Operations Research Forum (2024) 5:9
https://doi.org/10.1007/s43069-023-00288-3

RESEARCH

Temporal-Like Bivariate Fay-Herriot Model: Leveraging Past Responses and Advanced Preprocessing for Enhanced Small Area Estimation of Growing Stock Volume

Aristeidis Georgakis[1] · Vasileios E. Papageorgiou[2] · Demetrios Gatziolis[3] · Georgios Stamatellos[1]

Received: 25 October 2023 / Accepted: 12 December 2023 / Published online: 6 January 2024

Abstract
Forest inventories are crucial for effective ecosystem management but often lack precision for smaller geographical units due to limited sample sizes. This study introduces an enhanced temporal-like bivariate Fay-Herriot model, improving upon its univariate counterpart. The model incorporates field data and auxiliary data, including canopy height metrics from WorldView stereo-imagery and past census data, sourced from the University Forest of Pertouli in Central Greece. The model aims to estimate the growing stock volume for 2008 and 2018, focusing on enhancing the precision of the 2018 estimates. The 2008 dependent variable is used as auxiliary information by the model for more reliable 2018 small area estimates. A novel preprocessing pipeline is also introduced, which includes outlier identification, cluster analysis, and variance smoothing. Compared to direct estimates and the standard univariate Fay-Herriot model, our bivariate approach shows a percentage variance reduction of 96.58% and 13.52%, respectively. The methodology not only offers more reliable estimates with reduced variance and bias but also contributes to more accurate decision-making for sustainable forest management.

Keywords Multivariate area-level model · EBLUP · Clustering · Outliers · Repeated forest inventories · Remote sensing data

Mathematics Subject Classification 62J20 · 62P12 · 62H30 · 62H11

1 Introduction

Small area estimation (SAE) is recognized as a powerful statistical methodology designed to generate accurate information for specific subpopulations, particularly when sample sizes are limited [1]. SAE serves as a critical tool for informed

This article is part of the Topical Collection on *Mathematical Models and Optimization for Environmental Engineering and Sustainable Technologies*

Extended author information available on the last page of the article decision-making across various scientific domains. Applications of SAE are diverse, ranging from mapping poverty in economic studies and epidemiological research, to crop yield estimation in agriculture [1–3]. One of the most important applications lies in the field of forestry, where accurate biometrical attribute assessments are crucial for sustainable management [4–6].

Forest inventories (FIs) serve as the cornerstone for collecting data and estimating key forest attributes such as wood volume and above-ground tree biomass. These inventories are broadly categorized into national and management forest inventories (MFIs). MFIs play a vital role in the sustainable management of forest ecosystems. Their objective is not only to provide informed estimates for the entire forest population but also to offer precise information for geographic subpopulations such as forest management units (FMUs) or forest stands with limited sample sizes. The scope of MFIs includes also areas that have not been intentionally sampled due to the structured nature of the sampling process, primarily due to laborious and costly field surveys.

In the context of FIs, a sample unit or a plot represents a portion of forested land that contains a cluster of measured trees. The primary variable of interest in MFIs is the growing stock volume (GSV), which is vital for various forest management objectives including timber production, carbon sequestration, understanding of ecological health and structural complexity, and wildlife habitat provision. However, relying solely on direct estimates obtained from sampling surveys frequently results in substantial fluctuations in the sampling intensity of small areas of interest and, ultimately, to imprecise or unreliable estimations. Additionally, direct estimators are unable to provide estimates for areas without sample. To address this challenge, traditional strategies such as increasing the sample size or adopting more efficient sampling methods, like big basal area factor (big BAF) [7–10] have been employed. However, these approaches frequently are practically constrained owing to increased field data collection expenses.

To address the above difficulties, a contemporary and comprehensive strategy entails incorporating advanced statistical modeling methodologies, such as SAE. Specifically, the model-based SAE presents with a solution. Operating within the framework of two-stage linear mixed models, this approach leverages existing auxiliary data to improve estimations for “small areas,” small geographic regions, or domains, where direct estimates are inadequate [1]. Model-based SAE borrows strength from auxiliary information, including remote sensing and historical data, to enhance the sampling procedure. Model-based SAE is generally classified into two main categories based on the types of auxiliary variables utilized: unit-level models and area-level models [1, 6]. Firstly, unit-level models leverage variables accessible at the level of the individual sample unit [2], typically corresponding to field plots [11–13]. Generally, models utilizing unit-level data belong to the known area-based approach (ABA) [14–16].

Second, area-level models, also known as Fay and Herriot (FH) models, were initially introduced by Fay and Herriot [17]. These models establish a relationship between direct estimates like the mean and aggregated area-specific auxiliary information, among the small areas of interest. Area-level models have demonstrated notable efficacy within forestry [18, 19], especially in cases where the coordinates of sample plot centers are unavailable or when significant positioning errors lead to suboptimal correlations [20, 21]. Furthermore, this approach is computationally more efficient [6].

Auxiliary variables are crucial components for the success of SAE implementation. In the realm of FIs, remote sensing contributes most of the auxiliary variables. SAE is well supported by detailed 3D remote sensing data, derived mainly from laser scanners (light detection and ranging (LiDAR) or airborne laser scanning (ALS)) [18, 22], and 3D point clouds gathered from digital aerial photogrammetry [20, 23]. High-resolution remotely sensed data covering the whole population are useful for both unit and area-level models. Other data such as censuses can be useful covariates as well. In general, it is important to make a detailed assessment of all available data and to select appropriate sources of information to help improve the accuracy and reliability of small area estimates.

In FIs, the definition of a small area of interest varies, contingent upon the primary intent of the inventory and the spatial precision of estimations. For MFIs, these areas could encompass forest stands, compartments, or FMUs. Recent research has illustrated that after employing cluster analysis, the definition of small areas can be expanded to encompass more homogenous regions and, as a consequence, enlarge the sample size of the redefined domains (small areas) [25]. This preprocessing step is mandatory in FH models every time the assumption of a strong linear correlation between the auxiliary data and response variable is not met. Recent research in heterogenous uneven-aged fir proved that the direct estimates GSV for domains with 1–3 sample plots do not correlate strongly with the available auxiliary data, and the FH model could be applied only after the above process [24, 25].

An extension of the univariate FH (UFH) model is the multivariate FH (MFH), which can provide more efficient estimators by considering the correlations between the response variables, for both empirical best linear unbiased prediction (EBLUP) and hierarchical Bayes (HB) estimation approaches [1, 26–28]. The MFH models can use the residual maximum likelihood (REML) method to estimate the random effects and their correlations [29]. Similar to MFH, multivariate nested error regression was introduced by Fuller and Harter [30] for the unit-level SAE approach. Our emphasis in this article is on a new, multivariate SAE methodology for modeling GSV.

A temporal bivariate area-level linear mixed model with independent time effects was used to estimate small area socioeconomic indicators based on EBLUP predictors [31]. Another instance of a Bivariate mixed-effects model based on a conditionally specified model offers a practical approach to SAE for bivariate data with binary and Gaussian components, permitting frequent inference based on empirical Bayes (EB) predictors [32]. Recently, a new multivariate mixed-effects model for SAE of mixed-type response variables subject to item nonresponse was introduced, showing its improvements compared to direct estimators and univariate models [33]. By employing bivariate SAE models, researchers can significantly enhance the accuracy and reduce the variance of estimates derived from smaller household surveys in the USA, without the need for additional regression covariates. This improvement is facilitated by capitalizing on the strong correlations between population characteristics estimated in smaller surveys and those in the American Community Survey (ACS) [34]. In the context of discontinuities arising from changes in the design of repeated surveys, van den Brakel and Boonstra [35] introduced a bivariate hierarchical Bayesian model, which offers advancements over traditional UFH models and incorporates an adjusted step-forward selection procedure to mitigate overfitting.

In this study, we introduce the term “bivariate temporal-like” to differentiate it from “temporal univariate” Fay-Herriot models commonly employed in SAE. Our primary focus lies in evaluating the applicability of the proposed model for handling repeated measures and leveraging temporal correlations or trends. Although both models aim to enhancing the accuracy and efficiency of small area estimates, they differ in their nature of temporal data and underlying assumptions.

The bivariate temporal-like Fay-Herriot model emphasizes the significance of repeated surveys or responses from disparate time periods (e.g., GSV for 2008 and 2018). Unlike traditional temporal models, it does not model the entire time series of a given variable. Instead, it concentrates on capturing the correlation between two specific time points, making it particularly useful when only limited temporal data are available.

An innovative aspect of this model involves the use of past response variables as explanatory data. This approach captures the inherent temporal correlation between different time points, thereby minimizing the need for external covariates and offering a unique perspective on SAE. The bivariate approach primarily focuses on capturing the correlation between two specific time references or years, rather than modeling the temporal dynamics over extended periods. A critical assumption to maintain is that past responses serve as reliable predictors of future responses, particularly when the variable of interest remains relatively stable over time. In the context of sustainable management of a single-tree selection system in uneven-aged forests, and given consistent management practices across units, it can be assumed that the GSV remains stable across different FMUs.

In contrast, the typical temporal univariate Fay-Herriot model, which follows an autoregressive process, leverages the entire temporal correlation structure in the data by incorporating time-dependent auxiliary variables. This provides a more comprehensive approach to modeling temporal dynamics [36]. SAE under a multivariate linear model for repeated measures data has been addressed using random effects growth curve models and accounting for the correlation structure among repeated measures within each area [37, 38]. Both models can account for repeated measures, where surveys from different periods are considered as repeated observations of the same small area, allowing the model to “borrow strength” both across small areas and over time.

One of the first examples of SAE application in agriculture was with a multivariate linear regression model for repeated data measurements to produce district-level estimates of crop yield in Rwanda for the agricultural seasons of 2014 [39] based on [38] with covariates being available at the district level. Additionally, the sampling design that response data are surveyed with can vary [40]. For the example above, the Crop Assessment Survey was based on a two-stage stratified sampling design where strata were administrative districts [39]. In the first stage, primary sampling units are selected with probability proportional to the area size. For the second stage, secondary sampling units are selected with simple random sampling.

Consequently, in this paper, we propose the utilization of a temporal-like bivariate Fay-Herriot (BFH) model on FI data derived from a pipeline preprocessing methodology, targeting the estimation of the GSV for the years 2008 and 2018. This method presents with various benefits over conventional approaches. Initially, it facilitates the integration of supplementary variables like remotely sensed data and historical records, thereby enhancing the accuracy and precision of estimations. Secondly, it accommodates the interrelation between GSVs and auxiliary variables, capturing the intricate connections within the forest ecosystem.

Through the implementation of the innovative bivariate Fay-Herriot model—a novel approach in the field of forestry—we seek to improve the precision and dependability of estimating GSV, thereby supporting well-informed decision-making regarding sustainable forest management. We utilized remote sensing and historical data sourced from the University Forest of Pertouli. Additionally, we have incorporated a preprocessing methodology that combines clustering, variance smoothing, variable selection, and outlier detection techniques. The proposed preprocessing scheme significantly enhances the robustness of the estimates of the multivariate framework. The performance of the proposed BFH approach is rigorously evaluated concerning bias and the coefficient of variation (CV). Our findings are compared to the widely-used UFH model and the precision of the direct estimates, leading to noteworthy insights about the practicality and effectiveness of these statistical models in the context of SAE for FIs.

2 Materials and Methods

2.1 Study Area and Data

This study was conducted in the University Forest of Pertouli in Central Greece (Fig. 1). This ecosystem, characterized by uneven-aged forest, is primarily dominated by the hybrid fir species, *Abies borisii-regis* Mattf. The forested area spans 2260 ha within a larger territory of 3297 ha, which includes other land uses like meadows. The forest is divided into 174 FMUs, 160 of which were included in the analysis after eliminating unmanaged FMUs and those lacking adequate auxiliary data. The field data comprises 239 sample plots, each measuring 0.1 ha for the year 2018. Of the 239 sample plots, 218 were common with those from 2008 and used in this study. The sampling frequency within the forest land was close to 1%. Nearly half of the FMUs have a single sample plot, and the other half have two plots. Only a few FMUs consist of three or none [41].

Our dataset combines tree-level information with pixel-level height metrics extracted from remote sensing data. Specifically, we generated a 2-m canopy height model (CHM) and used it to compute the distribution of tree height metrics, including L-moments, for each FMU and small area. Figure 2 presents the distribution of the mean values from the CHM at the FMU level (left panel) accompanied by boxplots of height at the cluster level or small areas of interest (right panel). The CMH was computed by subtracting the available LiDAR-based digital terrain model (DTM) from the digital surface model (DSM) generated from the WorldView 2 satellite data stereo pair.

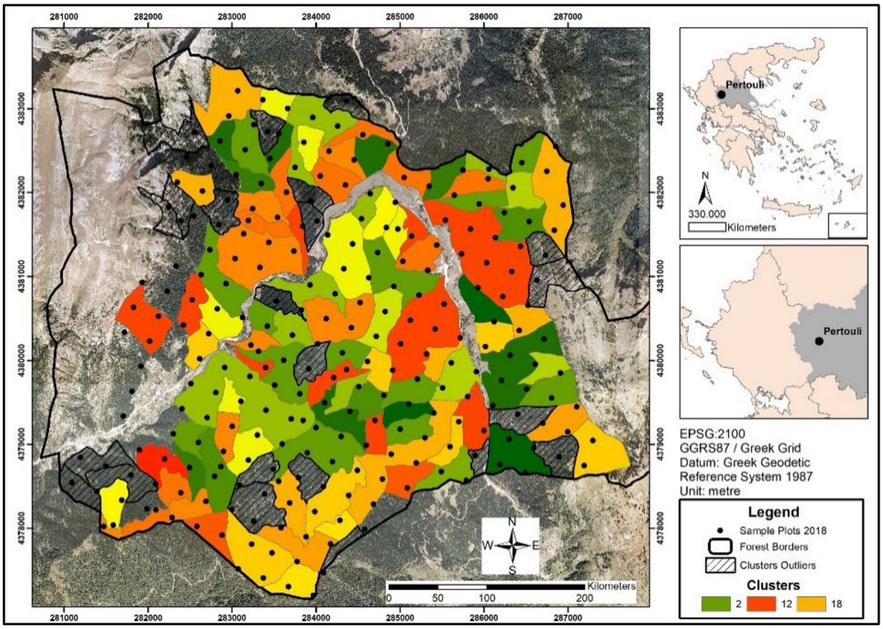

Fig. 1 Pertouli forest study area: sampling design and cluster distribution

We incorporated tree-level density and volume data from past inventories. These attributes of the forest were used as covariates in the multivariate Fay-Herriot models that we investigated, to obtain dependable statistics at the area level. Despite

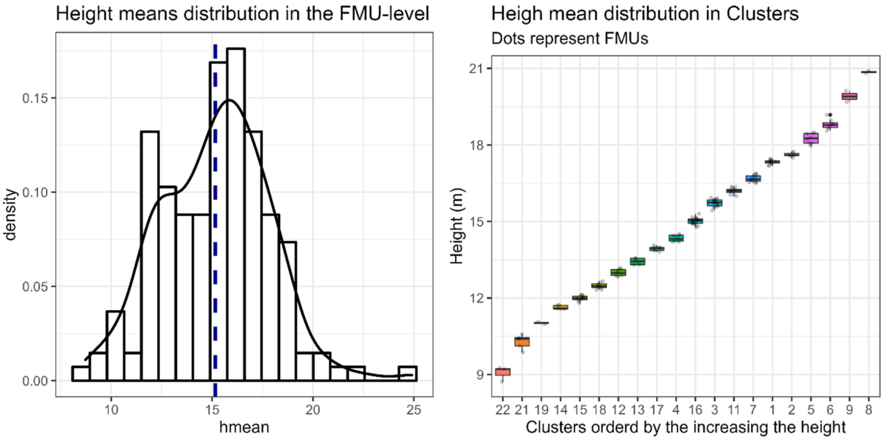

Fig. 2 Distribution of canopy height mean in Pertouli forest. In the left panel, the distribution of height means is illustrated in the forest management unit (FMU) level. The right panel displays the distribution of the height mean of the FMUs at the cluster level

employing systematic sampling for the sample plots, we used direct estimators based on simple random sampling with replacement. No small areas beyond the sampled ones were considered, and each domain contained a minimum of two sample plots to facilitate variance estimation. The auxiliary variables of the dataset, along with descriptive metrics and units, are presented in Table 1.

2.2 Advanced Preprocessing Pipeline

In the present analysis, small areas are redefined through the aggregation of FMUs or stands, accomplished via cluster analysis centered on height metrics. This process is conducted to establish a linear correlation between the variable of interest and auxiliary data. The term "small areas" pertains to consolidated FMUs. The rationale behind clustering FMUs emerges from the necessity—to adhere to the foundational model assumption—of a robust linear association between covariates and the variables of interest. The data are obtained from multi-story, uneven-aged forests, and as a result, the modest sample size of 1–3 sample plots per FMU cannot accurately represent the overall status. This situation led to a feeble correlation, rendering the utilization of area-level models infeasible. Grouping FMUs that share similar characteristics addresses this challenge. During the clustering procedure, the choice of clustering variables holds significance as they are expected to encapsulate the heterogeneity among FMUs. For this reason, we opted for tree height, a variable strongly linked to volume. Specifically, we utilized the mean height metric calculated by aggregating pixel heights to the FMU level.

In the process of conducting clustering analysis, we utilized the hierarchical single-linkage clustering algorithm [42], with the Euclidean distance serving as the measure for calculating distances [43]. We set the minimum cluster count at 8, which, on average, corresponds to over 10 sample plots suggested for expanded small areas [44]. As an indicator for the clustering process, we used the aggregated canopy height mean at the FMU level. The challenge in clustering for SAE purposes involved achieving a robust correlation, while concurrently generating as many domains as feasible to enhance predictions at a more refined spatial resolution [25].

Table 1 Auxiliary data derived from remote sensing data and previous census inventories

Data type	Descriptive statistics	Abbreviations	Unit metric	Year
Height	Quantiles	h25; h50; h75; h90	Meters (m)	2022
Height	Central tendency	hmean; hmode	m	2022
Height	Dispersion metrics	hsd; hcv	m; dimensionless	2022
Height	Shape distribution; L-moments	hLskew, hLcv; L3, L4	Dimensionless	2022
Census	Central tendency	FirTreeDensityt97ha	Trees/hectare (ha)	1997
Census	Central tendency	ForestDensity97ha	Trees/ha	1997
Census	Central tendency	ForestGSV97ha	m^2/ha	1997
Census	Central tendency	FirGSV88ha	m^3/ha	1988

To achieve the aim, we relied on the Calinski-Harabasz index [45] to identify the optimal number of clusters, discernible by the maximum value of the index. A study with simulated forest plots found that the Calinski-Harabasz index consistently performed better at detecting the optimal number of clusters in comparison to other methods [46]. This is aligned with recent extensive research on the optimal number of clusters based on different clustering schemes and auxiliary variables for the estimation of forest attributes in the context of SAE [25].

Using cluster analysis, we established 36 expanded domains, each encompassing an average of 8.25 sample plots. The collective average relative standard error (RSE) among all domains was 13.31%. Among these domains, the 7 that contained a single sample plot were excluded from the analysis due to the inability to calculate the variance of the sampling mean. Another 18 sample plots in 9 domains, considered correlation outliers or exhibited covariance instability, were also eliminated, yielding an 8.26% reduction in the total number of sample plots.

The exclusion of this small percentage of plots led to a significant increase in correlation, as illustrated in the left panel of Fig. 3. This also resulted in the elimination of extreme co-variance estimates. The right panel in Appendix Fig. 10 shows the enhanced, post-clustering correlation, calculated without the removal of outliers, compared to the correlation diagram at the FMU level on the left panel (prior to clustering FMUs). Finally, 20 domains were identified as sampled areas, averaging 9.65 sample plots each, as detailed in Table 2. We observed that systematic sampling ensures consistent sampling intensity across each small area, as indicated in Table 2. Following cluster formation, auxiliary variables were aggregated within the newly defined smaller regions, from which we derived direct estimates using a design-based approach.

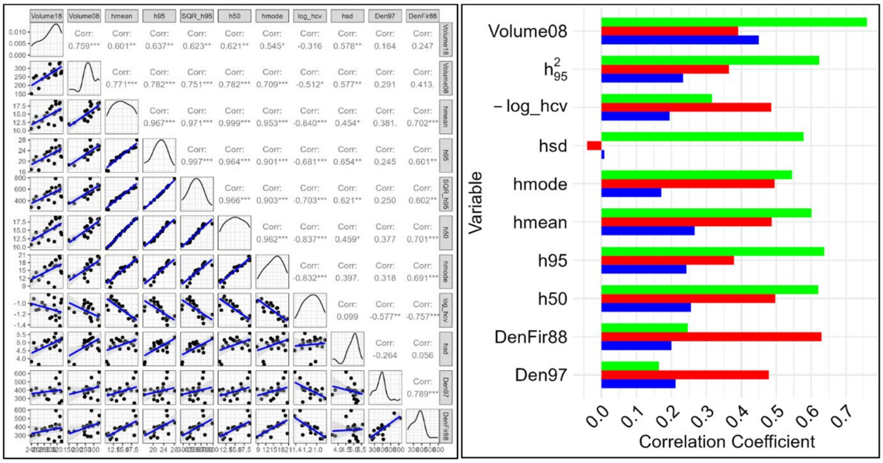

Fig. 3 Left panel: correlations at the cluster level—excluding outliers—for the response variables (volume 18 and 08) and auxiliary variables are displayed. Right panel: diagrammatic comparison of correlations before (blue color) and after the clustering process (red color) and the extraction of outliers (green color)

Table 2 Summary statistics of clusters, number of sample plots, area size, participation of primary forest management units (FMUs)

	Number of clusters	Total nPlots[a]	Mean of nPlots	Total area in hectares (ha)	Mean domain area (ha)	Mean of FMUs	Sampling intensity
All data	36	218	6.42	2076.05	61.06	6.80	0.95%
No outliers	20	193	9.65	1815.99	90.80	4.64	0.94%

[a] "nPlots" refers to the number of sample plots or sample size, and "Total" refers to the total number of population units

Before proceeding with the temporal-like FH model, a correlation analysis was conducted utilizing the Pearson correlation coefficient and assessing the significance of the coefficients obtained. Outliers were identified using scatterplots, and their impact on correlation was further validated using a robust multivariate estimator, the Donoho-Stahel estimator of multivariate location and scatter [47]. Moreover, we conducted a supplementary outlier evaluation by comparing the standard Pearson and the weighted correlation, with sample sizes serving as weights. Both types of correlation must show minimal disparities, given that domains featuring limited sample sizes could disproportionately affect the correlation outcomes. From the analysis, we observed that specific domains, predominantly comprising just two sample plots, exhibited extreme direct estimates and variances.

Overall, we have identified three types of outliers during the model-building process through the following steps: (I.) in correlation analysis, specifically focusing on the dependencies between the GSV of 2018, the other response (GSV of 2008), and auxiliary variables; (II.) in covariance matrices; and (III.) in the residuals of random area effects. An alternative to withdrawing outliers is to employ a robust approach for outlier treatment in SAE; more about this approach is included in the discussion. The right panel of Fig. 3 displays a comparative analysis of correlations, focusing on the GSV for 2018, compared to the Volume08 (GSV of 2008) auxiliary response variable, along with the auxiliary variables.

The analysis spans three distinct categories: (1) correlations for the initial FMUs (FMU-level, colored with blue), (2) after clustering of FMUs to form larger domains (colored with red), and (3) following the removal of outliers from these clusters (colored with green). The left panel represents the correlation values after the application of the preprocessing pipeline.

The final step of the preprocessing pipeline involves a variance-smoothing procedure. The FH model presupposes known sampling variances, an assumption that is frequently untenable in real-world applications. The model tends to be unstable, leading to computational issues, zero random area effects, and a degradation in the performance of the multivariate estimator. Smoothing the sampling variances using the generalized variance function (GVF) can mitigate this issue [48], especially for domains with small sample sizes. In forestry applications, a specific type of smoothing is implemented by weighting the sampling variance based on the small area size A_d [18, 19, 49]. This implies allocating greater weight to larger domains, as they are expected to yield more precise estimates.

The smoothed sampling variances $\widetilde{\sigma}^2_{\varepsilon d}$ are defined as

$$\widetilde{\sigma}^2_{\varepsilon d} = \frac{V_\varepsilon}{n_d}, \tag{1}$$

$$V_\varepsilon = \frac{\sum_{i=1}^{D} A_d \widehat{\sigma}^2_{\varepsilon d}}{\sum_{i=1}^{D} A_d} \tag{2}$$

where V_ε is a constant weighted mean of variances, n_d denotes the number of units in domain d, A_d is the respective total area, and D is the number of domains. The term $\widehat{\sigma}^2_{\varepsilon d}$—calculated directly from the sample plots—is utilized to estimate the $\sigma^2_{\varepsilon d}$, whereas $\widetilde{\sigma}^2_{\varepsilon d}$ presents a smoothed version of the original sampling variances. The following equation presents the unbiased sample variance under simple random sampling (SRS) with replacement, without the finite population correction factor

$$\widehat{\sigma}^2_{\varepsilon d} = \sum_{j=1}^{n_d} \left(y_{dj} - \overline{y}_d\right)^2 / \left(n_d - 1\right) \tag{3}$$

Furthermore, we calculated the covariances by taking the square root of the smoothed variances and multiplying by the correlation of the variables under study,

$$\widetilde{\sigma}_{\varepsilon d12} = \rho \bullet \sqrt{\widetilde{\sigma}^2_{\varepsilon d1} \bullet \widetilde{\sigma}^2_{\varepsilon d2}} = \rho \bullet \widetilde{\sigma}_{\varepsilon d1} \bullet \widetilde{\sigma}_{\varepsilon d2} \tag{4}$$

Specifically, $\widetilde{\sigma}^2_{\varepsilon d1}$ and $\widetilde{\sigma}^2_{\varepsilon d2}$ represent the smoothed variances of the two examined response variables. This is based on SRS, which assigns equal inclusion probabilities to all units within a given domain. When different probabilities are assigned to each domain based on the domain area, this results in a weighted sampling variance. In such cases, the weighted variances are typically smaller, but the smoothed variances exhibit fewer fluctuations. Details about the weighted variance estimation are provided in [50].

In the left panel of Fig. 4, we present the variances of the GSV 2018 direct estimates under three conditions: using SRS (Variance_SRS), after applying the selected smoothing procedure (Smoothed_Variance), and with varying inclusion probabilities of sample plots (Weighted_Variance) based on the domain area. It is important to note that the sampling error variance is not assumed to follow a normal distribution. After applying the smoothing procedure, our analyses reveal a shift from initially non-normally distributed sampling variances to a normal distribution, as shown in Fig. 4's right panel part and confirmed by the Shapiro–Wilk test.

After all the preprocessing steps, the variable selection process was explored based on a range of techniques. We initially tested lasso regression with cross-validation [51, 52], random forests' variable importance [53], recursive feature elimination [54], PCA-based selection [55], Bayesian model averaging (BMA) [56], variance inflation factor (VIF) assessments [57], and the branch and bound method [58]. Among these, we adopted the EBLUP-FH method, employing stepwise

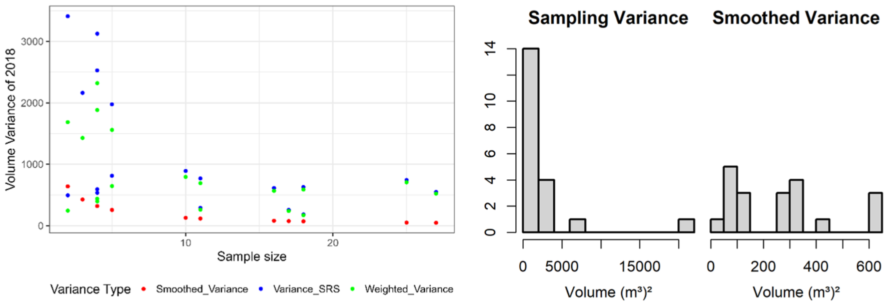

Fig. 4 Left panel: three types of sampling variances for direct estimates of growing stock volume in 2018—variance calculated using simple random sampling (SRS) (Variance_SRS), variance after application of selected smoothing procedure (Smoothed_Variance), and variance with varying inclusion probabilities of sample plots based on domain area (Weighted_Variance). Right panel: frequency distributions representing both typical sampling variances and smoothed sampling variances for growing stock Volume

selection based on the AIC criterion using the “saeBest” package [59]. This method outperformed the others in suitability. More specifically, the sets of auxiliary variables suggested by the aforementioned methodologies generated higher MSE and CV values compared to the set of variables provided by the EBLUP-FH method. Consequently, the final MSE and CV values were the selection criterion of the EBLUP-FH method. The incorporation of many covariates can cause computation challenges including convergence problems; therefore, the model should be kept as simple as possible, aiming to avoid overfitting issues. For this reason, after the variable selection, we executed and then interpreted the model’s coefficients and variance of random area effects.

2.3 Multivariate Fay-Herriot Model

MFH models incorporate the correlation among response variables and borrow strength from auxiliary variables, resulting in more efficient parameter identification. This fact leads to results of reduced variance, compared to direct estimates and the UFH model. Moreover, compared to the standard univariate approach, the MFH models incorporate the linear dependencies between the chosen response variables, alleviating the requirement of auxiliary variables with significant relationships with all the characteristics under study.

FH models can be described in two stages, consisting of a sampling design stage and a linking model. In the sampling design stage, direct survey estimates (Y_{dk}) account for the sampling variability or random sampling errors (ε_{dk}) of the true area values (Y_{dk}). The direct estimates of the domain means are used as responses in the area-level model. The sampling model can be described as

$$y_{dk} = Y_{dk} + \varepsilon_{dk}, d=1,\ldots, D, k=1,\ldots, K \tag{5}$$

The covariance estimator of the sampling means of the response variables is

$$\widehat{cov}\left(\overline{y}_{dk},\overline{y}_{dl}\right)=\widetilde{\sigma}_{\varepsilon dkl}=\rho\bullet\sigma_{\overline{y}_{dk}}\bullet\sigma_{\overline{y}_{dl}} \tag{6}$$

Based on this formula, where the smoothed variances are employed, the covariance matrix of sampling errors ($\Sigma_{\varepsilon d}$) is constructed.

The linking model relates the true population means (y_{dk}) to the auxiliary variables ($\mathrm{x}_{\mathrm{dk}}^{\mathrm{T}}$) through the parameter vector (b_{k}). As a result, the linking model is represented by the linear relation

$$Y_{dk}=\boldsymbol{x}_{dk}^{T}\boldsymbol{b}_{k}+\ u_{dk},\ d=1,...,D,k=1,...,K \tag{7}$$

We note that D represents the cardinality of domains or small areas in the population, while K represents the number of linearly correlated target variables. In addition, $\boldsymbol{y}_{dk}$ corresponds to the unbiased direct survey estimate of the population's means Y_{dk} for each domain d, and target variable k, where $d=1,\dots,D$ and $k=1,\dots,K$. Also, $\boldsymbol{x}_{dk}$ represents a vector of p auxiliary variables that are linearly correlated with Y_{dk},$d=1,\dots,D$, $k=1,\dots,K$. Finally, $\boldsymbol{b}_{k}=\left(b_{k1},\dots,b_{kp_{k}}\right)$ is a vector of coefficients describing the relationship between the auxiliary variables and the direct estimates. The subscript p_k corresponds to the cardinality of auxiliary variables $\boldsymbol{x}_{dk}=\left(x_{dk1},\dots,x_{dkp_{k}}\right)$ that are incorporated in the linear model that describes Y_{dk}. More specifically, in the bivariate case ($K=2$) formulas (2) and (3) can be written as

$$\begin{pmatrix}y_{d1}\\ y_{d2}\end{pmatrix}=\begin{pmatrix}Y_{d1}\\ Y_{d2}\end{pmatrix}+\begin{pmatrix}\varepsilon_{d1}\\ \varepsilon_{d2}\end{pmatrix},d=1,\dots,D \tag{8}$$

and

$$\begin{pmatrix}Y_{d1}\\ Y_{d2}\end{pmatrix}=\begin{pmatrix}\boldsymbol{x}_{d1}&0\\ 0&\boldsymbol{x}_{d2}\end{pmatrix}\begin{pmatrix}\boldsymbol{b}_{1}^{T}\\ \boldsymbol{b}_{2}^{T}\end{pmatrix}+\begin{pmatrix}u_{d1}\\ u_{d2}\end{pmatrix},d=1,\dots,D \tag{9}$$

or

$$\begin{pmatrix}Y_{d1}\\ Y_{d2}\end{pmatrix}=\begin{pmatrix}x_{d11}\dots x_{d1p_{1}}&0&\dots&0\\ 0&\dots&0&x_{d21}\dots x_{d2p_{2}}\end{pmatrix}_{2\times p}\begin{pmatrix}b_{11}\\ \vdots\\ b_{1p_{1}}\\ b_{21}\\ \vdots\\ b_{2p_{2}}\end{pmatrix}_{p\times 1}+\begin{pmatrix}u_{d1}\\ u_{d2}\end{pmatrix},d=1,\dots,D \tag{10}$$

where $p=p_1+p_2$, shows the sum of the auxiliary variables used for both Y_{d1} and Y_{d2}.

By combining Eqs. (5) and (7), we obtain an area-level mixed random effect model where y_{dk} is modeled as a linear combination of x_{dk}, a domain-specific random effect u_{dk}, and the sampling error $\boldsymbol{\varepsilon}_{\mathrm{dk}}$. Hence, the area-level random effect model is given by

$$\boldsymbol{Y}_{dk} = \boldsymbol{x}_{dk}^T\boldsymbol{b}_k + \ u_{dk}, \ \varepsilon_{dk} = 1, ..., D, k = 1, ..., K \tag{11}$$

while for the BFH, we have

$$\begin{pmatrix} y_{d1} \\ y_{d2} \end{pmatrix} = \begin{pmatrix} x_{d11} \ldots x_{d1p_1} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & x_{d21} \ 0 \ x_{d2p_2} \end{pmatrix}_{2\times p} \begin{pmatrix} b_{11} \\ \vdots \\ b_{1p_1} \\ b_{21} \\ \vdots \\ b_{2p_2} \end{pmatrix}_{p\times 1} + \begin{pmatrix} u_{d1} \\ u_{d2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{d1} \\ \varepsilon_{d2} \end{pmatrix}, d = 1, \ldots, D \tag{12}$$

The MFH model for the d-th domain can be expressed in matrix form as

$$\boldsymbol{Y}_d = \boldsymbol{x}_d^T\boldsymbol{b} + \ z_d \ \boldsymbol{ud}, \ \boldsymbol{\varepsilon}_d, \ d = 1, ..., D \tag{13}$$

where $\boldsymbol{y}_d = \left(y_{d,1}, y_{d,2}, \ldots, y_{d,K}\right)^T$ is a vector of direct survey estimates for the domain d, $\boldsymbol{x}_d$ is a matrix of auxiliary variables for the domain d, $\boldsymbol{b} = \left(\boldsymbol{b}_1^T, \boldsymbol{b}_2^T, \ldots, \boldsymbol{b}_K^T\right)_{p\times 1}^T$ is a parameter matrix, $\mathrm{z_d}$ is a design matrix, $\boldsymbol{u}_d$ is a vector of random effects for the domain d, and ε_{d} is a vector of sampling errors. Both vector of sampling errors and random effects follow normal distributions of zero mean, where $\boldsymbol{\varepsilon}_d \sim N(0, \boldsymbol{\Sigma}_{\varepsilon d})$ and $\boldsymbol{u}_d \sim N(0, \boldsymbol{\Sigma}_{ud})$.

The covariance matrix of $\boldsymbol{y}_d$ is given by

$$\boldsymbol{V}_d = z_d \ \boldsymbol{\Sigma}_{ud} \ z_d^T + \boldsymbol{\Sigma}_{\varepsilon d} \tag{14}$$

where $\boldsymbol{\Sigma}_{\varepsilon d}$ represents a known $K \times K$ sampling covariance matrix of the direct estimates. Matrix $\boldsymbol{\Sigma}_{ud} = diag\left(\sigma_{uk}^2\right), 1 \leq k \leq K$, is associated with homoscedastic and uncorrelated random effects. Alternative structures/formulas for $\boldsymbol{\Sigma}_{ud}$ can be found in [29] to account for correlated and heteroscedastic effects. For the UFH case, $\boldsymbol{\Sigma}_{\varepsilon d}$ also takes a diagonal form.

By aggregating the D area-level models, Eq. (13) can be described in matrix form as

$$\boldsymbol{y} = \mathrm{Xb} + \mathrm{Zu} + \varepsilon \tag{15}$$

This formulation represents the integration of the D area-level models into a matrix form for the MFH. The covariance matrix of $\boldsymbol{y}$ is now represented by

$$\boldsymbol{V} = \mathrm{Z}\boldsymbol{\Sigma}_{\mathrm{u}} \, \mathrm{Z}^{\mathrm{T}} \ + \ \boldsymbol{\Sigma}_{\varepsilon} \tag{16}$$

Matrices $\boldsymbol{\Sigma}_{\varepsilon}$ and $\boldsymbol{\Sigma}_u$ have a block diagonal form, assuming independence between sampling errors and random effects across domains.

To estimate the target variables for different domains, we employ the multivariate empirical best linear unbiased predictors (EBLUP) of $\boldsymbol{Y}$,

$$\widehat{\boldsymbol{Y}}_E = \boldsymbol{X}\widehat{\boldsymbol{b}}_E + \boldsymbol{Z}\widehat{\boldsymbol{u}}_E \tag{17}$$

The respective estimate for the covariance matrix of the EBLUP estimators is produced by

$$\widehat{\boldsymbol{V}} = \boldsymbol{Z}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}\boldsymbol{Z}^{T} + \boldsymbol{\Sigma}_{\varepsilon} \tag{18}$$

In addition, the model's coefficients and random effect estimates are obtained according to

$$\widehat{\boldsymbol{b}}_{E} = \left(\boldsymbol{X}^{T}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{y} \tag{19}$$

and

$$\widehat{\boldsymbol{u}}_{E} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}\boldsymbol{Z}^{T}\widehat{\boldsymbol{V}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{b}}_{E}\right) \tag{20}$$

The above quantities are generated based on an iterative residual maximum likelihood (REML) algorithm, resulting in the construction of the Fisher information matrix. The results of this algorithm are later incorporated into the estimation of the covariance matrix of random effects $\widehat{\Sigma}_{ud}$. The estimated covariance matrix participates in the computation of $\widehat{\boldsymbol{V}}$, since $\widehat{\boldsymbol{V}} = \boldsymbol{\Sigma}_{ud} + \boldsymbol{\Sigma}_{\varepsilon d}$, according to EBLUP. More details about the REML algorithm can be found in [29]. The statistical analysis was conducted using the open-source statistical software R, with the R package "msae" [60].

To compare the performance of two estimators, we use the relative efficiency (RE) and the percentage variance reduction gain [6]. The direct estimates are denoted as y_{dk}^{DIR} and the EBLUP model-based estimates of the temporal-like BFH and the UFH as y_{dk}^{EBLUP}. The RE is computed as the ratio of the variance of the EBLUP estimates to the variance of the direct estimates

$$RE = \frac{Var\left(y_{dk}^{EBLUP}\right)}{Var\left(y_{dk}^{DIR}\right)} \tag{21}$$

If RE is less than one, the EBLUP estimator in the numerator is more efficient than the direct estimator in the denominator. The RE can be expressed as a percentage variance reduction gain using the formula $100\%(1-RE)$. A positive percentage gain indicates that the EBLUP estimator is more efficient, i.e., has lower variance.

In addition to the bias diagnostic scatter plot, we assess potential bias by comparing model estimates with direct estimates in large domains, using direct estimates derived from all available observations as the gold standard [61]. This comparison serves as a "calibration diagnostic" for evaluating and potentially correcting bias in small area estimation models.

The difference between modeled and direct estimates, when aggregated to larger domains, helps identify if any specific large domain is estimated less accurately than others [62]. The expectation is that model estimates should closely align with direct estimates when aggregated at appropriate large domain levels. The calibration value, expressed as a percentage difference (Eq. 22), serves as an indicator of average relative bias (ARB). A value close to zero suggests unbiased model-based estimates, while a significant deviation may signal bias, warranting further investigation or adjustments.

$$\mathrm{ARB}\% = \frac{1}{D}\sum\nolimits_{i=1}^{D}\left(\frac{y_{dk}^{EBLUP} - y_{dk}^{DIR}}{y_{dk}^{DIR}}\right) \times 100 \tag{22}$$

Lastly, validating small area estimates poses challenges distinct from conventional statistical methods like in-sample or cross-validation, especially in data-sparse domains [61]. To address this, we performed a simulation resembling leave-one-out cross-validation (LOOCV) to evaluate the model's robustness in the absence of particular domains. The validation process involved assessing the model's performance by examining the estimated variance of random area effects and mean squared error (MSE) when excluding specific domains.

3 Results

The objective of the FIs is to provide updated and precise estimates regarding forest attributes. Therefore, we present the estimates from the most recent sampling in 2018, as this point in time is more relevant for current FI decision-making and management activities. After the model selection procedure based on the minimization of the AIC criterion, we obtained the following variable set, where

$$\begin{pmatrix} \text{Volume18} \\ \text{Volume08} \end{pmatrix} \sim \begin{pmatrix} h\text{mean} + h95 + h50 + \log(h\text{mode}) + log(hLcv) \\ h\text{mean} + hsd + Den97 + DenFir88 + h95^2 \end{pmatrix} \tag{22}$$

Table 3 displays the beta coefficients, standard errors, t-statistics, and p-values after the application of the temporal-like BFH according to the formula 18. We noted once again, that after the application of the clustering and outlier detection procedures, there remained a total of $D = 20$ domains, a number that is considered acceptable for FH modeling. We underline that for FH models of higher complexity, a small number of domains may lead to convergence issues [63].

In forestry, estimates that do not exceed a relative standard error (RSE) of the mean or CV threshold of 10–15% are typically considered reliable [64]. Both the temporal-like BFH and UFH models performed exceptionally well, with the BFH producing 20 domains and UFH producing 19 domains with a CV of less than 5% (Table 4). These results indicate improvement over the direct estimates, which produced five domains with a CV higher than 15%, while other 3 domains displayed a CV greater than 10%. The boxplots of Fig. 5, left panel, present a comparative distribution of the error results for the direct estimates, the univariate, and finally, the temporal-like bivariate FH approach. Figure 5, right panel, showcases the confidence intervals for both UFH and BFH. The confidence intervals are illustrated with black solid and red dotted lines, respectively. The estimates of the three SAE approaches are markedly different. Figure 6 showcases maps depicting the spatial distribution of the BFH estimates for the GSV of 2018 (left panel) and CVs (right panel), correspondingly.

In summary, the results showed substantial variance reduction, with gains of 96.58% for BFH and 96.05% for UFH. When comparing the two EBLUP estimators, the temporal-like BFH model demonstrated 13.52% greater efficiency in variance reduction than the standard UFH approach.

Table 3 Beta coefficients, standard errors, *t*-statistics, and *p*-values for the utilized UFH and MFH models

	Beta UFH	Beta BFH	Std. error UFH	Std. error BFH	*t*-statistics UFH	*t*-statistics BFH	*p*-value UFH	*p*-value BFH
(intercept)	−224.710	−199.170	248.580	214.040	−0.904	−0.931	3.66e−01	3.52e−01
hmean	−451.460	−423.710	99.354	88.632	−4.544	−4.781	5.52e−06	1.75e−06
h95	96.323	91.399	30.888	26.560	3.119	3.441	1.81e−03	5.79e−04
h50	338.580	321.010	60.256	55.346	5.619	5.800	1.92e−08	6.63e−09
log(hmode)	−333.320	−324.760	64.041	61.686	−5.205	−5.264	1.94e−07	1.40e−07
log(hcv)	−773.640	−696.050	264.450	229.500	−2.926	−3.033	3.44e−03	2.42e−03
(intercept)	−561.030	−593.670	165.410	110.320	−3.392	−5.381	6.95e−04	7.40e−08
hmean	65.751	65.309	17.316	11.895	3.797	5.490	1.46e−04	4.01e−08
hsd	97.855	103.520	25.309	17.919	3.866	5.777	1.10e−04	7.61e−09
Den97	0.312	0.258	0.113	0.099	2.769	2.594	5.62e−03	9.48e−03
DenFir88	−0.421	−0.388	0.136	0.123	−3.085	−3.155	2.04e−03	1.61e−03
$h95^2$	−1.161	−1.124	0.388	0.254	−2.991	−4.435	2.78e−03	9.22e−06

Table 4 CV% distribution in classes, regarding direct, UFH, and temporal-like BFH estimates

CV% range	Direct (SRS)	UFH	BFH
0–5.0	1	19	20
5.1–10.0	11	1	0
10.1–15.0	3	0	0
>15.0	5	0	0

Ensuring the fulfillment of model assumptions holds paramount importance in the process of model construction. According to Fig. 7, the residuals generated by the implementation of the BFH model for the GSV of 2018 follow the assumptions of normality and homoscedasticity. The model's residuals seem to take place inside the confidence interval of the displayed qqplot (left panel) [65], while the standardized residuals are distributed without showing significant trends to the GSV 2018 predictions derived from the application of the temporal-like BFH.

As per its design, model-based SAE inherently introduces a degree of bias, which serves the purpose of diminishing estimator variance; a common outcome of shrinkage toward the mean [66]. Nonetheless, excessive bias can lead to erroneous estimations. The black regression line depicted in the bias diagnostic scatter plot (Fig. 8) portrays the connection between direct estimates and temporal-like BFH estimates. Line alignment with the red diagonal bisector ($y=x$) indicates unbiased BFH estimates [62]. This tendency suggests that the BFH estimations exhibit consistency with the direct estimates.

The additional bias validation, employed for calibration diagnostics by comparing model-based estimates and direct estimates (Eq. 22), yielded the following results. For the total population of 193 samples, the average relative bias (ARB%) was 0.022%, almost zero. However, for the larger six domains, each with an average of 32

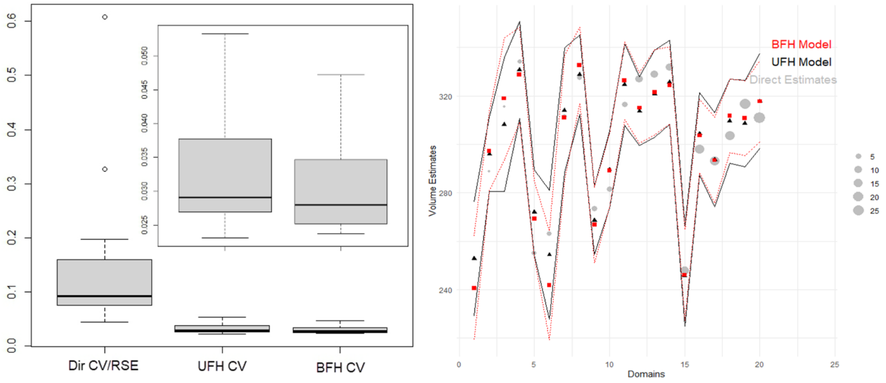

Fig. 5 On the left panel: distribution of coefficient of variation (CV) produced by the direct estimates, univariate, and temporal-like bivariate Fay-Herriot model. Inside the square on the upper right part of the figure, we display an enlargement of the boxplots corresponding to the CV of the UFH and the temporal-like BFH. On the right panel: direct sampling-based domain estimates of volume means for 2018 (gray points), univariate (black points), and bivariate Fay-Herriot (red points) estimates. Confidence intervals for the model-based estimates are illustrated with solid (univariate) and dotted (bivariate) lines

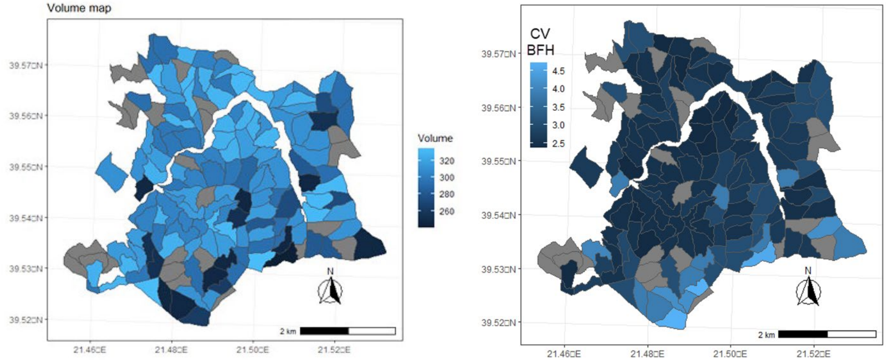

Fig. 6 On the left panel: spatial distribution for the GSV of 2018 estimation across forest management units (FMUs) based on the BFH model. Brown color indicates excluded FMUs outlier FMUs mostly with one sample plot. On the right panel: coefficient of variation for each forest management unit based on the temporal-like BFH estimates for the GSV of 2018. The brown color indicates excluded outlier FMUs mostly with one sample plot

sample plots, the ARB% values were as follows: 0.126%, 1.398%, 0.725%, – 1.293%, 0.175%, and 0.169%. These values indicate that the EBLUP model estimates are nearly unbiased compared to the assumed unbiased direct estimates.

Finally, the simulation, similar to leave-one-out cross-validation, indicated that 90% of the models (18 out of 20) exhibited comparable MSEs and non-zero model variances. This indicates model stability in the absence of fitted data. In contrast, the remaining 10% resulted in zero random area effects due to convergence or computational issues, causing zero model variances. Additionally, the estimates generated from these two models were biased and could potentially lead to misleading conclusions. Specifically, domains 2 and 12 (depicted in Fig. 9, left panel) had the largest sample sizes, constituting 12.96% (25/193) and 13.99% (27/193) of the total number of plots, respectively. The exclusion of one of these domains led to zero random area effects and smaller MSEs (Fig. 9, right panel).

However, this phenomenon is misleading, as the model's assumption of unbiased direct estimates is violated due to non-normally distributed random area effects.

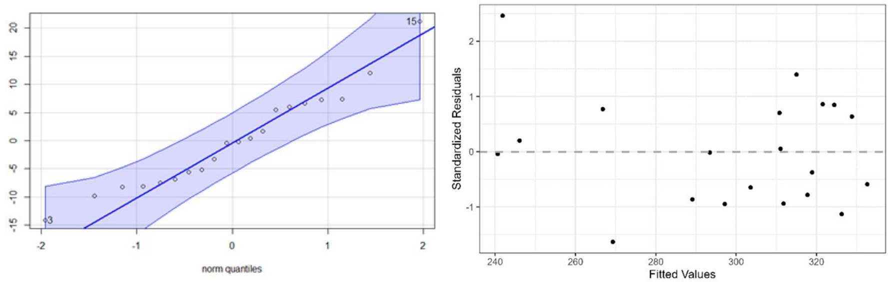

Fig. 7 Normality (left panel) and homoscedasticity (right panel) plots for the residuals generated from the BFH after the estimation for the GSV of 2018

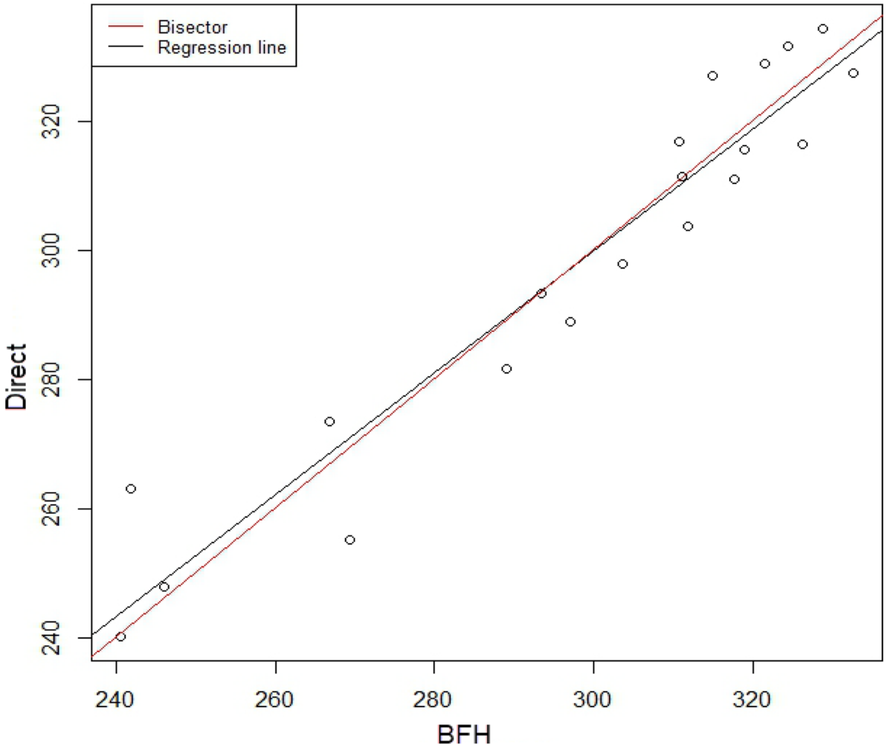

Fig. 8 Bias diagnostic plot. Bivariate Fay-Herriot (BFH) estimates (x-axis) versus direct estimates (y-axis) for the growing stock volume of 2018. The red line represents the $y=x$ bisector, and the black line represents the regression line

This highlights the significant influence of domains with large sample sizes. To address these issues, testing for multicollinearity and adjusting the covariates for increased informativeness or reducing their number could be considered. Moreover,

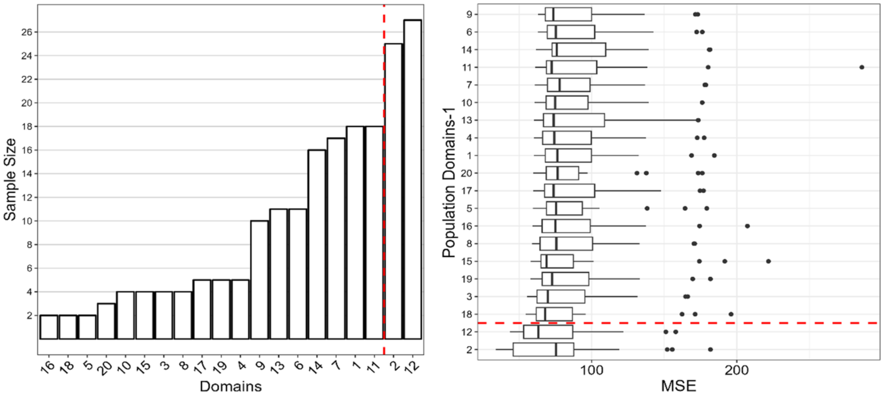

Fig. 9 Sample sizes for each domain (left panel) and distribution of MSEs (right panel) across twenty simulations, each excluding one domain (Domains-1)

such adjustments may contribute to the development of a more robust model for domain changes, albeit with an associated increase in MSEs. If challenges persist, alternative estimation methods such as hierarchical Bayes or empirical Bayes could be explored as substitutes for the EBLUP employed in this study. Another option involves using an adjusted form of maximum likelihood for the MFH to overcome issues related to zero model variance [67, 68].

4 Discussion

This study introduces an innovative technique for conducting multivariate small area estimations in the context of FIs. Our principal goal is to achieve precise estimations of the GSV for both the years 2008 and particularly 2018. To achieve this, we employed remote sensing data and census auxiliary data. The proposed statistical estimation methodology leverages a temporal-like bivariate Fay-Herriot model, effectively integrating information gleaned from both auxiliary and response variables, in this case, the GSV of 2008. Moreover, our model considers the sampling errors inherent in real data, as well as those associated with the underlying regression model. To enhance the linear relationship between the response and auxiliary variables, we introduce a new preprocessing pipeline, shown to significantly enhance the robustness of the domain estimates.

Several studies have emphasized the benefits of utilizing Fay-Herriot modeling to generate small area estimates in FIs [18, 19, 49, 69–71]. In our research, we expanded upon these insights by introducing a novel approach that arguably incorporates more sophisticated and suitable statistical methodologies. The introduced MFH model leverages the interconnected relationships among various elements: the target variables, direct survey evaluations, auxiliary variables, latent variability through random effects, and sampling errors that address the variability stemming from the sampling process itself. By concurrently considering these diverse components, we enhance the accuracy of parameter estimates and elevate the overall reliability of the estimation process. The multivariate model inherently exploits the temporal correlation between the two-time points, using one of the two response variables as auxiliary information. This feature is necessary when the response variable demonstrates strong auto-correlation. If past response variables offer sufficient predictive power for future instances, the need for external auxiliary variables may be reduced, thereby simplifying the modeling process [72]. This is especially pertinent when suitable auxiliary data are limited or when issues regarding the quality or relevance of available auxiliary data arise. We emphasize that this methodology is particularly well-suited for repeated surveys or MFIs, such as those conducted on a decadal basis, making it an ideal strategy for consistent, long-term SAE.

As previously highlighted, a unique contribution of this paper is the newly introduced preprocessing pipeline. The cluster analysis applied to FMUs significantly improves the correlation between auxiliary and response variables, resulting in the creation of a robust statistical model. Simultaneously, the outlier detection step identifies and removes extreme domain values, further enhancing the model's linear relationships. Another key step for enhancing both reliability and performance is variance smoothing. Although the FH model assumes known sampling variances, this assumption often does not hold in real-world scenarios. In the context of SAE, it is common to encounter substantial variability due to small sample sizes for each small area. This variability can lead to significant discrepancies between the estimated covariances of small areas, potentially causing computational issues related to the ill-posed inversion of the covariance matrix $\boldsymbol{V}$. Consequently, the variance smoothing step effectively addresses this challenge, facilitating the model's convergence process.

The comparison between the MFH and UFH methodologies reveals several notable advantages associated with the former approach. Our findings illustrate that the MFH technique effectively reduces both the MSE and CV, thereby enhancing the model's credibility. An intriguing aspect of our investigation lies in the strong correlation observed among the target variables. This attribute accentuates the improved performance of the MFH approach when compared to UFH, particularly in scenarios where sampling errors are small [34]. In cases where this correlation is less pronounced, both UFH and MFH models could potentially yield similar results [73]. Furthermore, the multivariate approach amalgamates insights from auxiliary variables that exhibit linear correlations with data derived from the two response variables. This amalgamation contributes to increased estimation accuracy, manifesting in reduced bias and enhanced precision, as in the GSV estimates for the two-time points. Additionally, the inclusion of information from correlated variables mitigates the influence of outliers in one variable on the overall estimation process, fostering increased stability and reliability.

In this study, we recognize the critical importance of accurately identifying and appropriately treating outliers to ensure optimal model performance. We have identified three types of outliers that may occur: (I.) in correlation analysis, (II.) in covariance matrices, and (III.) in the residuals of random area effects. To mitigate the adverse impact of outliers, we propose two primary strategies. The first involves the use of robust estimators [74]. In this scenario, estimates are available for all small areas.

The second strategy entails identifying and removing outliers. If the outliers are domains rather than individual sampling units, these domains are considered unsampled. Outlier domains, particularly those with only one unit or, in extreme cases, two units with problematic covariances, are either excluded or estimated using the synthetic component of the FH model, utilizing known calculated regression parameters. Recent advancements in modeling have enhanced the efficiency of the synthetic component in SAE. For example, some models provide estimates of random area effects for unsampled or outlier-removed areas based on clustering information from similar small areas [75, 76]. In cases of missing data, a newly proposed empirical best predictor (EBP) offers estimates for unsampled domains [77].

A third option may emerge from the estimation of zero variance of the random area effects. In such cases, an adjusted form of maximum likelihood for the multivariate Fay-Herriot model can be employed to circumvent zero model variance [67, 68]. However, it is imperative to validate the model assumptions before proceeding with this approach.

It is crucial to underscore the beneficial impact of sampling variance smoothing techniques on the estimation of random area effects variance in our study. In the absence of smoothing, extreme covariance values pose computational and convergence challenges, highlighting the importance of this preprocessing step. While variance smoothing can often obviate the need for other strategies, it is not a universal solution. Specifically, it may not be sufficient in cases with minimal data, such as those involving only two sample plots for each small area.

The application of univariate SAE techniques, particularly the FH approach, necessitates a robust collection of auxiliary variables [29] with strong linear correlations to the examined variable. This prerequisite is seldom met in heterogeneous forests with few sample plots per small area that are unable to describe the desired response variable. In contrast to the regression models that have been extensively employed in literature, the FH model considers the survey's sampling design. Furthermore, many of the abovementioned regression analyses are accompanied by restrictive assumptions regarding the covariance structure, while the FH models provide a more flexible alternative as the variances of the response variables are estimated through formula (1).

The implementation of the multivariate Fay-Herriot model incorporates additional response variables that could provide useful insights, enabling simultaneous estimates for density, height, or area features of interest. However, the number of response variables should be determined with great care, as the cardinality of variance components increases significantly. More specifically, a MFH model containing K response variables requires the estimation of $K + \binom{K}{2}$ components. Hence, a relatively small number of domains D might lead to convergence issues. Based on the high correlation values that are generated from the preprocessing pipeline (even greater than 0.6), we believe that the implementation of more complex FH models will provide interesting conclusions regarding sustainable forest management.

The presented methodology can simultaneously explore indicators associated with official statistics [34, 73, 78], epidemiology [79, 80], and poverty [1, 3, 81]. This methodology not only improves the estimation efficiency but also nurtures a comprehensive understanding across a range of scientific fields. It is important to underline that the proposed preprocessing and SAE approach is not confined to our specific dataset. Rather, it can readily be adapted for any analysis involving SAE with area-level models.

5 Conclusions

This study presents a new methodology regarding SAE in forest inventories applied to a dataset that combined past census and remote sensing data. It aims to present a statistical methodology for the precise estimation of GSV in the Pertouli University Forest for the year 2008 and, more importantly, for the year 2018. Before the implementation of the proposed temporal-like bivariate Fay-Herriot model, aiming at the production of reliable area estimations, a novel preprocessing pipeline is displayed including clustering, outlier detection, variable selection through correlation analysis, and variance smoothing steps. This preprocessing setup culminates in a significant enhancement of the overall linear dependencies between auxiliary and response variables, resulting in more robust GSV estimations.

In summary, the BFH model adeptly combines auxiliary data, encompassing historical inventory and remote sensing data, to jointly predict the target variables for both the 2008 and 2018 inventories. Outperforming direct sampling-based estimates, the MFH model showcases improved effectiveness when compared to the UFH model-based technique. By exploiting auxiliary information and correlations between variables, BFH offers a more reliable and accurate assessment process, improving the decision-making in forestry. It is important to note that the presented methodology—pipeline preprocessing steps and the temporal-like BFH/MFH model—is not restricted to our dataset but can be easily employed for any SAE attempt, leading to the sustainable management of any forest ecosystem. In the future, more complex MFH models that simultaneously take into account different forest characteristics such as height, tree density, woody volume, and basal area will be considered, with the aim of better and simultaneous estimates.

Appendix

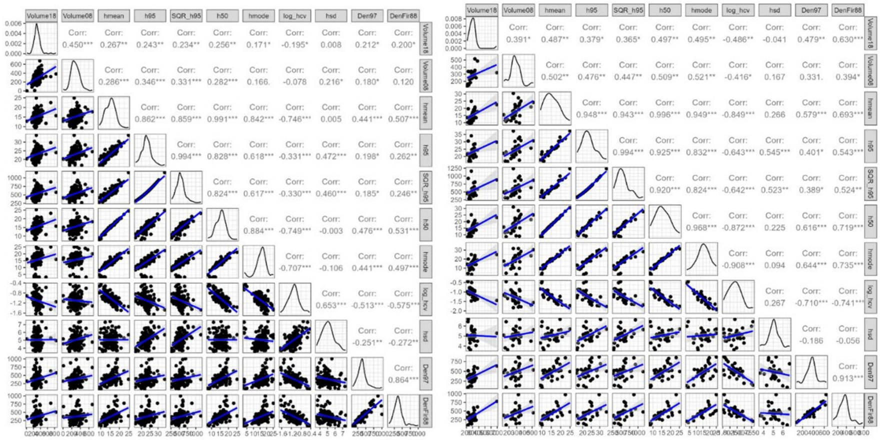

Fig. 10 Correlations diagram at FMU level (left panel) and cluster level (post-clustering) with outliers (right panel)

Acknowledgements The WorldView stereo imagery was obtained via the End User License Agreement (EULA) for the U.S. Federal Civil Government (Title 5 U.S. Code).

Author Contribution A.G. and G.S. conceptualized the study; methodology, A.G. and V.P.; software, A.G.; validation, A.G. and V.P.; formal analysis, A.G. and V.P.; investigation, A.G.; resources, A.G., D.G., and

G.S.; data curation, A.G.; writing—original draft preparation, A.G. and V.P.; writing—review and editing, A.G., V.P., D.G., and G.S; visualization, A.G.; supervision, G.S.; project administration, A.G.; funding acquisition, A.G. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by HEAL-Link Greece. This study is part of A.G.'s doctoral thesis that has been financially supported by the Hellenic Scholarship Foundation (IKY) and the European Social Fund – ESF) through the operational programme "Human Resources Development, Education and Lifelong Learning" in the context of the Act "Enhancing Human Resources Research Potential by undertaking a Doctoral Research Sub-action 2: IKY Scholar-ship Programme for PhD candidates in the Greek Universities."

Data Availability Sample survey data and digital maps used in the study are available upon request by the University Forest Administration and Management Fund at Aristotle University of Thessaloniki (https://www.auth.gr/en/university_unit/tameio-uniforest-en/, accessed on 22 December 2023). The Digital Surface Model extracted from the WorldView imagery is available from the authors upon request.

Declarations

Competing Interests The authors declare no competing interests.

References

1. Rao JN, Molina I (2015) Small area estimation. John Wiley & Sons Inc
2. Battese GE, Harter RM, Fuller WA (1988) An Error-components model for prediction of county crop areas using survey and satellite data. J Am Stat Assoc 83(401):28–36
3. Pratesi M (2016) Analysis of poverty data by small area estimation
4. Georgakis A (2019) Small area estimation in forest inventories. Seventh International Conference On Environmental Management, Engineering, Planning And Economics (CEMEPE 2019) And SECOTOX Conference. Mykonos island, Greece
5. Guldin RW (2021) "A systematic review of small domain estimation research in forestry during the twenty-first century from outside the United States." 4(96)
6. Dettmann GT, Radtke PJ, Coulston JW, Green PC, Wilson BT, Moisen GG (2022) "Review and synthesis of estimation strategies to meet small area needs in forest inventory." 5
7. Diamantopoulou MJ, Georgakis A (2023) Exploration of big-BAF sampling potential for volume estimation in Abies borisii-regis Matff. forest stands. Operations Research Forum 4(4):71
8. Diamantopoulou MJ, Georgakis A (2023b) Assessing reliable wood volume estimation of forest stand, through the application of the big-BAF sampling methodology. Tenth International Conference on Environmental Management, Engineering, Planning and Economics (CEMEPE 2023) and SECOTOX Conference. Skiathos Island, Greece
9. Georgakis A, Stamatellos G (2019) Two Contemporary and efficient two-stage sampling methods for estimating the volume of forest stands: a brief overview and unified mathematical description. Open Journal of Forestry 09(03):13
10. Iles K (2012) Some current subsampling techniques in forestry. Mathematical and Computational Forestry & Natural Resource Sciences 4(2):77

11. Breidenbach J, Astrup R (2012) Small area estimation of forest attributes in the Norwegian National Forest Inventory. Eur J Forest Res 131(4):1255–1267
12. Goerndt ME, Monleon VJ, Temesgen H (2013) Small-area estimation of county-level forest attributes using ground data and remote sensed auxiliary information. Forest Science 59(5):536–548
13. Mauro F, Monleon V, Temesgen H (2015) Using small area estimation and Lidar-derived variables for multivariate prediction of forest attributes. Forest Inventory and Analysis (FIA) symposium 2015, Portland, Oregon
14. Breidenbach J, Næsset E, Lien V, Gobakken T, Solberg S (2010) Prediction of species specific forest inventory attributes using a nonparametric semi-individual tree crown approach based on fused airborne laser scanning and multispectral data. Remote Sens Environ 114(4):911–924
15. Corona P, Fattorini L (2008) “Area-based lidar-assisted estimation of forest standing volume.” Can J Forest Res 38:2911+
16. White JC, Wulder MA, Varhola A, Vastaranta M, Coops NC, Cook BD, Pitt D, Woods M (2013) A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. For Chron 89(06):722–723
17. Fay RE, Herriot RA (1979) Estimates of income for small places: an application of James-Stein procedures to census data. J Am Stat Assoc 74(366):269–277
18. Goerndt ME, Monleon VJ, Temesgen H (2011) A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. Can J For Res 41(6):1189–1201
19. Magnussen S, Mauro F, Breidenbach J, Lanz A, Kändler G (2017) Area-level analysis of forest inventory variables. Eur J Forest Res 136(5):839–855
20. Breidenbach J, Magnussen S, Rahlf J, Astrup R (2018) Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. Remote Sens Environ 212:199–211
21. Mauro F, Monleon VJ, Temesgen H, Ford KR (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):14
22. Frank BM (2020) Aerial laser scanning for forest inventories: estimation and uncertainty at multiple scales. Oregon State University, PhD diss
23. Magnussen S, Mandallaz D, Breidenbach J, Lanz A, Ginzler C (2014) National forest inventories in the service of small area estimation of stem volume. Can J For Res 44(9):1079–1090
24. Georgakis A (2021) Further improvements of growing stock volume estimations at stratum-level with the application of Fay-Herriot model. 33rd PanHellenic statistics conference. Statistics in the Economy and Administration, Larissa, Greece, Greek Statistical Institute and the Departments of Business Administration and of Economics, University of Thessaly
25. Georgakis A, Gatziolis D, Stamatellos G (2023) “A primer on clustering of forest management units for reliable design-based direct estimates and model-based small area estimation.” Forests 14. https://doi.org/10.3390/f14101994
26. Fay RE (1987) “Application of multivariate regression to small domain estimation.” Small area statistics: 91–102
27. Ghosh M, Datta GS, Fay RE (1991) Hierarchical and empirical multivariate Bayes analysis in small area estimation. Proc 7th Annu Res Conf Bur Cens 63–79
28. Ghosh M, Nangia N, Kim DH (1996) Estimation of median income of four-person families: a Bayesian time series approach. J Am Stat Assoc 91(436):1423–1431
29. Benavent R, Morales D (2016) Multivariate Fay-Herriot models for small area estimation. Comput Stat Data Anal 94:372–390
30. Fuller W, Harter R (1987) “The multivariate components of variance model for small area estimation.” Small Area Stat
31. Benavent R, Morales D (2021) Small area estimation under a temporal bivariate area-level linear mixed model with independent time effects. Stat Methods Appl 30(1):195–222
32. Sun H, Berg E, Zhu Z (2022) Bivariate small-area estimation for binary and gaussian variables based on a conditionally specified model. Biometrics 78(4):1555–1565
33. Sun H, Berg E, Zhu Z (2023) “Multivariate small-area estimation for mixed-type response variables with item nonresponse.” J Surv Stat Methodol smad018

34. Franco C, Bell WR (2022) Using American Community Survey Data to improve estimates from smaller U.S. surveys through bivariate small area estimation models. Journal of Survey Statistics and Methodology 10(1):225–247
35. van den Brakel JA, Boonstra H-J (2021) Estimation of domain discontinuities using Hierarchical Bayesian Fay-Herriot models. Surv Methodol 47(1):151–190
36. Marhuenda Y, Molina I, Morales D (2013) Small area estimation with spatio-temporal Fay-Herriot models. Comput Stat Data Anal 58:308–325
37. Ngaruye I (2017) Contributions to small area estimation : using random effects growth curve model doctoral thesis, comprehensive summary, Linköping University Electronic Press
38. Ngaruye I, Nzabanita J, Rosen DVd, Singull M (2017) "Small area estimation under a multivariate linear model for repeated measures data." Commun Stat - Theory Method 46(21):10835–10850
39. Innocent N, Dietrich VR, Martin S (2016) Crop yield estimation at district level for agricultural seasons 2014 in Rwanda. African Journal of Applied Statistics 3(1):69–90
40. Georgakis A, Stamatellos G (2020) Sampling design contribution to small area estimation procedure in forest inventories. Modern Concepts & Developments in Agronomy 7(1):694–697
41. UFAMF (2018) Pertouli University Forest Management Plan 2019–2028, University Forest Administration and Management Fund (UFAMF)
42. Saligkaras D, Papageorgiou VE (2023) Seeking the truth beyond the data. AIP Publishing, An unsupervised machine learning approach, p 2812
43. Saligkaras D, Papageorgiou VE (2022) “On the detection of patterns in electricity prices across European countries: an unsupervised machine learning approach.” AIMS Energy 10(6)
44. Westfall JA, Patterson PL, Coulston JW (2011) Post-stratified estimation: within-strata and total sample size recommendations. Can J For Res 41(5):1130–1139
45. Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat 3(1):1–27
46. Corral GR (2020) Investigating selection criteria of constrained cluster analysis: applications in forestry. Statistical Methods and Applications in Forestry and Environmental Sciences. G. Chandra, R. Nautiyal and H. Chandra. Singapore, Springer Singapore 161–180
47. Maronna RA, Yohai VJ (1995) The behavior of the Stahel-Donoho robust multivariate estimator. J Am Stat Assoc 90(429):330–341
48. Wolter KM (2007) Generalized variance functions. Introduction to Variance Estimation. K. M. Wolter. New York, NY, Springer New York 272–297
49. Ver Planck NR, Finley AO, Kershaw JA, Weiskittel AR, Kress MC (2018) Hierarchical Bayesian models for small area estimation of forest variables using LiDAR. Remote Sens Environ 204:287–295
50. Särndal CE, Swensson B, Wretman JH (1992) Model assisted survey sampling. Springer-Verlag
51. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc: Ser B (Methodol) 58(1):267–288
52. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. 40th International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada
53. Breiman L (2001) Random Forests. Mach Learn 45(1):5–32
54. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1):389–422
55. King JR, Jackson DA (1999) Variable selection in large environmental data sets using principal components analysis. Environmetrics 10(1):67–77
56. Jennifer AH, David M, Adrian ER, Chris TV (1999) Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. Stat Sci 14(4):382–417
57. O’brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quant 41(5):673–690
58. Narendra, Fukunaga (1977) “A branch and bound algorithm for feature subset selection.” IEEE Transa Comput C-26(9):917–922
59. Ubaidillah A, Aziz SD (2021) saeBest: selecting auxiliary variables in small area estimation (SAE) model
60. Permatasari N, Ubaidillah A (2022) msae: an R package of multivariate Fay-Herriot Models for small area estimation. The R Journal 13:28
61. Srebotnjak T, Mokdad AH, Murray CJL (2010) A novel framework for validating and applying standardized small area measurement strategies. Popul Health Metrics 8(1):26

62. Brown G, Chambers R, Heady P, Heasman D (2001) Evaluation of small area estimation methods—an application to unemployment estimates from the UK LFS. Proc Stat Can Symp
63. Esteban MD, Lombardía MJ, López-Vizcaíno E, Morales D, Pérez A (2020) Small area estimation of proportions under area-level compositional mixed models. TEST 29(3):793–818
64. Mauro F, Molina I, García-Abril A, Valbuena R, Ayuga-Téllez E (2016) Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels. Environmetrics 27(4):225–238
65. Almeida A, Loy A, Hofmann H (2018) ggplot2 compatible quantile-quantile plots in R. R J 10(2):248
66. Chandra H, Salvati N, Chambers R (2017) Small area prediction of counts under a non-stationary spatial model. Spatial Statistics 20:30–56
67. Angkunsit A, Suntornchost J (2020) "Bivariate Fay-Herriot models with application to Thai socio-economic data." Naresuan Univ J: Sci Technol (NUJST) 29:(1)
68. Angkunsit A, Suntornchost J (2022) Adjusted maximum likelihood method for multivariate Fay-Herriot model. Journal of Statistical Planning and Inference 219:231–249
69. Ver Planck NR, Finley AO, Huff ES (2017) Hierarchical Bayesian models for small area estimation of county-level private forest landowner population. Can J For Res 47(12):1577–1589
70. Green PC, Burkhart HE, Coulston JW, Radtke PJ (2019) "A novel application of small area estimation in loblolly pine forest inventory." Forestry: Int J For Res
71. Temesgen H, Mauro F, Hudak AT, Frank B, Monleon V, Fekety P, Palmer M, Bryant T (2021) Using Fay-Herriot models and variable radius plot data to develop a stand-level inventory and update a prior inventory in the Western Cascades, OR, United States. Frontiers in Forests and Global Change 4(157):17
72. Franco C, Maitra P (2023) Combining surveys in small area estimation using area-level models. WIREs Comput Stat 15(6):18
73. Guha S, Chandra H (2022) Measuring and mapping micro level earning inequality towards addressing the sustainable development goals – a multivariate small area modelling approach. Journal of Official Statistics 38(3):823–845
74. Jiang J, Rao JS (2020) "Robust small area estimation: an overview." 7(1):337–360
75. Torkashvand E, Jozani MJ, Torabi M (2017) Clustering in small area estimation with area level linear mixed models. J R Stat Soc A Stat Soc 180(4):1253–1279
76. Desiyanti A, Ginanjar I, 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 11. https://doi.org/10.3390/math11010135
77. Burgard JP, Morales D, Wölwer A-L (2022) Small area estimation of socioeconomic indicators for sampled and unsampled domains. AStA Advances in Statistical Analysis 106(2):287–314
78. Papageorgiou V, Tsaklidis G (2021) "Modeling of premature mortality rates from chronic diseases in Europe, investigation of correlations, clustering and granger causality." Commun. Math Biol Neurosci 67
79. Papageorgiou VE, Tsaklidis G (2023) "An improved epidemiological-unscented Kalman filter (hybrid SEIHCRDV-UKF) model for the prediction of COVID-19. Application on real-time data." Chaos Solit Fractals 166:112914
80. Papageorgiou VE, Tsaklidis G (2023) "A stochastic SIRD model with imperfect immunity for the evaluation of epidemics." Appl Math Model
81. Yilema SA, Shiferaw YA, Zewotir T, Muluneh EK (2022) Multivariate small area estimation of undernutrition for children under five using official statistics. Stat J IAOS 38:625–636

Authors and Affiliations

Aristeidis Georgakis[1] · **Vasileios E. Papageorgiou**[2] · **Demetrios Gatziolis**[3] · **Georgios Stamatellos**[1]

✉ Aristeidis Georgakis
arisgeorg@for.auth.gr

Vasileios E. Papageorgiou
vpapageor@math.auth.gr

Demetrios Gatziolis
demetrios.gatziolis@usda.gov

Georgios Stamatellos
stamatel@for.auth.gr

[1] School of Forestry and Natural Environment, Aristotle University of Thessaloniki, Thessaloniki, Greece

[2] Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki, Greece

[3] USDA Forest Service, Pacific Northwest Research Station, Portland, OR, USA