Canadian Journal of Forest Research

OPEN ACCESS | Research Article

Model-based small-area estimation with area-effects for sampled and non-sampled domains

Annika Kangas[a], **Mari Myllymäki**[b], **and Petteri Packalen**[b]

[a]Natural Resources Institute Finland (Luke), Bioeconomy and environment, Yliopistokatu 6, 80101 Joensuu, Finland; [b]Natural Resources Institute Finland (Luke), Bioeconomy and environment, Latokartanon kaari 9, 00790 Helsinki, Finland

Corresponding author: **Annika Kangas** (email annika.kangas@luke.fi)

Abstract

Previous studies recommend the empirical best linear unbiased predictor (EBLUP) for small-area estimation. However, EBLUP estimation requires at least one observation from each small area, while most of the areas may be non-sampled. One approach to overcome this problem is to predict the area-effects for the non-sampled areas with a model developed using the estimated area-effects from the sampled areas. Another approach is to cluster the small areas to larger groups and introduce a cluster-effect into the prediction model. We tested these approaches in a set of simulated small areas (domains). When observations from all or most domains were available, EBLUP with a domain-effect, or combined cluster- and domain-effect were the most reliable calibration methods. When the sampling fraction and the size of the domains were smaller, calibrating with the cluster-effect only was the most reliable method. Without any calibration, the model-based estimates for the domains with the highest volumes were severely underestimated. When observations were available, the EBLUP calibration improved the results in the high-end of the distribution. With the smallest sampling fractions and domains, also the predicted area-effects reduced the underestimation. However, the modelled area-effects were estimated from the population data, rather than from a sample.

Key words: mixed model, area-effect, group-effect, EBLUP, non-sampled area

1. Introduction

In Finland, forest management inventory for stand-level decision-making is model-based small-area estimation based on stratified sampling of field plots, aerial laser scanning data, and aerial images (Maltamo et al. 2021). The pixel-level results can be calculated using either parametric models (Astrup et al. 2019) or non-parametric models (Maltamo and Packalen 2014). The stand-level results are means and sums of predictions at pixels belonging to the stand. Likewise, model-based small-area estimation is used also in larger scales, to calculate estate-level or municipality-level inventory results. In what follows, all small areas are called domains, irrespective if the size of these areas.

In several studies it has been noted that empirical best linear unbiased predictor (EBLUP) using either an area-level or a unit-level mixed model is a good option for small-area estimation (Breidenbach et al. 2016; Magnussen and Breidenbach 2017; Frescino et al. 2022; Kangas et al. 2025*a*, 2025*b*). This means assuming constant within-area correlations of the model prediction errors (Magnussen and Breidenbach 2017; Astrup et al. 2019). It would also be possible to assume continuous correlations and utilize kriging or block-kriging estimates (Wadoux and Heuvelink 2023). However, approximation with fixed area-effect is often sufficient.

The area-effects model can also be seen as a composite estimator, where the initial sample statistic within the small domain is adjusted using an EBLUP prediction for the unsampled units (Militina et al. 2007). Similar ideas can also be used in the context of k nearest neighbors (KNN) modelling. For instance, Bell et al. (2022) made a bias correction to KNN small-area estimates by calculating a design-weighted mean of the observed errors. Kangas et al. (2025*b*) and Nothdurft et al. (2009) used a mixed model in combination with a KNN to calibrate the predictions. Another alternative to the area-effect model is a hierarchical Bayesian model, where the area-effect is replaced with prior information regarding the targeted domains (see White et al. 2021).

To utilize the area-effects for calibration, at least one observation should be available from each domain. According to Magnussen and Breidenbach (2017) more than one observation should be available from all domains. However, it is highly unlikely that observations would be available for all domains, if they are very small, such as forest stands. Typically, not even all municipalities have observations in the Finnish National Forest Inventory (NFI). Thus, there is a need to develop methods by which the calibration can also be used for non-sampled domains.

One possibility to overcome the problem of lacking calibration data 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. However, the problem of this approach is that the true area-effect is unknown, and the area-effects needed for such a model would have to be estimated from a sample, typically meaning fairly small number of observations. Thus, the resulting model would include also the estimation uncertainty of the random effects. Moreover, if there is a correlation between the random area-effect and an available explanatory variable, such variable likely should be included in the model as a fixed effect in the first place. Calibration with a prediction model might be a reasonable option, if the model for the random effects comes from independent data, for instance from another area.

A second possibility would be to introduce a multivariate model for two or more forest attributes (e.g., volume and height), and to utilize the secondary attributes through cross-model correlations of errors to estimate the area-effect of the key attribute of interest (e.g., Burgard et al. 2021). For instance, if both volume and height were modelled based on same data, then observed errors in height predictions could be used to estimate the area-effect for volume model and vice versa. That approach, in turn, would require that there are additional observations from each of the domains of the height, and that is unlikely. It might be possible to utilize as the additional variable e.g., a height observed using single-tree detection with (independent) photogrammetric or lidar point-cloud data, but the accuracy might not be sufficient for cross-model calibration. Using multiple response variables in the modelling can also in itself improve the reliability of (area-level) small-area estimation (Georgakis et al. 2025 and references therein).

A third possible strategy would be to assume that the nearby domains share similarities, meaning that the area-effects of the nearby domains are spatially correlated. For instance, Saei and Chambers (2005), Sikov and Cerda-Hernández (2023), and Benedetti et al. (2024) estimated the area-effects using the spatial correlation between domains. Chung and Datta (2022) proposed a Bayesian spatial model for the area-effects. To be useful, this approach would require that there are between-domain correlations in addition to within-domain correlations as assumed in the area-effect approach. In addition, there should be sampled domains nearby each non-sampled domain, as the spatial correlations typically have a short range (Breidenbach et al. 2016).

The last option found from the literature is to cluster similar domains to groups and use the cluster-effect to calibrate the non-sampled areas (e.g., Anisa et al. 2014; Desiyanti et al. 2023). This approach requires that the domains are similar enough that a shared cluster effect would improve the results. The performance of the method is likely to depend on the number and the size of clusters in relation to the number of domains, and the information available for clustering.

The aim of this paper is to test two options of estimating the area-effect for both sampled and non-sampled units, namely utilizing the predicted domain-effects and introducing the cluster-effects. We will test the performance of domain-level estimation is simulated dataset utilizing calibration with domain- and cluster-effects and with predicted domain-effects. We will also test the effect of (1) the size of clusters, (2) the variables used to cluster the domains, (3) the homogeneity of the domains, and (4) the sampling fraction on the performance of these different approaches.

2. Materials

We made a simulation experiment to test the performance of the small-area estimators. For this purpose, we utilized wall-to-wall data on a region of size about 5900 ha, a small part of an earlier airborne laser scanning (ALS) campaign. This data has been used also in other simulation studies (see Kangas et al. 2023, 2025*a*, 2025*b*). This test data were available on a grid of 231 824 population elements of size 16 m × 16 m, for which coordinates, and 17 ALS features are available. To have 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 simulated 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 m (see details from Kangas et al. 2023; Kangas et al. 2025*b*).

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, as well as a set of 190 ALS features (see Tuominen et al. 2017 and Balazs et al. 2022 for details). 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, 2020), using the ALS features available for the whole grid (Table 1). This external model had residual standard error = 0.232 m^3/ha and multiple $R^2 = 0.897$ (Kangas et al. 2023). This seven-predictor model was used only in the generation of the simulated population, not in the estimators used in the simulation test (Table 2).

3. Methods

3.1. Model-based small-area estimators using mixed linear models

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

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

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 fixed effects in domain j, $\boldsymbol{\beta}$ is the vector of fixed model coefficients, v_j is the random area-effect (later called domain-effect) for domain j ($v_j \sim N(0, \sigma_v^2)$), e_{ji} is the residual error in 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 (in case weighting is needed e.g., for heteroscedastic residuals) and n_j is the sample size within domain j. In matrix

Table 1. The available airborne laser scanning features in the population data.

Name	Explanation
HMax	The maximum height of the points (m, first echo)
Hq45_f, Hq55_f, Hq70_f, Hq90_f, Hq95_f	Height at which given percentiles (45%, 65%, 70%, 90%, and 95%) of first echos of vegetation points are accumulated (m)
Hq20_l	Height at which given percentiles (20%) of last echos of vegetation points are accumulated (m)
Pveg_f	Proportion of vegetation points relative to all points (%, first echo)
Pveg_l	Proportion of vegetation points relative to all points (%, last echo)
Iskew_f	Skewness of the vegetation point heights (first echo)
PAbov	Proportion of points above mean height
P20_l	Proportion of points having cumulated at 20% of the height from all points (%, last echo)
Rmpl_ind	Rumple index
Volin	Inner volume (Vega et al. 2016)
CSum	Sum Entropy (Haralick et al. 1973) of canopy surface model
I_mean	Average intensity of ALS echoes

Table 2. 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
HMax	0.02380686	0.007498333	3.174953
Hq55_f	0.02799505	0.01001328	2.795793
Hq90_f	0.03482004	0.01273452	2.734304
Pveg_f	0.0104878	0.0005722577	18.32706
P20_l	0.02489417	0.00371768	6.696159
I_mean	−0.05024336	0.007689848	−6.533726
CSum	0.3666484	0.02072759	17.68891

Note: The definitions of the variables are given in Table 1.

form this is

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}v + \mathbf{e} \quad (2)$$

where $\mathbf{Z}$ is an (N x 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)

$$\boldsymbol{V} = \sigma_v^2 \mathbf{Z}\mathbf{Z}^{\mathbf{t}} + \sigma_e^2 \mathbf{C}^{-1} \quad (3)$$

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

$$\text{cor}\left(e_{ji}, e_{jk}\right) = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2} \quad (4)$$

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

$$\mu_j = \bar{\boldsymbol{x}}_{\boldsymbol{j}}\boldsymbol{\beta} + v_j + \frac{1}{N_j}\sum_{i=1}^{N_j} e_{ji} \quad (5)$$

where $\bar{\boldsymbol{x}}_{\boldsymbol{j}}$ is the population mean vector of the values $\mathbf{x}_{\boldsymbol{ji}}$ in domain j, N_j is the total number of units within the domain j, v_j is the domain-effect, and e_{ji} is the residual error in unit i within domain j. In eq. 5, the domain-effect vector v is assumed to be estimable from data. However, if there are no observations from the domain (i.e., $n_j = 0$), the estimate of domain-effect $\widehat{v}_j = 0$. 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

$$\widehat{\mu}_{\text{MB},j} = \bar{\boldsymbol{x}}_{\boldsymbol{j}}\widehat{\boldsymbol{\beta}} \quad (6)$$

When this estimator is calibrated using the observations from the domain j to estimate the domain-effect $\widehat{v}_j$, the estimator of the mean is of the form (Mauro et al. 2017)

$$\widehat{\mu}_{\text{EBLUP},j} = \bar{\boldsymbol{x}}_{\boldsymbol{j}}\widehat{\boldsymbol{\beta}} + \widehat{v}_j \quad (7)$$

In this study, the domains were clustered to groups of similar domains based on the explanatory variables available. Then, it was possible to estimate a model with a cluster-effect and a nested domain-effect as

$$y_{hji} = \mathbf{x}_{hji}\boldsymbol{\beta} + \gamma_h + v_{hj} + e_{hji},\ h = 1, \ldots, H \quad (8)$$
$$\text{j} = 1, \ldots, J_h;\ i = 1, \ldots, n_{hj}$$

where y_{hji} is the volume in unit i within domain j (m^3/ha) located within cluster h, $\mathbf{x}_{\boldsymbol{hji}}$ is the vector containing all observed predictor values for fixed effects in domain j and cluster h, $\boldsymbol{\beta}$ is the vector of fixed model coefficients, γ_h is the random cluster-effect for cluster h ($\gamma_h \sim N\left(0, \sigma_\gamma^2\right)$), v_{hj} is the random domain-effect for domain j in cluster h ($\upsilon_{hj} \sim N\left(0, \sigma_\upsilon^2\right)$), and e_{hji} is the residual error of unit i in domain j located within cluster h ($e_{hji} \sim N\left(0, \sigma_e^2\right)$). With this model (eq. 8), the cluster-effect can be used in calibration if observations are available from a cluster but not from the domain, and both cluster and domain effects can be used if observations are

available both from the cluster and the domain as

$$\widehat{\mu}_{\mathrm{EBLUP},hj} = \bar{\boldsymbol{x}}_{\boldsymbol{hj}}\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_h + \widehat{v}_{hj} \quad (9)$$

The fixed part of the model used for calculating the results was estimated with leaps R package (Thomas Lumley based on Fortran code by Alan Miller, 2024). It was the three best predictors model with Proportion of vegetation points relative to all points (%, first Pveg_f and last Pveg_l echo), and Inner volume (Volin) as predictors (Table 1). The RMSE of this model, calculated from all units in the population, was 36.11 m3/ha and R2 0.87. From the three predictors, only Pveg_f was a shared variable with the "true" model (Table 2).

3.2. Generation of small domains and clusters of domains

To have spatially contiguous and fairly homogeneous small areas mimicking forest stands or forest estates, we divided our study large area of 5900 ha to J domains with a k-means clustering approach. In the first set of domains, we used coordinates of the pixels and the maximum height (HMax, Table 1) variable to cluster the pixels to get somewhat homogeneous small domains. To analyse the effect of homogeneity, we generated also another set of domains, where in addition to the previous variables, we used the Sum Entropy (CSum) to create more homogeneous small areas. We used values of $J = 500$ (on average about 12 ha domains) and $J = 2500$ (on average about 2.5 ha domains). These different options for dividing the pixels to domains are gathered in Fig. S1 in supplementary material.

The J domains were further clustered to $H = 30$ clusters for utilizing the cluster-effects (Fig. S1 in supplementary material). We used as clustering variables the same variables that were used to form the small domains (HMax, CSum) as well as the size of the small domains. In addition, we tested the performance of clustering with one of the fixed predictors of the model, namely Inner Volume (Volin).

3.3. Predicting the domain-effects using a model

The "true" domain effects for each domain were estimated using the general EBLUP estimation procedure of lme4 R-package (Bates et al. 2015). These domain-effects were treated as "true" effects, as they were estimated from a model fitted to the whole population data (i.e., not from a sample). They were estimated separately for each four sets of small domains and modelled as a function of domain means of available predictors as

$$v_j = \bar{\boldsymbol{x}}_{\boldsymbol{j}}\boldsymbol{b} + \boldsymbol{e}_{\boldsymbol{j}} \quad (10)$$

The predictors $\boldsymbol{x}_j$ in model (10) can be either same, partly same or different than variables used for unite-level models (1–9). Since the models here predict the "true" domain-effects, they form an upper limit to what is achievable with a prediction model with these predictors in this dataset. In a more realistic setting, these domain-effects would have to be estimated from a sample, and the resulting models would therefore include more uncertainty.

3.4. Simulation experiment

For the analysis, we first used a sample size m of 5000, meaning an expected number of observations per domain to be 10 ($J = 500$) or 2 ($J = 2500$). To analyze the effect of sampling fraction, we calculated the results also with $m = 1000$ (expected number of observations per domain 2 or 0.4) and $m = 500$ (the expected value of observations per domain 1 or 0.2). With the $m = 1000$ sample size we also tested the effect of the number of clusters using $H = 10$ and 50. The sample was selected either with simple random sampling (SRS) or pseudo-systematic sampling (SYS, Brus 2019). The systematic sampling was carried out so that the area was divided to 1000 blocks using the coordinates, and from each block 5 or 1 observations were randomly selected. For $m = 500$, the number of blocks was also set to 500.

The observations available from the domains and the clusters were then used in the calibration of the area results. The tested calibrations were:

NoE = Not calibrated, only fixed effects model prediction used (eq. 6).

DomE = Calibrated with a domain-effect for sampled domains and fixed effect prediction for non-sampled domains (eq. 7).

CluE = Calibrated with cluster-effect for sampled clusters, and fixed effect prediction for non-sampled clusters (eq. 9).

CDE = Calibrated with both cluster-effect and domain-effect for sampled domains or with a cluster-effect for non-sampled domains (eq. 9).

PreE = Calibrated with a regression prediction of a domain-effect (eq. 10).

3.5. Evaluation of the results

For each of the three cluster size options, two sampling design options, three sample size options, two domain size options, and two homogeneity options we simulated $S = 1000$ realizations. We estimated for each method and every domain j the mean and standard error with the corresponding estimators. Then, we estimated the bias as the average difference between estimated domain means $\widehat{\bar{y}}_{sj}$ and the actual domain mean μ_j from samples $s = 1, \ldots, S$, i.e.,

$$\mathrm{bias}_j = \frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\bar{y}}_{sj} - \mu_j\right) \quad (11)$$

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

$$\mathrm{SE}_j = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(\widehat{\bar{y}}_{sj} - \overline{\widehat{\bar{y}}_{sj}}\right)^2} \quad (12)$$

and the true RMSE as

$$\mathrm{RMSE}_j = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(\widehat{\bar{y}}_{sj} - \mu_j\right)^2} \quad (13)$$

Table 3. The model for predicting the true domain effects in a case of 2500 domains formed based on HMax and coordinates or based on HMax, CSum, and coordinates.

	Maximum height			Maximum height and Sum Entropy		
Variable	Estimate	Standard error	*t*-value	Estimate	Standard error	*t*-value
Intercept	44.205	2.204	20.058	3.276	1.836	1.784
HMax	4.149	0.138	30.035	3.452	0.115	30.072
PAbov	−1.040	0.039	−26.708	−0.866	0.038	−23.074
CSum	−12.561	0.618	−20.326	−2.578	0.446	−5.785
Size	−0.010	0.014	−0.763	0.047	0.013	3.493

Table 4. The average domain-level results with large and small domains, formulated using maximum height (HMax) or maximum height and Sum Entropy (CSum), sample size 5000, and the cluster number $H = 30$ using simple random sampling.

		Large domains ($J = 500$)			Small domains ($J = 2500$)		
Clustering	Cal	Se (m^3/ha)	Bias (m^3/ha)	RMSE (m^3/ha)	Se (m^3/ha)	Bias (m^3/ha)	RMSE (m^3/ha)
HMAX	NoE	1.44	−0.35	12.34	1.08	−0.54	12.43
	DomE	6.40	−0.35	7.25	7.08	−0.54	10.60
	CluE	2.08	**−0.05**	9.47	2.10	**−0.17**	11.11
	CDE	5.79	**−0.05**	**6.75**	6.51	**−0.17**	**9.75**
	PreE	**1.40**	−0.20	13.34	**1.05**	−0.36	15.06
CSUM	NoE	1.31	−0.10	9.87	1.05	−0.56	11.53
	DomE	4.95	−0.10	5.80	6.47	−0.56	9.90
	CluE	1.73	**−0.04**	7.75	1.95	**−0.10**	10.20
	CDE	4.55	**−0.04**	**5.41**	5.97	**−0.10**	**9.01**
	PreE	**1.22**	0.36	12.39	**0.96**	0.64	15.75

Note: The best average results for both HMax and CSum are bolded.

4. Results

4.1. Models for domain-effects

The "true" domain effects for the case of $J = 500$ and domains formed using the coordinates and HMax variable are shown in the supplementary material (Fig. S2). The positive effects (describing the underestimates of the domain means) were at largest about 120 m^3/ha, and the negative effects (describing overestimates of the domain means) about −40 m^3/ha. Thus, even when the model was estimated from the population data rather than a sample, the un-calibrated model produced large over- and underestimates of the true domain means.

The model for predicting these "true" domain effects (eq. 10) as a function of the domain mean values of predictor variables included as predictors HMax, CSum, Size, and PAbov. Of these, the HMax and CSum were also used for forming the homogeneous small domains and Size to forming the clusters of small domains. When the domains were formed with the maximum height and coordinates (HMax), the model predicting the true domain effects was able to explain 27% of the variation with $J = 2500$ domains (Table 3). When more homogeneous domains were formed using in addition to maximum height also the Sum Entropy (CSum), the model was able to predict 31% of the variation. Thus, it was easier to predict the domain-effects when the domains were more homogeneous. The homogeneity of the domains also had an effect on the calibration results.

In the selected model, PAbov was additional information not utilized in the unit-level models nor in the clustering (eqs. 1 - 9). If the predictor PAbov was removed, the explained variation dropped from 31% to 17% with the more homogeneous domains, and from 27% to 8% with the less homogeneous domains. Thus, the efficiency of this approach is dependent on having additional information to the original model for the predictions.

4.2. Performance of different calibration approaches

When the domains were formed based on maximum height and coordinates, and the clusters of domains with maximum height and size of the domains (HMax), the best option in terms of minimizing the standard error was the PreE, for bias either the CE or the CDE, and for RMSE the CDE (Table 4). With the more homogeneous domains (CSum), the results were very similar in terms of the ranking of the calibration methods. With the small domains ($J = 2500$), the standard errors for CSum domains were 2%–9% smaller than with the less homogeneous HMax domains and with the large domains ($J = 500$) 4%–23% smaller (Table 4). Even in the NoE case the standard error thus was reduced a little, as the estimation of the domain-effects also affects to the estimation of the fixed effects. Moreover, also the RMSE was 6%–8% smaller for CSum domains than for HMax domains in the small domains and 18%–20% smaller in the large domains except with the PreE predicted domain-effect. In CSum domains the PreE

Fig. 1. The bias as the function of the true mean volume of domains with the different predictor-options, domains formed with maximum height (HMax), sample design simple random sampling, sample size $m = 5000$, and cluster number $H = 30$.

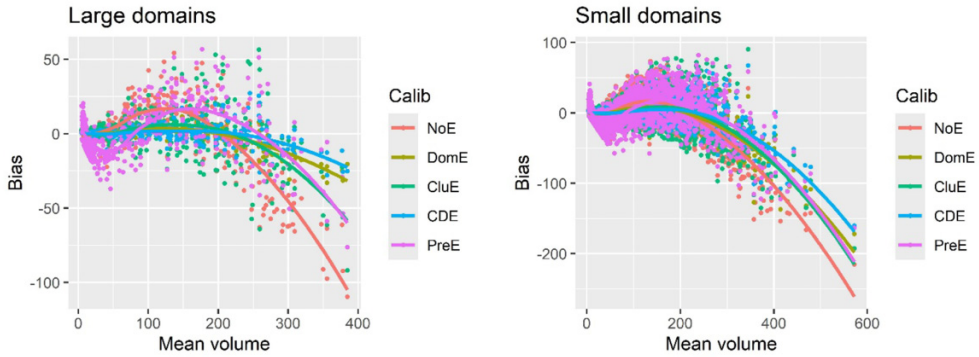
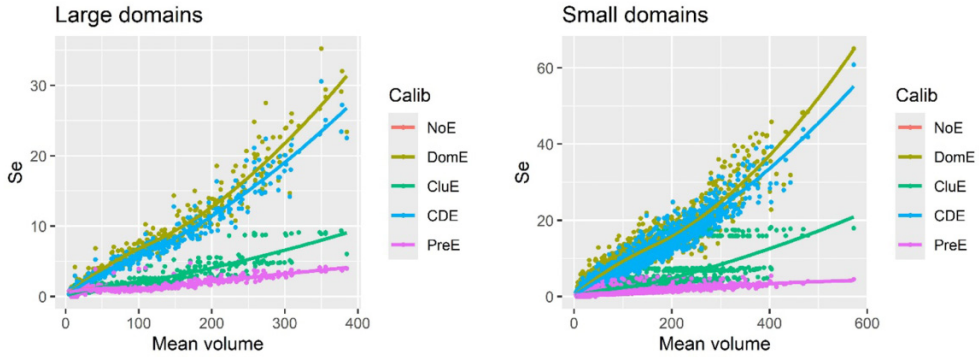


Fig. 2. The standard error as the function of the true mean volume of domains with the different predictor-options, domains formed with maximum height (HMax), sample design simple random sampling, sample size $m = 5000$, and cluster number $H = 30$.



improved the results 7% compared to HMax domains in the large domains, but worsened them 4% for the small domains, even though the model used for CSum domains explained more of the variation (Table 4).

For both large and small domains, the cluster-effect improved the average RMSEs, but using the domain-effects when that was possible, was the best option (Table 4). The main difference between the small and large domains was that the mean bias was higher in the smaller domains, and the relative improvement available from calibration was on average smaller. The differences between the sampling designs were small, and therefore the results for systematic sampling are presented only in the supplementary material (Table S1).

A quite surprising result was that the average standard error was always higher with the calibrated options than with the non-calibrated options (Table 4). It appears that even when the expected number of observations from each domain was as high as 10, the number of observations was not high enough to provide stable improvements. Thus, the calibration increased the variation among estimates from the simulated samples for the same domains rather than reduced it.

When looking how the model bias behaves, the situation is different. With both domain size options and domains formulated using maximum height (HMax), the largest bias in the high-end volume domains occurred with a non-calibrated model NoE (Fig. 1). With the larger domains ($J = 500$), the domain-level calibrations produced nearly unbiased results also for these domains. The cluster-effect calibration (CluE) improved the results, while not quite as much. In the smaller domains, where there was a few or zero observations to calibrate at domain-level, the improvement available from the cluster-effect was nearly as good. The predicted domain-effect (PreE) was worst of the calibration options regarding the high-end bias. The results with the more homogeneous clustering of domains were quite similar to these.

However, while the calibration was able to decrease the bias with the high-end volume domains, it increased the standard errors compared to a case of no calibration (Fig. 2). This is especially true for the calibrations in domain-level (DomE and CDE) and high-end volume domains. Also, the cluster-effect (CluE) with 30 groups ($H = 30$) increased variation among estimates across the simulated samples, but not as much as the domain-level calibration.

4.3. Effect of sample size

When the sample size m was reduced, first to 1000 and then to 500, the calibration results were worse overall, as could be expected (Table 5 and for systematic sampling

Table 5. The average domain-level results with small domains ($J = 2500$), formulated using maximum height (HMax), sample size 1000 or 500, and the cluster number $H = 30$ using simple random sampling.

	$m = 1000$			$m = 500$		
Cal	Se (m^3/ha)	Bias (m^3/ha)	RMSE (m^3/ha)	Se (m^3/ha)	Bias (m^3/ha)	RMSE (m^3/ha)
NoE	2.43	−0.84	13.34	3.54	−1.32	13.89
DomE	5.81	−0.84	13.29	5.93	−1.32	14.13
CluE	4.57	**−0.29**	12.75	6.34	**−0.72**	**13.65**
CDE	6.50	**−0.29**	**12.74**	7.60	**−0.72**	13.85
PreE	**2.35**	−0.61	16.20	**3.42**	−1.03	16.72

Note: The best average results for both sample sizes are bolded.

Fig. 3. The bias as the function of the true volume of the domains with the different predictor-options, domains formed with maximum height (HMax), sample design simple random sampling, sample size 1000 (left) or 500 (right), and cluster number 30 for the small domains ($J = 2500$).

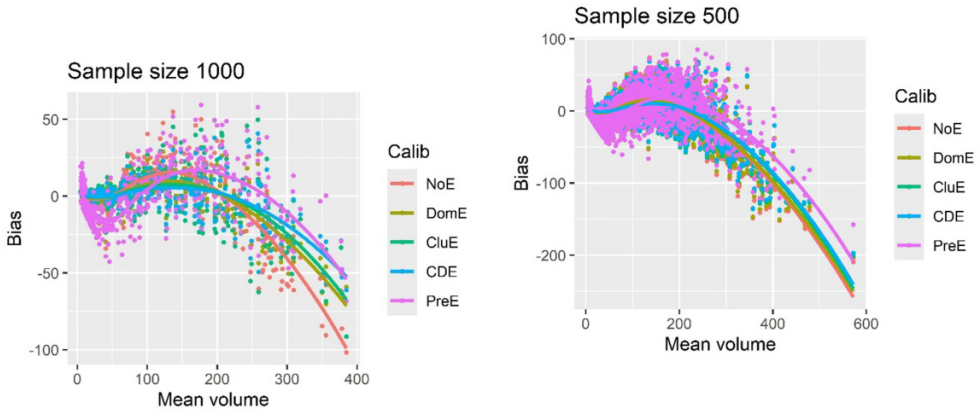


Table 6. The average domain-level results with large domains ($J = 500$), formulated and clustered using Inner volume (Volin), sample size $m = 5000$ and the cluster number $H = 30$ using simple random sampling, and the relative differences (Δ) to the HMAX results (Table 2).

Cal	Se (m^3/ha)	Bias (m^3/ha)	RMSE (m^3/ha)	ΔSe%	ΔBias%	ΔRMSE%
NoE	1.43	−0.28	10.45	−0.6	−21.4	−15.3
DomE	6.73	−0.28	8.17	5.1	−21.4	12.7
CluE	2.12	**−0.21**	9.95	2.2	303.9	5.1
CDE	6.69	**−0.21**	**8.12**	15.5	303.9	20.3
PreE	**1.43**	−0.28	11.15	2.0	35.3	−16.4

Note: The best average results are bolded.

Table S2). However, the smaller the sample fraction, the better the calibration with a cluster-effect was compared to the other calibration options. When looking at the bias in the high-end domains, the calibration with cluster-effect (CluE) seems to be a good option. The prediction model approach (PreE) looks also useful, especially with the smallest sampling fraction (Figure 3). However, in a more realistic setting, where the modelled random effects must be estimated from a sample, the improvements are not expected to be as good as in the case where known values of the domain effects were used to fit the model.

4.4. Effect of clustering variables and cluster number

In the results above, it was assumed that additional information to the explanatory variables of the models were available for formulating the domains and the clusters of domains. Often this is not the case, and therefore, we tested for a case where the domains were clustered to groups with one of the explanatory variables in the fixed part of the model, the Inner volume (Volin). It turned out that the cluster-effect calibration on average still improved the results, but the efficiency was not as good as with clustering that involved additional information to the fixed effects of the model (Table 6 and for systematic sampling Table S3). With Volin used for the cluster formulating instead of HMax or CSum, the non-calibrated (NoE) results and results calibrated with a model (PreE) were even more accurate than when using HMax. Regarding PreE this is a little surprising, as in this case the model predicting the true random effects was only able to explain 12% of the variation of the true domain effects.

Table 7. The average domain-level results in the small domains ($J = 2500$), with sample size $m = 1000$, the domains based on maximum height and groups based on maximum height and size using simple random sampling.

	$H = 10$			$H = 30$			$H = 50$		
Cal	Se (m^3/ha)	Bias (m^3/ha)	RMSE (m^3/ha)	Se (m^3/ha)	Bias (m^3/ha)	RMSE (m^3/ha)	Se (m^3/ha)	Bias (m^3/ha)	RMSE (m^3/ha)
NoE	2.42	− 0.69	13.52	2.43	− 0.84	13.34	**2.39**	**− 0.59**	13.30
DomE	6.09	− 0.69	13.49	5.81	− 0.84	13.29	**5.85**	**− 0.59**	13.24
CluE	**3.63**	**− 0.24**	13.50	4.57	**− 0.29**	12.75	5.21	− 0.27	12.65
CDE	6.51	**− 0.24**	**13.42**	**6.50**	**− 0.29**	**12.74**	6.87	− 0.27	**12.63**
PreE	2.34	− 0.45	16.46	2.35	− 0.61	16.20	**2.32**	**− 0.39**	16.24

Note: For each calibration method, the group size H producing best results are bolded.

The number of clusters had only a small effect on the average results (Table 7 and for systematic sampling Table S4). However, the larger number of groups improved the results to some extent, especially regarding the RMSE. The smaller number was beneficial when only the cluster-effect was used to calibration. We tested also a larger number of clusters, but with larger values of H the estimation often turned to be infeasible.

5. Discussion

The results show that it is possible to improve the small-area-estimation results with calibration, and when suitable information is available, cluster-effect for a group of small areas is useful. Model-based predictions always tend towards mean, meaning that the exceptionally large true values are underestimated (e.g., Mehtätalo and Lappi 2020, p. 110, see also Ståhl et al. 2024). That results in biased estimates in the small domains having the largest true volumes. The calibration is especially useful in reducing the underestimation problem in these high-end volume domains.

In this study, we attempted to formulate reasonably homogeneous domains and cluster them to homogeneous groups based on HMax and CSum. With the smaller domain size ($J = 2500$), the domains formulated with Maximum height can be assumed to mimic forest stands delineated using the height of the trees. The Sum Entropy, describing the texture of the canopy surface, can be assumed to describe the structure of the stand canopy. The average size, 2.5 ha, mimics the average size of forests stands in Finland, which is about 2.0 ha. The larger domains ($J = 500$) are much larger than forest stands in Finland on average and thus can be assumed to be more heterogeneous than the smaller domains. Forest stands of over 10 ha can be found mostly in the northern parts of Finland, and in southern Finland 12 ha can be a size of a whole forest estate.

Suitable information is important for clustering the domains successfully. In this study, the clustering (HMax, CSum) was based on the domain-level means of such lidar features that were not already utilized in the fixed part of the model. We tested clustering the domains also based on the domain-means of the features also used as fixed effects in the unit-level model (Volin), but that reduced the efficiency of the clustering, as the clusters did not bring additional information into the analysis. If it can be assumed that domains near each other are more alike than domains further away, also the coordinates could be used for the clustering. Such an approach is also a special case of spatial calibration. Suitable independent information could also include e.g., site information or dominant tree species information from available forest databases, or information from new data, e.g., satellite images or photogrammetric point cloud data collected after lidar acquisition used in model development.

However, in all cases tested, the clustering improved the results. With the larger domains ($J = 500$) and large sample size the improvement from clustering was not large. When there are on average 10 observations from each domain, the domain-effects could be estimated with such accuracy that the cluster-effect was not truly useful. With the smaller domains ($J = 2500$), the cluster-effect was more useful, as domain effects could not be estimated, or the estimates had a high variation. The benefits of cluster-effects were best demonstrated in the cases where the sampling fraction was smallest, i.e., when there were scant observations available for domain-level calibration.

Even though the domain-effects predicted with a model (eq. 10) rather than EBLUP approach were predicting "true" domain effects and used additional information to the fixed part of the unit-level model (Table3), they were still less useful than the calibration with EBLUP, provided that observations for the calibration were available from the domain. However, when the sampling fraction reduced, and the proportion of non-sampled domains increased, also this approach showed potential. Thus, whenever EBLUP is possible, it is recommended over the PreE approach.

Moreover, in real life the random effects have to be estimated from a sample, and thus the uncertainties included increase, and the potential of this method is reduced. Therefore, using a regression to predict the random effect might be useful in cases where there are several observations available from each sampled domain to get a good estimate of the random effects, and zero observations from a large majority of the domains, where the random effects are predicted. Suitable strategies for this approach require further studies. For instance, for this approach to work in reality, it might be a good option to first sample a set of small domains, and then acquire several observations from each of them, rather than acquiring one observation from a majority of stands. If a large proportion of sampled domains have 1–2 observations, this approach is not likely useful. The same also applies, if

spatial proximity is used in predicting the random effects in the non-sampled areas.

The number of groups may also be important. Low number of groups means that the calibration effect may be small, but there may be observations from all groups for a stable estimate (i.e., low variance of the estimated random effects). On the other hand, larger number of groups means more homogeneous groups, resulting higher within-group correlation and thus likely also more accurate calibration. However, with larger number of groups identifiability problems occurred. Thus, selecting the best way to cluster the domains is a complicated optimization problem, and the optimal number of clusters, for instance, likely depends on the homogeneity of the original domains.

The results also show that calibration can be surprisingly unstable, even when the sample size is moderately high. The benefits of calibration in the low-end domains might be fairly small, and therefore the best strategy might be to utilize calibration only in high-end volume cases, i.e., when the predicted domain mean is higher that e.g., 100 m^3/ha and therefore benefit from calibration.

6. Conclusions

Grouping the small domains into homogeneous groups and adding a cluster-effect to a mixed model is a promising approach in solving the non-sampled domain problem. It is inevitable that a large proportion of the small domains has no data for calibration, when these domains are as small as forest stands. For predicting the random errors in the non-sampled domains based on the random effects in the sampled domains can also improve the domain-level results, provided that an accurate estimate of these random effects is available for modelling.

Article information

History dates

Received: 12 January 2026
Accepted: 4 March 2026
Accepted manuscript online: 12 March 2026
Version of record online: 27 April 2026

Copyright

Data availability

All codes and the modelling data are available from the corresponding author with request.

Author information

Author ORCIDs

Annika Kangas https://orcid.org/0000-0002-8637-5668

Author contributions

Conceptualization: AK, MM, PP
Formal analysis: AK
Funding acquisition: AK, MM, PP
Methodology: AK, MM, PP
Software: AK
Writing – original draft: AK
Writing – review & editing: MM, PP

Competing interests

The authors declare there are no competing interests.

Funding information

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).

Supplementary material

Supplementary data are available with the article at https://doi.org/10.1139/cjfr-2025-0310.

References

Anisa, R., and Kurnia, A., Indahwati. 2014. Cluster information of non-sampled area In small area estimation. IOSR J. Math. (IOSR-JM), **10**(1): 15–19. doi:10.9790/5728-10121519.

Astrup, R., Rahlf, J., Bjørkelo, K., Debella-Gilo, M., Gjertsen, A.K., and Breidenbach, J. 2019. Forest information at multiple scales: development, evaluation and application of the Norwegian forest resources map SR16. Scand. J. For. Res. **34**(6): 484–496. doi:10.1080/02827581.2019.1588989.

Balazs, A., Liski, E., Tuominen, S., and Kangas, A. 2022. Comparison of Neural Networks and k-nearest neighbors methods in forest stand variable estimation using airborne laser data. ISPRS Open J. Photogramm. Remote Sens. doi:10.1016/j.ophoto.2022.100012.

Bates, D., Mächler, M., Bolker, B., and Walker, S. 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. **67**(1): 1–48. doi:10.18637/jss.v067.i01.

Bell, D.M., Wilson, B.T., Werstak, C.E., Jr, Oswalt, C.M., and Perry, C.H. 2022. Examining *k*-nearest neighbor small area estimation across scales using national Forest inventory data. Front. For. Glob. Change, **5**: 763422. doi:10.3389/ffgc.2022.763422.

Benedetti, R., Piersimoni, F., Pratesi, M., Salvati, N., and Suesse, T. 2024. Handling out-of-sample areas to estimate the unemployment rate at local labour market areas in Italy. Int. Stati. Rev. doi:10.1111/insr.12596.

Breidenbach, J., McRoberts, R.E., and Astrup, R. 2016. Empirical coverage of model-based variance estimators for remote sensing assisted estimation of stand-level timber volume. Remote Sens. Environ. **173**: 274–281. doi:10.1016/j.rse.2015.07.026. PMID: 28148972.

Brus, D.J., 2019. Sampling for digital soil mapping: a tutorial supported by R scripts. Geoderma, **338**: 464–480. doi:10.1016/j.geoderma.2018.07.036.

Burgard, J.P., Morales, D., and Wölwer, A-L. 2021. Small area estimation of socioeconomic indicators for sampled and unsampled domains, AStA Adv. Stat. Anal. **106**(2): 287–314. doi:10.1007/s10182-021-00426-4.

Chung, H.C., and Datta, G.S. 2022. Bayesian spatial models for estimating means of sampled and non-sampled small areas. Survey Methodology, 463: **48**(2): 463–489.

Desiyanti, A., Ginanjar, I., and Toharudin, T. 2023. Application of an empirical best linear unbiased prediction Fay–Herriot (EBLUP-FH) multivariate method with cluster information to estimate average household expenditure. Mathematics, **2023**(11): 135. doi:10.3390/math11010135.

Frescino, T.S., McConville, K.S., White, G.W., Toney, J.C., and Moisen, G.G. 2022. Small area estimates for national applications: a database to dashboard strategy using FIESTA. Front. For. Glob. Change, **5**: 779446. doi:10.3389/ffgc.2022.779446.

Georgakis, A., Papageorgiou, V.E., and Stamatellos, G. 2025. A new approach to small area estimation: improving forest management unit estimates with advanced preprocessing in a multivariate Fay–Herriot model. For. Int. J. For. Res. **98**(4): 605–622. doi:10.1093/forestry/cpae061.

Haralick, R.M., Shanmugam, K., and Dinsteein, J. 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. **3**: 610–621. doi:10.1109/TSMC.1973.4309314.

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

Kangas, A., Myllymäki, M., and Packalen, P. 2025b. The effect of sampling design in model-based small area estimation. Forestry. doi:10.1093/forestry/cpaf076.

Kangas, A., Myllymäki, M., and Mehtätalo, L. 2023. Understanding uncertainty in forest resources maps. Silva Fennica, **57**: 22026. doi:10.14214/sf.22026.

Kilkki, P., and Lappi, J. 1987. Estimation of taper curve using stand variables and sample tree measurements. Scand. J. For. Res. **2**: 121–126.

Magnussen, S., and Breidenbach, J. 2017. Model-dependent forest stand-level inference with and without estimates of stand-effects forestry. Int. J. For. Res. **90**: 675–685. doi:10.1093/forestry/cpx023.

Maltamo, M., and Packalen, P. 2014. Species specific management inventory in Finland. *In* Forestry applications of airborne laser scanning –concepts and case studies. Vol. 27. *Edited by* M. Maltamo, E. Naesset and J. Vauhkonen. Springer. pp. 241–252.

Maltamo, M., Packalen, P., and Kangas, A. 2021. 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. **51**: 257–266. doi:10.1139/cjfr-2020-0322.

Mauro, F., Monleon, V.J., Temesgen, H., and Ford, K.R. 2017. Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. PLoS One, **12**(12): e0189401. doi:10.1371/journal.pone.0189401. PMID: 29216290.

Mehtätalo, L., and Lappi, J. 2020. Biometry for forestry and environmental data. CRC Press. 411p. doi:10.1201/9780429173462.

Militino, A.F., Ugarte, M.D., and Goicoa, T. 2007. A BLUP synthetic versus an EBLUP estimator: an empirical study of a small area estimation problem, J. Appl. Statist. **34**:(2): 153–165. doi:10.1080/02664760600994893.

Nothdurft, A., Saborowski, J., and Breidenbach, J. 2009. Spatial prediction of forest stand variables. Eur. J. For. Res. **128**(2009): 241–251. doi:10.1007/s10342-009-0260-z.

Saei, A., and Chambers, R. 2005. Out of Sample Estimation for Small Areas using Area Level Data (S3RI Methodology Working Papers, M05/11) Southampton, UK. Southampton Statistical Sciences Research Institute, University of Southampton 23 pp. doi:10.1007/s10260-023-00698-x.

Sikov, A., and Cerda-Hernández, J. 2023. Estimating the prevalence of anemia rates among children under five in Peruvian districts with a small sample size. J. Ital. Stat. Soc. **32**(7).

Ståhl, G., Gobakken, T., Saarela, S., Persson, H.J., Ekström, M., Healey, S.P., et al. 2024. Why ecosystem characteristics predicted from remotely sensed data are unbiased and biased at the same time—and how this affects applications. For. Ecosyst. **11**(2024): 100164. doi:10.1016/j.fecs.2023.100164.

Tuominen, S., Pitkänen, T., Balázs, A., and Kangas, A. 2017. Improving multi-source national forest inventory by 3D aerial imaging. Silva Fennica, **51**(4). doi:10.14214/sf.7743.

Véga, C., Renaud, J.-P., Durrieu, S., and Bouvier, M. 2016. On the interest of penetration depth, canopy area and volume metrics to improve Lidar-based models of forest parameters. Remote Sens. Environ. **175**: 32–42. doi:10.1016/j.rse.2015.12.039.

Wadoux, A.C., and Heuvelink, G.B.M. 2023. Uncertainty of spatial averages and totals of natural resource maps. Methods Ecol. Evol. **14**: 1320–1332. doi:10.1111/2041-210X.14106.

White, G.W., McConville, K.S., Moisen, G.G., and Frescino, T.S. 2021. Hierarchical Bayesian small area estimation using weakly informative priors in ecologically homogeneous areas of the Interior western forests. Front. For. Glob. Change, **4**: 752911. doi:10.3389/ffgc.2021.752911.