Forestry: An International Journal of Forest Research, 2025, 1–13
https://doi.org/10.1093/forestry/cpaf076
Original Article

The effect of sampling design in model-based small area estimation

Annika Kangas[1,*], Mari Myllymäki[2], Petteri Packalen[2]

[1]Natural Resources Institute Finland (Luke), Bioeconomy and environment, Yliopistokatu 6, Joensuu 80100, Finland
[2]Natural Resources Institute Finland (Luke), Bioeconomy and environment, Latokartanonkaari 9, Helsinki 00790, Finland

*Corresponding author. Natural Resources Institute Finland (Luke), Bioeconomy and environment, Yliopistokatu 6, Joensuu 80100, Finland.
E-mail: annika.kangas@luke.fi

Abstract

It has been shown that simple random sampling is not necessarily the best option, when data is collected for modelling and mapping forest resources. Instead, other sampling designs like systematic or stratified sampling may be better options for those purposes. Furthermore, it has been shown that for stratified sampling, Neyman allocation based on the influence of a given model predictor will produce the smallest estimation error for its coefficient. In this study, we explore if the small-area estimation can be improved by the selection of sampling design, and how that depends on the properties of the small areas. We tested four different sampling designs (simple random sampling, pseudo-systematic sampling, spatially balanced sampling, and stratified sampling) for small area estimation. We also tested two different versions of Neyman allocation: traditional Neyman allocation based on remote sensing variables, and another based on their influences on the estimated regression coefficients. The results show that the model-based small-area estimates were seriously underestimated for the domains with largest volume with all modelling methods, due to the model predictions not capturing exceptionally large values. This could only slightly be alleviated with the choice of a sampling design. On the other hand, the designs weighting the high-end volume domains produced less accurate results for the middle and low-end volume domains. The model-based estimation without field plots for calibrating the model is not capable of identifying the domains with largest values of target variables nor producing unbiased estimates for them. Thus, it is important to develop calibration methods applicable also for non-sampled domains.

Keywords: model-based; estimation; sampling design; small-area; simulation

Introduction

It was noted already in 1970's that simple random sampling is not necessarily the best sampling method for model-based inference. Royall (1970) proved in his classic paper that the optimal sampling design for model-based inference with a linear model through origin is purposive sampling of the units with the largest values of the predictor. Later, it was noted that while such a design is optimal if the model is correctly formulated, it is usually better to take a balanced sample (Royall and Herson 1973). If the model was not correctly formulated, there was a high risk of biased estimates. It has also been noted that the simple random sampling is not very efficient for mapping purposes, but designs such as systematic sampling or balanced sampling might be more efficient (Brus 2019).

The municipality-level inventory results in Finland are based on systematic cluster sampling and model-based small-area estimation with k-nearest neighbour (KNN) approach using satellite images as auxiliary data (e.g. Mäkisara et al. 2019). For the largest municipalities, it is possible to use design-based estimation such as post-stratification (Haakana et al. 2020), but the smallest municipalities might not have any sample plots. The forest management inventory for stand-level decision making, on the other hand, is model-based small-area estimation based on stratified sampling of field plots, aerial laser scanning data, and aerial images. The sampling design for these field plots does not aim for representative sample of the whole population, but rather aims at accurate models for predicting the forest variables in wall-to-wall fashion (Maltamo et al. 2021). The results can be calculated using either parametric models such as linear mixed-effect models (Astrup et al. 2019) or non-parametric models such as KNN (Maltamo and Packalen 2014). Bayesian approaches have also been used (Junttila et al. 2008).

Our hypothesis is that mapping and small-area estimation set additional requirements for sampling design compared to large-area estimation. If a design that is better for modelling purposes can be derived, it can be assumed such a design is also better for model-based small-area estimation, as the validity of the results solely lies on the quality of the model. The potential improvement is assumed to depend on the size of the small areas and the proportion of between-area variation from the total variation: the more the small areas differ from each other, the more likely borrowing information from nearby areas would result in biased estimates and vice versa. The possible improvements are also assumed to depend on the quality of the explanatory variables in regard to explaining the variation.

Sampling designs possibly better suited for small-area estimation are systematic designs or pseudo-systematic designs with

Handling editor: Dr. Tzeng Yih Lam
Received 18 June 2025; **revised** 30 September 2025; **accepted** 9 October 2025

one sample unit selected from a systematic grid (Brus 2019) or designs striving for a spatially balanced sample such as local pivotal method (LPM, Grafström et al. 2014). Stratification may also be used to obtain samples that are better fitted for modelling purposes than SRS samples.

If stratified sampling is used, the sample needs to be allocated to the strata. The allocation can be carried out in different fashions, for instance proportionally to the strata size. The allocation of the sample can also be optimized based on the size of the stratum h (N_h), the measurement cost within the stratum h (c_h) and the variation of the variable of interest (S_h) within stratum h using so-called optimal allocation (Cochran 1977, Eq 5,23). Neyman allocation is a special case of this allocation, assuming the measurement costs are equal in all strata as

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} \quad (1)$$

However, Neyman (1934) showed that if we wish to select the sample that would provide the smallest variance of a regression coefficient of one predictor of a model, an optimal design would be a stratified sampling using Neyman allocation based on the variance of influence of this predictor variable rather than the variance of the variable. These influences describe the effect of excluding a specific observation from a sample on the coefficient

$$INF_{in} = \hat{\beta}_n - \hat{\beta}_{(i)n} \quad (2)$$

where $\hat{\beta}_n$ is the coefficient estimated from all n observations, and $\hat{\beta}_{(i)n}$ is the coefficient estimated with all but observation i (e.g. Mehtätalo and Lappi 2020 p. 103). The influence for observations at the mean of a given predictor variable is zero and increases towards the extremes. Unfortunately, the influence of any observation is unobservable before the sample is taken, and thus the truly optimal sample is unattainable. However, it is possible to approximate the influence function to get efficient sampling designs (Chen and Lumley 2022).

The aim of the paper is to test the performance of model-based small-area estimators under four different sampling designs, in order to evaluate the effect of sampling designs on model-based estimation. Our aim is to find out what kind of design would be most efficient, i.e. produce the most accurate results for a given sampling effort. We also analyze how this depends on the size of the small areas (and size of the sample within it) as well as the homogeneity of the small areas, measured with the proportion of the between-area variation of the total variation in the population. We utilize a unit-level linear model, empirical best linear unbiased prediction (EBLUP) approach with a mixed linear model, a k-nearest-neighbor model and a k-nearest-neighbor model adjusted with an EBLUP approach. The study is carried out as a simulation study where the 'ground truth' at population level is known. The designs and modelling methods are compared using empirical standard error, bias and RMSE and the empirical coverage of the estimated confidence intervals.

Materials

The ground truth data for the simulation experiment was prepared as follows (Fig. S1 in supplementary material): We utilized wall-to-wall airborne laser scanning (ALS) features on a region of ~5900 ha (Kangas et al. 2023, 2025). Data contained ALS features on a grid of 231 824 square cells of 16 m × 16 m. To generate a 'ground truth' we simulated for each pixel i a volume with $y_i = \exp(\mu_i + e_i)$, where μ_i is the predicted logarithm of volume from an external model and e_i is the simulated random error. The errors were assumed to be autocorrelated and stem from a zero-mean Gaussian random field with exponential semivariogram model having variance $\sigma^2 = 0.0538$, nugget effect $\tau^2 = 0.0292$ and range parameter $\phi = 337$, resulting in a practical range of 1011 meters (Kangas et al. 2023).

The external model was based on an independent modelling dataset that had 1044 observations with field-measured values of total plot volume, basal area, mean diameter, and mean height. The modelling dataset contained ALS features that were also available in the wall-to-wall data (i.e. population). The modelling data including details of ALS features are documented in Tuominen et al. (2017) and Balazs et al. (2022). We modelled the plot-specific $\ln(y)$ using the best seven-predictor model estimated with leaps package in R (Thomas Lumley based on Fortran code by Alan Miller 2024), using the following 17 ALS features: The maximum height of the points, Height at which given percentiles (20% last echo, 45%, 55%, 65%, 70%, 90% first echo) of vegetation points are accumulated (m), Proportion of vegetation points relative to all points (%, first and last echo), Skewness of the vegetation point heights, Proportion of points above mean height, Proportion of points having cumulated at 20% of the height from all points (%, last echo), Rumple index, Inner volume (Véga et al. 2016), SumEntropy (Haralick et al. 1973) of canopy surface model and Average intensity of ALS echoes. Unless otherwise stated, the features were calculated from the first echoes. The 'true' model had residual standard error = 0.232 and multiple $R^2 = 0.897$ (Kangas et al. 2023, Table 1). This model was only used for generating the population, but not for calculating the simulation results.

Methods

Model-based estimators using a mixed linear model

For indirect model-based estimators we assumed an unit-level linear mixed model

$$y_{ji} = \mathbf{x}_{ji}\boldsymbol{\beta} + \upsilon_j + e_{ji},\ \ j = 1,\ldots,J; i = 1,\ldots,n_j \quad (3)$$

where y_{ji} is the volume in unit i within domain j (m^3/ha), $\mathbf{x_{ji}}$ is the vector containing all observed predictor values for the fixed effects in unit i and domain j, $\boldsymbol{\beta}$ is the vector of fixed model coefficients, υ_j is the random area effect for domain j ($\upsilon_j \sim N(0, \sigma_\upsilon^2)$), e_{ji} is the random error for unit i in domain j ($e_{ji} \sim N(0, \sigma_e^2 c_{ji}^{-1})$), c_{ji}^{-1} is the weight of the unit i in domain j (used in case of heteroscedastic residuals), and n_j is the sample size within domain j. In matrix form this is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\upsilon} + \mathbf{e} \quad (4)$$

where $\mathbf{Z}$ is an (N × J) indicator matrix of units belonging to domain j. This leads to a block-diagonal variance–covariance matric of the errors (Militino et al. 2007)

$$\mathbf{V} = \sigma_\upsilon^2 \mathbf{Z}\mathbf{Z}^\mathbf{t} + \sigma_e^2 \mathbf{C}^{-\mathbf{1}} \quad (5)$$

where $\mathbf{C}$ is an (N × N) diagonal weight matrix with elements c_{ji}, resulting a constant correlation between units i and k within each domain j, namely

$$\mathrm{cor}\left(e_{ji}, e_{jk}\right) = \frac{\sigma_\upsilon^2}{\sigma_\upsilon^2 + \sigma_e^2} \quad (6)$$

With an alternative specification, it would also be possible to assume a continuous correlation depending on the distance

Table 1. The parameter values for the 'true' model $\log(y) = \beta_0 + \beta x_1 + \beta x_2 + \beta x_3 + \beta x_4 + \beta x_5 + \beta x_6 + \beta x_7 + \varepsilon$.

	Estimate	Std. Error	T Value
Intercept	1.536607	0.141549	10.85565
Maximum height	0.02380686	0.007498333	3.174953
Height where 55% of points are cumulated	0.02799505	0.01001328	2.795793
Height where 90% of points are cumulated	0.03482004	0.01273452	2.734304
Prop. of vegetation points	0.0104878	0.0005722577	18.32706
Prop. of points cumulated at 20% of height	0.02489417	0.00371768	6.696159
Average Intensity	−0.05024336	0.007689848	−6.533726
SumEntropy	0.3666484	0.02072759	17.68891

between the units considered (Kangas et al. 2023, Wadoux and Heuvelink 2023).

The mean for domain j is (Militino et al. 2007, Mauro et al. 2017)

$$\mu_j = \bar{\mathbf{x}}_j\boldsymbol{\beta} + \upsilon_j + \frac{1}{N_j}\sum_{i=1}^{N_j} e_{ji} \tag{7}$$

where $\bar{\mathbf{x}}_j$ is the average vector of the values $\mathbf{x}_{ji}$ in domain j, N_j is the total number of units within the domain j, υ_j is the area (or domain) effect and e_{ji} is the error related to the unit i within domain j. In (7), the area effect vector υ is assumed to be estimable from data. However, if there are no observations from the domain (i.e. $n_j = 0$), the estimate of area-effect $\hat{\upsilon}_j = 0$. Moreover, if the N_j is large enough, the last term will be approximately zero. In such case the model-based estimator will simply be

$$\hat{\mu}_{MB,j} = \bar{\mathbf{x}}_j\hat{\boldsymbol{\beta}} \tag{8}$$

and the estimator of its variance (Breidenbach et al. 2016, Kotivuori et al. 2020) is $\mathrm{var}(\hat{\mu}_{MB,j}) = \bar{\mathbf{x}}_j'\,\mathrm{var}(\hat{\boldsymbol{\beta}})\,\bar{\mathbf{x}}_j$, estimated from the estimated variances of the parameters. Estimator for the domain mean $\hat{\bar{y}}_{MB,j}$ coincides with the estimation of the superpopulation parameter $\hat{\mu}_{MB,j}$. However, when we wish to estimate the variance of the domain mean $\hat{\bar{y}}_{MB,j}$ rather than the superpopulation parameter $\hat{\mu}_{MB,j}$, the formula is written as (Breidenbach et al. 2016)

$$\mathrm{var}\left(\hat{\bar{y}}_{MB,j}\right) = \bar{\mathbf{x}}_j'\,\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right)\bar{\mathbf{x}}_j + \frac{1}{N_j^2}\sum_{i=1}^{N_j}\sum_{l=1}^{N_j}\widehat{\mathrm{cov}}\left(e_{ji}, e_{jl}\right) \tag{9}$$

where the last term consists of the estimated covariances (or variances in case $i = l$) of the errors of all N_j pixels or units within the domain. In the case of an area-effect model, the last term can be estimated with $\frac{1}{N_j^2}\left(\left(\sum_{i=1}^{N_j}\sigma_\upsilon\right)^2 + \sum_{i=1}^{N_j}\sigma_e^2\right)$.

When the model-based estimator is calibrated using the observations from the domain j to estimate the area effect $\hat{\upsilon}_j$, the estimator of the mean is of the form (Mauro et al. 2017)

$$\hat{\mu}_{EBLUP,j} = \bar{\mathbf{x}}_j\hat{\boldsymbol{\beta}} + \hat{\upsilon}_j \tag{10}$$

Its MSE under the assumed model (3) is estimated with (Militino et al. 2007, see also Breidenbach et al. 2018, Mauro et al. 2017, Rao and Molina 2015 chapter 5.2.6)

$$MSE\left(\hat{\mu}_{EBLUP,j}\right) = g_{\ j,1} + g_{j,2} + 2g_{j,3} + g_{j,4} \tag{11}$$

where

$$g_{j,1} = \left(1 - \hat{\gamma}_j\right)\hat{\sigma}_\upsilon^2 \tag{12}$$

$$g_{j,2} = \left(\bar{X}_{j,r} - \hat{\gamma}_j\bar{x}_j\right)^2\mathrm{var}\left(\hat{\beta}\right) \tag{13}$$

$$g_{j,3} = c_{j.}^{-2}\left(\hat{\sigma}_\upsilon^2 + \hat{\sigma}_e^2/c_{j.}\right)^{-3}$$
$$\left[\hat{\sigma}_e^4\widehat{\mathrm{var}}\left(\hat{\sigma}_\upsilon^2\right) + \hat{\sigma}_\upsilon^4\widehat{\mathrm{var}}\left(\hat{\sigma}_e^2\right) - 2\hat{\sigma}_e^2\hat{\sigma}_{e\upsilon}^2\widehat{\mathrm{cov}}\left(\hat{\sigma}_e^2, \hat{\sigma}_\upsilon^2\right)\right] \tag{14}$$

$$g_{j,4} = \frac{\hat{\sigma}_e^2}{N_j^2}\sum_{i\in j_r} c_{ji}^{-1} = \frac{\left(N_j - n_j\right)\hat{\sigma}_e^2}{N_j^2}\bar{X}_{j,r} \tag{15}$$

where $\hat{\sigma}_e^2$, $\hat{\sigma}_\upsilon^2$, are estimates of the corresponding variances, $\gamma_j = \frac{\sigma_\upsilon^2}{\sigma_\upsilon^2 + \sigma_e^2/c_{j.}}$ and $c_{j.}$ is the sum of weights c_{ji} within domain j. Here $MSE\left(\hat{\mu}_{MB,j}\right) = E\left(\hat{\mu}_{MB,j} - \mu_{MB,j}\right)^2$ where the expectation is taken with respect to the model (3) (Rao and Molina 2015 p. 98), but it does not describe the bias under the design, as the design-bias is unobserved and unobservable, unless the whole population is measured.

Model-based estimators using KNN

The KNN estimator for one unit i in domain j, $\hat{y}_{ji}$, is

$$\hat{y}_{ji} = \sum_{k=1}^{K} w_k^{ji} y_k^{ji} \tag{16}$$

where y_k^{ji} is measured value of variable y in the kth nearest neighbor unit of i in domain j, and K is the number of the used neighbors. The weights w_k^{ji} are calculated based on the similarity between the target unit ji and its neighbors k, and this similarity is defined by a distance metric in predictor space. In this study, we use the Euclidean distance metric.

In the model-based framework, the estimator for the mean in domain j is the mean of the predictions

$$\hat{\mu}_{KNN,j} = \frac{1}{N_j}\sum_{i=1}^{N_j}\hat{y}_{ji} = \frac{1}{N_j}\sum_{i=1}^{N_j}\sum_{k=1}^{K} w_k^{ji} y_k^{ji} \tag{17}$$

However, suitable estimators for the variance and the MSE are missing. Kangas et al. (2024) proposed to use the KNN approach to predict the error variance for each unit i within domain j as

$$\hat{e}_{ji}^2 = \sum_{k=1}^{K} w_k^{ji}\left(e_k^{ji}\right)^2 \tag{18}$$

That is, $\hat{e}_{ji}^2$ is the weighted average of the squared observed residuals of the K nearest neighbors of that unit. We approximated

the between-unit covariances using a fixed correlation assumption with correlation (6) estimated from the area-effect model (3). These assumptions lead to an estimator

$$\widehat{\mathrm{var}}\left(\hat{\mu}_{KNN,j}\right) = \frac{1}{N_j^2}\sum_{i=1}^{N_j}\hat{e}_{ji}^2 + 2\frac{1}{N_j^2}\sum_{i=1}^{N_j}\sum_{l>i}^{N_j}\hat{\mathrm{cov}}\left(\hat{e}_{ji},\hat{e}_{jl}\right) = \frac{1}{N_j^2}\left\{\left(\left(\sum_{i=1}^{N_j}\hat{e}_{ji}\right)^2 - \sum_{i=1}^{N_j}\hat{e}_{ji}^2\right)\cdot\hat{\rho} + \sum_{i=1}^{N_j}\hat{e}_{ji}^2\right\}, \quad (19)$$

where $\hat{e_{ji}} = \sqrt{\hat{e}_{ji}^2}$. Another alternative is to combine an indirect KNN estimator and an EBLUP estimator using mixed model. The composite of these two estimators can be obtained by using the mixed model (3) to predict the residuals of the KNN estimate for each (non-sampled) unit i in domain j, and calculating the result as a sum of the KNN mean (17) and the EBLUP estimator (10) for the mean error as

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N_j}\hat{y}_{ji} + \bar{\mathbf{e}}_j\hat{\boldsymbol{\beta}} + \hat{v}_j \quad (20)$$

This approach enables us to use the EBLUP analytical (11–15) formulas to calculate the variance of the estimator, and to utilize the within-area correlations estimable from the area effects $\hat{v}_j$. Bell et al. (2022) made a bias correction to KNN small-area estimates by calculating a design-weighted mean of the observed errors. The approach of combining KNN and an area-effect model can be seen as a generalization of that approach (e.g. Nothdurft et al. 2009).

Generation of small domains

To have spatially contiguous and non-overlapping sets of small domains (in what follows, the term domain refers specifically to the simulated small areas, and term small area refers to generic concept of small area), the large area was divided to J domains with a k-means clustering approach with the coordinates as the sole predictor (called HET from now on, Figs. S1 and S2 in supplementary material). To make a more homogeneous set of small domains, we used Maximum height (hmax_f) variable in addition to the coordinates (HOM, Figs. S1 and S3). We used values of $J = 100$ (resulting on average ∼59 ha domains) and 500 (resulting on average ∼12 ha domains). As the average size of a forest stand in Finland is ∼2 hectares and average forest estate size ∼30 hectares, even the more homogeneous domains are not as homogeneous as forest stands would be, as each domain typically would include several stands.

The linear model (3) results of the simulation experiment were calculated using the three best predictors found with leaps R package. The best predictors for a linear model were Proportion of vegetation points relative to all points (%, first echo Pveg_f and last echo Pveg_l), and Inner volume (Volin). The RMSE of this model, calculated from all units in the population, was 36.11 m^3/ha and R^2 0.87. In the simulations, the model was estimated from each observed sample (Fig. S1), and is thus less accurate than the model estimated from the whole population data. In KNN, same predictors were used. The model estimated from the whole population data was used for estimating the true influences of the observations and for assessing the heterogeneity of the generated small domains.

With $J = 100$ and only coordinates used for clustering, a linear area-effect model (3) with the above-mentioned explanatory variables resulted in as a between-domain variation (i.e. variance of the area-effect σ_v^2) 9.3% of the total variation ($\sigma_v^2+\sigma_e^2$), and when Maximum height (hmax_f) variable was also used for clustering, the between-domain variation was 18.3% of total variation. With $J = 500$ the between-domain variation was 16.5% for the heterogeneous area division and 32.2% for the homogeneous division. The between-domain variation represents the variation that the fixed part of the model could not explain. When clustering was made solely on coordinates (Fig. 1 heterogeneous), the mean volumes of the produced domains had a symmetric distribution. However, when hmax_f was introduced to the area division, the corresponding distribution was markedly skewed (Fig. 1 homogeneous). With 500 areas, both distributions were skewed.

Estimators and sampling designs

In this study, we compared (i) simple random sampling (SRS), (ii) local pivotal method (LPM) using the explanatory variables of the model to balance the sample, (iii) pseudo-systematic sampling with one unit selected from m (= total sample size) clusters defined by k-means clustering (SYS), and (iv) stratified sampling (STR). In stratified sampling we first selected strata based on the quantiles (0%, 20%, 40%, 60%, 80%, and 100%) of the Inner volume variable (Volin). Then, the stratum based on quantiles 0–20% was kept as such and the other four strata were further divided to two strata based on variable Proportion of vegetation points relative to all points calculated from first echos (Pveg_f). This resulted in nine strata with 46 365, 23 159, 23 206, 23 180, 23 184, 23 180, 23 185, 23 146, and 23 219 units.

While the stratification is based on two variables, the allocation of the sample to these strata can still be carried out in several fashions. Here, we tested Neyman allocation based on the variation of Influence (Eq. 2) of Volin and Pveg_f separately and the sum of the influences of these two variables (later called STR_{Inf}, see Fig. S1 in supplementary material). The sampling allocations and results for all these variations were almost identical, so only the allocation based on the sum of the influences is shown. This allocation resulted in a sample size of 13, 3, 37, 20, 108, 64, 148, 205 and 402 units in the nine strata, respectively. The influence was calculated from a model fitted to the whole data set, and thus it is a true (but in reality, unattainable) influence in this dataset.

Finally, Neyman allocation was calculated in traditional way (Eq. 1), based on the variance of the same variables and their sum (later called STR_{Ney}, see Fig. S1 in supplementary material). Also in this case, the sampling allocations for all these variations were almost identical, so only the allocation based on the sum of the variables is shown, resulting a sample size of 3, 10, 26, 94, 100, 141, 135, 177 and 314 units in the nine strata, respectively. Especially, using the influence as a basis for stratification very heavily weight the largest values of the predictors.

We generated $S = 500$ random samples s of size $m = 1000$ from the simulated population data. With $J = 100$ the expected sample size within each domain was 10 and with $J = 500$ it was 2. It means that for many of the domains no data for EBLUP calibration was available, especially in the low-end volume domains with stratified sampling and Neyman allocation (either STR_{Inf} or STR_{Ney}). The sampling fraction was 0.43%.

The tested estimators were model-based estimator using predictions from a linear model estimated for the large area (MBLA) and model-based estimator with predictions from a KNN model. In addition, we used model-based estimator where the predictions were calibrated with observations from the domain with EBLUP and a combination of KNN and EBLUP approach (KEB) (See Fig. S1 in supplementary material).

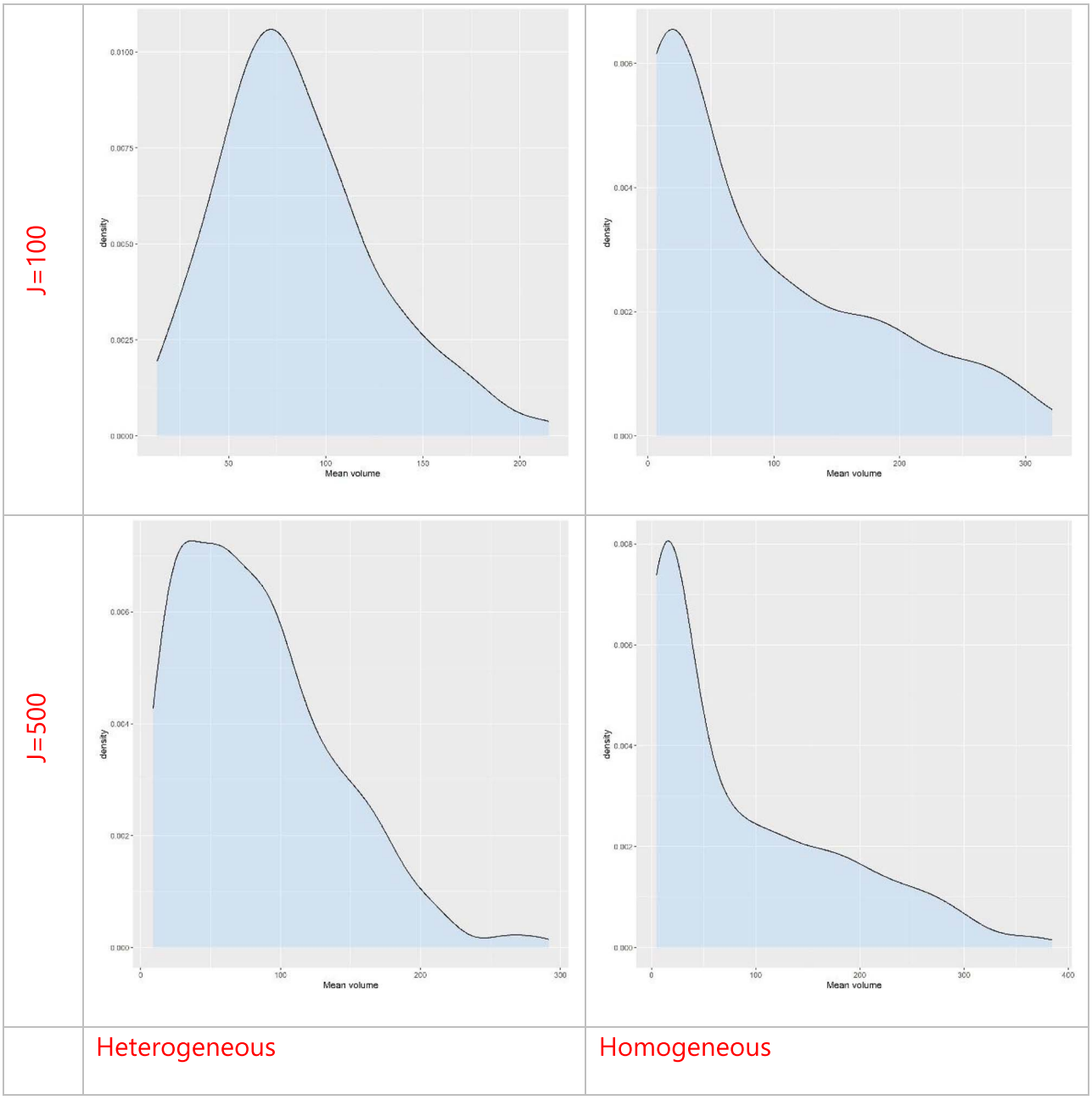

Figure 1. The distribution of true mean volumes across domains with 100 areas (upper) and 500 areas (lower).

Performance evaluation

We estimated the bias as a difference between the mean of mean estimates $\hat{\bar{y}}_{j,s}$ from samples $s = 1, \ldots, S$ and true mean $\bar{Y}_j$, i.e.

$$Bias_j = \frac{1}{S}\sum_{s=1}^{S}\hat{\bar{y}}_{j,s} - \bar{Y}_j \tag{21}$$

the true standard deviation as the standard deviation between the mean estimates of the S simulations, i.e.

$$Se_j = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\bar{y}}_{j,s} - \bar{\bar{y}}_{j,s}\right)^2} \tag{22}$$

and the true RMSE as

$$RMSE_j = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\bar{y}}_{j,s} - \bar{Y}_j\right)^2} \tag{23}$$

We also calculated for each of the estimators the empirical coverage of the confidence interval (CI). First, we estimated for each method the mean $\hat{\bar{y}}_{j,s}$ (Equations 8, 10, 17, and 20) and standard deviation $\hat{\sigma}\left(\hat{\bar{y}}_{sj}\right)$ with the corresponding estimators (Equations 9, 11, 19, and 11). Then, for each domain, we counted the proportion of samples s where the true mean was included in the estimated confidence interval (Breidenbach et al. 2016)

$$CI = \hat{\bar{y}}_{sj} \pm 1.96\,\hat{\sigma}\left(\hat{\bar{y}}_{sj}\right) \tag{24}$$

We used the R-package JoSAE (Breidenbach et al. 2018) for calculating the EBLUP predictions and their variances and the R package yaImpute (Crookston and Finley 2007) for calculating the KNN predictions. The whole simulation experiment is described in the Fig. S1.

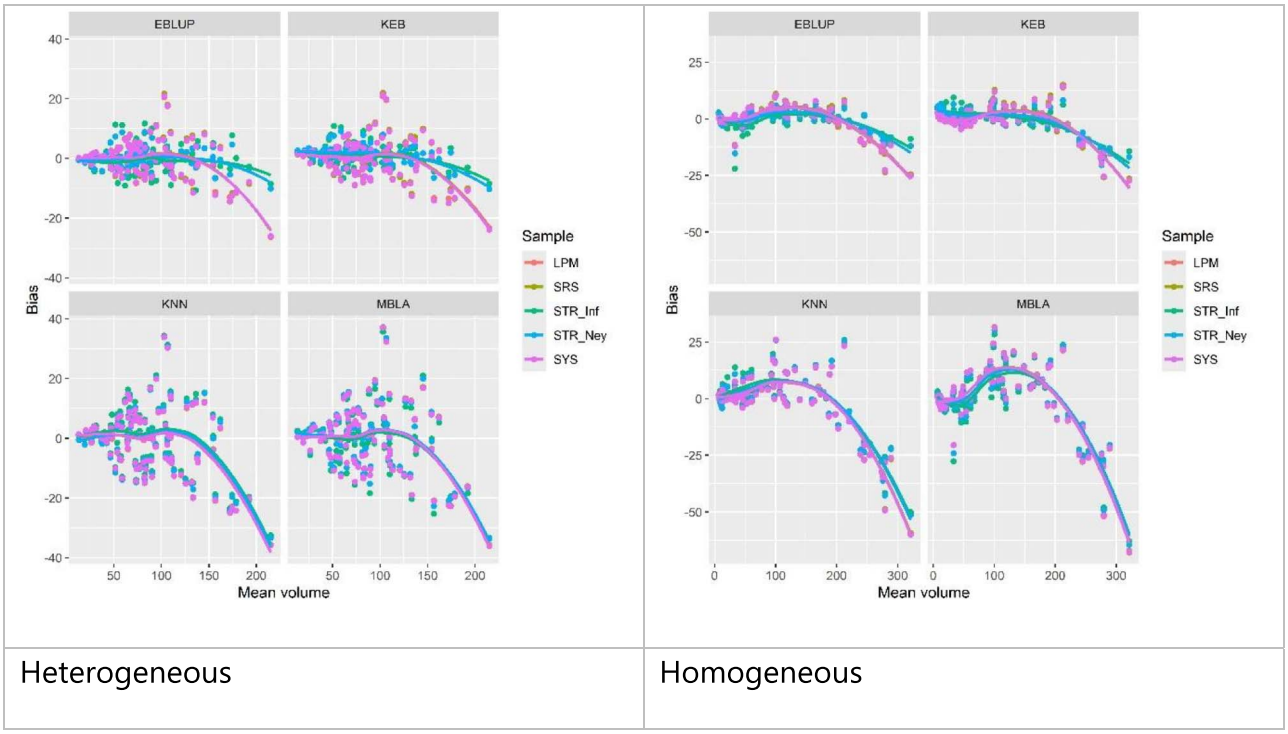

Figure 2. The bias of 100 domains as a function of the mean volume of the domains with the four sampling designs and four modelling approaches. Stratified sample is allocated with respect to variation of the sum of the explanatory variables Volin and Pveg_f (STR_{Ney}) or their influences (STR_{Inf}).

Results

The four sampling designs and two allocation schemes for the stratified sampling, SRS, LPM, SYS, STR_{Inf}, and STR_{Ney} produced fairly similar results, but differences between the designs could be seen in domains having the highest mean volumes. In both sets of small domains (HOM and HET) the results with respect to bias were fairly good for domains with mean volume less than 150 m^3/ha, but large underestimates were observed for the domains with larger volumes (Fig. 2). In the more heterogenous domains, the largest biases were smaller. It can be assumed that the errors in heterogeneous domains somewhat cancelled each other out, while this was not the case in the homogeneous areas. The increasing bias is due to the fact that a model (i.e. an expected value conditional on the used predictors) can never capture the exceptionally large (or small) values, unless the model is perfect, i.e. can explain all the variation. This phenomenon is clear from any residual plots against the true values (cf. Ståhl et al. 2024).

The biases at the high-end domains were generally smallest when using EBLUP and KEB methods with Neyman allocation based on the influences (STR_{Inf}), but also the traditional Neyman allocation (STR_{Ney}) produced good results regarding bias. Partly this might be due to higher number of observations to be used for calibration in these domains, but the EBLUP and KEB results only included the domains from which there was at least one observation in the sample. The worst results regarding to bias on the high-end volume domains were produced by SYS, SRS, and LPM. With MBLA and KNN the differences in bias between the sampling methods were negligible.

The large biases in the domains with high volumes can also be seen from the empirical coverage of the confidence intervals. On average, the empirical coverages for both the heterogeneous and homogeneous domains were good, varying from 92% to 98% for MBLA and from 85% to 94% with KNN. With both MBLA and KNN the empirical coverage reduced as a function of the true mean volume of the domain with all sampling designs (Fig. 3). The poorest sampling designs in this respect were the SRS, LPM and SYS. With the EBLUP methods and the composite of EBLUP and KNN (KEB) the decrease was generally not as steep, which was due to the possibility of calibrating the estimate. These results are only shown for areas from which at least one sample unit is available for calibration. On the other hand, the empirical coverage in the low-end volume domains was also poor for stratified sampling designs (STR_{Inf} and STR_{Ney}), as for these domains suitable observations for calibration were scarce.

The results were very similar with smaller domains ($J = 500$), except that the largest mean volume increased from ~200 to 300 m^3/ha in heterogeneous domains and from ~300 to 400 m^3/ha in homogeneous domains. Therefore, the biases at the high-volume end were even more pronounced (Fig. 4), the largest biases being ~100 m^3/ha underestimates. Also, the empirical coverages for $J = 500$ (Fig. 5) were fairly similar than with $J = 100$ areas (Fig. 3), and therefore, in the following only figures for $J = 500$ are presented. The estimator of standard error for KNN (Eq. 17) can inherently deal with a heteroscedastic variance: when the residuals are small for lower volumes, their standard errors also tend to be small in the low-end domains. For the estimator of the standard error for MBLA (Eq. 11), on the other hand, homoscedastic variance was assumed. This can be seen from the empirical coverages where the coverage is smaller for KNN than for linear model in the low-end domains and higher in the high-end domains.

Regarding to the RMSEs, the stratified samples again produced the best results in the domains with the highest mean volumes (Fig. 6). The homogeneous domains had much higher RMSEs than heterogeneous domains. Partly this is due to heterogeneous domains having much shorter tail in the mean

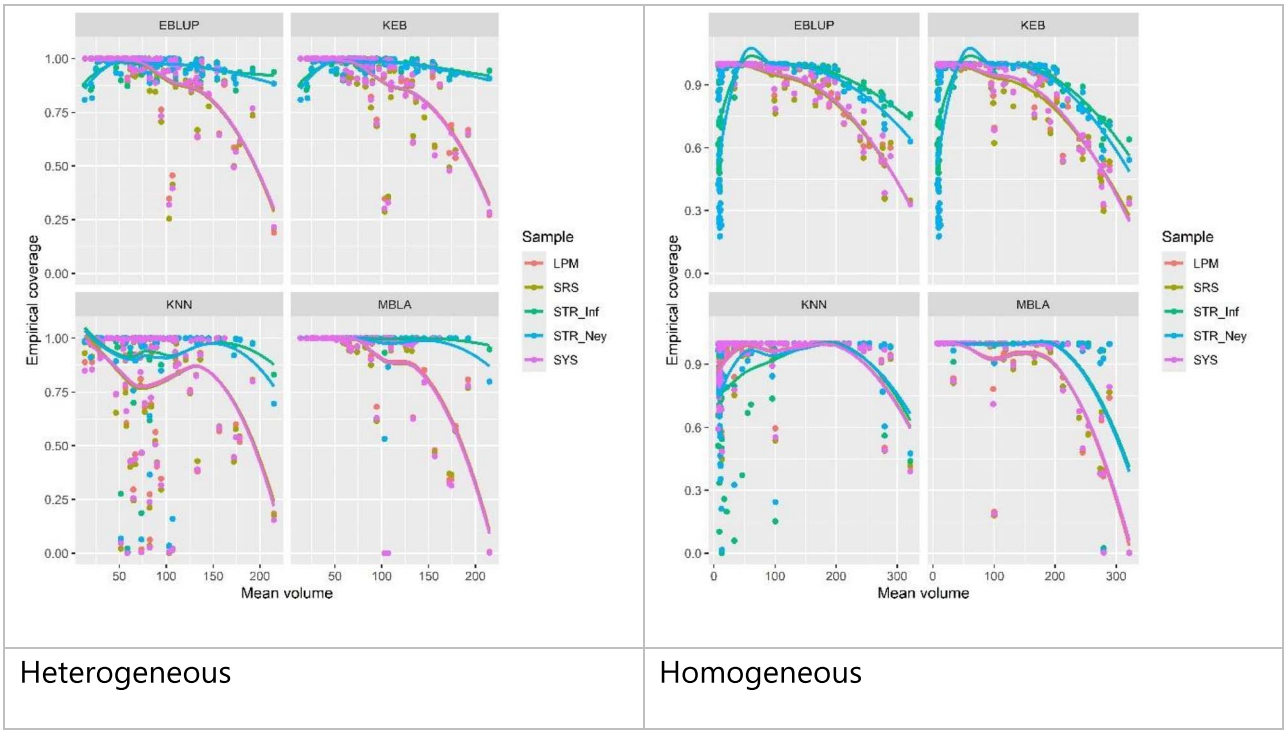

Figure 3. The empirical coverage of 100 domains as a function of the mean volume of the domains with the four sampling designs and four modelling approaches. Stratified sample is allocated with respect to variation of the sum of the explanatory variables Volin and Pveg_f (STR_{Ney}) or their influences (STR_{Inf}).

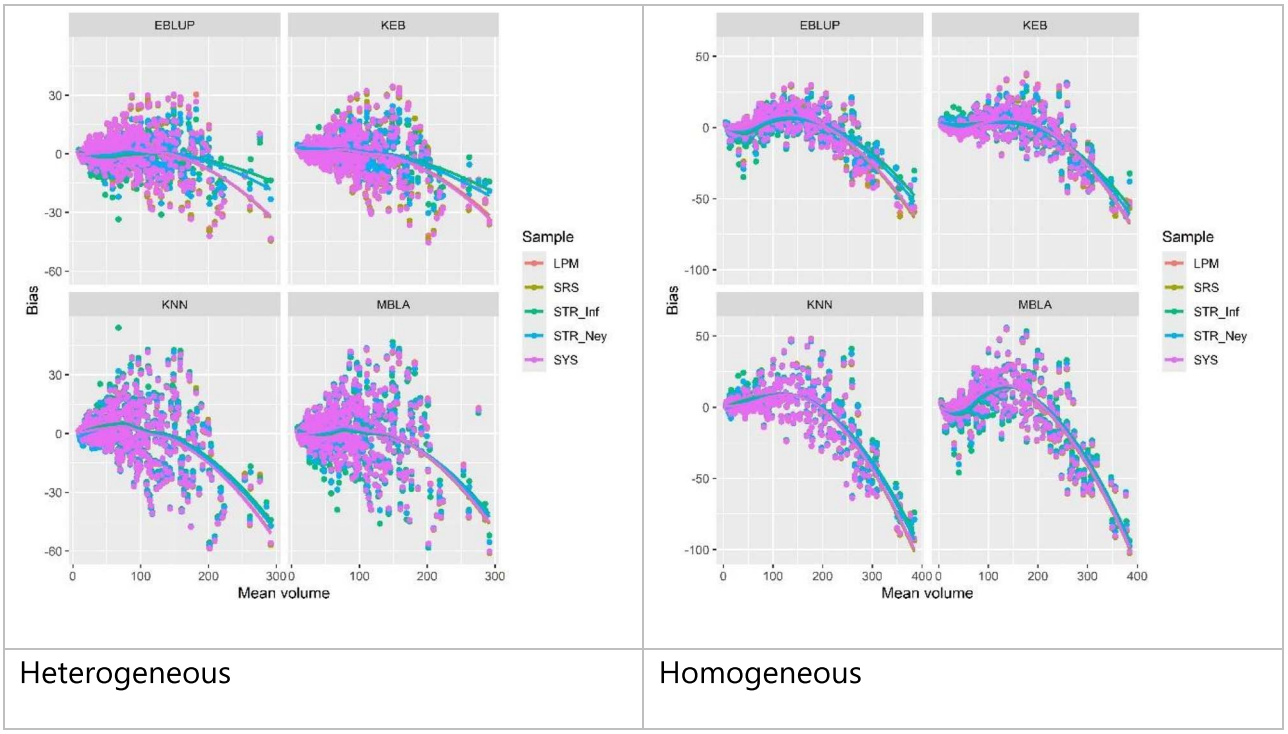

Figure 4. The bias of 500 domains as a function of the mean volume of the domains with the four sampling designs and four modelling approaches. Stratified sample is allocated with respect to variation of the sum of the explanatory variables Volin and Pveg_f (STR_{Ney}) or their influences (STR_{Inf}).

volume distribution (Fig. 1, heterogeneous). Moreover, calibrated EBLUP and KEB had generally smaller RMSEs than the uncalibrated KNN and MBLA. It is notable that in the heterogeneous domains and small or medium mean volumes SRS with EBLUP and KEB were better than STR_{Inf} or STR_{Ney}. This is likely due to calibration working better with less weighted data, i.e. with allocation better balanced across the low-end domains. In all other occasions STR_{Inf} or STR_{Ney} was better than SRS, meaning that stratification could reduce the RMSE in the high-end domains, but the effect was quite small.

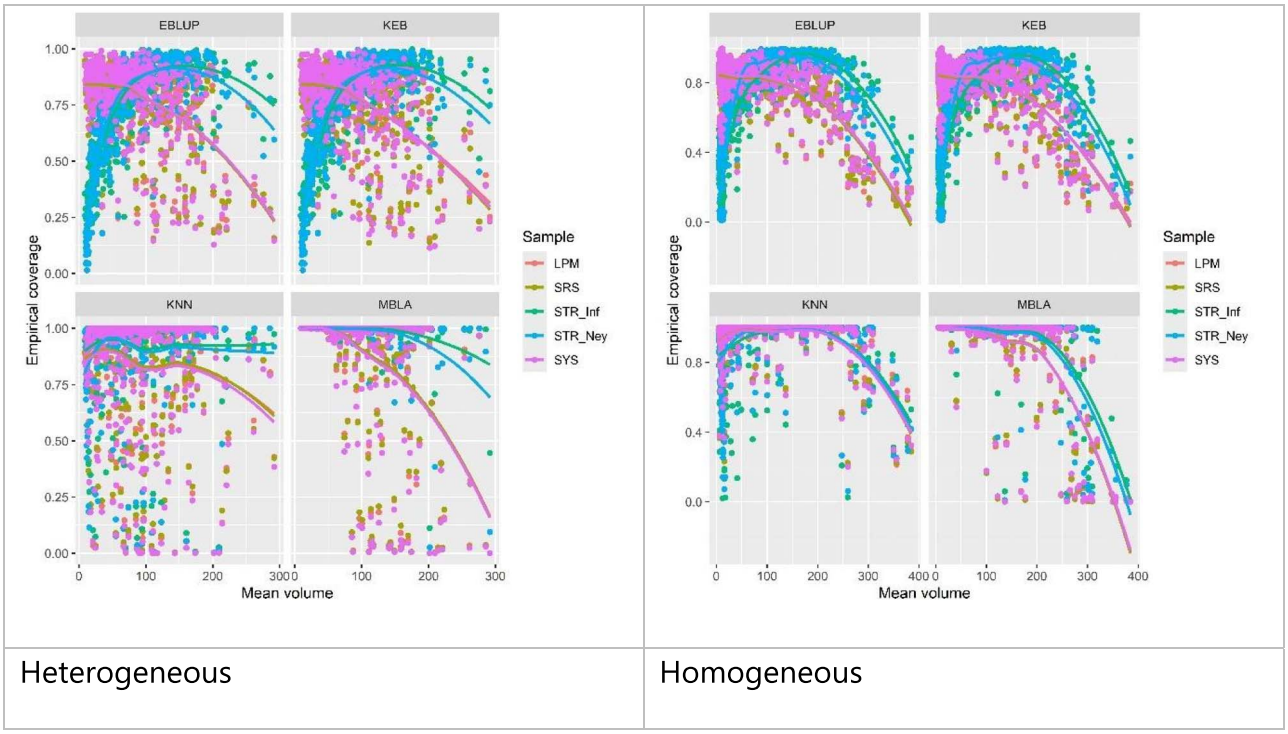

Figure 5. The empirical coverage of 500 domains as a function of the mean volume of the domains with the four sampling designs and four modelling approaches. Stratified sample is allocated with respect to variation of the sum of the explanatory variables Volin and Pveg_f (STR_{Ney}) or their influences (STR_{Inf}).

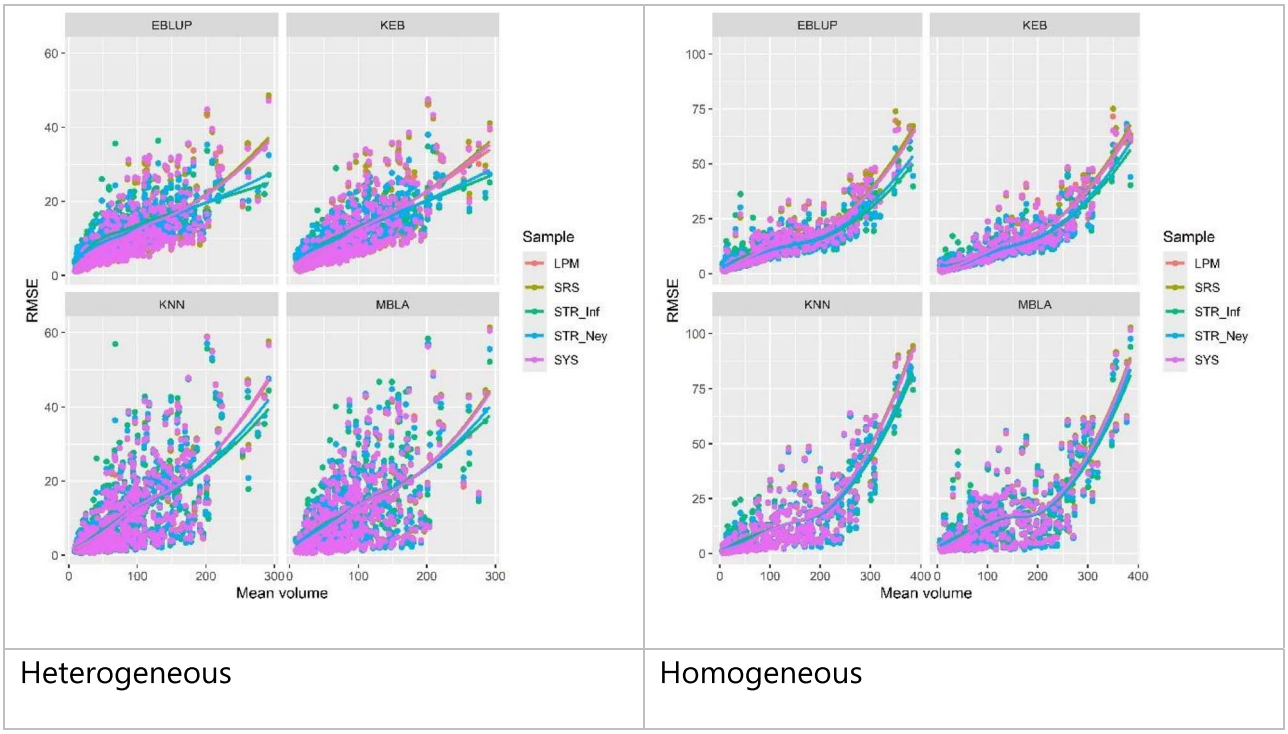

Figure 6. The RMSE of 500 domains as a function of the mean volume of the domains with the four sampling designs and four modelling approaches. Stratified sample is allocated with respect to variation of the sum of the explanatory variables Volin and Pveg_f (STR_{Ney}) or their influences (STR_{Inf}).

As a function of the within-domain variance and homogeneous areas, the EBLUP and KEB estimates were nearly unbiased with both of the stratified designs (STR_{Inf} and STR_{Ney}, Fig. 7). With MBLA and KNN the bias increased towards the high-end volume domains basically irrespective of the sampling design, but less in homogeneous domains than in heterogeneous domains. Regarding to the RMSE, the stratified designs were again better in the high-end domains, but SRS and SYS were often better in the

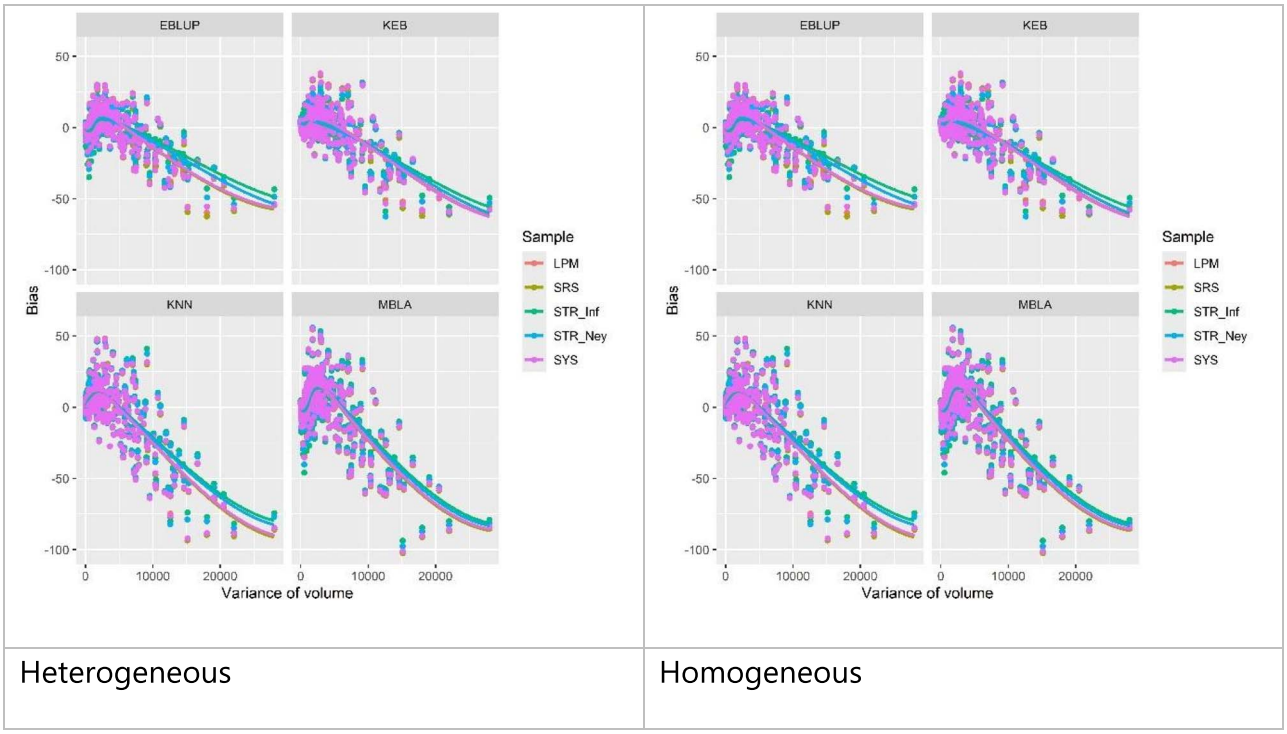

Figure 7. The bias of 500 domains as a function of the variance of volume within the domains with the four sampling designs and four modelling approaches. Stratified sample is allocated with respect to variation of the sum of the explanatory variables Volin and Pveg_f (STR_{Ney}) or their influences (STR_{Inf}).

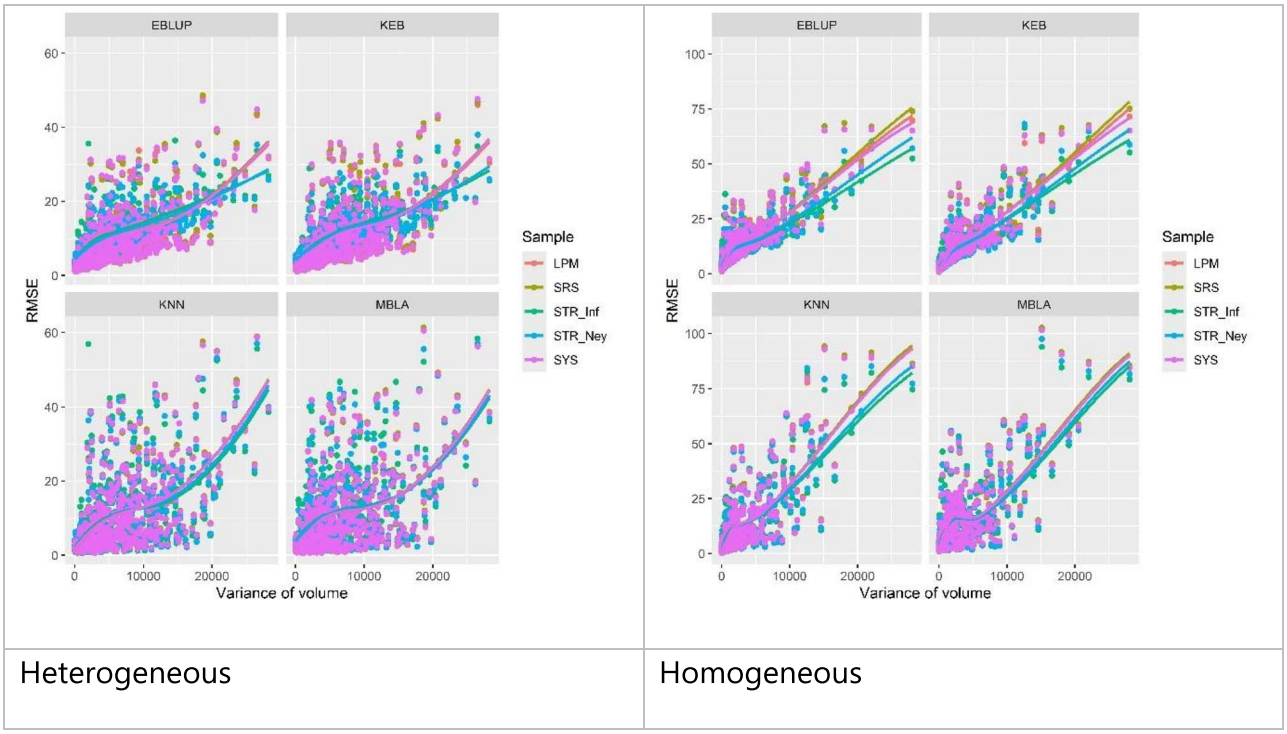

Figure 8. The RMSE of 500 domains as a function of the variance of volume within the domains with the four sampling designs and four modelling approaches. Stratified sample is allocated with respect to variation of the sum of the explanatory variables Volin and Pveg_f (STR_{Ney}) or their influences (STR_{Inf}).

lower end domains, especially when calibration was used (Fig. 8). However, the differences were small.

On average over all methods, both with 100 and 500 domains, if the domains were heterogeneous, systematic sampling produced the smallest standard error (22) and RMSE (23). The largest Se and RMSE were produced with STR_{Ney} (Table 2). In the homogeneous domains the STR_{Inf} allocation based on the variance of the influence of the two predictors and STR_{Ney} allocation based

Table 2. The averages of Se (Eq. 22), bias (Eq. 21) and RMSE (Eq. 23) over the 100 and 500 small areas with homogeneous and heterogeneous area division using Volin and Pveg (first echo pveg_f) for stratification and allocated according to the variance of the sum of them STR_{Ney} or the sum of their influences STR_{Inf}.

		$J = 100$						$J = 500$					
		Homogeneous			Heterogeneous			Homogeneous			Heterogeneous		
	Method	Se	Bias	RMSE	Se	Bias	RMSE	Se	Bias	RMSE	Se	Bias	RMSE
LPM	**EBLUP**	5.36	−0.44	6.51	5.33	−0.22	7.33	7.15	−0.71	10.11	6.19	−0.05	10.14
	KEB	4.87	−0.44	6.36	5.34	−0.14	7.59	6.32	−0.69	9.54	6.15	0.11	10.24
	KNN	2.25	−1.05	7.38	1.92	−0.75	8.69	2.77	−1.43	10.00	2.39	−0.28	10.64
	MBLA	2.58	−0.44	9.02	1.56	−0.22	8.53	2.42	−0.90	11.26	1.75	−0.06	10.72
SRS	**EBLUP**	5.43	−0.26	6.61	5.35	−0.13	7.31	7.20	−0.79	10.11	6.21	−0.05	10.12
	KEB	4.90	−0.24	6.44	5.31	−0.14	7.54	6.35	−0.74	9.52	6.24	0.06	10.26
	KNN	2.36	−1.00	7.40	1.94	−0.77	8.70	2.76	−1.57	10.02	2.44	−0.39	10.67
	MBLA	2.73	−0.26	9.11	1.62	−0.14	8.55	2.40	−0.99	11.25	1.81	−0.05	10.75
STR_{Inf}	**EBLUP**	5.46	−0.40	6.16	8.78	−0.18	9.20	7.03	−1.10	9.61	9.78	−0.35	11.28
	KEB	4.73	0.87	6.52	8.71	1.19	9.16	5.96	0.17	9.32	9.27	1.15	11.11
	KNN	1.92	−0.57	7.29	1.50	−0.46	8.44	2.35	−1.13	9.98	2.00	−0.09	10.53
	MBLA	3.10	−0.39	9.34	2.07	−0.16	8.79	2.81	−1.26	11.73	2.21	−0.39	11.27
STR_{Ney}	**EBLUP**	5.20	−0.72	5.99	9.48	−0.83	10.31	6.89	−1.22	9.52	10.55	−0.88	11.94
	KEB	4.71	0.80	6.51	9.44	0.89	9.90	6.00	0.28	9.19	10.01	1.01	11.44
	KNN	2.00	0.45	7.53	1.55	0.69	8.57	2.42	0.06	10.13	2.02	1.08	10.63
	MBLA	2.90	−0.71	8.98	1.94	−0.82	8.80	2.81	−1.29	11.66	2.27	−0.92	11.56
SYS	**EBLUP**	5.31	−0.36	6.46	4.56	−0.17	6.98	7.06	−0.69	10.05	5.14	−0.11	10.02
	KEB	4.87	−0.38	6.38	4.70	−0.12	7.29	6.23	−0.73	9.50	5.25	−0.14	10.11
	KNN	2.31	−0.97	7.48	1.75	−0.82	8.61	2.79	−1.45	10.03	2.24	−0.60	10.58
	MBLA	2.57	−0.36	9.01	1.41	−0.17	8.50	2.46	−0.85	11.26	1.69	−0.11	10.68

MBLA = a linear model estimated for the large area; KNN = model-based estimator with predictions from a k nearest neighbours model; EBLUP = model-based estimator where the predictions were calibrated with observations from the domain; KEB = a KNN prediction calibrated with the EBLUP approach.

on the variance of the predictor variables often produced the best results regarding standard error, but worst results regarding RMSE. SRS and SYS produced the best results regarding RMSE. The differences were however small: with 100 domains SYS was on average 2.2% better than SRS with respect to RMSE in the heterogeneous domains, and 0.7% better in the homogeneous domains (Table 2). With 500 domains, the SYS was better by 1.0% in heterogeneous domains and 0.1% better in homogeneous domains than SRS.

The results were also be compared with respect to the pixel-level accuracy (Fig. 9). In both the heterogeneous and homogeneous domains and both domain sizes the average RMSEs of pixel-level predictions were smallest with SYS for MBLA and with STR_{Ney} for KNN. When calibration was possible, SYS was also the best for EBLUP and LPM for KEB. However, for EBLUP and KEB it was not possible to calculate the average RMSEs at all for all cases. This was due to the results being only available from the areas with observations, and when using Neyman allocation it meant that many of the low-end domains did not have observations in the iterations. In these areas, KNN and MBLA would need to be used. As the need for calibration is highest in the high-end domains, this is not a serious problem. In all cases, the allocation with respect to influence (STR_{Inf}) produced less accurate results than allocation with respect the variance of the variable itself (STR_{Ney}). Stratification did not improve the estimate of the whole area mean except for KNN in the heterogeneous domains but often produced biased results. Regarding to the overall mean, LPM and SYS with either EBLUP or KEB produced the least biased results.

Discussion

In model-based estimation and inference, sampling design is unimportant in a sense that any design is allowable. Yet, sampling design can influence model quality, and through that affects inferences, as the validity of inferences is solely based on the validity of the model. In this study the sampling design had a clear effect on the pixel-level RMSEs that on average increased when using stratification. Stratification also had an effect on the whole-area mean estimates, for which the stratification with Neyman allocation produced biased results. This is due to not acknowledging the sampling weights in the modelling. With proper design weights (in design-based setting) the Neyman allocation produces unbiased results (e.g. Särndal et al. 1992).

Regarding to the small area estimation, the effect of designs was on average fairly small. The largest difference among the sampling design was observed with KNN, where the stratification markedly reduced the average Se across the areas with both sample allocations (Table 2). Yet, RMSE only improved a little, and mainly in the heterogeneous domains. Contrary to this, in model-based estimation with a linear model (MBLA) the stratification and Neyman allocation based on either the variance of predictor variables or their influence (STR_{Ney} and STR_{Inf}) increased the average RMSE by 2–15% compared to SRS. LPM did not provide notable improvements over SRS. Overall, SYS was the best method, but even for SYS the improvements were typically less than 1% in RMSE. This result was somewhat surprising, as the Neyman allocation based on the influence of given predictor produces the smallest variance of the estimate of this predictor. It is evident that the potential improvements in the variances and covariances of the parameters are less important than the concentration of similar errors to the same domains, i.e. the spatial correlations. Moreover, the results were quite similar irrespective if the influence was calculated from the inner volume, the proportion of the vegetation points or from both. In preliminary tests the allocations were done both with respect to the single predictor variables and to their sum, and the variations between them were negligible.

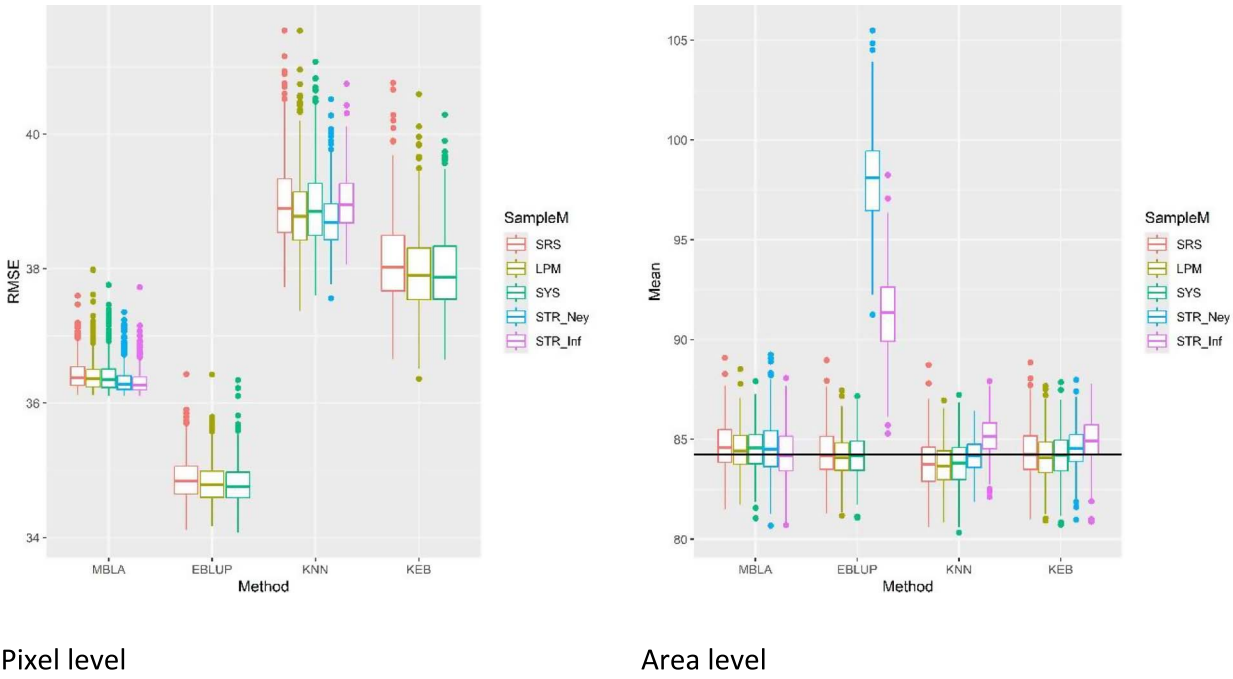

Figure 9. The average RMSE at pixel level and the whole area mean with $J = 100$ for heterogeneous domains. The results for homogeneous areas are identical except for the calibration for the area-effect.

Since the Neyman allocation either according to a variance of predictor (STR_{Ney}) or its influence (STR_{Inf}) heavily weighted the strata with largest values of the predictors, the results improved mostly with domains with highest true mean volume. With LPM or stratification using proportional allocation (results not shown) the results at these high-volume domains did not improve on average, or only improved very little. The mean volumes in these domains were generally heavily underestimated, especially when the domains were small and homogeneous. Thus, even though the linear model is unbiased for the whole area, the model-based estimates in part of the domains were clearly biased. This can be seen as an example of the design-bias in the model-based estimation (Ståhl et al. 2024). For the cases when there were observations that could be used for calibration (EBLUP and KEB) the underestimation was to some extent lower with the Neyman allocation.

However, the smaller underestimation at the high-end domain came with a cost in domains with a small or moderate mean volume: in these domains the RMSE was highest with the stratification and Neyman allocation. In these domains there also was a limited opportunity to calibrate with observations, as the number of units selected from such domains was very small. This also reduces the efficiency of EBLUP and KEB estimators in such domains, which can be seen as higher RMSE (e.g. Fig. 6). Even more clearly it can be seen from the empirical coverages of the empirical confidence intervals (Figs. 3 and 5), which gave especially poor results in the low-end domains. Overall, using EBLUP calibration whenever it is possible seems advisable, confirming the earlier results with larger small domains (Breidenbach et al. 2016, Magnussen and Breidenbach 2017, Frescino et al. 2022, Kangas et al. 2024).

The importance of the spatial correlations within the domains was very clear seen in the empirical coverages. When the empirical coverages were calculated ignoring the spatial correlations, i.e. only based on the parameter errors, the empirical coverages of MBLA dropped below 0.6 already when the mean volume in the domain was 50 m^3/ha with $J = 100$ and homogeneous division of domains. With 500 domains and homogeneous domains, the empirical coverage dropped below 0.25 when the true means of the domains were over 100 m^3/ha. Thus, the assumption of the within-domain spatial correlation is the decisive factor regarding the successful estimation of the accuracy of the small area estimates (cf. Kangas et al. 2024 for KNN).

Part of the importance of the spatial correlation obviously comes from the assumption of a correlation of the errors in the true model, which was used to generate the population used in this study. However, the spatial correlations were important even if the errors of the true model were assumed independent. Assuming independent errors of the true model, the within-area correlation for the heterogeneous domains varied from 0.031 ($J = 100$) to 0.066 ($J = 500$) and for the homogeneous domains from 0.160 ($J = 100$) to 0.253 ($J = 500$). This can be explained by the fact that the true model had seven explanatory variables and the model used in the study had only three, even though the difference in R^2 was quite small (0.897 versus 0.87). The missing predictors compared to the 'true' model can thus introduce spatial correlation into the predictions even if the error of the 'true' model were assumed independent.

The smaller correlations caused smaller average RMSEs. For instance with heterogeneous domains and $J = 100$ the RMSE reduced from 8.55 to 5.24 and with homogeneous domains from 9.11 to 8.73 (average results are shown in supplementary material Table S1). The smaller correlation also caused smaller estimated variances, narrower CIs and thus reduced the empirical coverages. This was especially clear in the heterogeneous domains, where e.g. the average empirical coverage of linear model with SRS reduced from 93.4% to 86.7% (Average empirical coverages are shown in the supplementary material Table S1). Thus, the correlation assumption was decisive also when the true model errors were assumed independent when generating the true population.

It can be concluded that heavy weighting of the high-end predictor values improves the predictions and reduces the bias in the domains with high volumes, but with high risk of poor results elsewhere. Thus, if the main purpose of the inventory is to be consistently good through all domains, SYS can be recommended, but if the purpose is to locate domains with high volumes, the stratified designs using Neyman allocation might be useful. Since the homogeneity of the domains has a clear effect on the results, it can be assumed that the bias with respect to the high-end volumes would be even more pronounced, if the size of the domains were further reduced and they were even more homogeneous. Such could be the case with forest stands.

Therefore, it would be very important to be able to improve the predictions in the high-end domains. However, as the models never really capture the exceptional values, it can be assumed that improvements with better modelling techniques are likely to be small. If the poor results with STR_{Inf} and MBLA were partly due to prevailing non-linearities in the relationships, other type of modelling method than linear models (like boosted trees or copula prediction models) could possibly improve the results. For instance, Toivonen et al. (2024) found that tree boosting with random effects improved notably the results in the high-end tail of stand age. Cheng et al. (2025) used copula regression, and that also is promising regarding to the high-end values.

In practical forest inventory, systematic sampling often used in National Forest Inventories appears to be a good choice regarding the small-area estimation. While stratified sample with Neyman allocation is optimal for the large-area estimation, for small domains it may be a risky option. The best option for modelling could be a stratified sampling, where part of the sample were equally allocated to strata, and the rest were allocated using Neyman allocation. This would mean that there are some observations also for the low-end volume domains, but plenty of observations from the high-end volume domains for EBLUP calibration where the calibration is most effective.

It seems that calibration, i.e. obtaining an accurate estimate of the area-effect for all domains, is the likeliest way to improve the results in the high-end domains. It is also highly unlikely that NFI plots would be available for all small domains, especially if they are as small as stands. One possibility for this could be that the area-effects of the sampled domains were modelled and then predicted for the non-sampled domains. For instance, Kilkki and Lappi (1987) predicted the random effects of taper curve using regressions, and Saei and Chambers (2005) estimated the area-effects using the spatial correlation between domains. The potential of such approaches remains to be studied in the future.

Conclusion

Model predictions as expected values conditioned on predictors never can capture the exceptionally large true values of the response variable. Therefore, it is clear that model-based estimation does not fit well for the cases where the purpose is to locate the small areas with high values, like most valuable stands in stand-level inventory. Having field data for calibration using EBLUP helps and is thus highly recommendable whenever it is feasible. Sampling designs providing some sample units for most small areas, but yet assigning more calibration data for the most interesting small areas could be the optimal approach. Such approach could be used to improve the results in high-end domains. In the low-end and middle-end domains the sampling designs had minor effect. When the domains are as small and homogeneous as stands, improvements in the prediction of the area effects might be the optimal solution.

Author contributions

Annika Susanna Kangas (Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Software, Writing—original draft), Mari Myllymäki (Funding acquisition, Methodology, Writing—review & editing), and Petteri Packalen (Funding acquisition, Methodology, Writing—review & editing)

Supplementary data

Supplementary data are available at *Forestry* online.

Conflict of interest: None declared.

Funding

The study was supported by the Research Council of Finland through the project 'Is climate smart forestry a utopia if the preferences of landowners are not considered? (UTOPIA)' under Grant 352782, the European Union Horizon Europe (HORIZON) Research & Innovation programme under the Grant Agreement no. 101056907, and the Research Council of Finland's flagship 'Forest-Human-Machine Interplay – Building Resilience, Redefining Value Networks and Enabling Meaningful Experiences' (UNITE) (Grant number 357909).

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Astrup R, Rahlf J, Bjørkelo K. *et al.* Forest information at multiple scales: Development, evaluation and application of the Norwegian forest resources map SR16. *Scandinavian Journal of Forest Research* 2019;**34**:484–96. https://doi.org/10.1080/02827581.2019.1588989.

Balazs A, Liski E, Tuominen S. *et al.* Comparison of neural networks and k-nearest neighbors methods in forest stand variable estimation using airborne laser data. *ISPRS Open Journal of Photogrammetry and Remote Sensing* 2022;**4**:100012, 1–15.

Bell DM, Wilson BT, Werstak CE Jr. *et al.* Examining k-nearest neighbor small area estimation across scales using National Forest Inventory Data. *Front For Glob Change* 2022;**5**:763422. https://doi.org/3389/ffgc.2022.763422.

Breidenbach J, Magnussen S, Rahlf J. *et al.* Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. *Remote Sens Environ* 2018;**212**:199–211. https://doi.org/1016/j.rse.2018.04.028.

Breidenbach J, McRoberts RE, Astrup R. Empirical coverage of model-based variance estimators for remote sensing assisted estimation of stand-level timber volume. *Remote Sens Environ* 2016;**173**:274–81. https://doi.org/1016/j.rse.2015.07.026.

Brus DJ. Sampling for digital soil mapping: A tutorial supported by R scripts. *Geoderma* 2019;**338**:464–80. https://doi.org/1016/j.geoderma.2018.07.036.

Chen T, Lumley T. Optimal sampling for design-based estimators of regression models. *Stat Med* 2022;**41**:1482–97. https://doi.org/1002/sim.9300.

Cheng X, Hou Z, Kangas A. *et al.* Alleviating small sample problem in continuous Forest monitoring with remote sensing-assisted copulas. *Ecol Indic* 2025;**171**:113132. https://doi.org/1016/j.ecolind.2025.113132.

Cochran WG. *Sampling Techniques*. 3rd edition. New York: Wiley, 1977.

Crookston NL, Finley AO. yaImpute: An R package for kNN imputation. *J Stat Softw* 2007;**23**:1548–7660. https://doi.org/18637/jss.v023.i10.

Frescino TS, McConville KS, White GW. *et al.* Small area estimates for National Applications: A database to dashboard strategy using FIESTA. *Front For and Glob Change* 2022;**5**:779446. https://doi.org/3389/ffgc.2022.779446.

Grafström A, Saarela S, Ene L. Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Can J For Res* 2014;**44**:1156–64. https://doi.org/1139/cjfr-2014-0202.

Haakana H, Katila M, Heikkinen J. *et al.* Precision of exogenous post-stratification in small area estimation based on a continuous national forest inventory. *Can J For Res* 2020;**50**:359–70. https://doi.org/1139/cjfr-2019-0139.

Haralick RM, Shanmugam K, Dinsteein J. Textural features for image classification. *IEEE Trans Syst Man Cybern* 1973;**3**:610–21.

Junttila V, Maltamo M, Kauranne T. Sparse Bayesian estimation of Forest stand characteristics from airborne laser scanning. *For Sci* 2008;**54**:543–52. https://doi.org/1093/forestscience/54.5.543.

Kangas A, Myllymäki M, Mehtätalo L. Understanding uncertainty in forest resources maps. *Silva Fennica* 2023;**57**:22026.

Kangas A, Myllymäki M, Packalen P. Small area composite estimators in a simulation test. *Can J For Res* 2025;**55**:1–17. https://doi.org/10.1139/cjfr-2024-0070.

Kilkki P, Lappi J. Estimation of taper curve using stand variables and sample tree measurements. *Scand J For Res* 1987;**2**:121–6. https://doi.org/1080/02827588709382451.

Kotivuori E, Kukkonen M, Mehtätalo L. *et al.* Forest inventories for small areas using drone imagery without in-situ field measurements. *Remote Sens Environ* 2020;**237**:111404. https://doi.org/1016/j.rse.2019.111404.

Lumley, Thomas based on fortran code by Alan Miller. 2024. *The ´Leaps´ Package*. Regression Subset Selection. https://cran.r-project.org/web/packages/leaps/leaps.pdf.

Magnussen S, Breidenbach J. Model-dependent forest stand-level inference with and without estimates of stand-effects. *Forestry: An International Journal of Forest Research* 2017;**90**:675–85. https://doi.org/1093/forestry/cpx023.

Mäkisara K, Katila M, P J. The multi-source national forest inventory of Finland – Methods and results 2015. *Natural Resources and Bioeconomy Studies*, Vol. **8**. Helsinki: Natural Resources Institute Finland (Luke), 2019.

Maltamo M, Packalen P. Species specific management inventory in Finland. In: Maltamo M, Naesset E, Vauhkonen J, (eds.). *Forestry Applications of Airborne Laser Scanning –Concepts and Case Studies. Managing Forest Ecosystems*, Vol. **27**. Switzerland: Springer, 2014, 241–52 https://doi.org/1007/978-94-017-8663-8_12.

Maltamo M, Packalen P, Kangas A. From comprehensive field inventories to remotely sensed wall-to-wall stand attribute data—a brief history of management inventories in Nordic countries. *Can J For Res* 2021;**51**:257–66. https://doi.org/1139/cjfr-2020-0322.

Mauro F, Monleon VJ, Temesgen H. *et al.* Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. *PloS One* 2017;**12**:e0189401. https://doi.org/1371/journal.pone.0189401.

Mehtätalo L, Lappi J. *Biometry for Forestry and Environmental Data*. Boca Raton: CRC Press, 2020, 411 https://doi.org/1201/9780429173462

Militino AF, Ugarte MD, Goicoa T. A BLUP synthetic versus an EBLUP estimator: An empirical study of a small area estimation problem. *Journal of Applied Statistics* 2007;**34**:153–65. https://doi.org/1080/02664760600994893.

Neyman J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* 1934;**97**:558–625. https://doi.org/2307/2342192.

Nothdurft A, Saborowski J, Breidenbach J. Spatial prediction of forest stand variables. *Eur J Forest Res* 2009;**2009**:241–51.

Rao JNK, Molina I. *Small Area Estimation*. 2nd edition. New Jersey: JohnWiley & Sons, Inc., 2015 https://doi.org/1002/9781118735855

Royall RM. On finite population sampling theory under certain linear regression models. *Biometrika* 1970;**57**:377–87. https://doi.org/1093/biomet/57.2.377.

Royall RM, Herson JH. Robust estimation in finite populations I and II. *J Am Stat Ass* 1973;**68**:880–9 https://doi.org/1080/01621459.1973.10481440.

Saei A, Chambers R. *Out of Sample Estimation for Small Areas Using Area Level Data (S3RI Methodology Working Papers, M05/11)*, Vol. **23**. Southampton, UK: Southampton Statistical Sciences Research Institute, University of Southampton, 2005.

Ståhl G, Gobakken T, Saarela S. *et al.* Why ecosystem characteristics predicted from remotely sensed data are unbiased and biased at the same time – And how this affects applications. *Forest Ecosystems* 2024;**11**:100164. https://doi.org/1016/j.fecs.2023.100164.

Särndal CE, Swensson B, Wretman J. 1992. Model assisted survey sampling. Springer-Verlag. 694 p.

Toivonen J, Kangas A, Pitkänen TP. *et al. Forest Age Prediction in Northern Finland Using Airborne Lidar and Satellite Images*, 2024 Preprint. Available at SSRN 4886824.

Tuominen S, Pitkänen T, Balázs A. *et al.* Improving Multi-Source National Forest Inventory by 3D aerial imaging. *Silva Fennica* 2017;**51**. https://doi.org/10.14214/sf.7743.

Véga C, Renaud J-P, Durrieu S. *et al.* On the interest of penetration depth, canopy area and volume metrics to improve Lidar-based models of forest parameters. *Remote Sensing of Environment* 2016;**175**:32–42. https://doi.org/10.1016/j.rse.2015.12.039.

Wadoux AC, Heuvelink GBM. Uncertainty of spatial averages and totals of natural resource maps. *Methods in Ecology and Evolution* 2023;**14**:1320–32. https://doi.org/1111/2041-210X.14106.

Forestry: An International Journal of Forest Research, 2025, 1–13
https://doi.org/10.1093/forestry/cpaf076
Original Article