areas on the order of inventory reporting units that are amenable to probability-based (design-based) inferential methods.

The objectives of the study were twofold: (1) to optimize a multivariate, k-NN approach for simultaneously predicting multiple forest inventory variables; and (2) to compare multivariate model-assisted generalized regression (GREG) estimators using optimized k-NN predictions to post-stratified (STR) estimators with respect to inferences in the form of confidence intervals for multiple forest inventory parameters. For both the stratified and model-assisted estimators, the auxiliary information was in the form of metrics derived from ALS data. The study area was in north central Minnesota in the USA and is characterized by both lowland and upland forest areas interspersed with wetlands and lakes.

2. Data

2.1. Study area

The 7583-km^2 study area consisted of the entirety of Itasca County in north central Minnesota in the USA (Fig. 1). Land cover includes water, wetlands and forest consisting of uplands with deciduous mixtures of pines (*Pinus* spp.), spruce (*Picea* spp.), and balsam fir (*Abies balsamea* (L.) Mill.) and lowlands with spruce (*Picea* spp.), tamarack (*Larix laricina* (Du Roi) K. Koch), white cedar (*Thuja occidentalis* (L.)), and black ash (*Fraxinus nigra* Marsh.).

2.2. Forest inventory data

Data were obtained for plots established by the Forest Inventory and Analysis (FIA) program of the U.S. Forest Service which conducts the NFI of the USA. The FIA program has established field plot centers in permanent locations using a systematic unaligned sampling design that is regarded as producing an equal probability sample (McRoberts et al., 2010). The entire array of plots for Minnesota is subdivided into five systematic interpenetrating panels, and one panel is selected on a rotating basis for measurement each year. Each FIA plot consists of four 7.32-m (24-ft) radius circular subplots that are configured as a central subplot and three peripheral subplots with centers located at 36.58 m (120 ft) and azimuths of 0°, 120°, and 240° from the center of the central subplot. Field crews visually estimate the proportion of each subplot that satisfies the FIA definition of forest land: (i) minimum area 0.4 ha (1.0 ac), (ii) minimum tree cover of 10%, (iii) minimum width of 36.58 m (120 ft), and (iv) forest land use. For plots on forest land, field crews also observe species and measure diameter at breast-height (dbh, 1.37 m, 4.5 ft) and height for all trees with dbh of at least 12.7 cm (5 in.) on each subplot. Allometric model predictions of individual tree stem volumes are aggregated at subplot-level. For this study, uncertainty associated with the allometric model predictions was ignored. Species-level specific gravities are used to convert tree volumes to aboveground live tree stem biomass. Subplot-level response variables for this study included proportion forest area (A), basal area (BA, m^2/ha), growing stock volume (V, m^3/ha), aboveground live tree stem biomass (AGB, Mg/ha), tree density (D, stems/ha), and mean live tree height (HT, m).

Data were used for only the central subplots of the 242 plots measured in 2014 and 2015, because these were the only subplots and years for which plot coordinates were obtained using survey grade GPS receivers with sub-meter accuracy. For further reference, use of the term *plot* refers to the central subplot.

2.3. Airborne laser scanning data

Wall-to-wall ALS data were acquired in April 2012 with a nominal pulse density of 0.67 pulses/m^2. Ground returns were classified by the provider and were used to construct a digital terrain model via interpolation using the Tiffs (Toolbox for Lidar Data Filtering and Forest Studies) software (Chen, 2007). For this study that uses relatively small plots and ALS data characterized by small pulse densities, all pulse returns were used.

Distributions of all pulse return heights were constructed for the 168.3-m^2 plots and for the 169-m^2 square cells that tessellated the study area and served as population units. ALS metrics for each plot and cell included the mean (h_{mn}), standard deviation (h_{sd}), skewness (h_{sk}), kurtosis (h_{ku}), and quadratic mean height (h_{qm}) of the distributions of heights for all pulse returns (Lefsky et al., 1999; Chen et al., 2012). In addition, heights corresponding to the 10th, 20th, ..., 100th percentiles (h_{10}, h_{20}, ..., h_{100}) of the distributions were calculated as were canopy densities expressed as

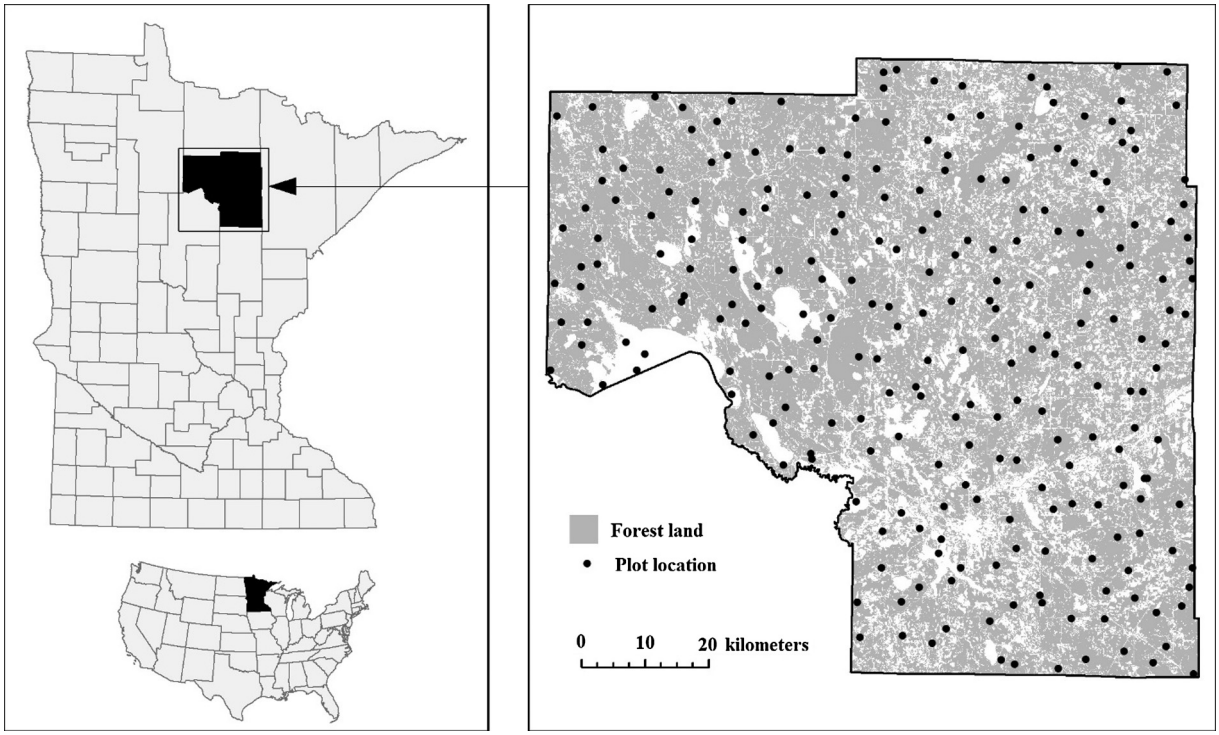

Fig. 1. Study area: Itasca County, Minnesota, USA.

the proportions of pulse returns with heights greater than 10%, . . ., 90%, 95% (cd_{10}, . . ., cd_{90}, cd_{95}) of the range between a minimum ALS above ground height threshold and the 95th height percentile (Gobakken and Næsset, 2008).

2.4. Outliers

Partly as a result of the FIA program's 5-year remeasurement cycle, two factors affected the utility of the observations for some FIA plots. First, because the ALS data were acquired in 2012 but the plots were not measured until 2014 and 2015, some plots were harvested or otherwise substantially disturbed between the two dates. To alleviate this discrepancy, plots were deleted from further analyses if they simultaneously satisfied three criteria: (i) 2009 or 2010 AGB greater than the 20th percentile of distribution of observed 2014/2015 AGB for forest plots, (ii) 2014 or 2015 AGB = 0, and (iii) h_{qm} greater than the 20th percentile of the distribution of observed h_{qm} for plots satisfying the first two criteria. Together the first two criteria indicate plot-level disturbance such as harvest, and given satisfaction of the first two criteria, the third criterion indicates the disturbance was after the ALS acquisition but before the plot measurement. Second, the FIA program classifies plots with respect to forest use, not forest cover. Therefore, plots classified as non-forest use but with measurable trees (e.g., orchards, parkland, residential property) would not be measured in the field and would have all tree-based attributes arbitrarily set to 0. To alleviate this discrepancy, plots were deleted from further analyses if they simultaneously satisfied two criteria: (i) classified as non-forest land use in 2014 or 2015, and (ii) h_{qm} greater than the 20th percentile of the distribution of observed h_{qm}. Together these two criteria indicate non-forest use but with measurable trees. Selection of the 20th percentile is arbitrary, albeit conservative because it leads to fewer deletions than smaller percentiles. With this approach, observations for 19 of the 242 plots were deleted and were considered to be missing at random (Rubin, 1976).

3. Methods

3.1. Optimization criterion

For forest inventory purposes, the ultimate analytical objective is a statistical inference in the form of a confidence interval for the population mean or total for each response variable. However, because observations of the multiple response variables are positively correlated as a result of being based on the same underlying tree data, estimates of the means will also be positively correlated. When only a single response variable is of interest, the inferential criterion would be minimization of the confidence interval width, but when multiple response variables are simultaneously of interest, the analogous criterion is minimization of the hypervolume (multi-dimensional volume) of the joint confidence region (Fig. 2). Because this hypervolume is proportional to the square root of the determinant of the estimated parameter covariance matrix, minimization of the square root of the determinant was the technical optimization criterion for this study. Optimization details are provided for specific estimators in the sections that follow. However, regardless of the optimization approach, confidence intervals for the means of individual response variables are still reported and are still expressed as $\hat{\mu}^r \pm t \cdot SE(\hat{\mu}^r)$ where r denotes the response variable, $\hat{\mu}^r$ is the estimate of the mean, and $SE(\hat{\mu}^r) = \sqrt{\hat{Var}(\hat{\mu}^r)}$.

3.2. Simple random sampling estimators

With equal probability sampling designs, the simplest approach to inference is to use the familiar *simple random sampling (SRS)* estimators for means and their variances,

$$\hat{\mu}_{SRS} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1a}$$

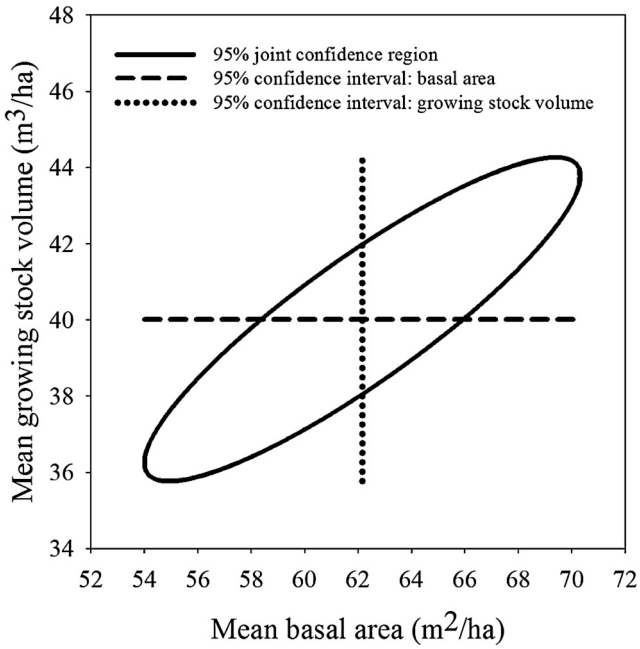

Fig. 2. Confidence intervals and joint confidence region illustrating that the confidence intervals do not necessarily reflect the shape or hypervolume of the confidence region.

and

$$\hat{\mathrm{Var}}(\hat{\mu}_{SRS}) = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_{SRS})^2}{n(n-1)}, \tag{1b}$$

where i indexes the n sample units, and y_i is the observation for the ith sample unit. The primary advantages of the SRS estimators are that they are intuitive, simple, and unbiased when used with an SRS design; the disadvantage is that variances are frequently large, particularly for small sample sizes and/or populations with large variability among population unit observations. Although $\hat{\mathrm{Var}}(\hat{\mu}_{SRS})$ from Eq. (1b) may be biased when used with systematic sampling, it is usually conservative in the sense that it overestimates the variance (Särndal et al., 1992). For this study, finite population correction factors were ignored because of the small sampling intensity of approximately one plot per approximately 183,000 population units.

3.3. Stratified estimators

Stratified estimates of means are calculated using the estimator provided by Cochran (1977) as,

$$\hat{\mu}_{STR} = \sum_{h=1}^{H} w_h \hat{\mu}_h \tag{2a}$$

where

$$\hat{\mu}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi},$$

h = 1,..., H denotes strata; y_{hi} is the ith sample observation for the hth stratum; w_h is the weight for the hth stratum calculated as the proportion of population units assigned to the stratum; n_h is the number of plots assigned to the hth stratum; $\hat{\mu}_h$ and $\hat{\sigma}_h^2$ are the sample estimates of the within-stratum mean and variance, respectively; and *STR* denotes the stratified estimators.

NFIs often use permanent plots whose locations are based on systematic grids or tessellations and use sampling intensities that are constant over large geographic areas. In such cases, even though stratified sampling is not possible, estimator precision can still be increased by using stratified estimation subsequent to the sampling, a technique characterized as post-sampling stratification or simply *post-stratification*. Cochran (1977, p. 135) provides a stratified estimator of the variance for use with post-stratification and the resulting random within-strata sample sizes,

$$\hat{\mathrm{Var}}(\hat{\mu}_{STR}) = \sum_{h=1}^{H} w_h \frac{n_h}{n}\frac{\hat{\sigma}_h^2}{n_h} + \frac{1}{n}\sum_{h=1}^{H}(1 - w_h)\frac{n_h}{n}\frac{\hat{\sigma}_h^2}{n_h}, \tag{2b}$$

where n is the total sample size over all strata and

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \hat{\mu}_h)^2.$$

Stratifications that are most effective with respect to minimizing variances are based on variables that are closely related to the response variable or variables of interest. When multiple response variables are to be estimated simultaneously, the same stratification must be used for all response variables to ensure compatibility. For this study, BA was considered as an integrator of all the response variables and was used as the basis for constructing strata. To facilitate stratification, a model of the relationship between plot-level BA (m^2/ha) and the ALS metrics was constructed as,

$$y_i = (\beta_0 \cdot x_{1i}^{\beta_1}) \cdot e^{\beta_2 x_{2i} + \cdots + \beta_p x_{pi}} + \varepsilon_i, \tag{3}$$

where i indexes plots, p is the number of predictor variables, y_i is plot-level BA, the x_i are ALS metrics, the βs are parameters to be estimated, and ε_i is a random residual assumed to follow a normal distribution, $N(0, \sigma_i^2)$. The form of the model is based on previous experiences whereby a single ALS metric, often h_{mn} or h_{qm}, accounts for half or more of the variation in the response variable. Thus, the initial power component of the model incorporates this single ALS metric and the exponential component of the model includes additional metrics that incrementally increase or decrease predictions. Quality of fit of the model to the data was assessed using pseudo-R^2,

$$R^{2*} = \frac{SS_{mn} - SS_{res}}{SS_{mn}}, \tag{4}$$

where SS_{mn} is the sum of squared differences between observations and their mean and SS_{res} is the sum of squared differences between observations and their corresponding model predictions.

The model prediction of BA for each population unit was divided by the greatest model prediction for all population units and multiplied by 100, thereby scaling all predictions to the [0,100] interval. Each population unit was then assigned to one of the 101 standardized classes [0,0], (0,1], ..., (99,100], and the proportion of units assigned to each class was calculated. Stratifications were constructed by aggregating adjacent classes into $n_{str} = 4$, $n_{str} = 6$, and $n_{str} = 8$ larger strata with each representing an approximately equal proportion of the study area. The rationale for selecting $n_{str} = 4$ as the minimum number of strata was that this is the number used by the FIA program for the study area; the rational for selecting $n_{str} = 8$ as the maximum number of strata was based on Cochran's (1977, p. 134) recommendation that more than 6–8 strata are not likely to produce additional gain in precision beyond smaller numbers of strata. All stratifications were subject to two constraints: (i) only adjacent classes could be aggregated into a larger stratum, and (ii) a minimum of 10 plots per stratum was required (Särndal et al., 1992, p. 267, 407; Cochran, 1977, p. 134; Westfall et al., 2011). Breidt and Opsomer (2008) coined the term *endogenous post-stratification* to describe this approach to post-stratification and established that for monotonic, parametric models such as Eq. (3), the adverse consequences

of using the same data to construct strata as are used as the response variable are minimal.

3.4. Model-assisted, generalized regression estimators

A *synthetic* estimator of the population mean for the rth response variable is,

$$\hat{\mu}^r_{Syn} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}^r_i, \tag{5a}$$

where N is the population size and $\hat{y}^r_i$ is the model prediction for the ith population unit. Hansen et al. (1983) note that models that do not "represent the state of nature" induce bias into this estimator which, for equal probability samples, can be estimated as,

$$\hat{\text{Bias}}(\hat{\mu}^r_i) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon^r_i, \tag{5b}$$

where $\varepsilon^r_i = \hat{y}^r_i - y^r_i$. The *model-assisted, generalized regression (GREG)* estimator is then,

$$\hat{\mu}^r_{GREG} = \hat{\mu}^r_{syn} - \hat{\text{Bias}}(\hat{\mu}^r_{Syn})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\hat{y}^r_i - \frac{1}{n}\sum_{i=1}^{n}\varepsilon^r_i \tag{5c}$$

(Särndal et al., 1992; Särndal, 2011). For $\bar{\varepsilon}^r = \frac{1}{n}\sum_{i=1}^{n}\varepsilon^r_i$, the estimator of the covariances for the estimates of the means from Eq. (5c) is,

$$\hat{\text{Cov}}(\hat{\mu}^{r_1}, \hat{\mu}^{r_2}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\varepsilon^{r_1}_i - \bar{\varepsilon}^{r_1}) \cdot (\varepsilon^{r_2}_i - \bar{\varepsilon}^{r_2}), \tag{5d}$$

and is a straightforward multivariate extension of the univariate estimator for a single variable as provided by Särndal et al. (1992). Optimization for the GREG estimators consisted of minimizing the square root of determinant of the covariance matrix from Eq. (5d).

3.5. Nearest neighbors techniques

For notational purposes, **Y** denotes a possibly multivariate vector of response variables observed for a sample, and **X** denotes a vector of auxiliary variables with observations for the entire population. In the terminology of nearest neighbors techniques, the auxiliary variables are designated *feature variables*; the space defined by the feature variables is designated the *feature space*; the sample of population units for which observations of both response and feature variables are available is designated the *reference set* with size denoted n; and the set of population units for which predictions of response variables are desired is designated the *target set* with size denoted N.

For a continuous response variable, r, the nearest neighbors prediction, $\hat{y}^r_i$, for the ith target unit is calculated as,

$$\hat{y}^r_i = \sum_{j=1}^{k} w_{ij} y^r_{ij}, \tag{6}$$

where $\left\{y^r_{ij}, j = 1, 2, \ldots, k\right\}$ is the set of response variable observations for the k reference units that are most similar or nearest to the ith target unit in feature space with respect to a distance metric, d, and w_{ij} is the weight assigned to the jth nearest neighbor with $\sum_{j=1}^{k} w_{ij} = 1$.

Implementation of the k-NN technique requires selection of a distance metric, the particular feature variables, the number, k, of nearest neighbors, and a method for weighting the neighbors. McRoberts et al. (2016, 2017) compared multiple combinations of these factors and concluded that the particular selections are not as important as optimization of the selections. Therefore, for ease of optimization, the *canonical correlation analysis distance metric* (CCA) and the *Dudani* neighbor weighting options were selected.

With the CCA metric, first proposed by Moeur and Stage (1995), a system of linear equations is solved to obtain estimates of coefficient vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, that maximize the correlation between $U = \alpha_1 \cdot Y^1 + \cdots + \alpha_p \cdot Y^p$ and $V = \beta_1 \cdot X_1 + \cdots + \beta_q \cdot X_q$ where Y^r denotes the rth response variable, X_j denotes the jth feature variable, and p and q are the numbers of response and feature variables, respectively. The solutions are obtained using canonical decompositions for which the eigenvectors, also designated canonical correlation coefficients, are denoted Γ, and the corresponding eigenvalues, also designated canonical correlations, are denoted λ. Feature space distances are then expressed as,

$$d_{ij} = \sqrt{(X_i - X_j)'\Gamma\Lambda^2\Gamma'(X_i - X_j)}, \tag{7}$$

where the elements of the diagonal matrix, Λ, are the squares, λ^2, of the canonical correlations.

Dudani (1976) proposed a weighting scheme which, slightly modified, bases the weight for the jth neighbor on the ratio of the distance between the jth and k^{th} neighbors and the distance between the first and the kth neighbors,

$$w_{ij} = \frac{d_{i,k+1} - d_{i,1}}{W}, \tag{8}$$

where $W = \sum_{j=1}^{k}\frac{d_{i,k+1}-d_{i,j}}{d_{i,k+1}-d_{i,1}}$. McRoberts et al. (2016, 2017) discusses this weighting scheme in detail, illustrates its implementation, and notes a slight modification that was used for this study.

With nearest neighbors techniques, unlike with regression models, prediction accuracy is adversely affected by feature variables unrelated to the response variable. These unrelated feature variables, characterized as *irrelevant* variables by Langley and Iba (1993), introduce randomness into distance calculations, contribute to selection of spurious neighbors, and produce less accurate predictions. Although the CCA distance metric theoretically weights feature variables relative to their importance, McRoberts et al. (2016, 2017) showed that prediction accuracy can be increased with this metric if irrelevant feature variables are first eliminated.

Optimization, therefore, consisted of elimination of irrelevant variables and use of the CCA metric and Dudani neighbor weighting together with the joint selection of the feature variables and the k-value that minimized the square root of the determinant of the joint covariance matrix for estimates of the population means for the six response variables. For each number of feature variables, all combinations were evaluated using a leave-one-out approach (Elisseeff and Pontil, 2002) with the reference data to select the particular combination and k-value that minimized the square root of the determinant.

4. Results and discussion

4.1. k-Nearest neighbors technique

The technical criterion for selection of k-NN feature variables, or equivalently identification and elimination of irrelevant features variables, was the square root of the determinant of the joint covariance matrix for the GREG estimates of the six response variable means. Minimization of this criterion was achieved for eight k-NN feature variables in the form of ALS metrics (Fig. 3). Thus, even though the CCA metric weights feature variables,

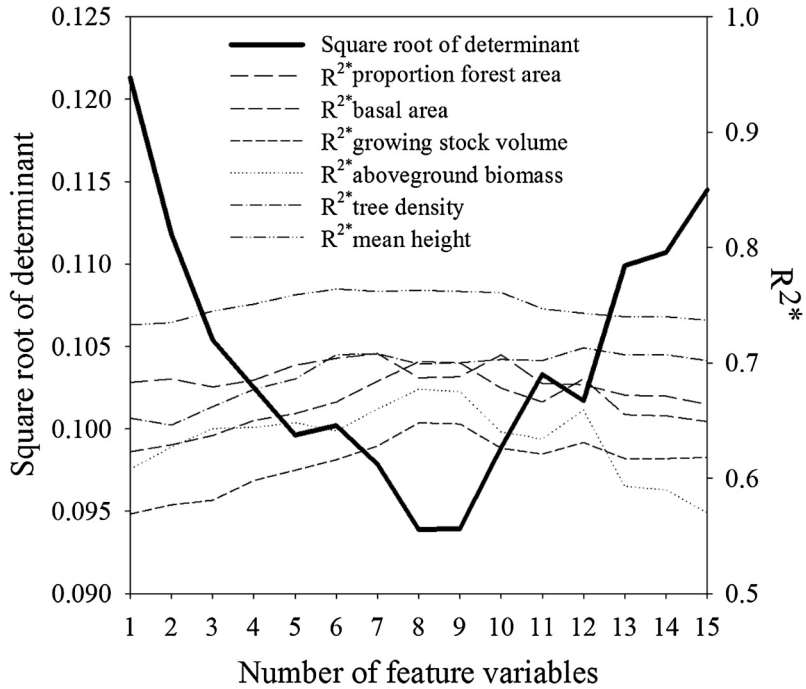

Fig. 3. Optimization of the k-Nearest Neighbors technique.

identification and elimination of irrelevant feature variables had beneficial effects. R^{2*} values for the individual response variables were also generally largest for approximately the same number of feature variables that minimized the determinant.

4.2. Stratified (STR) estimator

For the model of Eq. (3) used to predict BA, the initial power component of the model with h_{qm} as the predictor variable produced $R^{2*} = 0.668$ which was the greatest among all ALS metrics. Three additional mid-range height percentile metrics all statistically significantly improved the quality of the fit of the model to the data and produced $R^{2*} = 0.734$. Of importance, lack of fit of the model to the data does not induce bias into the stratified estimator but only reduces the degree to which estimator precision is increased. Relative to the SRS estimators that used no stratification, use of four strata decreased SEs by 29.0%–38.1%, depending on the response variable; relative to four strata, use of six strata reduced SEs by 2.3%–7.3%; and relative to six strata, use of eight strata reduced SEs by no more than 4% and for two response variables actually increased SEs (Table 1). Therefore, stratified estimates for the six response variables, based on six strata, were used for comparison to GREG estimates.

4.3. Generalized regression estimator (GREG)

Bias estimates obtained for the GREG estimator were small relative to estimates of the means, ranging in absolute value from 1.0% to 7.0% (Table 2). Further, the absolute values of all bias estimates were less than 1.5 SEs, suggesting lack of statistical significance. Subtraction of the bias estimate in the formulation of the GREG estimator of the mean is particularly useful for the k-NN predictor because of its tendency, particularly with large values of k, to under-predict large values and to over-predict small values (McRoberts, 2012, Section 3.4). Subject to probability sampling, subtraction of the bias makes the GREG estimator of the mean asymptotically unbiased, regardless of the prediction technique.

4.4. Multivariate versus univariate comparisons

GREG estimates for the six response variables obtained using the multivariate application of the k-NN technique were compared to GREG estimates for the six variables obtained using six independent univariate applications of the k-NN technique. Optimization for each response variable for the latter analyses entailed selecting the optimal combination of feature variables and k-value for use with the CCA metric and the Dudani neighbor weighting method.

Table 1
Stratified estimates.

Response variable[a]	Number of strata					
	4 strata		6 strata		8 strata	
	$\hat{\mu}_{STR}$	$SE(\hat{\mu}_{STR})$	$\hat{\mu}_{STR}$	$SE(\hat{\mu}_{STR})$	$\hat{\mu}_{STR}$	$SE(\hat{\mu}_{STR})$
A (proportion)	0.648	0.021	0.658	0.020	0.673	0.019
BA (m^2/ha)	10.639	0.621	10.495	0.596	11.014	0.596
V (m^3/ha)	66.579	4.773	65.707	4.663	69.154	4.578
AGB (Mg/ha)	41.164	2.590	40.392	2.414	42.817	2.458
D (stems/ha)	295.783	14.614	297.447	14.230	309.044	16.848
HT (m)	9.996	0.342	10.045	0.317	10.378	0.305

[a] A = proportion forest area, BA = basal area, V = growing stock volume, AGB = aboveground live tree stem biomass, D = tree density, HT = mean height.

Table 2
Comparing estimators.

Response variable[a]	Estimator						
	Simple random (SRS)		Stratified (STR)[b]		Generalized regression (GREG)		
	$\hat{\mu}_{SRS}$	$SE(\hat{\mu}_{SRS})$	$\hat{\mu}_{STR}$	$SE(\hat{\mu}_{STR})$	$\hat{Bias}(\hat{\mu}_{Syn})$	$\hat{\mu}_{GREG}$	$SE(\hat{\mu}_{GREG})$
A (proportion)	0.598	0.033	0.658	0.020	0.007	0.657	0.018
BA (m^2/ha)	10.174	0.941	10.495	0.596	−0.529	11.254	0.514
V (m^3/ha)	64.574	6.716	65.707	4.663	−4.576	70.268	3.974
AGB (Mg/ha)	39.517	3.888	40.392	2.414	−3.047	43.412	2.124
D (stems/ha)	277.952	21.875	297.447	14.230	3.097	314.243	12.436
HT (m)	9.231	0.560	10.045	0.317	−0.092	10.157	0.273

[a] A = proportion forest area, BA = basal area, V = growing stock volume, AGB = aboveground live tree stem biomass, D = tree density, HT = mean height
[b] Using six strata as per Table 1.

Table 3
Multivariate versus univariate estimates.

Response variable	R^{2*}	$\hat{Bias}(\hat{\mu}_{Syn})$	$\hat{\mu}_{GREG}$	$SE(\hat{\mu}_{GREG})$
Multivariate[a]				
A (proportion)	0.687	0.007	0.657	0.018
BA (m^2/ha)	0.702	−0.529	11.254	0.514
V (m^3/ha)	0.650	−4.576	70.268	3.974
AGB (Mg/ha)	0.702	−3.047	43.412	2.124
D (stems/ha)	0.677	3.097	314.243	12.436
HT (m)	0.763	−0.092	10.157	0.273
Univariate[b]				
A (proportion)	0.763	−0.002	0.681	0.016
BA (m^2/ha)	0.703	0.450	9.865	0.512
V (m^3/ha)	0.660	1.830	68.086	3.904
AGB (Mg/ha)	0.771	1.542	38.545	1.857
D (stems/ha)	0.677	−2.800	319.017	12.436
HT (m)	0.793	0.012	10.273	0.252

[a] Optimization: minimize confidence region hypervolume.
[b] Optimization: maximized individual variable R^{2*}.

As expected, univariate R^{2*} values were larger than multivariate R^{2*} values by as much as 11.1% (Table 3). One effect of the larger univariate R^{2*} values is that the univariate SEs were smaller than the multivariate SEs by as much as 12.6%. The univariate means ranged from 12.4% smaller to 3.7% greater than the multivariate means with the greatest percentage differences for the three diameter-based response variables, BA, V, and AGB. Because the response variables are strongly correlated and their means are reported simultaneously, multivariate inference is the preferred statistical approach. Consequences of failure to use the multivariate approach were that estimates of means differed by as much as 12.4%.

4.5. Inventory consequences

Use of the GREG estimators reduced SEs by 10.0%–13.8% relative to the stratified estimators (Table 2). The impact of these reductions can be expressed in terms of relative efficiency, calculated as,

$$RE = \frac{\hat{Var}(\hat{\mu}_{STR})}{\hat{Var}(\hat{\mu}_{GREG})}. \quad (9)$$

Because variances are inversely proportional to sample sizes, RE can be interpreted as the factor by which the sample size would have to be increased for estimator $\hat{\mu}_{STR}$ to achieve the same variance as was achieved with the estimator $\hat{\mu}_{GREG}$. For $\hat{\mu}_{STR}$ with six strata, REs for the six response variables were in the rather narrow range of 1.23–1.35. These results suggest that for these six response variables, use of the $\hat{\mu}_{GREG}$ estimators with ALS auxiliary data could possibly reduce required sample sizes by 23%–35% with no loss of precision relative to the $\hat{\mu}_{STR}$ estimators with six BA-based strata.

For the study area, the FIA program currently uses post-stratified estimators with four strata derived from a tree canopy cover product which, in turn, is partially based on Landsat imagery (Gormanson et al., 2017; Homer et al., 2015). A multivariate test of significance was used to compare the post-stratified estimates obtained using the FIA stratification and the GREG estimates obtained for this study. The Hoteling T^2 test indicated no statistically significant difference between the two sets of estimates at the $\alpha = 0.05$ significance level, although the GREG estimates were all larger than the post-stratified FIA estimates (Hotelling, 1931). The latter result is attributed to plots with trees on land with non-forest use for which tree-based attributes were incorporated into the GREG estimates with positive values but into the FIA estimates with values of 0. The individual FIA SEs were larger than the GREG SEs by 38%–57% with corresponding REs ranging from 1.91 to 2.48. The latter result suggests that if the FIA program were to use the GREG estimators with ALS auxiliary data, the sampling intensity could be reduced by approximately half with no loss of precision. Further, when considering that the FIA estimates were based on data for all four subplots of each plot, whereas the GREG estimates were based on data for only a single subplot for the same plots, the sampling intensity could be reduced even more. A 50% reduction in sampling intensity could be achieved by measuring only half as many plots each year, by measuring only two of the four subplots at each plot location, or by using a single plot with twice the area of the current central subplot at each plot location. With the latter option, the FIA plot size would be approximately 335 m^2 which would still be less than the much more commonly used 500-m^2 NFI basic sampling unit (Tomppo et al., 2010, NFI Reports). Further, the larger plot size would be more amenable for use with remotely sensed auxiliary data (Vauhkonen et al., 2014, p. 6).

5. Conclusions

Two primary conclusions were drawn from the study. First, optimization of the k-NN technique for multivariate prediction by minimizing the square root of the determinant of the covariance matrix of response variable means was generally satisfactory. Optimization of this criterion was only slightly more computationally difficult than optimization of a criterion such as root mean square error. In addition, for the six response variables considered, multivariate optimization produced little loss in prediction accuracy relative to individual univariate optimizations.

Second, the generalized regression estimators reduced standard errors relative to those obtained using the post-stratified estimators with six strata by 10.0%–13.5%. Further, increases in sample sizes for use with the post-stratified estimators necessary to achieve the same precision as with the generalized regression estimators were in the range of 23%–35%. For most national forest inventories, sample size increases of this magnitude are not financially feasible.

From an operational perspective, the results of the study suggest that multivariate, airborne laser scanning-assisted inventories could be fairly easily implemented, subject to availability of the airborne laser scanning data. If such data are not available, then consideration could be given to reducing the sampling intensity to offset the cost of acquiring the airborne laser scanning data. For the FIA program, reduction of the sampling intensity by as much as half would likely entail no loss of precision.

Acknowledgements

The authors thank Mr. Dale Gormanson, Forest Inventory and Analysis, Northern Research Station, U.S. Forest Service, St. Paul, MN USA, for assistance with calculation of the FIA post-stratified estimates.

References

Bickford, C.A., 1953. Increasing the efficiency of airphoto forest surveys by better definition of classes. Station Paper No. 58, Northeastern Forest Experiment Station, Forest Service, U.S. Department of Agriculture. Upper Darby, Pennsylvania, 9 pp.

Bickford, C.A., 1960. A test of continuous inventory for National Forest management based upon aerial photographs, double sampling, and remeasured plots. In: Anonymous (Ed.), Proceedings of the Society of American Foresters Meeting, November 15–19, 1959, San Francisco, California. Society of American Foresters, Washington, DC, pp. 143–148.

Breidenbach, J., Næsset, E., Vegard, L., Gobakken, T., Solberg, S., 2010. Prediction of species specific forest inventory attributes using a nonparametric semi-individual tree crown approach based on fused airborne laser scanning and multispectral data. Remote Sens. Environ. 114, 911–924.

Breidt, F.J., Opsomer, J.D., 2008. Endogenous post-stratification in surveys: Classifying with a sample-fitted model. Ann. Statistics 36 (1), 403–427.

Chen, Q., 2007. Airborne lidar data processing and information extraction. Photogrammetric Eng. Remote Sens. 73 (12), 1355–1365.

Chen, Q., Vaglio Laurin, G., Battles, J.J., Saah, D., 2012. Integration of airborne lidar and vegetation types derived from aerial photography for mapping aboveground live biomass. Remote Sens. Environ. 121, 108–117.

Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E.O., Waser, L., Travaglini, D., McRoberts, R.E., 2016. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. Remote Sens. Environ. 176, 282–294.

Cochran, W.G., 1977. Sampling Techniques. Wiley, New York.

Dash, J.P., Marshall, H.M., Rawley, B., 2015. Methods for estimating multivariate stand yields and errors using k-NN and aerial laser scanning. Forestry 88, 237–247.

Dudani, S.A., 1976. The distance-weighted k-Nearest-Neighbor rule. IEEE Trans. SMC 6 (4), 325–327.

Elisseeff, A., Pontil, M., 2002 Leave-one-out error and stability of learning algorithms with applications. In: Advances in learning theory: methods, models, and applications. NATO Advanced Study Institute on Learning Theory and Practice, pp. 111–1130.

Frazer, G.W., Magnussen, S., Wulder, M.A., Niemann, K.O., 2011. Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. Remote Sens. Environ. 115, 636–649.

Gobakken, T., Næsset, E., 2008. Assessing effects of laser point density, ground sampling intensity, and field plot sample size on biophysical stand properties derived from airborne laser scanner data. Can. J. Forest Res. 38, 1095–1109.

Gormanson, D.D., Pugh, S.A., Barnett, C.J., Miles, P.D., Morin, R.S., Sowers, P.A., Westfall, J., 2017. Statistics and quality assurance for the Northern Research Station Forest Inventory and Analysis Program, 2016. General Technical Report NRS-166. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station, 23 p. Available at: https://www.nrs.fs.fed.us/pubs/53730. (Last accessed: March 2017).

Hansen, M.H., Madow, W.G., Tepping, B.J., 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. J. Am. Stat. Assoc. 78 (384), 776–793.

Hansen, M.H., Wendt, D.G., 2000. Using classified landsat thematic mapper data for stratification in a statewide forest inventory. In: McRoberts, R.E., Reams, G.A., Van Deusen, P.C. (Eds.), Proceedings of the First Annual Forest Inventory and Analysis Symposium. 1999 November 2-3; San Antonio, Texas. General Technical Report NC-213. U.S. Department of Agriculture, Forest Service, North Central Research Station, St. Paul, MN, pp. 20–27. Available at: http://fia.fs.fed.us/symposium/proceedings/. Last accessed: March 2017.

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., Megown, K., 2015. Completion of the 2011 national land cover database for the conterminous United States – representing a decade of land cover change information. Photogrammetric Eng. Remote Sens. 81, 345–354.

Hotelling, H., 1931. The generalization of Student's ratio. Ann. Math. Stat. 2 (3), 360–378.

Langley, P., Iba, W., 1993. Average-case analysis of a nearest neighbor algorithm. In: Proceedings International Joint Conference on Artificial Intelligence-93. Chambery, France, pp. 889–894.

Lefsky, M.A., Harding, D., Cohen, W.B., Parker, G., Shugart, H.H., 1999. Surface lidar remote sensing of basal area and biomass in deciduous forests of eastern Maryland, USA. Remote Sens. Environ. 67, 83–98.

LeMay, V., Temesgen, H., 2005. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Sci. 51 (2), 109–119.

Li, Y., Andersen, H.-E., McGaughey, R., 2008. A comparison of statistical methods for estimating forest biomass from light detection and ranging data. Western J. Appl. Forestry 23 (4), 223–231.

McRoberts, R.E., 2009. A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. Remote Sens. Environ. 11, 532–545.

McRoberts, R.E., 2012. Estimating forest attribute parameters for small areas using nearest neighbors techniques. Forest Ecol. Manage. 272, 3–12.

McRoberts, R.E., Gobakken, T., Næsset, E., 2012. Post-stratified estimation of forest area and growing stock volume using lidar-based stratifications. Remote Sens. Environ. 125, 157–166.

McRoberts, R.E., Hansen, M.H., Smith, W.B., 2010. Development of the national forest inventory of the United States of America. In: Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E. (Eds.), National Forest Inventories: Pathways to Common Reporting. Springer, p. 612.

McRoberts, R.E., Næsset, E., Gobakken, T., 2013. Inference for lidar-assisted estimation of forest growing stock volume. Remote Sens. Environ. 128, 268–275.

McRoberts, R.E., Tomppo, O., Finley, A.O., Heikkinen, J., 2007. Estimating areal means and variances of forest attributes using the k-Nearest Neighbors technique and satellite imagery. Remote Sens. Environ. 111, 466–480.

McRoberts, R.E., Wendt, D.G., Nelson, M.D., Hansen, M.H., 2002. Using a land cover classification based on satellite imagery to improve the precision of forest inventory area estimates. Remote Sens. Environ. 81, 36–44.

McRoberts, R.E., Chen, Q., Domke, G.M., Næsset, E., Gobakken, T., Chirici, G., Mura, M., 2017. Optimizing nearest neighbour configurations for airborne laser scanning-assisted estimation of forest volume and biomass. Forestry 90, 99–111.

McRoberts, R.E., Domke, G.M., Chen, Q., Næsset, E., Gobakken, T., 2016. Using genetic algorithms to optimize k-Nearest Neighbors configurations for use with airborne laser scanning data. Remote Sens. Environ. 184, 387–395.

McRoberts, R.E., Holden, G.R., Nelson, M.D., Liknes, G.C., Gormanson, D.D., 2006. Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis programof the USDA Forest Service. Can. J. Forest Res. 36, 2968–2980.

Moeur, M., Stage, A.R., 1995. Most similar neighbor — an improved sampling inference procedure for natural resource planning. Forest Sci. 41 (2), 337–359.

Nothdurft, A., Saborowski, J., Breidenbach, J., 2009. Spatial prediction of forest stand variables. Eur. J. Forest Res. 128, 241–251.

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., Gobakken, T., 2008. Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. Remote Sens. Environ. 112, 3079–3090.

Poso, S., 1972. A method of combining photo and field samples in forest inventory. Communicationes Instituti Forestalis Fenniae 76.

Poso, S., Paananen, R., Simila, M., 1987. Forest inventory by compartments using satellite imagery. Silva Fennica 21, 69–94.

Poso, S., Hame, T., Paananen, R., 1984. A method of estimating the stand characteristics of a forest component using satellite imagery. Silva Fennica 18, 261–292.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.
Särndal, C.-E., 2011. Combined inference in survey sampling. Pakistan J. Stat. 27 (4), 359–370.
Särndal, C.-E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer, New York.
Temesgen, H., LeMay, V.M., Froese, K.L., Marshall, P.L., 2003. Imputing tree lists from aerial attributes for complex stands of south-eastern British Columbia. Forest Ecol. Manage. 177, 277–285.
Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E. (Eds.), 2010. National Forest Inventories: Pathways for Common Reporting. Springer, Dordrecht.
Vauhkonen, J., Maltamo, M., McRoberts, R.E., Næsset, E., 2014. In: Maltamo, M., Næsset, E., Vauhkonen, J. (Eds.), Forestry Applications of Airborne Laser Scanning: Concept and Case Studies. Springer, Dordrecht.
Westfall, J.A., Patterson, P.L., Coulston, J.W., 2011. Post-stratified estimation: within-strata and total sample size recommendations. Can. J. Forest Res. 41, 1130–1139.
Zhao, K., Popescu, S., Nelson, R., 2009. Lidar remote sensing of forest biomass: A scale-invariant estimation approach using airborne lasers. Remote Sens. Environ. 113, 182–196.